iX SPECIAL 2022

Green IT

Ressourcen schonen und Kosten sparen

Nachhaltig programmieren
Designfehler vermeiden
Komplexität reduzieren
Ressourcen effektiv einsetzen

Strom und Kühlung
Zukunftsfähiges Energiemanagement
Natürliche Kältemittel nutzen
Mit Flüssigkeit bis zum Server
Systemnahe Wärmeübertragung für luftgekühlte Server

Der Weg zum klimaneutralen RZ
Den CO_2-Fußabdruck von RZs beurteilen
Ungenutzte Rechenressourcen heben
Ökozertifikate für Rechenzentren
Sparsames Design: Open Compute Project

CO_2-Fußabdruck der Cloud
Einzelne IT-Dienste messen • Die eigenen Emissionen visualisieren

14,90 €
Österreich 16,40 €
Schweiz 27.90 CHF
Luxemburg 17,10 €
www.ix.de

NEUE WEGE STATT AUS-GETRETENER PFADE.

Cordaware **bestzero**®: Mit Sicherheit einfach besser.

Remote Zugriff auf lokale Ressourcen **schnell** und **einfach** bereitstellen.

Keine offenen eingehenden Ports erforderlich => **Zero-Firewall-Config**.

Cordaware
bestzero®

Verfügbar für Windows, macOS, Linux und Android

Cordaware GmbH Informationslogistik +++ Fon +49 8441 8593200 +++ info@cordaware.com +++ www.cordaware.com

Keine Zeit zum Ausruhen

Über den Kampf gegen die Klimakatastrophe wird in der IT viel geredet und geschrieben – nur getan wird auch hier viel zu wenig. Doch warum ist das so? Eine der Antworten, die gerade in der Informationstechnik erstaunen mag: Oft fehlt das notwendige Wissen. Das Wissen darüber, wie sich Software nachhaltiger und weniger ressourcenverschwenderisch gestalten lässt. Das Wissen darüber, wie sich die hinlänglich bekannten Energieverschwender wie Netzteile und Lüfter ersetzen oder reduzieren lassen. Das Wissen darüber, wie sich die RZ-Klimatisierung vom Klimakiller zum Wärmespender verwandeln lässt. Oder das Wissen, wie sich kritische oder umweltschädliche Stoffe ersetzen lassen.

Zu den oft fehlenden Handlungsstrategien gesellt sich die zunehmende Komplexität. Denn wer glaubt, technischer Fortschritt gehe automatisch mit dem Schließen von Wissenslücken einher, irrt: Lieferketten, wechselseitige Abhängigkeiten von Teilsystemen und Wirkungsweisen komplexer Softwaresysteme werden immer unüberschaubarer, Beschaffungsrisiken und die Folgen von Ausfällen immer unkalkulierbarer. Das führt vielerorts zu Überforderungen und Ratlosigkeit, die schnell in Handlungsunfähigkeit mündet.

Eine weitere Antwort auf das Warum hat gleich mehrere Teile. Der erste lautet: In der IT dreht sich alles um genau zwei Dinge: Kosten und Gewährleistungen, genannt SLAs. Was nichts anderes heißt als Geld und die Sorge, den eigenen Kopf hinhalten zu müssen. Und im Zweifelsfall sind geringere Investitionen jetzt wichtiger als spätere Einsparungen bei den Betriebs- und Folgekosten. Der zweite Teil: Den Status quo beizubehalten ist immer einfacher, als Änderungen herbeizuführen. Gewohnheiten, Bequemlichkeiten, mangelnde Courage und der fehlende Wille zur Eigeninitiative stehen Veränderungen am häufigsten im Weg – auch in der als innovativ geltenden IT-Branche.

Doch werden weder Klimakatastrophe, Umweltzerstörung noch Artensterben darauf Rücksicht nehmen und warten, bis Menschen ihre Bequemlichkeiten abgelegt haben. Handeln tut jetzt not – je länger wir abwarten, desto ungemütlicher wird die Zukunft. Deshalb will das Sonderheft nicht nur Wissen vermitteln und Antworten geben, sondern auch Anregungen zum Handeln.

Susanne Nolte

Risiken und Abhängigkeiten

Digitalisierung ist der Inbegriff des Fortschritts, doch sie schafft vor allem neue Abhängigkeiten. Den gewonnenen Komfort und die kurzfristigen Erfolge bezahlt nicht nur die IT, sondern die ganze Gesellschaft mit nachlassender Resilienz. Dabei hätte die IT ein enormes Potenzial, die Wegbereiterin für eine nachhaltigere Zukunft zu sein. Doch derzeit weisen ihr steigender Energiebedarf und ihre Energieverschwendung ebenso wie ihr Umgang mit Rohstoffen in eine andere Richtung.

ab Seite 7

Quo vadis, IT?

Überblick
Dimensionen der Nachhaltigkeit 8

Ökodaten
Der Nutzen von Umweltinformationssystemen 12

Versorgungssicherheit
Lieferketten mit und für die IT nachhaltiger und resilienter gestalten 16

Lieferabhängigkeiten
Knappe Rohstoffe als Richtungsweiser 22

Ressourcennutzung
Elektroschrott vermeiden 30

Recht
Vorschriften zu einer nachhaltigeren IT 36

Grüne Software

Kontrollverlust
Technisches Versagen in der Softwareentwicklung 40

Strategie
Organisatorische Einbettung nachhaltiger Software 48

Design
Nachhaltig programmieren mit Bedacht 54

Abschalten
Zombies im Rechenzentrum 58

Im Detail
Energieeffizienz von Software messen 62

C++
Ressourcensparend programmieren: Lernen von der Embedded-Entwicklung 67

Energieeffiziente Rechenzentren

Kennzahlen
SIEC – Die Energieverschwendung ungenutzter Rechenressourcen offenlegen 72

Messverfahren
Mit KPI4DCE die Energie- und Ressourceneffizienz von Rechenzentren beurteilen 78

Alternatives Design
Open Compute Project: Energie- und Raumeffizienz zu niedrigsten Kosten? 84

Nachhaltig programmieren

Unhaltbare Ressourcenverschwendung ist nur ein Vorwurf, den sich Softwareentwickler gefallen lassen müssen. Weit größere Risiken für die Gesellschaft bergen Designfehler und die wachsende Komplexität von Software, die ihre Schöpfer die Kontrolle darüber verlieren lässt. Eine Umkehr würden genügsame Funktionen, durchdachtes Design, höhere Qualität, gute Wartbarkeit, neue Narrative und langfristige Strategien bringen.

ab Seite 39

Energieeffizienz in Rechenzentren

Diskussionen über den Energiebedarf der IT drehen sich oft um Rechenzentren, denn dort, wo sich IT-Ressourcen konzentrieren, sind Einsparungen besonders wirksam. Potenzial ist hier reichlich vorhanden. Doch wer der Verschwendung Einhalt gebieten will, braucht Messmethoden, Maßstäbe und Standards.

ab Seite 71

Cloud-Services
Den Fußabdruck einzelner IT-Dienste messen **92**

Dashboards
Die CO_2-Emissionen der eigenen Cloud-Nutzung visualisieren **96**

Normen
Ökozertifikate für nachhaltige Rechenzentren **100**

Kommentar
Welche Bezugspunkte braucht eine Umweltkennzahl? **105**

Umweltgerechte RZ-Infrastruktur

Stromversorgung
Zukunftsfähiges Energiemanagement im Rechenzentrum **110**

Stromausfall
Alternativen zur unterbrechungsfreien Stromversorgung mit Blei-Gel-Akkus **118**

HFKW-Alternativen
Natürliche Kältemittel für die RZ-Klimatisierung **122**

Supercomputer
Direkte Warmwasserkühlung von HPC-Chips **128**

Racks und Reihen
Systemnahe Wärmeübertragung für luftgekühlte Server **134**

Abwärme
Flüssigkeitskühlungen für Server **140**

Projekte
Abregelungen und Abwärme: Ungenutzte Energie nutzen **146**

Medien

Buchmarkt
Literatur zu nachhaltiger IT **151**

Rubriken

Editorial: Keine Zeit zum Ausruhen **3**

Impressum **127**

Klimafreundlichere RZ-Infrastruktur

Einen nicht unwesentlichen Teil des CO_2-Fußabdrucks von Rechenzentren hat deren Infrastruktur zu verantworten. Dabei ist sie unerlässlich für den geregelten und zuverlässigen IT-Betrieb. Das aber darf kein Freifahrtschein für ihre klimaschädliche Wirkung bleiben. Ansätze, das zu ändern, gibt es viele.

ab Seite 109


enterJS
Die Konferenz für
Enterprise-JavaScript
22. und 23. Juni 2022 – Darmstadt
Jetzt Tickets sichern!
www.enterjs.de
+++ Workshops zu Node.js, Angular, Playwright und JavaScript am 21. und 24. Juni +++
Silbersponsor
adesso
business. people. technology.
Veranstalter
iX MAGAZIN FÜR PROFESSIONELLE INFORMATIONSTECHNIK
heise Developer
dpunkt.verlag
© Copyright by Heise Medien.

Quo vadis, IT?

Die IT ist eine energieintensive Branche, die aber nicht als solche gesehen wird. Ein Fehler, wie ein genauer Blick zeigt. Dabei stehen nicht nur Energiebedarf und -verschwendung der IT in der Kritik, sondern auch die von ihr geschaffenen Abhängigkeiten und ihr Umgang mit Rohstoffen. Sie ist aber auch eine Branche mit enormem Potenzial und Wegbereiterin für eine nachhaltigere Zukunft. Doch welche dieser beiden Seiten die Richtung vorgeben wird, hängt auch und vor allem davon ab, welchen Weg die Gesellschaft künftig einschlägt, den der fortgesetzten Zerstörung oder den der Einsicht.

Überblick – Dimensionen der Nachhaltigkeit 8

Ökodaten – Der Nutzen von Umweltinformationssystemen 12

Versorgungssicherheit – Lieferketten mit und für die IT nachhaltiger und resilienter gestalten 16

Lieferabhängigkeiten – Knappe Rohstoffe als Richtungsweiser 22

Ressourcennutzung – Elektroschrott vermeiden 30

Recht – Vorschriften zu einer nachhaltigeren IT 36

Dimensionen der Nachhaltigkeit

Durchschritten

Dr. Alexander Schatten

In einer nachhaltig organisierten Gesellschaft muss auch und besonders ihr digitales Nervensystem, die IT, nachhaltig gestaltet sein. Dabei hat Nachhaltigkeit nicht nur eine ökologische, sondern auch eine technische, eine soziale sowie eine ökonomische und administrative Dimension.

Dass sich die Informations- und Kommunikationstechnik zum digitalen Nervensystem der Gesellschaft entwickelt hat, verleiht ihr das Potenzial, sie auch sicherer, nachhaltiger und resilienter zu machen. Allein die Covid-Pandemie hätte vor zwanzig Jahren mit Sicherheit wesentlich mehr wirtschaftliche und gesellschaftliche Verwerfungen verursacht. Auch das stetige Wachstum von Städten und die damit verbundene dichtere und gleichzeitig ökologischere Lebensweise, moderne Produktionsmethoden und Umweltinformationssysteme wären ohne moderne IT nicht denkbar.

Zugleich schafft die schnell fortschreitende Digitalisierung aber wechselseitige, teilweise sogar zirkuläre Abhängigkeiten. Die Stromversorgung ist beispielsweise in komplexer Weise abhängig von der IT: Unterschiedliche technische Komponenten des Systems kommunizieren über digitale Netze, der Verbrauch wird modelliert, Systeme werden gesteuert und kontrolliert; aber auch Mitarbeiter verwenden Computer zum Abstimmen. Um ihrer Arbeit nachgehen zu können, benötigen sie Mobilitätsdienstleistungen wie Tankstellen oder den öffentlichen Verkehr, deren Logistik an IT gebunden ist. Die IT selbst ist abhängig vom Strom und von der Mobilität von Mitarbeitern und Gütern. Damit schließt sich der Kreis: Fällt eines dieser Systeme aus, sind alle anderen gefährdet. Ein Neustart nach einem größeren Ausfall ist aufgrund dieser komplexen Abhängigkeiten und der Unplanbarkeit eines solchen Vorfalls alles andere als trivial.

Ein ähnliches Bild lässt sich von fast allen essenziellen Diensten moderner Gesellschaften zeichnen, etwa der Nahrungsmittel- und Wasserversorgung, dem Finanzwesen, Gesundheitswesen und den Blaulichtorganisationen. Den Vorteilen der Digitalisierung stehen also erhebliche Risiken gegenüber. Über IT-Systeme können sich Störungen blitzschnell global ausbreiten und kritische gesellschaftliche Systeme in Gefahr bringen. Zudem verbraucht die IT-Infrastruktur selbst große Mengen an Ressourcen, während die Globalisierung von Produktion und Entsorgung mit Umweltverschmutzung und Menschenrechtsverletzungen einhergeht.

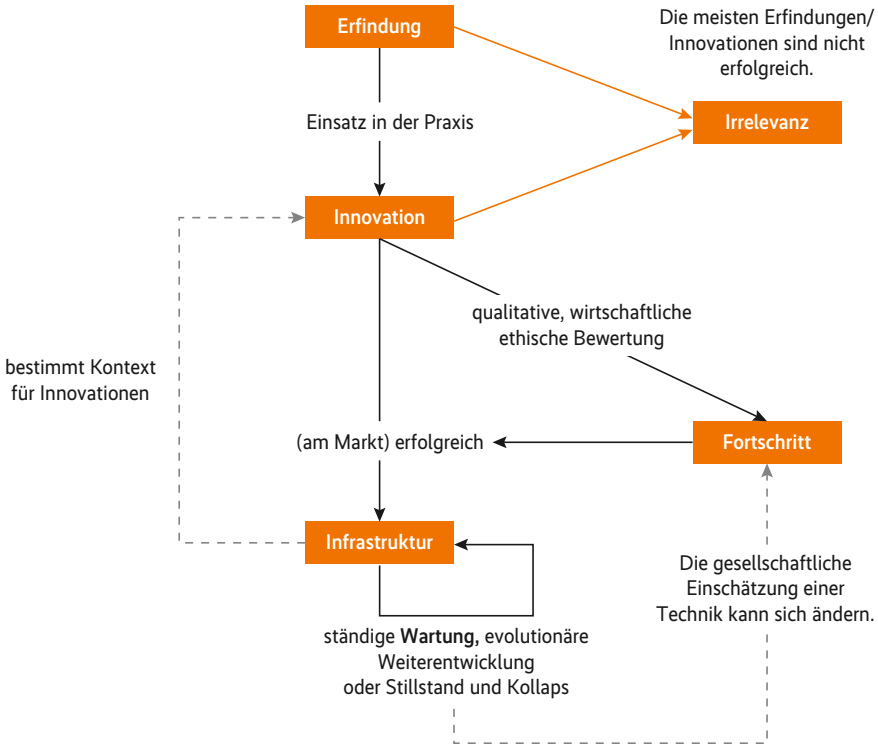
Komplexität und Wechselwirkungen

Hat die Menschheit also eine Technik mit existenziellen Risiken auf die Welt losgelassen oder technische Systeme, die Risiken

reduzieren und die Nachhaltigkeit der Gesellschaft verbessern können? Wie so häufig ist die Antwort nicht einfach, sondern hängt von vielen Faktoren ab. Dieses Sonderheft stellt sie vor.

Die IT hat sich über die letzten Jahrzehnte von einem komplizierten lokalen zu einem komplexen und global vernetzten System entwickelt. Durch Vernetzung kann die Stabilität in einem System steigen; die Zunahme an Komplexität führt aber ab einem bestimmtem Maß zu Effekten, die die Kontrolle dieser Systeme erschweren und zu unerwartetem Verhalten führen können.

Komplexität und Kompliziertheit, wenngleich umgangssprachlich oft synonym verwendet, meinen etwas völlig Unterschiedliches: Die Kompliziertheit bezieht sich auf das Verständnis eines Systems durch den Menschen, wohingegen die Komplexität eine systemische Eigenschaft ist. Sie entsteht in der dynamischen Interaktion zahlreicher Akteure, wobei die Interaktion den inneren Zustand der Akteure sowie die Verknüpfung der Akteure verändern kann. Genau in dieser Situation befindet sich die IT in ihrem Zusammenspiel mit den gesellschaftlichen, wirtschaftlichen und politischen Systemen.



Die technische Entwicklung selbst folgt einem Kreislauf, in dem Innovationen zur Infrastruktur heranwachsen, die wiederum neue Innovationen ermöglicht. Doch: Was Innovation ist, muss noch lange kein Fortschritt sein.

Die Folge ist ein Systemverhalten, das aus dem Verständnis einzelner Akteure nicht ableitbar oder gar vorhersagbar ist. Es kann sehr schwer sein, Änderungen im System vorzunehmen, die zu einem klar definierten Zustand führen. Oftmals ist schon die Beschreibung des aktuellen oder gewünschten Systemzustandes eine kaum zu handhabende Herausforderung (siehe Artikel „Ein schwerer Anfang“ ab Seite 40). Man spricht dann von Wicked Problems. Wer also die moderne Gesellschaft nachhaltiger und resilienter gestalten will, darf den Umgang mit solchen komplexen IT-Systemen – Entwicklung, Management, Wartung, Finanzierung – nicht vernachlässigen.

Die unterschiedlichen Dimensionen der Nachhaltigkeit

Die Begriffe Nachhaltigkeit und Resilienz waren bisher im Zusammenhang mit IT eher ungebräuchlich. Bedenkt man aber das immer kritischer werdende Zusammenspiel von Software respektive Technik, Gesellschaft und Wirtschaft, sollten diese Begriffe im Softwareengineering und in der IT-Infrastruktur einen festen Platz einnehmen.

Die Nachhaltigkeitsforschung unterscheidet in der Regel vier Dimensionen der Nachhaltigkeit, die in den weiteren Artikeln vertieft werden. An dieser Stelle ein kurzer Überblick:

Die **ökologische Dimension** ist den meisten am geläufigsten. Bei der IT sind direkte Effekte wie Wasser-, Energie- und Ressourcenverbrauch in der Produktion, der Stromverbrauch von Rechenzentren und Endgeräten, der Rohstoffverbrauch und das Verwerten alter Geräte und Komponenten, von Elektronikschrott, zu bedenken. Sie meint aber auch indirekte Effekte, etwa das Entwickeln von Geschäftsmodellen oder Techniken, die ohne IT nicht denkbar waren, wie Just-in-Time-Lieferketten, die immer einfachere Produktion minderwertiger Produkte oder die Ausbeutung von Ressourcen durch Analyse und Modelle etwa bei der Erdöllagerstättenverwaltung oder der Beobachtung von Fischschwärmen.

Daneben zeitigt IT aber auch positive Effekte, indem sie etwa einen besseren Umgang mit der Umwelt ermöglicht beziehungsweise Lebensweisen, die umweltbewusster sind als frühere. Umweltinformationssysteme und Transparenzmaßnahmen in Lieferketten beispielsweise verbessern das Verständnis über die Auswirkungen unseres Verhaltens auf die Umwelt. IT erlaubt ein Leben von Menschen in Großstädten, das unter ökologischen Aspekten dem Leben auf dem Land deutlich überlegen ist. Auch virtuelle Arbeitsformen können die Resilienz unserer ökonomischen Praktiken erhöhen, und dies bei gleichzeitiger Reduktion der Umweltfolgen.

Die **technische Dimension** beschäftigt sich mit der Frage, wie immer komplexer werdende Systeme unter Kontrolle gehalten werden können. Gesellschaft und Wirtschaft stellen immer kritischere Anforderungen und die zunehmende Vernetzung führt zu emergentem, kaum planbarem Verhalten. Alte technische Metaphern haben ausgedient und sollten durch neue – der Komplexität, Nachhaltigkeit und Resilienz folgende – ersetzt werden. Dies gelingt bisher kaum.

X-TRACT

- IT-Infrastruktur ist heute ein komplexes und global vernetztes System, das in enger Abhängigkeit zu anderen komplexen Systemen steht und das gesellschaftliche Geschehen mitbestimmt.
- Technischer Fortschritt findet immer nur im Kontext gesellschaftlicher Bewertung und vorhandener Infrastruktur statt.
- Wer die Gesellschaft nachhaltiger und resilienter gestalten will, muss auch die Entwicklung, Wartung, Finanzierung und das Management der IT nachhaltig gestalten.
- Nachhaltigkeit hat vier Dimensionen: eine ökologische, eine technische, eine soziale und eine ökonomisch-administrative.

Definitionen

Softwareengineering ist der Prozess zum Erstellen von Softwaresystemen unter Berücksichtigung technischer, administrativer, wirtschaftlicher und gesellschaftlicher Aspekte.

Komplexität ist eine systemische Eigenschaft, die aus der Interaktion individueller Akteure entsteht. Das Verhalten komplexer Systeme lässt sich nicht aus dem Verhalten einzelner Akteure ableiten und wird auch emergentes Verhalten genannt.

Wicked Problems wurden erstmals im Jahr 1969 beschrieben: Probleme weisen in einer komplexen und technisch vernetzten Gesellschaft eine andere Struktur auf als in früheren technischen Systemen. Solche Probleme sind schwer fassbar, oftmals ist weder der aktuelle Zustand noch der gewünschte Zielzustand klar beschreibbar und ebenso wenig eine Lösung im klassischen Sinn.

Techno-soziale Systeme entstehen aus der engen Interaktion von Technik und Gesellschaft einschließlich Wirtschaft und Politik und lassen sich nicht isoliert, also rein technisch verstehen.

Resilienz ist die Fähigkeit eines Systems, sich von Störungen in angemessener Zeit zu erholen und dabei die wesentlichen Funktionen wiederherzustellen. Es muss aber nicht notwendigerweise den ursprünglichen Zustand wiederherstellen.

Nachhaltigkeit ist ein Prinzip, das die langfristige Nutzung von Ressourcen und gesellschaftlichen Systemen unter Erhaltung der regenerativen Fähigkeiten ermöglicht. Dies schließt Wirtschaft, Umwelt, Gesellschaft und Technik mit ein.

Im Zusammenspiel mit neuen Managementmethoden ist der Herausforderung zu begegnen, dass klassische Lebenszyklusmodelle von Software kaum mehr eine Rolle spielen. Bei vielen kritischen Vorfällen wie Datenverlust oder Cyberangriffen werden daher gerne Symptom und Ursache verwechselt. Ihre Ursache liegt oftmals in der sehr geringen Qualität von Software und Infrastruktur und dem daraus folgenden Kontrollverlust (siehe Artikel „Ein schwerer Anfang“ ab Seite 40).

Die **soziale Dimension** umfasst die Interaktion von IT und Gesellschaft und mit ihr die Frage: Wo stiftet IT tatsächlich Nutzen und wo überwiegen negative Effekte? Unter dem Stichwort digitaler Humanismus werden dabei die Folgen sozialer Netze auf die Politik, die Überwachung durch staatliche und private Akteure oder die psychologischen und soziologischen Effekte von Smartphones diskutiert.

Viele dieser Fragen sind ebenso globaler Natur: Die Produktion und Entsorgung von Hardware findet häufig in Ländern statt, die nicht mit europäischen Arbeits- oder Umweltstandards operieren. Cloud-Services und Managementdienste für Smartphones oder Notebooks werden in Ländern betrieben, die nicht über europäische Datenschutzstandards verfügen, um zwei Beispiele zu nennen.

Die **ökonomische und administrative Dimension** überlappt sich mit den anderen. Technische Risiken wie Blackouts oder IT-Ausfälle können große ökonomische Auswirkungen haben. Aber auch die Produktion und die Wartung von Software selbst hat enorme wirtschaftliche Konsequenzen, wie gescheiterte Großprojekte, Softwareerneuerungen oder die erzwungene Pflege alter Systeme zeigen, deren Umstellung oder Weiterentwicklung mangels nachhaltiger Entwicklungs- und Managementpraktiken scheitert. Der Einsatz neuer Management- und Planungsprozesse, die diesen neuen Gegebenheiten Rechnung tragen, ist herausfordernd. Erste Schritte sind durch den Einsatz von Methoden wie Scrum, SAFe, OKRs und Beyond Budgeting zu erkennen.

Auch der Umgang mit Wissen im digitalisierten und global vernetzten Unternehmen, Recruiting und Weiterbildung überfordert viele Unternehmen und öffentliche Stellen. Automatisierung führt meist zu einem Verlust an operativem Wissen, das aber für die langfristige Wartung ebendieser digitalen Systeme notwendig ist. Die Weiterentwicklung komplexer IT-Landschaften erfordert Mitarbeiter mit viel Erfahrung und hoher Qualifikation. Genau diese Personen haben aber heute die Möglichkeit, international, teilweise auch ohne vor Ort zu sein, an sehr gut bezahlten und spannenden Projekten zu arbeiten.

Zudem ist die IT immer stärker aus dem Blickwinkel des Infrastrukturmanagements zu betrachten. Die Forderung nach kurzfristiger Effizienz wird häufig zum Gegenspieler nachhaltiger Resilienz – auch der Geschäftsmodelle.

Innovation oder Fortschritt?

Der Lebenszyklus von Technik folgt einer bestimmten Logik (siehe Abbildung): Aus der Erfindung – einer neuen Idee – wird Innovation und ein darauf basierendes Produkt wird auf den Markt gebracht. Dieses Produkt, diese Innovation ist, auch wenn Marketingabteilungen das anders sehen, noch kein Fortschritt. Fortschritt erfordert eine qualitative gesellschaftliche Bewertung, die sich auch über den Lebenszyklus einer Technik verändern kann – erinnert sei an die Effekte sozialer Netze.

Eine am Markt erfolgreiche Innovation wird letztlich zur Infrastruktur, die eng mit anderen technischen Systemen und der Gesellschaft interagiert und deshalb über lange Zeiträume kontinuierlich gewartet, finanziert und weiterentwickelt werden muss. Die IT-Infrastruktur zeigt dabei Verhaltensmuster, die eher biologischen Ökosystemen als klassischen technischen Systemen früherer Jahrhunderte ähneln.

Erst gut funktionierende und gewartete IT-Infrastruktur ermöglicht Innovationen, da sie, sobald sie skaliert, immer mit bestehender Infrastruktur interagieren muss. Ist die Innovation erfolgreich, wird sie letztlich selbst zur Infrastruktur – und der Kreis schließt sich.

Innovation und Infrastruktur agieren aber auf sehr unterschiedlichen Zeitachsen und unter technisch sehr unterschiedlichen Rahmenbedingungen. Dies erfordert ebenso unterschiedliche Personen, Management, Finanzierung und ein Modell, das aufzeigt, wie der Übergang von der sich schnell ändernden Innovation zur Infrastruktur gelingen kann.

Aufgrund dieser Komplexität der heutigen IT und der daraus folgenden Verhaltensmuster ist eine Planbarkeit und Vorhersage erwünschter wie unerwünschter Effekte neuer Erfindungen allerdings nur in sehr begrenztem Umfang gegeben. Nachhaltigkeit und Resilienz dieses digitalen Nervensystems unserer Gesellschaft ist daher eine Rahmenbedingung dafür, dass IT-Systeme tatsächlich Fortschritt und nicht neue existenzielle Risiken hervorbringen. (sun@ix.de)

Dr. Alexander Schatten

ist Senior Researcher bei SBA-Research, Management-Berater und Podcaster: https://podcast.zukunft-denken.eu

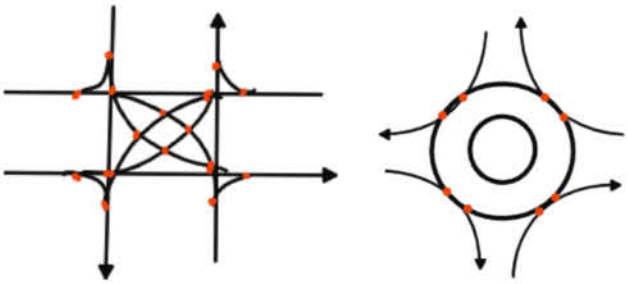
»» Cloudanwendungen in der KfW

Die KfW Bankengruppe (im Weiteren KfW) ist, gegründet 1948, eine Bank, die großen Wert auf Stabilität und Nachhaltigkeit legt. Vor diesem Hintergrund ist die Cloud als solches ein eher undurchsichtiges und mit zahlreichen Fragezeichen versehenes Konstrukt. Dennoch gibt es gute Gründe gewisse User-Workloads in der Cloud statt Onpremise zu etablieren.

Die KfW ist sich dessen bewusst und hat sich zunächst entschieden technische Leuchtturmprojekte umzusetzen. So sind in den letzten zwei Jahren Projekte in die Produktion übernommen worden, welche Machinelearning-Services nutzen, serverless Microservices oder Kubernetes.

In allen Projekten lag ein unterschiedliches Maß an Wissen im Cloudnative-Umfeld vor. Auch der klassische Anwendungsbetrieb hatte viel zu lernen. Nicht zuletzt unterliegt die KfW der Bafin und damit notwendiger und sinnvoller Regulatorik. Jedes der Leuchtturmprojekte wurde daher ein wenig anders umgesetzt. Mal mehr CI/CD, mal mehr Logging und Monitoring, mal weniger händische Config.

Um jedoch nachhaltig und skalierbar im Cloud-umfeld aktiv zu sein und der hohen Geschwindigkeitserwartung, welche an die Cloudgestellt wird, gerecht zu werden, braucht es eine andere Herangehensweise.



Ein bekanntes Bild ist die gleichberechtigte Kreuzung, an welcher bei wenig Verkehr nur selten Konflikte auftreten. Bei höherem Verkehrsaufkommen nehmen die Konflikte zu. Es ist also abzusehen, dass der Teilnehmer- und Geschwindigkeitszuwachs an einer solchen Kreuzung begrenzt ist.

An einem Kreisverkehr gibt es weniger Berührungspunkte und damit weniger Konflikte. Damit kann eine höhere Anzahl an Teilnehmern schneller ihre gewünschte Richtung erreichen.

Bestehende Prozesse und Arbeitsweisen haben ihre guten Gründe. Meist durch regulatorische Anforderungen begründet sind viele Schnittstellen entstanden, welche eine schnelle Reaktion auf äußere Einflüsse Onpremise erschweren.

In der KfW sind wir auf dem Weg, die Möglichkeiten einer Cloud-nativen Kultur durch konsequente Vermeidung von Komplexität auszuschöpfen. Wir verfolgen das Ziel mit iterativem und inkrementellem Voranschreiten auch auf zukünftige Veränderungen schneller und flexibler reagieren zu können.

Über den Autor

Daniel Kape
Productowner
Cloud Plattformen

Der Nutzen von Umweltinformationssystemen

Bestandsaufnahme

Dr. Alexander Schatten

Bewerten, vorhersagen, angemessen reagieren – dazu wollen alle Umweltinformationssysteme durch das Sammeln und Auswerten der unterschiedlichsten Ökodaten ihre Anwender befähigen.

Neben den direkten Effekten, mit denen die IT die Umwelt belastet, vor allem durch ihren Ressourcenverbrauch und den durch sie verursachten Abfall, gehen von der Informationstechnik auch zahlreiche indirekte Effekte aus, die zu einem nachhaltigeren Wirtschaften beitragen können. Die wesentliche Klasse der Techniken bilden dabei die Umweltinformationssysteme, die in den letzten Jahren eine immer wichtigere Rolle spielen, auch deshalb, weil sie sich von den klassischen Anwendungsfeldern hin zu immer stärker vernetzten und operativ wirksamen Werkzeugen entwickelt haben.

Klassisch definiert dienen Umweltinformationssysteme zur computergestützten Beobachtung der Umwelt mit dem Zweck, eine Ressourcenübernutzung zu vermeiden, das Management natürlicher Ressourcen zu vereinfachen und Informationen weiterzugeben. Sie verwenden oder stellen Datenbanken zum Austausch bereit, integrieren Systeme und Sensoren zum Erfassen, Verarbeiten und Verwalten von Umweltdaten. Geoinformationssysteme dienen häufig dazu, Daten geografisch verorten zu können.

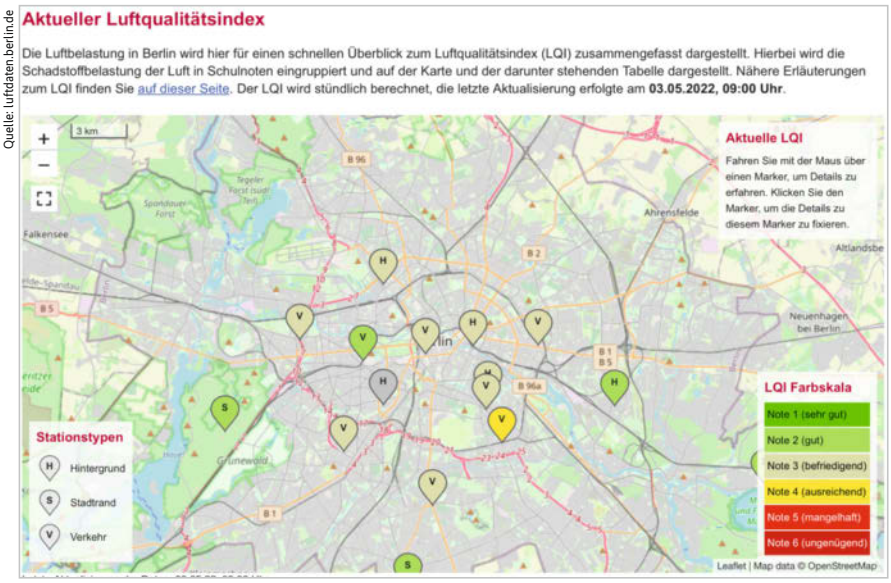
Durch die zunehmende Vernetzung und Kommunikation sowie die Integration in andere Systeme verschwimmen in der Praxis die Grenzen zwischen klassischen Umweltinformationssystemen, Modellierung sowie Berichterstattung einerseits und der Echtzeitinteraktion mit verschiedenen Interessengruppen, der Umsetzung politischer Rahmenbedingungen und Vorgaben sowie der Unterstützung nachhaltiger Geschäftsmodelle andererseits. Heute finden sich Umweltinformationssysteme in der Unterstützung von Gesetzgebung, etwa bei Fragen der Luftreinhaltung, Abwasserqualität oder beim Management von Biodiversität und Fischerei, genauso wie in der Unternehmens-

iX-TRACT

- Umweltinformationssysteme bilden eine breite Klasse an Software, die in diversen Bereichen zum Einsatz kommt.
- Sie liefern und verarbeiten Klima-, Schadstoff- und Verschmutzungsdaten ebenso wie Daten zur Biodiversität oder zu Wildtierwanderungen.
- Umweltinformationssysteme generieren nicht nur Statistiken, sondern werden vor allem für Vorhersagen, Szenarienentwicklung und für das Krisen- und Katastrophenmanagement benötigt.
- In der Industrie können sie zudem als Ökobilanzsysteme die Nachhaltigkeit der Produktion verbessern helfen und einheitliche Reports generieren, wie sie vielleicht schon bald verpflichtend sind.

leitung, beispielsweise beim Risikomanagement von Lieferketten (siehe Artikel „Zerreißprobe“ ab Seite 16).

Quelle: luftdaten.berlin.de



luftdaten.berlin.de gibt Einblicke in die Qualität der Berliner Luft (Abb. 1).

Breite und tiefe Datenbasen

Zunehmend nutzen auch Privatpersonen Informationen aus solchen Systemen, etwa bei Kaufentscheidungen, der Auswahl von Urlaubszielen oder des Wohnorts. Aktivisten verwenden Umweltdaten zur Unterstützung oder als Basis ihrer Arbeit und Journalisten nutzen sie für Recherchen. In politischen oder Planungsprozessen werden Umweltverträglichkeitsprüfungen zu einem wichtigen Teil neuer Vorhaben. Eine breite Datenbasis sowie neue Verfahren der Modellierung und Szenarienbildung begleiten diese Prozesse. Zudem nutzen Personen nicht nur Umweltinformationssysteme, sondern werden zu Datenlieferanten für Umweltinformationssysteme – solche Projekte bezeichnet man als Citizen Science oder Bürgerwissenschaft.

Zu den schon länger etablierten Umweltinformationssystemen gehören Luftgütemessungen, vor allem in Städten. Strukturierte Luftgütemessungen gehen im deutschsprachigen Raum mindestens auf die 1970er-Jahre zurück. In den letzten Jahrzehnten wurden diese Messnetze sukzessive in Umweltinformationssystemen gebündelt und verwaltet. Aus regelmäßigen Berichten wurde eine Darstellung von Echtzeitdaten, wie die auf luftdaten.berlin.de, die Feinstaub, bodennahes Ozon, Stickoxide und Kohlenmonoxid sowie einen Gesamtgüteindikator aufführt (siehe Abbildung 1). Wie in anderen Domänen üblich, gibt es mittlerweile auch bei den Umweltdaten zahlreiche Aggregationsservices, die frei verfügbare Daten zusammenführen, auf Karten verorten und damit einer breiteren Öffentlichkeit zugänglich machen (siehe Abbildung 2 und ix.de/zby8).

Mit Luftgütemessungen in einem engen Zusammenhang stehen meteorologische Beobachtungen. Die systematische Erfassung von Basisinformationen geht an einigen Standorten viele Jahrhunderte zurück. Heute existiert ein dichtes globales Netz an Messstationen, die Temperatur, Niederschlag, Wind und eine Reihe anderer Messdaten wie Wasserstände von Flüssen, manchmal auch ozeanografische Informationen erfassen und deren Daten heute digital verarbeitet und gespeichert werden. Diese Systeme generieren nicht nur Statistiken, sondern bilden die Basis von Vorhersagen und des Krisen- und Katastrophenmanagements etwa bei Hochwasser, Starkregen, Stürmen oder Dürren.

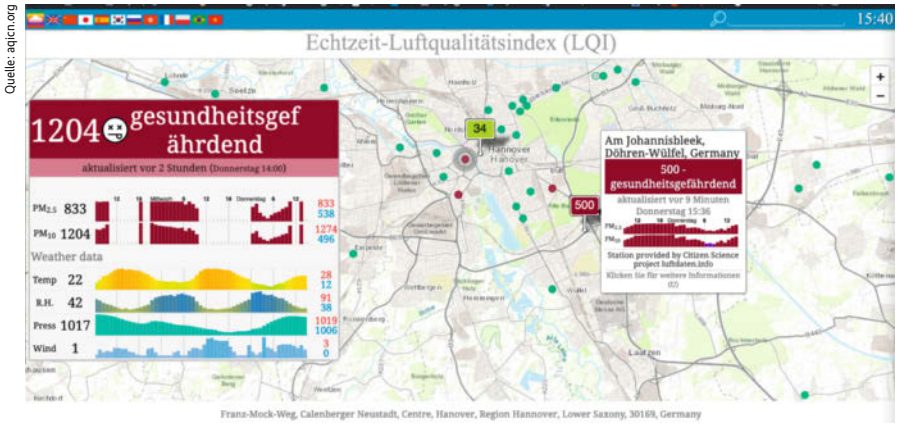
Eingreifen können und verstehen lernen

Eine weitere Anwendungsdomäne, die in den letzten Jahren stark an Bedeutung gewonnen hat, bilden Systeme, die der Überwachung der Biodiversität und dem Management von Ökosystemen dienen. Auch sie arbeiten mit geografischen Informationssystemen zusammen. Ein schönes Beispiel dafür ist das Mara Elephant Project: Es beobachtet die Bewegung der Elefantenherden in der Region unter anderem mit Sensoren. Diese Daten werden ergänzt mit der Position von Fahrzeugen oder Helikoptern der Ranger und Beobachtungsdaten von Personen vor Ort. Die Software Earthranger verarbeitet und visualisiert die Daten in Echtzeit (siehe Abbildung 3). Sie hilft dem Team, das Verhalten der Wildtiere besser zu verstehen und zu erkennen, wann diese definierte Orte verlassen oder betreten, damit die Ranger Konflikte mit der Bevölkerung vermeiden, rechtzeitig eingreifen und Wilderei unterbinden können (siehe ix.de/zby8).

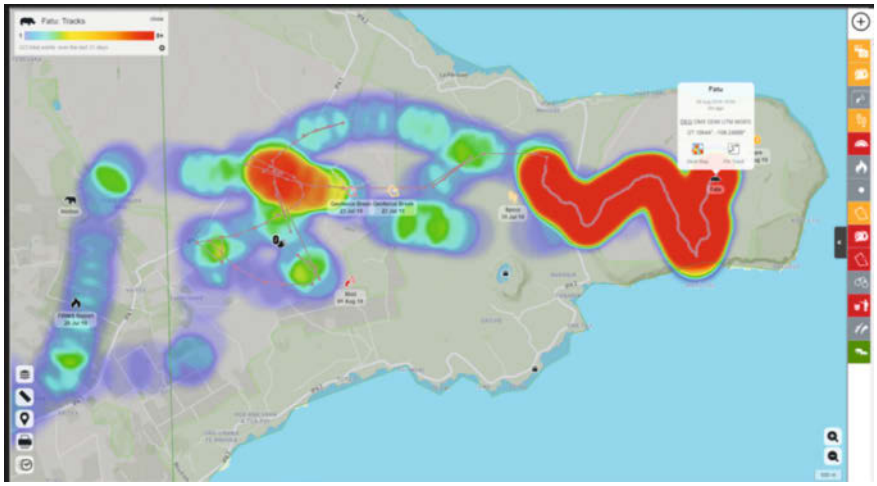
Daten aus derartigen Systemen sind auch für die Forschung zunehmend interessant. Sie helfen den Wissenschaftlern, Ökosysteme besser zu verstehen und Veränderungen zu dokumentieren. Heute sammeln weltweit unterschiedlichste Initiativen Umweltdaten für die wissenschaftliche Arbeit und stellen sie für die Forschung bereit, darunter die Marine Scientific Research Data Website der US-Behörde National Oceanic and Atmospheric Administration (siehe ix.de/zby8). In der Metadatenbank sammelt und strukturiert die NOAA zahlreiche Umweltdaten und Projekte, um sie Wissenschaftlern aus aller Welt zur Verfügung zu stellen.

Einige der Projekte, die mit diesen Daten operieren, haben eine konkrete politische oder geschäftliche Entscheidungsfindung zum Ziel. Beispielsweise wird der Tourismus als einer der wesentlichen Wirtschaftsbereiche vieler Regionen in

Quelle: aqicn.org



Die Visualisierungen der Luftgütemessungen wie hier in Hannover bergen manchmal Überraschungen (Abb. 2).



Quelle: earthranger.com

Earthranger verarbeitet und visualisiert unter anderem die vom Mara Elephant Project gesammelten Daten zu den Bewegungen der Elefantenherden in Echtzeit (Abb. 3).

Zukunft immer stärker gefordert sein, sowohl den eigenen Einfluss auf die Umwelt zu begrenzen als auch sich den sich ändernden Gegebenheiten anzupassen, etwa dem Einfluss des Klimawandels auf Skiregionen (siehe ix.de/zby8). Diese Projekte befinden sich zum Teil noch in der Startphase, aber sie zeigen die zu erwartende stärkere Nutzung von Umweltdaten in Entscheidungsprozessen.

Citizen Science – Beteiligung von Laien

Ein bemerkenswertes Randphänomen in diesem Bereich sind Tätigkeiten von Bürgerwissenschaftlern oder Citizen Scientists. Solche Projekte erfassen wissenschaftlich relevante Daten, indem sie auf Laien zurückgreifen. Oftmals sind es sehr engagierte Menschen, die in der Lage sind, Daten in einem Umfang zu sammeln, der die Möglichkeiten wissenschaftlicher Einrichtungen und Behörden übersteigt.

Beispielsweise sammelt das Coastwatch-Projekt Daten, die es erlauben, den ökologischen Zustand von Küsten besser zu beurteilen. Amateurwissenschaftler besuchen dabei systematisch Strände, beobachten die Biodiversität und dokumentieren die Verschmutzung von Stränden mit eigens bereitgestellten Apps wie Microlitter oder berichten über die Sichtung invasiver Spezies mit Plant Tracker (siehe ix.de/zby8). Aber auch mit eigenen Sensoren tragen Amateure zu dichteren Messnetzen bei: Inzwischen existiert eine eigene Szene, die meteorologische und Luftschadstoffsensoren im eigenen Garten bereitstellt, wartet und die Daten zur allgemeinen Verwendung freigibt.

Das wirft die Frage auf, wie die Zuverlässigkeit der von Amateuren gelieferten Daten etwa durch die Kalibrierung der Messgeräte gewährleistet werden kann. Auch der Missbrauch wäre grundsätzlich denkbar. Insgesamt aber können diese Projekte zumindest ergänzende Informationen liefern, die mit professionellen Messstationen in diesem Umfang nicht leicht und vor allem nicht günstig zu erlangen wären, darunter auch Informationen, die Wissenschaftlern oder Behörden Hinweise geben, wo sie genauer hinsehen sollten.

An einigen Beispielen zeigen sich die zunehmenden Überschneidungen und die fließenden Übergänge zwischen Erfassung, Darstellung und Modellierung und zwischen Umweltinformationssystemen, Forschung und Entscheidungsunterstützung. Während etwa Wetterstationen die aktuellen Daten erfassen, darstellen und archivieren, operieren Modelle mit diesen Daten, um bessere Vorhersagen zu treffen oder bestehende Modelle zu verbessern.

Die Klimaforschung versucht aus derartigen Daten längerfristige Entwicklungen herzuleiten. Sie kombiniert sie mit der Rekonstruktion historischer Daten und entwickelt daraus eine Basis oder Referenz für die Modellierung und Szenarioentwicklung, wie man sie als SSP1-1.9 bis SSP5-8.5 aus den IPCC-Berichten kennt. Die Kombination von Klima- und Wettermodellen wiederum fließt zunehmend in Entscheidungsprozesse ein. Auch für andere Umweltdaten, etwa bodennahes Ozon, werden Umweltinformationssysteme mit der Modellierung verbunden, um daraus Szenarien und Maßnahmen abzuleiten.

Industrielles Reporting im Aufschwung

Auch Unternehmen greifen verstärkt auf Daten von Umweltinformationssystemen zurück oder etablieren eigene Systeme, um umweltrelevante Daten auf Konzernebene zu sammeln. Dazu zählen Daten zu Energie- und Rohstoffverbrauch, Emissionen, Fuhrparkdaten, Daten zur Landnutzung oder Daten, die aus Abwasser- oder Abgasmessungen stammen oder von Drohnen, die den Bewuchs messen.

Es gibt einige Anbieter, die verschiedenste internationale Daten wie Umweltdaten, Wetterberichte und politische Nachrichten kombinieren. Diese werden mit den Daten des Supply-Chain-Managements eines Unternehmens zusammengebracht, um Nachhaltigkeitsrisiken in der Lieferkette frühzeitig zu erkennen: Gibt es Streiks oder Überschwemmungen, die Lieferungen blockieren könnten? Drohen Dürren, die die landwirtschaftliche Produktion beeinträchtigen werden, oder politische Konflikte, die in einen Engpass bestimmter Rohstoffe münden könnten?

Auch in der Energiegewinnung werden zunehmend Onlinesysteme eingesetzt, die die Effizienz und CO_2-Bilanz aktueller Produktion und Nutzung etwa auf regionaler Ebene abbilden können (siehe ix.de/zby8). Derartige Dienste können politisch relevant werden, wenn man beispielsweise die CO_2-Intensität der deutschen Stromproduktion nach einer extrem teuren Energiewende mit der CO_2-Intensität der Nachbarstaaten vergleicht.

Zukünftig besonders relevant kann diese Datenintegration für ein vereinheitlichtes Nachhaltigkeitsreporting werden. Die GRI (Global Reporting Initiative) und das WEF (World Economic Forum) sehen etwa Wasser- und Biodiversitätsrisikobewertungen in den Unternehmensberichten vor. Die Berichterstattung gestaltet sich einfacher, effizienter und präziser, wenn sich die Berichte automatisch mithilfe eines Umweltinformationssystems erstellen lassen. Derartige Reports werden immer mehr durch staatliche Regulative gefordert. In Großbritannien etwa gibt es entsprechende Transparenzanforderungen bereits, in der Europäischen Union sind sie in Planung.

Eine der wesentlichsten Initiativen in diesem Kontext ist die Task Force on Climate Related Financial Disclosures im Rahmen der G20. Sie bildet Klimarisiken auf finanzielle Effekte ab. Unternehmen erstellen zu diesem Zweck Unternehmensszenarien, die auf den etablierten IPCC-Klimaszenarien aufbauen.

Zuletzt soll noch ein Bereich genannt werden, der derzeit im Wesentlichen in Form von Studien und über in Intervallen ak-

tualisierte Datenbanken abgehandelt wird: Ökobilanzen und das Lifecycle Assessment (LCA) von Produkten (siehe ix.de/zby8). Lifecycle Assessment meint die Untersuchung der Umweltauswirkungen eines Produktes oder einer Produktklasse entlang des gesamten Lebenszyklus von der Produktion über die Nutzung bis zu Entsorgung oder Recycling. Zahlreiche Datenbanken bieten Ökobilanzen und Informationen zu Produkten unterschiedlicher Branchen an: Die Palette reicht dabei von Energie- und Ressourcenverbrauch bis zur Dokumentation von Menschenrechtsverletzungen. Die Integration solcher Datenquellen in LCA-Anwendungen wie openLCA erleichtert die Bewertung des Lebenszyklus einzelner Produkte.

Eine sehr grundsätzliche Schwierigkeit dieser Systeme liegt in der enormen Komplexität des Unterfangens, die sich aus vielfachen Wechselwirkungen eines Produktes mit Gesellschaft und Natur ergeben. Allein die Eingrenzung, welche Teile in einen Lebenszyklus mit eingerechnet werden, bietet zahlreiche Grauzonen und Interpretationsspielraum. Dennoch, die zunehmende Digitalisierung all dieser Bereiche legt nahe, dass in Zukunft derartige LCAs näher an die Echtzeit und aktuelle Daten der Prozesse heranreichen und nicht nur auf Durchschnitte oder gelegentlich aktualisierte Datenbanken bezogen werden. Das könnte die Beurteilung der am Markt befindlichen Produkte in wesentlich höherer Präzision und auf Basis der tatsächlichen Nutzung ermöglichen.

Open-Data-Initiativen fordern häufig die weitreichende Öffnung vorhandener Daten zur besseren Integration, aber auch, um Unternehmen und der öffentlichen Hand besser auf die Finger sehen zu können. Jede Transparenzmaßnahme – und für Open Data gilt dies im Besonderen – hat allerdings nicht nur Vorteile, sondern birgt auch Risiken. Ein Blick auf einige der im Artikel genannten Beispiele macht dies deutlich.

Die Schattenseiten der Offenheit

Das Beobachten von Wildtieren mit Sensoren ist leider nicht nur für Biologen und Parkmanager von Vorteil, sondern auch für Wilderer sehr hilfreich, wenn sie Zugriff auf diese Daten erhalten. Offene Daten können hier sehr schnell ein sinnvolles System ins Gegenteil verkehren. Auch das immer einfachere Beobachten von Fischschwärmen hilft nicht nur denjenigen, die sich an die gesetzlichen Vorgaben halten. Nicht zuletzt sind transparente Daten aus Unternehmenslieferketten ein gefundenes Fressen für Akteure aller Art, die damit relativ einfach Schwachstellen finden und Unternehmen erheblichen Schaden zufügen können. (sun@ix.de)

Quellen

Alle Quellen siehe: ix.de/zby8

Dr. Alexander Schatten

ist Senior Researcher bei SBA-Research, Managementberater und Podcaster: podcast.zukunft-denken.eu

Lieferketten mit und für die IT nachhaltiger und resilienter gestalten

Zerreißprobe

Dr. Alexander Schatten

Nicht nur die Politik, auch jedes einzelne Unternehmen muss an der Resilienz der Lieferketten arbeiten. Dabei müssen IT und Organisation Hand in Hand gehen.

Immer häufiger schafft es die mangelnde Resilienz und Nachhaltigkeit von Lieferketten in die Abendnachrichten. Und die schlechten Nachrichten reißen nicht ab: Von der Ransomware-Attacke auf Maersk 2017 über die Auswirkungen der chinesischen Null-Covid-Strategie auf den Frachtverkehr und Zwischenfälle wie den der Ever Given 2021 im Suez-Kanal bis zum Krieg in der Ukraine schaukeln sich die Störungen in den Lieferketten immer weiter hoch.

Die Covid-Krise hat unter anderem deutlich gemacht, dass Europa selbst einfache Produkte wie medizinische Masken und Schutzausrüstung nicht mehr regional produzieren kann und von globalen Lieferketten abhängig ist. Aber auch die Produktion unterschiedlichster Elektronik- und Haushaltsprodukte ist durch Lieferschwierigkeiten von Mikroprozessoren ins Trudeln geraten. Der Ukrainekrieg wiederum führt zu Verwerfungen bei der Energieversorgung, Rohstoffen und Nahrungsmitteln.

Weniger bekannt ist, dass die Ukraine auch ein global bedeutender Produzent des Edelgases Neon ist, das für die Halbleiterproduktion benötigt wird. Das setzt die Industrie weiter unter Druck. Die Produktion von Neon mag für viele eine Kuriosität sein, aber sie zeigt, wie unglaublich komplex die globale Produktion und Lieferung wesentlicher Produkte geworden ist. Eine Störung am anderen Ende der Welt kann nicht vorhersagbare Effekte nach sich ziehen.

So ist es nicht verwunderlich, dass der Allianz Risiko Barometer Cyber-Incidents und Lieferkettenfehler auf Platz eins und zwei listet, also als die bedeutendsten Risiken 2022 (siehe ix.de/zxg5). Aber weder lässt sich dies mit schnellen Maßnahmen in den Griff bekommen noch kommt das überraschend. Bereits vor fast 10 Jahren warnten etwa Apotheker in deutschen Medien da-

iX-TRACT

- Heute sind die Lieferketten von mangelnder Resilienz und Nachhaltigkeit geprägt.
- Erst langsam setzt sich die vergessene Erkenntnis wieder durch, dass Lieferketten gemanagt werden müssen.
- Auch die IT hat zum besorgniserregenden Zustand der Lieferketten beigetragen.
- Eine hohe Priorität nimmt das Thema Lieferketten und Versorgungssicherheit auch in der Politik ein.
- Unternehmen sind in der Lage, ihre eigenen Lieferketten resilienter zu gestalten. Dabei müssen sie aber auch die sich ändernden Regularien im Blick behalten.

vor, dass wesentliche Arzneimittel oder deren Vorläufersubstanzen weltweit nur mehr in wenigen, fallweise einzelnen Produktionsstätten erzeugt werden. Sie sprachen 2014 von untragbaren Zuständen. Auch in anderen Bereichen sollte diese selbst geschaffene Abhängigkeit nicht besonders überraschen.

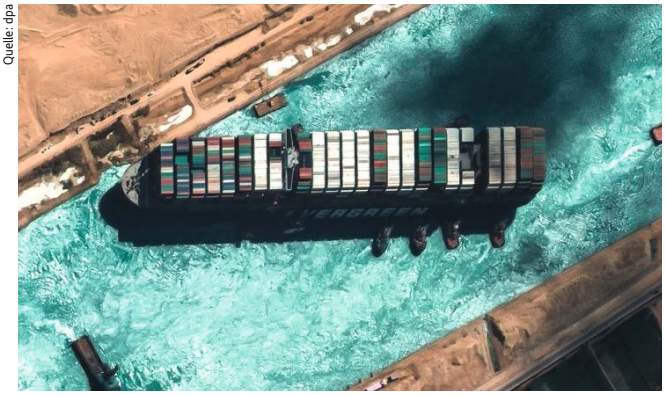
Lieferketten müssen gemanagt werden

Neu ist also weder die Abhängigkeit moderner Gesellschaften von globalen Lieferketten noch das Wissen darüber. Nur hat man die damit verbundenen Risiken bisher ignoriert. Kurzfristige Optimierung und Effizienzsteigerung wurde auch in diesem Fall vor Nachhaltigkeit und Resilienz gesetzt. Krisen und Vorfälle, die die Lieferketten vor immer neue Zerreißproben stellen, legen nur die Schwachstellen offen, die im Kern seit Langem bekannt sind. Die Erkenntnis, dass Lieferketten gemanagt werden müssen, ist ebenfalls keine neue Idee. Der Begriff des Supply-Chain-Management ist mindestens 40 Jahre alt.

Studien großer Beratungsfirmen legen nahe, dass das Risikobewusstsein langsam bei den Unternehmen ankommt. Rund 90 Prozent der Befragten geben an, lokaler beschaffen und die Resilienz der Lieferketten erhöhen zu wollen. Dabei haben laut Befragung aber rund zehn Prozent gar keinen Einblick in ihre Lieferkette, nur rund die Hälfte hat einen Überblick über die erste Stufe und weniger als ein Viertel über die zweite Stufe. Darüber hinaus ist so gut wie kein Unternehmen informiert. Zudem ist in vielen Bereichen eine zu starke Abhängigkeit von einzelnen Lieferanten gegeben.

Da die Geschäftsrisiken in nicht resilienten Lieferketten massiv steigen, mit allen schon jetzt beobachtbaren Folgen, steht die Bedeutung von Nachhaltigkeit und Resilienz in Lieferketten bereits auf der politischen Agenda zahlreicher Staaten. Auf internationaler Ebene findet sich der Hinweis auf nachhaltige Lieferketten schon länger in den Sustainable Development Goals 9 „Industrie, Innovation und Infrastruktur“ und 12 „Verantwortungsvoller Konsum und Produktion“ der UNO.

Quelle: dpa



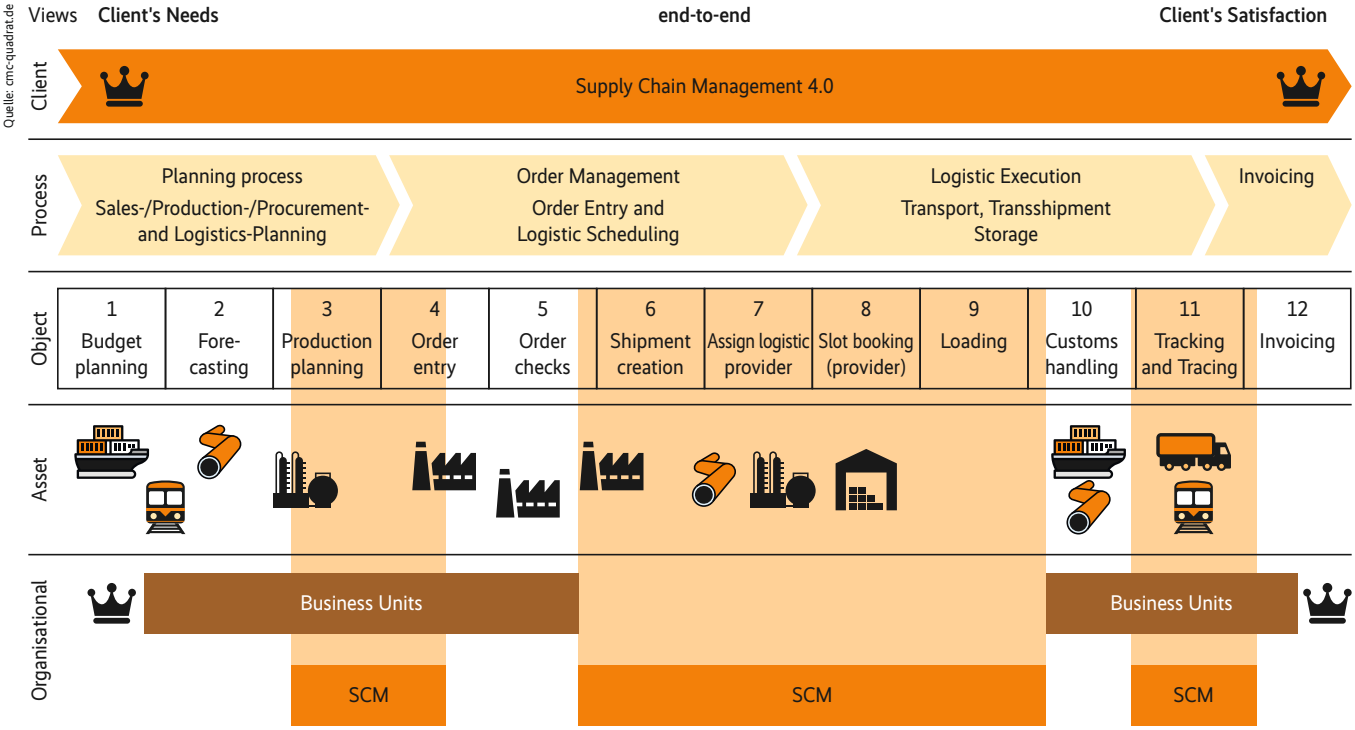
Ein einzelnes Schiff blockiert die Hauptroute zwischen Asien und Europa und verzögert den Nachschub damit um Wochen (Abb. 1).

> „Nachhaltigkeit in der Lieferkette meint das Management der ökologischen, sozialen und wirtschaftlichen Auswirkungen sowie die Förderung guter Unternehmensführung über den gesamten Lebenszyklus von Produkten und Dienstleistungen.“ *United Nations Global Compact, BSR, Nachhaltigkeit in der Lieferkette (2012)*

Hohe Priorität auch in der Politik

US-Präsident Biden erklärt 2022 die Resilienz von Lieferketten zu einer Top-Priorität seiner Regierung und fordert in einer gemeinsamen Stellungnahme mit Politikern anderer Nationen Frühwarnsysteme, mehr Transparenz, das Teilen von Daten sowie mehr Diversität, Offenheit und Sicherheit (siehe ix.de/zxg5). Die EU arbeitet ebenfalls an Regeln für Unternehmen, um Menschenrechte, Umwelt und Nachhaltigkeit in globalen Lieferketten zu verbessern. Teilweise sind sie bereits in nationale Gesetzgebung eingeflossen (siehe ix.de/zxg5).

Quelle: cmc-quadrat.de



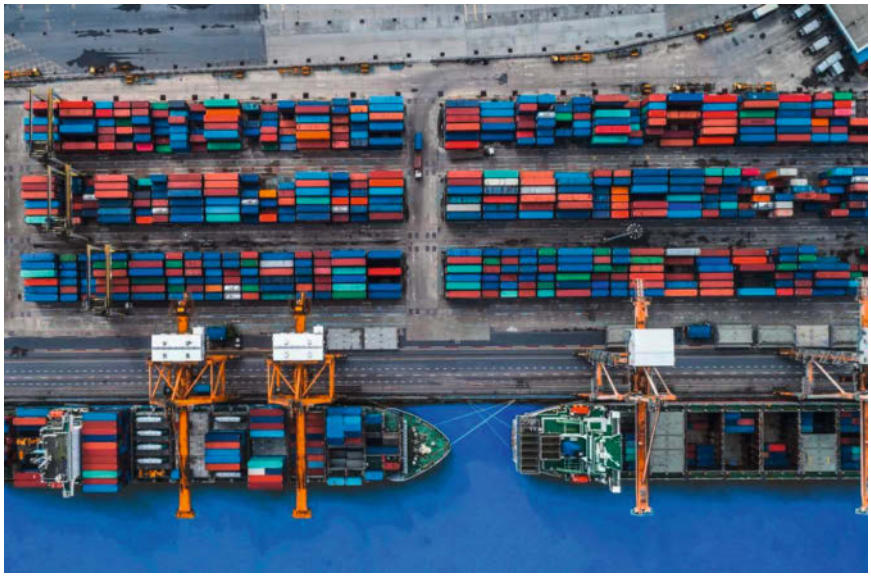
Das Supply-Chain-Management umfasst nicht nur viele Schritte in einer Kette, sondern auch unterschiedliche Ebenen (Abb. 2).

„[Unternehmen] müssen nachteilige Auswirkungen ihrer Aktivitäten auf die Menschenrechte, wie Kinderarbeit und Ausbeutung von Arbeitnehmern, und auf die Umwelt, beispielsweise Umweltverschmutzung und Verlust der biologischen Vielfalt, erkennen und – falls erforderlich – verhindern, beenden oder mindern. Für Unternehmen bringen diese neuen Vorschriften Rechtssicherheit und gleiche Wettbewerbsbedingungen. Für Verbraucher und Investoren werden sie mehr Transparenz schaffen.“ *Europäische Kommission*

In der Definition von Nachhaltigkeit im Kontext von Lieferketten finden sich dieselben Dimensionen wieder wie im Artikel „Durchschritten“ ab Seite 8: Umwelt, soziale Aspekte, Technik und Ökonomie. Beim Bemühen, die geforderten Ziele zu erreichen, rückt auch die Rolle der IT immer stärker in den Fokus. Denn Nachhaltigkeit und Resilienz in Lieferketten haben eine immer stärkere IT-Komponente, und das in wechselseitiger Abhängigkeit: Möchte man die Produkt- und Leistungsgruppen in der Lieferkette besser abbilden, die Materialflüsse verfolgen, Informationen über die gesamte Lieferkette transparent machen und mit politisch und ökonomisch Betroffenen austauschen, muss die IT die entsprechenden Hilfsmittel zur Verfügung stellen. Sie gerät hier aber selbst in die Falle nicht nachhaltiger Lieferketten, wie etwa am Chipmangel 2021 und 2022 zu beobachten war, aber auch durch strategische Fehlentscheidungen (siehe die Artikel „Ein schwerer Anfang“ und „Nur gemeinsam“ ab Seite 40 und 48).

Lieferketten und IT – systemische Verstrickungen

Die IT ist also sowohl Betroffene und Verursacherin als auch Verbündete: Just-in-Time-Lieferketten und moderne Logistik wurden erst durch eine weitreichende Digitalisierung, Geo-Services, RFIDs und dergleichen möglich. Neben einer höheren Effizienz führt dies auch zur höheren Transparenz der Herkunft der Produkte. Diese weiter reichende Transparenz ist allerdings, wie die Umfragen zeigen, bisher auf wenige Produzenten und kurze Lieferketten beschränkt.



Quelle: IBM

Ohne weitreichende Digitalisierung, Geo-Services, RFIDs und dergleichen ist eine derart verdichtete Logistik wie die heutige gar nicht möglich (Abb. 3).

Komplexe Produkte benötigen so viele verschiedene Rohstoffe, Zwischenschritte und Vorproduzenten, dass eine weitreichende Nachverfolgung und Risikoabschätzung bisher kaum möglich ist. Dies wird sich nicht wesentlich verändern. Nach hiesigen Regeln produzierende Hersteller sind kaum in der Lage, ihren direkten Zulieferern zumindest die formelle Zustimmung zu einem Verhaltenskodex zu entlocken, geschweige denn, dessen Einhaltung zu überwachen oder etwas über die Arbeitsbedingungen von deren Zulieferern in Erfahrung zu bringen.

Oft scheitert es schon an kulturellen Unterschieden. Ein Beispiel: Kinderarbeit, das Gefangenhalten vietnamesischer Arbeitssklaven oder Arbeitswochen von mehr als 60 Stunden bereiten den Verantwortlichen in manchen chinesischen Gegenden keinerlei schlechtes Gewissen. Hier überhaupt ein Unrechtsbewusstsein zu schaffen, ist über die geografische und kulturelle Distanz kaum möglich.

In einem Mikroprozessor – und dies ist nur ein Teil eines modernen Computers, Smartphones oder einer Waschmaschine – findet man mit etwa 60 chemischen Elementen schon das halbe Periodensystem wieder, wenn auch zum Teil nur in Spuren. Jedes dieser Elemente ist aber erforderlich und kommt aus einer Mine, hat eine eigene Liefer- und Verarbeitungskette. Hinzu kommt eine Vielzahl an Substanzen wie das besagte Neon, Maschinen und Know-how für die Produktion. Die IT ermöglicht also auf der einen Seite die komplexen Lieferketten, die die moderne Gesellschaft bestimmen, ist aber gleichzeitig selbst von diesen Lieferketten abhängig und kann die Krise verschärfen.

Auch die IT verursacht Störungen der Lieferketten

In den letzten Jahren gab es eine Reihe von der IT verursachter Störungen in den Lieferketten, deren Dimensionen verdeutlichen, dass sie immer mehr zu existenziellen Bedrohungen heranwachsen. Allein der NotPetya-Angriff auf den globalen Logistikdienstleister Maersk, der rund 20 Prozent des Welthandels in Schiffscontainern abdeckt, im Jahr 2017 hatte zur Folge, dass weltweit rund 45 000 Clients und 4000 Server neu installiert werden mussten, was einer Neuinstallation der kompletten IT-Infrastruktur gleichkam. Während dieser Krise arbeitete Maersk zehn Tage lang analog, also offline, und verzeichnete einen Schaden von 250 bis 300 Millionen US-Dollar (siehe ix.de/zxg5).

Die dabei aber systemisch wichtigere Frage für die Zukunft scheint zu sein: Lassen sich diese immer stärker digitalisierten Prozesse in der Zukunft noch manuell oder analog aufrechterhalten? Lassen sich nach einem Ausfall der IT-Systeme überhaupt noch die Containerinhalte auf den Schiffen ermitteln? Möglicherweise wird es verstärkt autonom navigierende Schiffe und davon abhängiges Inventarmanagement geben. Auch in großen automatisierten Lagerhäusern werden Waren schon heute häufig nicht mehr so angeordnet, dass Menschen sie ohne die entsprechenden Datenbanken finden würden. Ab diesem Zeitpunkt ist ein manueller Eingriff, ein analoger Notfallprozess, überhaupt nicht mehr möglich.

Im Jahr 2021 kam es ebenfalls zu einem Cyberangriff, diesmal auf die Colonial Pipeline in den USA, die etwa 9000 km lang ist und 3 Millionen Barrel Öl pro Tag zwischen Texas und New York transportiert (siehe ix.de/zxg5). Auch in diesem Fall war es eine Ransomware-Attacke, die dazu führte, dass die Colonial Pipeline Company das Pipelinesystem für fast eine Woche herunterfuhr, was zu Lieferengpässen an der Ostküste der Vereinigten Staaten führte.

Insgesamt nehmen Cyberangriffe stetig zu. Drei Viertel aller Organisationen geben an, 2020 unter einem Cyberangriff gelitten zu haben – von Phishing über Denial of Service bis zu Ransomware. Rund zehn Prozent der Organisationen beziffern den Schaden durch Cyberangriffe kumulativ auf mehr als eine Million Euro. Störungen der Lieferkette sind dabei häufig – wie die obigen Beispiele nahelegen – Seiteneffekte solcher Angriffe.

Lieferketten sind keine vorhersagbaren Systeme

Lieferketten sind aufgrund ihrer Komplexität und ihrer Eigendynamiken nicht mehr als statische, vorhersagbare Systeme zu verstehen, sondern vielmehr als komplexe, global verteilte und adaptive Systeme. Für Supply-Chain-Manager sind die Zeiten vorbei, in denen zukünftige Zustände und Risiken prinzipiell vorhersagbar waren. Die Hoffnung der Biden-Regierung und der internationalen Partner auf die Vorhersagbarkeit wird sich daher nicht erfüllen. Um zu vermeiden, dass selbstverstärkende Effekte in solchen systemischen Konstellationen katastrophale Wirkungen entfalten, müssen neue Managementprinzipien her (siehe Artikel „Nur gemeinsam“ ab Seite 48).

Neben den Hindernissen, die sich für einzelne Unternehmen ergeben, gibt es auch strategische Aspekte, die etwa die Europäische Union im Ganzen treffen: Die digitale Infrastruktur Europas ist bei Soft- und Hardware abhängig von wenigen Lieferanten aus politisch instabilen Ländern. Japan hat als Gegenreaktion in den letzten Jahren die Möglichkeit ausländischer Unternehmen eingeschränkt, an öffentlichen Ausschreibungen teilzunehmen, und fördert lokale Unternehmen, die ihre Produktion wieder aus China zurück nach Japan verlagern. Dies soll auch explizit die Lieferketten sicherer machen (siehe ix.de/zxg5).

Allerdings bleibt es nicht bei den klassischen Risiken der Verfügbarkeit, Lieferbarkeit und der Kosten einzelner Komponenten durch strategisch ungünstige Konzentration von Lieferketten. Die Komplexität moderner Infrastruktur macht es auch immer leichter, versteckte Features in Systeme einzubauen und damit Spionage zu ermöglichen oder Cyberangriffe vorzubereiten (siehe Artikel „Ein schwerer Anfang“ ab Seite 40).

Für die resilientere Gestaltung der eigenen Lieferkette ergeben sich eine Reihe konkreter Handlungsfelder. Am Anfang steht das Verständnis der Risiken im eigenen Unternehmen oder der eigenen Organisation: Was soll geschützt werden und warum? Von welchen Lieferanten ist man in welchem Maße abhängig? Daraus folgt der Versuch einer ersten Bewertung des Risikos, das von den Lieferketten ausgeht. Mindestanforderungen an Lieferanten sind zu beschreiben und mitzuteilen.

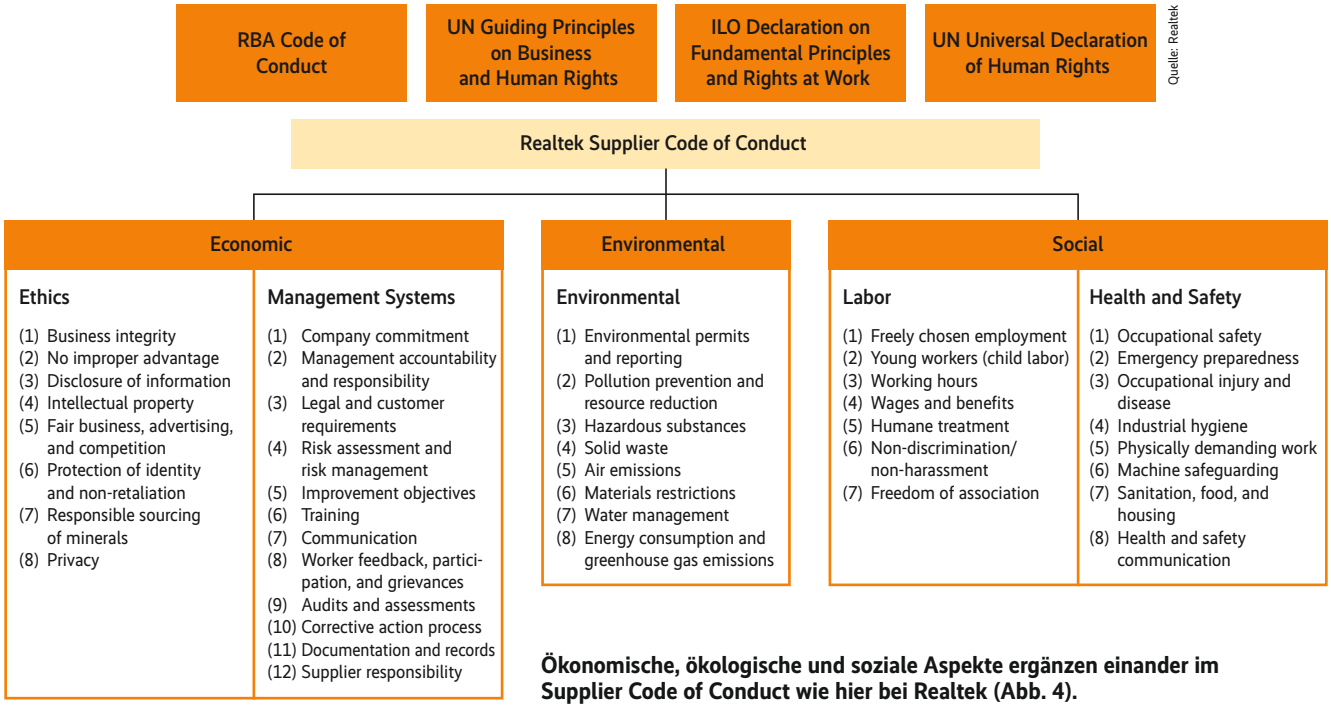
Der zweite Schritt besteht darin, die Kontrolle zurückzuerlangen und Verantwortung für die Lieferkette zu definieren. Diese darf sich nicht allein in einer Supply-Chain-Abteilung wiederfinden. Auch die Softwareentwicklung hat in den letzten Jahrzehnten gelehrt, dass sich Qualität nicht durch die alleinige Schaffung einer Abteilung verbessern lässt. Dies trifft in verstärktem Maße für Software-Security-Aspekte und Lieferkettenresilienz zu. Koordinierung und Reporting lassen sich in einer Abteilung zentralisieren, das Bewusstsein muss aber im gesamten Unternehmen verankert sein. Sind die Rahmenbedingungen geklärt, sind konsequente Kontrollprozesse, interne und externe Audits der Lieferanten einzuführen; eine interne Lieferantendatenbank sollte ständig aktualisiert werden.

Das Unternehmen trägt die Risiken, die sich aus Security-Incidents und der Lieferkette ergeben, im Wesentlichen selbst. Versicherungen spielen keine nennenswerte Rolle, weder gegen Cyberangriffe noch gegen Suppy-Chain-Risiken. Die Risiken sind schlicht zu komplex und die möglichen Folgen zu schwerwiegend beziehungsweise nicht einfach in Policen zu formulieren.

Ebenso wichtig ist es zu analysieren, wie stark die Abhängigkeit von einzelnen Lieferanten oder Regionen ist. Redundanz und Vielfalt rechnet sich mittelfristig besser als kurzfristige Effizienz mit hohem Clusterrisiko. Bei Kernprozessen kann es ein sehr kluges Ziel sein, Prozesse so zu gestalten, dass sie zwar digital effizient ablaufen, aber im Notfall auch analog abgearbeitet werden können.

Reporting verbessern und Risiken abschätzen

Das Schaffen von Transparenz in den Lieferketten hat mehrere Dimensionen: Aus strategischer Sicht folgen Unternehmen heute der Triple Bottom Line oder dem Shared Value Concept (siehe ix.de/zxg5). Bei dieser speziellen Ausprägung der Balanced Scorecard steht nicht mehr nur die ökonomische Performance im Vordergrund, sondern auf gleicher Ebene die Aus-



Ökonomische, ökologische und soziale Aspekte ergänzen einander im Supplier Code of Conduct wie hier bei Realtek (Abb. 4).

wirkungen des Unternehmens auf soziale Aspekte und die Umwelt (siehe Abbildung 4).

Damit einher geht meist auch die Verbesserung des Unternehmensreportings, in der Regel mit Standards wie GRI (Global Reporting Initiative) oder den WEF Stakeholder Capitalism Metrics. Diese Reporting-Vorlagen bilden auch Lieferketten und Nachhaltigkeitsrisiken ab. Ist dies als Unternehmensstrategie verankert, sind auch die dafür notwendigen Transparenzmaßnahmen leichter zu etablieren.

Der Versuch, die eigenen Lieferketten top-down unter Kontrolle zu halten, ist vor allem bei größeren Unternehmen mit Zehntausenden Lieferanten und vielschichtigen Lieferketten nur schwer zu bewerkstelligen. Deshalb gibt es zahlreiche Dienstleister, die globale Lieferantendatenbanken mit Risikobewertungen und -profilen aufbauen und aktualisieren, White- und Blacklists führen und zusätzlich die Bewertung durch Rückgriff auf offene Daten, soziale Netzwerke und Nachrichtenagenturen erweitern. Zusätzlich kann aus ökologischen Gesichtspunkten der Einsatz von Lifecycle-Analysen angebracht sein (siehe Artikel „Bestandsaufnahme“ ab Seite 12).

Allerdings hat jede Maßnahme in einem komplexen Umfeld mehrere Seiten; dies trifft auch auf Transparenzmaßnahmen zu. Auch wenn heute ein Mehr an Transparenz in der Öffentlichkeit immer als wünschenswert gilt, werden dabei gerne die Schattenseiten übersehen: Sichtbarkeit ist ja nicht nur für konstruktive Akteure gegeben, sondern auch für destruktive. Liegen Lieferketten tatsächlich offen, ist dies natürlich auch für Kriminelle eine wertvolle Ressource, um Angriffe besser planen und Schwachstellen aufspüren zu können. Dies betrifft sowohl Schwachstellen der IT-Infrastruktur wie auch der Lieferketten selbst. Die Partner, mit denen Informationen ausgetauscht werden, sollten daher gründlich ausgewählt werden. Vor einer generellen und unbedingten Transparenzoffensive etwa im Unternehmensreporting ist daher abzuraten.

Lieferketten, besonders diejenigen größerer Unternehmen, sind im Grunde immer international. Damit ist eine Reihe von Regularien zu beachten, etwa der UK Modern Slavery Act, der moderne Sklaverei im Visier hat, aber auch der California Transparency in Supply Chain Act und der Australian Modern Slavery Act 2018 entfalten internationale Wirksamkeit. Auch die EU initiiert gerade eine neue Gesetzgebung. Kurz gesagt: Die politischen Akteure sind durch die Krisen und Vorfälle der letzten Jahre hellhörig geworden und schaffen neue Regularien.

Diese Richtlinien lassen sich in aller Regel nur durch konsequente Prozesse im Unternehmen und den Einsatz von IT und Informationsdienstleistern in der Lieferkette einhalten. Die für Software im engeren Sinne und von Software abhängige Prozesse bestehenden Herausforderungen beschreibt der Artikel „Ein schwerer Anfang“ ab Seite 40.

Fazit

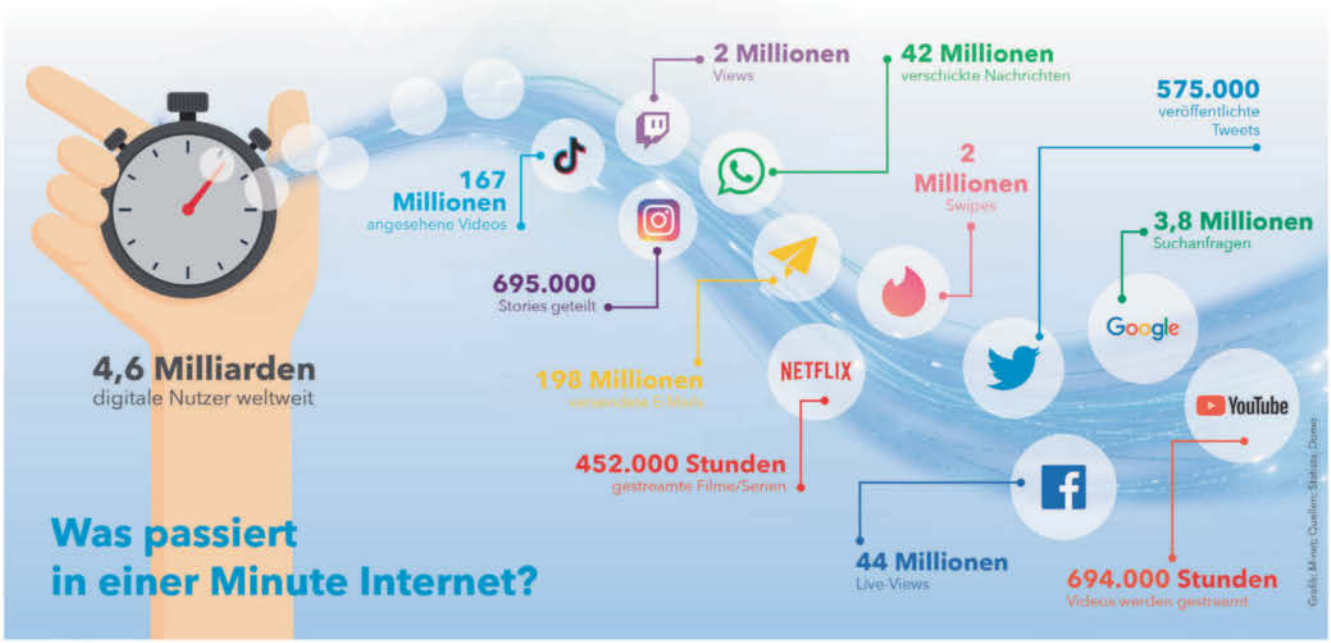
Nachhaltige Lieferketten sind eine ungeheuer schwierige und zugleich dringliche Herausforderung in der modernen Welt. Dabei treffen komplexe IT-Systeme auf komplexe Produktions- und Lieferprozesse und wirken wechselseitig aufeinander. Helfen kann hier der kluge Einsatz von Digitalisierung und Transparenzmaßnahmen. Ein Zurück zu der Idee einer deterministischen Sicherheit und Vorhersagbarkeit gibt es aber nicht mehr. Dies ist eine falsche Idee, die nur das mittelfristige Risiko drastisch erhöht. Vielmehr spielen langfristige geopolitische und strategische Überlegungen eine immer größere Rolle, ebenso wie die Erkenntnis, dass Diversität, Redundanz, die Vermeidung von Clusterrisiken und Graceful Degradation, also die Möglichkeit, wesentliche Prozesse auch analog zu betreiben, unverzichtbar sind. (sun@ix.de)

Quellen

Alle Quellen siehe ix.de/zxg5

Dr. Alexander Schatten

ist Senior Researcher bei SBA-Research, Managementberater und Podcaster: podcast.zukunft-denken.eu



Business-Trend 2022: Klimafreundliches Internet

Was haben Homeoffice, Industrie 4.0 und Cloud-Computing gemeinsam? Richtig, sie verursachen einen rasanten Anstieg beim Datenaufkommen deutscher Unternehmen. Doch was vielen nicht bewusst ist: Wer das Internet nutzt, verursacht unmittelbar auch Treibhausgase.

Grüner Surfen per Glasfaser

Allein die rund 3,8 Millionen Google-Suchanfragen einer einzigen Minute verursachen etwa 769 Kilogramm CO_2. Umgerechnet auf den Verbrauch eines Autos entspricht das rund 2.360 gefahrenen Kilometern. Mit Blick auf die Klimabilanz macht es daher einen großen Unterschied, welche digitale Infrastruktur ein Unternehmen nutzt und wie diese betrieben wird. So werden in einem FTTH-Glasfasernetz, bei dem die Glasfaserkabel bis ins Firmengebäude oder das Büro reichen, Daten komplett durch Lichtimpulse übertragen. Anders ist das bei Netzen mit Kupferkabeln, beispielsweise bei herkömmlichen VDSL- oder Kabelanschlüssen. Dort muss das Signal in elektrische Impulse umgewandelt werden, was bis zu 15-mal mehr Energie benötigt. Echte FTTH-Anschlüsse, wie sie auch der Münchner Glasfaseranbieter M-net für Geschäftskunden anbietet, tragen daher maßgeblich zu einer klimaschonenderen Nutzung des Internets im gewerblichen Umfeld bei.

Gelebter Klimaschutz: M-net ist der erste klimaneutrale Telekommunikationsanbieter in Deutschland

Seit 2018 hat Bayerns führender Glasfaseranbieter mit einer ambitionierten Klimastrategie daran gearbeitet, seinen CO_2-Fußabdruck zu verringern. So hatte M-net im ersten Halbjahr 2021 als erster deutscher Telekommunikationsanbieter die Klimaneutralität erreicht. Diese wurde nun durch den TÜV Rheinland für das gesamte Geschäftsjahr 2021 bestätigt. Betrachtet werden dabei alle direkten Emissionen des Unternehmens sowie Emissionen durch beschaffte Energie. Bereits in den vergangenen Jahren konnte M-net rund 90 Prozent seiner vermeidbaren Treibhausgasemissionen einsparen. So werden inzwischen alle Standorte mit eigenem Strombezug ausschließlich mit zertifiziertem Grünstrom betrieben. M-net hat zudem seit dem Sommer 2021 einzelne Glasfaserverteilerschränke mit Photovoltaikanlagen zur eigenen Energieversorgung ausgestattet.

Die vergleichsweise geringe Menge an CO_2-Emissionen, die im Moment noch nicht ganz vermieden werden können, kompensiert M-net mit der Förderung eines internationalen Klimaschutzprojekts nach Goldstandard in Madagaskar und mit einem regionalen Projekt in Poing bei München. Mit seiner ambitionierten Klimastrategie leistet M-net einen wichtigen Beitrag zum Pariser Klimaschutzabkommen und gleichzeitig zu einer nachhaltigeren Wirtschaft und Zukunft in der Region.

M·net Mehr Infos unter: m-net.de/klimaneutral
Alles zum Thema Glasfaser-Internet unter:
m-net.de/geschaeftskunden/business-internet

Vorteile der Glasfaser

Glasfaseranschlüsse für Unternehmen punkten mit einer Reihe von Vorteilen:

- Bandbreite: Höchste Performanz mit skalierbaren Bandbreiten bis zu 100 Gbit/s.
- Geschwindigkeit: Besonders schnelle Reaktionszeiten durch geringe Signalverzögerung (Latenz) – perfekt für Echtzeitanwendungen wie Video-Telefonie und Cloud-Services.
- Zuverlässigkeit und Ausfallsicherheit: Stabiler und sicherer als alternative Technologien.
- Zukunftsfähige Technologie: Wegbereiter für die Digitalisierung mit leistungsstarken Internet-, Telefonie- und Vernetzungslösungen.
- Nachhaltigkeit: Bis zu 15-mal geringerer Energieverbrauch und bessere CO_2-Bilanz als kupferbasierte Netze.

Knappe Rohstoffe als Richtungsweiser

Nicht in den Himmel

Bernd Schöne

Wie sich die IT weiterentwickeln wird, hängt auch vom Umgang der Hightechbranchen mit den von ihnen benötigten Rohstoffen ab.

Energie- und Verkehrswende, Digitalisierung und Industrie 4.0, Clouds und 5G: Nichts davon geht ohne IT, und doch gerät sie in immer größere Konkurrenz zu ihnen. Denn sie alle verlangen nach wertvollen und knappen Rohstoffen, und zwar mehr und mehr: Benötigte ein Fahrzeug mit Verbrennungsmotor noch 20 kg Kupfer, verlangt ein Elektroauto 80 kg. Dazu kommen bis zu 3 kg seltene Erden für die Magnete. Zudem soll die Erzeugung von Strom durch die erneuerbaren Energien 30- bis 40-mal kupferintensiver sein als die konventionelle, so die Hans-Seidel-Stiftung in ihrer Studie „Versorgungssicherheit bei kritischen Rohstoffen“. Doch damit stehen Europa und die USA nicht allein; vor allem China steigert seinen Bedarf unaufhörlich (siehe Abbildung 1).

Die letzte große Rohstoffkrise ist schon einige Jahrzehnte her. Zeitzeugen erinnern sich vor allem an die autofreien Sonntage im November 1973, die letztlich völlig wirkungslos blieben. Als Folge der Ölkrise von 1973/1974 wurde die IEA (Internationale Energie Agentur) gegründet, als selbstständige Organisation innerhalb der OECD (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung). Längst beschäftigt sich die IEA nicht mehr nur mit Öl. Sie beobachtet auch andere wichtige Rohstoffe. Denn Elektrofahrzeuge benötigen viel Kobalt, Lithium und Nickel; Windkraft- und Fotovoltaikanlagen große Mengen an Kupfer.

X-TRACT

- Alle Hightechbranchen benötigen seltene oder knappe Rohstoffe.
- Der Ressourcenhunger der Industrienationen wächst und damit auch die Abhängigkeit von Ländern mit Lagerstätten, insbesondere von China.
- Die nicht immer gesicherte Verfügbarkeit wird durch Störungen in den Lieferketten verschärft.
- Auch die Rohstoffspekulation nimmt zu und führt zu noch größerer Preisunsicherheit.

Ressourcenhunger

Die IEA schätzt, dass sich der jährliche Kupferbedarf in den kommenden 20 Jahren verdoppelt, der von Nickel verdreifacht und der von Kobalt versechsfacht. Noch weit dramatischer sieht es bei Lithium aus. Energie- und Verkehrswende haben den Bedarf des früheren Außenseitermetalls explodieren lassen. „Ein typisches Elektroauto benötigt sechsmal so viel Minerale wie ein Verbrenner, eine Offshore-Windkraftanlage sogar dreizehn-

mal mehr als ein Gaskraftwerk mit der gleichen Leistung“, so die Experten der IEA. Wenn die Menschheit die selbst gesteckten Klimaziele bei gleichbleibendem Wohlstand erreichen will, werden im Jahr 2040 mindestens viermal mehr Minerale benötigt, als heute zur Verfügung stehen und nachgefragt werden. Das Ziel „null CO_2-Emissionen“ bis zum Jahr 2050 würde die sechsfache Menge erfordern.

Das wirft die Frage auf: Werden alle ambitionierten „grünen“ Projekte genug Ressourcen zur Verfügung haben, um sich rasch zu entwickeln? Glaubt man den Experten, sieht es wohl eher schlecht aus. Denn anders als beim Öl, das längst nicht mehr nur in arabischen Ländern gefördert wird, hält sich die Auswahl bei den Abbaugebieten mineralischer Rohstoffe in Grenzen.

Dazu gesellen sich die Störungen im Transportwesen: hohe Krankenstände, lokale Lockdowns und anderweitig bedingte Verzögerungen in der Frachtabwicklung. 12 Prozent der weltweit zirkulierenden Container befinden sich laut Kieler Institut für Weltwirtschaft auf einem unbewegten Schiff. Sie dümpeln vor amerikanischen oder chinesischen Häfen, weil an Land keine Kapazität zum Löschen der Fracht vorhanden ist. 200 000 Seeleute warteten zudem pandemiebedingt fern der Heimat auf ihre Rückreise.

Alles ist knapp

Knapp ist heute alles. Die Container selbst, freie Ladebuchten an Bord der Schiffe und die Besatzung. Die Frachtraten haben sich auf wichtigen Strecken zwischen Januar 2020 und Januar 2022 verfünffacht. Kunden schätzen sich glücklich, überhaupt einen Reeder mit freier Kapazität zu finden. Das wird die Rohstoffe weiter verteuern und die Lieferzeiten in die Länge ziehen. Die Zeiten, als Produzenten Schiffe und Containerterminals als preiswertes Zwischenlager der Just-in-Time-Wirtschaft nutzten, sind vorbei (siehe Artikel „Zerreißprobe“ ab Seite 16).

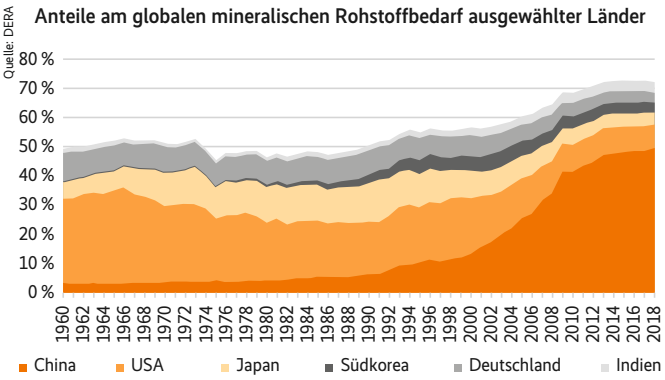
Nach Meinung der Reeder müssen deshalb noch größere Schiffe her. 400 Meter lange Schiffe mit Raum für 20 000 Standardcontainer (TEU) sind heute bereits die Regel, die ersten Schiffe für 24 000 TEU sind seit 2021 unterwegs (siehe Abbildung 2). Die 500-Meter-Marke wird anvisiert. Doch schon an den heutigen 400-Meter-Schiffen gibt es Kritik. Für noch größere Schiffe existieren zudem kaum Trockendocks; im Falle eines Defekts müssten sie quer über die Ozeane geschleppt werden, eventuell bis zur Werft ihrer Entstehung.

Spätestens seit Januar 2019 wird in den Niederlanden und in Deutschland auch über die Sicherheit von Frachtschiffen diskutiert. Damals verlor der Frachtriese MSC Zoe im Sturm 342 seiner 8062 Container vor Borkum im sensiblen Wattenmeer, was umfangreiche Such- und Säuberungsmaßnahmen zur Folge hatte (siehe Abbildung 3). Allein 2021 gingen laut Allianz-Versicherung 3000 Container über Bord. Schlechtes Wetter mit extremem Seegang, marode Schiffe und schlecht ausgebildete Mannschaften gelten als Ursache. Verlorene Container unterbrechen die Lieferketten, sie verschmutzen die Umwelt und gefährden andere Schiffe. Es ist durchaus möglich, dass irgendwann auch die Qualität und Sicherheit des Schiffstransports in Ökosiegel einfließt.

Rohstoffabhängigkeit unter Beobachtung

Auch die Europäische Kommission beobachtet die angespannte Situation in der Rohstoffversorgung aufmerksam. Seit 2011 lässt sie die Märkte analysieren. Sie sucht „kritische“ Rohstoffe,

Quelle: DERA



Insbesondere der Ressourcenhunger der chinesischen Industrie ist in den letzten Jahren stark gewachsen (Abb. 1).

also solche, deren Verfügbarkeit fraglich ist oder werden könnte, die aber benötigt werden. Die jüngste Studie „Study on the EU's list of Critical Raw Materials (2020)“ verheißt nichts Gutes (siehe Abbildung 4).

Stufte die EU 2011 noch 14 von 41 für die europäische Wirtschaft wichtigen in der Erde vorkommenden Rohstoffen als kritisch ein, waren es 2014 schon 20 von 54, 2017 27 von 78 und nun 30 von 83 (siehe Liste der kritischen Rohstoffe). 10 dieser 83 Rohstoffe fasst die EU in drei Gruppen zusammen: die leichten und schweren seltenen Erden sowie die Metalle der Platingruppe (siehe Tabelle „Derzeit in der EU beobachtete Elemente der Platingruppe und der seltenen Erden“).

Ernüchternd ist auch der Blick auf die von der EU erstellte Grafik der Lieferländer (siehe Abbildung 5). Die Abhängigkeit von China ist augenfällig, vor allem bei seltenen Erden. Die sind zwar gar nicht so selten, sondern reichlich in der Erdkruste vorhanden, aber sie sind schwer und nur mit Aufwand zu isolieren. Bei diesen Zwischenprodukten hat sich China eine Sonderstellung erarbeitet.

Seit 2008 bemüht sich die EU um mehr Unabhängigkeit von Importen, allerdings ohne die Bürger mit einzubeziehen. Der Versuch, den Lithiummangel durch Abbau der sehr großen Vorräte in Jadar, Serbien, zu lindern, führte im Januar 2022 zu gewaltsamen Protesten der Anwohner, die die serbische Regierung dazu veranlassten, dem britisch-australischen Bergbaukonzern Rio Tinto die Lizenz zu entziehen. Alle Hoffnungen ruhen derzeit auf Österreich, denn im Bundesland Kärnten werden auf der Koralpe 18 Millionen Tonnen des begehrten Metalls vermutet.

Spielball der Spekulationen

Längst sind Spekulanten auf den Nachschubmangel aufmerksam geworden. Riechen die Anleger Lunte, investieren sie in die

Quelle: eBlue Economy

23 992 Container transportiert die 400 Meter lange Ever Ace der Reederei Evergreen, angetrieben von einem Elf-Zylinder-Dieselmotor mit 96 465 PS. Sie lief 2021 als erstes von sechs geplanten Schiffen dieser Klasse vom Stapel (Abb. 2).

Quelle: picture alliance/Jan Spoelstra/dpa

Auch auf Ameland spülte die Nordsee die Reste der von der MSC Zoe über Bord gegangenen Fracht an, darunter zerbeulte Kühlschränke. Auf Borkum strandeten ganze Fernseher, zerstörte Container und leere Säcke, die an Bord noch das giftige Dibenzoylperoxid enthalten hatten. Nach Berechnung des Wasserstraßen- und Schifffahrtsamts Emden soll ein Viertel der verlorenen Fracht noch auf dem Meeresgrund liegen (Abb. 3).

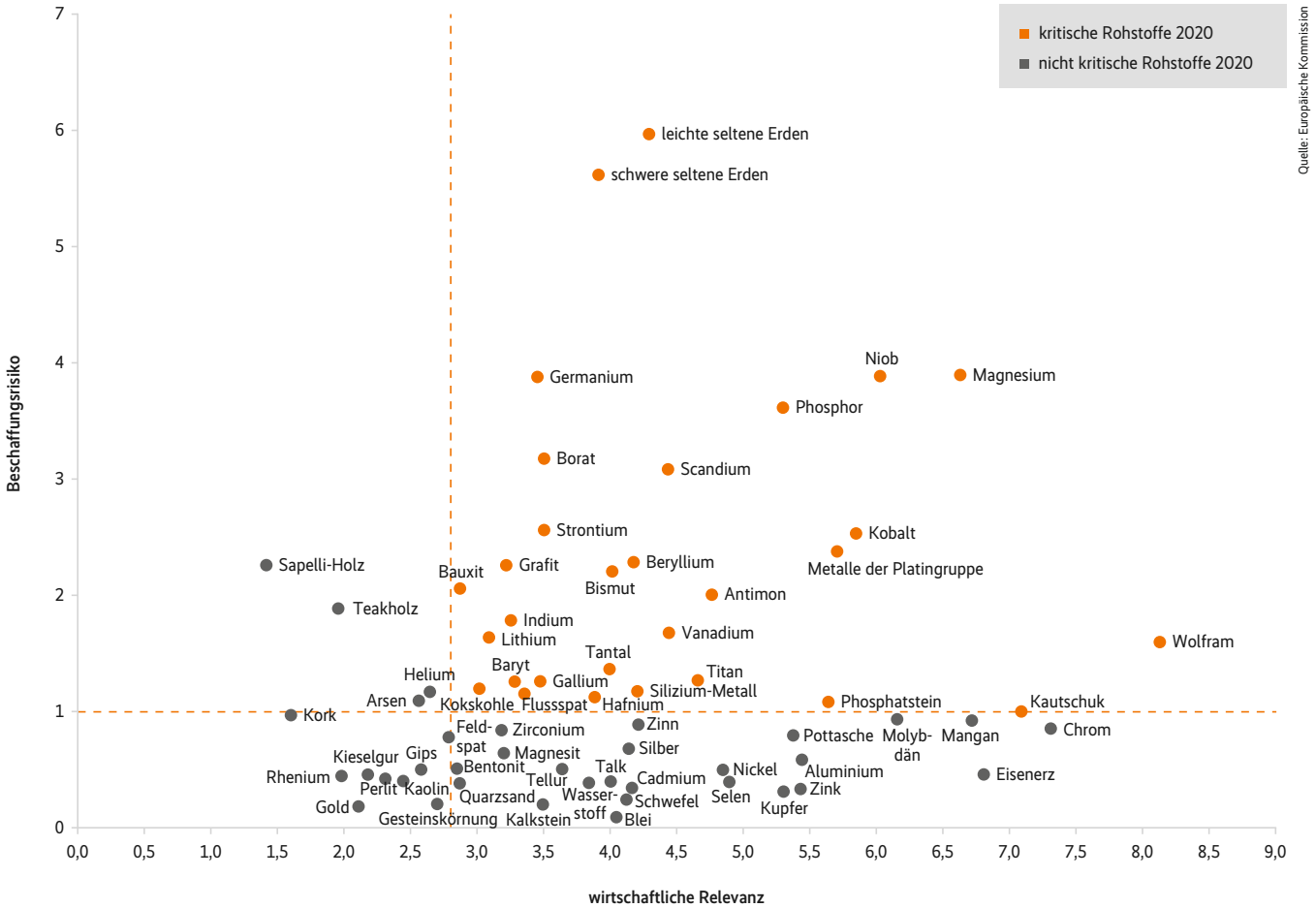
entsprechenden Werte und treiben die Rohstoffnotierungen so noch weiter in die Höhe, was meist weitere Spekulanten anlockt. Noch sind die Spekulanten zurückhaltend, sie fürchten den Hunt-Effekt, denn noch immer sind die Erschütterungen der 70er-Jahre an den Rohstoffbörsen zu spüren. Damals wollten die Brüder Hunt der Welt den Silberpreis diktieren. Die Milliardäre kauften einfach alles auf: reales Silber, aber auch Silberkontrakte, die an Rohstoffbörsen gehandelt werden. Der Silberpreis schnellte 1973 von 2 Dollar auf 50 Dollar hoch. Sie machten immer weiter, auch als ihnen das Geld ausging und sie ihre Kontrakte verpfänden mussten. Am Ende kontrollierten sie mehr als 10 000 t Silber.

Schließlich zog die Börsenaufsicht der COMEX (New York Commodities Exchange) die Reißleine, änderte die Spielregeln und untersagte den Verkauf weiterer Kontrakte. Der Silberpreis fiel rapide. Um von den horrend gestiegenen Preisen zu profitieren, ließen Privatleute massenhaft Silberbesteck einschmelzen. In Deutschland wurde die silberhaltige 5-DM-Münze durch ein Exemplar ohne Silber ersetzt. Die Blase platzte, der Preis stürzte auf unter 10 Dollar und die Brüder Hunt waren pleite.

Auch die vom Silber abhängige Foto- und Filmindustrie reagierte: Kodak, Fuji und Ilford entwickelten neue Filme mit Emulsionen, die deutlich weniger Silber enthielten. Gleichzeitig investierten die Laboratorien in Entsilberungssysteme, die das im Fixierbad gelöste Silber zurückgewinnen – eines der ersten Beispiele für eine global erfolgreiche Kreislaufwirtschaft teurer Rohstoffe.

Fehlende Rohstofflager

Marktbeobachter geben den nachfragenden Konzernen eine Mitschuld an der jetzigen Situation. Während der Zeit der Globalisierung galt Lagerhaltung als Verschwendung von Kapital. Man scheute langfristige Lieferverträge und spekulierte auf noch



Quelle: Europäische Kommission

Versorgungsrisiko und ökonomische Relevanz: Je weiter rechts ein Rohstoff steht, desto bedeutsamer ist er, und je höher, desto unsicherer ist die Versorgung. Alle orange markierten Ressourcen galten 2020 als kritische Rohstoffe (Abb. 4).

Liste der kritischen Rohstoffe		
Antimon	Baryt	Bauxit
Beryllium	Bismut	Borat
Flussspat	Gallium	Germanium
Grafit	Hafnium	Indium
Kautschuk	Kobalt	Kokskohle
Lithium	Magnesium	Niob
Phosphat	Phosphor	Scandium
Silizium-Metall	Strontium	Tantal
Titan	Vanadium	Wolfram
schwere seltene Erden	leichte seltene Erden	Metalle der Platingruppe

niedrigere Preise. Beteiligungen an Minen wurden veräußert. Das wenig lukrative Geschäft mit Basismaterialien überließen die Hightechkonzerne lieber anderen. Mächtige Nachfrager bündelten ihre Kräfte, um den Produzenten ihre Konditionen aufzuzwingen. Ökonomen sprechen von einem Nachfragemarkt. Die Kunden bestimmen bis hin zu „just in time“ die Konditionen.

Dadurch wurden Kapazitäten reduziert und es fand eine Konzentration der Anbieter statt, kurz darauf explodierte die Nachfrage. Inzwischen gilt der Rohstoffmarkt als Musterbeispiel eines Lieferantenmarkts, bei dem die Anbieter die Konditionen diktieren. Heute sind viele Firmen gezwungen, zu stark schwankenden Tagespreisen auf Spotmärkten einzukaufen. Damit sind sie sowohl den Spekulanten als auch der Tagespolitik ausgesetzt.

Bei vielen Rohstoffen mangelt es aber schon am Bestand. Ihr Abbau ist mühsam, die Lieferländer sind oft weit entfernt und politisch instabil, die natürlichen Ressourcen können zur Neige gehen. Neue Abbaugebiete zu finden ist langwierig und mit immensen Kosten verbunden. Zehn bis zwanzig Jahre dauert es, bis eine neue Rohstoffquelle gefunden und erschlossen ist.

Die Versorgungslage ist also immer eine genaue Analyse wert. Seit 2010 beobachtet die auf Erlass des Bundeswirtschaftsministeriums gegründete Deutsche Rohstoffagentur (DERA) in der Bundesanstalt für Geowissenschaften und Rohstoffe 60 Rohstoffe und Handelsprodukte. Gleichzeitig analysiert man die Zukunftstechniken.

Zukunftsbedarf abhängig von der Einsicht

Ein Beispiel: Samarium wurde lange Zeit nur als magnetischer Tonabnehmer für E-Gitarren verwendet, heute sind es Hightechmagnete für Hochtemperaturanwendungen. Andere seltene Erden sind durch die drastisch gestiegene Nachfrage nach Generatoren und elektrischen Antriebseinheiten in den Fokus von Wirtschaft und Politik geraten. Gerade die Digitalisierung stellt die Rohstoffbeschaffer vor Herausforderungen. Deshalb ließ die DERA die Fraunhofer-Institute ISI und IZM den Rohstoffbedarf von 33 Zukunftstechniken für das Jahr 2040 in drei Szenarien ermitteln. Die Ergebnisse liegen in der umfangreichen Studie „Rohstoffe für Zukunftstechnologien 2021“ vor.

Die Autoren wählten zwei extreme Entwicklungen: einen fossilen Pfad und ein nachhaltiges Szenario, bei dem die Dekarbonisierung die oberste Priorität hat. Im mittleren Zukunftsszenario folgt die weitere Entwicklung dem historischen

Derzeit von der EU beobachtete Elemente der Platingruppe und der seltenen Erden

Platingruppenelemente	leichte seltene Erden	schwere seltene Erden	
Platin	Lanthan	Yttrium	Europium
Iridium	Cer	Gadolinium	Lutetium
Palladium	Neodym	Terbium	Ytterbium
Rhodium	Praseodym	Dysprosium	Erbium
Ruthenium	Samarium	Holmium	Thulium

Muster, das bedeutet einen Anstieg im Trend der Digitalisierung. Im nachhaltigen Szenario sind Effizienzsteigerungen digitaler Techniken eingerechnet, die zu Energie- und Rohstoffeinsparungen des Sektors führen. Im fossilen Pfad steigen Digitalisierung und digitaler Konsum ungebremst – mit dem daraus resultierenden hohen Rohstoffbedarf. Aufgrund des Datenwachstums wächst in dem Szenario etwa der Markt für Festplatten weiter, für die man Metalle der Platingruppe wie Platin und Ruthenium benötigt.

Die untersuchten Szenarien sind angelehnt an die Shared Socioeconomic Pathways, die der 5. Sachstandsbericht des Weltklimarates IPCC 2011 definiert hat. Koordiniert hat die Studie DERA-Projektkoordinatorin Viktoriya Tremareva: „Der Rohstoffbedarf für die 33 Zukunftstechnologien in den drei Szenarien unterscheidet sich deutlich, sodass eine generelle Voraussage für das Jahr 2040 nicht möglich ist."

Steigender Bedarf auch bei nachhaltigem Wirtschaften

Im fossilen Szenario ergibt sich zum Beispiel im Jahr 2040 ein neunzehnfacher Bedarf an Ruthenium gegenüber 2018, bei Platin ist es der vierfache Bedarf. Im nachhaltigen Szenario überstieg der Rutheniumbedarf für 2040 die Produktion 2018 um mehr als das Zweifache, bei Platin liegt der Bedarf der betrachteten Branchen unter dem von 2018. In die Ergebnisse nicht eingeflossen ist allerdings der Bedarf anderer Branchen, sodass der Gesamtbedarf gegebenenfalls auch steigen kann.

Die Studienautoren erwarten bei bestimmten Rohstoffen Engpässe, wenn nicht schnell neue Quellen erschlossen werden. Das betrifft neben Platin und Ruthenium für Festplatten auch Metalle, die in Halbleiterprodukten und Glasfasern in winzigen Spuren eingesetzt werden, um deren Eigenschaften zu optimieren. Unverzichtbar sind auch seltene Erden, die zum Beispiel in den Festplattenmagneten zu finden sind. Verglichen mit 2018 müsste sich der Abbau im fossilen Pfad bis 2040 bei den leichten seltenen Erden verdoppeln und bei den schweren seltenen Erden versechsfachen, um den Rohstoffbedarf der betrachteten Branchen zu bedienen.

Die DERA-Rohstoffliste 2021 bewertet mögliche Versorgungsrisiken bei 34 Metallen, 27 Industriemineralen und weiteren Handelsprodukten. Das Ergebnis der aktuellen Erhebung: Fast 45 Prozent der untersuchten Bergwerks-, Raffinade- und Handelsprodukte unterliegen erhöhten Lieferrisiken. Entscheidend für die Industrie ist die Verfügbarkeit der jeweiligen Raffinadeprodukte, also der gereinigten und veredelten Stoffe. „Bei 25 der 27 in der Erhebung untersuchten Raffinadeprodukte dominiert China die Weiterverarbeitung. Dies zeigt die Bedeutung Chinas für die internationale Rohstoffversorgung", erläutert Maren Liedtke, Mitautorin der DERA-Rohstoffliste.

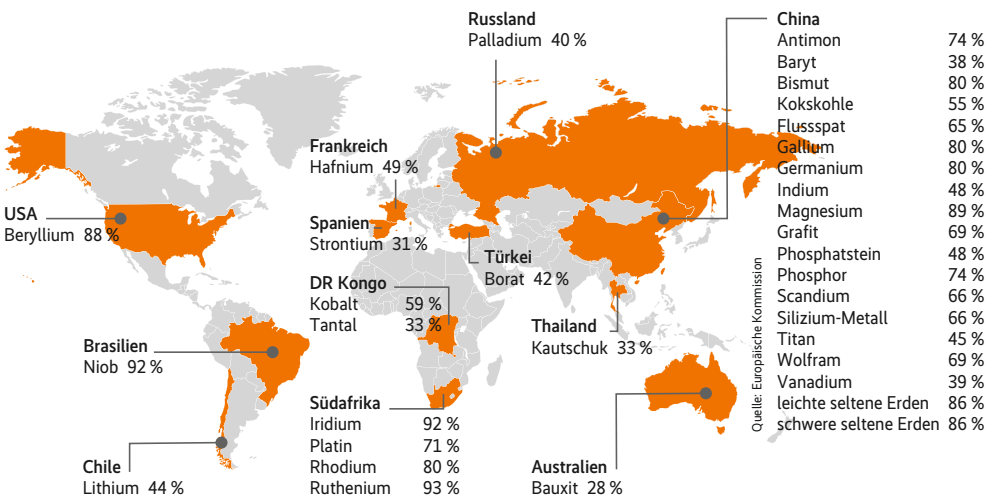
Versorgungsstolperer inklusive

Das Land hat sich in den letzten 30 Jahren eine schier erdrückende Übermacht bei der Versorgung der übrigen Industrienationen mit den begehrten Grundstoffen erarbeitet. Seit dem Beginn des Handelskrieges mit den USA unter Trump liefert China aber längst nicht mehr alles und nicht in jeder Menge. Gründe seien die Energieknappheit und der Eigenbedarf.

Im September 2021 stoppte China die Magnesiumlieferungen mit Hinweis auf den hohen Energiebedarf bei der Produktion. Ohne diesen Stoff lässt sich kein Aluminium herstellen. Der großen Aufregung folgte im November die Entwarnung, doch die Angst vor ähnlichen Schachzügen bleibt. Denn Europa importiert 95 Prozent des benötigten Minerals aus der Volksrepublik; 13 Prozent der Weltproduktion stammen vom russischen Hersteller Rusal. Der Krieg in der Ukraine und die Sanktionen könnten nach Meinung von Experten zu Preisen von über 4000 Dollar pro Tonne führen, was Weltrekord wäre.

Neben Marktmacht und Lagerstätten geht in die Risikobewertung auch die politische Stabilität der Lieferländer ein. Basis ist der Worldwide Governance Indicator der Weltbank, mit dem diese seit 1996 die Regierungssysteme der Welt bewertet. Dass der wichtigste Lieferstaat für das unverzichtbare Kobalt ausgerechnet die Demokratische Republik Kongo ist, die 64 Prozent der weltweit geförderten 118 500 t liefert und zu den politisch instabilsten Ländern gehört, trägt nicht gerade zur Beruhigung der Importeure bei, denn das zweitwichtigste Land ist Russland mit nur 4,6 Prozent. Tagespolitische Überraschungen kann aber auch der Worldwide Governance Indicator nicht abbilden.

In Chile wurde im Dezember 2021 mit Gabriel Boric ein Präsident gewählt, dem Umweltschutz beim Abbau von Rohstoffen am Herzen liegt. Chile ist, neben Peru, das wichtigste Bergbauland



Bei seltenen Erden besteht eine große globale Abhängigkeit. Insbesondere China verfügt über sehr große Vorräte (Abb. 5).

für Kupfer mit einem globalen Anteil von etwa 28 Prozent. Das Metall lässt sich nur durch das ungleich teurere Silber ersetzen und ist nicht in großen Mengen verfügbar. Allerdings wird Kupfer auch in zahlreichen Ländern dieser Welt gefördert. Den jährlich geförderten 20 Millionen Tonnen stehen erschlossene Reserven von 870 Millionen Tonnen gegenüber. Aber auch die befinden sich überwiegend in Chile.

Raritäten in den Rechenzentren

Festplatten speichern die Daten auf einer dünnen magnetischen Kobalt-Chrom-Platin-Schicht. Die Schreib- und Leseköpfe arbeiten mit Neodym-Eisen-Bor-Magneten. In manchen Fällen kommt Ruthenium als nicht magnetische Trennung der Datenspuren zum Einsatz. Neben Ruthenium stehen vor allem Kobalt und Platin auf der Liste der kritischen Ressourcen.

Ein Festplattenmangel infolge einer Verknappung etwa von Platin ist laut DERA trotzdem nicht zu befürchten, dazu seien die erforderlichen Mengen zu gering. Eine rasche Erschließung neuer Lagerstätten sei aber auch nicht möglich. Bleibt den Kunden nur die Hoffnung, die Hauptlieferanten für Platin, Südafrika und Russland, mögen politisch stabil und lieferwillig bleiben. Bei Ruthenium suchen Festplattenhersteller nach Alternativen.

Eine beliebte Alternative sind SSDs. Sie werden nach der SONOS- (Silizium-Oxid-Nitrid-Oxid-Silizium) oder TANOS-Bauweise (Tantalnitrid-Aluminiumoxid-Nitrid-Oxid-Silizium) gefertigt. Ihre Herstellung wird erst dann schwierig, wenn Silizium knapp wird. Relativ entspannt sieht die Situation bei Magnetbändern aus. Ihre magnetischen Beschichtungen enthalten zwar Eisen, Chrom und Strontium oder Barium, die Schichtdicke ist mit 10 nm aber minimal und der Bedarf an Metalloxiden pro Kassette mit 0,7 Gramm winzig.

Die Gesamtschau sieht allerdings weit dramatischer aus, wenn man das ressourcenintensivste Szenarium des fossilen Pfades betrachtet. Die Studie sieht für 2040 bei ungebremster Digitalisierung einen Bedarf von 26 Milliarden Datenträgern weltweit voraus. War der Anteil von Kobalt, Chrom, Platin, Ruthenium, Neodym und Tantal am weltweiten Bedarf im Jahr 2018 zu vernachlässigen, wird er bis 2040 explodieren. Die Hersteller von Speichermedien begnügten sich 2018 mit 180 t Neodym und mit weniger als 100 kg der anderen genannten Elemente. 2040 werden im fossilen Szenario bis zu 1479 t Kobalt, 133 t Chrom, 813 t Platin, 592 t Ruthenium sowie 9220 t Neodym und 649 t Tantal benötigt.

Indium tritt auf

Politik und Märkte müssen entscheiden, wohin das begehrte Neodym wandern soll, oder schnellstmöglich neue Lieferquellen erschließen. Ähnliche Verteilungskämpfe könnten sich um Indium entwickeln. Noch 1924 soll es gerade einmal 1 g des Metalls in isolierter Form gegeben haben. Heute finden sich Hunderte Tonnen Indiumzinnoxid in LCDs.

Displayhersteller werden laut DERA-Studie ihren Bedarf an Indiumzinnoxid von 185 t im Jahr 2018 bis auf 297 t 2040 steigern, die Produzenten von Solarzellen von 22 auf 127 t. Damit würden beide Segmente zusammen mehr als die Hälfte der heute zur Verfügung stehenden Raffinadeproduktion für sich beanspruchen.

5G und Internetausbau werden in den nächsten Jahren den Bedarf an Laserelementen für die optische Kommunikation erhöhen. Sie benötigen Gallium, Indium, Phosphor und Arsen. Zwar gilt derzeit nur Indium als kritischer Rohstoff, doch haben einige Experten auch ein waches Auge auf Gallium, dessen wichtigstes Lieferland China mit 80 Prozent ist.

Hoffnungsträger und Lichtblicke

Doch stecken inzwischen jede Menge der gefragten Rohstoffe in Altgeräten. Laut einer VDMA-Fraunhofer-Studie ist 2030 mit 230 000 t verbrauchten Lithium-Ionen-Batterien zu rechnen. Man könnte sie recyceln und so zumindest einen Teil der Rohstoffe zurückgewinnen. In jeder Fahrzeugbatterie stecken wiederverwertbare Rohstoffe im Wert von 600 bis 1300 Euro. Urban Mining ist zwar schon etliche Jahre alt, doch bleibt es die Ausnahme – ebenso wie der geschlossene Rohstoffkreislauf.

In den Mülleimern der westlichen Welt landen zwar Unmengen wertvoller Rohstoffe, sie zurückzugewinnen ist jedoch teuer und mit erheblichem Aufwand verbunden. Und ob die so gewonnenen Stoffe den hohen Qualitätsanforderungen der Produzenten gerecht werden, ist oft zweifelhaft.

Nur beim Kupfer konnte mit 70 Prozent eine recht hohe Recyclingrate erreicht werden. Allerdings ist hier der Bedarf auch besonders groß und die Gewinnung durch Demontage von Kupferleitungen oder Zerlegen von Kabeln recht einfach. Weltweit werden bis 2040 vier bis acht Millionen Tonnen Kupfer allein für neue Stromleitungen gebraucht.

Neben vielen Risiken sehen die europäischen Rohstoffspezialisten auch Lichtblicke. Bei Quantencomputern ist der Bedarf an Rohstoffen mäßig. Im Jahr 2040 werden für alle Quantencomputer zusammen maximal 240 kg Kupfer benötigt – bei einer Produktion von 24 Millionen Tonnen eine vernachlässigbare Menge.

Fazit

In 20 Jahren ungezügelter Globalisierung galten Vorratshaltung und Kreislaufwirtschaft als Zeit- und Geldverschwendung. Transportkapazitäten und Waren schienen unbegrenzt vorhanden zu sein. Doch schon die Studie „Global 2000 Bericht an den Präsidenten“ belegt, dass Ressourcen endlich sind und nicht immer dort bereitstehen, wo Hightechproduzenten sie benötigen.

Deshalb ist es auch unwahrscheinlich, dass bei einem „Weiter so“ zukünftig alle Zutaten für die Hightechprodukte in ausreichender Menge zur Verfügung stehen werden. Realistischer wäre es, unerwartete Schwierigkeiten einzurechnen. 2017 rechnete niemand mit einer Pandemie, die die Lieferketten durcheinanderwirbelt, oder dass die sechstägige Blockade des Suez-Kanals durch den 400-Meter-Riesen Ever Given weltweit den Reedern die Schweißperlen auf die Stirn treiben würde.

Vom Krieg in der Ukraine ganz zu schweigen. Noch vor dem Inkrafttreten von Sanktionen gegen Russland meldete das ifo Institut der Deutschen Wirtschaft im Februar, dass 74,6 Prozent der Firmen über Engpässe und Schwierigkeiten bei der Beschaffung von Vorprodukten und Rohstoffen klagten. Dass zukünftige Ereignisse eine Besserung herbeiführen, ist leider unwahrscheinlich. (sun@ix.de)

Bernd Schöne

ist freier Journalist.

AMD RYZEN™ PRO – OFFEN FÜR GUTE ZUSAMMENARBEIT

Die effiziente Verwaltung von PCs ist eine der wichtigsten Aufgaben der Unternehmens-IT. Die Admins müssen Netzwerke aufbauen, verwalten und sichern, die mit der steigenden Zahl der Mitarbeitenden und Gerätetypen Schritt halten können. Ein neues Konzept erlaubt es den IT-Teams, dieses Ziel ohne Mehraufwand zu erreichen, obwohl ihre Arbeit ständig umfangreicher und komplexer wird.

Die IT braucht neue Wege, um den gesamten Hardware-Lebenszyklus verwalten zu können. Systeme müssen abgebildet, bereitgestellt und implementiert werden. Die Verantwortlichen müssen dafür sorgen, dass sich die Systeme über Updates auf dem neuesten Stand befinden – von Firmware und Anwendungen bis hin zu Viren-Signaturen und anderen Sicherheitskorrekturen. Auch wenn alle Geräte pausenlos laufen, muss die IT-Abteilung in der Lage sein, den Anlagenbestand und den Systemzustand zu verfolgen.

Eine schwere Aufgabe, selbst an nur einem Standort. Mehrere Standorte und Mitarbeitende im Homeoffice können die Schwere der Herausforderung schnell vervielfachen, und jeder möchte Probleme so schnell wie möglich lösen. Damit die IT-Abteilung das leisten kann, benötigt sie eine flexible Plattform mit einer zentralen Verwaltung aller Systeme ¬– sowohl In- als auch Out-of-Band. Und sie braucht eine einheitliche Reihe von Tools, die unabhängig vom Hersteller des Prozessors oder des OEM-Geräts gute Dienste leisten.

Ausweg aus der Zwickmühle

Mit proprietären Werkzeugen ist so eine Mammutaufgabe kaum zu stemmen. Nur eine herstellerübergreifende Zusammenarbeit mit offenen Standards führt aus diesem Dilemma heraus. Darum engagiert sich AMD in der Distributed Management Task Force (DMTF). Das ist eine gemeinnützige Vereinigung, die sich der Förderung von Systemverwaltung und Interoperabilität durch Entwicklung von Standards verschrieben hat. Die DMTF wird dabei in der Branche breit unterstützt. Zu den über 200 Mitgliedern gehören auch andere wichtige Chip-Anbieter wie Intel, Nvidia, Qualcomm und ARM sowie führende Unternehmens-OEMs wie HP, Lenovo oder Dell.

Zu den wichtigsten Standards, die unter dem Dach der DMTF entstanden, zählt DASH (Desktop and mobile Architecture for System Hardware). Er ermöglicht die sichere Fernverwaltung (einschließlich der Out-of-Band-Verwaltung) von Desktops und mobilen Systemen unterschiedlicher Erstausrüster – sowohl für AMD- als auch für Intel-Prozessoren.

Der Standard definiert einen gemeinsamen Rahmen für die Verwaltung der meisten Out-of-Band-Verwaltungsaufgaben:

- Fernsteuerung der Stromversorgung
- Boot-Kontrolle, Patching
- Ferndiagnose
- Inventarisierung
- Sicherheit

Neben der Festlegung von Standards arbeitet AMD auch eng mit seinen Partnern in der DMTF an der Implementierung von Tools, die die Arbeit von Admins erleichtern. So sind beispielsweise aus unserem Engagement beim DASH-Standard ein Konsolen-SDK (Software Development Kit) und eine Referenzimplementierung (AMD Management Console, AMC) hervorgegangen, die kostenlos als Open-Source-Tools weitergegeben werden.

Das bietet der offene DASH-Standard

- DASH schafft langfristige Stabilität und fördert die Interoperabilität zwischen den Lösungen verschiedener Anbieter. Client-PCs, die den DASH-Standard nutzen und damit wesentliche Funktionen ermöglichen, sorgen für mehr Auswahlmöglichkeiten bei einem breiteren Spektrum von Anbietern.
- IT-Abteilungen können damit flexibler auf sich verändernde geschäftliche Anforderungen reagieren, ohne ihre Umgebung unnötig komplexer zu machen.
- Standardbasierte Anwendungen und Technologien senken die Verwaltungskosten durch die Vereinheitlichung von Tools und die Vereinfachung von Aufgaben.
- Mithilfe von Standards können sich IT-Abteilungen darauf konzentrieren, geschäftliche Anforderungen zu erfüllen, statt sich mit proprietären Tools für die Verwaltung spezifischer Systeme herumschlagen zu müssen.
- Anbieter profitieren von der Flexibilität des offenen

DASH-Standards, denn sie können ein breiteres Angebot an Lösungen mit einem unterschiedlichen Maß an Funktionen bereitstellen.

Diese DASH-Tools bietet AMD an

- AMD Management Plugin für SCCM (AMPS) – AMPS erweitert den Microsoft System Center Configuration Manager (SCCM) und den Microsoft Endpoint Configuration Manager (MECM). Damit lässt sich DASH in bestehende Verwaltungssysteme einbinden.
- AMD Management Console (AMC) – AMC ist eine eigenständige GUI-Anwendung, die auf die Umgebung kleiner Unternehmen abzielt. Mit ihr lassen sich bis zu 500 DMTF-DASH-kompatible Clients verwalten. AMC unterstützt auch KVM-Umleitung für grafikbasierte BIOS-Set-up-Bildschirme.
- AMD DASH Command Line Interface (CLI) – Mit dem AMD DASH CLI können DASH-Ziele Out-of-Band via Shell verwaltet werden. Damit ist die Integration in Workflows möglich.

Zusammenfassung

Die AMD-PRO-Plattform bietet zusammen mit DASH ideale Voraussetzungen, um einen PC-Gerätepark verwalten und produktiv halten zu können – unabhängig von der Größe. DASH nutzt einen offenen, dem Industriestandard entsprechenden Satz von Zugriffsprotokollen für die Verwaltung von OOB-Desktop- und mobilen Clients. So gelingt es, wichtige Verwaltungsaufgaben wie Fernsteuerung der Stromversorgung, Updates und Patches, Ferndiagnose sowie Bestandsaufnahme sicher durchzuführen. Da es sich bei DASH um ein erweiterbares Management-Framework handelt, kann es mit der Unterstützung neuer Funktionen zum Client-Management wachsen.

AMD PRO Manageability baut auf diesem Versprechen auf und stellt Unternehmen wesentliche Verwaltbarkeitsfunktionen bereit, mit denen sie heute und morgen Geschäftsziele einfach und flexibel unterstützen können. Mit einem Design nach offenen Standards, das nicht nur innerhalb einer bestehenden Umgebung funktioniert, sondern auch echte Wahlfreiheit ermöglicht, bieten AMD-Ryzen™-Prozessoren mit AMD-PRO-Technologien gewerblichen Kunden eine moderne Performance. Kunden erhalten die Sicherheits- und Verwaltungsfunktionen, die sie in ihren anspruchsvollen Geschäfts- und Technologieumgebungen benötigten.

Wichtige Anwendungsfälle für DASH

ANWENDUNGSFALL	VORTEILE
FERNSTEUERUNG DER STROM-VERSORGUNG	• Verwaltet den Stromverbrauch • Hilft, Stromkosten zu senken und den Energieverbrauch zu steuern • Hilft bei der Verwaltung von Emissionszertifikaten • Hilft aktiv bei der Steigerung der Produktivität • Sie können sich auf die Erfüllung von Geschäftsanforderungen konzentrieren
PATCH-SATURIERUNG BESCHLEUNIGEN	• Patch-Saturierung schneller erreichen • Aktualisieren Sie jedes DASH-fähige System, unabhängig vorn Hersteller • Standort, Systemstatus und Energiestatus sind irrelevant
REMOTE-DIAGNOSE UND RE-IMAGING	• Hilft, die Notwendigkeit von Besuchen am Schreibtisch zu reduzieren • Fehlerbehebung und Reparatur aus der Ferne • Hilft, Besuche am Schreibtisch zu verringern oder sogar zu vermeiden • Hilft, Ausfallzeiten für Benutzer zu verringern und die Produktivität zu optimieren
AUDITING – BESTANDSAUFNAHME	• Sammeln von Informationen unabhängig vom Client-System oder Stromversorgungsstatus • Hilft, manuelle Plattform-Audits zu verringern oder zu eliminieren • Inventarinformationen an einer zentralen Stelle speichern • Hilft, Softwarelizenzen effizient und effektiv zu verwalten • Hilft die Compliance mit Geschäftspraktiken zu verbessern

Hier erfahren Sie mehr zum Thema

DMTF: https://www.dmtf.org/
DASH: https://www.dmtf.org/standards/dash
Tools für DMTF DASH: https://developer.amd.com/tools-for-dmtf-dash/
AMD PRO Manageability: https://www.amd.com/de/technologies/security-manageability
AMD-PRO-Technologien: https://www.amd.com/de/technologies/pro-technologies

Elektroschrott vermeiden

Ohne Ende

Ariane Rüdiger

Noch ist die Elektronikbranche weit von einer echten Kreislaufwirtschaft entfernt. Selbst dass Elektroschrott auch nur legal und umweltgerecht entsorgt wird, ist nicht immer garantiert. Bleibt als einziger Ausweg, möglichst wenig davon zu erzeugen.

Was mit dem ausrangierten IT-Equipment passieren soll, hat für IT-Manager bislang kaum Bedeutung. Doch nun wächst der Druck auf Unternehmen, nachhaltiger zu werden; das betrifft auch die IT. Möglichst geschlossene, emissionsarme Methoden der Leistungserbringung sind deshalb aus Umwelt- und wirtschaftlichen Erwägungen das Gebot der Stunde.

Das von den Herstellern gern vorgetragene Argument, der Löwenanteil der Treibhauslast der IT entstünde während der Nutzung, ist längst obsolet. Neuere Studien kommen zu ganz anderen Ergebnissen (alle Quellen siehe ix.de/znur). Danach werden bis zu 75 Prozent der Kohlendioxid-Last eines IT-Produktes schon bei seiner Produktion in die Umwelt entlassen (siehe Abbildung 1).

Das zeigt auch Fujitsus Analyse eines vergleichsweise umweltfreundlich hergestellten Produkts, eines Esprimo-Rechners (siehe Abbildung 2). Etwas mehr als 400 kg seines CO_2-Fußabdrucks von etwas über 700 kg Kohlendioxid entstehen während der Produktion und 60 Prozent des Energieaufwands fallen allein für die Produktion und das Inverkehrbringen des Geräts an.

Das liegt unter anderem an den vielen Stoffen, die in der Hardware stecken – die Materialliste reicht quer durchs Periodensystem (siehe Artikel „Nicht in den Himmel“ ab Seite 22). Die globalen Produktionsketten enthalten viele Langstreckentransporte, der Abbau vieler Grund- und Funktionsstoffe geht oft einher mit Umweltzerstörungen und Menschenrechtsverletzungen, die Produktion mit viel Chemie und hohem Energieaufwand.

Recyclingquoten noch zu niedrig

Und wie sieht es mit dem Recycling aus? Nach neueren Statistiken von Eurostat ist das kleine Liechtenstein europäischer E-Waste-Recyclingmeister: Dort werden knapp 90 Prozent des Elektroschrotts regulär entsorgt. Deutschland bewegt sich leicht unterhalb des Durchschnitts der EU-Mitgliedsstaaten, der bei weniger als 40 Prozent liegt. Das ist kein Ruhmesblatt, sondern ein Armutszeugnis, denn der Wert sollte bereits bei 65 Prozent liegen.

Ausführliches Zahlenmaterial liefert eine Studie der UNU (United Nations University) und des UNCTAR, des UN-Instituts für Bildung (siehe ix.de/znur). Sie wertet Daten aus dem Jahr 2018 aus und betrachtet zwei Zahlen: POM (Placed on Market) sind die auf den Markt gebrachten EEE-Produkte (Electrical and Electronic Equipment) und WEEE (Waste of Electrical and Electronic Equipment) die generierten Abfälle. Das POM-Sammelziel liegt bei 65, das für E-Waste bei 85 Prozent. Erreicht werden beide nicht. Wie weit die jeweiligen Länder oder Regionen hinter den POM-Sammelzielen zurückbleiben, zeigt Abbildung 3. Zwischen privat und beruflich genutzten Produkten differenziert die Studie nicht, auch nicht zwischen IT- und anderen elektronischen Geräten.

Aufgeschlüsselt hat die Studie auch, wohin die Stoffströme fließen, die sich außerhalb geregelter und gesetzeskonformer

Recyclingkreisläufe bewegen: 2,1 kg pro Einwohner und Jahr landen im Metallabfall und werden dort eventuell wiederverwertet, 1,4 kg landen im Hausmüll und sind damit fürs Recycling erst einmal verloren, ein bis zwei Kilogramm werden zu Zwecken des Wiedereinsatzes exportiert und 0,5 bis 1,4 kg verlassen auf illegalen Wegen die EU (siehe Abbildung 4). Die großen Unsicherheiten bei den letztgenannten Kategorien zeigen bereits, dass möglicherweise viele EEE-Güter, die als gebrauchsfähig gelabelt werden, Elektroschrott sind und eigentlich in Europa entsorgt werden müssten. Die jährliche Elektroschrottmenge schwankt je nach europäischer Region zwischen circa 12 kg pro Person in Osteuropa und etwa 23 kg pro Person im digitalisierten Nordeuropa. Westeuropa, zu dem Deutschland gehört, landet bei etwas über 20 kg pro Person.

Zahlen zu den Anteilen des IT-Schrotts am E-Waste gibt es beispielsweise von Eurostat: 2020 bestanden 14,2 Prozent des Elektroschrotts aus Consumer- und professionellem IT- und TK-Schrott. Destatis liefert für Deutschland und 2018 folgende Werte: 853 100 t Elektro- und Elektronikaltgeräte, davon machte die IKT 125 000 t aus. Zwischen professioneller und Consumer-IKT wird nicht unterschieden.

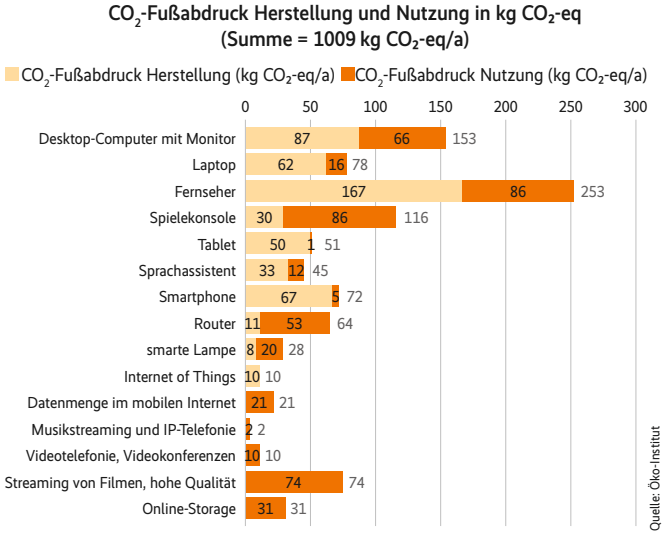
Recycling formell und informell

Wie düster es beim End-of-Life-Recycling in Deutschland aussieht, offenbart auch ein Forschungspapier von Barbara Reck. Danach erreichen zwar die wichtigsten Metalle aus Elektronik und IT-Equipment, etwa Eisen, Kupfer, Silber, Aluminium, Recyclingraten von über 50 Prozent. Damit ist aber noch nichts darüber gesagt, wie viel von den Materialien, die ins Recycling gegeben werden, tatsächlich am Ende brauchbar wieder herauskommen und den Rohstoffbedarf tatsächlich verringern.

Im Idealfall trennt man beim Recycling das Brauchbare vom Unbrauchbaren, schlachtet das teilweise Brauchbare aus und baut die brauchbaren Komponenten aus größtenteils unbrauchbaren Geräten in größtenteils brauchbare ein. Dann folgt die möglichst sortenreine Zerlegung des Rests zumindest teilweise in Handarbeit und das Inverkehrbringen der nunmehr wieder intakten Geräte oder Schrottfraktionen.

Zu den Verwertern der auf diese Weise nicht weiter nutzbaren Reste gehören spezialisierte Aufbereitungsunternehmen wie Umicore. Das belgische Unternehmen unterhält eine große Wiedergewinnungsanlage unter anderem mit einem Schmelzofen in Hoboken bei Antwerpen und bietet eine ganze Reihe von Metallen als Sekundärrohstoffe an. Dort werden die nicht verwertbaren Reste geschreddert, verwertbare Metalle chemisch-physikalisch abgesondert, mit unterschiedlichen Methoden behandelt, bis sie die gewünschte Reinheit haben, und als Sekundärrohstoffe wieder in den Handel gebracht. Der Rest landet in der Schlacke, die bis heute nicht wiederverwertet wird. Allerdings könnte es durchaus sein, dass sie im Rahmen des Urban Mining, wie die Wiedergewinnung von Rohstoffen in Großstädten genannt wird, als Ressourcenquelle entdeckt wird.

Läuft es schlecht, werden eigentlich nicht mehr funktionsfähige Geräte als funktionsfähig deklariert und ausgeführt, um auf riesigen Schrotthalden wie Olsosun in Nigeria zu landen. Dort werden die Geräte oder Reste in informeller Arbeit ohne jede Sicherheitsvorkehrung auseinandergenommen, Kabelhüllen auf offenem Feuer verbrannt und Kupferdraht oder andere wertvolle Bestandteile wieder in den Handel gebracht (siehe ix.de/znur). Die extrem gesundheitsschädliche Handarbeit bringt für ein Kilogramm Aluminium laut UNEP (United Nations Environmental Program) 0,61 US-Dollar, für ein Kilogramm Messing 1,94 Dollar und für Kupfer immerhin 4,17 Dollar pro Kilogramm. Große Hersteller unterhalten inzwischen auch eigene Recyclingfabriken in Nigeria und anderen IT-Schrott-Hotspots, um die Grundlage für geordnete Recyclingprozesse zu legen und den Ruf der Branche wiederherzustellen.



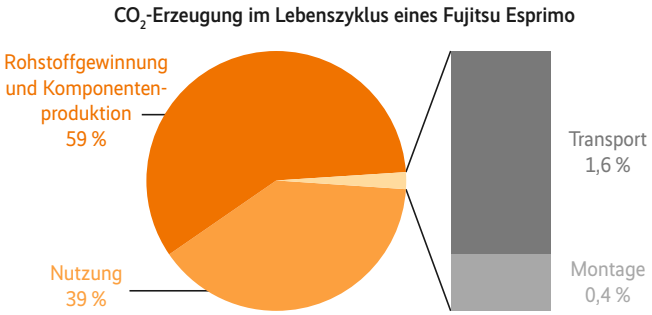
Ein großer Teil des CO_2-Fußabdrucks entsteht bereits bei der Herstellung der Geräte (Abb. 1).

Recyclingverfahren fehlen: seltene Erden und mehr

Viele nur in geringen Mengen vorhandene Elemente, besonders die seltenen Erden, werden allerdings so gut wie gar nicht wiederverwertet. Von den vielen Seltenerdmetallen landen im Durchschnitt weit mehr als 90 Prozent feinverteilt in der Umwelt und sind damit nicht mehr zurückzugewinnen. Man spricht hier von dissipativen Verlusten. Für das Rückgewinnen solcher Spuren aus den Geräten, aus Müllhalden oder gar aus der Umwelt hat man schlicht noch keine Verfahren (siehe Kasten „Seltene Erden und ihr Recycling“). Das ist insofern ökonomisch fragwürdig, als es die Abhängigkeit der IT- und der Hightechbranche von bestimmten Lieferanten zementiert

iX-TRACT

- In Deutschland fallen pro Bürger und Jahr derzeit ungefähr 20 kg Elektroschrott an.
- Ein halbes bis knapp eineinhalb Kilogramm Elektroschrott pro EU-Bürger und Jahr verlassen auf illegalen Wegen die EU und werden auf nicht genehmigten Abraumhalden ungeregelt und auf umwelt- und gesundheitsschädliche Weise entsorgt oder recycelt.
- Viele Materialien lassen sich nicht einfach aus alten IT-Komponenten und -geräten zurückgewinnen, da das bislang komplexe und teure Prozesse erfordert.
- Alternativen zum schnellen Austausch von IT-Hardware sind die verlängerte Inhouse-Nutzungsdauer oder der Kauf gebrauchter Hardware, die qualifizierte Refurbisher aufarbeiten und mit Gewährleistung verkaufen.



Quelle: Fujitsu

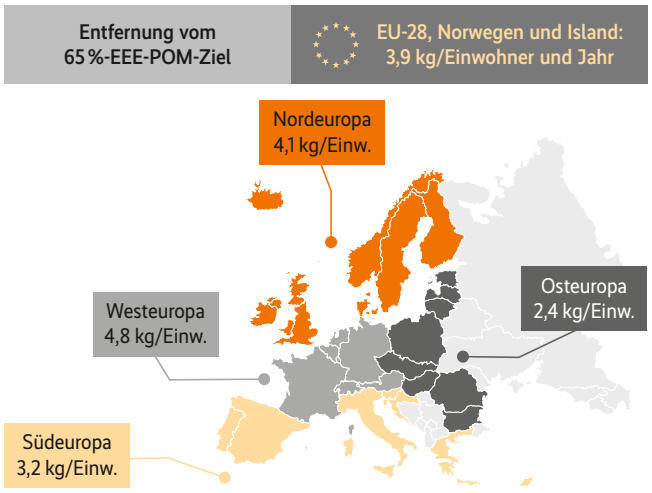
Der größte Anteil der CO_2-Erzeugung während der Fertigung eines Fujitsu Esprimo entfällt auf die Gewinnung der Rohstoffe (Abb. 2).

(siehe die Artikel „Zerreißprobe“ und „Nicht in den Himmel“ ab Seite 16 und 22).

Wie viel der jeweiligen Stoffe die einzelnen Elektrogerätesektoren benötigen, zeigt die EU-Studie „Report on Critical Raw Materials and the Circular Economy“ (siehe Abbildung 5). Dabei bleiben indirekte Zuflüsse wie Kabelummantelungen unberücksichtigt, da nicht alle Kabel für IT-Systeme verwendet werden. Insbesondere Produzenten kleiner, starker Magnete für Festplatten oder die Mobiltechnik sind auf Stoffe wie Indium, Gallium, Neodym, Dysprosium et cetera angewiesen.

Nur bei Silber und Blei stammt die Hälfte aus dem Recycling. Bei Zinn, Kobalt, Eisen, Nickel und Zink sind es 30 bis 40 Prozent, bei Gold 20 Prozent. Platin- und andere Metalle stammen zu weniger als 20 Prozent aus der Wiedergewinnung, viele Funktionsmetalle und seltene Erden wie Gallium, Germanium und Indium stammen fast ausschließlich aus der Primärproduktion (siehe Kasten „Materialbilanz am Beispiel von Gallium“).

Industrie und Forschung versuchen, Recyclingkreisläufe zu entwickeln, die kritischen Materialien durch andere zu substituieren, ihre Einsatzmengen in den jeweiligen Produkten zu verringern oder Produkte und Dienstleistungen zu ersinnen, die dasselbe leisten, aber keine oder kaum kritische Materialien enthalten. Die EU fördert das mit ihrer Circular-Economy-Strategie. Beispielsweise machen Halbleiterspeicher die Magneten in den bisherigen Festplatten obsolet. Außerdem experimentiert man mit neuen magnetverstärkenden Techniken, die seltene Erden ersetzen könnten.



Quelle: UNCTAD, UNU

Europa blieb 2018 weit unter den angestrebten Recyclingraten für Elektroschrott, gemessen in Kilogramm pro Person und Jahr, die fehlen, um die angestrebte Recyclingrate von 65 Prozent bei auf den Markt gebrachten Produkten zu erreichen (Abb. 3).

Eigene IT: so wenig und so umwelteffizient wie möglich

Doch was können IT-Abteilungen tun? Denkbar sind in der Reihenfolge ihrer Umwelteffektivität das weitgehende Einsparen eigener IT-Infrastruktur, nachhaltige Beschaffung, Verwendung von Second-Use-Geräten, längere Nutzung – auch einzelner Komponenten – und dann erst das klassische Recycling.

Reduzieren kann man die eigene IT heute vor allem durch den Griff zu Cloud-Techniken. Mit einer Private Cloud lässt sich die eigene IT konsolidieren und die Auslastung erhöhen, mit dem Griff zu Public-Cloud-Services verlagert man den Hardwareeinsatz und den Energieverbrauch an die Provider. Was auf den ersten Blick nach Greenwashing klingt, hat insofern einen Einsparungseffekt, als die Cloud-Hardware idealerweise weitaus besser ausgelastet ist als ausschließlich selbst genutzte. Deren Auslastung liegt nach aktuellen Auskünften aus dem Umweltbundesamt auch bei gut geführten und weitgehend virtualisierten Infrastrukturen bei kaum über 20 Prozent, einstellige Auslastungsraten sind keine Seltenheit. Dennoch spielen bei der Wahl zwischen Provider und on Premises viele Faktoren eine Rolle.

In Zukunft soll es aber einfacher sein, beim Einkauf von RZ-Dienstleistungen oder Cloud-Services auf Umweltfreundlichkeit zu achten. Metriken, die die Umweltwirkungen eines bestimmten Cloud-Service wiedergeben, sind in Arbeit (siehe Artikel „Flocke für Flocke“ ab Seite 92). Entstehen sollen RZ-Kataster, aus denen Anwender alle wichtigen Informationen auch über Umweltwirkungen infrage kommender Provider beziehen können. Doch das wird wohl noch drei bis fünf Jahre dauern. Bis dahin empfiehlt es sich, auf Zertifikate zu achten und nachzufragen (siehe Artikel „Dickicht der anderen Art“ ab Seite 100).

In jedem Fall zu empfehlen ist der Kauf möglichst umweltfreundlicher Hardware. Energiebedarf und Effizienz sollten beim Kauf ebenso berücksichtigt werden wie die lange Nutzbarkeit, Nachrüstbarkeit oder die Möglichkeit, bei Upgrades alte Komponenten in der neuen Umgebung weiterzuverwenden. Beispielsweise hat eine mangelnde Performance meist eine leicht behebbare Ursache, etwa in Gestalt eines in der Standardkonfiguration viel zu klein dimensionierten Arbeitsspeichers.

Zahlreiche Stellschrauben für Nachhaltigkeit

Helfen kann auch die Modularität: Wer seine IT nach Bedarf auf- und abbauen kann, muss selten Ressourcen überprovisionieren, die, wenn sie endlich eingesetzt werden, veraltet sind. Beim Einkauf berücksichtigen oder vorschreiben lassen sich auch Labels wie TCO Certified oder Energy Star. TCO befasste sich anfangs mit Endgeräten wie Bildschirmen oder Headsets und prüfte vorwiegend ergonomische Kriterien. Heute werden aber auch Server in die Zertifizierung einbezogen, die auch Kriterien der Nachhaltigkeit umfasst.

Dazu zählen die Forderung an die Hersteller, Metalle aus nachhaltigen, verantwortlichen Quellen zu beziehen, die Nutzung eines gewissen Anteils an recyceltem Material, beispielsweise beim Gehäuse, die Offenlegung der Methoden der CO_2-Bilanzierung bei der Geräteproduktion oder die Nennung der wiederverwendbaren Anteile der Geräte nach der Nutzungsphase, Langlebigkeit, eine Mindestgarantiedauer und anderes

Seltene Erden und ihr Recycling

Seltene Erden, genauer gesagt die Metalle der seltenen Erden, sind nicht etwa selten, sondern seltsam, das heißt, sie haben teils ungewöhnliche Eigenschaften, beispielsweise können sie den Magnetismus von Eisenmagneten erheblich verstärken. Zudem kommen sie niemals allein vor, sondern sind immer mit anderen Metallen vergesellschaftet, darunter strahlende wie Thorium oder Uran. Aufgrund der komplexen chemisch-physikalischen Prozesse bei ihrer Trennung vom übrigen Erz und des strahlenden Abfalls, den ihre Produktion häufig zurücklässt, ist ihre Gewinnung meist sehr umweltschädlich.

Der Großteil gerade der seltener vorkommenden schweren Seltenerdmetalle stammt aus China. Das entsprechende Know-how ist in Europa und den USA teils nicht aufgebaut worden, teils durch die Auslagerung der Rohstoffproduktionsprozesse nach Asien gewandert. Schwankende Weltmarktpreise gefährden viele Investitionen. Daraus resultiert eine hohe Rohstoffabhängigkeit von China. Mehr zu Bedeutung, Förderung, Verarbeitung, Einsatzfeldern, Recycling und Substitution von Seltenerdmetallen liefert das Werk „Seltene Erden: Umkämpfte Rohstoffe des Hightech-Zeitalters“ von Luitgard Marschall und Heike Holdinghausen (siehe die Buchrezensionen „Fit im Kopf“ ab Seite 151).

Derzeit gibt es einige Projekte, die sich mit dem Recycling seltener Erden und anderer Funktionsmetalle beispielsweise aus Permanentmagneten befassen (siehe ix.de/znur). Ziel ist es, die in Abraumhalden und Altprodukten gebundenen Materialien durch Recycling oder Urban Mining zurückzugewinnen.

An der TU Clausthal, Lehrstuhl für Rohstoffaufbereitung und Recycling, arbeitet das Projekt **Semarec** (Seltenerd-Magneten-Recycling) an Verfahren zur Wiedergewinnung von Neodym aus Neodym-Eisen-Bor-Magneten. Sie stecken etwa in Elektro- und Servomotoren. Zunächst versucht das Projekt, geeignete Stoffströme herauszufinden. Kooperationspartner ist die Siemens AG.

Das Fraunhofer-Leitprojekt **Seltene Erden** will die Abhängigkeit Deutschlands und Europas von dieser Stoffklasse insgesamt verringern. Zum Projekt gehört auch der Bereich **Design for Recycling,** in dem die Forschenden die Recyclingmöglichkeiten von Komponenten und Materialien aus Elektromotoren untersuchen.

Das Horizon-2020-Forschungsprojekt **Susmagpro** der EU will eine krisensichere, auf Recycling basierende Lieferkette für Seltenerdmagneten in Europa aufbauen. Koordiniert von der Hochschule Pforzheim, beteiligen sich 19 europäische Partner aus Wissenschaft und Industrie.

Das von der EU geförderte Projekt **ADIR** am Institut für Lasertechnik der RWTH hat neue, lasergestützte Verfahren entwickelt, um Metalle wie Tantal, Wolfram und seltene Erden wie Neodym aus Elektroschrott zurückzugewinnen. Es setzt auf die Materialhandhabung durch Roboter, hochauflösende Bildgebungsverfahren, lasergestütztes Entlöten, 3-D-Lasermessung und Lasermaterialidentifikation, automatische Sortierung von Modulen und metallurgische Behandlung der einzelnen Stofffraktionen.

Das Fraunhofer-Institut für Fertigungstechnik und Angewandte Materialforschung IFAM hat ein **hydrometallurgisches Verfahren** mit vorherigem physikalisch-mechanischem Aufschluss zur Wiedergewinnung seltener Erden entwickelt.

Die Ludwig-Maximilians-Universität in München und andere arbeiten an **bakteriellen Verfahren** zur Wiedergewinnung seltener Erden aus Elektroschrott. Man hat entdeckt, dass bestimmte Bakterien mit Vorliebe bestimmte seltene Erden verzehren, weil sie für diverse biologische Prozesse in ihnen notwendig sind. Bringt man solche Bakterien in Behälter mit entsprechend vorbereitetem Material, reichern sie die gewünschten Materialien an. Bei der Ernte entzieht man ihnen das entsprechende Element chemisch. Bislang funktioniert das im Labor mit Lanthan.

Das KIT entwickelt im Rahmen der Digitalisierungs-Landesstrategie von Baden-Württemberg eine energie- und abfalloptimierte Ultraeffizienz-Fabrik. Dazu gehört ein Roboter, der selbstständig Elektromotoren auseinandernehmen soll, um die darin enthaltenen Komponenten und Materialien für die Weiternutzung oder das Recycling zu erschließen.

mehr. Außerdem wird Verantwortlichkeit für die Lieferkette und die Einhaltung grundlegender Arbeitsrechte in den Fabriken verlangt.

Schließlich müssen Hersteller, die das Label anstreben, auch ihre Entsorgungsmethodik und -kette offenlegen. Diejenigen, die all dies am Anfang nicht bewältigen, können sich einer vollen Zertifizierung schrittweise annähern, solange sie jährlich Verbesserungen nachweisen. Mit der TCO-Zertifizierung RZ-tauglicher Systeme sieht es derzeit aber noch mau aus. Die Kategorien Netzwerkequipment und Storage sind im Internetverzeichnis der zertifizierten Systeme leer, von seinen Servern hat HPE eine Reihe die Zertifizierung durchlaufen lassen. Größer ist die Auswahl bei Desktops, Laptops oder Projektoren (siehe ix.de/znur).

Second Use von Hyperscaler-Equipment

Einen weiteren Ansatz liefert das Open Compute Project (OCP). Sein von Hyperscalern inspiriertes Ziel ist standardisierte, hocheffiziente Hardware mit entsprechenden Schnittstellen. Sie soll die Anwender von Herstellerbindungen befreien, weil jede OCP-Hardware zumindest im Ideal zu jeder anderen passt, unabhängig vom Hersteller.

Das Start-up IT Renew will aufbereitete OCP-Hyperscaler-Systeme für Anwender aus anderen Bereichen zugänglich machen. Der Dienstleister baut diese Geräte aus, arbeitet sie auf und verkauft sie mit Garantie an andere Kunden weiter. Außerdem hilft IT Renew bei der Integration und beim Troubleshooting. Besonders in Edge-Datacentern soll sich aussortiertes Hyperscaler-Equipment bewähren. Typischerweise sind die Geräte zwei oder drei Jahre alt und können noch viele Jahre

Quelle: UNCTAD, UNU

Von Hausmüll bis illegalem Export – ein Teil des Elektroschrotts landet nicht dort, wo er hinsoll. Dazu kommt ein hier nicht gezeigter Anteil, über dessen Verbleib nichts bekannt ist (Abb. 4).

einwandfrei funktionieren. IT Renew gibt standardmäßig drei Jahre Gewährleistung, möglich sind aber auch sechs Jahre. Ab Werk fehlerhafte Hardware, ein recht verbreitetes Übel, ist dann bereits aussortiert. Im Programm hat IT Renews auch Komponenten, etwa Memory für Laptops. Die Mutterfirma des Integrators Boston IT, 2CRSI, verkauft zudem Equipment, bei dem zumindest das Gehäuse aus aufgearbeiteten OCP-Systemen stammt.

OCP-Racks sind allerdings 21 statt die üblichen 19 Zoll breit und bieten damit Platz für drei Server nebeneinander. Optimiert ist in ihnen außerdem die Kühlung, der Materialeinsatz und die Stromversorgung: Zentrale Netzteile versorgen die Server über Stromschienen (siehe Artikel „Ein eigenes Ökotop“ ab Seite 84).

Refurbished-Hardware: viele Bezugsquellen

Doch auch wer beim 19″-Format bleiben will, kann sich mit gebrauchten Servern, Netzwerksystemen und Storage vom Markenhersteller ausrüsten. Beim Refurbisher sollte man sich erkundigen, woher die Hardware stammt, wer sie aufarbeitet und wer für den Service zuständig ist. Bei der Gewährleistung sollte es keine Abstriche vom Üblichen geben. Manche Anbieter haben sich auf den Vertrieb aufgearbeiteter Geräte bestimmter Hersteller spezialisiert, beispielsweise von Cisco-Netzwerkequipment, manche bieten ein buntes Programm. Häufig ist das Angebot gebrauchter Systeme mit Systemhausdienstleistungen gekoppelt. Kurz: Wer dort sucht, kann Geld, CO_2 und Elektroschrott einsparen.

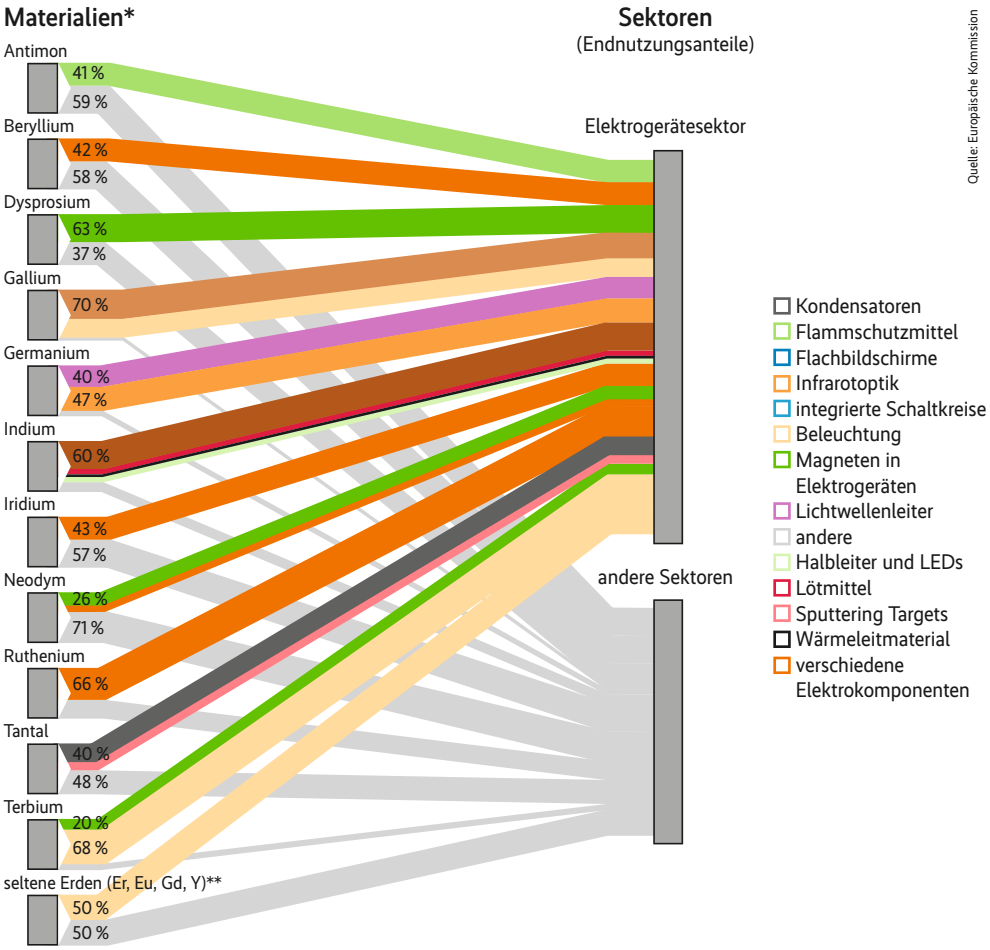
Gerade die großen Systemhäuser legen viel Wert auf Rücknahme und Aufarbeitung ihrer Produkte. Unternehmen mit One-Vendor-Infrastrukturen können bei ihnen oft die Rücknahme ihres Equipments nach der Nutzungsphase und den Austausch gegen neue Hardware zum Vertragsbestandteil machen. HPE etwa konstruiert die Produkte von vorneherein auf Recycelbarkeit hin. Ein Recyclability Assessment Tool soll die Wiederverwertung der Produkte am Ende ihrer Einsatzphase steigern. HPE nennt für seinen Server ProLiant DL380 einen Wert von 98,5 Prozent recycelbarer Anteile – ob sie tatsächlich wiederverwertet werden, ist eine andere Frage.

In zwei Total Repair Centers weltweit, eins davon in Schottland, werden Geräte aus dem eigenen Haus, aber auch die anderer Hersteller, so sie aus einer HPE-dominierten Infrastruktur stammen, zurückgenommen, auf die Möglichkeit einer zweiten Nutzungsphase hin untersucht, gegebenenfalls ausgeschlachtet oder ergänzt und als Refurbished-Systeme auf den Markt gebracht.

Wer liefert, muss auch recyceln

Wer seine Systeme von kleineren Herstellern bezieht oder sich gar individuell zusammenbauen lässt, kann sich auf derlei Service nicht verlassen und muss die gesetzlich festgelegten Wege des WEEE-Recyclings beschreiten. Hier wird es aber Firmenkunden in Zukunft leichter gemacht. Stand keine Rücknahmepflicht im Liefervertrag, mussten sie sich bislang an einen Entsorger wenden, der ihre Geräte kostenpflichtig zurücknahm und den Recycling- oder Refurbishment-Prozess übernahm.

Privatkunden können dagegen ihre Geräte schon seit einiger Zeit im Rahmen der Regeln des Elektrogesetzes (ElektroG) entweder im Laden zurückgeben oder auf Wertstoffhöfen abladen, wo sie von Beauftragten der Hersteller abgeholt werden (siehe Artikel „Kein grüner Zweig“ ab Seite 36). Gemäß den europäischen Kreislaufwirtschaftsplänen wurde jetzt zur Umsetzung der europäischen WEEE-Richtlinie 2012/19/EU das ElektroG verschärft. Die im Januar 2022 in Kraft getretene Version 3 verlagert die Verantwortung auch für kommerziell genutzte Geräte weg vom Kunden zum Hersteller oder Lieferanten.



* Dies ist lediglich ein Ausschnitt kritischer Rohstoffe in europäischen Elektrogeräten. Weitere sind Cerium, Flussspat, Hafnium, Helium, Kobalt, Lanthan, Mangan, Naturkautschuk, Palladium, Platin, Praseodym, Rhodium, Samarium, Silicium, Wolfram und Vanadium.
** Durchschnittlicher Anteil für Erbium, Europium, Gadolinium und Yttrium.

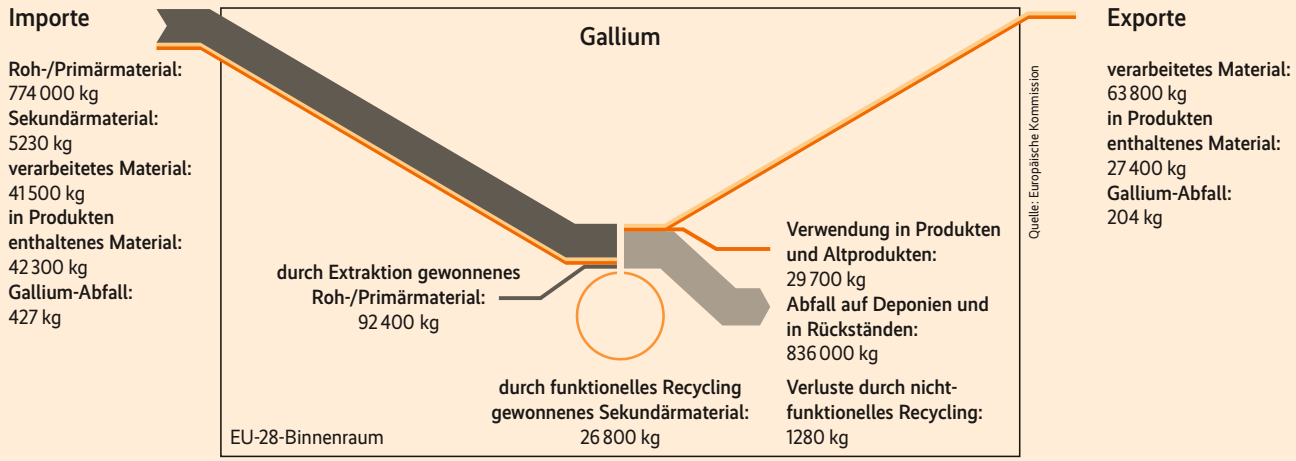
Die Studie belegt eindrücklich, wie viele Stoffe in welchen Anteilen heute in Hightechmaterialien stecken. Nur für einige ist Ersatz in Sicht (Abb. 5).

Materialbilanz am Beispiel von Gallium

Beispielhaft sei anhand von Gallium die EU-28-Materialbilanz eines für die Elektronik wichtigen Metalls erläutert (siehe Abbildung 6). Gallium steckt beispielsweise in schnellen Galliumarsenid-Schaltkreisen. Nach Europa fließen 774 t Rohmaterial (braun), 5 t Sekundärrohstoffe, gut 41 t verarbeitetes Material (gelb) und gut 42 t in Form von Produkten (orange). Außerdem wird die minimale Menge von 427 kg Gallium-Abfall importiert. Durch Extraktion erzeugt Europa 92 t Roh- und Primärmaterial, durch funktionelles Recycling knapp 27 t Gallium.

Auf der anderen Seite werden knapp 64 t verarbeitetes Material (gelb) und gut 27 t Produkte (orange) exportiert, der exportierte Gallium-Abfallanteil liegt bei 204 kg. Knapp 30 t stecken in Altprodukten und Produkten, die ihr Nutzungsende erreicht haben, 836 t landen auf Deponien (grau). Nur ein eher winziger Anteil von 27 t wird für die Wiederverwendung aufbereitet.

Auffällig: Das importierte Gallium wird zu ganz überwiegenden Teilen selbst verbraucht und dann weggeworfen. Mögen die Anteile des funktionalen Recyclings bei den unterschiedlichen Materialien mal höher, mal noch kleiner sein – meist sind sie aber erheblich geringer als das, was deponiert wird, irgendwo versickert oder ungenutzt in Altprodukten schlummert.



Die Materialbilanz von Gallium zeigt die hohe Abhängigkeit vom Import und die sehr geringe Recyclingrate (Abb. 6).

Zuständig für das Organisieren des Recyclings ist wie bisher die Stiftung EAR mit Sitz in Fürth. Sie entstand mit den Rücknahmeverpflichtungen des ElektroG, beauftragt vom Umweltbundesamt, finanziert von den Herstellern. Auch die Verbände Bitkom und ZVEI (Zentralverband der Elektronik-Industrie) sind in der Stiftung vertreten. Sie organisiert die Registrierung von Herstellern und Inverkehrbringern, die in Deutschland IT und Elektronik verkaufen und auch zurücknehmen müssen. Außerdem koordiniert sie die Geräteabholung beispielsweise auf Wertstoffhöfen.

Überfällig: Wegwerfen verboten!

Seit 2022 ist die EAR auch für gewerblich genutzte IT-Produkte zuständig. Anbieter müssen sich bei der EAR registrieren und einen Ansprechpartner für Deutschland benennen. Außerdem müssen sie für Kunden eine „zumutbare Möglichkeit zur Rückgabe“ schaffen, dokumentieren, durch die EAR genehmigen lassen und die finanziellen und organisatorischen Mittel dafür bereithalten. Eine Verpflichtung der Kunden, dem Hersteller die Altgeräte zu überlassen, besteht aber nicht.

Auch das Kreislaufwirtschaftsgesetz (KrWGes) sieht seit 2020 mehr Recycling vor. Danach dürfen beispielsweise Wiederverkäufer intakte Geräte nicht mehr entsorgen, sondern müssen sie vermarkten. Über alle Retouren und ihren Verbleib sind Register zu führen. Das gilt auch, wenn sie zwar nicht mehr intakt sind, nach Reparatur oder Aufbereitung aber wieder einsetzbar wären. Das regelmäßige Verschrotten von Retouren wie beim Onlinegiganten Amazon ist damit verboten. Ob das gegen zu viel Schrott hilft, wird mit Sicherheit von der Kontrolldichte abhängen.

Dass ein besseres Recycling von Elektroschrott zunächst Investitionen erfordert, zeigt sich darin, dass Italien hier in den nächsten Jahren 150 Millionen Euro aus dem Corona-Fonds der EU investieren möchte. Man darf gespannt sein, was die neue deutsche Regierung auf diesem Gebiet unternimmt. Im Koalitionsvertrag (siehe ix.de/znur) steht von verlängerten Nutzungsdauern und E-Schrott-Recycling nichts, wie überhaupt das Thema „Nachhaltige IT“ in wenigen Zeilen auf Seite 18 abgehandelt wird. Gefordert wird in Bezug auf Nutzung und Recycling lediglich, dass Updates und Ersatzteile für die Anwender erkennbar über die reguläre Nutzungsdauer verfügbar sein müssen. Darüber, wie lang die ist oder sein sollte, schweigt das Papier. (sun@ix.de)

Quellen

[1] Vince Beiser; Sand; Reihe Stoffgeschichten; Oekom-Verlag; München 2021, S. 112 ff.

[2] Die im Text angeführten Projekte, Studien und weitere Quellen sind über ix.de/znur zu finden.

Ariane Rüdiger

ist freie IT-Journalistin.

Vorschriften zu einer nachhaltigeren IT

Kein grüner Zweig

Tobias Haar

Energie- und Ressourcenschonung sind Grundpfeiler einer nachhaltigen IT. Auf deutscher und EU-Ebene sind weitere Vorgaben geplant.

Erstaunlicherweise ist das Bundesamt für Seeschifffahrt und Hydrographie bislang die einzige Bundesbehörde, deren RZ mit dem Blauen Engel für Energieeffizienz ausgezeichnet wurde. Daran hat sich seit 2016 nichts geändert. Insgesamt betrieb die Bundesverwaltung Ende 2020 ganze 177 Rechenzentren. Hinzu kommen die RZs der Länder und Kommunen.

Auf der Webseite des Blauen Engels finden sich lediglich zwei weitere ausgezeichnete Rechenzentren: der Green IT Cube des GSI Helmholtz-Zentrums für Schwerionenforschung und das Höchstleistungsrechenzentrum Stuttgart (HLRS). „Ein mit dem Blauen Engel zertifiziertes Rechenzentrum bietet seine Leistung umweltschonend an“, heißt es auf der Webseite. Umso verwunderlicher, dass hier nicht mehr passiert. Nicht einmal die RZs des Bundesumweltministeriums (BMUV) sind zertifiziert.

„Mit gutem Beispiel voranzugehen, ist nicht nur der beste Weg, andere zu beeinflussen, es ist der einzige.“ *Albert Einstein*

Eine Absichtserklärung findet sich zumindest im Koalitionsvertrag der Ampelkoalition vom Herbst 2021: „Wir werden Rechenzentren in Deutschland auf ökologische Nachhaltigkeit und Klimaschutz ausrichten, unter anderem durch Nutzung der Abwärme. Neue Rechenzentren sind ab 2027 klimaneutral zu betreiben. Öffentliche Rechenzentren führen bis 2025 ein Umweltmanagementsystem nach EMAS (Eco Management and Audit Scheme) ein. Für IT-Beschaffungen des Bundes werden Zertifizierungen wie der Blaue Engel Standard. Ersatzteile und Softwareupdates für IT-Geräte müssen für die übliche Nutzungsdauer verpflichtend verfügbar sein.“ Allerdings müssen die gesetzlichen Vorgaben dafür auf nationaler und EU-Ebene erst noch geschaffen werden.

Die Webseiten des BMUV nennen drei Säulen einer grünen IT: energieeffiziente Rechenzentren, nachhaltige Beschaffung von Hardware und die genannte Zertifizierung mit dem Blauen Engel. Insgesamt machen die Webseiten nicht den Eindruck, als seien sie seit dem Regierungswechsel im Herbst 2021 aktualisiert worden. Konkrete Vorhaben zur Umsetzung der im Koalitionsvertrag geäußerten Absichten sucht man vergebens.

Grundsätzlich gibt es drei Ansätze, IT-Betreiber durch staatliche Maßnahmen zum Energiesparen zu bewegen: Gesetze, Förderungen und die Steuerung der Nachfrage durch Preiseingriffe. Zudem wirken Unternehmen mit einer nachhaltigen Geschäftspolitik attraktiver auf Kunden und Anleger. Betreiber nicht staatlicher Rechenzentren sind schon aus Kostengründen an einem energieeffizienten Betrieb interessiert. Fachleute erwarten, dass der Strompreis weiter steigen wird. Mit dem zunehmenden Preisdruck dürften staatliche Vorgaben an Relevanz verlieren. Gerade deshalb muss aber auch die staatliche Verwaltung zu einem energieeffizienten Betrieb ihrer Rechenzentren verpflichtet werden. Private RZ-Betreiber könnten mit einer Selbstverpflichtung staatlichen Vorgaben zuvorkommen.

Marktregulierung oder Gesetz?

Der Leitfaden „Energieeffizienz in Rechenzentren“ des Branchenverbands Bitkom e. V. bietet hier zusätzliche Hilfestellung an (alle Leitfäden siehe ix.de/z9bw). Er umfasst im Wesentlichen die Bereiche Energiemanagement und Optimierungspotenziale. Gesetzeskraft haben solche Leitfäden nicht. Allerdings beschreiben sie häufig den Stand der Technik. In diesem Fall ist aber Zurückhaltung angezeigt, denn der Leitfaden stammt von 2015 und ist sicherlich überarbeitungsbedürftig.

Deutlich aktueller ist der Leitfaden „Nachhaltige Rechenzentren“ des Forschungsverbundes „Nachhaltige Rechenzentren Baden-Württemberg“ (EcoRZ) vom Juni 2020. Er beleuchtet anschaulich die Faktoren, die bei der Energieeffizienz eine Rolle spielen. Sie reichen von der ökologischen Standortbewertung bis zur Abwärmenutzung. Weitere Aspekte sind etwa soziale Indikatoren wie Mitarbeiterwohlergehen oder das Vermeiden von Rohstoffen, die bewaffnete Konflikte provozieren. Denkbar wäre, diese Vorgaben stärker in Gesetzen zu verankern.

In einigen Bereichen gibt es ökologisch motivierte gesetzliche Vorgaben beim Einsatz von IT. Ein Beispiel ist das Elektro- und Elektronikgerätegesetz mit Stand vom 1. Januar 2022. Offiziell lautet die Bezeichnung „Gesetz über das Inverkehrbringen, die

Rücknahme und die umweltverträgliche Entsorgung von Elektro- und Elektronikgeräten“, kurz ElektroG 3. § 1 ElektroG 3 legt die abfallwirtschaftlichen Ziele fest:

„Es bezweckt vorrangig die Vermeidung von Abfällen von Elektro- und Elektronikgeräten und darüber hinaus die Vorbereitung zur Wiederverwendung, das Recycling und andere Formen der Verwertung solcher Abfälle, um die zu beseitigende Abfallmenge zu reduzieren und dadurch die Effizienz der Ressourcennutzung zu verbessern. Um diese abfallwirtschaftlichen Ziele zu erreichen, soll das Gesetz das Marktverhalten der Verpflichteten regeln.“ *§ 1 ElektroG 3*

Hersteller müssen die Wiederverwendung, Demontage und Verwertung bei der Konstruktion ihrer Elektro- und Elektronikgeräte möglichst berücksichtigen. Auch sollen „Altbatterien und Altakkumulatoren durch Endnutzer problemlos und zerstörungsfrei entnommen werden können“. Die Geräte sind „möglichst so zu gestalten“, dass dieses Ziel erreicht werden kann. Was genau unter „möglichst so zu gestalten“ zu verstehen ist, regelt das Gesetz nicht. Es kommt – wie oft in der Juristerei – auf den Einzelfall an. Mitunter muss aber der Umweltschutz hinter Sicherheitsaspekten zurückstehen. Dies gilt auch für die Vorgabe in § 4 Absatz 1 Satz 2 und 3, Batterien so in Geräte einzubauen, dass sie entnehmbar sind. Denn:

„Absatz 1 Satz 2 und 3 gilt nicht für Elektro- und Elektronikgeräte, in denen aus Gründen der Sicherheit, der Leistung, aus medizinischen Gründen oder aus Gründen der Vollständigkeit von Daten eine ununterbrochene Stromversorgung notwendig und eine ständige Verbindung zwischen dem Gerät und der Batterie oder dem Akkumulator erforderlich sind.“
§ 4 Absatz 3 ElektroG 3

Bringt der Hersteller Elektro- und Elektronikgeräte in Verkehr, die nicht in privaten Haushalten zum Einsatz kommen, ist er zum Vorlegen eines Rücknahmekonzepts verpflichtet. Hinzu kommen Entsorgungspflichten, etwa die Trennung von Altbatterien und -akkus. Relevant sind auch die Rücknahmepflichten der Hersteller und Vertreiber – zu Letzteren zählen auch Onlineshops. Teilweise bestehen hier die Rücknahmepflichten auch dann, wenn ein Kunde kein Neugerät erwirbt.

Klein und groß: Batterien und Akkus

Seit 2021 gilt die aktuelle Fassung des Batteriegesetzes, also des „Gesetzes über das Inverkehrbringen, die Rücknahme und die umweltverträgliche Entsorgung von Batterien und Akkumulatoren“, kurz BattG 2. Es gilt „auch für Batterien, die in andere Produkte eingebaut oder anderen Produkten beigefügt sind“, und damit nicht nur für Gerätebatterien wie die verbreiteten AA- oder AAA-Modelle. Die zugelassenen privaten oder herstellereigenen Rücknahmesysteme müssen eine Sammelquote gebrauchter Batterien von 50 Prozent erreichen. Das Gesetz enthält etliche weitere Vorgaben, etwa Kennzeichnungs- oder Informationspflichten der Batteriehersteller. Einen guten Überblick bietet die Webseite batteriegesetz.de.

BattG 2 dürfte bald erneut überarbeitet werden, spätestens mit dem Inkrafttreten der derzeit in finaler Abstimmung befindlichen EU-Batterieverordnung. Sie sieht einige Neuerungen vor, zum Beispiel die ab 2024 geltende Pflicht, Batterien oder Akkus in Smartphones so einzubauen, dass Nutzer sie „einfach und sicher selbst“ entnehmen können.

Der Leitfaden des Forschungsverbundes „Nachhaltige Rechenzentren Baden-Württemberg“ liefert zahlreiche Anregungen für ein energieeffizientes Rechenzentrum.

Was gilt beim Batterie- und Akkueinsatz im Rechenzentrum, etwa im Rahmen der unterbrechungsfreien Stromversorgung, kurz USV? Auch beim industriellen Einsatz sollen die Vorgaben des BattG weitreichend greifen:

„Dieses Gesetz gilt für alle Arten von Batterien, unabhängig von Form, Größe, Masse, stofflicher Zusammensetzung oder Verwendung. Es gilt auch für Batterien, die in andere Produkte eingebaut oder anderen Produkten beigefügt sind.“
§ 1 Absatz 1 BattG 2

§ 2 BattG klärt, dass auch Akkus als Batterien gelten:

„‚Batterien‘ sind aus einer oder mehreren nicht wiederaufladbaren Primärzellen oder aus wiederaufladbaren Sekundärzellen bestehende Quellen elektrischer Energie, die durch unmittelbare Umwandlung chemischer Energie gewonnen wird.“ *§ 2 Absatz 2 BattG 2*

Rechenzentren verwenden in der Regel Blei-Gel-Akkus, mitunter auch Lithium-Ionen-Akkus. Beide gelten als Industriebatterien und unterliegen dem Batteriegesetz, wenn keine Ausnahmen nach § 1 Absatz 2 BattG vorliegen, etwa, wenn es sich um Batterien in Rechenzentren des Bundes handelt, „die mit dem Schutz der wesentlichen Sicherheitsinteressen der Bundesrepublik Deutschland in Zusammenhang stehen“. Auch im militärischen Bereich und beim Einsatz im Weltraum greifen die Regeln des BattG nicht. Solche Batterien dürfen deshalb einen höheren Quecksilber- oder Cadmiumanteil enthalten.

Auch Industriebatterien darf der Hersteller nur in Verkehr bringen, wenn er oder seine Bevollmächtigten die geltenden Rücknahmepflichten einhalten. Allerdings muss er für sie kein Rücknahmesystem vorhalten, sondern lediglich „eine zumutbare und kostenfreie Möglichkeit der Rückgabe“ anbieten und eine gesetzeskonforme Verwertung gewährleisten. Händler und Verwerter müssen Altbatterien nicht an den Hersteller zurückgeben. Sie sind „nur“ zur kostenfreien Rücknahme und deren ordnungsgemäßer Entsorgung verpflichtet. Hersteller von Industriebatterien müssen trotz dieser Ausnahme die erforderlichen finanziellen und organisatorischen Mittel zur Verfügung stellen, damit Händler und Verwerter stellvertretend ihren Rücknahmepflichten nachkommen können.

Für Fahrzeug- und Industriebatterien bestehen derzeit noch keinen Sammelquoten. Allerdings sind sie zu kennzeichnen mit der Marke zur Herstelleridentifikation, Angaben zur Kapazität gemäß EU-Verordnung 1103/2010 und dem Symbol des durchgestrichenen Mülleimers, gegebenenfalls ergänzt um Gefahrstoffkennzeichnungen. Diese Kennzeichnung muss gut sichtbar, lesbar und dauerhaft sein.

Verkehrsverbote im Industriebatteriebereich bestehen derzeit lediglich für den Quecksilberanteil. Dieser darf 0,0005 Pro-

zent nicht übersteigen. Für die Anteile an Cadmium und Blei bestehen keine entsprechenden Begrenzungen. Im Rahmen administrativer Pflichten sind die Vorgaben über Registrierung, Garantiestellung und Reporting einzuhalten.

Pflichten der Rechenzentren

Über Pflichten auf Anwenderseite schweigt sich das Batteriegesetz aus. Dazu das Umweltbundesamt: „Geräte-, Fahrzeug- und Industrie-Altbatterien können weiterhin unentgeltlich bei den jeweiligen Vertreibern dieser Batteriearten zurückgegeben werden." Es besteht danach ein Rückgaberecht, aber keine Pflicht. Das Rückgaberecht sollte aber Anreiz genug sein. Denn gebrauchte Batterien so zwischenzulagern, dass sie gegebenenfalls Schadstoffe abgeben, kann teuer werden.

Nach § 25 des Kreislaufwirtschaftsgesetzes kann eine Rechtsverordnung erlassen werden, die verpflichtend vorschreibt, „dass die Besitzer von Abfällen diese den nach Absatz 1 verpflichteten Herstellern, Vertreibern oder nach Absatz 1 Nummer 2 eingerichteten Rücknahmesystemen zu überlassen haben". Das umfasst auch Industrie-Altbatterien: „Abfälle im Sinne dieses Gesetzes sind alle Stoffe oder Gegenstände, derer sich ihr Besitzer entledigt, entledigen will oder entledigen muss." Über die „Rücknahme gebrauchter Industriebatterien" und den zu beachtenden Rechtsrahmen informiert ein gleichlautender Flyer des ZVEI.

In Falle einer Havarie hat ein Unternehmer nicht nur für den Schaden und die Entsorgung aufzukommen. Im Raum stehen zudem Straftaten mit Freiheitsstrafe oder Geldstrafe wegen Gewässerverunreinigung oder Bodenverunreinigung. Bis zu fünf Jahre Freiheitsstrafe steht auf den „unerlaubten Umgang mit Abfällen". Ein Lagern von Abfällen außerhalb „vorgeschriebener oder zugelassener Verfahren" kann bereits ausreichen. Und wie immer im Strafrecht gilt: Unwissenheit schützt vor Strafe nicht.

Überdies gelten Nachweispflichten auch für Besitzer von Industriebatterien als „gefährlichen Abfällen", wenn sie nicht mehr dem eigentlichen Zweck dienen. Mitunter sind Batterien als Gefahrgut nach der Abfallverzeichnis-Verordnung zu kennzeichnen. Bleibatterien ist beispielsweise der AVV-Schlüssel 16 06 01 zugewiesen. Weitere Informationen hat das Bayerische Landesamt für Umwelt in einem „InfoBlatt Batterien und Akkumulatoren" zusammengestellt. Neben der Lagerung ist danach auch der Transport von Industriebatterien geregelt.

Viele RZ-Betreiber haben mit ihren Lieferanten vertragliche Vereinbarungen über die Rücknahme alter Industriebatterien getroffen. Das ist ein Grund dafür, dass das Batteriegesetz keine allzu strikten Vorgaben über das Wie der Rücknahme macht. In derartigen Verträgen sollten auch die Rollen und Verantwortlichkeiten der Beteiligten nach den einschlägigen Rechtsvorschriften klar beschrieben werden. Der unsachgemäße Umgang mit Industriebatterien kann teuer und mitunter strafbar sein. Je näher die Versorgung eines RZ mit Akkus und dergleichen an ein Battery-as-a-Service-Modell kommt, desto einfacher für seinen Betreiber. Rechtlich entlässt ihn ein entsprechender Vertrag jedoch nicht aus der abfallrechtlichen Verantwortung. Wie in anderen Outsourcing-Fällen auch verbleiben gewisse Überwachungspflichten beim Nutzer.

Nicht gleich wegschmeißen!

Derzeit laufen auf EU-Ebene auch Diskussionen über das Recht auf Reparatur von Elektro- und Elektronikgeräten. Für Fernseher, Spül- und Waschmaschinen sowie Kühlschränke gibt es bereits seit 2021 Vorgaben gemäß der Ökodesign-Richtlinie, die besagen, dass die Geräte reparierbar sein müssen. Das bedeutet auch, dass Hersteller Ersatzteile für einen bestimmten Zeitraum vorhalten müssen. Außerdem sollen Produkte so gestaltet sein, dass sie sich ohne Spezialwerkzeug und zerstörungsfrei öffnen lassen. Auch für kleinere Geräte soll ein Recht auf Reparatur kommen, das sich an diesen Prinzipien orientiert. Im Ampel-Koalitionsvertrag heißt es dazu:

> „Die Lebensdauer und Reparierbarkeit eines Produktes machen wir zum erkennbaren Merkmal der Produkteigenschaft (Recht auf Reparatur)." *Koalitionsvertrag 2021*

Dies soll den Zugang zu Ersatzteilen und Reparaturanleitungen umfassen. Gemeint sind damit auch Softwarekomponenten wie Middleware oder Betriebssysteme:

> „Herstellerinnen und Hersteller müssen während der üblichen Nutzungszeit Updates bereitstellen." *Koalitionsvertrag 2021*

Möglicherweise kommt es auch zu einer Verlängerung der gesetzlichen Gewährleistungszeiten bei Produkten, deren „jeweilige Lebensdauer" diese Gewährleistungsfristen übersteigt. Um die längere Nutzung von Geräten zu fördern, ist auch eine verringerte Umsatzbesteuerung für Gerätereparaturen im Gespräch, ebenso wie eine Pflicht zur Veröffentlichung von Reparaturanleitungen. Ein Reparaturindex auf der Produktverpackung soll Verbraucher zudem künftig über die „Reparaturfreundlichkeit" von Produkten informieren.

Unternehmensvertreter, Verbraucher- und Umweltschützer plädieren für möglichst einheitliche Regelungen auf EU-Ebene. Es ist nachvollziehbar, dass allzu unterschiedliche einzelstaatliche Regelungen ineffizient und letztlich teuer werden.

Fazit

„Die Begrenzung des IT-bedingten Energie- und Ressourcenverbrauchs durch Green-IT-Ansätze ist eine der Hauptaufgaben bei der Gestaltung der Digitalisierung." So beschreibt es das Umweltministerium Baden-Württemberg. Andererseits kann Digitalisierung Ressourcenschonung auch erst ermöglichen. In den letzten Jahren tat sich allerdings erstaunlich wenig im Bereich der Green IT – zumindest auf der Ebene der Gesetzgeber. Jetzt kommt etwas mehr Bewegung in das Thema. Auch auf EU-Ebene kommen die Diskussionen allmählich voran.

Bei Elektro- und Elektronikgeräten, aber auch Batterien gelten spezifischere Vorgaben als in anderen Bereichen. Ihnen gemein ist der Fokus auf der Vermeidung von Müll und Umweltverschmutzung. Über die möglichst energiearme Produktion oder den effizienten Betrieb von Rechenzentren fehlen konkretere Vorgaben – noch. (ur@ix.de)

Quellen

Die genannten Leitfäden und einige informative Webseiten sind über ix.de/z9bw zu finden.

Tobias Haar, Rechtsanwalt, LL.M. (Rechtsinformatik), MBA,

ist Rechtsanwalt mit Schwerpunkt IT-Recht bei Vogel & Partner in Karlsruhe.

Grüne Software

Häufig zeichnen Designfehler, unnötige Komplexität und Ressourcenverschwendung heutige Software aus. Bis zur Nachhaltigkeit hat sie damit noch einen langen und steinigen Weg vor sich, gepflastert mit genügsamen Funktionen, durchdachtem Design, höherer Qualität, guter Wartbarkeit, neuen Narrativen und langfristigen Strategien.

Kontrollverlust – Technisches Versagen in der Softwareentwicklung 40

Strategie – Organisatorische Einbettung nachhaltiger Software 48

Design – Nachhaltig programmieren mit Bedacht 54

Abschalten – Zombies im Rechenzentrum 58

Im Detail – Energieeffizienz von Software messen 62

C++ – Lernen von der Embedded-Entwicklung 67

Technisches Versagen in der Softwareentwicklung

Ein schwerer Anfang

Dr. Alexander Schatten

Solange die Komplexität der Softwaresysteme die Kompetenz der Produzenten übersteigt, lässt sich keine nachhaltige Software entwickeln. Diesen Kontrollverlust gilt es als Erstes in den Griff zu bekommen.

Die Nachhaltigkeit heutiger Softwaresysteme ist in allen Dimensionen mangelhaft. Angriffe durch Ausnutzen kritischer Sicherheitslücken, Datenklau im großen Stil oder Erpressungen mit Ransomware bilden nur die Spitze des Eisbergs, der diesen Mangel an Nachhaltigkeit offenlegt. In vielen Fällen ist das nur symptomatisch für eine fundamentalere Entwicklung: Die Komplexität der Infrastruktur überfordert inzwischen fast alle Organisationen (siehe Kasten „Die Softwarekrise“).

Das hat den Kontrollverlust über weite Bereiche der IT-Infrastruktur zur Folge und birgt ein erhebliches Schadenspotenzial – indirekt etwa durch Angriffe von außen oder direkt durch riesige gescheiterte Projekte in der Industrie. In ihnen werden oft Hunderte Millionen Euro abgeschrieben; man denke nur an die gescheiterte Altsystemablösung bei Lidl.

Auf geradezu komische Weise illustriert der Aufruf des Gouverneurs von New Jersey diese Situation mitten in der Covid-Krise: „If you know how to code COBOL, the state of New Jersey wants to hear from you.“ Man war kaum mehr in der Lage, wesentliche Funktionen der Behördensoftware zu warten, und sah sich gezwungen, im Fernsehen COBOL-Entwickler zu suchen.

Nahezu alle gesellschaftlichen Systeme wie die der Logistik oder der Energieversorgung sind inzwischen von der IT abhängig geworden, die sich wieder in zirkulären Abhängigkeiten befindet (siehe „Durchschritten“ ab Seite 8). Dieser Kontrollverlust ist bei Weitem kein rein technisches Problem, sondern ist nur im Zusammenspiel mit Management- und ökonomischen Mechanismen zu verstehen, die der Artikel „Nur gemeinsam“ ab Seite 48 beschreibt (siehe auch Kasten „Gründe für den zunehmenden Kontrollverlust“).

Mechanismen des Kontrollverlusts über die Systeme

Eine wesentliche Rolle spielt in allen komplexen Systemen die Skalierung. Unter dem Aspekt der Nachhaltigkeit kann die Skalierung positive Effekte zeitigen, etwa durch Effizienzgewinne

Die Softwarekrise

Seit den 1960er-Jahren steht der Begriff Softwarekrise im Raum. Bereits vor über 50 Jahren – 1968 – hieß es in dem Bericht, der auf einer NATO-Konferenz gehalten wurde:

„Es gibt eine wachsende Kluft zwischen den Ambitionen und Errungenschaften im Softwareengineering. Diese Kluft tritt in mehreren Dimensionen auf: zwischen den Versprechungen gegenüber den Nutzern und der von der Software erzielten Leistung, zwischen dem, was letztendlich möglich erscheint, und dem, was erreicht wurde, und zwischen den Schätzungen der Softwarekosten und den tatsächlichen Ausgaben. Die Lücke entsteht in einer Zeit, in der die Folgen von Softwarefehlern in all ihren Aspekten immer gravierender werden. Besonders alarmierend ist die scheinbar unvermeidliche Fehlbarkeit großer Software, da eine Fehlfunktion in einem fortschrittlichen Hardware-Software-System eine Frage von Leben und Tod sein kann, nicht nur für Einzelpersonen, sondern auch für Transportmittel mit Hunderten von Menschen und letztendlich auch für Nationen."

Obwohl die Software gut 50 Jahre später eine völlig andere Dimension erreicht hat, gilt der Text nahezu wörtlich auch heute noch. Sehr deutlich wird dies, wenn man sich die zunehmenden Herausforderungen bei der Wartung und Weiterentwicklung gewachsener Systeme ansieht oder die Zahl an gravierenden Schwierigkeiten bei großen Softwaresystemen betrachtet.

Es vergeht kaum ein Tag, an dem nicht eine kritische Sicherheitslücke, ein kritischer Softwarefehler etwa bei Flugzeugen oder ein großer Datenverlust zutage tritt: 2020 wurde beispielsweise bekannt, dass große Teile von US-Infrastruktur für längere Zeit von Hackern übernommen worden waren. Einer der führenden Securityexperten, Bruce Schneier, teilt die Einschätzung des damaligen Homeland Security Advisor Thomas Bossert: „Die Dimension dieser nationalen Securitylücke kann kaum übertrieben werden" (siehe ix.de/z7d4).

Auch Datenverluste in riesigem Umfang sind an der Tagesordnung, die ÖBB verliert 2020 Zehntausende Kundendaten, die US-Tochter der Deutschen Telekom bestätigt einen Angriff auf Kundendaten mit großem Schadenspotenzial, EasyJet verliert neun Millionen Kundendaten et cetera et cetera. Dazu kommt die Erpressung nach der Installation von Ransomware. So muss durch einen Ransomware-Angriff eine US-Pipeline geschlossen werden, große deutsche Unternehmen wie MediaMarkt verlieren den Zugriff auf ihre Systeme und können Kunden nur eingeschränkt bedienen, 2021 bleiben in Schweden 800 Supermarktfilialen geschlossen.

und eine Optimierung des Ressourceneinsatzes. Sie verändert aber das Verhalten von Systemen auf nicht einfach zu durchschauende Weise.

Dies betrifft gesellschaftliche oder ökonomische Systeme wie Unternehmen und in ähnlicher Weise Softwaresysteme, besonders wenn sie miteinander interagieren. Das Verhalten kann auf unterschiedlichen Skalierungsebenen variieren, dadurch lässt sich das Verhalten höherer Skalierung nicht vom Verhalten niedrigerer Skalierung ableiten. Um es mit einem Beispiel von Nassim Taleb zu illustrieren: „Ein Land ist keine große Stadt, eine Stadt keine große Familie und – Entschuldigung – die Welt ist kein großes Dorf." [1]

Mit Skalierung und wechselseitiger Interaktion steigt im Allgemeinen auch die technische Komplexität des Systems an: Mehr Stakeholder interagieren und definieren größere Anwendungen, die immer komplexere Aufgaben bewältigen. Bei Softwaresystemen sind zwei Arten von Komplexität zu unterscheiden: die externe und die interne.

iX-TRACT

- Die Komplexität der Softwaresysteme ist schneller gewachsen als die Fähigkeit, sie im Griff zu halten.
- Skalierungen, das Wachsen der Aufgaben, Übermodellierung, architektonische Mängel, Abhängigkeiten von externen Bibliotheken und Diensten, die zunehmende Anwendung neuer nichtdeterministischer Algorithmen und kurzfristige Effizienzsteigerungen forcieren die Komplexität der Systeme.
- Daraus resultiert ein Kontrollverlust, der die Gesellschaft teuer zu stehen kommt.
- Doch gibt es Strategien, mit denen man dem Kontrollverlust entgegenwirken kann.

Die externe Komplexität ist schwer vermeidbar und wächst aus kybernetischen Prinzipien heraus, auch genannt Requisite Variety: Ein Kontrollsystem – die Software – muss über ähnlich viele Freiheitsgrade verfügen wie das System, das es kontrollieren soll, also die Anwendungsdomäne. Eine hochkomplexe Domäne führt zu hochkomplexer Software. Die Akteure im System, seien es Mitarbeiter oder Softwarekomponenten, müssen die Fähigkeit und Freiheitsgrade beziehungsweise Diversität haben, sich an die externe Komplexität anzupassen. Schlichte Prozesse und Regeln reichen für komplexe Umgebungen nicht aus.

Ist diese Fähigkeit, auf die externe Komplexität zu reagieren, nicht gegeben, spricht die Kybernetik von Complexity Mismatch. Dann muss das System der niedrigeren Komplexität komplexer werden oder es kollabiert. Grundsätzlich droht heute auch der gegenläufige Fall: Wird die Software deutlich komplexer als die Realität, man denke an Finanz- und Börsensysteme oder die globale Vernetzung durch soziale Netzwerke, geht der Druck in Richtung der Welt, sich anzupassen oder eben zu kollabieren.

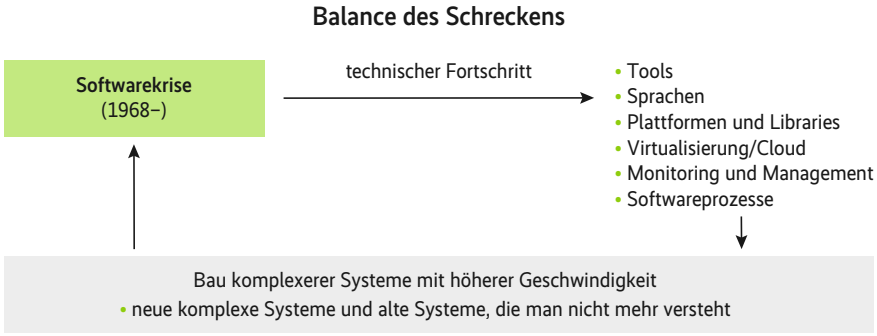
Unnötige Komplexität durch Übermodellierung und Architekturfehler

Teilweise vermeidbar ist hingegen die interne Komplexität. Sie kann der Übermodellierung der Welt oder einem Mangel an Softwareengineering oder technischer Architektur entspringen. Als – systemische – Folge dieser steigenden Komplexität zeigen die Maschinen kein klassisch deterministisches, also vorhersagbares Verhalten mehr, sondern ein emergentes.

Die zunehmende Anwendung neuer nichtdeterministischer Algorithmen wie des Machine Learning verschärft die Situation. Hier wird ja explizit nicht vorgegeben, wie die Regeln aussehen, nach denen sich die Maschine verhält, stattdessen soll sie die Regeln aus der Beobachtung der Welt erlernen. Damit steht die Maschine, die Software, in einem noch komplexeren Wechselverhältnis mit der Welt, mit der sie interagiert, denn das Ver-

halten dieser Softwaresysteme verändert ja wieder die Welt, von der sie lernen.

All diese Zusammenhänge kann man auch unter einem anderen Aspekt betrachten, nämlich dem von Interaktionen und Abhängigkeiten. Moderne Softwaresysteme sind abhängig von Personen, die technisches und fachliches Wissen über die Anwendungen besitzen. Fachliches Wissen geht häufig verloren durch die Digitalisierung und fehlt dann bei späteren Änderungen oder Migrationen, technisches Wissen durch die Spezialisierung auf einzelne Mitarbeiter und mangelnde Verbreitung im Team, besonders auch über längere Zeiträume.



Der Kontrollverlust führt zu einem Teufelskreis, in dem die Weiterentwicklung der Werkzeuge und Techniken zu immer komplexeren Systemen führt, die die Softwarekrise weiter befeuern (Abb. 1).

Daneben gibt es technische Abhängigkeiten, statische sowie dynamische. Statische Abhängigkeiten entstehen etwa durch Bibliotheken, ohne die keine moderne Software zu entwickeln wäre, außerdem durch Basissysteme wie Betriebssysteme oder Datenbanken. Verwendete Bibliotheken haben häufig wieder Abhängigkeiten zu anderen Bibliotheken, und damit entsteht in kurzer Zeit ein Abhängigkeitsgraph, der fast unüberblickbar und teilweise sogar widersprüchlich ist, etwa wenn die Teile einer Anwendung Abhängigkeiten zur selben Bibliothek, aber zu unterschiedlichen Versionen haben.

Unüberschaubare Abhängigkeiten

Diese Abhängigkeiten sind nach der Ansicht vieler Experten mittlerweile außer Kontrolle geraten. Das Betriebssystem eines Smartphones hat heute bis zu 10 GByte, eine gigantische Größe, wenn man bedenkt, dass etwa Windows 95 30 MByte umfasste. Google Search füllt auf Android etwa 350 MByte, Microsoft Outlook und Excel auf macOS jeweils etwa 2 GByte, Signal – ein simpler Messenger – benötigt auf dem Desktop fast ein halbes Gigabyte. Wie kann es sein, dass eine E-Mail-Anwendung circa 60-mal größer ist als ein gesamtes Betriebssystem vor rund 25 Jahren?

Dies ist ein genereller Trend bei aktueller Software. In vielen Fällen ist der Code, der durch externe Abhängigkeiten in das Projekt hineingezogen wird, wesentlich umfangreicher als der eigentliche Code der Anwendung. Dieses Bloating ist zum Teil auch durch den Versuch der Modularisierung und Einhegung von Komplexität zu erklären, weil die Infrastruktur, die zur Modularisierung verwendet wird, selbst immer umfangreicher und komplexer wird. Es ist aber auch eine Manifestation des Kontrollverlustes. Kein Entwickler kann mehr durchblicken, was sich in 2 GByte großem Programmcode einer einzelnen Anwendung abspielt. Positive Gegenbeispiele gibt es dagegen leider nur sehr wenige.

Man steckt nun in einem Dilemma: Entweder hält man alle Abhängigkeiten aktuell, etwa weil man bekannte Sicherheitslücken schließen muss. Das aber führt häufig zu instabilem Verhalten, weil sich das Verhalten der eigenen Anwendung durch die enorme Zahl an Abhängigkeiten laufend ändert. Oder man fixiert die Abhängigkeiten, sobald die Software einen stabilen Zustand erreicht hat, schleppt dann aber die zahlreichen Sicherheitslücken und Fehler von Bibliotheken mit, die man nicht aktualisiert. Letzteres ist in der Praxis häufig zu sehen, da man von der schieren Dynamik der Änderungen überfordert ist.

Der Mangel an automatisierten Tests, die seit 20 Jahren Stand der Technik sind, verschärft die Situation: Abweichendes Verhalten wird dann oft erst nach dem Release erkannt. Je länger man aber wartet, diese Aktualisierungen vorzunehmen, umso schwieriger wird es, die dadurch aufgebaute technische Schuld abzuarbeiten. Zu diesen statischen Abhängigkeiten gesellen sich dynamische Abhängigkeiten, beispielsweise wenn sich Systeme aktualisieren, auf die man keinen unmittelbaren Einfluss hat, etwa Cloud-Systeme oder Services von anderen Unternehmen oder Abteilungen.

Software altert

Software altert – dies ist die bildlichere Beschreibung dieses Verhaltens, wie Kevin Kelly es ausdrückt: „Neue Computer versteinern. Apps werden schwächer. Code erodiert. Je komplexer die Werkzeuge werden, desto mehr, nicht weniger Aufmerksam-

Gründe für den zunehmenden Kontrollverlust

Skalierung: Durch das Skalieren verändert sich das Verhalten von Systemen auf nicht immer durchschaubare Weise.

Wachsende Aufgaben: Komplexere Anwendungsdomänen bedingen komplexere Kontrollsysteme.

Übermodellierung und architektonische Mängel: Die unnötige interne Komplexität aufgrund von Designfehlern verursacht zusammen mit externer Komplexität ein emergentes Verhalten.

Nichtdeterministische Algorithmen: Die zunehmende Anwendung von ML-Algorithmen versetzt die Software in ein noch komplexeres Wechselverhältnis mit der Welt, von der sie lernt und die sie verändert.

Interaktionen und Abhängigkeiten zwischen Mensch und Software: Fachwissen geht durch Digitalisierung und Spezialisierung verloren und fehlt bei Änderungen oder Migrationen.

Technische Abhängigkeiten: Moderne Software lässt sich nicht ohne Bibliotheken, Datenbanken und Betriebssysteme entwickeln, deren Aktualisierungen aber eigenen Regeln folgen.

Kurzfristige Effizienzsteigerungen: Das Vernachlässigen von Wartung und Weiterentwicklung verringert die Resilienz und Nachhaltigkeit der Systeme.

keit benötigen sie“ [2]. Als Teil von immer größer werdenden Metasystemen sind die Systeme ständig an die Veränderungen der Metasysteme anzupassen. Anderenfalls altern und sterben sie. Webseiten von vor 20 Jahren funktionieren nur noch bedingt in modernen Browsern, noch schlimmer umgekehrt, alte Apps nicht mehr in neueren Betriebssystemversionen, und API-Änderungen in der Cloud müssen nachgezogen werden.

„Neue Computer versteinern.
Apps werden schwächer. Code erodiert.“
Kevin Kelly

Das Verhalten von Software im Betrieb lässt sich folglich prinzipiell nicht mehr in Teststellungen replizieren und nicht mehr einfach beschreiben. Das bedeutet nicht, dass das klassische Testen – selbst das meist mangelhaft umgesetzte – damit wertlos wäre, es reicht nur bei Weitem nicht mehr aus. Das hat ernsthafte Konsequenzen: Wie soll die Funktionsbeschreibung moderner Software aussehen? Was bedeutet Stabilität? Oder umgekehrt: Was bedeutet Fehler? Wie müssen Qualitätssicherung, Wartung und Monitoring aussehen, damit sie den Herausforderungen genügen?

Moderne Softwaresysteme sind, bildlich ausgedrückt, keine klassischen Maschinen mehr, sondern ähneln vielmehr biologischen oder kybernetischen Systemen. Sie sind in Ökosysteme eingebettet und damit einer ständigen Evolution unterworfen. Eine der wesentlichen Herausforderungen der nächsten Jahrzehnte für Softwareingenieure, Manager und Gesellschaft besteht deshalb darin, diesen Wandel zu verstehen und alle Prozesse danach auszurichten. Damit verbunden ist auch die Notwendigkeit, in mehreren Zeitdimensionen parallel zu denken und zu arbeiten.

„Die meisten modernen Effizienzmaßnahmen
sind nur verzögerte Bestrafung.“
Nassim Taleb

In den letzten Jahrzehnten lag der Fokus häufig auf kurzfristigen und simplen Effizienzsteigerungen. Diese führen aber in der Regel zu geringerer Resilienz und Nachhaltigkeit der Systeme, sie sind, um beim Bild zu bleiben, nicht ökosystemkonform. Wieder hat Nassim Taleb den Nagel auf den Kopf getroffen: „Die meisten modernen Effizienzmaßnahmen sind nur verzögerte Bestrafung.“ Das heißt, was man kurzfristig – vielleicht – an Effizienz gewinnt, zahlt man mittelfristig mehrfach zurück.

So geht es nicht weiter

Kurzfristiges Denken hat bisher einigermaßen funktioniert, weil es in der IT wenig an bestehender Infrastruktur gab. Dies ist nun vorbei. Jede Innovation muss mit der bestehenden Infrastruktur zusammenspielen und ist von ihr abhängig. Die kurzfristigen Gewinne und das Vernachlässigen der Wartung und Weiterentwicklung von Legacy Systems – bereits der Name spricht Bände – führt zu dem heute überall sichtbaren Kontrollverlust. Die Folge sind zum Teil extrem steigende Kosten, um aktuelle Systeme am Laufen zu halten, bei gleichzeitig steigenden Hürden, neue Funktionen einzubringen – von den Mängeln der Qualität, der Sicherheit, der Resilienz und Stabilität ganz zu schweigen (siehe Abbildung 2).

Gesellschaftlich sind damit existenzielle und strategische Risiken verbunden. Betrachtet man die ubiquitäre Abhängigkeit der Gesellschaft von Software, die häufig mangelhafte Qualität und die genannten komplexen Abhängigkeiten der Systeme untereinander, wird es offensichtlich, dass Software eine strategische Dimension bekommen hat. Das wirft die Frage auf, warum sich in den gut 50 Jahren seit der NATO-Konferenz nicht viel bewegt hat. Verändert hat sich nur der Wirkungsradius der Softwarekrise. Wenngleich die Softwarebranche sehr wohl dazugelernt hat, hat sie die zunehmende Fähigkeit, Software zu entwickeln, aber nicht dazu genutzt, bestehende Anforderungen besser umzusetzen, sondern immer komplexere und größere Projekte in Angriff genommen, die ständig über der Schwelle der Überforderung liegen. Mit anderen Worten: Die Komplexität der Softwaresysteme ist schneller gewachsen als die Fähigkeit, sie im Griff zu halten (siehe Abbildung 3).

„Der menschliche Verstand ist nicht dazu geschaffen,
komplexe Systeme zu verstehen .“ *Rupert Riedl*

Auch wenn die Aufgabe, den Kontrollverlust der IT wieder einzufangen, alles andere als trivial ist, scheint das Ziel klar zu sein: Von Softwaresystemen, die alle wesentlichen Funktionen der modernen Welt steuern, kann man erwarten, dass sie resilient und im Idealfall antifragil sind: Stress sollte die Systeme also ständig besser machen. Robustheit, das Zurückziehen auf einen klar definierten Systemzustand, ist in dem sich ständig ändernden Techno-Ökosystem meist kein taugliches Mittel mehr. Vielmehr sind Prozesse notwendig, die das stete Prüfen, Anpassen und Weiterentwickeln aller Teile und Module sicherstellen.

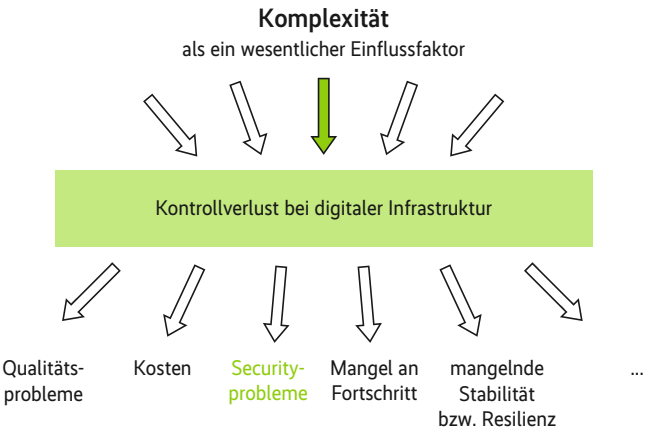
Schneller reagieren und durchhalten

Auch sollten die Systeme möglichst lose gekoppelt sein. Im Idealfall darf sich kein Teil darauf verlassen, dass die anderen zuverlässig funktionieren. Denn die Annahme, dass Stabilität und Sicherheit erreichbare Zustände sind, ist bei verteilten, komplexen Systemen bereits im Kern irrig. Manche Securityexperten geben daher den Rat: „Assume breach“ oder „Assume you are hacked“ – nehmen Sie an, dass Ihr System bereits gehackt ist oder jederzeit gehackt werden kann. Dann stellt sich die Frage: Wie lassen sich katastrophale Fehler oder Datenverluste unter dieser Unsicherheit minimieren und wie Systeme sich resilient gestalten, wenn Planungssicherheit in keiner Weise gegeben ist?

„There is no law against building something
without understanding it.“
George Dyson [3]

Daraus ergibt sich die Notwendigkeit, in mehreren Zeitachsen zu denken. Das Motto eines großen US-Unternehmens „Be fast and break things“ ist eine verheerende und der Gesellschaft gegenüber respektlose Idee. Das richtige Motto müsste heißen: „Be fast and don’t break things.“ Es gilt also, einerseits langfristig zu denken, denn wesentliche IT-Infrastruktur und Daten leben viel länger als häufig angenommen und müssen über lange Zeiträume evolutionär gepflegt werden. Hier verbietet es sich, ständig den neuesten technischen Hypes nachzuspringen. Andererseits muss, wenn man Innovation und evolutionären Fortschritt möchte, eine schnellere, explorativere Zeitachse der Entwicklung möglich sein.

Beide Zeitachsen müssen aber technisch zusammenspielen und ineinander übergehen. Die Infrastruktur benötigt Reifung und ständige Wartung. Wer nicht bereit ist, das zu finanzieren, zahlt später das Mehrfache für die Folgefehler. Innovation und



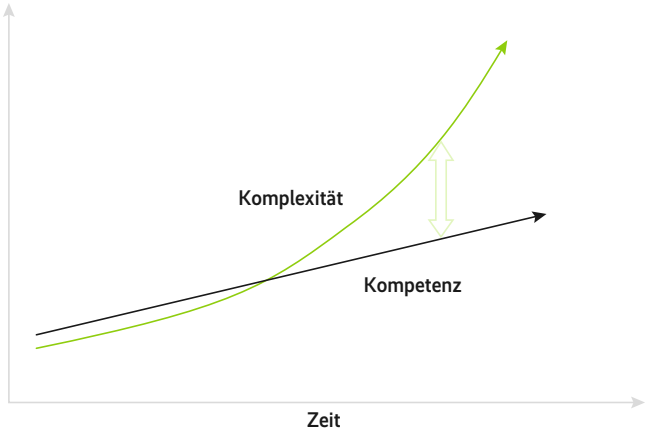
Der Kontrollverlust über die digitale Infrastruktur hat vielfältige Folgen (Abb. 2).

neue, explorative Funktionen müssen auf reifer und gut gewarteter Infrastruktur aufsetzen. Zum langfristigen Denken gehört es auch, die strategische Dimension von Software zu erkennen. Architektonische Entscheidungen – etwa ob und welche Cloud zu verwenden ist, welche Abhängigkeiten man zulassen möchte, wie Security und Qualitätssicherung zu implementieren sind – sind keinesfalls rein technische Fragen, sondern haben eine politische und strategische Dimension.

Was tun?

Funktionsweisen technischer Systeme müssen verstehbar und von Menschen interpretierbar bleiben. Hat man das Verhalten der eigenen Systeme nicht mehr im Griff, kann man kaum von Ingenieurwesen sprechen, geschweige denn formale oder rechtliche Zusagen einhalten. Emergentes Verhalten ist bei komplexen Systemen allerdings unvermeidbar. Das aber bedarf neuer Steuerungsmethoden. Daraus folgen technisch-architektonische und soziale Überlegungen (siehe Kasten „Strategien, um dem Kontrollverlust entgegenzuwirken“).

Aus technischer Sicht ist es wesentlich, dass man auf die inhärente Komplexität der zu bewältigenden Aufgabe nicht noch zusätzlich eine technische Komplexität packt. Davon ist die Praxis aber oft weit entfernt: Teams suchen oft nach einfachen und schnellen Lösungen, die sich aber im Nachhinein als komplex und kaum verständlich herausstellen.



Während die Kompetenz linear steigt, nimmt die Komplexität der Systeme exponentiell zu (Abb. 3).

Den Grund hat Rich Hickey auf der Strange Loop 2011 in seinem Vortrag „Simple Made Easy“ herausgearbeitet, in dem er simpel und einfach gegenüberstellt (siehe ix.de/z7d4): Einfach oder easy ist das, was den Entwicklern einfach oder naheliegend erscheint, etwa wenn sie auf Frameworks, Programmiersprachen oder Clouds zurückgreifen, die sie kennen. Simpel dagegen bedeutet, ein System zu bauen, in dem die technischen und architektonischen Mittel der Aufgabe und nicht dem aktuellen Kenntnisstand des Teams angepasst sind. Dies ist in der Praxis eine erhebliche Herausforderung, vor allem wenn Altsysteme zu berücksichtigen sind. Der schnelle und einfache Weg erweist sich dann meist schon mittelfristig als teuer und schwer wartbar.

Zu vermeiden ist aber nicht nur die Komplexität, weil sie zu unerwarteten Effekten führt, sondern auch die Kompliziertheit: Was Menschen nicht mehr verstehen, können sie auch nicht warten und verändern. Besonders wenn Systeme wachsen, sind häufig größere Refactorings durchzuführen und zu budgetieren, um die Gesamtverständlichkeit weiter zu gewährleisten. Denn Quick Fixes und Workarounds bekommt man kaum mehr in den Griff. Programmierpraktiken wie Pair Programming und Code Reviews helfen, weil sie das Wissen so besser im Team verbreiten und die Gefahr reduzieren, dass sich komplizierte Lösungen, die nur Einzelne verstehen, etablieren.

Besonders bei neuen Aufgabenstellungen ist nicht zu erwarten, dass ein Team ein simples Ergebnis in der ersten Iteration entwirft. Prototypen auf verschiedenen Ebenen – Papier, nicht funktional, funktional – sind hier ein möglicher Zugang, um sich dem Ziel zu nähern. Dabei darf man einer Verlockung nicht verfallen: Aus dem Prototyp will man für die Implementierung lernen. Nimmt man Prototypen in Betrieb – nach dem Motto „funktioniert schon weitgehend“ –, hat man meist eine schwere technische Schuld mitgenommen, die man später teuer bezahlt.

Entwicklungsgeschwindigkeit und Lebensdauer zusammenbringen

Eine soziale Herausforderung, der sich Entwickler und Betreiber gemeinsam stellen müssen, ist: Wie kann es gelingen, die unterschiedliche Zeitlichkeit abzubilden? Innovation auf dem Markt benötigt Geschwindigkeit und Agilität. Das Zusammenspiel mit langsameren etablierten Systemen muss dabei stets gewährleistet bleiben. Laufen diese beiden auseinander, führt dies zu großen technischen und menschlichen Reibungsverlusten.

Zwar gibt es Ansätze wie DevOps, die Entwicklung und Betrieb einander näher bringen sollen, doch dürfte das letzte Wort hier noch nicht gesprochen sein. Derzeit arbeiten die einen mit eher schwergewichtigen und langsamen Prozessen – so wenig Änderung wie möglich –, die anderen möchten am liebsten weitgehend ohne formale Prozesse ihre neuen Features in Betrieb nehmen, notfalls an den entsprechenden Betriebsabteilungen vorbei. So kann es keine nachhaltige Softwareentwicklung geben.

Google bezeichnet das als Konflikt zwischen Initial und Sustained Velocity, also zwischen der initialen Geschwindigkeit, ein Feature oder Produkt auf den Markt zu bringen, und der nachhaltigen Geschwindigkeit, dieses Produkt auch langfristig warten und weiterentwickeln zu können. Letzteres ist nur möglich, wenn frühzeitig Aspekte wie Wartbarkeit, Sicherheit und Zuverlässigkeit bedacht werden.

In den letzten Jahren erleben – aus guten Gründen – alte Programmierkonzepte eine Renaissance, die in neue Programmiersprachen gegossen werden, etwa simplere Sprachen mit stärkerer Rigidität wie Go und Rust oder Sprachen wie Clojure,

die sich stark an klassische funktionale Sprachen anlehnen. Die Form der Objektorientierung, die in den 1990er- und 2000er-Jahren modern war, stellt sich für viele Anforderungen als Irrweg und Überdesign heraus, ebenso wie die als Gegenreaktion begreifbaren „schlampigen" Skriptsprachen.

Auch die datenorientierte Programmierung tritt wieder stärker in den Vordergrund aus der Erkenntnis, dass Daten zumeist bedeutsamer, langlebiger und besser verständlich sind als Code. Operationen finden dabei funktional auf Daten statt, um möglichst wenig innere Variabilität zu verursachen und so leicht verständlich und testbar zu bleiben. Dies soll auch beim State Management helfen, einem schwer zu lösenden Problem, das besonders in verteilten Anwendungen für ein großes Maß an technischer Komplexität verantwortlich ist.

Strategien, um dem Kontrollverlust entgegenzuwirken

Versuchungen widerstehen: Schnelle und einfache Wege wie der Rückgriff auf vertraute Werkzeuge, Quick Fixes und das Inbetriebnehmen von Prototypen bezahlt man später teuer.

Den Konflikt von Entwicklung und Betrieb auflösen: Die Fähigkeit, das Produkt auch langfristig zu warten und weiterzuentwickeln, ist trotz des Bestrebens der schnellen Markteinführung während der Entwicklung zu berücksichtigen.

Programmierkonzepte überdenken: Funktionale und datenorientierte Programmierung mit rigideren Sprachen sollte in vielen Fällen objektorientiertes Überdesign ablösen.

Mit Maß modularisieren: Form und Ausmaß der Modularität haben einen hohen Einfluss auf die Komplexität und den Ressourcenverbrauch.

Die Fragilität der Clouds berücksichtigen: Public Clouds als zentralisierte Angebote führen zu neuen Single Points of Failure und zu neuen Abhängigkeiten.

Sich dem Dilemma der Wiederverwendung stellen: Das Verwenden von Bibliotheken spart Arbeit und Fehler, erzeugt aber Abhängigkeiten und Kontrollverlust.

Software Gardening betreiben: Statt an mechanistischen Bildern festzuhalten, sollte man Softwaresysteme als Ökosysteme begreifen.

Modularität richtig dosieren

Die richtige Form und Skalierung der Modularisierung zu finden und zu entscheiden, welche Module man selbst in der Hand haben möchte und welche man als externe Abhängigkeiten hinzunimmt, ist eine der größten Herausforderungen der Softwareentwicklung. Docker für die Modularisierung zu verwenden, bringt zahlreiche Vorteile für Entwicklung und Betrieb. Mit der Containerisierung lassen sich Module sehr stark voneinander entkoppeln und Entwicklung, Konfigurationsmanagement und Tests vereinfachen. Auch können die Container in unterschiedlichen Clouds laufen und damit ein Vendor Lock-in verhindern. Auf der anderen Seite handelt man sich mit Docker und den notwendigen Tools eine hohe Komplexität ein, die die Anwendung erst rechtfertigen muss. Außerdem benötigen all diese Komponenten erhebliche zusätzliche Systemressourcen.

Auch ist die richtige Granularität an Modularisierung nicht einfach zu finden. In den letzten Jahren – Stichwort Microservices – begingen viel Teams den Fehler, die Services zu fein zu schneiden. Dadurch mag zwar der individuelle Service einfach zu verstehen sein, aber die Gesamtanwendung ist eine hochkomplexe Interaktion zahlreicher Microservices. Damit führen Microservices zu einer Makrokomplexität. Erschwerend kommt hinzu, dass sich die Form der Modularisierung bei wachsenden Anwendungen anpassen muss.

Das Nutzen von Cloud-Computing zum Modularisieren und Auslagern von Funktionen, vor allem wenn es in enger Kopplung verwendet wird, ist ein äußerst zweischneidiges Schwert. Cloud-Computing ist ohne Zweifel ein einfacher Weg, Anwendungen zu gestalten, es ist aber auch ein Weg, der oft zu einer hohen Komplexität und in kaum zu handhabende technische und ökonomische Abhängigkeiten führt.

Die Idee der Internetprotokolle war eine weitsichtige und resiliente: Kein Knoten sollte den Ausfall des gesamten Netzes oder großer Teile verursachen. Neue Teilnehmer lassen sich einfach an dieses Netz anschließen. Mit diesen Protokollen wurden mit der Public Cloud wieder zentralisierte Angebote aufgebaut, dominiert von wenigen Monopolisten. Was an der Basis resilient ist, ist auf den höheren Ebenen fragil. Um ein Beispiel zu nennen: 2019 hatte Cloudflare große Schwierigkeiten mit seiner Serverkapazität. Da viele andere Dienstleister wie Dropbox, Shopify oder Zendesk Cloudflare nutzen, hat dessen Ausfall kaskadenartig zahlreiche andere Services mitgerissen.

Abhängig von anderen

Ähnliche Ausfälle gab es in der jüngeren Vergangenheit bei Amazon AWS und Microsoft Azure [4]. Damit hat man in vielerlei Hinsicht Resilienz und Nachhaltigkeit gegen Einfachheit und Stabilität im Normalfall getauscht – vor allem in globaler Hinsicht. Statt der vielen kleinen Ausfälle riskiert man heute den Totalausfall großer Teile der IT-Infrastruktur durch solche Abhängigkeiten. Zudem findet mit der Nutzung anbieterspezifischer Cloud-Dienste und -APIs ein technischer Lock-in statt, denn das Wechseln bestimmter Module zu einem anderen Anbieter ist nicht oder nur mit großem Aufwand möglich.

Der letzte Teil der genannten Abhängigkeiten betrifft Bibliotheken. Einerseits möchte man Basisfunktionen nicht zum hundertsten Mal neu schreiben. Das Wiederverwenden über Open-Source-Bibliotheken ist deshalb als Fortschritt zu sehen. Würde jeder Entwickler seine eigene SSL-Bibliothek schreiben, wären Sicherheitslücken an allen Ecken und Enden die Folge.

Andererseits sind noch keine sauberen Prozesse dieser Wiederverwendung etabliert. Dadurch entstehen Großrisiken: Zeigt eine der viel verwendeten Bibliotheken eine kritische Lücke, sind auf einen Schlag Millionen von Produkten betroffen, wie kürzlich bei Log4j zu beobachten. Auch sind manche millionenfach verwendeten Bibliotheken nur von ganz wenigen Programmierern abhängig – mit allen damit verbundenen Risiken.

Auch die Verwendung von Bibliotheken ist in den letzten Jahrzehnten in vielen Projekten außer Kontrolle geraten. Das Warten und Testen wird zum kaum beherrschbaren Albtraum. Die Frage, welche Funktion man sich als externe Abhängigkeit ins Projekt nimmt, wird also zur eminent strategischen Ent-

scheidung, die die Technik gemeinsam mit dem Produktmanagement zu treffen hat.

Dass komplexe Systeme emergente Effekte zeigen können, verändern auch die Art und Weise der Qualitätssicherung und der Systembeschreibung. Klassisches Testen in Stages und unterschiedlichen Ebenen wie Unit, Modul, Integration ist nach wie vor wichtig, aber nicht mehr ausreichend. Da das Verhalten des Systems erst zur Laufzeit in Gänze erkennbar wird, genügen auch die klassischen Betriebsparameter nicht mehr.

Software als Ökotop

Ein modernes Softwaresystem wäre vielmehr als ein Ökosystem zu betrachten, das mit Sensoren im Betrieb ständig funktional überwacht wird. Management und Technik müssen dafür Gesundheitsindikatoren definieren, die zur Laufzeit geprüft werden können. Diese Indikatoren benötigen aber eine Datenbasis. Dies setzt etwa eine zentrale Logging-Strategie und ein geeignetes Event-Handling voraus. Diese Daten helfen später beim Erstellen dieser Indikatoren und Metriken, ebenso beim Verständnis, wie bestimmte Funktionsweisen im System abgebildet sind, zumal es die Verteilung moderner Anwendungen oft schwer macht, überhaupt zu verstehen, von welchen Modulen ein bestimmter Prozess abhängt.

Ist Antifragilität und Resilienz erwünscht, muss dieses Verhalten zur Laufzeit stärker in den Vordergrund rücken: durch das Monitoring der Gesundheit, aber auch durch Herausforderung des Systems im Betrieb. Wenn der Ausfall einzelner Komponenten die Stabilität des Gesamtsystem nicht gefährden soll, ist dies regelmäßig zur Laufzeit zu trainieren. Netflix hat mit dem Chaos Monkey gezeigt, in welche Richtung der Weg geht [5, 6]. Recovery-Pläne sind in dieser Hinsicht ebenso wenig belastbar, wie ein Muskel nicht durch Theorie trainiert wird. Wenn eine Komponente eine andere nicht kompromittieren darf, müssen Privilegien und Vertrauen zwischen Komponenten minimiert werden. Trust no one – weder im Sinne der Funktionsweise noch der Sicherheit – ist jedenfalls ein guter Startpunkt, um ein weiteres Beispiel zu nennen.

Nur auf Basis eines klugen und strategischen Abhängigkeitsmanagements, der genannten automatisierten Test- und Qualitätssicherungspraktiken von der Entwicklung bis zur Laufzeit und der konsequenten Weitergabe und Dokumentation des Wissens ist eine evolutionäre Weiterentwicklung der gesamten Anwendung möglich. Wenn einzelne Legacy-Module verstauben, bis nach Jahren niemand mehr genau weiß, was diese Module bewirken, ist eine Ablösung oder Weiterentwicklung – wie zahlreiche gescheiterte Großprojekte zeigen – nur mehr mit enormem Aufwand möglich.

> „The Code is a liability, not an asset.“ *Google* [7]

Damit steht am Ende die Ablösung von Modulen, das Entfernen von Code, Deprecation genannt. Google pflegt intern das Mantra: „Code is a liability, not an asset“, Code wird als Belastung gesehen, nicht als Gewinn. Je weniger Code für eine benötigte Funktion, desto besser. Überflüssiges ist daher zu entfernen. Alter kann ein Indikator sein, aber kein alleiniges Kriterium für das Entfernen eines Systems. Im Zentrum steht die Frage, ob man ein Modul noch im Griff hat oder ob man die Aufgabe effizienter mit einem neuen lösen kann.

Dies ist aber einer der schwierigsten Aspekte, mit dem die Softwareindustrie noch zu wenig Erfahrung hat. Was man aber beobachten kann, ist, dass Zombie-Systeme, die nicht abgeschaltet werden, weil niemand genau weiß, ob und in welcher Form sie noch benötigt werden, die Folge mangelnden Komponenten- und Infrastrukturmanagements sind und erhebliche Risiken darstellen (siehe Artikel „Licht aus“ ab Seite 58). Im Sinne einer Betrachtung von Software als Ökosystem ist der Tod einer Komponente ebenso wesentlicher Teil der Weiterentwicklung wie das Hinzufügen neuer Funktionen.

Fazit

Die technischen Herausforderungen, resiliente und antifragile Softwaresysteme zu bauen, sind erheblich. Zwar hat die Softwarebranche seit den 1960er-Jahren einiges dazugelernt, aber die Komplexität ihrer Lösungen hat die wachsende Fähigkeit ständig übertroffen. Dies hat zu Kontrollverlust und zum Teil zu extrem schlechter Softwarequalität geführt. Um diese kritische Situation unter Kontrolle zu bekommen, sind aber nicht nur technische Antworten notwendig, sondern auch neue Ansätze in Management und Planung sowie neue Narrative (siehe Artikel „Nur gemeinsam“ ab Seite 48).

Geeigneter und der aktuellen Situation angemessener als mechanistische Bilder ist der Begriff des Software Gardening, also die Idee, Softwaresysteme als Ökosysteme zu begreifen. Auch der Projektbegriff sollte langsam ausgedient haben. Dieser Wandel scheint essenziell, ist aber gleichzeitig wenig verstanden. Auch fehlen bisher teilweise die Werkzeuge und Prozesse, um mit Softwaresystemen als Ökosystem umzugehen. Es ist dringend geraten, sich diesmal die Zeit zu nehmen, die eigenen Praktiken zu verbessern, denn es steht wesentlich mehr auf dem Spiel als vor 50 Jahren. (sun@ix.de)

Quellen

[1] Nassim Nicholas Taleb; Das Risiko und sein Preis: Skin in the Game; Penguin Verlag 2018

[2] Kevin Kelly; Inevitable: Understanding the 12 Technological Forces That Will Shape Our Future; Viking 2016

[3] George Dyson; Analogia: The Emergence of Technology Beyond Programmable Control; Farrar Strauss & Giroux; 2020

[4] Susanne Nolte; Nicht vorhergesehen; Ausfall bei AWS: Wenn Managementclients Produktivsysteme lahmlegen; iX 2/2022, S. 82

[5] Philipp Steevens; Chaos Engineering als Resilienzkonzept; iX 6/2022, S. 118

[6] Casey Rosenthal, Nora Jones; Chaos Engineering: System Resiliency in Practice; O'Reilly 2020

[7] Titus Winters, Tom Manshreck, Hyrum Wright; Software Engineering at Google; O'Reilly 2020

[8] Heather Adkins, Betsy Beyer, Paul Blankinship, Piotr Lewandowski, Ana Oprea, Adam Stubblefield; Building Secure and Reliable Software Systems; O'Reilly 2020

[9] Andy Hunt, Dave Thomas; The Pragmatic Programmer: From Journeyman to Master; Addison-Wesley 1999

[10] Onlinequellen und weiterführende Literatur siehe ix.de/z74d

Dr. Alexander Schatten

ist Senior Researcher bei SBA-Research, Management-Berater und Podcaster: https://podcast.zukunft-denken.eu

ZENTRAL- UND LANDESBIBLIOTHEK BERLIN

WISSEN | BEGEGNUNG | INSPIRATION

Die Stiftung Zentral- und Landesbibliothek Berlin (ZLB) ist die größte öffentliche Bibliothek in Deutschland, die am besten besuchte Kultur- und Bildungseinrichtung Berlins und Teil des Verbundes der Öffentlichen Bibliotheken Berlins (VÖBB). Mit ihren ca. 350 Mitarbeiter*innen bietet die Bibliothek in allen Segmenten ihrer Arbeit ein innovatives und partizipatives Medien-, Beratungs- und Veranstaltungsangebot und entwickelt sich dabei zunehmend zu einer Plattform für die Communities der Stadtgesellschaft. Hier teilen die Bürger*innen ihr Wissen, hier wird die Teilhabe am digitalen Leben genauso wie an informierten städtischen Diskursen ermöglicht. Eine moderne und serviceorientiert ausgerichtete Informationstechnologie mit entsprechenden Digitalen Services werden dabei genauso benötigt wie Methoden der laufenden Organisationsentwicklung und des Changemanagements.

Werden Sie Teil unseres Teams!
Verstärken Sie unsere Direktion Digitale Entwicklung und Verbundangelegenheiten mit Ihrer Expertise als

IT-Projektleiter*in & Informationssicherheits-spezialist*in

Ihr Aufgabengebiet:

- Mitarbeit in und Leitung von informations- und bibliothekstechnologischen Projekten
- Weiterentwicklung von Maßnahmen der Informationssicherheit
- Erstellen von Informationssicherheitskonzepten nach BSI-Grundschutz
- Aufbau und Einführung eines Informationssicherheits-Managementsystem (ISMS)
- Beurteilung von sicherheitsrelevanten Vorfällen und Risiken
- Durchführung Sicherheitsaudits und internen Schulungsmaßnahmen

Leiter*in des Referats IT-Anwendungen

Ihr Aufgabengebiet:

- fachliche Leitung und personelle Führung des Referats
- Neukonzeption von elektronischen Dienstleistungsangeboten in enger Zusammenarbeit mit den Fachbereichen der ZLB
- Strukturelle und konzeptionelle Weiterentwicklung von nutzerfreundlichen und zukunftsfähigen IT-Dienstleistungen
- Weiterentwicklung der Schnittstellen zur IT-Infrastruktur und Bereitstellung der Clientsysteme
- Organisation und Gewährleistung des internen Service-Desk
- Koordination und Leitung bereichsübergreifender Projekte

Für diese Aufgaben suchen wir fachlich sowie persönlich überzeugende Persönlichkeiten, die aktiv die Zukunftsstrategie der ZLB mitgestalten. Sie haben Ihr Hochschulstudium der Informatik oder Informationswissenschaft bzw. in einer vergleichbaren Fachrichtung erfolgreich abgeschlossen und konnten idealerweise berufliche Erfahrungen in einer fachlich entsprechenden Position sammeln. Dabei arbeiten Sie gerne in Netzwerken und stark dienstleistungsorientiert. Konzeptionelle Stärken und Organisationstalent runden Ihr Profil ab.

Sind Sie so jemand?
Dann brauchen wir Sie! Wir freuen uns auf Ihre aussagekräftige Bewerbung per E-Mail an **stellenausschreibung@zlb.de**. Die detaillierten Stellenausschreibungen mit den Anforderungen und weiterführenden Informationen finden Sie auf unserer Homepage **www.zlb.de** unter der Rubrik „Über uns/Jobs".

Organisatorische Einbettung nachhaltiger Software

Nur gemeinsam

Dr. Alexander Schatten

Moderne Techniken allein führen nicht zu einer nachhaltigen Softwareentwicklung, wenn die organisatorischen und politischen Ebenen nicht mitziehen.

Software hat über die Jahrzehnte eine stetig zunehmende strategische Dimension erhalten, und dies auf mehreren Ebenen – von der Unternehmens- über die staatliche bis zur geopolitischen Ebene. Damit richtet sich der Blick zunehmend auf die mittel- und langfristigen ökonomischen Effekte und Schadenspotenziale, die von der wachsenden Komplexität und den globalen technischen und ökonomischen Abhängigkeiten ausgehen. Das Erstellen nachhaltiger Software erlangt damit eine existenzielle Bedeutung für eine moderne Gesellschaft. Hier sind aber nicht nur Techniker gefragt, sondern auch Politik und Management.

Dass Aspekte der Nachhaltigkeit großer Softwaresysteme in der Vergangenheit meist nicht hinreichend beachtet wurden, macht das Umsteuern zunehmend aufwendiger. Bei vielen Unternehmen und Organisationen hat sich eine stetig größer werdende technische und organisatorische Schuld aufgebaut. Solche Schulden müssen wie ökonomische Schulden beglichen werden. Anderenfalls behindern sie Innovation und Weiterentwicklung der IT-Landschaft bis zum völligen Stillstand bei gleichzeitig stetig steigenden Kosten (siehe Abbildung 1).

Der daraus resultierende Verlust der Kontrolle über essenzielle technische Systeme birgt Risiken, die in einer von Software abhängigen Welt nicht mehr als akzeptabel gelten dürfen. Dazu kommt, dass das Beheben dieser technischen Schuld exponentiell schwerer und teurer wird, je länger man wartet. Diese Erkenntnis lässt sich durchaus aus der Erfahrung mit unterschiedlichsten Systemen und Infrastrukturen generalisieren [1, 2].

Neue Narrative für Entscheidungsträger

Das Management denkt häufig immer noch in klassischen Lifecycle- und Planungsmodellen, die mittlerweile aus Softwareengineering-Lehrbüchern verschwunden sind und auch in der Praxis keine Rolle mehr spielen sollten. Der Wandel gestaltet sich aber in der Praxis aufwendiger als erwartet. Software verhält sich nicht wie andere technische Geräte, die beschafft und nach der Abschreibungsfrist ausgetauscht werden. Dieses Bild passt in keiner Weise zu modernen Softwaresystemen, die hochintegriert und verknüpft sind und sich stetig verändern müssen.

Abschalten und Austauschen sind seltene und hoch komplizierte, manchmal auch komplexe Vorgänge, vor allem wenn man die Wartung vernachlässigt hat.

Aber nicht nur die technischen Systeme sind hochintegriert und verknüpft, dies trifft auch auf die Integration in die Organisation zu. Die Einführung agiler Softwareentwicklung allein – um ein Beispiel zu nennen – ändert daher nichts an der Situation, wenn nicht die gesamte Organisation ihren Umgang mit Softwaresystemen verändert. Das erfordert nicht nur einfache Prozessänderungen, sondern einen systemischen Ansatz. Diese schmerzliche Erkenntnis mussten viele Unternehmen in den letzten Jahren machen.

Planung im konventionellen Sinne funktioniert nicht mehr; Innovation und Agilität lassen sich nicht auf eine konventionelle Organisation aufstülpen. Auch die Arten von Risiken verändern sich dramatisch: Isolierte und eingrenzbare Ausfälle weichen zunehmend Kaskadeneffekten, die häufig die Organisation verlassen oder von außerhalb ausgelöst werden. Das gesamte Unternehmen kann zum Stillstand kommen, mit enormen Folgekosten beziehungsweise Datenverlusten. Risiken werden teilweise an kaum einschätzbare und oftmals auch nicht bekannte Partner externalisiert.

Kurz gesagt, Veränderungen in der Softwarelandschaft unter den Projektbegriff zu fassen, wird weitgehend illusorisch. Fast jedes Charakteristikum, das ein klassisches Projekt definiert, ist bei moderner Software nicht mehr gegeben: Es gibt keinen klaren Start- und schon gar keinen Endpunkt. Fix definierte Budgets sind in den meisten Fällen Potemkinsche Dörfer.

Ökosysteme statt technischer Geräte

Ein Bild aus der Jahrtausendwende scheint wesentlich besser auf moderne Software zu passen: Software Gardening [3]. Die Idee: Komplexe Softwaresysteme ähneln eher einem biologischen Ökosystem als einem deterministischen technischen System – also eher einem Teich als einer Dampfmaschine. Hardware, Software und Anwender formen ein System mit komplexen Wechselwirkungen zwischen allen Beteiligten. Daraus folgt, dass Software altert und ständig gepflegt und gewartet werden muss, selbst wenn sich an der Funktionsweise der Software nichts ändert. Der Austausch einzelner Komponenten bleibt selten auf diese beschränkt. Das ändert die Rolle aller Beteiligten und vor allem die des Managements.

Der wichtigste erste Schritt dürfte gegangen sein, wenn das – technische und vor allem fachliche – Management bei Software nicht mehr an Bits und Bytes oder an Maschinen denkt, sondern an einen riesigen und komplexen Garten, an ein Ökosystem, das man nur teilweise unter Kontrolle hat. Die Konsequenzen: Das Systemverhalten im Betrieb ist kaum mehr mit traditionellen Methoden zu beschreiben und vor allem nicht von klassischen Testsystemen hinreichend ableitbar.

Umso schwerer wird es auch, einen Zielzustand zu definieren, im Besonderen, wenn dieser weiter in der Zukunft liegt. Es braucht neue technische und organisatorische Wege, die es erlauben, Softwaresysteme im Betrieb zu beobachten und schnell zu reagieren, wenn Betriebsparameter abweichen. Parameter, die keinesfalls rein technischer Natur sein dürfen, sondern das gewünschte Makroverhalten widerspiegeln.

Das zweite Umdenken: Software und fachliche Dimension sind nicht zwei getrennte Teile, die IT kein Dienstleister der Fachabteilung wie das Fuhrparkmanagement. Stattdessen ist – fast – jedes moderne Unternehmen de facto ein Softwareunternehmen und ist wie ein solches zu führen. Klassisches Requirements-Management hat schon in der Vergangenheit viele Projekte zum Scheitern gebracht und muss sich unter diesen Rahmenbedingungen neu erfinden. Evolutionäre Weiterentwicklung des Gesamtsystems bedarf anderer Managementparadigmen und Budgetierungsprozesse.

Drittens sind Zeitlichkeiten neu zu denken: Auch wenn es gern vergessen wird, der Großteil des Aufwands fließt in Wartung, nicht in Neuentwicklung oder Innovation. Dabei gilt Wartung als Tätigkeit mit niedrigem Status – dieses generelle Phänomen betrifft nicht nur die IT. Dies hat oft übersehene negative Konsequenzen für Innovationen: Neue Funktionen müssen letztlich mit bestehender Infrastruktur interagieren und in sie integriert werden. Hat man die notwendige evolutionäre Wartung des Gesamtsystems nicht im Griff, geht das nicht nur zulasten der Sicherheit, Zuverlässigkeit und der Kosten. Das einzig Nachhaltige an der IT ist dann die Tatsache, dass die Innovation im Keim erstickt wird.

iX-TRACT

- Eines der wesentlichsten Fundamente für nachhaltige Entwicklung von Software dürfte das Etablieren neuer Bilder und Narrative für die organisatorische Einbettung und das Management von Software sein. Geeignet ist etwa das Software Gardening.
- Bereits bei der Entwicklung ist die Frage zu berücksichtigen, wie langfristige Wartung und Handlungsfähigkeit zu gewährleisten sind.
- Lang-, mittel- und kurzfristige Planungen benötigen unterschiedliche Detailgrade und Flexibilität. Planungs- und Budgetierungsrhythmen sollten sich an der Situation und nicht am Kalender orientieren.
- Kurzfristige Effizienzmaßnahmen verursachen häufig langfristige Schäden, da sie die Resilienz und Agilität reduzieren und die Fragilität erhöhen. Anzustreben ist stattdessen ein antifragiles Verhalten, damit die Organisation auch unter wechselnden Rahmenbedingungen erfolgreich sein kann.
- Die stetige Zunahme an Komplexität erhöht den Bedarf an hoch qualifizierten Mitarbeitern, die nur durch ein Bündel an Maßnahmen zu finden und zu halten sind.

Die Fallen der Automatisierung

Digitalisierung ist ein Spezialfall der Automatisierung und bedeutet nicht nur, dass Maschinen die Tätigkeit von Menschen übernehmen, sondern das in aller Regel auch noch in anderer Weise, denn eine kluge Automatisierung bildet nicht einfach analoge Prozesse ab. Die vormaligen Abläufe und das dafür notwendige Wissen werden in Folge zwischen Mitarbeitern kaum mehr weitergegeben. Die Automatisierung soll zwar die Produktivität und Qualität verbessern, führt aber gleichzeitig zum Verlust an internem Know-how und einer Zunahme an externer Abhängigkeit. Verlust an internem Know-how deshalb, weil die Automatisierung häufig von externen Dienstleistern zugekauft wird und das interne Know-how damit an diese ausgelagert wird. Je komplexer und verteilter die Automatisierungslogik ist, desto schwieriger wird es, die strategische Bedeutung einzuschätzen.

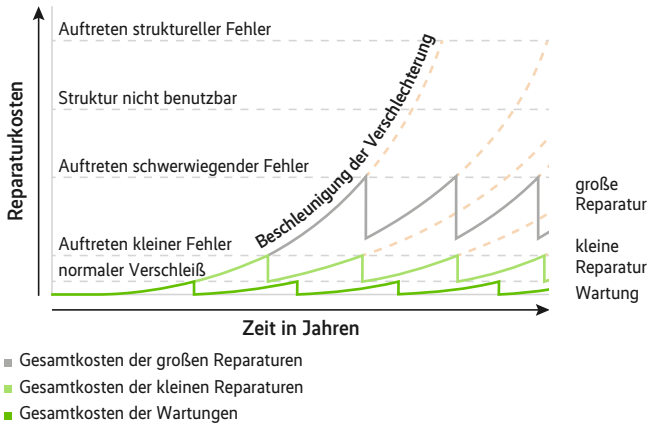
Jedes Digitalisierungsprojekt muss also auch von strategischen Maßnahmen und Überlegungen begleitet werden, ob und wie dieser Know-how-Verlust abgefedert oder verhindert werden soll, um langfristige Wartung und Handlungsfähigkeit zu gewährleisten. Dabei kann es Prozesse geben, die über ein Unternehmen oder eine Organisation hinaus so standardisiert sind, dass es unsinnig wäre, das entsprechende Wissen im eigenen Haus zu halten. Darunter fallen die typische Office-Software, Datenbanken, allgemeine Buchhaltungslogik und dergleichen. Hier ist aber auf politischer Ebene zu beachten, dass es zu keinen geopolitisch bedrohlichen Abhängigkeiten kommt.

Dann wird es aber Prozesse geben, die im Kerngeschäft des Unternehmens oder der Organisation liegen, durch die man sich vom Mitbewerb abgrenzt. Mit vielen Digitalisierungsprojekten kann man vergleichsweise einfach schnelle Erfolge erzielen, aber nur selten davon langfristig profitieren. Man läuft sogar Gefahr, die kurzfristigen Erfolge langfristig teuer zu bezahlen. Werden Prozesse automatisiert, die im Kerngeschäft der Organisation liegen, sind Mechanismen zu etablieren, die sowohl das technische als auch vor allem das fachliche Know-how erhalten. Anderenfalls ist mittel- und langfristige Wartung und spätere Ablösung nicht möglich.

Neue Machine-Learning-Ansätze und breite Skalierungen verschärfen die Situation, weil Entscheidungsfindung und Lernprozesse oft nicht mehr unmittelbar nachvollziehbar sind. Googles Techniker stellen aus mittelfristiger Erfahrung fest: „Die Entwicklung und das Deployment von ML-Systemen ist relativ schnell und billig möglich. Die Wartung über die Zeit ist allerdings schwierig und teuer.“ [4]

Planerische Komplexität durch unterschiedliche Zeitebenen

Überflüssig ist eine langfristige Detailplanung in komplexen Strukturen, die in einer sich schnell ändernden Umgebung operieren müssen. Der Versuch führt zu Potemkinschen Planungsdörfern, die in der Praxis eher Schaden verursachen. Diese immer noch etablierte Form der Planung und Budgetierung benötigt zwar viele Ressourcen, schafft aber nur eine Scheinsicherheit. Dabei werden Strukturen etabliert, innerhalb derer man kaum in der Lage ist, auf neue Herausforderungen zu reagieren, und die das tatsächliche Verhalten im Unternehmen nicht abbilden.



Je weiter Wartungen und Reparaturen aufgeschoben werden, desto größer die Folgen, vom technischen Schaden zum strukturellen Ausfall (Abb. 1).

Im Zentrum modernen Managements steht daher ein geschickter Umgang mit unterschiedlichen Zeitebenen: Lang-, mittel- und kurzfristige Planungen benötigen unterschiedliche Detailgrade und Flexibilität. Nur in zeitlicher Nähe lassen sich einigermaßen belastbare Details festlegen. Dabei müssen sich auch unternehmerische Steuerungswerkzeuge den neuen Methoden anpassen. Die Umstellung auf agile Softwareentwicklung mit Prozessen und Prinzipien wie Scrum oder eXtreme Programming hat bereits vor fast 20 Jahren begonnen. Über die Zeit wurde aber vielen Unternehmen schmerzhaft bewusst, dass diese Prozesse weder in etablierte Management- noch in Reporting-Mechanismen passen und auch nicht gut skalieren. Was für ein kleines Team oder Start-up passt, funktioniert nicht zwangsläufig in großen Organisationen.

Sobald mehrere Teams für eine Softwarelandschaft zusändig sind, kommt die Komplexität wieder durch die Hintertür hereinspaziert. Abstimmungen zwischen Teams sind notwendig, um das Gesamtsystem konsistent weiterzuentwickeln. Dies wirkt gegen die Kernideen von Scrum. In den letzten Jahren wurden große Anstrengungen unternommen, diese agilen Prozesse zu skalieren. Prozesse wie das SAFE-Framework und agile Zieldefinitionen über OKRs (Objectives and Key Results) versuchen die unterschiedlichen Ansprüche zu integrieren [5].

Das scheint in die richtige Richtung zu weisen. Dennoch gibt es zahlreiche Hürden zu überwinden, etwa die in Unternehmen etablierten Standards wie COBIT und ITIL an agile Softwareentwicklung, evolutionäre Wartung, Software Gardening und DevOps anzupassen. Im Augenblick prallen hier noch sehr unterschiedliche Welten aufeinander. Dabei könnte es sich als sinnvoll erweisen, sich von lieb gewordenen Begriffen zu trennen, etwa vom Begriff Projekt – anderenfalls ist das Kind bereits in den Brunnen gefallen, bevor das „Projekt“ überhaupt begonnen hat.

Jenseits der Budgets

Mit der Planung geht in aller Regel die Budgetierung Hand in Hand. In zahlreichen Unternehmen gehört das Erstellen von Jahresbudgets immer noch zum Standard. Damit will man drei Ziele erreichen, die miteinander im Widerspruch stehen: Prognose, Zieldefinitionen und die Allokation von Ressourcen. Zudem treffen Prognosen und Zieldefinitionen in komplexen und von sich ändernden äußeren Bedingungen geprägten Umfeldern langfristig so gut wie nie zu. Die Initiatoren der agilen Softwareentwicklung haben dies bereits vor mehr als 20 Jahren erkannt. Sie schreiben im Manifesto for Agile Software Development: „Reagieren auf Veränderung ist wichtiger als das Befolgen eines Plans“ (siehe ix.de/zf9k).

Diese Erkenntnis bleibt nicht auf Softwareentwicklung beschränkt. Es gibt seit Längerem eine Beyond-Budget-Bewegung, die in den letzten Jahren deutlich an Schwung gewonnen hat. Ihr Ziel ist im Grunde ein ähnliches wie das der agilen Softwareentwicklung: Statt mit großem Aufwand Pläne zu erstellen, die der Realität nicht standhalten, kombiniert man die Planung in vernünftigem Detailgrad mit zwölf definierten Managementprinzipien.

Es würde zu weit führen, die Prinzipien an dieser Stelle zu detaillieren, aber die Überschneidungen mit den Ideen der agilen Softwareentwicklung sind bemerkenswert: Ziel, Werte und Kundenfokus stehen im Vordergrund, zudem die Transparenz, verbunden mit der Autonomie aller Akteure. Hinzu gesellen sich schlanke Planung und Planungsrhythmen, die sich an der Situation und nicht am Kalender orientieren, außerdem Per-

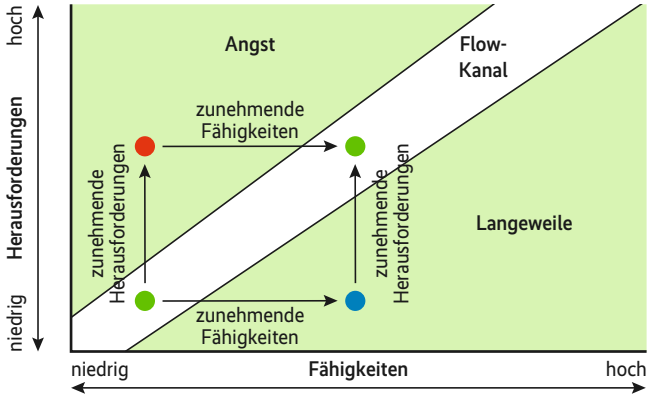
formanceevaluierungen, die sich nicht in trivialisierten und kontraproduktiven KPIs (Key Performance Indicator) erschöpfen.

Jedes Unternehmen oder jede Organisation wird hier einen Weg finden müssen, der zu den eigenen Herausforderungen passt. Es soll an dieser Stelle betont werden, dass Planung und Management von Softwaresystemen nur als Teil einer Gesamtstrategie erfolgreich sein kann. Nur ein gut gepflegtes Softwareökosystem kann die bestehenden Anforderungen effizient erfüllen und neue schnell umsetzen. Budgetierung von Software, die dem Bild des Software Gardening folgt, erfordert andere Prinzipien als Planungszyklen, die in der entfernteren Vergangenheit möglicherweise noch angemessen und funktional waren.

Langfristige Schäden kurzfristiger Effizienzmaßnahmen

Viele Unternehmen verfolgten in den letzten Jahrzehnten das Ziel der Effizienzsteigerung. Die aber ist häufig ein zweischneidiges Schwert und viel schwerer umzusetzen, als es auf den ersten Blick scheinen mag. Häufig falsch gemacht, indem man etwa auf wenige Spezialfälle und auf kurzfristige Effekte optimiert, führt die Effizienzsteigerung zu einer Reduktion von Resilienz und Agilität. Dabei wird kurzfristig gespart, langfristig aber das Vielfache an Kosten generiert und oftmals der Organisation schwerer Schaden zugefügt.

Eine Organisation, die unter wechselnden Rahmenbedingungen erfolgreich sein möchte, strebt antifragiles Verhalten



Der Flow nach Mihály Csíkszentmihályi bildet den Grat zwischen Über- und Unterforderung (Abb. 2).

an. Ziel ist es, dass Stress das System ständig besser macht. Dafür benötigt sie gute Rückkopplungsmechanismen, die es erlauben, aus Erfahrung zu lernen. Doch theoretische Resilienz ist ein Papiertiger. Deshalb ist dieser Stress in die täglichen Abläufe zu integrieren.

Auf technischer Ebene macht es das Chaos Engineering vor. Es testet die Maßgabe, dass keine einzelne Komponente das Gesamtsystem nennenswert beeinträchtigen darf. Der von Netflix entwickelte Chaos Monkey etwa deaktiviert zufällige Komponenten zu zufälligen Zeitpunkten. Jeder Entwickler und Manager weiß daher, dass es keine Ausreden gibt, wenn die eigene Komponente keine geeigneten Fallback-Mechanis-

men implementiert hat. Andere technische Verfahren, komplexe Systeme resilienter zu gestalten, bietet die Zero-Trust-Security.

Aus übergeordneter Sicht ist es dabei wesentlich, dass kein Teil eines resilienten Systems zu stark von anderen abhängt. Dafür ist ein gutes Maß an Diversität und Redundanz notwendig. Steht die Organisation vor neuen Herausforderungen und Krisen, erlauben Diversität, Redundanz und Entkopplung, mit diesen flexibel umzugehen und alternative Pfade zu bestehenden Prozessen zu finden. Anders hingegen Organisationen, die zur Steigerung der Effizienz jede Redundanz entfernt und sich der Stabilität verschrieben haben: Stabilität ist in einem dynamischen Umfeld nicht nur unerreichbar, sondern kontraproduktiv, sie führt zu Erstarrung.

Diversität und kritische Personalfragen

Diversität, Redundanz und alternative organisatorische Pfade sind nur mit Mitarbeitern möglich, deren Wissen sich nicht zu eng auf einzelne Details beschränkt. Auch hier gilt: Kurzfristig mag die Effizienz höher sein, wenn sich Mitarbeiter hochgradig spezialisieren, mittel- und langfristig ist dies gefährlich. Deshalb gilt etwa beim eXtreme Programming die Vorgabe des Pair Programming [6]. Hier arbeiten zwei Programmierer am selben Computer gemeinsam an einer Aufgabe. Das garantiert zwei Sichtweisen und verteilt das Know-how im Team.

Stress spielt auch bei den Fragen der Personalführung eine wesentliche Rolle. Das Ideal ist wohl der von Mihály Csíkszentmihályi definierte Flow-Zustand: Die Fähigkeit eines Mitarbeiters sollte sich mit den Anforderungen im Gleichgewicht befinden (siehe Abbildung 2). Das aber ist leichter gesagt als getan [7].

Den neuen Gegebenheiten müssen auch das Recruiting und der Umgang mit den Mitarbeitern Rechnung tragen. Die stetige Zunahme an Komplexität erfordert hoch qualifizierte Mitarbeiter, die aber schwer zu finden sind. Die Folge davon ist ein weltweiter Wettbewerb um die besten Köpfe. Unternehmen müssen in der Lage sein, Mitarbeiter selbst zu qualifizieren und dann auch zu halten.

Hoch qualifizierte Mitarbeiter erwarten aber zu Recht andere Arbeitsbedingungen, als viele Unternehmen zugestehen wollen. Selbstständig agierende Mitarbeiter wollen keine Vorschriften, wann sie im Homeoffice arbeiten oder ob sie auf eine Konferenz fahren dürfen. Solche Vorschriften lassen eher auf tiefer liegende kritische Probleme in der Organisation schließen, wenn sich das Management genötigt fühlt, Mitarbeiter wie Kleinkinder zu behandeln.

Die Open-Source-Community und die Covid-Pandemie haben vielfach gezeigt, dass auch verteilte und heterogene Teams erfolgreich arbeiten können. Klassisches Management hat begonnen, die Lehren daraus zu ziehen. Das sollte aber nicht ins Gegenteil umschlagen, denn die Verteilung des Wissens und organisatorische Fähigkeit sind wesentlich wichtiger als die Kompetenz Einzelner. Google nennt dies den Genius Myth, den Genie-Mythos. Mit Genies lässt sich keine nachhaltige Software betreiben, mit gut geführten Organisationen schon.

Risiken verteilter Arbeit und Abhängigkeiten

Dennoch löst sich der eklatante Fachkräftemangel nicht durch einfache organisatorische Maßnahmen in Luft auf: Besonders betroffen sind Organisationen und Unternehmen, die scheinbar langweilige, aber für die Gesellschaft wesentliche Software entwickeln und betreiben. Die meisten relevanten Systeme benötigen keine KI, virtuellen Räume oder Gamification, sondern klassisches, gutes Softwareengineering, um Krankenhäuser zu organisieren, Mahnwesen zu betreiben, Buchhaltung abzuwickeln oder Häuser und Mieten zu verwalten.

Beheben lässt sich der Fachkräftemangel vermutlich nur durch ein Bündel an Maßnahmen: Die Kosten für Software werden steigen – schneller, wenn sie nicht nachhaltig entwickelt wird –, Teams werden international aufgestellt sein und Arbeitsbedingungen flexibler werden, und dies nicht nur für große Konzerne, sondern auch für kleine und mittlere Betriebe. Möglicherweise ergeben sich hier auch neue Formen der Zusammenarbeit betroffener Unternehmen.

Die zunehmende Verteilung der Arbeitsleistung, aber auch die Abhängigkeit von externen Systemen und Bibliotheken werfen völlig neue Fragen und Risiken für Technik und Management auf: Bei heutigen Softwareprojekten bildet die Eigenentwicklung oftmals nur einen Bruchteil des gesamten Anwendungscodes. Den Rest liefern Bibliotheken und die Infrastruktur in Form von Komponenten-Frameworks, Datenbanken, Message-Brokern und dergleichen.

Eine immer größere Rolle spielt Open-Source-Software, die vieles in der Softwareentwicklung vereinfacht und standardisiert, aber auch neue Risiken eröffnet. Dabei muss das Management einige Fragen beantworten und die Antworten in Prozesse gießen: Wer entscheidet, welche Softwarelizenzen für die interne Verwendung zulässig sind? Wer entscheidet über das Verwenden neuer Bibliotheken und trägt die architektonischen Konsequenzen? Gerade unerfahrene Entwickler neigen dazu, für jede Kleinigkeit auf neue Bibliotheken zurückzugreifen, ohne darauf zu achten, ob diese stabil, sicher und gewartet sind.

Sicher ist, dass jede externe Abhängigkeit auch – zum Teil erhebliche – Risiken birgt und daher ständig gewartet und beobachtet werden muss. Aber wer stellt sicher, dass all diese Abhängigkeiten aktualisiert und gewartet werden? Je mehr Abhängigkeiten im Unternehmen, desto komplexer wird diese Frage.

Eigen- und Fremdentwicklung – eine Frage der Strategie

In der Praxis sieht man viele Projekte, die über lange Zeit ihre Abhängigkeiten nicht aktualisieren, weil sie entweder keinen sauberen Entwicklungsprozess haben oder weil sie Angst vor den Seiteneffekten auf den eigenen Code haben. Es fehlen vielleicht gute – automatisierte – Regressionstests, die Probleme mit Infrastruktur- oder Bibliotheks-Updates aufzeigen würden. Stattdessen bleibt man aus Angst lieber auf einem völlig veralteten Softwarestand. Eine fatale Situation, die mit der Zeit nicht besser wird, die Softwarequalität beschädigt und große Sicherheitsrisiken birgt. Und nicht zuletzt: Sind stabile Prozesse etabliert, die nicht nur die Richtlinien formulieren, sondern diese auch prüfen?

Die fundamentale Frage, die das Management entscheiden muss: Welche Teile entwickelt man selbst, etwa weil sie von strategischer Bedeutung sind, und bei welchen Funktionen erfindet man das Rad klugerweise nicht neu? All dies betrifft nicht nur Bibliotheken, sondern im selben Maß alle anderen Abhängigkeiten von eingekauften Softwareteilen, Lieferanten, Cloud-Services, aber auch ganz neuen Mechanismen wie Codebibliotheken und KI-Tools für Entwickler, die dem Management nicht entgehen dürfen: Wer ist der Urheber eines Programms, wenn Entwickler mit Copy-and-paste Codeteile aus Stack Overflow kopieren oder ein Tool wie Git Copilote verwenden, ein KI-

Tool, das aus einer großen Codebasis lernt und dem Entwickler Code-Snippets vorschlägt?

Das ist erst der Anfang einer Entwicklung, die Urheberschaft infrage stellt und entsprechende rechtliche, inhaltliche und qualitative Fragen aufwirft, auf die das Management Antworten finden muss. Wer haftet im Konfliktfall? Wie lassen sich SLAs einhalten, wenn große Teile der Funktionen nicht in der eigenen Hand liegen? Die Sache wird nochmals komplexer durch den zunehmenden Kontrollverlust bei Altsystemen und das daraus folgende emergente Verhalten verteilter und skalierter Systeme mit unklaren wechselseitigen Abhängigkeiten.

Fazit: Die strategische und politische Dimension

Software hat in der modernen Welt eine strategische Dimension. Die europäische Politik beginnt erst langsam, die Konsequenzen zu verstehen. IT wird regelmäßig als Innovator gesehen, es zählt das Motto: Digitalisierung zuerst, Bedenken später. Kein verantwortungsbewusster Politiker aber würde bei der Verteidigung oder der Energieversorgung auf diese Strategie setzen.

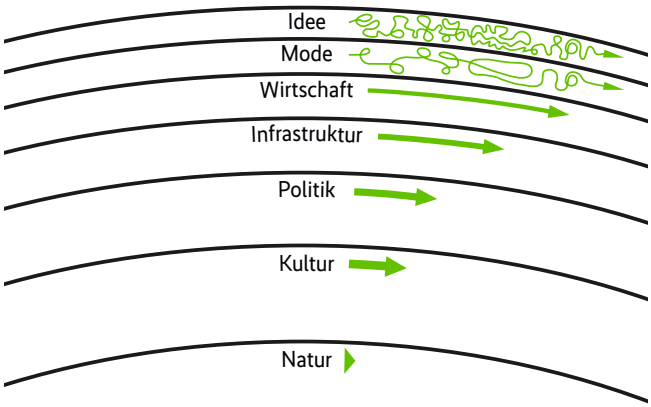
Selbstredend treibt die IT häufig die Innovation an, aber ebenso wichtig ist heute die Infrastruktur. Innovationen werden, wenn sie erfolgreich sind, zur Infrastruktur und sind als solche jahrzehntelang zu warten und zugleich Fundament neuer Innovationen (siehe Abbildung 3). Anders gesagt: Gelingt es nicht, die bestehende IT-Infrastruktur gut zu warten und nachhaltig zu betreiben, wird damit auch jede Innovation im Keim erstickt und das Gesamtsystem zu einem existenziellen Risiko für die Gesellschaft.

> „Das Schnelle lernt, das Langsame erinnert. Schnell schlägt vor, Langsam verwirft. Schnell und Klein leiten Langsam und Groß durch angehäufte Innovation und gelegentliche Revolution. Langsam und Groß kontrolliert Klein und Schnell durch Rahmenbedingungen und Konstanz. Schnell hat unsere Aufmerksamkeit, aber Langsam hat all die Macht."
> *Stewart Brand* [8]

EU und europäische Staaten haben in der Vergangenheit zahlreiche schwere Fehler begangen. Die Gesellschaft steckt heute in Sachen Software und Hardware in einer kritischen Abhängigkeit von zum Teil politisch wenig vertrauenswürdigen Partnern. Ein gutes Beispiel ist die Cloud-Strategie der Europäischen Union. Sie hat sich vor einigen Jahren vom traditionellen Softwarebetrieb verabschiedet und die Cloud als strategisches Ziel definiert. Daraufhin hat die EU DIGIT (Generaldirektion Informatik) Verträge über EU-Services mit vier Cloud-Providern geschlossen, darunter mit den drei US-Firmen AWS, Microsoft und IBM.

In einem Werbevideo für AWS erklärt ein hochrangiger EU-Vertreter, dass unter anderem die Software, die die EU verwendet, um mit allen EU-Bürgern zu kommunizieren, auf AWS läuft (siehe ix.de/zf9k). Offensichtlich haben die entscheidenden EU-Behörden die strategische und langfristige Dimension von Software nicht verstanden. Anders sind solche Fehlentscheidungen kaum zu erklären. Man stelle sich den umgekehrten Fall vor: Ein Spitzenbeamter des US-Department of Homeland Security würde vorschlagen, wesentliche Daten und Funktionen von US-Behörden in die Cloud eines französischen oder deutschen Anbieters auszulagern.

Zumindest scheint auf europäischer Ebene langsam ein Problembewusstsein Einzug zu halten. In den letzten Jahren gab es einige bemerkenswerte Initiativen, darunter die Datenschutz-Grundverordnung, den Chips Act oder den Versuch, KI-Anwendungen zu regulieren. Nun lässt sich trefflich darüber streiten, ob diese Ansätze in der Umsetzung gelungen sind, zumindest ist aber anzuerkennen, dass die geopolitische und strategische Dimension erkannt wurde.



Die unterschiedlichen Ebenen bewegen sich in unterschiedlichen Geschwindigkeiten: Neue Ideen und Moden benötigen eine langlebige Basis, darunter Wirtschaft, Infrastruktur und Politik. Dabei hat jede Ebene eine weitere Ebene über sich, die schneller ist, und eine unter sich, die sie bremst (Abb. 3).

Was im EU-Raum gilt, gilt im kleineren Maßstab für nationale Regierungen und für Unternehmen. Wer die strategische und langfristige Dimension von Software übersieht und daher nicht in der Lage ist, Softwaresysteme nachhaltig zu betreiben, wird in der Zukunft nicht auf der Seite der Gewinner stehen.

(sun@ix.de)

Quellen

[1] Lee Vinsel, Andrew L. Russell; The Innovation Delusion: How Our Obsession with the New Has Disrupted the Work That Matters Most; Currency 2020

[2] Steward Brand; How Buildings Learn: What Happens After They're Built; Viking Press 1994

[3] Andy Hunt, Dave Thomas; The Pragmatic Programmer: From Journeyman to Master; Addison-Wesley 1999

[4] D. Sculley, et al.; Hidden Technical Debt in Machine Learning Systems; NeurIPS Proceedings; Google 2015

[5] Ben Lamorte; Objectives and Key Results: Driving Focus, Alignment, and Engagement with OKRs; Wiley 2016

[6] Kent Beck; Extreme Programming Explained: Embrace Change; Addison-Wesley 2004

[7] Cal Newport; Deep Work: Rules for Focused Success in a Distracted World; Pitkus 2016

[8] Stewart Brand; Pace Layering: How Complex Systems Learn and Keep Learning; JoDS 2018

[9] Alle Onlinequellen siehe ix.de/zf9k

Dr. Alexander Schatten

ist Senior Researcher bei SBA-Research, Management-Berater und Podcaster: podcast.zukunft-denken.eu

Nachhaltig programmieren mit Bedacht

Gratwanderung

Rudolf Meier, Detlef Thoms

Schon mit wenigen eingesparten CPU-Sekunden lässt sich viel bewirken. Dabei ist aber zwischen Einsparung und Mehrverbrauch unterschiedlicher Ressourcen abzuwägen, außerdem zwischen den funktionalen Zielen und der Sparsamkeit der Anwendung.

Auch wenn Anwendungen durch das Bereitstellen wichtiger Kennzahlen dazu beitragen, den CO_2-Fußabdruck von Unternehmen und Organisationen zu reduzieren, ihre Entwicklung und ihr Betrieb generieren Emissionen (siehe Artikel „Bestandsaufnahme“ ab Seite 12). Eine umfassende Bilanz der Treibhausgasemissionen digitaler Systeme muss dabei unterschiedliche Faktoren berücksichtigen – nachzulesen etwa in der Dokumentation des Greenhouse Gas Protocol (siehe ix.de/zpfc). Nur ein Teil der Emissionen entsteht im Betrieb der verwendeten Hardware, einen weiteren verursachen Herstellung, Assemblieren und Transport der Hardware. Hinzu kommen weitere Emissionen, die nicht nur der Entwicklung der Software geschuldet sind, sondern auch der kontinuierlichen Betriebsfähigkeit des Unternehmens, etwa Reisen und Büroräume oder das Nutzen zugekaufter Produkte und Dienstleistungen.

Die Bilanzierung dieser Beiträge zum CO_2-Fußabdruck (Carbon Footprint Contributions) ist komplex und geht über den Rahmen dieses Artikels hinaus. Der Fokus liegt daher auf den Emissionen, die das Rechenzentrum beim Betrieb der Software generiert. Zu den primär genutzten Ressourcen gehören die Rechenleistung, Arbeitsspeicher, Netz und Massenspeicher, die Server und andere Hardware in Rechenzentren zur Verfügung stellen. Die IT-Systeme selbst verlangen zudem sekundäre Systeme, die die Infrastruktur eines Rechenzentrums ausmachen. Den Löwenanteil der zusätzlichen Emissionen produzieren hier die Kühlung und die unterbrechungsfreie Stromversorgung.

iX-TRACT

- Wollen Entwicklungsteams den CO_2-Fußabdruck ihrer Anwendungen reduzieren, können sie sich an die Richtlinien des Sustainable Programming halten.
- Einsparungen sind aber mit Bedacht vorzunehmen und müssen im Einklang mit den funktionalen Zielen der Anwendung stehen.
- Auf den ersten Blick erscheint es wenig, wenn eine Anwendung nur eine CPU-Sekunde und damit 10 Wattsekunden pro Transaktion einspart. Je nach Menge der Transaktionen, Benutzer und der Häufigkeit der Ausführungen kann sich das aber auf mehrere Gigawattstunden summieren.
- Spart man bei einer Komponente ein, kann das den Energiehunger anderer Komponenten steigern.

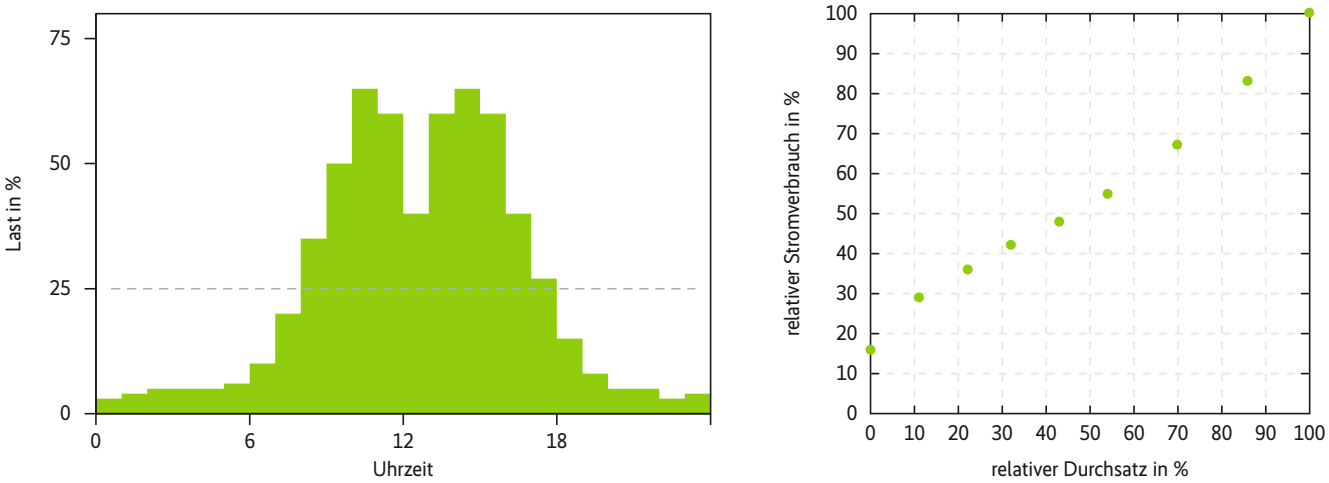
Den CO_2-Fußabdruck einer Software abschätzen

Dargestellt wird die Energieeffizienz eines Rechenzentrums meist mit der PUE (Power Usage Effectiveness) (siehe Artikel „Weiter gefasst“ ab Seite 78). Sie setzt den Stromverbrauch des Gebäudemanagements ins Verhältnis zu dem der IT. Bei einem gemittelten PUE-Wert von 1,5 verbraucht die Infrastruktur halb so viel Energie wie die IT.

$$PUE = \frac{Energieverbrauch\ des\ RZ}{Energieverbrauch\ der\ IT} = 1 + \frac{Energieverbrauch\ der\ Infrastruktur}{Energieverbrauch\ der\ IT}$$

Ein weiterer Wert, die Kohlenstoffemissionseffektivität CEI (Carbon Emission Intensity), ergibt sich aus dem vom Stromanbieter genutzten Mix aus konventionellen und erneuerbaren Energiequellen. Im Jahr 2019 betrug die globale durchschnittliche CO_2-Intensität des erzeugten Stroms 475 g CO_2/kWh (siehe ix.de/zpfc). Das Verknüpfen des elektrischen Energiebedarfs mit der Kohlenstoffemissionseffektivität ergibt den Ausstoß an Treibhausgasen.

Mit den eingeführten Größen lässt sich der CO_2-Fußabdruck einer Software abschätzen. Zur Orientierung dient ein einfaches Modell, das die Einflussfaktoren zeigt. Für das Abschätzen ist erforderlich: die PUE, die CEI, die Energieaufnahme des Servers bei voller Auslastung, das Auslastungsprofil des Servers und die Beziehung von Leistung und Last. Das Auslastungsprofil ist durch das Nutzen der betriebenen Software bestimmt und kann



Um den Energieverbrauch von Software auf einem Server einzuschätzen, braucht es ein Auslastungsprofil (links) und eine Leistung-Last-Beziehung (rechts) (Abb. 1).

für Geschäftssoftware von den Arbeitszeiten der Nutzer abhängen. Die Leistung-Last-Beziehung bestimmt, wie sich die Energieaufnahme des Servers mit dem Grad der Auslastung ändert.

Das Beispiel in Abbildung 1 verdeutlicht dies: Auf der linken Seite zeigt das Aktivitätsprofil für die Systemlast über einen Tag verteilt zwei Spitzenzeiten. Die durchschnittliche Auslastung beträgt 25 Prozent. Für eine Schätzung geht man davon aus, dass das die durchschnittliche Auslastung des Systems für jeden Tag ist. Die rechte Seite zeigt eine Leistung-Last-Beziehung, die aus dem Zertifikat 2016005 einer Ausführung des SAP-Power-Benchmarks stammt (siehe ix.de/zpfc). Der Verlauf dient als Beispiel, er hängt von der Ausstattung der Hardware mit Prozessoren, Arbeitsspeicher und anderen Komponenten sowie der Konfiguration des Betriebs ab. Dargestellt ist die Leistung relativ zur Volllast über der relativen Systemauslastung.

Nimmt man bei Volllast eine Leistung von 0,46 kW an, beträgt die Leistungsaufnahme bei 25 Prozent Systemauslastung 35 Prozent der Aufnahme bei Volllast oder 0,46 kW × 0,35 = 0,16 kW. Hinzu kommt ein angenommener Wert von 1,5 für die Energieeffizienz (PUE) und der Weltdurchschnitt von 475 g/kWh für die CO_2-Emissionsintensität (CEI). Für den kontinuierlichen Betrieb über ein Jahr kommt man damit auf einen geschätzten CO_2-Fußabdruck von einer Tonne CO_2.

$$0,16\ kW\ \times\ 8760\ h\ \times\ 1,5\ \times\ 475\ g\ CO_2/kWh\ =\ 1\ t\ CO_2$$

Diesen Fußabdruck gilt es durch Maßnahmen in der Softwareentwicklung zu verringern.

Energieeffizienz beginnt bei der Softwareentwicklung

Man erhält eine Vorstellung von den Energie- und Emissionsmengen, die das Ausführen von Programmcode braucht, wenn man die Frage aus der Sicht der Entwickelnden stellt: Welchen Fußabdruck hinterlässt das Nutzen einer CPU-Sekunde oder von 10 GByte RAM für eine Sekunde? Die genauen Daten hängen von unterschiedlichen Faktoren ab, unter anderem von der verwendeten Hardware und der Energieeffizienz des Rechenzentrums.

Eine grobe Schätzung ergibt für heutige Serverhardware einen Energiebedarf von je 5 bis 10 Wattsekunden sowohl pro CPU-Sekunde als auch für das Nutzen von 10 GByte RAM für eine Sekunde. Unter Annahme der globalen durchschnittlichen CO_2-Intensität entsprechen 5 bis 10 Wattsekunden etwa 0,7 bis 1,4 mg CO_2. Auch wenn diese Menge klein erscheint, summiert sie sich durch vielfache Nutzung schnell auf.

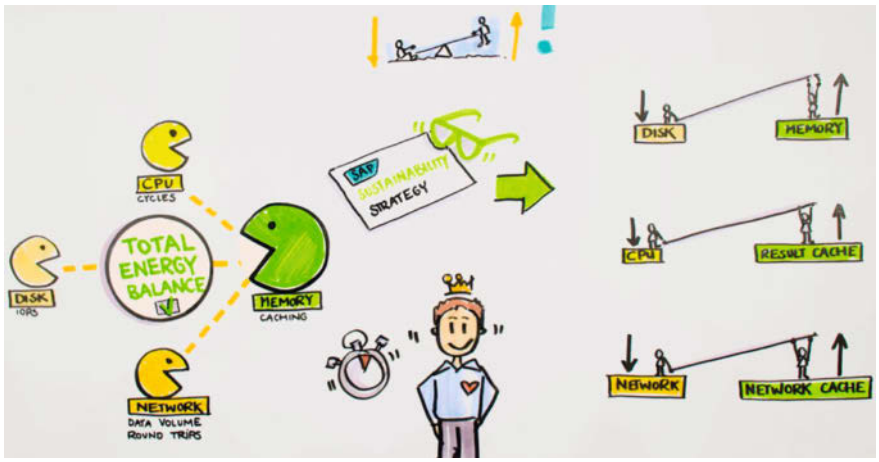
Organisationen sollten deshalb einige Regeln beachten, mit denen sich Emissionen reduzieren oder vermeiden lassen und die unter dem Begriff Sustainable Programming zusammengefasst sind. IT-Teams, die eine nachhaltige Programmierung umsetzen möchten, sollten fünf Regeln beachten:

1. Do what is necessary – nur das Notwendige tun: Im Programmablauf sollten nur die Informationen verwendet und nur die Module und Services aufgerufen werden, die tatsächlich erforderlich sind.
2. Do it once – identische Aufgaben nur einmal ausführen: Informationen, die an verschiedenen Stellen des Programmablaufs nötig sind, sollten nicht durch mehrfache identische Datenbankzugriffe oder Serviceaufrufe beschafft werden.
3. Do it right – von Anfang an richtig programmieren: Man sollte nur den Algorithmus verwenden, der sich am besten für die Aufgabe eignet.
4. Do it to the point – es auf den Punkt bringen: Die Genauigkeit des Algorithmus ist auf die Businessanforderungen abzustimmen.
5. Document – alle Schritte dokumentieren: Nur so lässt sich die Programmierung nachvollziehen und bei Bedarf ändern, ohne mit den Richtlinien 1 bis 4 in Konflikt zu kommen.

Herausforderungen für nachhaltiges Programmieren

Wie alle Prozesse bietet auch das ressourcenbewusste Programmieren einige Herausforderungen. Zum einen sollten die Anwendungen die IT-Ressourcen effizient und damit möglichst nachhaltig nutzen (siehe die Artikel „Durchdrungen“ und „Angepasst“ ab Seite 62 und 67). Das erfordert allerdings eine andere Herangehensweise an das Entwickeln von Applikationen, als viele es bisher gewohnt sind: Es gilt, die richtige Balance zwischen der End-to-End-Antwortzeit, dem Durchsatz, dem wirtschaftlichen und dem nachhaltigen Ressourcenverbrauch zu finden.

Zwar ergeben sich zwischen einer ressourcenbewussten Programmierung und der optimalen Skalierung der Systeme oft



Quelle: SAP SE

Maßnahmen wie In-Memory-Computing und Objektcaches fördern die nachhaltige Programmierung (Abb. 2).

Synergien, etwa für die Verbesserung der Anwendererfahrung und die Reduktion der Energieaufnahme. Allerdings ziehen Ressourcenreduktion und Performanceoptimierung nicht immer am selben Strang: Manchmal lässt sich die Verbesserung der Leistung nur durch einen höheren Ressourcenverbrauch erkaufen.

Zum anderen ist die richtige Balance zwischen funktionalen Zielen und einem sparsamen Umgang mit den Ressourcen zu finden. Unternehmen und ihre IT-Teams müssen hier ihre Prioritäten setzen. Beispielsweise müssen sie beim Entwickeln eines KI-Modells zwischen dem Energieverbrauch für das KI-Training und der Genauigkeit abwägen, die nur mit höherem Aufwand zu erreichen ist und damit mehr Energie bedarf.

Ein KI-Modell beispielsweise, das Anomalien in hochauflösenden Computertomografien möglichst genau erkennen soll, muss mit sehr vielen Falldaten gefüttert werden. Denn nur so erhält man möglichst detaillierte Angaben und kann im Ernstfall Menschenleben retten. Die Entscheidung der Entwickler würde in diesem Fall zugunsten einer höheren Genauigkeit des KI-Systems fallen.

Bei einem Validierungssystem für Dokumente könnte das anders aussehen. Hier ist in Betracht zu ziehen, ob ein energieintensiver Validierungsalgorithmus auf Grundlage eines Blockchain-Modells notwendig ist oder ob ein traditioneller, zentralisierter Ansatz ausreicht, der deutlich weniger Ressourcen benötigt. Die Beispiele zeigen bereits, in welcher Weise Entwickler zwischen dem Energieverbrauch und dem jeweiligen Detailgrad einer Anwendung abwägen müssen. Generelle Empfehlungen gibt es hier nicht, Unternehmen müssen immer im Einzelfall entscheiden.

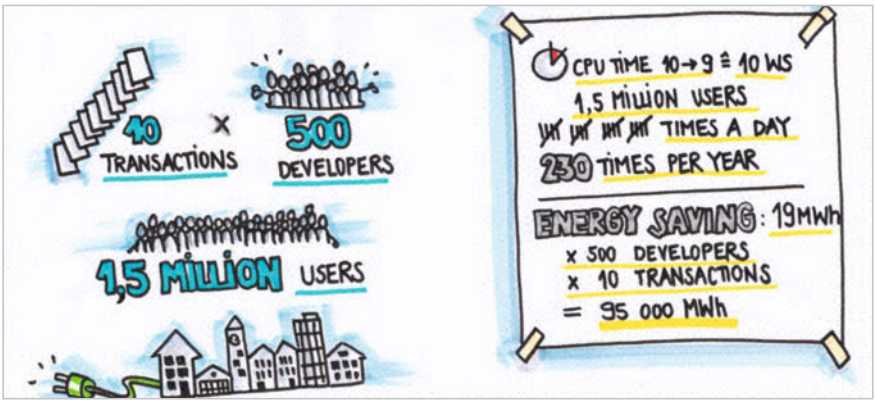
Anders sieht es beim Entwickeln neuer Softwarearchitekturen aus: Neben den funktionalen Zielen muss auch der Ressourcenverbrauch ein wichtiges Entscheidungskriterium sein. Es empfiehlt sich, während des Entwicklungsprozesses den Energieverbrauch verschiedener Softwarearchitekturen zu messen und zu bewerten.

Hier mehr, dort weniger

Der Energiehunger einer Software hängt wesentlich von der CPU-Last, der Memory-Nutzung, von Festplattenzugriffen und dem Netztraffic ab. Bei Unternehmenssoftware hat es sich bewährt, die Aktivitäten auf CPU, Disk und dem Netz durch den Einsatz von mehr Arbeitsspeicher zu reduzieren – diese vier Komponenten bedingen einander (siehe Abbildung 2). Wer beispielsweise den Browsercache nutzt, vermeidet Netztraffic. Indizierte Zugriffe reduzieren oder vermeiden CPU-intensive Datenscans. Objektcaches sparen CPU-Zyklen für den wiederholten Objektaufbau ein und das In-Memory-Computing vermeidet Zugriffe auf Festplatten.

Ein Beispiel: Eine Rechenzentrumsarchitektur für B2B-Applikationen soll nachhaltiger werden und weniger Energie verbrauchen. Wird mit einer entsprechenden Sustainable-Programming-Strategie eine CPU-Sekunde eingespart, entspricht das einer Energiereduktion von 10 Wattsekunden pro Transaktion. Dies mag auf den ersten Blick nicht signifikant sein. Hochgerechnet auf 500 Entwickler und Entwicklerinnen, die jeweils zehn Transaktionen so verbessern, die dann von jeweils 1,5 Millionen Anwenderinnen und Anwendern an 230 Werktagen zwanzigmal täglich durchgeführt werden, ergibt dies eine Energieersparnis eines Jahresverbrauchs von 30 000 Zweipersonenhaushalten – so der Stromspiegel der Regierung (siehe Abbildung 3 und ix.de/zpfc).

Eines sollten Unternehmen und IT-Teams beachten, wenn sie diese Empfehlungen umsetzen möchten: Die Energiebedarfe dieser Komponenten sind miteinander verbunden. Senkt man den Energieverbrauch an einer Stelle, beeinflusst das die anderen drei Bausteine und entscheidet über den Erfolg der Sustainability-Maßnahme. Denn während man auf der einen Seite durch Reduktion der Leistung Energie einspart, steigert man auf der anderen durch Erhöhen der Leistung den Energiehunger. Bevor IT-Teams sich für eine höhere Leistung oder für einen niedrigeren Energiebedarf entscheiden, sollten sie sich deshalb ein klares Bild von den Wechselwirkungen dieser vier Komponenten machen. Auf dieser Basis können sie dann eine faktenbasierte Entscheidung treffen und die Nachhaltigkeit ihrer Anwendungen verbessern.



Quelle: SAP SE

Eine CPU-Sekunde kann den Energiebedarf eines Jahres von 30 000 Zweipersonenhaushalten einsparen (Abb. 3).

Nicht nur durch die nachhaltige Programmierung, auch im Betrieb digitaler Systeme lassen sich die Emissionen reduzieren, etwa durch das gemeinsame Nutzen von Ressourcen: Nichts ist ineffizienter als ein Energieverbraucher, der nicht genutzt wird. Server im Idle-Modus ziehen zwar weniger Energie als unter Volllast, doch auch im unproduktiven Betrieb fällt eine nicht unerhebliche Grundlast an (siehe die Artikel „Licht aus“ und „Auszeit unter Aufsicht“ ab Seite 58 und 72). Um Ressourcen möglichst

effektiv produktiv zu nutzen, lassen sich verschiedene Strategien auch miteinander verweben. Prominentes Beispiel: Konsolidierung durch Cloud-Techniken. Bei vielen anfallenden Arbeitslasten lassen sich Ressourcen gemeinsam nutzen, etwa durch Resource Sharing oder Multi-Tenancy.

Zweites Beispiel: Das dynamische Bereitstellen von Ressourcen durch Elastizität. Teure Überversorgung lässt sich durch Kontrolle und Bedarfsvorhersage vermeiden. Die Bedarfsvorhersage erlaubt eine kontinuierliche Kapazitätsanpassung. Ungenutzte Ressourcen werden abgeschaltet. Nächstes Beispiel: Intelligente Rechenzentrumsverwaltung nutzen. Hierbei lassen sich Arbeitspakete über die Zeit oder über mehrere Rechenzentren verteilen und auf Basis verfügbarer erneuerbarer Energien koordinieren. Darüber hinaus bietet auch die Rechenzentrumshardware noch jede Menge Energiesparmöglichkeiten (siehe Artikel „Ein eigenes Ökotop“ ab Seite 84).

Fazit

Bei der nachhaltigen Programmierung gibt es viele Variablen zu bedenken. Je genauer Unternehmen die Optionen unter die Lupe nehmen, die sich durch nachhaltiges Programmieren ergeben, desto breiter wird ihr Handlungsspielraum. Organisationen müssen also in jedem Einzelfall abwägen, welchen Schwerpunkt sie setzen: einen möglichst geringen Energieverbrauch oder eine leistungsstarke Anwendung.

Ist diese Entscheidung gefallen, dürfen sie nicht vom Weg abkommen, denn es ist einfach, sich während der Umsetzung zu verzetteln. Daher steht am Anfang eine detaillierte Planung, die die Ziele festlegt. Die Reise dahin ist jedoch ein Lernprozess, der Zeit in Anspruch nimmt. Unternehmen wie SAP bieten dazu eine Reihe von Trainings in der eigenen Entwicklung an, die Teams beim Entwickeln energieeffizienter Applikationen unterstützen.

Ebenfalls wichtig ist die Unterstützung des Managements, denn in der gesamten Organisation, über alle Abteilungen hinweg, muss ein Umdenken zu mehr Nachhaltigkeit stattfinden. Denn Nachhaltigkeit ist in ihren verschiedenen Facetten immer häufiger ein Entscheidungskriterium für oder gegen einen Anbieter. Langfristig werden daher nur die Unternehmen erfolgreich sein, die nachhaltig agieren – nachhaltige Programmierung ist ein Baustein dabei. (sun@ix.de)

Quellen

Alle Quellen siehe ix.de/zpfc

Rudolf Meier

ist Head of Performance & Scalability im Product Engineering bei SAP.

Detlef Thoms

ist Chief Product Expert, Performance & Scalability im Product Engineering bei SAP.

Zombies im Rechenzentrum

Licht aus

Martin Lippert

Ungenutzte Systeme abzuschalten funktioniert leider nur in der Theorie – bisher. Mit einigen Anstrengungen ließe sich das aber auch in die Praxis umsetzen.

Rechenzentren auf erneuerbare Energien umzustellen, genügt bei Weitem nicht. Auch die Softwareentwicklung muss die von ihr verursachten Emissionen im Blick behalten und möglichst schnell eine emissionsfreie Softwarelandschaft anstreben. Die einfachste Möglichkeit, Energie, Hardware und damit auch Emissionen einzusparen, besteht sicherlich darin, die Software abzuschalten und gar nicht erst laufen zu lassen oder nur noch laufen zu lassen, wenn man sie tatsächlich benötigt. Das klingt fast schon zu trivial, um diesem Thema einen ganzen Artikel zu widmen.

Allerdings: Jon Taylor und Jonathan Koomey haben das Thema bereits 2015 umfassend erforscht und ihre Studie „Zombie/Comatose Servers Redux“ 2017 aktualisiert (siehe ix.de/zzny). Dazu haben die Autoren die Auslastung reservierter virtueller und physischer Maschinen in Rechenzentren untersucht. Die Studie gibt einen guten Eindruck von der Situation in Rechenzentren und zeigt, wie weit sie von der Wunschvorstellung entfernt ist, dass Software nur dann läuft, wenn sie tatsächlich verwendet wird.

Für die Studie beobachteten die Autoren insgesamt etwa 16 000 Server in zwölf Rechenzentren über zwei Zeiträume von je sechs Monaten. Eine Gruppe, die „Kohorte 2014“, läuft in fünf Rechenzentren und wurde sechs Monate lang im Jahr 2014 und sechs Monate lang im Jahr 2015 beobachtet. Die „Kohorte 2015“ läuft in sieben anderen Rechenzentren und wurde sechs Monate lang im Jahr 2015 beobachtet.

Die Autoren prüften alle fünf Minuten den Status der physischen Server, Hypervisoren und virtuellen Maschinen, fassten die Messdaten in 5760 Fünfzehn-Minuten-Intervallen zusammen und werteten die Netzwerk-, CPU-, Memory und Benutzeraktivitäten aus. Registriert eine Messung eine oder mehrere Aktivitäten dieser Art, gilt der Server zu diesem Zeitpunkt als aktiv. Über die Häufigkeit der gemessenen Aktivitäten teilten die Autoren die Server in drei Gruppen ein:

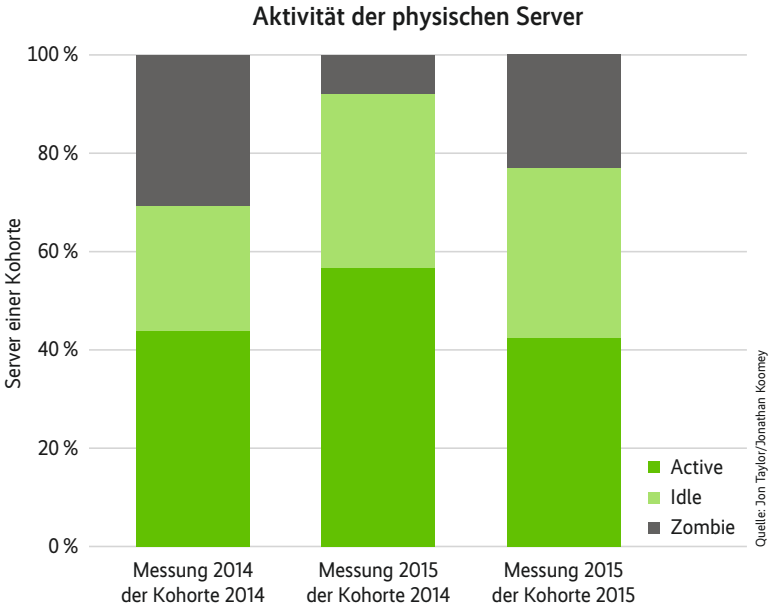
- **Komatöse Maschinen oder Zombies** sind die Maschinen, die über den gesamten Zeitraum keinerlei Anzeichen von Aktivität zeigten.
- Als **Idle** gelten alle Maschinen, bei denen nur weniger als 5 Prozent der Messungen Aktivität registrieren konnten.
- In die Gruppe **Active** kamen alle Maschinen, die bei mehr als 5 Prozent der Messungen aktiv waren.

Die Ergebnisse überraschen

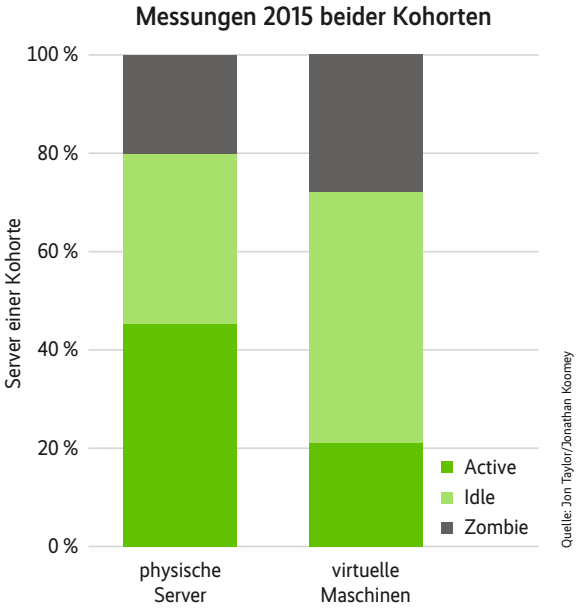
Die Zahl der in der Studie gefundenen Zombies ist überraschend hoch: Je nach untersuchter Kohorte gehören zwischen 23 und 30 Prozent der Maschinen in diese Gruppe – zeigten also über sechs Monate keinerlei Aktivität (siehe Abbildung 1). Es fällt auf, dass sich die Zahlen bei den virtuellen Maschinen nicht signifikant von denen physischer Maschinen unterscheiden (siehe Abbildung 2). Zwar leisten VMs einen großen Beitrag dazu, die Hardware besser auszulasten und ihre Menge

iX-TRACT

- Eine erschreckend hohe Zahl virtueller und physischer Systeme im Rechenzentrum wird gar nicht mehr oder nur selten benutzt.
- Vor allem die Softwareentwicklung kümmert sich meist wenig darum, dass nicht mehr genutzte Test- oder Altsysteme abgeschaltet werden.
- Hier ist eine regelmäßige – möglichst automatisierte – Inventur nötig. Auch müssen die Systeme regelmäßig auf eine Überprovisionierung hin untersucht werden.
- Selten genutzte Software sollte on demand starten. Die dazu nötigen Native Images mit Startzeiten von wenigen Millisekunden lassen sich selbst für Java-Anwendungen erstellen.



Bei den Erstmessungen der Kohorten 2014 und 2015 ist die Zahl der un- und selten benutzen Server erschreckend hoch (Abb. 1).



Noch schlimmer ist die Situation bei virtuellen Maschinen (Abb. 2).

damit zu reduzieren, trotzdem zeigen die Zahlen der Studie, dass auch unter den virtuellen Maschinen ein Zombie-Problem existiert.

Die Zahl der als Idle eingestuften Maschinen schwankt in den Kohorten zwischen 25 und 50 Prozent. In beiden Kohorten kommen die Maschinen der Gruppe „Active“ nicht über 45 Prozent hinaus. Das bedeutet, dass mehr als die Hälfte aller Server in den untersuchten Rechenzentren entweder komplett komatös oder zu mehr als 95 Prozent der Zeit inaktiv war.

Die Ergebnisse zeigen auch, welch enorme Menge Energie und Hardware sich einsparen ließe. Allein durch das konsequente Abschalten der Zombie-Maschinen ließe sich die Hardware und die von ihr verbrauchte Energie um 23 bis 30 Prozent reduzieren. Allerdings lässt sich der Strombedarf der Zombies nicht einfach aus ihrer Zahl ableiten. Das würde nur dann funktionieren, wenn die inaktiven Maschinen denselben Strombedarf hätten wie die unter Last laufenden Systeme, was aber nicht der Fall ist.

Leider sinkt der Energiebedarf inaktiver System aber nicht auf null. Eine Maschine verbraucht im Leerlauf immer noch 50 Prozent der Energie eines Servers unter Volllast – grob geschätzt. Entscheidend ist, dass man sich nicht der Illusion hingibt, dass nicht verwendete Software keine Energie verbrauchen würde.

Die möglichen Einsparungen bei den Idle-Maschinen lassen sich ähnlich berechnen, allerdings nicht durch einfaches Abschalten der Hardware, da die Software, wenn auch selten, so doch ab und zu verwendet wird. Hier sind gegebenenfalls Eingriffe an der Software selbst notwendig, verbunden mit dem Ziel, die Software nur dann zu betreiben, wenn sie auch verwendet wird. Je kongruenter diese beiden Zeiträume sind, desto effizienter der Betrieb.

Unterschiedliche Maßnahmen möglich, aber auch nachhaltig?

Die erste Maßnahme liegt auf der Hand: Inventur machen. Sie sollte die Frage beantworten, welche Software oder welche Serverkapazität nicht mehr benötigt wird und abgeschaltet werden kann. Damit ließen sich – legt man die Zahlen der Studie zugrunde – wahrscheinlich schon zwischen 23 und 30 Prozent der benutzten Maschinen und die entsprechende Menge an Energie einsparen. Eine einmalige Inventur kann zwar große Einsparungen bringen, allerdings treten ähnliche Situationen im

Laufe der Zeit immer wieder auf. Testsysteme, alte Software und Ähnliches werden weiter betrieben, aber nicht mehr verwendet, vergessen und so über die Zeit zum Zombie. Entgegenwirken kann man dem nur, wenn man die Systeme im Rechenzentrum kontinuierlich auf solche Zombies überprüft – und das am besten automatisiert.

Die nächste Frage ist, welche der verwendeten Software sich mit weniger Ressourcen betreiben lässt. Könnte die Software auch auf einer kleineren Instanz laufen? Könnte die Software mit weniger CPU oder weniger Speicher die gleiche Leistung erbringen? Wie viele Backup-Instanzen sind wirklich nötig? Muss die Software 24 Stunden durchlaufen oder lässt sie sich nachts abschalten oder zumindest herunterskalieren, wenn sie weniger oder gar nicht verwendet wird?

Nicht jeder dieser Aspekte lässt sich unbedingt automatisieren. Sie sollten aber nach Möglichkeit schon bei der Softwareentwicklung berücksichtigt werden. Die Zeiten, in denen man „zur Sicherheit" eine größere Maschine oder mehr Speicher reserviert hat, um möglichen Engpässen vorzubeugen, sind vorbei.

Wer Software entwickelt, trägt die Verantwortung für ihren Ressourcenverbrauch – den gilt es zu minimieren. Neben den Performancekriterien werden der Ressourcenverbrauch und der ökologische Fußabdruck in Zukunft ein wichtiger Faktor unter den nicht funktionalen Anforderungen – und wahrscheinlich sogar ein wichtiges Unterscheidungsmerkmal von Software werden.

Das angestrebte Ziel: Scale to Zero

Die automatische Skalierung von Software, je nachdem, wie stark sie verwendet wird, ist in heutigen Rechenzentren eigentlich Standard. Autoskalierung heißt das Zauberwort, mit dem sich Software automatisch hoch- und wieder herunterskalieren lässt. Cloud-Plattformen bieten diesen Mechanismus als Bestandteil an, sodass sich die Software relativ unabhängig davon entwickeln lässt.

Möchte man allerdings den in der Studie als „Idle" gruppierten Systemen an den Kragen, muss man ein „Scale to Zero" anstreben. Dabei wird die Software komplett heruntergefahren, wenn sie nicht verwendet wird – also auf „null" Instanzen herunterskaliert – und erst wieder gestartet, wenn sie benötigt wird.

Das lässt sich allerdings nicht mit jeder Software bewerkstelligen. Schaut man sich die Startzeiten heutiger Enterprise-Systeme an, findet man nicht selten Zeiten von vielen Sekunden, mitunter mehreren Minuten, bis eine Software nach dem Start in der Lage ist, eine Anfrage zu bearbeiten. Aber wie kann die Software erst bei einer eintreffenden Anfrage starten und gleichzeitig eine schnelle Antwortzeit erreichen, wenn sie erst nach mehreren Sekunden oder Minuten einsatzfähig ist?

In der Regel werden diese Systeme deshalb nicht auf null Instanzen heruntergefahren, sondern halten stets eine antwortbereite Instanz am Laufen. Sie kann schnell auf Anfragen reagieren und gegebenenfalls im Hintergrund weitere Instanzen hochfahren. Bei den Idle-Systemen verfehlt diese Strategie aber das Ziel. Die Software wird beispielsweise nur zu 5 Prozent der Zeit verwendet und soll für die restliche Zeit komplett heruntergefahren werden. Allerdings lässt sich unter Umständen nicht vorhersagen, wann genau diese 5 Prozent Aktivität auftreten werden – trotzdem sind die kurzen Antwortzeiten auf jeden Fall sicherzustellen.

Kurze Startzeiten implementieren

Ein möglicher Ausweg aus diesem Dilemma besteht darin, die Startzeit für die Software auf ein Maß zu bringen, das es erlaubt, die Software bei einer eintreffenden Anfrage on demand zu starten. Dazu müsste man die Startzeit aber drastisch reduzieren, und zwar in den Bereich weniger Millisekunden.

Normalerweise sind solche Startzeiten nur in Spezialanwendungen umsetzbar, die in entsprechenden Sprachen implementiert sind und direkt in Native Executables kompiliert werden. Die meisten Enterprise-Anwendungen sind dagegen mit Java und Spring Boot gebaut, laufen auf einer Java Virtual Machine und brauchen deutlich mehr als ein paar Millisekunden zum Hochfahren.

Dass aber Startzeiten von wenigen Millisekunden heute auch für Enterprise-Software erreichbar sind, zeigen neue Entwicklungen im Bereich Java. Die GraalVM-Technik zum Kompilieren von Native Images ist ein Beispiel dafür. Mit dieser Technik lassen sich Java-Anwendungen in plattformspezifische Native Executables kompilieren. Dadurch muss keine komplette JVM mehr zur Laufzeit gestartet und der Code nicht mehr durch einen JIT-Compiler on the fly optimiert werden, um optimale Performance zu erreichen. Die Anwendung liegt schon fertig kompiliert vor und startet in wenigen Millisekunden.

Diese Technik hält bereits Einzug in etablierte Java-Enterprise-Frameworks wie Spring und Spring Boot. Wer also bisher seine Anwendungen mit Spring Boot entwickelt, kann GraalVM nutzen, um die Startzeit seiner Anwendung zu reduzieren. Die Spring-Erweiterung Spring Native erlaubt es zudem, bestehende Spring-Boot-Anwendungen ohne viele Anpassungen GraalVM-Native-Image-tauglich zu machen – eine vielversprechende Entwicklung.

Fazit

Software und Softwareentwicklung müssen in den nächsten Jahren komplett emissionsfrei werden – daran führt kein Weg vorbei. Der Betrieb der Software ist dabei eine wesentliche Größe. Laut der Studie von Taylor und Koomey sind ungefähr ein Drittel der reservierten Maschinen in Rechenzentren Zombies, ein weiteres Drittel wird in weniger als 5 Prozent der Zeit tatsächlich verwendet.

Die Verschwendung von Ressourcen ist in Anbetracht dieser gewaltigen Zahlen enorm – und ihre Reduktion eine wichtige Aufgabe für die Zukunft. Dazu müssen nicht nur Rechenzentren effizienter und grüner werden – die Verschwendung selbst ist zu eliminieren. Dazu muss die Software selbst grüner werden. Denn wer Software entwickelt, trägt auch die Verantwortung für ihren Ressourcenverbrauch – den gilt es zu reduzieren. Es lohnt sich. (sun@ix.de)

Quellen

Die Studie ist unter ix.de/zzny zu finden.

Martin Lippert

arbeitet bei VMware und leitet dort die Projekte rund um die Spring Tools. Darüber hinaus engagiert er sich als Sustainability Ambassador bei VMware.

Es gibt 10 Arten von Menschen.

iX-Leser und die anderen.

3 × als Heft

Jetzt Mini-Abo testen:
3 Hefte + Bluetooth-Tastatur
nur 19,35 €

www.iX.de/testen

www.iX.de/testen | 49 (0)541 800 09 120 | leserservice@heise.de

iX MAGAZIN FÜR PROFESSIONELLE INFORMATIONSTECHNIK

Energieeffizienz von Software messen

Durchdrungen

Achim Guldner, Dr. Eva Kern, Dr. Sandro Kreten, Prof. Dr. Stefan Naumann

Die Energieeffizienz von Software lässt sich durch Änderungen einzelner Komponenten erhöhen. Welche das sind, können differenzierte Messungen offenlegen.

Die Frage, welchen Anteil Software am Ressourcen- und Energieverbrauch hat, führt die Fachwelt noch nicht allzu lange. Obwohl Software als immaterielles Gut zunächst nur indirekt wirkt, hat sie erheblichen Einfluss auf den Energie- und Rohstoffverbrauch: Durch ihren Umgang mit den Ressourcen des Rechners und die Update-Politik ihrer Hersteller hat sie nicht nur Einfluss auf seinen Energieverbrauch, sondern auch auf den Zeitpunkt seiner Neubeschaffung.

Basierend auf diesen Beobachtungen sind in den letzten Jahren eine Reihe Projekte gestartet worden, die die Nachhaltigkeit von Software analysieren, messen und evaluieren. Zudem wurden Kriterien entwickelt, die diese Nachhaltigkeitsaspekte von Software operationalisieren und zertifizieren. Seit 2020 existiert das Label „Blauer Engel für Softwareprodukte“, das in der ersten Version Desktop-Software auszeichnet (siehe ix.de/zksj). Die erste mit dem Blauen Engel ausgezeichnete Software ist der KDE-Dokumentenbetrachter Okular.

iX-TRACT

- Obwohl Software immateriell ist, lässt sich ihr Energieverbrauch messen.
- Die Messergebnisse zeigen, dass schon der Austausch von Teilen des Softwarestacks, von Bibliotheken oder Algorithmen die Energieeffizienz massiv erhöhen kann.
- Beispielsweise benötigen die in C implementierten NumPy-Funktionen nur etwa 2 Prozent der Energie, die die Python-Standardfunktionen erfordern.
- Nachhaltigkeitskriterien wie die des Blauen Engels für Software geben Handlungsempfehlungen auch für die Anwendungsentwicklung.

Zu betrachten ist der gesamte Lebenszyklus, also Herstellung, Nutzung und Entsorgung von Soft- und Hardware – Endgeräten, Netz und Cloud. Referenzmodelle wie das im Forschungsprojekt Green Software Engineering entstandene GREENSOFT-Modell bilden einen Rahmen für Forschung und Praxis und helfen bei der Entwicklung und Bewertung (siehe ix.de/zksj). Dieser Artikel wird zunächst anhand einiger Messergebnisse den Energiebedarf von Software darlegen und anschließend ausgewählte Kriterien des Blauen Engels für Software vorstellen.

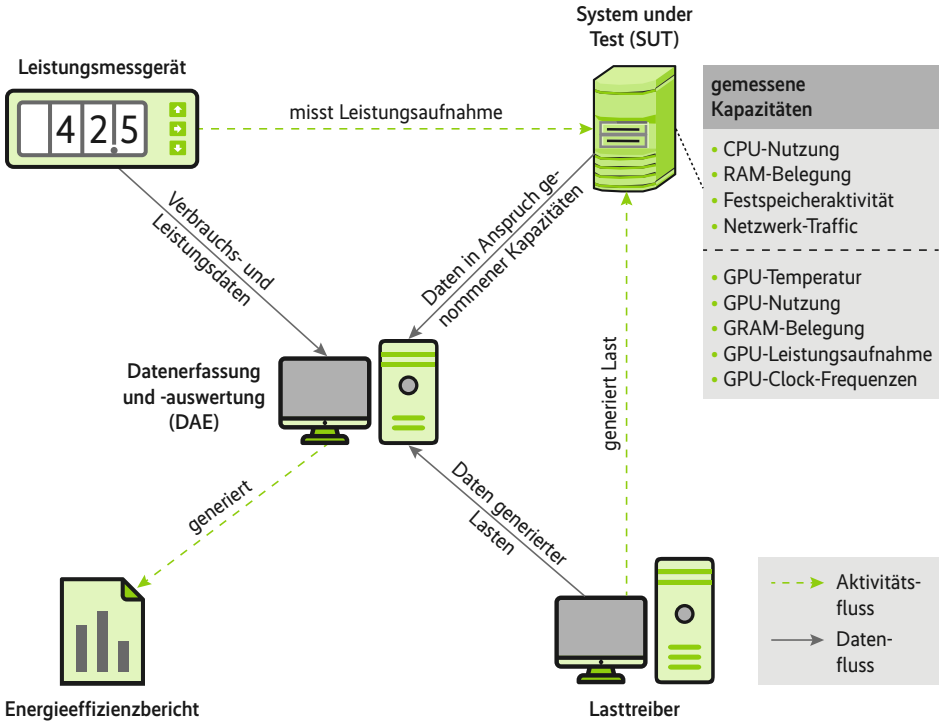
Den Nettoverbrauch von Software messen

Um den Energie- und Ressourcenverbrauch einer einzelnen Anwendung zu messen, muss man ihren Anteil an der Auslastung und der Leistungsaufnahme der verwendeten Hardware bestimmen. Dieser Nettoverbrauch berücksichtigt nicht die Tatsache, dass die Anwendung selbst auf Hardware, Betriebssysteme, womöglich auch Netzwerkkomponenten und externe Dienste angewiesen ist und deren Energieverbrauch mitverschuldet. Das Messverfahren wurde am Umwelt-Campus Birkenfeld der Hochschule Trier entwickelt, unter anderem auf Grundlage der ISO/IEC 14756:1999 „Information technology – Measurement and rating of performance of computer-based software systems“.

Zentraler Bestandteil des Messaufbaus ist ein System under Test, auf dem die zu messende Software ausgeführt und die Hardwarenutzung protokolliert wird (siehe Abbildung 1). Den Energieverbrauch zeichnet ein Leistungsmessgerät auf. Die Software durchläuft mehrere Szenarien, die ein Lasttreiber steuert. Das kann auf der Kommandozeile ein Skript sein, ein über ein Automatisierungstool gesteuertes UI-Szenario oder ein Client, der wiederholt eine API oder einen Dienst aufruft. Je nach Anwendung werden Leerlauf-, Standardnutzungs- und Lastszenarien entwickelt und gemessen, ähnlich dem SPECpower-Benchmark,

der die CPU-Last in 10-Prozent-Schritten erhöht.

Zusätzlich findet immer eine Baseline-Messung des Systems statt, ohne ausgeführte Software. Der durch die Software hervorgerufene Verbrauch der einzelnen Szenarien ergibt sich aus dem Delta zu den Baseline-Ergebnissen. Damit lässt sich etwa der Nettoverbrauch bestimmen, den das Ausführen von Okular verursacht: 0,07 Wh in 3,6 Minuten (siehe Abbildung 2 und ix.de/zksj). Das umfasst nur den Energiebedarf der Software ohne anteilige Anrechnung des Betriebssystems oder anderer Systemsoftware und ohne Berücksichtigung der Hardwareeffizienz. Durch das Wiederholen der Messungen lassen sich äußere Einflüsse, etwa durch Betriebssystemaktivitäten oder das zufällige Hochdrehen der Lüfter, minimieren, es entsteht ein Lastprofil für jedes Szenario. Analog ist auch die Nutzung anderer Ressourcen durch die Anwendung bestimmbar, etwa von Arbeitsspeicher, CPU oder Netz. Abbildung 3 zeigt die RAM-Belegung durch Okular.



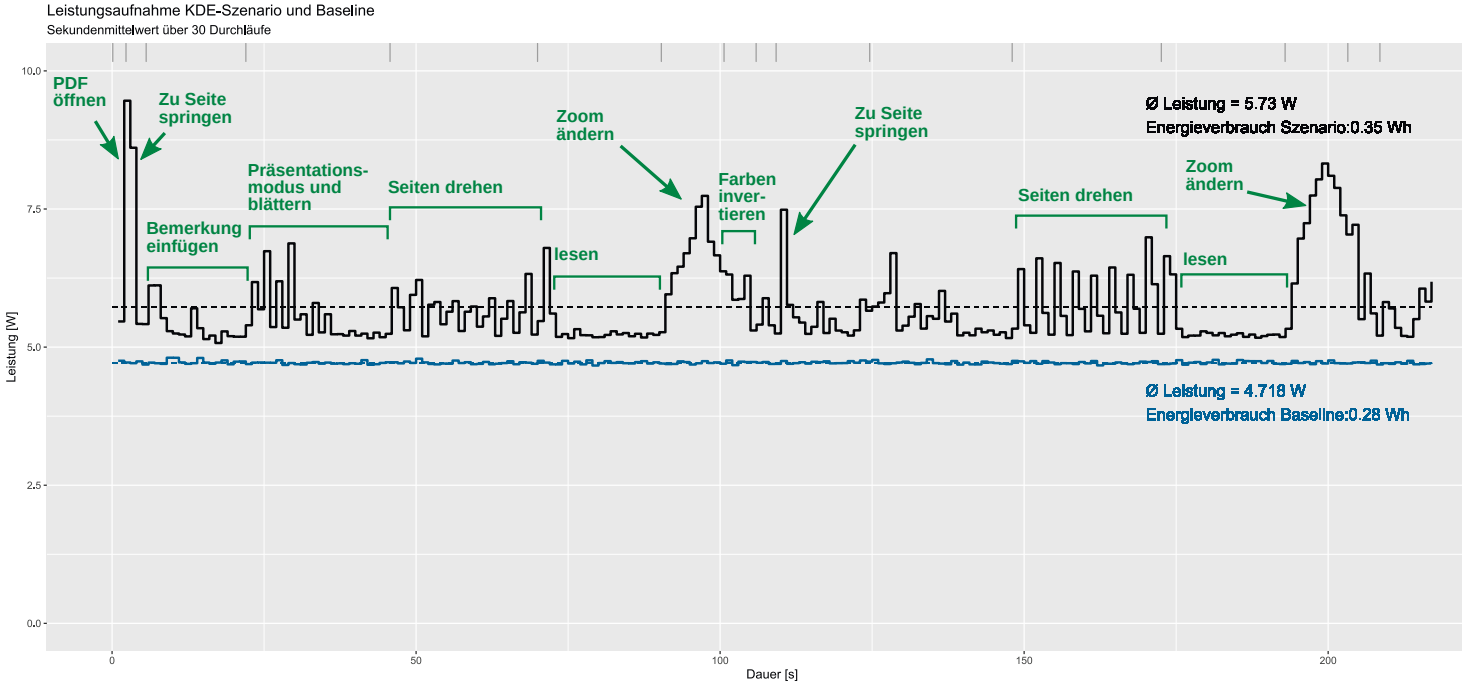
Im Messaufbau flankieren Lasttreiber und Leistungsmessgerät das Testsystem (Abb. 1).

Die Analyse der Details

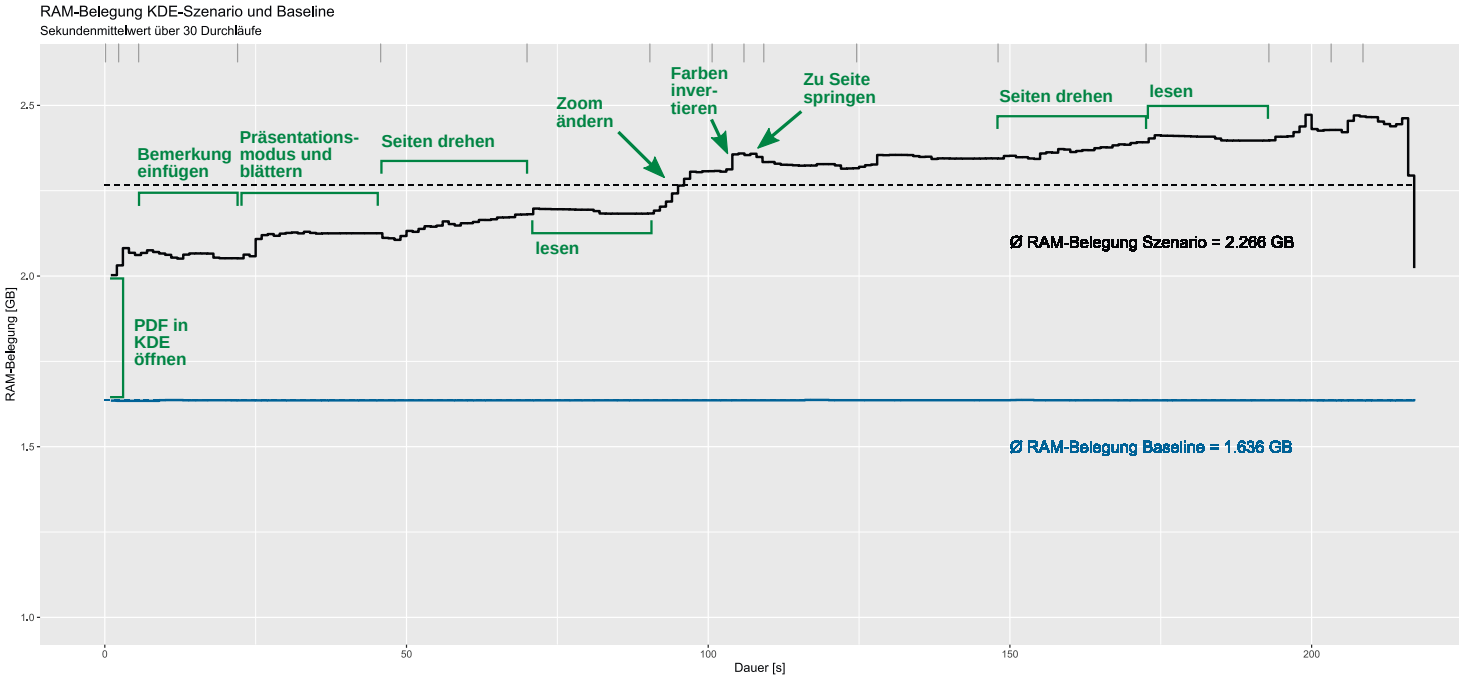
Dem Messen der Software können weitere Analysen folgen, die etwa der Frage nachgehen, wie sich die Effizienz steigern ließe. Infrage kämen dazu der Austausch von Teilen des Softwarestacks oder einzelner Bibliotheken, Algorithmen oder das Analysieren der Versionshistorie. Das Logging der Zeitstempel zu Beginn und Ende einer Softwareaktivität erlaubt zudem tiefer gehende Einblicke in die Lastprofile: Beispielsweise untersuchte das Team den Energieverbrauch auf Funktionsebene für grep oder die Entwicklung des Energieverbrauchs über die Trainingsepochen eines Convolutional Neural Network (siehe ix.de/zksj).

Wer ressourceneffiziente Software entwickeln möchte, kommt um eine Analyse ihrer Bestandteile kaum herum. Als Erstes ist die eingesetzte Programmiersprache zu betrachten. Ihre Wahl richtet sich nach Anwendungsgebiet, Zweck und Arbeitsumfang. Bereits hier können kleine Anpassungen große Auswirkungen auf die Ressourceneffizienz haben, etwa bei der Wahl der Bibliotheken.

Beispielhaft sollen in Python zuerst ein Array sortiert und dann Zufallszahlen generiert werden. Die erste Aufgabe bewältigen zum einen die Standard-Python-Funktion `range()` in einer `for`-Schleife und zum anderen die Funktionen `np.sum` und `np.arange` aus der Python-Bibliothek NumPy. NumPy ist auf die



Das Profil des Energieverbrauchs spiegelt die Aktivität des Okular-Benutzers wider (Abb. 2).



Anders als der Energieverbrauch steigt die RAM-Belegung während der Benutzung von Okular immer weiter an (Abb. 3).

einfache Handhabung großer mehrdimensionaler Arrays spezialisiert und stellt neben den Datenstrukturen auch effizient implementierte Funktionen für numerische Berechnungen bereit. Bei der zweiten Aufgabe treten die Python-Standardfunktion `random.randint()` und die NumPy-Funktion `np.random.randint()` gegeneinander an.

Die Messungen offenbaren große Unterschiede

Listing 1 zeigt alle vier Codeschnipsel, in denen die Funktionen je 100 Millionen Werte bearbeiten. In dem oben beschriebenen Messaufbau wurden die Codeschnipsel ausgeführt und die Leistungsdaten aufgenommen. Mit den NumPy-Funktionen beträgt der Energieverbrauch beim Addieren eines Arrays nur 2,3 Prozent und beim Generieren der Zufallszahlen sogar nur 1,18 Prozent dessen, was bei den Standardversionen anfällt (siehe Abbildung 4). Diese enorme Zeit- und Energieeinsparung ist der Tatsache geschuldet, dass NumPy in C implementiert ist und damit wesentlich besser mit den Hardwareressourcen haushalten kann als der native Python-Code.

Das zeigt, dass nicht nur spezifische Codeabschnitte, sondern auch die Wahl geeigneter Frameworks Vor- und Nachteile bieten kann. In Sachen Ressourceneffizienz bilden die Frameworks aber oft Blackboxes. Häufig steht zudem das Kriterium der Usability der Wahl des ressourcenschonendsten Frameworks entgegen.

Für einige Sprachen stehen Profiling-Werkzeuge zur Verfügung, mit denen sich ineffiziente Softwarebestandteile ausfindig machen lassen. Beispielsweise kann man mit dem Go Profiler einzelne Threads respektive Channels analysieren und damit einen im Prozessstack besser ausbalancierten Code gestalten.

Es existieren auch erste Plug-ins wie CodeCarbon, die die Energiekosten und den damit verbundenen CO_2-Ausstoß visualisieren. Das Projekt SoftAWere versucht unter anderem, den Verbrauch gängiger Bibliotheken und Tools etwa als Git-Badge sichtbar zu machen. Auch beim Machine Learning gibt es Unterschiede zwischen den Frameworks (siehe ix.de/zksj).

Zur Effizienz eines Produkts zählt aber auch der gesamte Softwarestack. Neben der selbst programmierten Software sind weitere Bestandteile wie Datenbanken oder Webserver in die Analyse des Systems einzubeziehen. Ob es bei der Betrachtung der eigenen Software Grenzen gibt und wo sie liegen, ist dabei diskutabel. Beispielsweise kann die Frage „Spielt es eine Rolle, ob meine Software auf eine ineffiziente API zugreift?“ weitere Fragen aufwerfen.

Kriterien für nachhaltige Software

Ein Grund für die vielen noch unbeantworteten Fragen liegt darin, dass ein einheitliches Verständnis von nachhaltiger Software bisher fehlt. Zwar beschäftigt sich die Forschung seit rund 15 Jahren mit den Charakteristika nachhaltiger Software, doch je nach Perspektive und Fokus priorisiert sie unterschiedliche Aspekte wie die Energieeffizienz, die Entwicklungsprozesse oder die Auswirkungen der Softwarenutzung auf Gesellschaft und Wirtschaft. Deshalb ist der Blaue Engel für Desktop-Software ein wichtiger Schritt zu einem einheitlichen Verständnis von nachhaltiger Software.

Die Entwicklung der Vergabekriterien folgte einem kombinierten Ansatz, in den wissenschaftliche Erkenntnisse, bestehende Softwarequalitätskriterien und softwarespezifische Kriterien aus allgemeinen Nachhaltigkeits- respektive Umwelt-

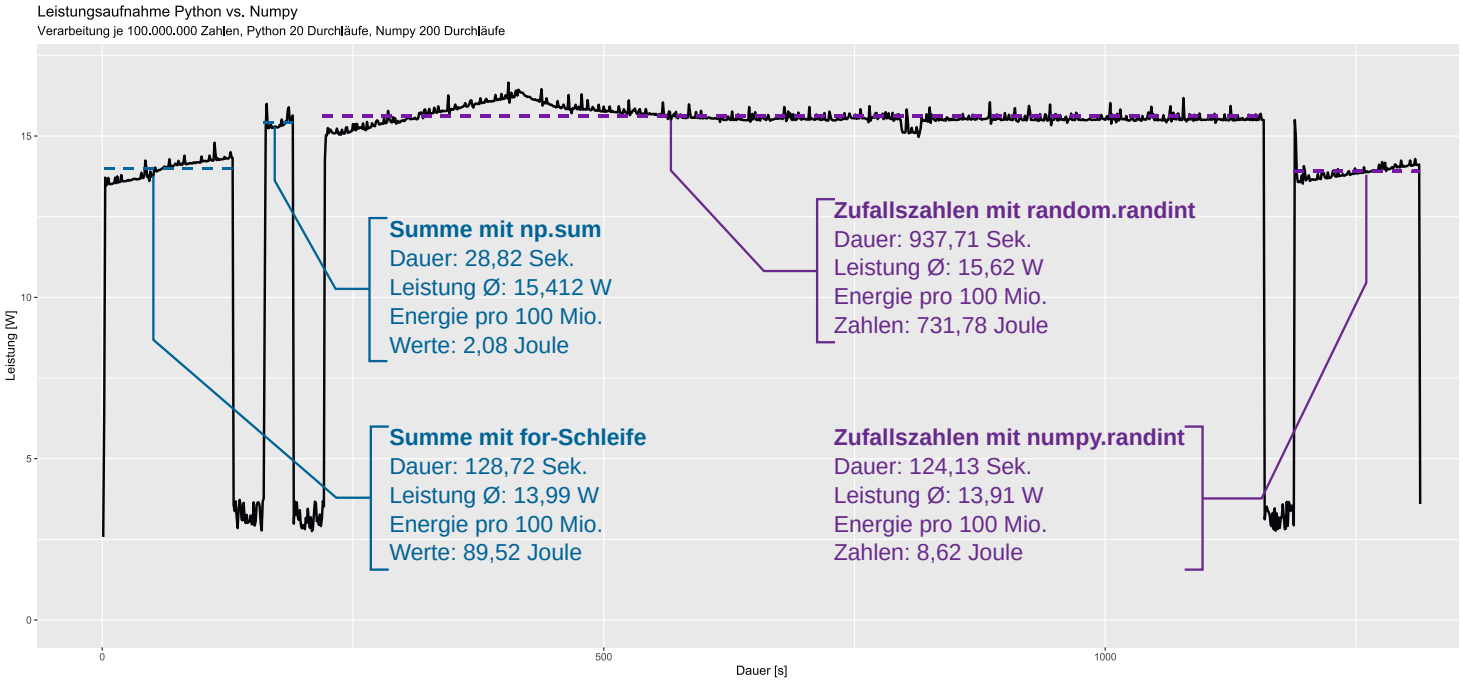
Listing 1: Die vier Codeschnipsel für die Leistungsmessung

```
# Standard-for-Schleife:
standard_sum = 0
for i in range(100000000):
 standard_sum = standard_sum + i

# NumPy-Funktionen sum und arange
import numpy as np
np_sum = np.sum(np.arange(100000000))

# Standard-Funktion randint
arr = [random.randint(0, 100) for p in range(1, 100000000)]

# NumPy-Funktion randint
np_arr = np.random.randint(low = 0, high = 100, size =
(100000000,))
```



Die Leistungsmessung verrät, wie effizient die Standard-Python- und NumPy-Funktionen arbeiten (Abb. 4).

indikatoren einflossen. Der Blaue Software-Engel beschränkt sich derzeit auf Desktop-Software. Diese enge Eingrenzung dient dazu, zuerst geeignete Evaluationsmethoden zu entwickeln und so einen Startpunkt in der Nachhaltigkeitsbewertung der immateriellen Produkte zu setzen. Der vollständige Katalog umfasst über 70 Indikatoren, die sich teilweise auch auf verteilte Systeme, mobile Apps und Embedded- beziehungsweise IoT-Geräte anwenden lassen und die in Zukunft weiterentwickelt werden.

Zum Kriterienkatalog gehören die Ressourcen- und Energieeffizienz, die potenzielle Hardwarenutzungsdauer und die Nutzungsautonomie. Dazu zählt die Verwendung einzelner Ressourcen wie CPU-Zeit, RAM-Belegung, Datenübertragung, Festplattenbelegung und je nach Anwendungsfall auch die Grafikkartennutzung, da sie direkte Hinweise auf die Energie- und Ressourceneffizienz liefern. Darüber hinaus müssen die Softwarehersteller für die Zertifizierung verschiedene Nachweise erbringen, darunter die Unterstützung eines Energiemanagements, die potenzielle Hardwarenutzungsdauer, Abwärtskompatibilität, Modularität, Offlinefähigkeit und Werbefreiheit.

Der Blaue Engel ist ein Verbraucherlabel und richtet sich in erster Linie an Softwarenutzende. Dennoch liefern seine Kriterien auch hilfreiche Orientierungspunkte, die Entwicklungsteams bereits während der Softwaregestaltung berücksichtigen können. Darüber hinaus liefern sie auch Anhaltspunkte für die Beschaffung von Softwareprodukten und die Beauftragung einer Neuentwicklung. Zudem gibt es weitere Gründe für die Kennzeichnung grüner Softwareprodukte: Erstens werden Informationen über nachhaltigkeitsrelevante Aspekte von Software verbreitet. Zweitens erhöht sie die Transparenz und trägt ein Thema aus der Wissenschaft in die Öffentlichkeit. Drittens ist der Ansatz für andere Produktarten erfolgreich erprobt.

Fazit

Software hat große Auswirkungen auf die Energieeffizienz der IT. Auch wenn unstrittig ist, dass sich Effizienz und Performance bedingen und gute Algorithmen sowie performante Software eine wichtige Grundlage bilden, die Vielzahl der Frame-

works, Programmiersprachen, Entwicklungsumgebungen und ihre Einbettung in verteilte Softwareumgebungen vereinfacht die Frage nach dem besten Weg dorthin nicht. Gerade in verteilten Umgebungen ist entscheidend, was wo gerechnet wird und welche Datenmengen wann übertragen werden.

Kriterien wie die des Blauen Engels für Software übernehmen damit auch die Aufgabe von Handlungsempfehlungen. Auch sollte das Kriterium des Energieverbrauchs häufiger Einzug halten in die Entwicklung und Auswahl von Anwendungen. Beim Autokauf lautet spätestens die zweite Frage: „Und was verbraucht der?“ Diese Frage sollte auch bei Softwareprodukten zur Regel werden. (sun@ix.de)

Quellen

Onlineliteratur und Projekte siehe ix.de/zksj

Achim Guldner

forscht seit 2013 im Bereich der Ressourcen- und Energieeffizienz von und durch IKT. Sein aktueller Fokus liegt auf der KI und cyberphysischen Systemen.

Dr. Eva Kern

forscht zur Bewertung und Kommunikation von Umweltaspekten von Software und bringt Forschungsergebnisse als Mitarbeiterin im Lüneburger Jugendumweltnetzwerk in die Bildungspraxis.

Dr. Sandro Kreten

ist CTO des Impact-Investors capacura und promovierte über die Energieeffizienz von Cloud-Systemen mit Schwerpunkt auf Containern.

Prof. Dr. Stefan Naumann

ist seit 2008 auf dem Umwelt-Campus Birkenfeld der Hochschule Trier im Bereich der Nachhaltigkeitsinformatik tätig. Sein Forschungsgebiet ist die Energieeffizienz von Software.

Der Go Profiler analysiert die CPU-Zeit einzelner Programmschritte (Abb. 5).

Ressourcensparend programmieren: Lernen von der Embedded-Entwicklung

Angepasst

Andreas Fertig

Mit geeigneten Konzepten lässt sich robusterer Code entwickeln, der auch noch die Rechnerressourcen schont.

Software erobert immer mehr Einsatzgebiete. Viele Geräte funktionieren heute nicht mehr ohne: Uhren, Waschmaschinen, Herde und allen voran Autos. Gemeinsam sind ihnen die Beschränkungen der Hardware. Deshalb ist es zwingend erforderlich, mit den begrenzten Ressourcen – vor allem CPUs, RAM und ROM – effizient umzugehen.

Anders bei Computern mit ihren zahlreichen CPU-Kernen, GByte an RAM und ROM sowie TByte an Festplatten: Ihre Software wird unabhängig von allen anderen laufenden Anwendungen programmiert. Thunderbird weiß nicht, ob Chrome gerade läuft und vielleicht noch Visual Studio. Hier schöpft jede Anwendung aus dem Vollen, der Gedanke an die Ressourcennutzung kommt oft zu kurz.

Dabei ist auch für Anwender ein ressourcenschonendes Programmieren auf Desktop-Rechnern wünschenswert: Die Anwendung startet schneller, läuft flüssiger und riskiert weniger Abstürze durch Out-of-Memory-Fehler. Treffend beschreibt Scott Meyers den Unterschied: „Gute Softwareentwicklung für Embedded Systems ist einfach gute Softwareentwicklung“ [1]. Wünschenswert wäre es also, davon zu lernen. Doch was lässt sich aus der Welt der eingebetteten Systeme in die der Desktops übertragen?

> „Gute Softwareentwicklung für Embedded Systems ist einfach gute Softwareentwicklung.“ *Scott Meyers*

Klein ist nicht immer sparsam

Ein Mythos, der sich noch immer hält, erzählt, man solle bei Schleifen einen möglichst kleinen Datentyp verwenden. Beispielhaft soll eine `for`-Schleife von `0` bis `5` zählen. In Listing 1 nutzt sie dazu eine Zählvariable vom Typ `int`. Gemäß diesem Narrativ braucht es keinen Typ `int`, der Wertebereich eines `char` reicht völlig aus (siehe Listing 2).

Ist die Variante in Listing 2 also ressourcenschonender? Die Antwort lautet Nein. Ein Character – `char` – ist ein einzelnes Zeichen und belegt deshalb nur ein Byte, während ein Integer – `int` – 4 Byte groß ist. Die Variable soll aber mit einer Zahl verglichen werden, deshalb muss der Compiler sie in den Typ `int` wandeln, um den Vergleich durchführen zu können. Das verursacht neben zusätzlichem Platzbedarf einen höheren Rechenaufwand für den Compiler. Der Compiler Explorer unter godbolt.org zeigt den daraus resultierenden Assemblercode, der nun eine Zeile mehr enthält. Auch C++ Insights macht diese implizite Konvertierung sichtbar (siehe Abbildung 1 und 2 und ix.de/zdrt).

Daten richtig anordnen

Daten optimal im Speicher zu halten, gehört zu den wichtigsten Aufgaben, wenn es um das Schonen der Ressourcen geht. Das folgende Beispiel setzt eine 32-Bit-Architektur voraus. Wenn Variablen unterschiedlichen Typs deklariert werden, fügt der Compiler in C und C++ zum Ausrichten der Datentypen Padding-Bytes ein. Im Beispiel von Listing 3 wäre das nach dem De-

iX-TRACT

- Mit einer Reihe von Strategien lässt sich nicht nur robusterer, sondern auch ressourcenschonender Code erstellen.
- Bei der Wahl der Variablentypen ist ihre Eignung wichtiger als ihr Platzbedarf.
- Sparen kann man auch bei Funktionen, die nicht zurückkehren.
- Was sich bereits zur Compilezeit erledigen lässt, beschleunigt die Anwendung während der Laufzeit.

Listing 1: for-Schleife mit int als Zählvariable

```
for(int i = 0; i < 5; ++i) {
// ...
}
```

Listing 2: for-Schleife mit char als Zählvariable

```
for(char i = 0; i < 5; ++i) {
// ...
}
```

klarieren der Variablen a und c der Fall: Auf einer 32-Bit-Architektur ist eine Variable vom Typ `int` mit eine Größe von 4 Byte auf einer durch vier teilbaren Adresse ausgerichtet. Die Variable vom Typ `char` hat eine feste Größe von 1 Byte und ist an jeder Adresse ausgerichtet. Um im vorliegenden Fall `b` korrekt auszurichten, sind 3 Padding-Bytes erforderlich. Die interne Sicht im Compiler erhält man wieder mit C++ Insights.

Die zweite Variable vom Typ `char` in dieser Struktur bringt weitere Padding-Bytes mit sich. Der Datentyp `char` selbst liegt hier korrekt ausgerichtet im Speicher. Da sich aber ein Array von `Data`-Objekten bilden lässt, muss auch `Data` selbst immer gleich ausgerichtet werden. Dafür werden 3 weitere Padding-Bytes nach `c` eingefügt. Die Größe der Struktur `Data` wächst damit auf 12 Byte, lediglich 6 davon sind aber Nutz-Bytes – ein Verschnitt von 50 Prozent.

Unnötigen Overhead reduzieren

Diesen Ballast schleppt das Programm ständig mit: bei einer Übergabe by value, bei einer Rückgabe by value, wenn der Datentyp `Data` in einem `std::optional` oder einem Container wie `std::vector` gespeichert wird. Auf einem Desktop-Rechner wirken 6 Byte sicherlich lächerlich wenig, auf Embedded-Geräten kann das dagegen eine ganze Menge sein. In beiden Fällen gilt: Je mehr dieser Daten das Programm hält oder transferiert, desto stärker fällt der Overhead ins Gewicht.

Auf Desktop-Rechnern kann man diesen Overhead lange ignorieren, da er mehr Ressourcen zur Verfügung stellt als ein Embedded-System. Laufen aber mehrere Applikationen und verplempert jede von ihnen die Ressourcen so, summiert sich das schnell, und ein neuer, größerer Rechner muss her. Beim Transfer zwischen unterschiedlichen Rechnern kommt noch eine stärkere Belastung des Netzes hinzu. Nutzt der Anwender WLAN, geht die Mehrbelastung zudem mit einer Verringerung der Batterielaufzeit einher.

Reduzieren lässt sich der Overhead durch eine bewusste Anordnung der Datentypen in `Data`. Listing 4 zeigt eine alternative, ressourcenschonende Anordnung. Als Faustregel gilt: Idealerweise sortiert man die Variablen der Größe ihrer Datentypen nach in absteigender Reihenfolge. Damit verschiebt sich ein mögliches Padding ans Ende, da das Padding zwischen den Deklarationen dadurch unwahrscheinlicher wird. Durch das Umsortieren reduziert sich das Padding in diesem Fall um 4 Byte auf 2 Byte. Die Gesamtgröße der Struktur `Data` sinkt von 12 auf 8 Byte. Auch hier visualisiert C++ Insights das interne Padding.

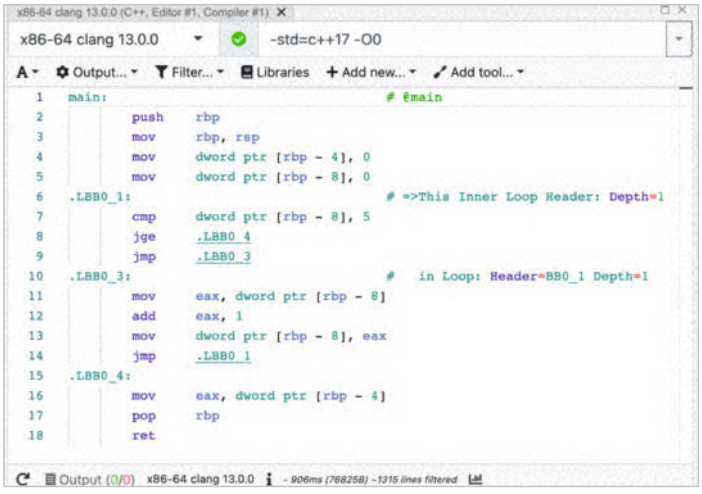
Callstack einsparen für Funktionen ohne Rückkehr

Seit C++11 steht eine Standardsyntax für Attribute zur Verfügung, die den Attributnamen in doppelte eckige Klammern packt. Mit `[[noreturn]]` steht ein Attribut bereit, das bei weiteren Ressourceneinsparungen helfen kann. Es markiert eine Funktion als nicht zurückkehrend. Das bedeutet, die Funktion besteht aus einer Endlosschleife oder beendet das Programm. Eine Kombination, wie Listing 5 sie zeigt, ist ebenfalls denkbar.

Für den Compiler bedeutet eine mit `[[noreturn]]` markierte Funktion, dass er den Callstack nicht erstellt. Das spart Assemblerinstruktionen und beschleunigt den Aufruf der Funktion. Auf godbolt.org lassen sich die Assemblerinstruktionen des in Listing 5 gezeigten Codes mit und ohne `[[noreturn]]` gegenüberstellen (siehe ix.de/zdrt). Schon ohne Optimierung braucht die mit `[[noreturn]]` markierte Variante eine Assemblerinstruktion weniger. Mit `-O3` sind es bereits sieben bei nur einem Aufruf von `Fun`. Da es aber im Code – logischerweise – nur eine solche Endlosschleife geben kann, hält sich die Ersparnis in Grenzen.

Der Horizont ist jedoch größer: weg von der Endlosschleife hin zu Funktionen wie `CustomAbort`. Diese Funktion generiert einen Log-Eintrag, bevor eine weitere Funktion `reset` aufgerufen wird, die den Neustart des Geräts oder des Programms veranlasst (siehe Listing 6). Auch bei der Funktion `CustomAbort` ist klar, dass sie nie zurückkehrt. Wieso also den Compiler Code generieren und zur Laufzeit ausführen lassen, wenn im Vorfeld bereits feststeht, dass dies toter Code ist?

Obwohl Variablen vom Typ char weniger Platz benötigen als Variablen vom Typ int, verursachen sie zusätzliche Transformationsarbeiten (Abb. 1).



Eine Variable vom Typ int belegt zwar mit ihren 4 Byte viermal so viel Platz wie eine Variable vom Typ char, muss aber nicht gewandelt werden (Abb. 2).

Listing 3: Die Struktur Data mit drei unsortierten Elementen in der Ansicht von C++ Insights

```
struct Data  // size: 12, align: 4
{
char a;                         /* offset: 0, size: 1
char padding[3];                           size: 3 */
int b;                          /* offset: 4, size: 4 */
char c;                         /* offset: 8, size: 1
char padding[3];                           size: 3 */
};
```

Listing 4: Die Struktur Data mit drei sortierten Elementen in der Ansicht von C++ Insights

```
struct Data  // size: 8, align: 4
{
int b;                          /* offset: 0, size: 4 */
char a;                         /* offset: 4, size: 1 */
char c;                         /* offset: 5, size: 1
char padding[2];                           size: 2 */
};
```

Listing 5: Die Funktion Fun mit [[noreturn]] markiert

```
[[noreturn]] void Fun()
{
try {
 while(true) { Process(); }

} catch(...) {}

std::exit(EXIT_FAILURE);
}
```

Listing 6: Funktion CustomAbort mit [[noreturn]] markiert

```
[[noreturn]] void
CustomAbort(int error,
  std::source_location loc =
  std::source_location::current())
{
LogError("{} from {}", error, loc);

reset();
}
```

Die Strategie besteht also darin, die Funktion `CustomAbort` im Code häufiger zu verwenden, beispielsweise an allen Stellen, an denen eine unerwartete Exception auftreten kann. Dadurch multipliziert sich – anders als bei `Fun` – die Ersparnis: Je mehr solcher Funktionen im jeweiligen Code existieren, desto höher die Einsparung.

In die Compilezeit verlagern

Mit `constexpr` stellt C++11 ein Mittel bereit, Berechnungen von der Laufzeit in die Compilezeit zu verschieben. Jede verschobene Berechnung bedeutet – während der Laufzeit – mehr Zeit für andere Dinge. Im besten Fall kommt die betreffende Funktion zur Laufzeit nie zum Einsatz, sodass der Compiler keinen Assemblercode für diese Funktion generieren muss.

Damit ergibt sich ein weiterer Vorteil: Kleinere Programme kosten nicht nur weniger Platz auf der Festplatte, sie sparen auch Ressourcen beim Download, was sich besonders bei Programmen mit hohen Downloadzahlen bemerkbar macht. Ein Beispiel, bei dem `constexpr` abhängig vom Compiler einen erheblichen Unterschied macht, zeigt Listing 7.

Die Klasse `Apple` hat einen Konstruktor, der ein Argument entgegennimmt. Der erhaltene Wert wird intern gespeichert. In `Use` wird ein `Apple`-Objekt mit dem Wert 9 angelegt, also mit einem zur Compilezeit bekannten Wert. Dennoch wird sie erst zur Laufzeit initialisiert. Ist der Konstruktor von `Apple` mit `constexpr` markiert, ergibt sich eine Verbesserung. Das gilt zumindest mit dem Compiler der GCC (Gnu Compiler Collection), der äußerst aggressiv beim Optimieren von mit `constexpr` markierten Funktionen vorgeht. Die Variante ohne `constexpr` benötigt 22 Assemblerinstruktionen, mit `constexpr` optimiert GCC diesen Code bereits bei `-O0` auf sieben Instruktionen, wie sich im Compiler Explorer nachvollziehen lässt (siehe ix.de/zdrt).

Listing 7: Eine Klasse Apple mit einem Konstruktor

```
class Apple
{
public:
explicit Apple(int i)
: x{i}
{}

private:
int x;
};

auto Use()
{
Apple a{9};

return a;
}
```

Listing 8: Exemplarisches Funktionstemplate Log

```
template<typename T>
void Log(T&& t)
{
std::clog << t << '\n';
}
```

Listing 9: Aufrufe von Log

```
int       i{};
const int ci{};

Log(i);
Log(2);
Log(ci);
```

Templates richtig verwenden

C++ bietet mit Templates eine hervorragende Möglichkeit, Code unter der Annahme von Eigenschaften eines Datentyps zu schreiben und dem Compiler das Ausfüllen für alle passenden Varianten von Datentypen zu überlassen. Compiler und die mit ihnen verbundenen Optimizer sind extrem gut, speziell bei Code, den der Compiler selbst erzeugt.

Dennoch gilt es, einige Muster zu beachten. Ein Beispiel hier ist die Übergabe von Parametern. Seit C++11 die Move-Semantik eingeführt hat, ist es verführerisch, für effizienten Code bei Templateparametern r-value-Referenzen zu nutzen. Damit las-

Listing 10: Transformation des Funktionstemplates Log mit C++ Insights

```
/* First instantiated from: insights.cpp:14 */
#ifdef INSIGHTS_USE_TEMPLATE
template<>
void Log<int&>(int& t)
{
std::operator<<(std::clog.operator<<(t), '\n');
}
#endif

/* First instantiated from: insights.cpp:15 */
#ifdef INSIGHTS_USE_TEMPLATE
template<>
void Log<int>(int&& t)
{
std::operator<<(std::clog.operator<<(t), '\n');
}
#endif

/* First instantiated from: insights.cpp:16 */
#ifdef INSIGHTS_USE_TEMPLATE
template<>
void Log<const int&>(const int& t)
{
std::operator<<(std::clog.operator<<(t), '\n');
}
#endif
```

sen sich Parameter verschieben, was auf den ersten Blick positiv erscheint. Ein Beispiel ist die Funktion `Log` in Listing 8. Die dort gezeigte Version ist sicherlich stark vereinfacht, da sie über nur einen Parameter verfügt. Sie zeigt aber, dass mit jedem an das Funktionstemplate `Log` übergebenen Datentyp eine neue Instanziierung einhergeht. Listing 9 zeigt drei Aufrufe von `Log`, alle nutzen den Datentyp `int`. Dadurch gibt es eine Instanziierung von `Log` für `int`.

Die Wahrheit sieht jedoch anders aus. Erneut bietet C++ Insights eine Möglichkeit, hinter die Kulissen zu schauen (siehe Listing 10). Hier sind gleich drei Instanziierungen zu erkennen: eine für `int&`, die zweite für `int` und die dritte für `const int&`. Der Grund: Das Template verhält sich durch die r-value-Referenz vollkommen korrekt und unterscheidet zwischen dem temporären Integer `2` und der konstanten beziehungsweise nicht konstanten Integer-Variablen.

An dieser Stelle ist die Move-Semantik kein Vorteil, da in `Log` selbst keine Daten verschoben, sondern lediglich Daten ausgegeben werden. In einem solchen Fall eignet sich der Parameter `const T&` wesentlich besser. Listing 11 zeigt diese Modifikation.

Mit `const T&` nimmt `Log` auch weiter Objekte sowohl vom Typ `const` als auch vom Typ nicht konstanter Integer entgegen – dank `&` effizient, da ohne Kopiervorgang (siehe Listing 12).

Fazit

Durch das gezielte Verwenden von C++-Features ist das ressourcensparende Programmieren nicht nur auf eingebetteten Geräten möglich. Durch Übertragen der Muster von eingebetteten Systemen auf Desktop-Anwendungen werden Programme dort flüssiger. (sun@ix.de)

Listing 11: Alternativer Aufruf von Log

```
template<typename T>
void Log(const T& t)
{
std::clog << t << '\n';
}
```

Listing 12: Transformation des alternativen Funktionstemplates Log mit C++ Insights

```
/* First instantiated from: insights.cpp:14 */
#ifdef INSIGHTS_USE_TEMPLATE
template<>
void Log<int>(const int& t)
{
std::operator<<(std::clog.operator<<(t), '\n');
}
#endif
```

Quellen

[1] Scott Meyers; Effektives C++ in Embedded Systems; aristeia.com 2011

[2] Vortrag von Scott Meyers und alle Beispiele auf godbolt.de und cppinsights.io siehe ix.de/zdrt

Andreas Fertig

beschäftigt sich als Trainer und Berater mit C++. Sein Fokus liegt auf Embedded-Software.

Energieeffiziente Rechenzentren

Einen großen Anteil am Gesamtenergiebedarf der IT haben die Rechenzentren. Hier steht oft das Verhältnis von IT-Verbrauch – von Server, Storage, Netz – zu dem der Infrastruktur – also Kühlung und Stromversorgung – im Mittelpunkt. Doch greift das zu kurz: IT und Infrastruktur müssen beide besser und effizienter werden, ganz gleich, wie sich das Verhältnis, die PUE, dadurch verschiebt. Damit einher geht auch ein Ringen um das richtige Messen, Rechnen und somit um geeignete Methoden und Kennzahlen.

Kennzahlen – Mit SIEC die Energieverschwendung ungenutzter Rechenressourcen offenlegen 72

Messverfahren – Mit KPI4DCE die Energie- und Ressourceneffizienz von Rechenzentren beurteilen 78

Alternatives Design – Open Compute Project: Energie- und Raumeffizienz zu niedrigsten Kosten? 84

Cloud-Services – Den Fußabdruck einzelner IT-Dienste messen 92

Dashboards – Die CO_2-Emissionen der eigenen Cloud-Nutzung visualisieren 96

Normen – Ökozertifikate für nachhaltige Rechenzentren 100

Kommentar – Welche Bezugspunkte braucht eine Umweltkennzahl? 105

SIEC – Die Energieverschwendung ungenutzter Rechenressourcen offenlegen

Auszeit unter Aufsicht

Dr. Ludger Ackermann, Dr. Dirk Harryvan

Nur wer die Energieverschwendung von Servern im Idle-Modus sichtbar macht, kann sie reduzieren.

Ist ein Server aktiv, also nicht vollständig im Idle-Modus, dann tut er irgendetwas Nützliches. Doch ist es bisher nicht gelungen, diesen Nutzen von IT-Leistung genauer zu messen und daraus eine Kennzahl zu generieren: Zu komplex ist das Zusammenspiel der Prozesse, die auf einer CPU laufen, zu ungenau die Abgrenzung, was Applikation und was Overhead ist – etwa das Betriebssystem, ohne das es aber nicht geht –; ganz zu schweigen von der Unmöglichkeit, den Energieverbrauch eines Servers aufzuteilen für Nützliches und für Unnützes.

Sicher ist also nur, wann ein Server keinen Nutzen erbringt, nämlich wenn er sich im Idle-Zustand befindet. Dann wartet er schlicht auf Arbeit – und benötigt dabei Energie, und zwar gar nicht wenig. Während die Diskussion um die Energieeffizienz von Rechenzentren lange Zeit auf den Energieverbrauch der technischen Anlagen der Gebäudetechnik, also Stromversorgung und Kälteerzeugung, abhob und Kennzahlen wie die PUE (Power Usage Effectiveness) allgemeine Akzeptanz gefunden haben, bleibt doch eine – berechtigte – Kritik daran, dass diese Sichtweise quasi die Ursache ignoriert, nämlich den Energieverbrauch der IT-Komponenten.

Ein Weg, die Sicht auf die Energieeffizienz von Rechenzentren zu erweitern und die IT-Komponenten in die Betrachtung einzubeziehen, besteht darin, den Energieverbrauch von Servern im Idle-Modus zu ermitteln. Die Kennzahl Server Idle Energy Coefficient (SIEC) macht das Verhältnis aus nutzbringend eingesetzter Energie und verschwendeter Energie sichtbar. Mit ihr kann die IT ihre Energieeffizienz ermitteln und verbessern und damit substanziell zur notwendigen Steigerung der Energieeffizienz in Rechenzentren beitragen.

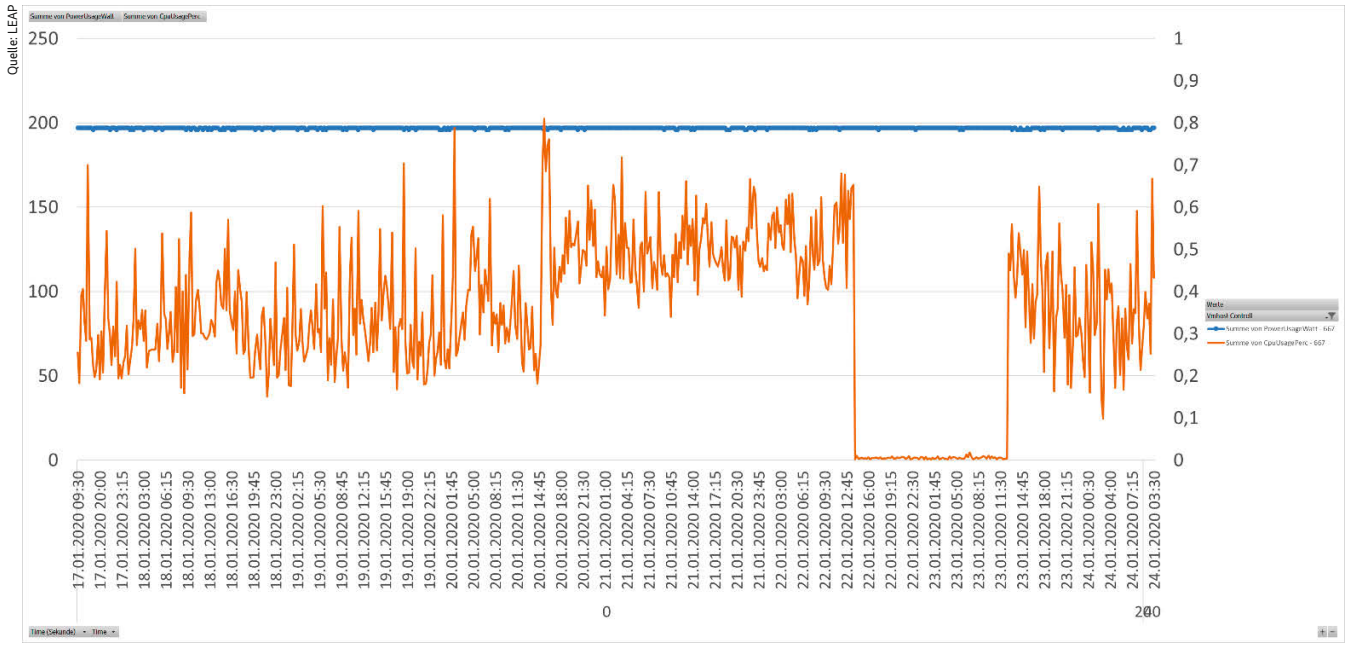
iX-TRACT

- Der Idle-Betrieb von Servern bedeutet pure Energieverschwendung. Er lässt sich aber nicht gänzlich abschaffen, sondern nur reduzieren.
- Die Kennzahl SIEC (Server Idle Energy Coefficient) weist den Anteil des Energieverbrauchs eines Servers aus, der im Idle-Modus verloren geht.
- Der SIEC kann jederzeit und kontinuierlich im produktiven Betrieb gemessen und ausgewiesen werden.
- Der Kennwert offenbart Möglichkeiten des Energiesparens im Idle-Modus.

Server Idle Energy Coefficient

Die Idee stammt aus den Niederlanden und wurde im Rahmen des Forschungsprojekts LEAP entwickelt und veröffentlicht (siehe ix.de/zsu6). Die Forscher maßen in Amsterdamer Rechenzentren den Energieverbrauch der Server bei unterschiedlichen Power-Management-Einstellungen. Server mit Power-Management-Einstellung High Performance ziehen eine konstant hohe elektrische Leistung, unabhängig von ihrer CPU-Last: In Abbildung 1 bleibt die blaue Linie konstant bei einem Energiekonsum von etwa 200 Watt, während die CPU-Last, die orange Kurve, um die 50 Prozent schwankt – schon ein recht hoher Wert für einen Server.

Besonders auffällig ist aber der Teil der Linie, in dem die CPU-Last über einige Stunden fast null beträgt: Der Server zieht ohne Aktivitäten auch hier 200 Watt. Das ist pure Energiever-



Trotz sich verändernder CPU-Last (orange) bleibt der Stromkonsum (blau) bei der Power-Management-Einstellung High Performance konstant hoch (Abb. 1).

schwendung, die sich mit der Zeit zu einer beachtlichen Summe addiert: 200 Watt sind 0,2 kW × 24 Stunden/Tag = 12 kWh/Tag oder 0,2 kW × 8760 Stunden/Jahr = 4380 kWh/Jahr. Diese Menge ist eigentlich noch mit dem PUE-Wert des Rechenzentrums zu multiplizieren, in einem guten Fall 1,3, dann ergeben sich 4360 kWh × 1,3 = 5694 kWh – fürs Nichtstun.

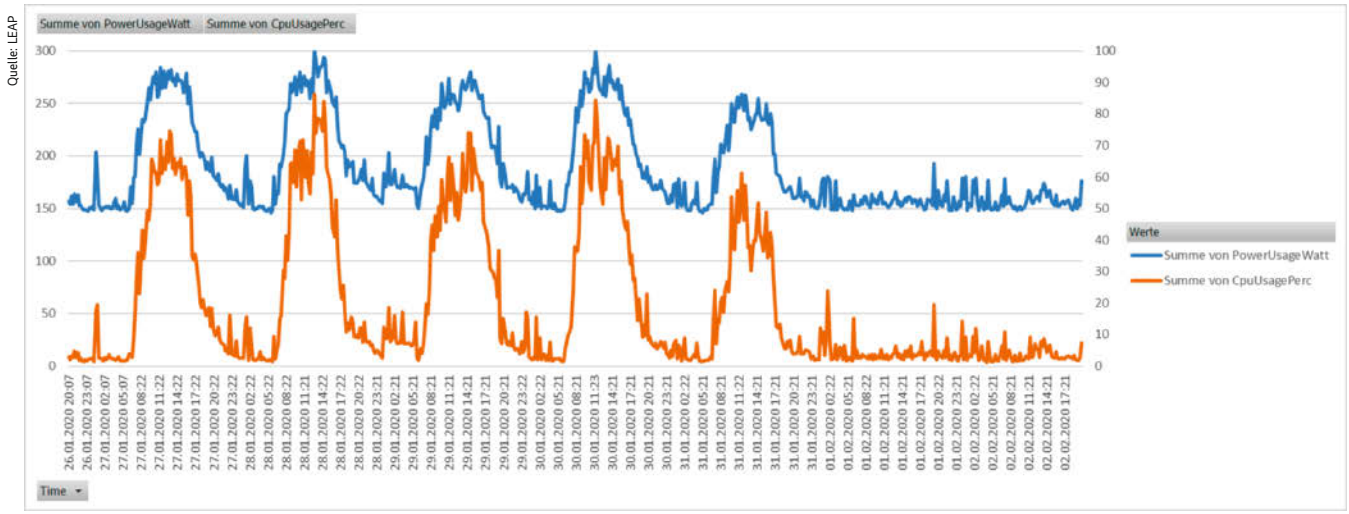
Andere Server mit der Power-Management-Einstellung Balanced passen dagegen ihre elektrische Leistung der CPU-Last an. Blau ist in Abbildung 2 wieder die elektrische Leistungsaufnahme dargestellt und orange die CPU-Last. Diesmal passt der Server die Stromaufnahme an die CPU-Last an. Im Idle-Modus zieht er etwa 150 Watt, im Active-Modus bis zu 300 Watt.

Sehr deutlich ist auch die Wochenkurve des Servers, der virtuelle Desktops bedient: Sie beginnt in der Nacht von Sonntag auf Montag mit einer niedrigen CPU-Last, steigt dann mit Beginn der Büroarbeitszeit und sinkt zur Nacht hin wieder ab. Das wiederholt sich an den anderen Werktagen, nur am Wochenende ist die CPU-Last weitgehend niedrig. Aber auch dieser VDI-Server (Virtual Desktop Infrastructure) verschwendet nicht wenig Energie: im Idle-Modus ganze 150 Watt. Das ist jedoch noch besser, als würde er die ganze Zeit über die 300 Watt des Active-Modus ziehen.

Auswirkungen auf die CPU-Performance

Der Server in der ersten Abbildung mit der konstanten Leistungsaufnahme befindet sich im Power-Management-Modus High Performance, das heißt, er verringert die Leistungsaufnahme nicht, wenn die CPU-Last sinkt. High Performance bedeutet allerdings nicht, dass der Server damit auch die höchste Leistung bringt. Durch die konstante CPU-Frequenz entsteht auch eine konstante Wärme. Die CPU-Frequenz wird daher begrenzt auf den TDP-Wert (Thermal Design Power), um ein Überhitzen zu verhindern.

Die beste Performance erreicht die CPU im Balanced-Modus. Werden die Kerne der CPU kaum oder gar nicht genutzt, regelt das System ihre Frequenzen herunter. Durch den geringeren



Bei der Power-Management-Einstellung Balanced hingegen passt der Server seinen Energiebedarf an die CPU-Last an (Abb. 2).

Nach dem Umschalten in den Balance-Modus sinkt die Leistungsaufnahme sofort um etwa 30 Watt (Abb. 3).

Strombearf sinkt auch die Wärmeabgabe, sodass der TDP-Wert unterschritten wird. Die dadurch entstehende thermische Reserve kann die CPU nutzen, um bei einer Leistungsanforderung kurzzeitig über den TDP-Wert hinauszugehen und die Turbofrequenzen einzusetzen. Wenn also wirklich Performance benötigt wird, ist die im Balanced-Modus höher als im High-Performance-Modus.

Im dritten Modus, dem Energy-Save-Modus, wird der CPU-Takt begrenzt, sodass die höheren Frequenzen nicht zur Verfügung stehen. Die Leistung im Idle-Zustand bleibt davon aber weitgehend unberührt, sodass sich nur ein geringer Energiesparvorteil gegenüber dem Balanced-Modus ergibt. Es gibt also keinen vernünftigen Grund, einen Server nicht im Balanced-Modus zu betreiben. Im Idle-Zustand braucht er weniger Energie, und im Active-Zustand steht ihm mehr Performance zur Verfügung.

Abbildung 3 zeigt wieder orange die CPU-Last und blau die Leistungsaufnahme. Nach der Hälfte der Messung wird der Server auf den Balanced-Modus umgeschaltet: Die Aufnahme sinkt sofort um etwa 30 Watt, die CPU-Last bleibt unverändert.

Den Idle-Verbrauch errechnen

Um den im Idle-Modus auftretenden Energieverbrauch zu ermitteln, wurde für den Server im Balanced-Modus die CPU-Last und die Leistungsaufnahme gemessen. Während die Messung der CPU-Last zum Standardrepertoire des Systemmanagements zählt, stellen fast alle Netzteile heute Messwerte zur Leistungsaufnahme bereit, die sich etwa unter Linux mit dem Befehl `powerstat` auslesen lassen (siehe ix.de/zsu6). Die Tabelle in Abbildung 4 zeigt einen kleinen Ausschnitt aus der Messung im LEAP-Projekt.

Der Tag der Messung, der 27.01.2020, war der Montag in der obigen Wochenkurve. An den Zeitstempeln lässt sich ablesen, dass der Anstieg der CPU-Last mit dem Arbeitsbeginn in den Büros einhergeht. Innerhalb der hier aufgeführten fünf Stunden streuen sich bereits die Werte der CPU-Last. Aus diesen Messungen lässt sich eine Kurve der Leistungsaufnahme über der CPU-Last zeichnen, die vom Idle-Zustand bis zum Active-Zustand reicht.

Abbildung 5 zeigt die gesamte Wochenkurve als blaue Punkte. Auffällig ist die Häufung bei CPU-Lasten kleiner als 10 Prozent. Die gepunktete grüne Linie ist eine lineare Näherungskurve, deren Formel in der Grafik zu sehen ist. Die Kurve schneidet die y-Achse bei $P_0 \approx 153$ Watt. Die gepunktete orange Linie ist eine quadratische Annäherung, die der Verteilung etwas besser entspricht. Sie ergibt einen Wert von $P_0 \approx 146$ Watt, der ungefähr dem obersten Wert in der Tabelle entspricht: $P_{idle} \approx 148$ Watt bei 1,8 Prozent CPU-Last. Alle Punkte streuen um die Linien, daher kommt es auch nicht so sehr auf Genauigkeit beim Ermitteln von P_0 an.

Der Energieverbrauch des Servers im Idle-Modus zu einem Zeitpunkt t bestimmt sich wie folgt: Immer dann, wenn die CPU-Last unter 100 Prozent liegt, wird die Differenz zu den 100 Prozent multipliziert mit der Zeit, die der Server im Idle-Modus verbringt. Das wiederum multipliziert mit P_0 ergibt den Energieverbrauch:

$$E_{idle}\ (t)\ =\ (100\% - CPU - Last) \times P_0 \times Messzeitraum$$

Der Messzeitraum wird in Stunden angegeben. Wenn wie in der Tabelle alle 15 Minuten gemessen wird, beträgt er also 0,25 h. Beispielsweise beträgt die CPU-Last um 8:22 Uhr 32,13 Prozent und P_0 150 Watt oder 0,15 kW:

$$E_{idle}\ (t)\ =\ (100\% - 32,13\%) \times 0,15 \times 0,25 = 0,68 \times 0,15 \times 0,25 \approx 0,06\ kWh$$

Sobald P_0 einmal bestimmt wurde, lässt sich E_{idle} also aus der Messung der CPU-Last berechnen. Als Nächstes soll daraus eine Kennzahl ermittelt werden, die die Energieverschwendung des Servers im Idle-Modus wiedergibt. Dafür ist die Rechnung von oben über alle Messpunkte n aufzuaddieren, was zu folgender Formel führt:

$$E_{idle} = \sum_{1}^{N} E_{idle}(t, n)$$

Nun fehlt noch der Energieverbrauch des Servers E_{gesamt} im gesamten Messzeitraum, also etwa in der Woche aus Abbildung 1. Den liefert wieder das Netzteil. Im Zweifel nimmt man den Mittelwert multipliziert mit dem Messzeitraum.

$$E_{gesamt} = \sum_{1}^{N} P(n) \times t(n)$$

Die Kennzahl, die sich daraus ergibt, soll Server Idle Energy Coefficient (SIEC) heißen. Sie ist in einem Artikel bei 4E TCP veröffentlicht, heißt dort aber noch Server Idle Coefficient oder SIC (siehe ix.de/zsu6). Der SIEC errechnet sich wie folgt:

$$SIEC = \frac{E_{idle}}{E_{gesamt}} \times 100\ \%$$

Er gibt also den Anteil der verschwendeten Energie E_{idle} zum gesamten Energieverbrauch des Servers E_{gesamt} an. Beim VDI-Server oben mit der Wochenkurve haben sich folgende Zahlen ergeben:

$$E_{idle} = 20,2\ kWh$$
$$E_{gesamt} = 31,9\ kWh$$

Daraus ergibt sich ein SIEC von 63,3 Prozent:

$$SIEC = \frac{20,2\ kWh}{31,9\ kWh} \times 100\ \% = 63,3\ \%$$

Obwohl der Server mit seiner Power-Management-Einstellung die Leistung an die CPU-Last anpasst, verschwendet er immer noch deutlich mehr als die Hälfte seiner Energie. Das macht deutlich, dass neben allen anderen Umweltwirkungen, die derzeit auch politisch in der Diskussion sind, die Energieeffizienz von IT-Komponenten und von Rechenzentren weiter im Vordergrund stehen sollte.

Andere Serverkennzahlen für Rechenzentren

Es sind schon einige Kennzahlen und KPIs (Key Performance Indicators) für Rechenzentren in der Welt – warum also noch eine? Die PUE gibt ähnlich wie ihre Verwandten, die Cooling Efficiency Ratio (CER), die Carbon Usage Effectiveness (CUE) und die Water Usage Effectiveness (WUE), ausschließlich Auskunft über die Gebäudetechnik, sie betrachtet die IT also gar nicht. Dagegen zielen die Kennzahlen $ITEE_{SV}$ und $ITEU_{SV}$ auf die Energieeffizienz und die Nutzung der IT-Komponenten.

Die $ITEE_{SV}$ (IT Equipment Energy Efficiency for servers) legt den Fokus auf die Energieeffizienz eines Servers im Active-Modus. Der Wert lässt sich durch einen frei gewählten Benchmark ermitteln, wodurch nur die Server vergleichbar sind, die dem gleichen Benchmark unterzogen wurden. Die Kennzahl lässt sich nicht im produktiven Betrieb ermitteln, sondern nur auf dem Prüfstand. Die $ITEU_{SV}$ (IT Equipment Utilization for servers) hingegen erfasst lediglich die CPU-Nutzung, unabhängig von der Leistungsaufnahme des Servers. Beide Kennzahlen geben also nur beschränkten Einblick in den Energieverbrauch des Servers im Betrieb und werden daher nicht häufig genutzt.

Quelle: LEAP

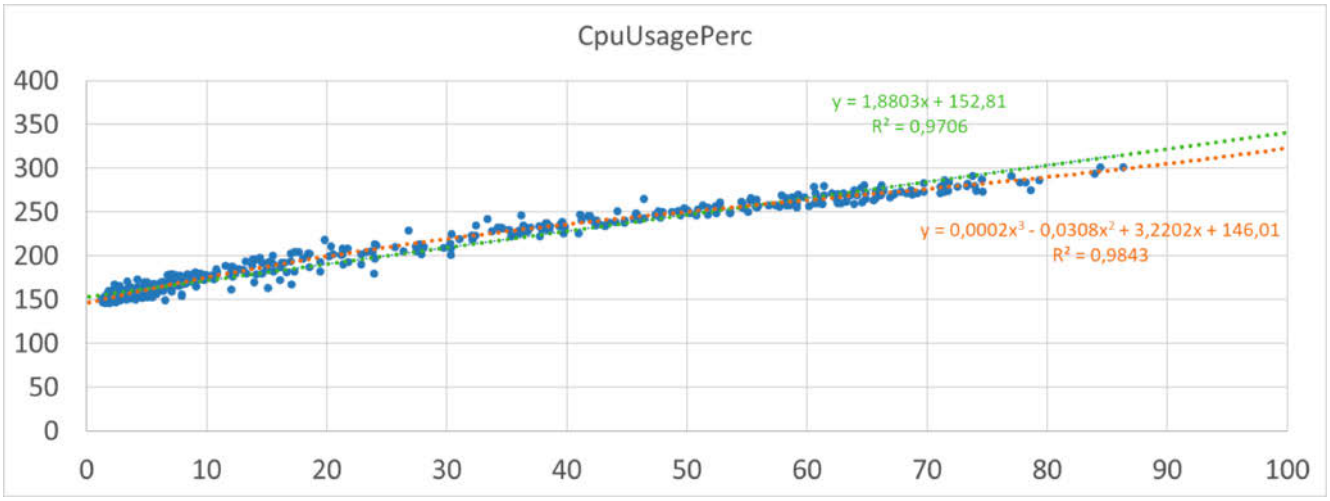
Server ID	Time stamp	Power (Watt)	CPU load (%)
citrix.amsterdam	27/01/2020 05:37	148	1,8
citrix.amsterdam	27/01/2020 05:52	147	1,92
citrix.amsterdam	27/01/2020 06:07	151	3,84
citrix.amsterdam	27/01/2020 06:22	171	3,38
citrix.amsterdam	27/01/2020 06:37	155	2,77
citrix.amsterdam	27/01/2020 06:52	179	8,43
citrix.amsterdam	27/01/2020 07:07	181	16,68
citrix.amsterdam	27/01/2020 07:22	211	28,06
citrix.amsterdam	27/01/2020 07:52	226	36,12
citrix.amsterdam	27/01/2020 08:07	202	22,88
citrix.amsterdam	27/01/2020 08:22	218	32,13
citrix.amsterdam	27/01/2020 08:37	235	43,08
citrix.amsterdam	27/01/2020 08:52	232	34,25
citrix.amsterdam	27/01/2020 09:07	236	39,48
citrix.amsterdam	27/01/2020 09:22	251	49,36
citrix.amsterdam	27/01/2020 09:37	265	46,41
citrix.amsterdam	27/01/2020 09:52	253	50,92
citrix.amsterdam	27/01/2020 10:06	266	65,73
citrix.amsterdam	27/01/2020 10:22	275	63,93
citrix.amsterdam	27/01/2020 10:37	269	61,39

Zum Arbeitsbeginn der Mitarbeiter steigen die CPU-Last und der Energiebedarf des Terminalservers (Abb. 4).

Kaum bekannt ist, dass es eine EU-Eco-Design-Verordnung für Server gibt (siehe ix.de/zsu6). Sie verlangt die Angabe der Leistungsaufnahme im Idle-Zustand und der Energieeffizienz im Active-Zustand (Eff_{Server}). Für den Idle-Zustand setzt die Verordnung zudem einen Grenzwert, der aber von den eingebauten Komponenten abhängt. Trotzdem: Wer eine Schätzung für den Energieverbrauch in der geplanten Lebenszeit vor- und in die TCO (Total Costs of Ownership) aufnimmt, kann mit der Integration der Eco-Design-Verordnung in den Einkaufsprozess gute Server auswählen. Und wer P_0 nicht durch eine Messung ermitteln will, kann die Herstellerangabe des P_{idle}-Wertes aus der Eco-Design-Verordnung als Näherung nutzen.

SIEC ist also die erste Kennzahl, die dem IT-Betrieb einen Einstieg in die Ermittlung der Energieeffizienz der IT-Komponenten im produktiven Betrieb ermöglicht, und damit den Be-

Quelle: LEAP



Die Messwerte drängen sich an der quadratischen Annäherung (orange), die der Verteilung etwas besser entspricht als die grün dargestellte lineare Näherungskurve (Abb. 5).

ginn einer Optimierung. Jede eingesparte Kilowattstunde IT-Energie wird zudem mit dem Faktor PUE belohnt, indem in den technischen Anlagen weitere Energie eingespart werden kann.

Die Kennzahl SIEC nutzen

Selbst wenn alle Server im Balanced-Modus laufen, ermöglicht es der SIEC, die dann weiterhin bestehende Verschwendung zu quantifizieren und den Druck auf die Hersteller zu erhöhen, den Wert P_0 oder P_{idle} zu verbessern. Je weniger Leistung ein Server im Idle-Zustand benötigt, desto effizienter werden die Rechenzentren.

Der Balanced-Modus lässt sich verbessern, indem die CPU-Zustände noch weiter heruntergefahren werden, wenn der Server keine Performance benötigt. Er sollte sich in einer Art Schlafmodus befinden, in dem er zwar nicht ausgeschaltet ist, sich aber darauf beschränkt, zu prüfen, ob er wieder geweckt werden soll. Aus dem Schlafmodus ist er dann in wenigen Millisekunden wieder bei voller Leistungsfähigkeit, sodass der Nutzer keine Verzögerung bemerkt.

Werden Server ausgetauscht, bietet sich die Möglichkeit, den SIEC zu optimieren. Werden beim Einkauf neuer Server die Gesamtkosten aus Investition beim Kauf, Wartung und Energieverbrauch betrachtet, sind energieeffiziente Server im Vorteil. Um den Energieverbrauch berechnen zu können, werden die Angaben aus der Eco-Design-Verordnung benötigt und das Verhältnis aus der zu erwartenden Zeit im Active-Zustand und im Idle-Zustand.

Wesentliche Verbraucher in Servern sind die kleinen Lüfter, die Abwärme beseitigen. Flüssiggekühlte Server benötigen keine Lüfter, sind effizienter und können erheblich mehr Leistung auf kleinem Raum anbieten. Dadurch ist P_{idle} geringer und der SIEC wird besser. Zudem benötigen solche Server, wie sie in HPC-Rechenzentren üblich sind, einen geringeren Overhead in der Kühlung, wodurch das Rechenzentrum insgesamt erheblich energieeffizienter wird. Allerdings muss das Rechenzentrum dazu mit der Flüssigkühlung umgehen können.

Darüber hinaus lässt sich der SIEC als Kennzahl auf das ganze Rechenzentrum ausweiten. Er könnte dann etwa $SIEC_{DC}$ heißen und gibt an, wie viel Energie insgesamt in Servern und der Gebäudetechnik durch Idle-Zeit verschwendet wurde. Dann kann man die Auswirkung von Verbesserungen auf beiden Seiten, IT-Komponenten und Gebäudetechnik, messen und nachweisen. Die Sustainable Digital Infrastructure Alliance (SDIA) bereitet eine Schnittstelle vor, über die Daten zur Energieeffizienz und Nachhaltigkeit von Rechenzentren zwischen Gebäudetechnik und IT über einen Open Data Hub ausgetauscht werden können.

Fazit

Der KPI Server Idle Energy Coefficient (SIEC) ermöglicht das Berechnen der Energie, die ein Server im Idle-Zustand benötigt, also verschwendet. Der SIEC lässt sich jederzeit und kontinuierlich im produktiven Betrieb messen und ausweisen. Der Balanced-Modus im Power-Management erlaubt es Servern, sowohl weniger Energie im Idle-Zustand zu verbrauchen, als auch eine höhere Performance im Active-Zustand bereitzustellen. Daher sollten alle Server schnellstmöglich in den Balanced-Modus umgestellt werden. Weiter lässt sich der SIEC durch die Anwendung der EU-Eco-Design-Verordnung verbessern, indem Server mit kleinen P_{idle}-Werten ausgewählt werden und eine Gesamtkostenbetrachtung inklusive Energieverbrauch vorgenommen wird. Wer den SIEC auf seinen Servern messen will, kann den Autoren die Daten oder das Ergebnis an siec@dc-e.de schicken. Mit den vertraulich behandelten Daten möchten sie zeigen, dass der SIEC eine praxisnahe Kennzahl ist. (sun@ix.de)

Quellen

Alle Dokumentationen siehe ix.de/zsu6

Dr. Ludger Ackermann

ist Senior Data Center Consultant bei der Data Center Excellence GmbH, Auditor für den Blauen Engel für Rechenzentren und Editor der Normenteile EN 50600-3-1, EN 50600-4-7 und EN 50600-5-1.

Dr. Dirk Harryvan

ist Senior Consultant und Certified Data Protection Officer (DPO) bei der Certios B. V. in den Niederlanden.

Mit KPI4DCE die Energie- und Ressourceneffizienz von Rechenzentren beurteilen

Weiter gefasst

Marina Köhn

Die bekannte RZ-Effizienzkennzahl PUE konzentriert sich auf die Infrastruktur. Das vom UBA erarbeitete Indikatorensystem KPI4DCE soll dagegen den gesamten ökologischen Fußabdruck von IT und Infrastruktur erfassen.

Sollen digitale Techniken dabei helfen, die Umwelt zu schützen, muss es gelingen, den Energieverbrauch der digitalen Infrastruktur zu senken und den Rohstoffverbrauch auf ein Minimum zu reduzieren. Die Rechenzentren als Herzstück der Digitalisierung haben hierbei eine zentrale Verantwortung.

Doch: Wie viele Rechenzentren existieren überhaupt in Deutschland, Europa und weltweit? Wie viel Energie verbrauchen sie und wie viele Rohstoffe – verbaut in Technik – stecken in ihnen? Und wie effizient werden Energie und Rohstoffe eingesetzt?

Es ist erstaunlich und erschreckend zugleich, dass die Antwort auf alle Fragen lautet: Wir wissen es nicht. Erstaunlich, weil es technisch und organisatorisch möglich wäre, diese Informationen zu erhalten. Erschreckend, weil die Gesellschaft in einem rasenden Digitalisierungszug sitzt, ohne dass ein Stellwerk die Weichen stellt. Die Auswirkungen auf die Klimaerwärmung und den Verbrauch wertvoller Rohstoffe nicht zu kennen, ist leichtsinnig und gefährlich.

Die Debatte über die notwendigen Maßnahmen für eine umweltverträgliche Gestaltung der digitalen Infrastruktur zeigt sich nicht zuletzt im Koalitionsvertrag, in dem es heißt: „Wir werden Rechenzentren in Deutschland auf ökologische Nachhaltigkeit und Klimaschutz ausrichten, unter anderem durch Nutzung der Abwärme. Neue Rechenzentren sind ab 2027 klimaneutral zu betreiben. Öffentliche Rechenzentren führen bis 2025 ein Umweltmanagementsystem nach EMAS (Eco Management and Audit Scheme) ein. Für IT-Beschaffungen des Bundes werden Zertifizierungen wie der Blaue Engel Standard.“

Bisherige Kennzahlen: Irrwege zum Teilerfolg

Neben Umweltmanagementsystemen und Zertifizierungen sind Kennzahlen ein wichtiges Instrument einer nachhaltigen Kli-

iX-TRACT

- Nur mit den richtigen Kennzahlen für Rechenzentren lässt sich die Effizienz der digitalen Infrastruktur bestimmen.
- Das Kennzahlensystem KPI4DCE bewertet anders als die PUE auch die Energieeffizienz der IT.
- Neben dem Stromverbrauch erfassen die Kennzahlen von KPI4DCE auch die Treibhausgasemissionen, den Rohstoffaufwand und den Wasserverbrauch, und zwar über den gesamten Lebenszyklus aller eingesetzten Komponenten hinweg.

maschutzstrategie. Denn nur mit den richtigen, für Rechenzentren geeigneten Kennzahlen lässt sich die Effizienz der digitalen Infrastruktur bestimmen. Die etablierten Kennzahlen erlauben es zwar, den aktuellen Zustand eines Rechenzentrums zu ermitteln, doch vernachlässigen viele von ihnen den wesentlichen Aspekt eines Rechenzentrums, nämlich die Leistung der IT, und sind deshalb nachweislich ungeeignet. Denn die IT hat bei allen vier Umweltwirkungen – abiotischer Rohstoffverbrauch (ADP), kumulierter Energieaufwand (KEA), globaler Treibhausgaseffekt (GWP) und Wasserverbrauch – den größten Anteil. Abbildung 1 zeigt das anhand dreier untersuchter Rechenzentren.

Am häufigsten verwendet wird die Power Usage Effectiveness (PUE). Sie beschreibt, in welchem Verhältnis der gesamte Energieverbrauch eines Rechenzentrums zum Energieverbrauch der IT steht, und wird gebildet, indem der jährliche Gesamtenergiebedarf des Rechenzentrums durch den der IT geteilt wird. Die PUE soll die Energieeffizienz eines Rechenzentrums ausweisen, kann aber genau das nicht leisten. Denn sie gibt nur Auskunft über die Energieeffizienz der Infrastruktur samt Klimatisierung, Energieumwandlung, Notstromversorgung und kann keine Aussagen über die Energieeffizienz der Kernaufgaben des Rechenzentrums machen, nämlich über die Rechen-, Speicher- und Übertragungsleistung. Noch schlimmer: Wenn die Kernaufgaben effizienter erledigt werden, also tatsächlich Energie und Hardwareressourcen einsparen, dann wird die PUE sogar schlechter. Das macht deutlich, dass ein neuer Ansatz nötig ist.

Abbildung 2 stellt die berechneten PUE-Werte und die durchschnittlichen CPU-Auslastungen der Server in zwölf untersuchten Rechenzentren gegenüber. Während sich die PUE dort im Jahresdurchschnitt zwischen 1,19 und 1,75 bewegt, liegt die mittlere CPU-Auslastung aller Server in den jeweiligen Rechenzentren zwischen 4 und 80 Prozent. Eine Korrelation zwischen niedrigem PUE-Wert und hoher CPU-Auslastung ist nicht erkennbar.

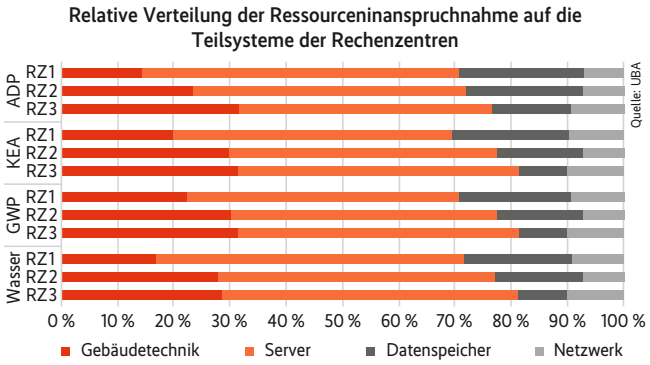
Mit der ganzheitlichen Sicht die Potenziale erkennen

Wer die Effizienz beschreiben will, muss das Verhältnis von Nutzen und Aufwand ermitteln – oder, auf Rechenzentren übertragen, das Verhältnis von Rechenzentrumsleistung und IT-Ressourcen. Mit der Frage, wie das zu bewerkstelligen ist, hat sich das Umweltbundesamt zusammen mit der TU Berlin und dem Öko-Institut in einem Forschungsprojekt beschäftigt und das Kennzahlensystem Key Performance Indicators for Data Center Efficiency (KPI4DCE) entwickelt, mit dem sich die Effizienz eines gesamten Rechenzentrums bewerten lässt (siehe ix.de/zxn5).

Beim KPI4DCE wird dem jeweiligen Nutzen des Rechenzentrums, also dem Rechnen, Speichern, Datenübertragen und der Infrastrukturleistung, der jeweils verursachte Umweltaufwand zugewiesen: also Rohstoff-, Energie- und Wasserverbrauch. Einbezogen wird der gesamte Lebenszyklus, also der Umweltaufwand für die Herstellung, Nutzung und Entsorgung der IT. Die KPI4DCE-Kennzahlen berechnen sich jeweils als Quotient aus Nutzen und Aufwand:

```
KPI4DCE = Nutzen/Aufwand
```

Der IT-Nutzen lässt sich anhand der jeweiligen IT-Leistung ermitteln: die Rechenleistung durch die durchgeführten Rechenoperationen, die Speicherleistung anhand des belegten Speichers und die Netzleistung anhand der übertragenen Datenmenge. Dagegen lässt sich der Nutzen der Gebäudetechnik nicht



In allen drei untersuchten Rechenzentren geht der größte Anteil bei allen vier Umweltwirkungen – abiotischer Rohstoffverbrauch (ADP), kumulierter Energieaufwand (KEA), globaler Treibhausgaseffekt (GWP) und Wasserverbrauch – zulasten der IT (Abb. 1).

direkt aus den Leistungsdaten, sondern nur über eine Hilfsgröße ableiten. Denn die Aufgabe der Gebäudetechnik liegt darin, den IT-Betrieb sicherzustellen. Sie hat keinen eigenen Nutzen und ist abhängig von den Qualitätsanforderungen und dem Leistungsvermögen der IT. Um dennoch den Infrastrukturnutzen darzustellen, zieht man den standardisierten Kennwert DCiRE (Data Center Infrastructure Ressource Efficiency) als Hilfsgröße hinzu. Der Infrastrukturnutzen wird als das Verhältnis der Ressourceninanspruchnahme der IT zur Ressourceninanspruchnahme des RZ über den Lebensweg bestimmt. Dabei ist der Nutzen der vier RZ-Teilbereiche einzeln zu berechnen.

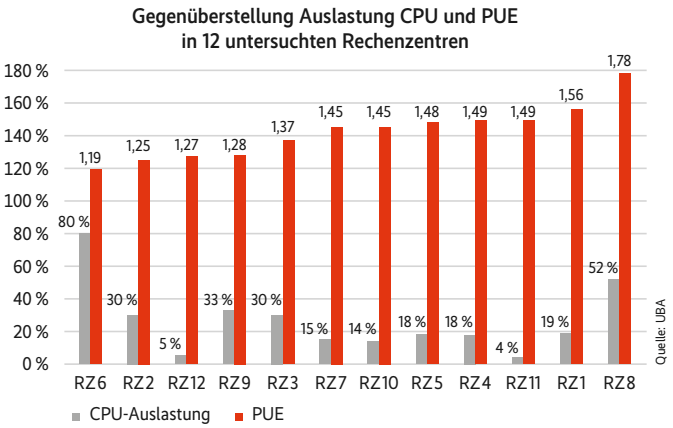
Den Nutzen berechnen

Der Rechennutzen ergibt sich aus der Frage, welche Rechenlast die Server pro Jahr bewältigen. Dazu rechnet man die mit den einschlägigen CPU-Benchmarks ermittelten Performancewerte aller installierten CPUs aufs Jahr hoch und multipliziert das Ergebnis mit der anhand des Monitorings gemittelte CPU-Auslastung pro Jahr:

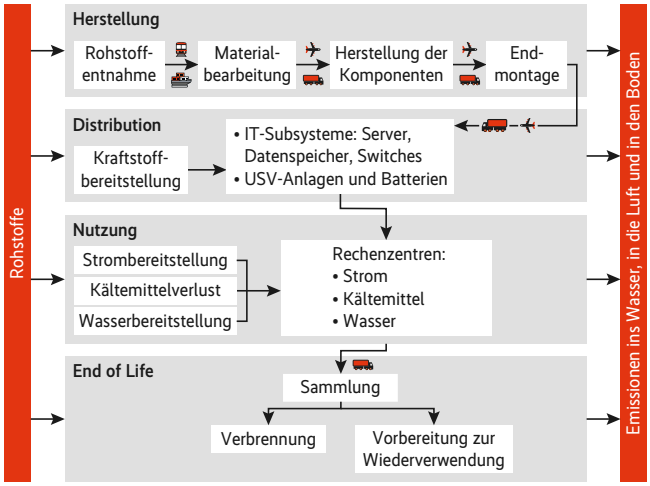
```
Rechennutzen/a = summierte Performance aller
CPUs × mittlere Serverauslastung
```

Der Speichernutzen definiert, wie viel Speicherplatz die Nutzdaten jährlich belegen. Dafür wird der installierte Speicherplatz mit der durchschnittlichen Speicherplatzbelegung multipliziert:

```
GByte/a = installierter Speicherplatz × mittlere
Speicherplatzbelegung
```



In den zwölf untersuchten Rechenzentren lässt sich zwischen der CPU-Auslastung der Server und der PUE keine Korrelation erkennen (Abb. 2).



Die Methode der Ökobilanz für die Berechnung der Umweltwirkungen der RZ-Komponenten erfasst den gesamten Lebenszyklus aller beteiligten Komponenten (Abb. 3).

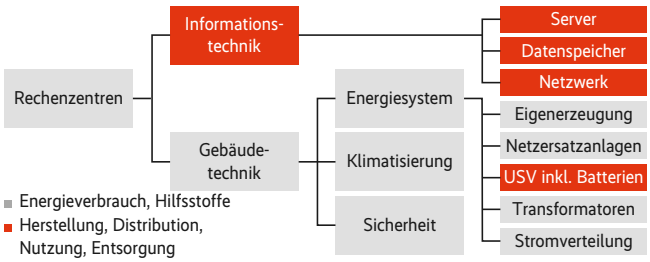
Der Netzwerknutzen gibt die Datenmenge wieder, die jährlich zwischen Rechenzentrum und Außenwelt übertragen wird.

```
GBit Datentransfer/s = mittlerer genutzter Durchsatz
externer Datenverkehr × s/a
```

Der Infrastrukturnutzen gibt an, welche Ressourcenmengen die IT selbst beansprucht. Dazu wird die Ressourceneffizienz der Gebäudetechnik in Anlehnung an die DCiE ermittelt. Sie wird als das Verhältnis der Ressourceninanspruchnahme der IT zu der Ressourceninanspruchnahme des RZ über den Lebensweg definiert.

```
Infrastrukturnutzen = bereitgestellte Ressourcen für
IT = Aufwand Server + Aufwand Speicher + Aufwand
Netzwerk
```

Zum Ermitteln der Ressourcen werden die in der Ökobilanzierung etablierten Ressourcenindikatoren verwendet. Nach diesen richtet sich auch die Einheit. Zu dem so errechneten Nutzen der jeweiligen Teilbereiche wird im nächsten Schritt der Umweltaufwand für die Herstellung und den Betrieb der IT-Hardware und der Infrastrukturtechnik des Rechenzentrums ermittelt. Hierfür kommt die Methode der Ökobilanzierung zum Zuge, die die Umweltwirkung durch die Entnahmen aus der Umwelt etwa von Rohstoffen und durch Einträge in die Umwelt etwa von CO_2-Emissionen bestimmt (siehe Abbildung 3).



Bilanzierungsgrenzen: Die Methode KPI4DCE nimmt für Komponenten mit relativ kurzen Erneuerungszyklen eine komplette Ökobilanz vor. Bei Komponenten mit langen Erneuerungszyklen, die den Großteil der Gebäudetechnik ausmachen, werden dagegen nur der Energieverbrauch und die verbrauchten Hilfsstoffe in der Nutzung des Rechenzentrums betrachtet, nicht aber ihr Herstellungsaufwand (Abb. 4).

Die Methode KPI4DCE zieht hierfür die vier Umweltwirkungskategorien abiotischer Rohstoffaufwand (ADP), Treibhausgasemissionen (GWP), kumulierter Energieaufwand (KEA) und Wasserverbrauch zurate.

- **Abiotischer Rohstoffverbrauch:** ADP (Abiotic Depletion Potential) bewertet die Inanspruchnahme von Mineralien und fossilen Rohstoffen; die Einheit ist kg Antimon-Äquivalent pro Jahr (kg Sb-eq/a).
- **Treibhausgaspotenzial:** GWP (Global Warming Potential) bewertet die Wirkung auf die Erderwärmung; die Einheit ist kg Kohlendioxid-Äquivalent pro Jahr (kg CO_2-eq/a).
- **Kumulierter Energieaufwand:** KEA (Cumulative Energy Demand) beschreibt den Verbrauch an energetischen Ressourcen, und zwar die erneuerbaren und die nicht erneuerbaren; die Einheit ist Megajoule pro Jahr (MJ/a).
- **Wasserverbrauch:** Die Water Usage bemisst sich in Kubikmeter pro Jahr (m^3/a).

Die große Herausforderung der Bilanzierung eines gesamten Rechenzentrums besteht darin, die Vielfalt an Komponenten und die Tiefe der Bilanzierung auf ein handhabbares Maß zu reduzieren. Eine komplette Ökobilanz inklusive Herstellung und Entsorgung wurde für Komponenten mit relativ kurzen Erneuerungszyklen vorgenommen. In Abbildung 4 sind diese Komponenten rot dargestellt. Für einen Großteil der Gebäudetechnik werden dagegen nur der Energieverbrauch und die verbrauchten Hilfsstoffe in der Nutzung des Rechenzentrums betrachtet. Sie sind in den grauen Kästchen aufgeführt. Der Herstellungsaufwand der Gebäudetechnik wie Klima- und Sicherheitsanlagen wurde aufgrund des geringen Einflusses auf das Endergebnis ignoriert.

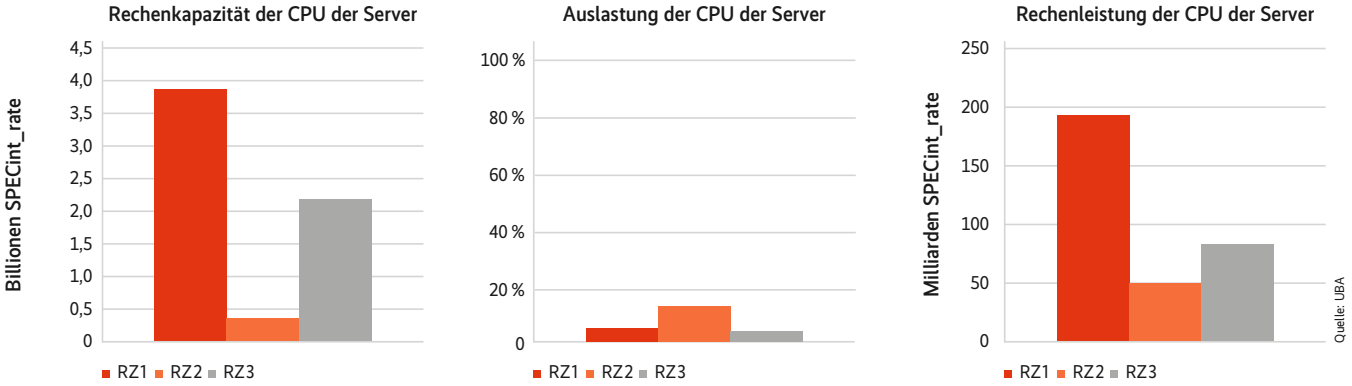
Die Sachbilanzdaten, die die Grundlage für die Bilanzierung der Wirkungsindikatoren bilden, stammen aus öffentlich verfügbaren Publikationen und Datenbanken. Zudem floss die Expertise von IT-Industrie und Wissenschaft mit ein. Die Bilanzierung erfasst grundsätzlich den Zeitraum eines Jahres. Nur so ist sichergestellt, dass wetterbedingte Einflüsse und Leistungsspitzen Berücksichtigung finden. Für die hier vorgestellten Ressourceneffizienzkennzahlen gilt: je größer der Wert, desto besser die Ressourceneffizienz.

Am Beispiel erklärt: KPI4DCE-Berechnung für die Servereffizienz

Die Methode klingt vermutlich kompliziert, das ist sie aber nicht. Beweisen soll das eine Energieeffizienzberechnung der Rechenleistung eines Rechenzentrums. Sie umfasst die Auslastung der installierten CPUs und die Energie, die beim Herstellen, Liefern, Nutzen und Entsorgen der Server aufzuwenden ist.

In der Regel geben die CPUs das Rechenleistungsvermögen der Server vor. Dadurch lässt sich die erbrachte Rechenleistung anhand der installierten Rechenleistung und der mittleren CPU-Auslastung ermitteln. Eine Ausnahme bildet das HPC (High-Performance Computing). Rechenzentren, die überwiegend dem HPC dienen, lassen sich mit dieser Methode nicht bewerten.

In diesem Beispiel sollen die Wirkungsindikatoren KEA und GWP der Server, also ihr kumulierter Energieaufwand und ihr Treibhausgaspotenzial, und damit der CO_2-Fußabdruck der RZ-Rechenleistung bestimmt werden. KEA und GWP weisen sehr häufig den gleichen Effizienzgrad aus. Sie unterscheiden sich jedoch in jenen Fällen, bei denen die Energieversorgung aus erneuerbaren Energiequellen kommt oder die Klimatisierung auf klimaschädliche Gase verzichtet.



Durch die höhere Auslastung von RZ2 relativiert sich dessen geringere installierte Rechenkapazität bei der tatsächlich erbrachten Rechenleistung (Abb. 5).

Die Rechenleistung der Server lässt sich mit Benchmarkdaten der SPEC CPU2006 charakterisieren (siehe Kasten „SPEC CPU2006“ und ix.de/zxn5). Mit dem Rückgriff auf veröffentlichte Benchmark-Ergebnisse kann man die potenzielle Rechenkapazität von Servern erfassen, ohne sie im Produktivbetrieb messen zu müssen. Multipliziert man die Benchmarkwerte eines Servers mit der im Betrieb gemessenen Auslastung, ergibt sich für diesen Server ein abstraktes Maß für die im Betrieb erbrachte Rechenleistung.

Vergleichbare Messergebnisse fürs eigene RZ verwenden

Ergebnisse, die den Regeln der SPEC entsprechen, sind auf deren Webseite veröffentlicht. Die INTrate-Ergebnisse von Servern mit derselben Zahl und Art der CPUs überträgt man auf die eigenen Systeme. Alle Systeme in einem RZ addiert ergeben die installierte Gesamtrechenkapazität des Rechenzentrums beziehungsweise das potenzielle Rechenleistungsvermögen. Über die durchschnittliche CPU-Auslastung lässt sich dann die Rechenleistung feststellen, die das Rechenzentrum in einem Jahr tatsächlich erbracht hat.

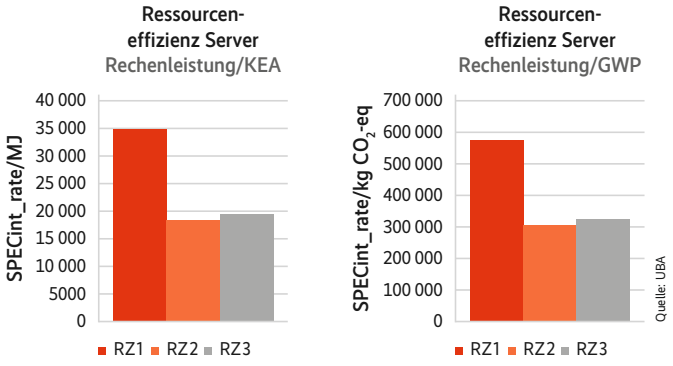
Abbildung 5 stellt die Rechenkapazität, Auslastung und Rechenleistung der Server aus drei vom Umweltbundesamt untersuchten Rechenzentren gegenüber. Zuerst wurde die installierte Rechenkapazität anhand der Datenblätter und vergleichbarer Benchmarkdaten erfasst (links). Die mittlere Grafik fasst die Monitoringdaten der durchschnittlichen Auslastung der installierten CPUs über ein Jahr zusammen. Daraus errechnet sich die erbrachte Rechenleistung (rechts). Zwar hat RZ1 eine zehnfach höhere Rechenkapazität als RZ2, erbringt aber aufgrund der geringen Auslastung der CPUs nur die vierfach höhere Rechenleistung.

Der Aufwand beispielsweise für die Herstellung, Distribution und Entsorgung eines Servers ergibt sich aus der Ökobilanz. Die spezifischen Datensätze für die Komponenten wie CPUs, Speicherchips, unbestückte Leiterplatten und HDDs stammen aus den Lebenszyklusdatenbanken ProBas und ecoinvent (siehe ix.de/zxn5).

Im letzten Schritt ist der Umweltaufwand mit dem IT-Nutzen in Verbindung zu bringen. Im Beispiel ist der Energieaufwand der Rechenleistung als Quotient aus der erbrachten Rechenleistung und den von den Servern verursachten Umweltaufwänden zu bilden (siehe Abbildung 6). RZ1 kommt auf die meisten SPECint_rate pro Megajoule, nämlich 35 000 (links). Ähnliche Abstände sind rechts für die potenzielle Klimaerwärmung ersichtlich. RZ1 kommt unter Freisetzung von einem kg CO_2-eq auf fast 600 000 SPECint_rate, RZ2 und RZ3 dagegen nur auf 300 000.

Mit dem KPI4DCE-Tool die Berechnungen vereinfachen

Auch wenn die Methode relativ einfach ist, ist das Ermitteln und Berechnen der Daten zeitaufwendig. Aus diesem Grund hat das UBA das Berechnungswerkzeug KPI4DCE-Tool entwickelt (siehe Abbildung 7). Es hilft beim Erfassen der spezifischen RZ-Daten durch standardisierte Eingabefelder und berechnet anhand der hinterlegten Umweltaufwände die Energie- und Ressourceneffizienz der jeweiligen Teilbereiche des Rechenzentrums.

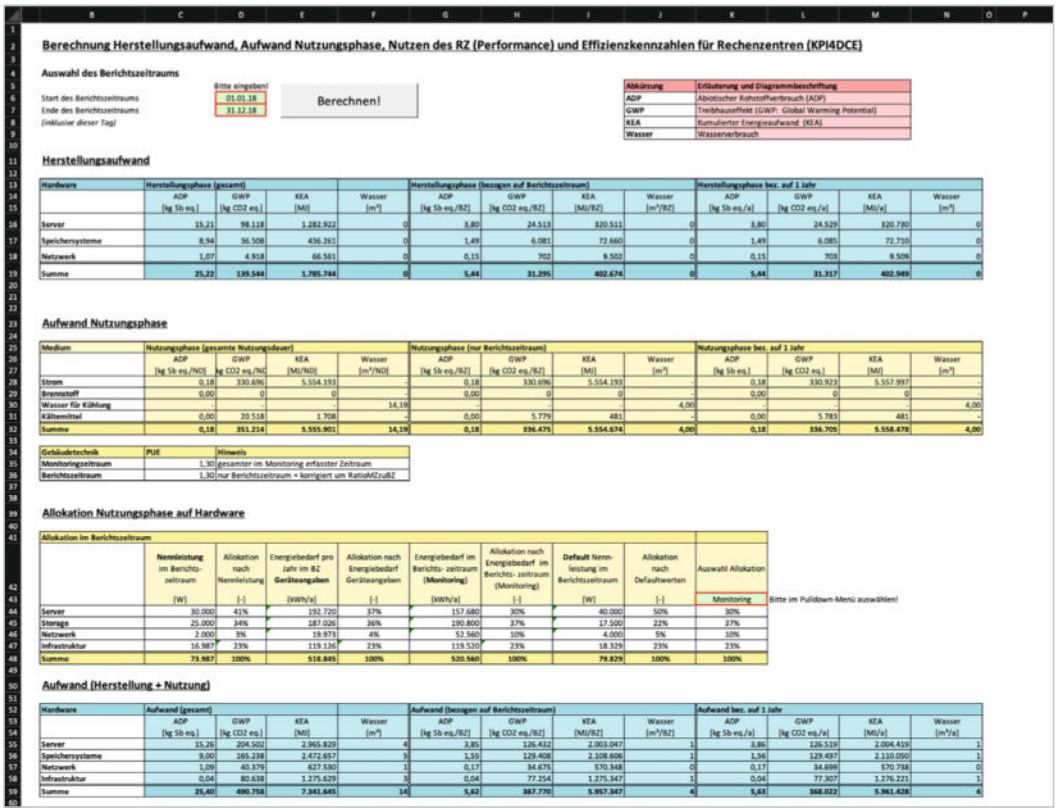


RZ1 kommt auf doppelt so viele SPECint_rate pro Megajoule wie RZ2 und RZ3. Ähnlich sieht es bei den SPECint_rate pro kg CO_2-eq aus (Abb. 6).

SPEC CPU2006

Die Benchmark-Suite SPEC CPU2006 enthält vier Benchmarksammlungen, die die CPU-Performance beim Ausführen von Ganzzahl- und Gleitkommaoperationen messen. Die beiden Sammlungen INTrate und FPrate simulieren einen Mehrbenutzerbetrieb mit parallelen Prozessen. INTspeed und FPspeed dagegen simulieren eine leistungshungrige Applikation, die ihre Arbeit auf viele Threads aufteilt. KPI4DCE verwendet die Ergebnisse von INTrate.

Die SPEC CPU2006 gibt keine mit anderen Benchmarks vergleichbaren Werte oder Rechenoperationen pro Sekunde wie der Linpack-Benchmark aus, sondern verwendet ihre eigenen Einheiten. Referenz ist eine Sun Ultra Enterprise 2 mit einem 296-MHz-UltraSPARC-II-Prozessor, getestet unter Solaris 10. Ein Ergebnis von 5 beim INTrate, also den Mehrbenutzertests mit Integer-Operationen, bedeutet eine fünffache Leistung gegenüber der Enterprise 2 oder die Rechenleistung von fünf solcher Sun-Maschinen. *Susanne Nolte*



Berechnung Herstellungsaufwand, Aufwand Nutzungsphase, Nutzen des RZ (Performance) und Effizienzkennzahlen für Rechenzentren (KPI4DCE)

Auswahl des Berichtszeitraums

Start des Berichtszeitraums: 01.01.18
Ende des Berichtszeitraums: 31.12.18
(inklusive dieser Tag)

Bitte eingeben!

Berechnen!

Abkürzung	Erläuterung und Diagrammbeschriftung
ADP	Abiotischer Rohstoffverbrauch (ADP)
GWP	Treibhauseffekt (GWP: Global Warming Potential)
KEA	Kumulierter Energieaufwand (KEA)
Wasser	Wasserverbrauch

Herstellungsaufwand

Hardware	Herstellungsphase (gesamt)				Herstellungsphase (bezogen auf Berichtszeitraum)				Herstellungsphase bez. auf 1 Jahr			
	ADP [kg Sb eq.]	GWP [kg CO2 eq.]	KEA [MJ]	Wasser [m³]	ADP [kg Sb eq./BZ]	GWP [kg CO2 eq./BZ]	KEA [MJ/BZ]	Wasser [m³/BZ]	ADP [kg Sb eq./a]	GWP [kg CO2 eq./a]	KEA [MJ/a]	Wasser [m³]
Server	15,21	98.118	1.282.922	0	3,80	24.513	320.511	0	3,80	24.529	320.730	0
Speichersysteme	8,94	36.508	436.261	0	1,49	6.081	72.660	0	1,49	6.085	72.710	0
Netzwerk	1,07	4.918	66.561	0	0,15	702	9.502	0	0,15	703	9.509	0
Summe	25,22	139.544	1.785.744	0	5,44	31.295	402.674	0	5,44	31.317	402.949	0

Aufwand Nutzungsphase

Medium	Nutzungsphase (gesamte Nutzungsdauer)				Nutzungsphase (nur Berichtszeitraum)				Nutzungsphase bez. auf 1 Jahr			
	ADP [kg Sb eq./ND]	GWP [kg CO2 eq./ND]	KEA [MJ/ND]	Wasser [m³/ND]	ADP [kg Sb eq./BZ]	GWP [kg CO2 eq./BZ]	KEA [MJ]	Wasser [m³]	ADP [kg Sb eq.]	GWP [kg CO2 eq.]	KEA [MJ]	Wasser [m³]
Strom	0,18	330.696	5.554.193	-	0,18	330.696	5.554.193	-	0,18	330.923	5.557.997	-
Brennstoff	0,00	0	0	-	0,00	0	0	-	0,00	0	0	-
Wasser für Kühlung	-	-	-	14,19	-	-	-	4,00	-	-	-	4,00
Kältemittel	0,00	20.518	1.708	-	0,00	5.779	481	-	0,00	5.783	481	-
Summe	0,18	351.214	5.555.901	14,19	0,18	336.475	5.554.674	4,00	0,18	336.705	5.558.478	4,00

Gebäudetechnik	PUE	Hinweis
Monitoringzeitraum	1,30	gesamter im Monitoring erfasster Zeitraum
Berichtszeitraum	1,30	nur Berichtszeitraum = korrigiert um RatioMZzuBZ

Allokation Nutzungsphase auf Hardware

Allokation im Berichtszeitraum

	Nennleistung im Berichtszeitraum [W]	Allokation nach Nennleistung [-]	Energiebedarf pro Jahr im BZ Geräteangaben [kWh/a]	Allokation nach Energiebedarf Geräteangaben [-]	Energiebedarf im Berichts- zeitraum (Monitoring) [kWh/a]	Allokation nach Energiebedarf im Berichts- zeitraum (Monitoring) [-]	Default Nenn- leistung im Berichtszeitraum [W]	Allokation nach Defaultwerten [-]	Auswahl Allokation
									Monitoring
Server	30.000	41%	192.720	37%	157.680	30%	40.000	50%	30%
Storage	25.000	34%	187.026	36%	190.800	37%	17.500	22%	37%
Netzwerk	2.000	3%	19.973	4%	52.560	10%	4.000	5%	10%
Infrastruktur	16.987	23%	119.126	23%	119.520	23%	18.329	23%	23%
Summe	73.987	100%	518.845	100%	520.560	100%	79.829	100%	100%

Bitte im Pulldown-Menü auswählen!

Aufwand (Herstellung + Nutzung)

Hardware	Aufwand (gesamt)				Aufwand (bezogen auf Berichtszeitraum)				Aufwand bez. auf 1 Jahr			
	ADP [kg Sb eq.]	GWP [kg CO2 eq.]	KEA [MJ]	Wasser [m³]	ADP [kg Sb eq./BZ]	GWP [kg CO2 eq./BZ]	KEA [MJ/BZ]	Wasser [m³/BZ]	ADP [kg Sb eq./a]	GWP [kg CO2 eq./a]	KEA [MJ/a]	Wasser [m³/a]
Server	15,26	204.502	2.965.829	4	3,85	126.432	2.003.047	1	3,86	126.519	2.004.419	1
Speichersysteme	9,00	169.238	2.472.697	3	1,39	129.408	2.108.606	1	1,56	129.497	2.110.050	1
Netzwerk	1,09	40.379	627.580	1	0,17	34.675	570.348	0	0,17	34.699	570.788	0
Infrastruktur	0,04	80.638	1.275.629	3	0,04	77.254	1.275.347	1	0,04	77.307	1.276.221	1
Summe	25,40	490.758	7.341.645	14	5,62	367.770	5.957.347	4	5,63	368.022	5.961.428	4

Quelle: UBA

Das Berechnungstool KPI4DCE hilft beim Erfassen und Berechnen der eigenen Daten (Abb. 7).

Die KPI4DCE-Kennzahlen machen insbesondere den Tradeoff zwischen Energie- und Materialeffizienz sichtbar. Zwar mag ein möglichst kurzer Erneuerungszyklus bei den Servern dem Energieaufwand beim Rechnen zugutekommen. Dem steht allerdings der energetische Aufwand vor allem bei der Herstellung der Hardware gegenüber.

Ein Rechenzentrumsbetreiber kann anhand der detaillierten Informationen auch erkennen, welche ungenutzten Ressourcen in seinem RZ schlummern. Dadurch können die Kennzahlen auch eine wertvolle Hilfe bei der Optimierung des Betriebs sein. Eine höhere Auslastung der Server beispielsweise verbessert das Verhältnis von Energieverbrauch und erbrachter Leistung. Die PUE kann diese Verbesserung der Ressourcennutzung nicht abbilden. Sie würde sich nur ändern, wenn sich durch den höheren Energieverbrauch des Servers der Energieverbrauch der Infrastruktur, also etwa der Kühlung und der unterbrechungsfreien Stromversorgung, nicht in gleicher Weise erhöhen würde.

Auch wenn die KPI4DCE-Kennzahlen die Energie- und Ressourceneffizienz des Rechenzentrums verlässlich wiedergeben, ist der politische Weg zu mehr Transparenz noch weit. Hierfür braucht es verpflichtende Vorgaben für Betreiber von Rechenzentren. Denn solange keine Vergleichswerte anderer Rechenzentren vorliegen, lassen sich die eigenen KPI4DCE-Werte nicht einordnen.

Ausblick: Register und Energiekennzeichnung für RZs

Da der Energieverbrauch deutscher Rechenzentren nirgends statistisch erfasst wird, existieren auch keine Daten darüber. Das macht es schwierig, Mindestanforderungen an Rechenzentren zu stellen und ihre Einhaltung zu überprüfen. 2021 hat deshalb das UBA ein Forschungsvorhaben mit dem Ziel gestartet, ein Register für Rechenzentren zu entwickeln und aufzubauen, das den jährlichen Energieverbrauch und weitere relevante Daten erfasst und veröffentlicht.

Das soll die nötige Transparenz für einen fairen Wettbewerb auch bei der Energieeffizienz der Rechenzentren schaffen. Zudem lassen sich mit einem solchen Register zukünftige Entwicklungen besser prognostizieren. Das kommt auch den Rechenzentren zugute. Als zweites Ziel des Vorhabens soll ein nationales Bewertungssystem für energieeffiziente Rechenzentren entstehen. Es soll einfach und verständlich sein und gleichzeitig die Unterschiede – etwa Geschäftsmodell, Alter oder regionale Gegebenheiten – berücksichtigen, um eine faire Bewertung sicherzustellen. Da derzeit die Aufnahme in das Rechenzentrumsregister freiwillig ist, ist das Umweltbundesamt auf die Mitwirkung der Verbände und der RZ-Betreiber angewiesen. Anderenfalls werden die Methoden und Werkzeuge ungenutzt bleiben und die Lenkungswirkungen nicht stattfinden.

Die Digitalstrategie der EU-Kommission sieht vor, dass die Rechenzentren in Europa bis zum Jahr 2030 klimaneutral werden sollen. Ein entsprechendes Maßnahmenpaket soll die Treibhausgasemissionen bis 2030 um mindestens 55 Prozent gegenüber dem Stand von 1990 senken. Energieeffizienzmaßnahmen spielen hierbei eine zentrale Rolle. Mit einem europäischen Register analog zum deutschen ließen sich die Maßnahmen und Ergebnisse überprüfen. Betreiber von Rechenzentren sollen verpflichtet werden, den Energieverbrauch und die Effizienz jährlich zu berichten. Denn nur, wenn der Istzustand bekannt ist, lassen sich wirkungsvolle Maßnahmen ergreifen, um das Klimaschutzziel zu erreichen. (sun@ix.de)

Quellen

KPI4DCE, SPEC-CPU-Ergebnisse und Umweltmanagementsysteme siehe ix.de/zxn5

Marina Köhn

ist Informatikerin und seit 1992 im Umweltbundesamt wissenschaftlich tätig. Mit dem Thema Green IT beschäftigt sie sich seit über 20 Jahren.

Open Compute Project: Energie- und Raumeffizienz zu niedrigsten Kosten?

Ein eigenes Ökotop

Hubert Sieverding

Mit dem 21"-Rack-Konzept will das Open Compute Project helfen, in Rechenzentren Energie, Platz und Kosten zu sparen. iX konnte zwei OCP-Server gegen zwei energieeffiziente Exemplare im klassischen Format antreten lassen.

Die Ursache für die Ineffizienz heutiger Rechenzentren ist 30 Jahre alt und misst 4,445 cm oder 1,75 Zoll. Dies ist die Höheneinheit (HE) eines 19-Zoll-Racks, also eines Stahlgestells, das Geräte mit einer Frontseitenbreite von 48,26 cm – umgerechnet 19 Zoll – aufnehmen kann. Zwar ist das 19″-Rack deutlich älter, doch erst 1992 einigte man sich auf besagte Höheneinheit. Mit Beginn des Internetzeitalters wurde der Platz in den Serverräumen enger und enger und die Server schrumpften in der Höhe auf diese Größe.

Ein Standard-Rack lässt sich mit 42 Servern dieser Bauart bestücken, in denen über 500 kleine Lüfter bis zu 84 CPUs Luft zufächeln und allein für diese Luftbewegung bereits bis zu 10 kW Leistung aufnehmen. Zwar findet eine derartige Vollauslastung niemals statt, doch sind die zudem redundanten Netzteile entsprechend groß dimensioniert, haben aber bei geringer Last einen schlechteren Wirkungsgrad als ohnehin schon. Ein bis zwei Kilowatt Verlustleistung führen allein die schlecht ausgelasteten PSUs (Power Supply Units) eines voll bestückten Racks an den Warmgang des Serverraums ab, der entsprechend Platz benötigt, um die Wärme über die Deckenlüftung abführen zu können – so nicht eine spezielle Warmgangeinhausung eingebaut ist (siehe Artikel „Nah dran“ ab Seite 134). Die Kaltluftzufuhr übernimmt ein Doppelboden ebenso wie die Kabellage, was wiederum besondere Anforderungen an die Gesamtarchitektur der Räumlichkeiten stellt.

Der Facebook-Konzern Meta muss all dies im Hinterkopf gehabt haben, als er 2011 einen radikal anderen Ansatz wählte: das 21″-Rack und eine dazu passende Serverhardware. Wichtigster Punkt des Lastenhefts: Eine x-beliebige Halle muss sich binnen kürzester Zeit zum Serverraum umrüsten lassen.

iX-TRACT

- Das Open Compute Project entwickelt offene Hardwarespezifikationen für Server, Racks und Rechenzentren.
- Das OCP will damit vor allem die Energie-, Raum- und Kosteneffizienz von Rechenzentren steigern.
- Zum Design von OCP-Racks und -Servern zählen eine zentrale Gleichstromversorgung, größere, dafür weniger Lüfter und ein erweiterter Betriebstemperaturbereich.
- Außerdem entfallen selten genutzte Bauteile und das Verschachteln geschlossener Gehäuse.

Die leere Schublade bietet Platz für drei Servereinschübe. Die erste Generation des Open Racks nutzt drei Gleichstromschienen, die zweite Generation verwendet eine Unterverteilung von der mittigen Stromschiene auf die drei Einschübe (Abb. 1).

Funktional: Einen ungewohnten Anblick bietet die Vorderseite eines Open Racks. Statt Zierblenden finden sich dort die Anschlusskabel und in der unteren Reihe die zentralen 3000-Watt-Netzteile. Die rötlichen Schaumstoffblöcke ganz unten verhindern einen thermischen Kurzschluss im nicht voll bestückten Rack (Abb. 2).

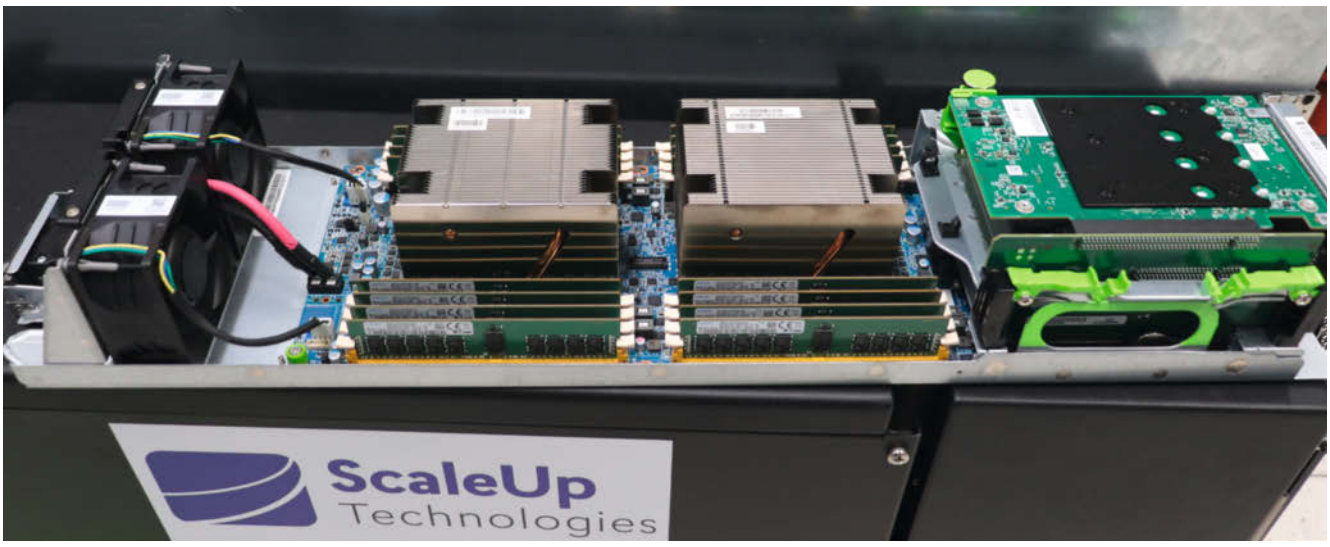
Herausgekommen ist das Konzept offener Hardware, das vor allem bei der Energie-, Raum- und Kosteneffizienz den etablierten Ansätzen voraus sein will. Denn die Spezifikationen des Open Compute Project (OCP) sind Open Source und damit allen zugänglich.

Ungewohnter Anblick

Das Open Rack ist wie sein 19″-Pendant 60 cm breit, jedoch 106 cm tief und 2,2 m hoch. Es steht auf Rollen, wird nur von vorne bestückt und verkabelt, da es auf der Rückseite eine zentrale Stromschiene trägt. Die erste Generation hatte noch derer drei, die kommende Generation nutzt 48 V statt der heutigen 12 V Gleichstrom. Die Strom- und Netzwerkkabel kommen von der Decke, ein Doppelboden ist nicht notwendig. Das Höhenraster beträgt 48 mm und ist damit 3,5 mm höher als die klassische HE. 41 OU (Open Units) stapeln sich im hohen Rack übereinander. Ein typischer Blechrahmen, also eine Schublade, die Servereinschübe trägt, nutzt 2 OU und ist 21 Zoll, also 538 mm breit.

Typischerweise trägt dieser Rahmen drei nebeneinanderliegende Serverschlitten im Format 2OU3N (siehe Abbildung 1). Teure Teleskopschienen sucht man vergebens. Der gewonnene Platz kommt voll dem Equipment zugute. Über Adapter lassen sich auch 19″-Komponenten, insbesondere Switches und Router, einbauen. Den Strom liefern 3000-Watt-Netzteile, je zu sechst in zwei zentralen Netzteileinschüben angeordnet. Beim Einsatz außerhalb abgesicherter Rechenzentren finden darin auch Lithium-Ionen-Akkus für die Notstromversorgung Platz. Ein voll bestücktes Open Rack trägt also 48 2OU3N-Server, 12 Netzteile/Akkueinheiten und drei Switches.

In den in Deutschland üblichen klimatisierten Serverräumen wirkt das Open Rack wie ein Fremdkörper. ScaleUp Technologies aus Hamburg erwarb vor einem Jahr zwei Racks mit runderneuerten Facebook-Servern. Die komplett bestückten schweren Racks passen gerade so durch Tor und Fahrstuhl und überraschen mit ihrem puristischen Outfit: Freie Einschübe hatten die Amerikaner mit Pizzakartons vor thermischem Kurzschluss geschützt und Seitenwände waren selbst von OCP-Rack-Lieferant Rittal nicht zu beschaffen. Eine benachbarte Metallwerkstatt maßschneiderte sie kostengünstig und ScaleUp ersetzte die Pappeinschübe durch feuerhemmenden Schaumstoff (siehe Abbildung 2). Die Stromzufuhr aus dem Doppelboden stellte die nächste Herausforderung dar, denn die mitgelie-



Der Facebook-Server der ersten Generation Wiwynn Tioga Pass ist bestückt mit zwei Intel-E3-CPUs. Auf der Vorderseite (rechts) stecken Erweiterungskarten und Festplatten. Hinten (links) die beiden großen Lüfter und der Anschluss zur Stromversorgung. Die grüne Farbe zeigt, wo der Admin Hand anlegen darf (Abb. 3).

Auf der Rückseite des Tioga Pass befindet sich mittig der Anschluss zur zentralen Gleichstromversorgung, dahinter die beiden 80-mm-Lüfter, die jeweils maximal 14,4 Watt ziehen (Abb. 4).

ferten Stromkabel waren dafür einen Meter zu kurz und mussten verlängert werden.

Sparsamkeit als Designziel

Die Energieeffizienz des 21″-OCP-Konzepts fußt auf vier Säulen: der zentralen Gleichstromversorgung, größeren, dafür weniger Lüftern, einem erweiterten Betriebstemperaturbereich und dem Entfallen selten genutzter und überflüssiger Bauteile wie geschlossener Rahmen. Die zentrale Gleichstromversorgung erhöht die Wahrscheinlichkeit, dass die Netzteile in einem Fenster hohen Wirkungsgrads arbeiten. Die OCP-Spezifikation schreibt in einem Bereich von 20 bis 90 Prozent Auslastung einen Wirkungsgrad von 94 Prozent vor. Dies ist deutlich strenger als die Platinum-Vorgabe der 80-plus-Initiative, die einen Wirkungsgrad von mindestens 89 Prozent verlangt.

Die Höhe der Servereinschübe von fast 10 cm erlaubt den Einbau großer Lüfter. Dies reduziert den zur Luftbewegung notwendigen Energiebedarf gegenüber 1-HE-Servern deutlich. Erstens sind 80-mm-Lüfter deutlich effizienter und zweitens reichen zwei davon für einen Servereinschub. Statt der maximal 42 1U-Server mit 500 kleinen Lüftern werkeln im voll bestückten Open Rack 48 Server mit insgesamt 96 Lüftern, die bei je 14,4 Watt zusammen maximal 1400 Watt Leistung aufnehmen. Die auf der Rückseite platzierten Lüfter saugen Kaltluft durch ein kühlungsoptimiertes schlankes Servergehäuse.

Sofern überhaupt eine Einhausung vorgesehen ist, muss der Warmgang nicht betreten werden und kann entsprechend zweckmäßig ausfallen. Dank der erlaubten Einlasstemperatur von bis zu 35 °C ist das Konzept prädestiniert für eine freie Kühlung. An klimatisch geeigneten Standorten reicht es, Racks in eine Halle zu stellen und für ausreichend Außenluft zu sorgen. Überall sonst verschiebt sich der Klimatisierungsschwellenwert deutlich. Faktisch ist eine Kälteanlage nur an wenigen heißen Sommertagen notwendig.

Das Konzept gemeinsamer Infrastruktur für mehrere Server ist nicht neu. Fast alle Serveranbieter stellen Blade-Systeme oder modulare Systeme her, bei denen sich viele Einschübe Backplane, Netzteile und Lüfter teilen. Dem OCP-Ansatz am nächsten kommt Supermicro mit seinen Twin-Servern, bestehend aus zwei oder vier Server-Nodes und gemeinsamen, zentral angeordneten Netzteilen. Das Bedienkonzept bleibt jedoch klassisch: Festplatten vorne, Kabelanschlüsse hinten.

iX will herausfinden, was die Versprechen der OCP-Anbieter wert sind. Dankenswerterweise stellte Circle B aus Amsterdam zwei OCP-Server zum Remote-Test zur Verfügung. Bei dem ersten handelt es sich um einen SV7220 von Wiwynn. Der Server mit Codenamen Tioga Pass ist ausgestattet mit zwei Xeon-6230-CPUs von Intel, gehört also zur zweiten Generation der Facebook-Server. Die erste verwendete Xeon-E3-Prozessoren (siehe Abbildung 3 und 4). Neu auf dem Markt ist die Maschine E8020 von Mitac mit Codenamen Capri und AMDs Epyc 7702P. Auch dieses Modell wartet bei Circle B auf Nutzer und ist Gegenstand der Untersuchung.

Antreten gegen energieeffiziente Konkurrenz

Gegenhalten wollen zwei 19″-Systeme, die für ihre Energieeffizienz bekannt sind, nämlich die Lenovo SR655 mit AMD Epyc 7702, bereitgestellt von Lenovo in Stuttgart, und Gigabytes R152-P30 mit Ampere Altra Q80, die iX Ende letzten Jahres testete [1]. Die Tabelle „Gemessene Systeme im Vergleich“ stellt die Systeme und ihre Ausstattung gegenüber: links zwei besonders effiziente Rackserver im 19″-Format, rechts die beiden Herausforderer aus dem OCP-Lager.

Mit SPECpower steht ein anerkannter Benchmark zum Messen der Energieeffizienz von Servern bereit. Jedoch verlangt dieses Werkzeug einen externen Strommesssensor, den die Software periodisch über USB oder die serielle Schnittstelle abfragt. Dieses Verfahren gewährleistet nachvollziehbare und unverfälschbare Ergebnisse, ist jedoch in der Cloud und bei OCP-Servern mit zentraler Stromversorgung nicht einsetzbar. Es bleiben also nur Werkzeuge übrig, die sich auch für eine Remote-Messung eignen.

Standard bei CPU-Messungen ist seit Jahren die SPEC CPU2017. Die Benchmark-Suite teilt sich in Messungen der Integer- und Floating-Point-Leistung sowie in RATE und SPEED auf. Die Rate-Tests simulieren einen Mehrbenutzerbetrieb, indem sie Prozesse parallel starten; die Speed-Tests simulieren eine leistungshungrige Applikation, die ihre Arbeit auf viele Threads aufteilt. Das I/O-Verhalten spielt bei der Suite aufgrund der langen Laufzeit der Einzeltests keine Rolle.

INTrate setzt sich aus 10 Einzeltests zusammen, die teils auf vorhandene Applikationen und Bibliotheken wie X264 zurück-

Listing: Ausgabe von ipmi-sensors

```
Lenovo SR655:# ipmi-sensors -t "Other_Units_Based_Sensor"
ID  | Name        | Type                     | Reading | Units | Event
73  | PSU1_PIN    | Other Units Based Sensor | 192.00  | W     | 'OK'
74  | PSU1_POUT   | Other Units Based Sensor | 180.00  | W     | 'OK'
78  | PSU2_PIN    | Other Units Based Sensor | 192.00  | W     | 'OK'
79  | PSU2_POUT   | Other Units Based Sensor | 188.00  | W     | 'OK'
83  | CPU_Power   | Other Units Based Sensor | 185.00  | W     | 'OK'
84  | MEM_Power   | Other Units Based Sensor | 77.00   | W     | 'OK'
85  | Total_Power | Other Units Based Sensor | 390.00  | W     | 'OK'

Lenovo SR655:# turbostat -S -q -n 1 -s 'Busy%,PkgWatt'
Busy%   PkgWatt
75.06   188.59

Mitac Capri:# ipmi-sensors -t "Other_Units_Based_Sensor"
ID  | Name            | Type                     | Reading | Units | Event
9   | CPU VR PIN      | Other Units Based Sensor | 86.00   | W     | 'OK'
10  | HSC Input Power | Other Units Based Sensor | 234.00  | W     | 'OK'
13  | CPU Pkg Power   | Other Units Based Sensor | 165.00  | W     | 'OK'
14  | CPU VR POUT     | Other Units Based Sensor | 72.00   | W     | 'OK'
17  | DIMM VR0 POUT   | Other Units Based Sensor | 38.88   | W     | 'OK'
20  | DIMM VR1 POUT   | Other Units Based Sensor | 37.80   | W     | 'OK'
```

greifen. Ihr mehrfacher Durchlauf dauert auf der Referenzmaschine, einer Sun Fire V490, etwas mehr als 4 Stunden. FPrate nutzt 13 Einzeltests, darunter die bekannte Applikation blender. Der Referenzlauf auf der Sun dauert knapp 8 Stunden. Die Speed-Tests spielen hier keine Rolle. Die SPEC-Leistungswerte aller Maschinen wurden nach dem iX-Standard unter Ubuntu 20.04 mit der GNU Compiler Collection und der Option -O3 gemessen, nur die Gigabyte-Maschine unter Ubuntu 21.04.

Wege, die Leistungsaufnahme zu messen

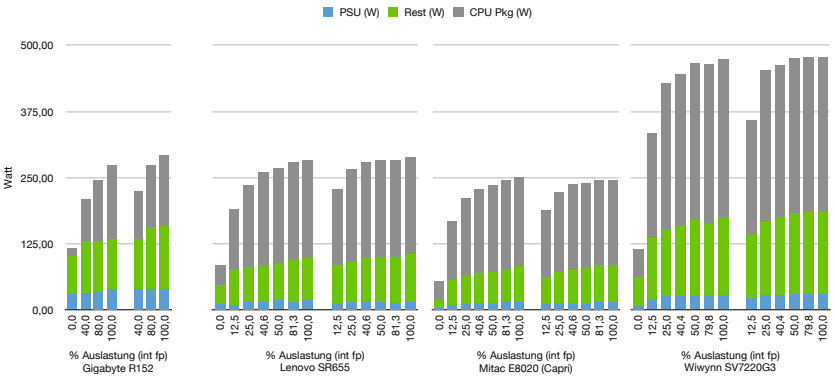
Für Messungen ohne externes Leistungsmessgerät bietet sich der BMC (Baseboard Management Controller) an. Mit den entsprechenden Privilegien kann man mit dem ipmitool darauf zugreifen. Dessen Kommando `ipmi-sensors` liefert eine Liste von Messwerten, die sich je nach BMC, CPU, BIOS und Protokollversion unterscheiden. Das Listing zeigt die Ausgaben des Befehls auf der Lenovo SR655 und der Mitac Capri. Bei letzterer fehlen die PSU-Sensoren, da das System keine eigenen Netzteile besitzt. Server für 21″-Racks nutzen einen HSC (Hot Swap Controller), um die Versorgung des Boards durch die zentrale Gleichstromversorgung zu überwachen. Dieser Sensor gibt auch Auskunft über den Stromverbrauch des Servereinschubs. HSC Input Power ist also mit den PSUx_POUT-Werten klassischer Server vergleichbar.

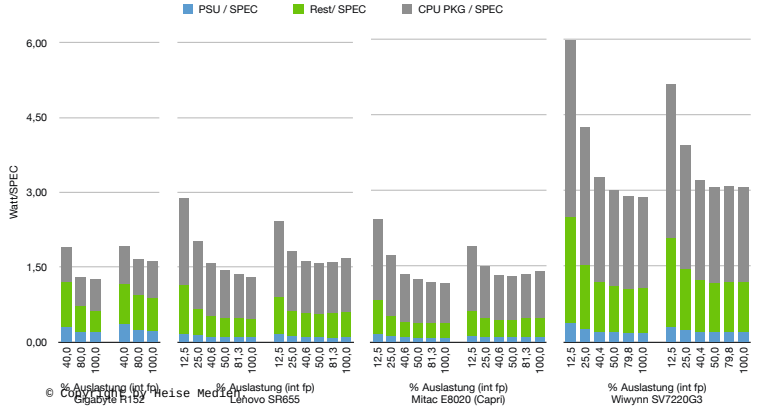
Zudem greift der BMC auch über das RAPL-Interface (Running Average Power Limit) auf einen Satz von Spezialregistern der CPU zurück. Intel führte RAPL 2011 zusammen mit der Sandy-Bridge-Architektur ein. Auch AMD- und ARM-Prozessoren neuerer Generation beherrschen diese Schnittstelle, allerdings ist die Dokumentation in allen Fällen bescheiden. Mit dem Linux-Kommando `perf` gibt RAPL Auskunft über die Auslastung, C-States und Energieaufnahme jedes Kerns. Das Kommando `turbostat -S -q -n 1` liefert die akkumulierten Werte und zusätzlich den augenblicklichen, leicht gemittelten Ge-

Auf der Gigabyte R152-P30 wurde die Energiemessung während der Läufe der SPEC CPU2017 händisch per Stichprobe durch Einsicht der BMC-Konsole erhoben, auf den anderen Maschinen konnte sie automatisiert alle 150 Sekunden durchgeführt werden. Anschließend wurden die Werte bereinigt und gemittelt. Der Wirkungsgrad der OCP-Netzteile ist einheitlich mit 94 Prozent angesetzt. CPU_Pkg umfasst je nach Hersteller die Gesamtenergieaufnahme von CPU einschließlich DRAM- und PCIe-Ansteuerung (Abb. 5).

Messergebnisse

%	CPU2017	SPECrate	Busy%	Total Power (W)	PSU (W)	Rest (W)	CPU Pkg (W)	Total / SPEC	PSU / SPEC	Rest/ SPEC	CPU PKG / SPEC
	Gigabyte R152										
0,0			0,10	**116,00**	32,00	70,00	14,00				
40,0	intrate	109,00	40,00	**208,00**	32,00	98,00	78,00	**1,91**	0,29	0,90	**0,72**
80,0	intrate	186,00	80,00	**244,00**	36,00	94,00	114,00	**1,31**	0,19	0,51	**0,61**
100,0	intrate	217,00	100,00	**272,00**	40,00	94,00	138,00	**1,25**	0,18	0,43	**0,64**
40,0	fprate	117,00	40,00	**224,00**	40,00	94,00	90,00	**1,91**	0,34	0,80	**0,77**
80,0	fprate	166,00	80,00	**272,00**	40,00	116,00	116,00	**1,64**	0,24	0,70	**0,70**
100,0	fprate	181,00	99,00	**292,00**	40,00	118,00	134,00	**1,61**	0,22	0,65	**0,74**
	Lenovo SR655										
0,0			0,02	**84,00**	12,00	33,03	38,97				
12,5	intrate	66,00	12,51	**189,75**	10,42	64,52	114,82	**2,88**	0,16	0,98	**1,74**
25,0	intrate	118,00	25,01	**236,31**	14,77	64,59	156,95	**2,00**	0,13	0,55	**1,33**
40,6	intrate	165,00	40,62	**259,80**	15,80	68,41	175,59	**1,57**	0,10	0,41	**1,06**
50,0	intrate	188,00	50,04	**269,42**	16,77	72,01	180,64	**1,43**	0,09	0,38	**0,96**
81,3	intrate	205,00	81,14	**279,00**	15,80	78,91	184,29	**1,36**	0,08	0,38	**0,90**
100,0	intrate	217,00	99,83	**283,14**	16,93	80,40	185,81	**1,30**	0,08	0,37	**0,86**
12,5	fprate	95,00	12,52	**228,20**	13,53	71,23	143,43	**2,40**	0,14	0,75	**1,51**
25,0	fprate	146,00	25,07	**266,00**	15,14	76,22	174,64	**1,82**	0,10	0,52	**1,20**
40,6	fprate	171,00	40,70	**278,69**	16,21	81,26	181,23	**1,63**	0,09	0,48	**1,06**
50,0	fprate	179,00	50,11	**282,39**	16,00	84,10	182,29	**1,58**	0,09	0,47	**1,02**
81,3	fprate	176,00	80,87	**283,54**	12,53	87,94	183,07	**1,61**	0,07	0,50	**1,04**
100,0	fprate	173,00	99,94	**288,37**	16,47	88,63	183,27	**1,67**	0,10	0,51	**1,06**
	Mitac E8020 Capri										
0,0			0,00	**53,94**	3,24	14,70	36,00				
12,5	intrate	69,30	12,45	**167,88**	10,07	46,59	111,22	**2,42**	0,15	0,67	**1,60**
25,0	intrate	123,00	24,76	**209,55**	12,57	51,02	145,96	**1,70**	0,10	0,41	**1,19**
40,6	intrate	169,00	40,51	**227,73**	13,66	55,47	158,60	**1,35**	0,08	0,33	**0,94**
50,0	intrate	189,00	49,81	**233,72**	14,02	57,26	162,44	**1,24**	0,07	0,30	**0,86**
81,3	intrate	207,00	80,53	**244,92**	14,70	63,82	166,41	**1,18**	0,07	0,31	**0,80**
100,0	intrate	218,00	99,52	**251,70**	15,10	66,93	169,67	**1,15**	0,07	0,31	**0,78**
12,5	fprate	98,90	12,48	**187,59**	11,26	50,27	126,07	**1,90**	0,11	0,51	**1,27**
25,0	fprate	150,00	24,71	**223,26**	13,40	57,89	151,97	**1,49**	0,09	0,39	**1,01**
40,6	fprate	177,00	40,29	**236,30**	14,18	62,36	159,76	**1,34**	0,08	0,35	**0,90**
50,0	fprate	184,00	49,60	**238,95**	14,34	63,83	160,78	**1,30**	0,08	0,35	**0,87**
81,3	fprate	180,00	80,17	**244,63**	14,68	69,13	160,83	**1,36**	0,08	0,38	**0,89**
100,0	fprate	177,00	98,27	**244,86**	14,69	69,12	161,05	**1,38**	0,08	0,39	**0,91**
	Wiwynn SV7220										
0,0			0,10	**113,22**	6,79	54,43	52,00				
12,5	intrate	55,00	12,50	**333,66**	20,02	115,69	197,96	**6,07**	0,36	2,10	**3,60**
25,0	intrate	101,00	24,76	**429,56**	25,77	126,87	276,92	**4,25**	0,26	1,26	**2,74**
40,4	intrate	137,00	40,28	**446,34**	26,78	133,87	285,69	**3,26**	0,20	0,98	**2,09**
50,0	intrate	155,00	49,89	**466,26**	27,98	141,34	296,94	**3,01**	0,18	0,91	**1,92**
79,8	intrate	160,00	79,21	**463,04**	27,78	137,10	298,16	**2,89**	0,17	0,86	**1,86**
100,0	intrate	165,00	99,19	**473,14**	28,39	146,82	297,94	**2,87**	0,17	0,89	**1,81**
12,5	fprate	70,30	12,50	**358,94**	21,54	121,98	215,42	**5,11**	0,31	1,74	**3,06**
25,0	fprate	116,00	24,89	**451,93**	27,12	138,11	286,71	**3,90**	0,23	1,19	**2,47**
40,4	fprate	144,00	40,00	**461,15**	27,67	149,39	284,09	**3,20**	0,19	1,04	**1,97**
50,0	fprate	156,00	49,48	**475,35**	28,52	152,98	293,85	**3,05**	0,18	0,98	**1,88**
79,8	fprate	155,00	78,88	**477,89**	28,67	155,86	293,36	**3,08**	0,18	1,01	**1,89**
100,0	fprate	157,00	98,20	**478,31**	28,70	155,45	294,16	**3,05**	0,18	0,99	**1,87**





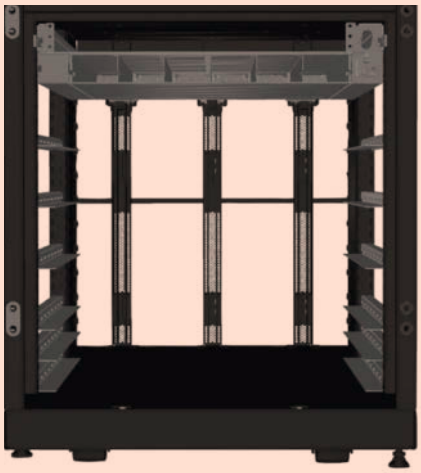
OCP-Anbieter im Überblick

Mit dem Einstieg von **Gigabyte** in den Markt der 21"-OCP-Hardware erhöht sich das Angebot offener Hardware gleich um eine ganze Palette neuer Produkte. Gaben bisher vor allem Facebook und das Open Compute Project die Spezifikationen vor, finden sich inzwischen Produkte fürs Open Rack, die nicht im OCP-Marketplace verzeichnet sind.

Gigabyte bietet zwei 21"-Racks der Version 1 mit drei Stromschienen in 41 OU und 12 OU Höhe mit zwei Netzteilebenen einfacher OU-Höhe an, dazu passend neue Doppel-CPU-Einschub-Server im 2OU3N-Format, wahlweise mit AMDs Epyc- oder Intels Xeon-CPU (siehe Abbildung 6). Es sind die ersten Epyc-Dual-Sockel-Server in diesem Gehäuseformat. Die Intel-Server verfügen serienmäßig über eine 10GE-NIC.

Ungewöhnlich ist der Rückfall auf das 1-OU-Format für die GPU-Server der T181-Baureihe. Zwar bietet das 537 mm breite Chassis Platz für vier GPUs doppelter Höhe, doch quälten sich hier zwölf kleine Lüfter nach bekanntem Muster, die in reichlicher Menge anfallende Wärme abzuführen. Auch hier entscheidet der Kunde zwischen den CPUs von AMD oder Intel. Dem bekannten 2OU3N-Muster folgt der Storage-Node T280. Drei Einschübe bieten jeweils Platz für fünfzehn 3,5"- oder 2,5"-Disks. Das Monitoring übernimmt eine separater BMC.

Pionier **Wiwynn** geht mit seinem SV7221G2-P in die Höhe und definiert das 4OU3N-Format, also einen Server mit vierfacher Höhe und einem Drittel Breite. Dadurch soll das System mehr Platz für Disks und PCIe-Karten bieten. Ergänzt wird das 4-OU-Angebot um Storage-Server mit bis zu 72 3,5"-Einschüben und dem Blade-Server Yosemite V3, bestückt mit zwölf Schlitten im 1OU3N-Format. Jedes Blade fasst eine Intel-CPU mit einer maximalen TDP von 165 Watt und hat Platz für minimale Erweiterungen wie eine M.2- und eine PCIe-Karte. Ein JBOF-System (Just a Bunch of Flash) im 2-OU-Format mit 30 NVMe-SSDs kann über NVMe-Expander beliebige OCP-Server bedienen.



OCP-Neuling Gigabyte bietet neben OCP-Servern und großen 41-OU-Racks auch ein kleines Rack mit 12 OU, quasi ein Einstiegsmodell. Mit seinen drei Stromschienen ist es auch mit älteren OCP-Servern kompatibel. Unten ist auf zwei Ebenen Platz für die Netzteile zur zentralen Stromversorgung und für einen Satz Akkus. Darüber finden vier 2-OU-Schubladen mit je drei 2OU3N-Servern Platz. Ganz oben findet ein 19"-Switch Platz in einem Adapterrahmen (Abb. 6).

Auch Eigentümer klassischer 19"-Racks können OCP-Hardware einsetzen, ohne alles umbauen zu müssen. Mitac bietet dazu einen Adapter an, der ins 19"-Rack geschraubt wird und wahlweise 8 oder 16 OU in der Höhe umfasst. Der Adapter verwendet eine zentrale Stromschiene gemäß Open-Rack-V2-Spezifikation. Aufgrund der geänderten Bauform passen noch zwei Servereinschübe nebeneinander in eine Schublade (Abb. 7).

samtenergiebedarf des Packages, bezeichnet als PkgWatt (siehe Listing).

Dieser Begriff ist jedoch nicht eindeutig definiert. Bei Intel umfasst er alle Elemente des Chips, jedoch nicht immer das DRAM. Bei AMDs Epyc wird auch die Spannung des DRAM vom integrierten Voltage Regulator vorgegeben. Dank dieses Tricks lässt sich der Energiebedarf der Speicherriegel auslastungsabhängig regeln und bei Bedarf auf ein Minimum senken. Deshalb umfasst das Package auch einen Teil des Speichers. Ob der Prozessor unter PkgWatt die gesamte Leistungsaufnahme der Speicherriegel inkludiert, ließ sich aber nicht klären, auch die zugehörigen Unterlagen halten sich diesbezüglich bedenkt. Fest steht jedoch, dass PkgWatt eine charakteristische Größe für den Hunger des Prozessors darstellt. Auf die separate Darstellung des DRAM-Energieverbrauchs wurde für die weiteren Auswertungen verzichtet.

Voraussetzung für `turbostat` ist ein Linux-Kernel der Version 5.13 oder höher, sofern ein AMD Epyc den Server antreibt. Diese Version stand auf dem OCP-Server von Mitac nicht zur Verfügung. Deshalb wurde die Auslastung stattdessen mit dem Befehl `top -b -n 1` gemessen und mit dem ipmitool die Pkg-Power ausgegeben. Sie entspricht PkgWatt. Die Energie wurde im 150-Sekunden-Zyklus während des SPEC-Benchmarks gemessen und später um auslastungsarme Zeiten bereinigt und gemittelt.

Abbildung 5 stellt die INTrate- und FPrate-Ergebnisse der Gesamtenergieaufnahme gegenüber und schlüsselt den Verbrauch in Relation zur SPECrate für verschiedene Auslastungsgrade der Maschinen auf. Die Auslastung ließ sich mit dem Starten zusätzlicher Benchmarkprozesse erhöhen. 40,6 Prozent Auslastung der AMD-CPU bedeuten, dass 52 der insgesamt 128 virtuellen Kerne beschäftigt werden. Für die beiden OCP-Server wurde der Wirkungsgrad der Netzteile auf Basis des PSU-Monitors des Racks einheitlich auf 94 Prozent festgelegt und die Total Power ausgehend von HSC Input Power hochgerechnet.

Mal fressen die CPUs, mal die Netzteile

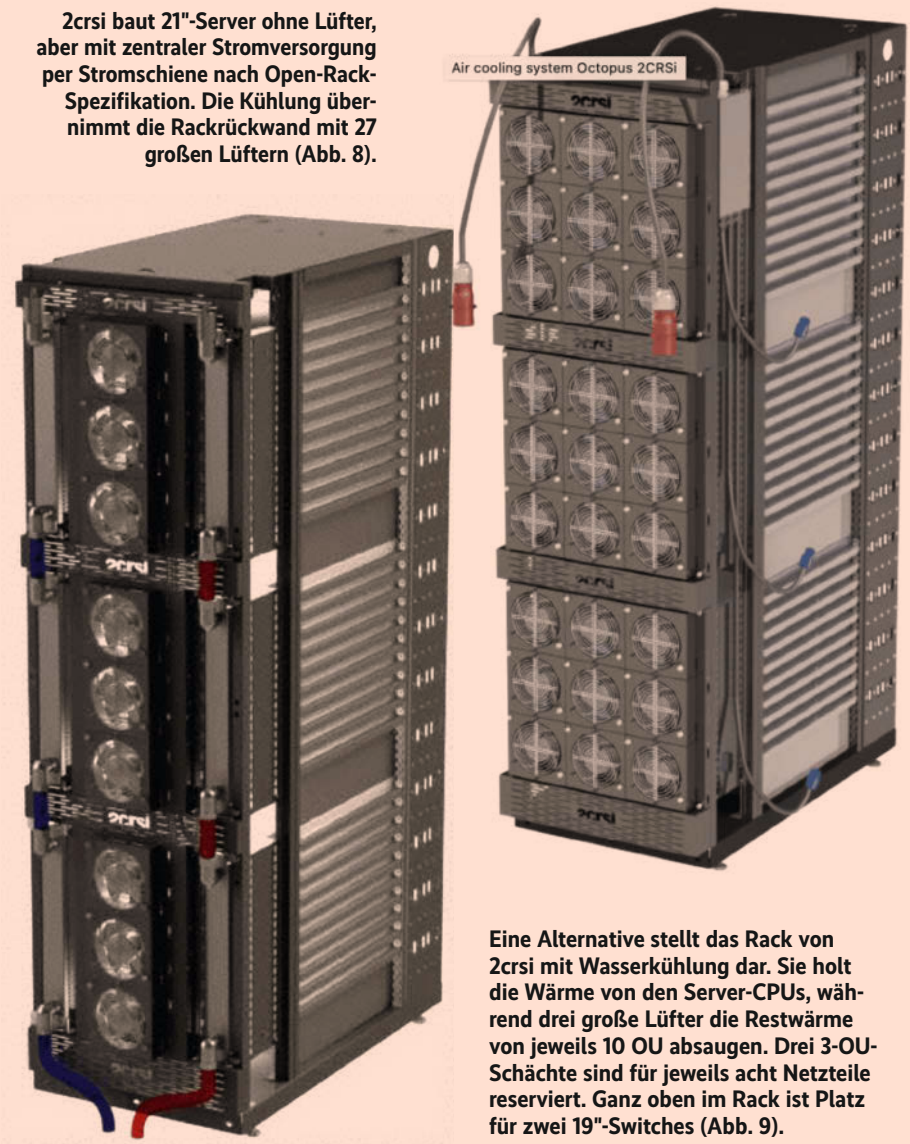
Ein Blick auf den absoluten Verbrauch der Maschinen zeigt: Unter Last determiniert die Energieaufnahme des Prozessors den Gesamtverbrauch. Die ARM-CPU beherrscht kein Multithreading und passt ihre Leistungsaufnahme sowohl beim Integer- als auch beim FP-Test der steigenden Last an. Die beiden anderen CPU-Typen greifen bereits ab 25 Prozent Auslastung in

Auch **Inspur** gehört zu den Facebook-Ausrüstern der ersten Stunde. Neben der Tioga-Pass-Implementierung ON5293M5 finden sich im Katalog ein Just-a-Bunch-of-GPU-Server ON5388M5 in dreifacher OCP-Höhe, der per NVMe Anschluss an die Server findet. Sogar 16 GPUs fasst der ON5488M5 in vierfacher Höhe nach dem gleichen Prinzip.

Das Open Rack bietet über Adapter auch 19"-Komponenten eine Heimat. **Mitac** geht den umgekehrten Weg und baut einen Adapter fürs 19"-Rack, genannt ESA (siehe Abbildung 7). Dorthinein passen nebeneinander zwei statt drei OCP-Einschübe. Eine zentrale Stromschiene liefert den Strom. Neben den getesteten Capri-Servern mit Epyc-CPU bietet Mitac auch die Xeon-Dual-Sockel-Server der Facebook-Tioga-Pass-Baureihe. Ebenfalls im Programm ist ein JBOF-Einschub für gedrittelte 2-OU-Rahmen. Darin finden sechzehn U.2-SSDs Platz, überwacht von einem eigenen BMC ASPEED 2520.

2crsi bestückt sein Open Rack mit Servern OCtoPus 3E im 1OU3N-Format und verzichtet dabei auf integrierte Lüfter. Siebenundzwanzig 16 cm große Lüfter nehmen die gesamte Rückwand des Racks ein und ziehen die Luft durch jede Ritze (siehe Abbildung 8). Alternativ unterstützen neun große Lüfter die Wasserkühlung (siehe Abbildung 9) – wahrlich eine Neuinterpretation des Open Racks. Auch belebt 2crsi den Server einfacher Höhe neu. Auf dem Blech des OCtoPus 5 bringt der Pariser Hersteller fünf CPU-Module fürs HPC unter. OCtoPus 1.x platziert bis zu acht GPUs links und rechts einer zentralen Epyc-Dual-Sockel-Einheit in einem 2-OU-Blech. Allen gemeinsam ist die zentrale Gleichstromversorgung.

2crsi baut 21"-Server ohne Lüfter, aber mit zentraler Stromversorgung per Stromschiene nach Open-Rack-Spezifikation. Die Kühlung übernimmt die Rackrückwand mit 27 großen Lüftern (Abb. 8).



Eine Alternative stellt das Rack von 2crsi mit Wasserkühlung dar. Sie holt die Wärme von den Server-CPUs, während drei große Lüfter die Restwärme von jeweils 10 OU absaugen. Drei 3-OU-Schächte sind für jeweils acht Netzteile reserviert. Ganz oben im Rack ist Platz für zwei 19"-Switches (Abb. 9).

die Vollen, auch deutlich erkennbar an der Energieaufnahme der FP-Einheiten.

Insgesamt beeindruckt das Potenzial des OCP-Servers von Mitac mit einem Gesamtenergiebedarf von unter 250 Watt. Im direkten Vergleich mit der Lenovo fällt auf, dass die gleiche AMD Epyc 7702 in der Mitac Capri eine niedrigere Energieaufnahme bei leicht höherer SPEC-Leistung zeigt. Ursache könnte sein, dass die SR655 mit doppelt so viel, aber langsamerem DDR4-Speicher bestückt ist. Die Lenovo SR655 verfügt über sehr effiziente Netzteile, deren Wirkungsgrad die 80-plus-Spezifikation Platinum erfüllt. Für den Test waren sie auf 1+1 konfiguriert, das heißt, das zweite Netzteil steht auf Stand-by und erhöht den Wirkungsgrad des ersten.

Die Vorteile ihrer ARM-CPU büßt die Gigabyte-Maschine durch ihre Netzteile und den Nachteil des 1U-Gehäuses bei der Kühlung wieder ein. Der getestete Server R152-P30 ist mit zwei 650-Watt-80-plus-Platinum-Netzteilen bestückt, die jedoch 2+0 statt 1+1 konfiguriert sind. Der Hersteller 3Y Power garantiert für das Netzteil einen Wirkungsgrad von 94 Prozent bei 50 Prozent Auslastung. Bei 10 Prozent Last sind es gerade noch 86,5 Prozent. Die verwendete Hardware benötigte nur circa 250 Watt und lastet beide Netzteile lediglich zu ineffizienten 18 Prozent aus. Auf Anfrage erklärte Gigabyte, dass die 80-plus-Wirkungsgrade zudem nicht die benötigte Lüfterleistung der Netzteile umfassen, womit sie dem zweiten OCP-Argument – sparsamere Lüfter – Vorschub leisten.

Energiefressende Lüfter

Der Energiekonsum der Lüfter geht wie der Strombedarf der sonstigen Komponenten im Anteil „Rest“ auf. Hier spielt die Mitac Capri die Vorteile des schlanken Chassis-Designs mit nur einer großen Wärmequelle und zwei großen 8-cm-Lüftern optimal aus. Ein thermisch optimierter, doppelt hoher Standardserver könnte wahrscheinlich mit der Mitac durchaus mithalten.

Die Lenovo kann es in der getesteten Konfiguration nicht. Das Testgerät ist mit sechs Hochleistungslüftern in der Größe 60 × 60 × 38 mm³ mit einer Leistungsaufnahme von jeweils 24 Watt bestückt, deren maximale Drehzahl 23 900 UpM beträgt. Zudem ist die Lüftersteuerung der Maschine recht statisch und richtet sich an der Bestückung bestimmter Komponenten aus. Ist beispielsweise die Bedingung „100-GE-Adapter installiert“ erfüllt, drehen die Lüfter zu mindestens 55 Prozent, und zwar unabhängig von der Einlasstemperatur. In Anbetracht dessen fällt der Mehrbedarf der SR655 in Höhe von 15 Watt beim Anteil „Rest“ sogar be-

Gemessene Systeme im Vergleich

	Gigabyte R152-P30 (ARM)	Lenovo ThinkSystem SR655	Mitac E8020 Capri	Wiwynn SV7220 Tioga Pass
Typ	ARM-Server im Standardformat	x64-Server im Standard-format	OCP-x64-Server	OCP-x64-Server
Formfaktor	19 Zoll, 1U	19 Zoll, 2U	21 Zoll, 2OU 1/3	21 Zoll, 2OU 1/3
CPU	Ampere Altra Q80-30 80C@2 GHz	AMD Epyc 7702 64C@2 GHz	AMD Epyc 7702P 64C@2 GHz	2 × Intel Xeon Gold 6230R; 2 × 26C@2,10 GHz
DRAM	16 × 16 GByte DDR4@3200 MHz	8 × 32 GByte DDR4@2933 MHz	8 × 16 GByte DDR4@3200 MHz	12 × 16GByte DDR4@2400 MHz
NIC	2 × GE I350	2 × 10 GE Mellanox MT27710	2 × 10 GE Mellanox MT27710	2 × 10 GE Mellanox MT27710
Speicher	2 × 500 GByte NVMe	2 × 480 GByte SATA	240 GByte M.2-SATA; 3,2 TByte NVMe	4 TByte SATA; 1 TByte M.2-NVMe; 3,2 TByte NVMe
Netzteile	2 × 650 W Platinum (2+0)	2 × 750 W Platinum (1 + 1)	Rack: 5 + 1 × 3000 W Platinum über Stromschiene	Rack: 5 + 1 × 3000 W Platinum über Stromschiene

scheiden aus. Dank des mustergültigen AMD-Prozessordesigns ist der 2U-Rackserver der mit Intel CPUs bestückten Wiwynn SV7220 in Sachen Effizienz und Leistung deutlich überlegen.

Das Bild ändert sich kaum, wenn man den Energiebedarf der Systeme mit der SPEC-Leistung in Bezug setzt. Die falsch konfigurierten Netzteile und die kleinen Lüfter vermasseln den Vorsprung, den die ARM-CPU mit ihrem Bedarf von 0,61 Watt/SPECint bei 80 Prozent Auslastung auf der Gigabyte-Maschine vorlegt. Bei der Bewertung der Gesamtsysteme hängt die Mitac Capri mit einem Bestwert von 1,15 Watt/SPECint alle Mitbewerber ab. Die Gigabyte benötigt für die gleiche Arbeit 1,25, die Lenovo 1,3 Watt/SPECint und der OCP-Server von Wiwynn zeigt, dass das Doppelprozessorkonzept in die Jahre gekommen ist. Einzig die Möglichkeit, mehr DIMMs unterzubringen, spricht noch für dieses Design.

Die Ausstattung als wenig beachteter Faktor

Allen getesteten Servern gemein ist ihre recht spartanische Ausstattung. Die Leistungsaufnahme ändert sich deutlich, wenn ein Server mit Zusatzkomponenten vollgestopft wird. Stromhungrige GPUs sind ein bekanntes Beispiel, schnelle Netzwerk- und Storage-Adapter und Disks ein anderes. Die Leistungsaufnahme der Lenovo steigt um circa 100 Watt, sobald man die PCIe-Slots mit zwei RAID-Controllern und einer 100-GE-Karte bestückt und in die Festplatteneinschübe jede Menge Disks und U.2-SSDs steckt. Neben den Komponenten ziehen auch Lüfter, ja selbst die CPU mehr Strom und auch die Netzteilverluste erhöhen sich. Dass sich auch die Leistungsaufnahme des Prozessors erhöht, erklärt sich, wenn man bedenkt, dass er auch die PCIe-Ansteuerung übernimmt.

In Sachen Energieeffizienz kann also das OCP-Prinzip überzeugen, sobald AMDs Epyc darin werkelt. Die Doppelprozessorausführung von Wiwynns Tioga-Pass-Server wirkt im Vergleich dazu ein wenig altbacken. Auch beim Platzbedarf überzeugt Mitacs Capri-Server, denn es passen drei davon nebeneinander ins 60 cm breite Rack. Dank der leistungsfähigen AMD-CPU verfügt jeder Einschub über Platz für zwei PCIe-4.0-x16-HHHL-Slots (halbe Höhe, halbe Länge), sechs SATA- und sechs U.2-NVMe-Drives, einen M.2-Riegel und eine OCP-2.0-Netzwerkkarte. Lediglich für PCIe-Slots ist weniger Platz als in Standard-2U-Servern. Deren Vorteil bei der umfangreichen Festplattenausstattung macht das OCP-Konzept durch geschickte Platzierung von Storage im Rack wett. Angeboten werden etwa NVMe-Laufwerke, die per PCIe-Verlängerung direkt mit den Servern sprechen. Statt eleganter Frontblenden zieren dutzende Kabel die Vorderseite des Open Racks.

Die entscheidende Frage nach dem Preis ist nicht einfach zu beantworten, zumal es in Deutschland bisher keinen Distributor für OCP-Hardware gibt. Die Hersteller selbst halten sich mit Preisauskünften bedeckt. Letztendlich dürfte der Unterschied nicht sehr groß sein: Zum einen bestimmen CPU und Speicher den Großteil der Kosten, zum anderen gleicht ein höherer Aufwand bei der Basisinfrastruktur von 21″-Systemen etwa durch zentrale Netzteile die Einsparungen bei den Systemkomponenten wieder aus. 21″-Technik ist eine Grundsatzentscheidung, die immer mehr Hersteller bejahen (siehe Kasten „OCP-Anbieter im Überblick“).

Fazit

Das 21″-Konzept ist nicht revolutionär, bringt aber mehrere Detailverbesserungen. Durch den Umstieg auf eine zentrale Gleichstromversorgung und den Verzicht auf viele kleine und somit ineffiziente und zudem laute Lüfter sinkt bei Verwendung moderner Prozessoren der Energiebedarf messbar. Hinzu kommt eine Raumökonomie, die beim 19″-Konzept mit höherem Kühlungsaufwand bezahlt werden muss. Der Verzicht auf den Doppelboden und ein klares Bedienkonzept, das alle Ports und Einschübe an die Rackfront verlegt und den Platzbedarf des Warmgangs minimiert, schafft neue Möglichkeiten zur Umnutzung bestehender Räumlichkeiten.

OCP-Server sind nicht per se effizienter. Altbackene Doppelprozessorsysteme im neuen Gewand sind auch konventionellen Servern mit effizienten AMD-Epyc- oder ARM-CPUs unterlegen. Leider wird das gradlinige 2OU3N-Konzept zunehmend durch Rückfall auf 1-OU-Server, also Server, die eine Einschubebene des Open Racks belegen, und ihre schlechten thermischen Eigenschaften verwaschen.

Auch die Energieeffizienz bestehender konventioneller Rackserver kann in vielen Fällen mit wenig Aufwand, ja teilweise ohne Downtime gesteigert werden. Ein einfaches Umstellen der Netzteile auf den 1+1-Betrieb verbessert den Wirkungsgrad zu lastarmen Zeiten erheblich. Das Entfernen ungenutzter Komponenten reduziert den Energiebedarf genauso wie das Entfernen des Skiträgers vom Autodach. Trotzdem stecken in vielen Servern inaktive Komponenten, weil verantwortliche Personen eine Downtime scheuen.

Auch in der Lüftersteuerung bestehender Systeme steckt Potenzial. Trotz Klimatisierung drehen die vielen kleinen Lüfter auf Hochtouren, weil der Algorithmus zur Steuerung zwar alle Messdaten vorliegen hat, diese jedoch nicht vernünftig auswertet. Ein BIOS- oder Firmware-Update kann Stromkosten senken. Gerade in Rechenzentren macht sich jede nicht in Wärme gewandelte Kilowattstunde doppelt bezahlt. (sun@ix.de)

Quellen

[1] Hubert Sieverding; Aufbruch; Gigabytes Rack-Server R152-P30 mit 80 ARM-Kernen im Test; iX 1/2022, S. 42

Hubert Sieverding

arbeitet nach langjähriger Tätigkeit in der Automobilbranche als freier Autor.

heise Academy

Die Webinar-Serie von Heise

SQL Server im Unternehmen

In **5 Online-Trainings** zum SQL-Server-Experten

Wie können Sie das meiste Potenzial aus Ihren Datenbanksystemen herausholen?

In fünf Online-Trainings lernen Sie die wichtigsten Grundlagen und Techniken, um erfolgreich mit dem SQL Server von Microsoft in der Praxis zu arbeiten.

DIE TERMINE:

23. Juni 2022
SQL Server in der Praxis

30. Juni 2022
Abfragetechniken für SQL Server

14. Juli 2022
Erweiterte Abfragetechniken für den SQL Server

21. Juli 2022
Indizes und Performance in SQL Server

28. Juli 2022
Entity Framework Core und SQL Server im Einsatz

Exklusiver Kombi-Preis: 595,-

Einzelpreis: 169,-

Jetzt Kombi-Rabatt sichern und 250,- sparen!

webinare.heise.de/sql-server

Den Fußabdruck einzelner IT-Dienste messen

Flocke für Flocke

Jens Gröger

Kennen Rechenzentrumsbetreiber den Umweltfußabdruck ihres RZ, können sie den auch auf einzelne Dienste oder Datenpakete herunterrechnen. Für eine Vergleichbarkeit bedarf es aber akzeptierter Methoden.

Umweltaspekte spielen auch in den Beschaffungsabteilungen der Unternehmen eine immer wichtigere Rolle. Während es bei physischen Produkten längst üblich ist, den Energieverbrauch oder andere Umwelteigenschaften als Einkaufskriterium heranzuziehen, ist dies bei Cloud-Diensten heute noch nicht möglich. Dabei haben auch Cloud-Rechenzentren eine physische Grundlage und einen messbaren Umweltfußabdruck.

iX-TRACT

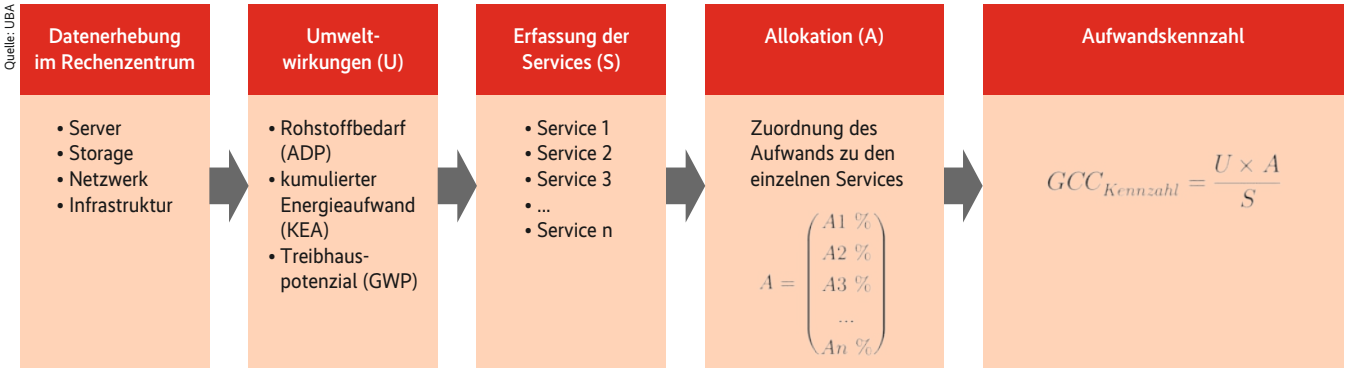
- Wer Cloud-Dienstleistungen unter Umweltgesichtspunkten genauer unter die Lupe nehmen will, benötigt Messverfahren und Kennzahlen.
- Im Auftrag des Umweltbundesamtes hat das Forschungsprojekt Green Cloud Computing solche Methoden und Kennzahlen entwickelt.
- Untersucht hat das Projekt einzelne Dienste wie Cloud-Storage, Videokonferenzen und Virtual Desktop Infrastructure.
- Tatsächlich ließe sich sogar der Fußabdruck einzelner Datenpakete bestimmen und in den TCP-Header eintragen.

Beispielsweise hat sich allein der Stromverbrauch von Cloud-Rechenzentren in Europa von 2010 bis 2020 mehr als versechsfacht und wird auch in Zukunft weiter steigen – und mit ihm die Treibhausgasemissionen [1].

Deshalb wird es immer wichtiger, Cloud-Dienstleistungen auch unter Umweltgesichtspunkten genauer unter die Lupe zu nehmen: Wie groß ist der Umweltfußabdruck eines Cloud-Service? Wie lassen sich die Angebote vergleichen? Unter welchen Umständen bietet das Auslagern von Rechen- und Speicherressourcen in die Cloud Vorteile? Solche Fragen hat sich das Umweltbundesamt gestellt und das Forschungsprojekt Green Cloud Computing (GCC) initiiert. Darin haben das Öko-Institut und das Fraunhofer IZM eine Methodik entwickelt, mit der sich Cloud-Dienstleistungen bilanzieren lassen [2].

Umweltbilanzierung von Rechenzentren

Die erste Herausforderung liegt darin, die Umweltwirkungen von Rechenzentren zu bestimmen. Bei einem einzelnen Produkt hat sich die Methode der Ökobilanzierung nach ISO 14040/14044 durchgesetzt (siehe ix.de/z1y8). Die Ökobilanz umfasst dessen gesamten Lebenszyklus, von der Rohstoffgewinnung über Herstellung, Transport und Nutzung bis hin zur Entsorgung, also „von der Wiege bis zur Bahre“.



Bei der Berechnung der GCC-Kennzahlen werden die beteiligten Komponenten auf ihre unterschiedlichen Umweltwirkungen hin untersucht und diese den bereitgestellten Services zugeordnet (Abb. 1).

Bei Rechenzentren beinhaltet die normgerechte Ökobilanz die komplette Vorkette zur Herstellung elektronischer Bauteile, Leiterplatten und Gehäusematerialien, die Fertigung von Halbleitern, IT-Equipment und Gebäudetechnik, den Transport der Geräte sowie den Betrieb des Rechenzentrums inklusive Energie- und Medienverbrauch. Schließlich gehören auch die Demontage veralteter Geräte, deren Recycling, Verbrennung oder Deponierung dazu. Entlang dieser gesamten Kette werden Umweltwirkungen berechnet und in die Bilanz einbezogen. Allerdings ist das Erstellen einer Ökobilanz für elektronische Produkte, die ihrerseits aus Hunderten von Bauteilen bestehen, ausgesprochen aufwendig. Für Rechenzentren multipliziert sich der Aufwand noch einmal durch die Vielzahl der darin verbauten Komponenten. Zudem sind die Hersteller von Elektronikprodukten sehr sparsam mit Informationen zu den sich rasant weiterentwickelnden Fertigungstechniken und deren Umweltwirkungen. Dennoch haben die Beteiligten des GCC-Projekts zumindest orientierende Ökobilanzen für Server, Storage und Netzwerkkomponenten erstellt, die für den größten Teil der Umweltwirkungen von Rechenzentren verantwortlich sind (siehe Artikel „Weiter gefasst“ ab Seite 78).

Eine Frage der Kategorisierung

Umweltwirkungen lassen sich in unterschiedlichen Umweltwirkungskategorien beschreiben. Das GCC-Projekt hat sich auf die Kategorien abiotischer Rohstoffverbrauch (ADP), kumulierter Energieaufwand (KEA) und das globale Treibhausgaspotenzial (GWP) beschränkt (siehe Abbildung 1). Andere Ressourcen wie Wasser und Flächen sind für Rechenzentren ebenfalls wichtig. Deren Relevanz ist jedoch stark vom jeweiligen Standort abhängig, weshalb sie nur im Einzelfall zu betrachten sind.

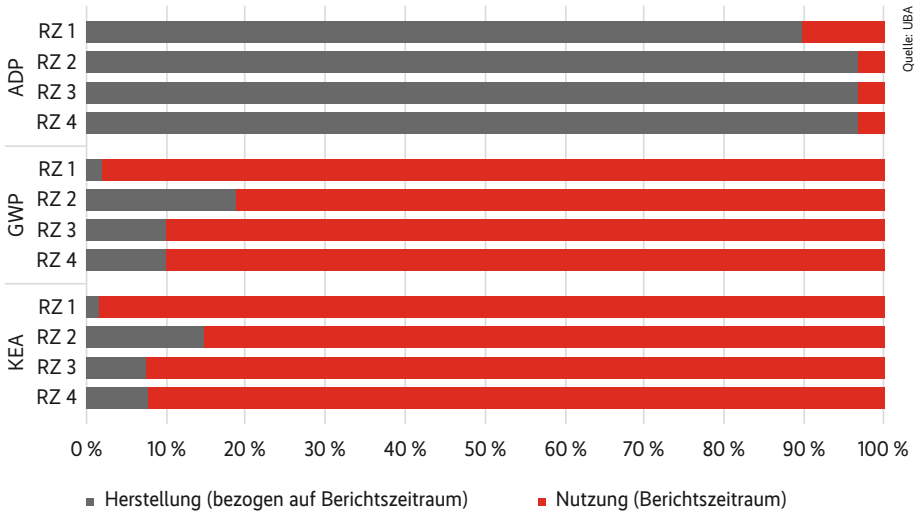
Nicht zufällig betreffen 90 bis 97 Prozent des Rohstoffverbrauchs die Herstellungsphase (siehe Abbildung 2). Der Energieaufwand und Treibhausgasemissionen treten dagegen zu rund 90 Prozent in der Nutzungsphase auf. Geschuldet ist das vor allem dem hohen Strombedarf der Rechenzentren. Beschränkt man den Blick auf diese beiden Umwelteffekte, kann man die Kalkulation vereinfachen und sich die komplette Ökobilanz ersparen. Der Stromzähler des Rechenzentrums verrät bereits den größten Teil des Energiebedarfs. Die Treibhausgasemissionen lassen sich daraus mit dem durchschnittlichen Strommix berechnen.

Sind KEA und GWP des Rechenzentrums bilanziert, sind die Daten auf die darin angebotenen Dienstleistungen herunterzurechnen. Einfach ist das immer dann, wenn ein RZ ausschließlich eine einzelne Dienstleistung anbietet.

$$GCC_{Kennzahl} = \frac{Umweltwirkung\ des\ RZ}{Zahl\ der\ Serviceeinheiten}$$

Online-Storage

Stellt ein Rechenzentrum beispielsweise ausschließlich Online-Storage mit insgesamt 10 PByte zur Verfügung und hat dabei einen CO_2-Fußabdruck von 2000 t pro Jahr, lässt sich daraus leicht ein spezifischer Umweltfußabdruck von 200 kg CO_2-eq/TByte (CO_2-Äquivalent/TByte) errechnen. Das GCC-Forschungsprojekt hat diese Rechnung für vier verschiedene Rechenzentren nachvollzogen. Das jährliche GWP für Online-Storage lag bei den untersuchten Rechenzentren zwischen 166 und 280 kg CO_2-eq/TByte. Der Wert gibt unmittelbar Auskunft über den CO_2-Fußabdruck, den der Betrieb des Cloud-Dienstes Online-Storage verursacht – ohne seinen Aufbau.



Die Phasen des RZ-Lebenszyklus werden von unterschiedlichen Umweltwirkungen geprägt. In der Herstellung dominiert der Materialaufwand, bei der Nutzung der Energiebedarf (Abb. 2).

Zuordnung von Umweltwirkungen auf Services

	Allokationsfaktor				
Umweltwirkung Teilbereich	Service 1	Service 2	...	Service n	Summe
$Server_{GWP}$	$A1_{sv}$	$A2_{sv}$	...	An_{sv}	100 %
$Storage_{GWP}$	$A1_{st}$	$A2_{st}$	...	An_{st}	100 %
$Network_{GWP}$	$A1_{nw}$	$A2_{nw}$	...	An_{nw}	100 %
$Infra_{GWP}$	$A1_{infr}$	$A2_{infr}$	...	An_{infra}	100 %

Quelle: UBA

Komplizierter wird die Berechnung, wenn das Rechenzentrum unterschiedliche Dienstleistungen erbringt. Dies ist fast immer der Fall, denn ein Rechenzentrum erfüllt in der Regel viele unterschiedliche Aufgaben parallel. Dann ist der Fußabdruck des Rechenzentrums auf mehrere Dienste aufzuteilen – in der Ökobilanzierung Allokation genannt. Dazu bedarf es geeigneter Allokationsregeln. Die GCC-Methodik schlägt vor, je nach Anwendungsfall entweder über Hardware- oder über virtuelle Ressourcen zu allozieren. Hardwareressourcen können entweder ganze Rechenzentren oder einzelne Systeme sein, etwa Server oder Storage-Systeme. Bei den virtuellen Ressourcen kommen insbesondere virtuelle Server, Container oder einzelne Softwarefunktionen als Aufteilungsgrößen infrage.

Jeder Service erhält einen prozentualen Anteil an den Umweltwirkungen des jeweiligen Teilbereiches, sodass sich das GWP auf alle angebotenen Dienste verteilt (siehe Tabelle „Zuordnung von Umweltwirkungen auf Services“). Beispielsweise rechnet man das GWP dem Service i zu, indem man mit einem Allokationsfaktor Ai das jeweilige GWP von Server, Storage, Netz und Infrastruktur alloziert und die einzelnen Ergebnisse addiert:

$$Service\ i_{GWP} = Ai_{server} \times Server_{GWP} + Ai_{netz} \times Netz_{GWP} + Ai_{infra} \times Infra_{GWP}$$

Videokonferenzen

Untersucht anhand physischer Ressourcen hat das GCC-Projekt das Bereitstellen von Onlinevideokonferenzen (siehe Tabelle „GCC-Kennzahl-Berechnung für Videokonferenzen“). Der Plattformbetreiber hat in einem Rechenzentrum vier Server gemietet, die er ausschließlich für seinen Onlinedienst verwendet. Die Herstellung der vier Server verursacht ein GWP von 1945 kg CO_2-eq, der sich bei einer vierjährigen Nutzungsdauer auf rund 9 kg pro Woche herunterrechnen lässt. Bei einem Rechenzentrums-PUE von 1,12 kam der Anbieter des Videokonferenzdienstes auf einen wöchentlichen Stromverbrauch von 112 kWh, was 52 kg CO_2-eq entspricht.

Als Serviceeinheit wurde für die Videokonferenzen die Einheit Teilnehmerstunden (Tlnh) gewählt, die sich als Produkt aus der Zahl der Teilnehmer und deren Verweildauer in einer Videokonferenz berechnet. Der Anbieter stellt pro Woche 27 133 Tlnh Videokonferenzen bereit. Die GCC-Kennzahl beträgt hier also 2,27 g CO_2-eq pro Teilnehmer und Stunde.

Virtuelle Desktop-Infrastruktur

Eine Allokation virtueller Ressourcen hat das Projekt bei der Untersuchung der Desktop-Virtualisierung als Cloud-Dienstleistung vorgenommen. Eine VDI (virtuelle Desktop-Infrastruktur) stellt die gesamte Clientsoftware bereit, die nicht der lokale Computer, sondern ein Server ausführt. Das Endgerät dient nur noch zur Dateneingabe über Tastatur und Maus sowie zur Datenausgabe über einen Monitor. Die GCC-Methode bilanzierte den Energieverbrauch und die CO_2-Emissionen, die durch eine VDI in einem Rechenzentrum entstehen. Dadurch ist es möglich, sie mit der lokalen Variante zu vergleichen.

Das untersuchte Rechenzentrum bot viele Dienste an, darunter Webressourcen, Datenbanken, Mail-, Print-, File- und Backup-Dienste, Videokonferenzen und Mobile-Device-Management. Die Bilanzierung des gesamten Rechenzentrums ergab, dass das GWP bei jährlich 447 t CO_2-eq lag. Etwas mehr als ein Zehntel der Rechenzentrumsressourcen ließ sich der VDI für 890 Arbeitsplätze zuordnen. Über das Virtualisierungsmanagement des Rechenzentrums werden die Mengen an insgesamt verfügbaren virtuellen Servern, Speichervolumina und Netzwerkkarten bestimmt und den unterschiedlichen Diensten zugewiesen (siehe Tabelle „Allokation der Treibhausgasemissionen über virtuelle Ressourcen“).

Für die VDI ergab sich dadurch ein Anteil an den virtuellen Ressourcen, der sich als Allokationsfaktor zur Berechnung des GWP heranziehen lässt. Durch die VDI werden zwölf Prozent der virtuellen Server, neun Prozent der virtuellen Storage-Kapazitäten und zehn Prozent der Netzwerkkarten genutzt. Für die Zuordnung der Gebäudetechnik auf den Cloud-Service wurde der gewichtete Mittelwert der drei IT-Komponenten verwendet.

Aus dieser Allokation ergibt sich ein GWP für alle VDI-Instanzen von 53 t CO_2-eq/Jahr. Als Serviceeinheit zählen hier die 890 Arbeitsplätze, die der Dienst bereitstellt. Dies ergibt eine GCC-Aufwandskennzahl von 59 kg CO_2-eq pro VDI-Arbeitsplatz und Jahr. Anhand dessen lässt sich die VDI-Implementierung mit Thin Clients mit PC-Arbeitsplätzen vergleichen: Dabei bietet die VDI-Variante eine jährliche Einsparung von 33 kg CO_2-eq pro Arbeitsplatz.

Fazit

Die drei Beispiele Online-Storage, Videokonferenzen und VDI zeigen, dass sich das GWP für unterschiedlichste Cloud-Dienstleistungen berechnen lässt. Prinzipiell ist dies auch für die Umweltwirkungskategorien ADP, KEA oder Wasser möglich. Bei Bedarf sind auch feiner graduierte Allokationen denkbar, etwa das GWP einer im Container ausgeführten Software oder einer einzelnen Rechenaufgabe, etwa eines Dienstes, der einen Audiostream in einen lesbaren Text umwandelt.

In dem vom Umweltbundesamt beauftragten Projekt entstand zunächst nur eine Mess- und Berechnungsmethode zur einheitlichen Berechnung des Umweltfußabdrucks eines Cloud-

Quelle: UBA

GCC-Kennzahl-Berechnung für Videokonferenzen

Herstellungsaufwand $Server_{GWP}$	**1945 kg CO_2-eq**
Nutzungsjahre pro Server	4 Jahre
Aufwand: Herstellung pro Woche	9,35 kg CO_2-eq/Woche
Energiebedarf Rechenzentrum	112 kWh_{el}/Woche
Aufwand: Nutzungsphase pro Woche (Energiebedarf)	52 kg CO_2-eq/Woche
Nutzen: Teilnehmerstunden pro Woche	27 133 Tlnh/Woche
Berechnung der GCC-Kennzahl (= Aufwand/Nutzen)	
GCC $Videokonferenz_{GWP}$ Herstellung	0,34 g CO_2-eq/Tln/h
GCC $Videokonferenz_{GWP}$ Nutzung	1,93 g CO_2-eq/Tln/h
GCC $Videokonferenz_{GWP}$	**2,27 g CO_2-eq/Tln/h**

Quelle: UBA

Allokation der Treibhausgasemissionen über virtuelle Ressourcen

Hardware	GWP Rechenzentrum	VDI-Anteil an virtuellen Ressourcen	GWP VDI
Server	320 t CO_2-eq/a	12 %	39 t CO_2-eq/a
Storage	41 t CO_2-eq/a	9 %	4 t CO_2-eq/a
Netzwerk	20 t CO_2-eq/a	10 %	2 t CO_2-eq/a
Gebäudetechnik	66 t CO_2-eq/a	12 %	8 t CO_2-eq/a
Summen	**447 t CO_2-eq/a**		**53 t CO_2-eq/a**

Dienstes. Damit diese Methode die ökologische Transparenz von Cloud-Diensten erhöhen kann, bedarf es einer breiten Anwendung. Ein wichtiger Baustein hierfür könnte die automatische Kennzeichnung solcher Dienste mit einer Umweltkennzahl sein.

Jede Datenübertragung sollte umweltrelevante Informationen enthalten. Beispielsweise könnten die optionalen Felder des TCP-Headers den ADP und das GWP des Datenpakets enthalten. Dadurch könnten Kunden den Umweltaufwand protokollieren und bei verteilten Diensten über Anbietergrenzen hinweg bilanzieren. Die vom Umweltbundesamt aufgeworfenen Fragen ließen sich damit leicht für die konkreten Fälle beantworten, zum Beispiel: Das Treibhausgaspotenzial der Rechenaufgabe beträgt bei Anbieter A 2,5 g, bei Anbieter B nur 1,5 g und bei einer lokalen Berechnung 3 g CO_2-eq.

Mit der GCC-Methodik liegt ein definiertes und reproduzierbares Verfahren vor, mit dem sich Cloud-Services bilanzieren und kennzeichnen lassen. Nun sind die Cloud-Anbieter an der Reihe, dieses Verfahren anzuwenden und ihre Dienstleistungen in Zukunft transparenter zu machen. Die Kennzeichnung mit einem GWP bietet für umweltbewusste Anbieter auch wirtschaftliche Vorteile. Denn mit einer Kennzeichnung entsteht ein Wettbewerb um den umweltverträglichsten Cloud-Dienst. Kunden können gezielt diejenigen auswählen, deren Anbieter ihre Verantwortung zum Umwelt- und Klimaschutz wahrnehmen. (sun@ix.de)

Quellen

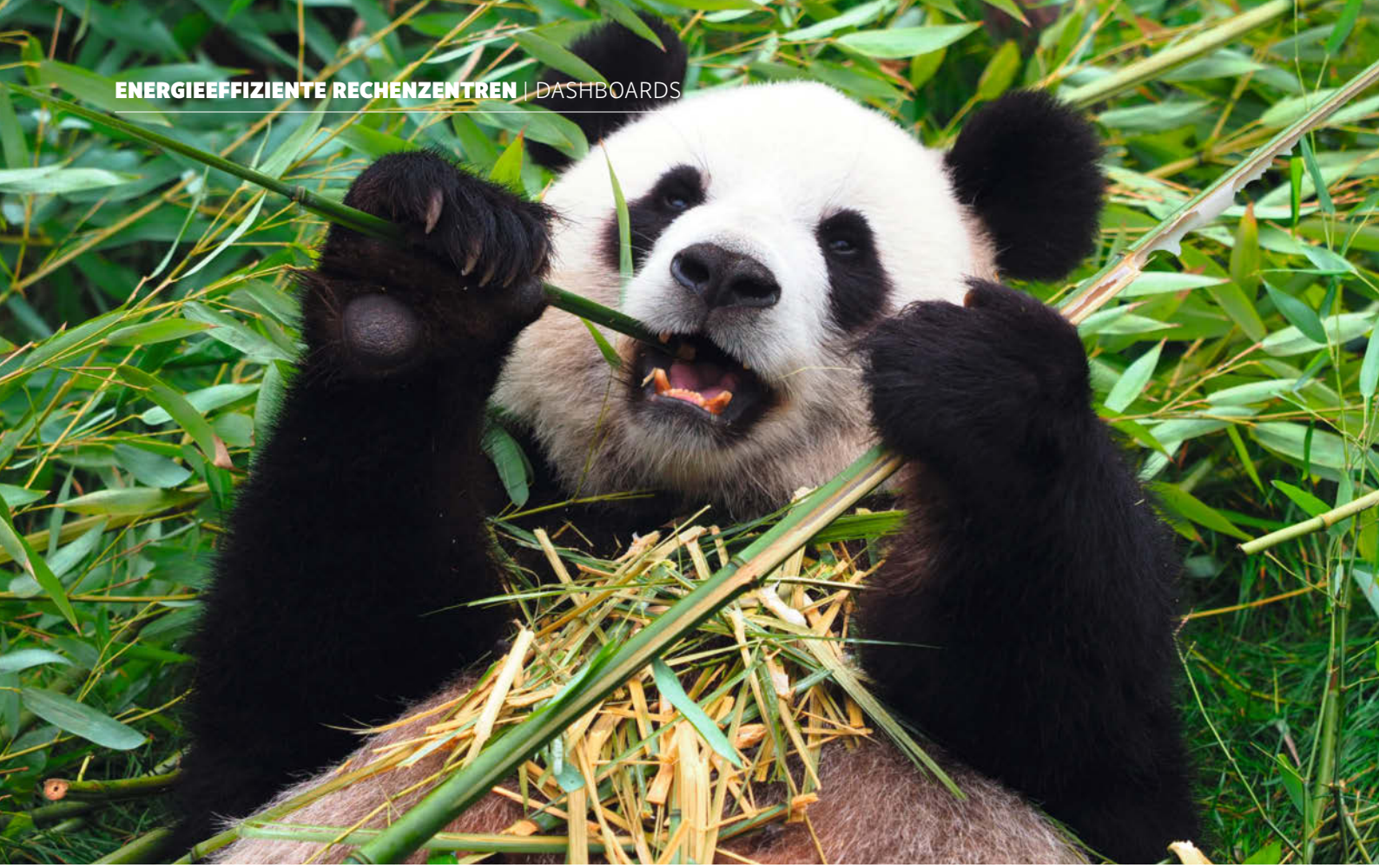
[1] F. Montevecchi, T. Stickler, R. Hintemann, S. Hinterholzer; Energy-efficient Cloud Computing Technologies and Policies for an Eco-friendly Cloud Market, Final Study Report; Wien 2020

[2] Jens Gröger, Ran Lui, Lutz Stobbe, Jan Druschke, Nikolai Richter; Green Cloud Computing; Lebenszyklusbasierte Datenerhebung zu Umweltwirkungen des Cloud Computing; 2021

[3] Links zu den referenzierten Papers: ix.de/z1y8

Jens Gröger

ist Senior Researcher beim Öko-Instituts e. V. im Fachbereich Produkte und Stoffströme. Sein Forschungsschwerpunkt liegt bei nachhaltiger Informations- und Kommunikationstechnik. Er entwickelt Methoden zur Bewertung der Energieeffizienz von Computern, Software, Telekommunikationsnetzen und Rechenzentren.



Die CO_2-Emissionen der eigenen Cloud-Nutzung visualisieren

Bewusst konsumieren

Frank Pientka

Nicht nur aus ökonomischen Gründen sind Cloud-Provider daran interessiert, klimaneutraler zu werden. Doch noch fehlen geeignete Bemessungsgrundlagen. Wie also können Kunden ihren CO_2-Fußabdruck in der Cloud abschätzen und optimieren?

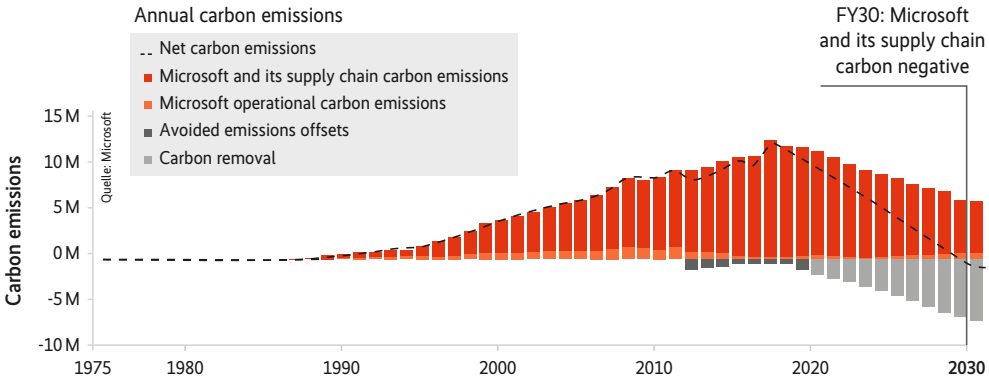
Bei der Optimierung sollte man nicht zu früh anfangen, denn schließlich ist sie oft nach Donald Knuth die „Wurzel allen Übels“. Trotzdem muss das Energieeffizienzziel gerade bei der Cloud-Nutzung eine frühere und größere Rolle spielen. Mit dem Greenhouse Gas Protocol lassen sich zwar zu den drei Scopes direkte, indirekte und sonstige Emissionen standardisierte Berichte erstellen, es ist jedoch kein verbindlicher und allgemein anerkannter Standard. Insofern sind die Marketingberichte der Cloud-Anbieter mit Vorsicht zu genießen, da es hier weder eine einfache Vergleichbarkeit noch allgemein anerkannte Standards gibt (siehe Abbildung 1 und ix.de/zd2e). Deswegen empfiehlt es sich, selbst geeignete Metriken zu finden, die in den unternehmensweiten Klimabericht einfließen können.

Der bekannteste Wert zum Ermitteln der Energieeffizienz ist die PUE (Power Usage Effectiveness). Ihn hat die Green Grid Initiative 2007 eingeführt und die ISO 2016 als ISO/IEC 30134-2:2016 veröffentlicht. Er teilt den Gesamtverbrauch des RZ durch den des IT-Equipments und gibt damit an, welchen Anteil die RZ-Infrastruktur wie Kühlung und unterbrechungsfreie Stromversorgung am Energieverbrauch hat. Im Idealfall ist die PUE = 1, dann nämlich fließt der gesamte RZ-Strom in die IT, und die Infrastruktur hätte keinen Anteil daran. Leider sagt die PUE nichts über die Effizienz und die Auslastung des IT-Equipments aus (siehe Artikel „Weiter gefasst“ und „Maß-Stäbe setzen“ ab Seite 78 und 105).

Da die PUE je nach Klimazone übers Jahr mehr oder weniger starken Schwankungen unterliegt, wird sie meist über zwölf Monate gemittelt. Eine solche TTM-PUE (Trailing Twelve-Month PUE) verwendet auch Google. Der IT-Dienstleister konnte die TTM-PUE für seine Rechenzentren von ursprünglich 1,21 auf derzeit 1,1 reduzieren (siehe Abbildung 2). Eine Verbesserung von 0,11 klingt zwar nicht nach viel, angesichts der Größe von Google ist das aber eine ganze Menge.

Googles zuständiger Senior-Vizepräsident Urs Hölzle hat das dahinterstehende energieoptimierte Rechenzentrumsdesign in seinem IEEE-Artikel beschrieben (siehe ix.de/zd2e). Er führt auch aus, dass die Rechenhardware für einen Hochlastbetrieb optimiert ist, die meisten Anwendungen jedoch diesen Bereich

Microsoft etwa will bis 2030 „CO_2-negativ“ und bis zum Jahr 2050 sogar „treibhausgasneutral“ werden, was bedeutet, alle Emissionen zu kompensieren, die es seit seiner Gründung verursacht hat, und will das auch von seinen Lieferanten verlangen (Abb. 1).



kaum erreichen und dadurch mehr Energie verbrauchen als nötig. Seine eigenen Arbeitslasten verteilt Google über Zeitzonen hinweg auf die Rechenzentren, die gerade freie Kapazitäten haben. Mit einer leichtgewichtigen Virtualisierung und Containerisierung können Cloud-Anbieter die Arbeitslasten feingranularer verteilen und so die Auslastung verbessern. Durch solche und andere Maßnahmen und die sich daraus ergebenden Skaleneffekte erreichen Cloud-Anbieter eine größere Energieeffizienz, als sie On-Premises-Rechenzentren jemals erreichen könnten.

Der Energieverbrauch der eigenen Cloud-Nutzung

Doch welche Faktoren bestimmen den Energieverbrauch der eigenen Cloud-Nutzung? Das ist zunächst das ausgewählte Rechenzentrum. Allerdings geben die Anbieter unterschiedlich gern Auskunft über die Effizienz ihrer Rechenzentren. Google Cloud weist den ökologischen Fußausdruck jedes seiner Rechenzentren aus (siehe ix.de/zd2e). In manchen Regionen, etwa in Asien, in denen der größte Teil des Stroms auch in Zukunft nicht aus regenerativen Quellen stammen wird, bleibt nur der Wechsel in eine Region mit einer nachhaltigeren Energieerzeugung. Bis 2030 wollen alle großen Anbieter klimaneutral werden, was nur geht, wenn sie selbst regenerative Energie erzeugen.

Auch wenn die Cloud-Anbieter daran arbeiten, ihre Effizienz zu steigern, sind auch die Kunden gefragt, verantwortungsvoll mit den von ihnen gemieteten Ressourcen umzugehen. Dazu lässt sich das verteilte Verantwortungsmodell von AWS auf die grüne Cloud übertragen (siehe Abbildung 3). Für den energieeffizienten Betrieb ist der Anbieter verantwortlich. Dafür, welche Dienste ein Kunde wie intensiv nutzt, ist der Kunde selbst verantwortlich. Nur wer seine Last- und Qualitätsanforderungen kennt, kann die geeigneten Cloud-Dienste auswählen.

Wer den Netzverkehr reduzieren will, lässt die Daten möglichst nahe und lange beim Nutzer. Hierzu gibt es für immer mehr Cloud-Dienste unterschiedliche Caches oder Edge-Konzepte. Die beste Optimierung ist aber die Vermeidung: Nur Daten, die man nicht versendet, benötigen keine Energie. Zudem sollte man den Datenverkehr möglichst Ost-West innerhalb des lokalen Netzes des Cloud-Anbieters belassen und den ein- und ausgehenden Nord-Süd-Verkehr in die Cloud minimieren. Es gibt sogar einen nicht ganz ernst gemeinten RFC 7511, der erklärt, wie man für das Routen des IPv6-Verkehrs den grünsten Weg nutzen kann.

Datensparsamkeit reduziert auch den Speicherverbrauch. Bei der Auswahl des persistenten Speichers muss man oft einen Kompromiss zwischen Verfügbarkeit, Zugriffszeit und Kapazität eingehen. Deswegen bietet es sich an, seine Daten je nach Nutzungsart und -dauer in Datenklassen einzuteilen, die man je nach Lebenszyklus auf günstigere und oft energiesparsamere Medien auslagert.

Bereits mit der Auswahl des Prozessors kann man den Energieverbrauch senken. Diesseits von HPC-Anwendungen kann man bei allen Cloud-Anbietern zwischen Intel- und AMD-Prozessoren wählen. Inzwischen werden ARM-Prozessoren auch auf Servern immer beliebter. Sie sind aufgrund ihrer Herkunft aus dem Embedded- und Mobile-Bereich auf einen geringen Energiebedarf optimiert, davon profitieren auch Serveranwendungen [1, 2]. AWS bietet inzwischen seine selbst entwickelten ARM-Prozessoren Graviton in der dritten Generation für kostengünstige EC2- oder RDS-Instanzen an. Der Graviton2-Prozessor hat eine um den Faktor 3,45 bessere Energieeffizienz als vergleichbare x86-Prozessoren [3]. Dadurch, dass viele Programme auf ARM portiert sind, steht einem Wechsel der CPU-Plattform nichts im Wege.

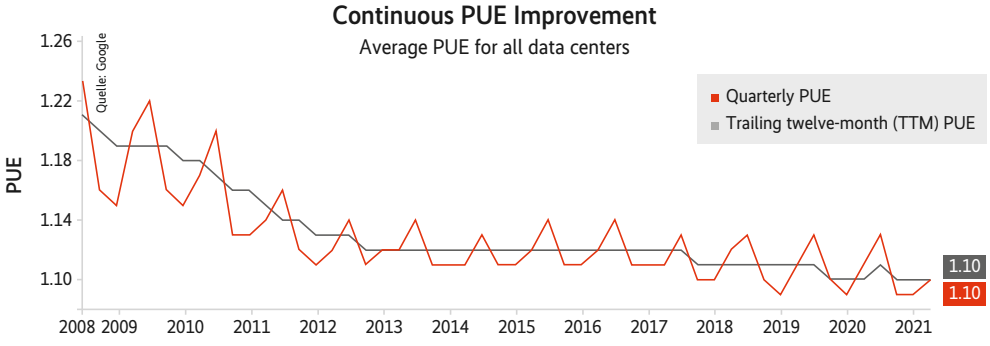
Dashboards zum Berechnen des Cloud-CO_2-Fußabdrucks

Die Energieeffizienz ist nicht nur eine Frage der Hardware, sondern auch der eingesetzten Software, ihres Lastprofils und ihrer Anforderungen. Inzwischen bieten die Cloud-Provider ein Dashboard an, mit dem man auf Grundlage seiner Abrechnungsdaten den CO_2-Fußabdruck berechnen kann. Da diese Programme jedoch noch in der Entwicklung sind, fließen noch nicht alle Dienste und Regionen in die Berechnungen ein.

Bei der GCP (Google Cloud Platform) ist das Dashboard direkt in die Managementkonsole integriert. Dadurch ist es möglich, die berechneten Daten auch weiter auszuwerten. Bei Microsofts Azure Emissions Impact Dashboard funktioniert das nicht.

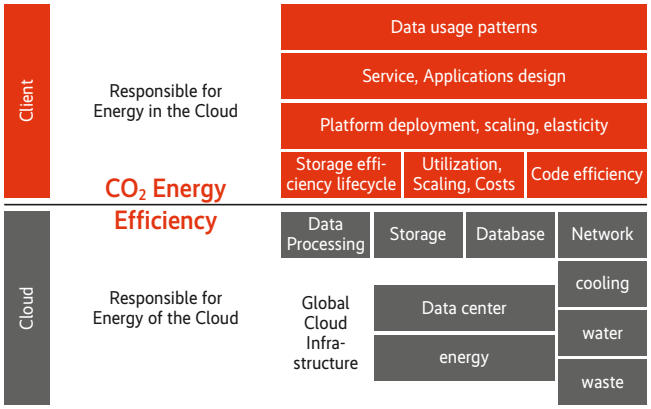
iX-TRACT

- Verantwortlich für die Klimabelastung durch die Cloud-Nutzung sind Anbieter und Kunden: Die Cloud-Anbieter müssen die Ressourcen möglichst effizient zur Verfügung stellen und die Kunden verantwortungsvoll mit ihnen umgehen.
- Erste Cloud-Provider stellen ihren Kunden Dashboards bereit, mit denen sie den CO_2-Fußabdruck ihrer eigenen Cloud-Nutzung berechnen können.
- Noch fehlen aber standardisierte Metriken, um die Daten der Provider-Dashboards zu vergleichen und in firmeneigenen Klimaberichten weiterzuverarbeiten.
- Einen ersten Überblick gibt das Multi-Cloud-Werkzeug Cloud Carbon Footprint, das unter der Lizenz Apache 2 steht.



Google hat die PUE seiner Rechenzentren vor allem in den Jahren zwischen 2008 und 2012 kräftig gesenkt (Abb. 2).

Die separate Power-BI-Anwendung kostet extra und ist nur für bestimmte Kontenarten einsetzbar. Amazon stellt derzeit nur die Best-Practice-Empfehlungen in der Nachhaltigkeitssäule des AWS-Well-Architected-Framework bereit, die in diesem Jahr noch in die Managementkonsole integriert werden sollen (siehe ix.de/zd2e). Leider werden die zugrunde liegenden Berechnungen nicht veröffentlicht, sondern es wird bestenfalls die verwendete grobe Methodik erläutert.



Quelle: AWS

AWS skizziert die Verantwortungsbereiche von Kunden und Betreibern für die Green Cloud (Abb. 3).

Listing: Cloud Carbon Footprint herunterladen, einrichten und starten

```
git clone -c core.autocrlf=false --branch latest https://
github.com/cloud-carbon-footprint/cloud-carbon-footprint.
git
[...
npx @cloud-carbon-footprint/create-app
yarn install
packages/api/.env
...
yarn start-with-mock-data
```

AWS-EC2-Instanzen mit ARM Graviton2, AMD Epyc und Intel Xeon im Vergleich

EC2-Typ	m6g	m5a	m5n
CPU-Plattform	Graviton2	Epyc 7571	Xeon Platinum 8259CL
Cores pro Socket	64	32	24
Takt	2,5 GHz	2,5–2,9 GHz	2,9–3,2 GHz
Architektur	Arm v8.2	x64 + AVX2	x64 + AVX512
Mikroarchitektur	Neoverse N1	Zen	Cascade Lake
TDP	80–110 W (geschätzt)	180 W	210 W
Preis	2,464 US-Dollar/Std.	2,752 US-Dollar/Std.	3,808 US-Dollar/Std.

Dadurch ist es wie bei den Cloud-Kosten kaum möglich, einen Cloud-übergreifenden Vergleich pro Dienst zu erstellen oder bei der Nutzung mehrerer Clouds eine einheitliche Grundlage für den eigenen CO_2-Bericht zu erhalten. Trotzdem sind diese Berichte kombiniert mit anderen Cloud-Verbrauchsmessungen ein erster Schritt zu einer energieeffizienten Cloud-Nutzung.

Ein unabhängiges Werkzeug für die Berechnung des eigenen Multi-Cloud-CO_2-Fußabdrucks ist der Cloud Carbon Footprint (CCF), den ThoughtWorks unter die Lizenz Apache 2.0 gestellt hat (siehe ix.de/zd2e). Das Werkzeug kann am besten mit AWS umgehen, lässt sich aber mit Abstrichen auch mit Azure und GCP einsetzen. Grundsätzlich errechnet es aus der täglichen Verbrauchsabrechnung den Energieverbrauch und die dadurch erzeugten CO_2-Emissionen pro Dienst und pro Konto. Dabei fließt die Energieeffizienz des verwendeten Cloud-Rechenzentrums ein. Die Befehle zum Herunterladen, Einrichten und Starten des Frontends mit Dashboard und Demodaten zeigt das Listing.

Das Dashboard lässt sich samt Demodaten unter http://localhost:3000 im Browser aufrufen, alternativ als Demoanwendung auch direkt ohne lokale Installation im Web unter demo.cloudcarbonfootprint.org. Es kann neben der Darstellung des CO_2-Emissionsverlaufs der analysierten Konten diesen auch in Flugstunden, zu pflanzende Bäume oder Akkukapazität des Smartphones umrechnen (siehe Abbildung 4). Die CO_2-Äquivalent-Emissionen berechnen sich wie folgt:

```
Cloud provider service usage × Cloud energy ↵
                                conversion factors [kWh] ×
Cloud provider PUE × grid emissions factors ↵
                                    [metric tons CO₂e]y
```

Umgang mit dem Cloud Carbon Footprint

AWS-Metriken bereitet Cloud Carbon Footprint über den nach S3 exportierten Kosten- und Verbrauchsbericht mit Amazon Athena auf, GCP-Daten über den exportierten und mit BigQuery aufbereiteten Kostenbericht und Azure-Werte über die Verbrauchs-API. Gruppiert man die Daten nicht nach Regionen, sondern nach Diensten, sieht man, dass es nicht sehr viele Dienste unterstützt. Grobe erste Empfehlungen, wie man seinen Energieverbrauch reduzieren kann, gibt das Werkzeug derzeit nur für GCP und AWS (siehe Abbildung 5). Auf jeden Fall hilfreich ist die Darstellung des Effizienzgrads der Cloud-Rechenzentren, um hier durch einen Regionswechsel CO_2 einzusparen, falls der nicht berücksichtigte Netzverkehr das wieder auffrisst.

Sinnvoll ist es jedoch, die CCF-Anwendung etwa auf einer AWS-EC2-Instanz zu hosten und die Verbindungen zu den Clouds der anderen Anbieter dort einzurichten. Falls das Recht zum Auslesen von Rechnungen nicht in der Hand einer zentralen Abteilung liegt, kann man die Verbrauchswerte in den JavaScript-Dateien anpassen oder noch fehlende Dienste in den Quellen ergänzen. Grundsätzlich ist es möglich, Verbrauchs-

Quelle: ThoughtWorks

Das Dashboard im Cloud Carbon Footprint rechnet den Ressourcenverbrauch ins CO_2-Äquivalent um (Abb. 4).

daten auch aus anderen Quellen zu integrieren, wenn man sie auf die bereits vorhandenen Felder abbilden kann. Über die Exportfunktion können die aggregierten Werte in den eigenen Klimabericht einfließen.

Die ThoughtWorks-Empfehlungen für eine grüne Cloud verteilen sich auf die drei Bereiche Konfiguration, Optimierung und Umbauen (siehe ix.de/zd2e). Hat man ein möglichst effizientes Rechenzentrum ausgewählt, gilt die Aufmerksamkeit dem geschätzten Energieverbrauch der genutzten Dienste: Neben dem Speicherverbrauch und der übertragenen Datenmenge sollte man auch die Berechnungskomplexität der eignen Programme reduzieren. Gerade Modethemen wie ML oder Blockchain sollten noch mehr mit Bedacht eingesetzt werden, da sie einen enormen Energieverbrauch haben und viele Anwendungsfälle auch mit traditionellen Verfahren auskommen. Zudem ist ein lastabhängiges Herunterskalieren oder ein zeitgesteuertes Herauf- und Herunterfahren von Ressourcen ein einfaches Mittel, um nicht mehr benötigte Ressourcen freizugeben und die vorhandenen elastischen Möglichkeiten der Cloud zu nutzen.

Fazit

Da die Kosten in der Cloud meist verbrauchsabhängig sind, kann man diese gemeinsam mit dem dahinterstehenden Energieverbrauch optimieren. Dabei profitiert man oft von den technischen Fortschritten der Cloud-Anbieter, da es einfacher ist, deren neuere, energieeffizientere Dienste zu nutzen, statt die einmal angeschaffte und noch nicht abgeschriebene Alt-Hardware auszutauschen. Doch auch hier gilt: Die grünsten Dienste sind die, die man nicht braucht. Für eine optimale Cloud-Nutzung muss man vor allem den Ressourcenbedarf seiner Anwendungen gut kennen und kann dann die dazu passenden Dienste und virtuellen Instanzklassen auswählen. Wenn der Ressourcenbedarf über die Zeit schwankt, sollte man die Infrastruktur dynamisch an den Bedarf anpassen und sowohl eine zu starke Unter- als auch eine Überprovisionierung vermeiden.

Zudem ist noch mehr Transparenz der Anbieter nötig (siehe ix.de/zd2e). Es ist zu begrüßen, dass die Anbieter erste CO_2-Verbrauchs-Dashboards für einige ihrer eigenen Dienste anbieten. Für einen unternehmensweiten Klimabericht gerade bei einer Multi-Cloud-Nutzung müssen sich die dort anfallenden Daten aber standardisieren und weiterverarbeiten lassen. Bisher gibt es dafür noch wenige ausgereifte Werkzeuge. Deshalb kann man sich mit dem Open-Source-Werkzeug Cloud Carbon Footprint selbst eine erste Grundlage schaffen, um einen Überblick über den eigenen CO_2-Verbrauch zu gewinnen. (sun@ix.de)

Quellen

[1] Hubert Sieverding; Aufbruch; Gigabytes Rack-Server R152-P30 mit 80 ARM-Kernen im Test; iX 1/2022, S. 42

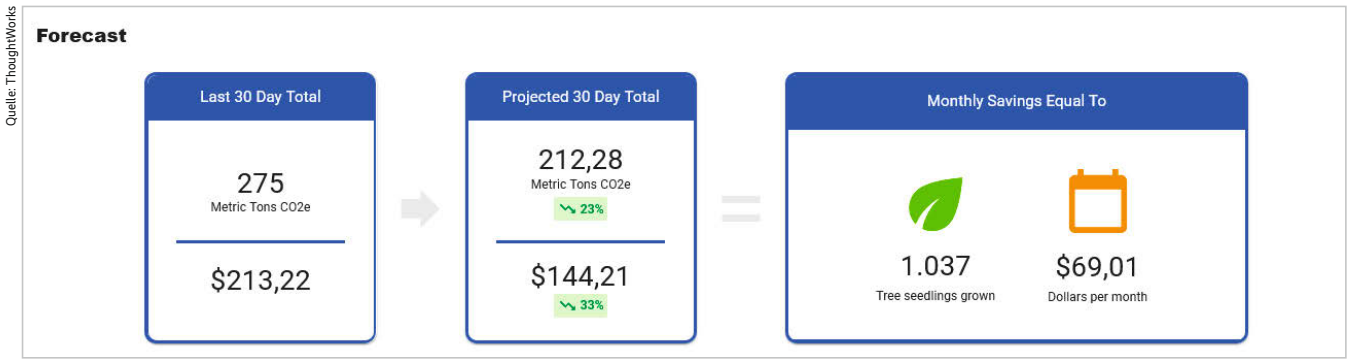
[2] Hubert Sieverding; Viele Wege, ein Ziel; ARM-Server im Überblick; iX 1/2022, S. 50

[3] Hubert Sieverding; Fernduell; ARM-Server aus der Cloud im Vergleich; iX 2/2022, S. 70

[4] Empfehlungen und weiterführende Literatur: ix.de/zd2e

Frank Pientka

begleitet als zertifizierter und erfahrener Architekt die Kunden bei ihrer erfolgreichen Reise in der Cloud.



Quelle: ThoughtWorks

Cloud Carbon Footprint gibt auch grobe Empfehlungen und berechnet die dadurch erzielten Einsparungen an Kosten und CO_2-Emissionen (Abb. 5).

Ökozertifikate für nachhaltige Rechenzentren

Dickicht der anderen Art

Dr. Béla Waldhauser

Eine breite Palette von Ökosiegeln für umweltgerechte Gebäude, Betriebe und Rechenzentren zertifizieren entweder einen kontinuierlichen Verbesserungsprozess oder das Erfüllen von Ausschlusskriterien.

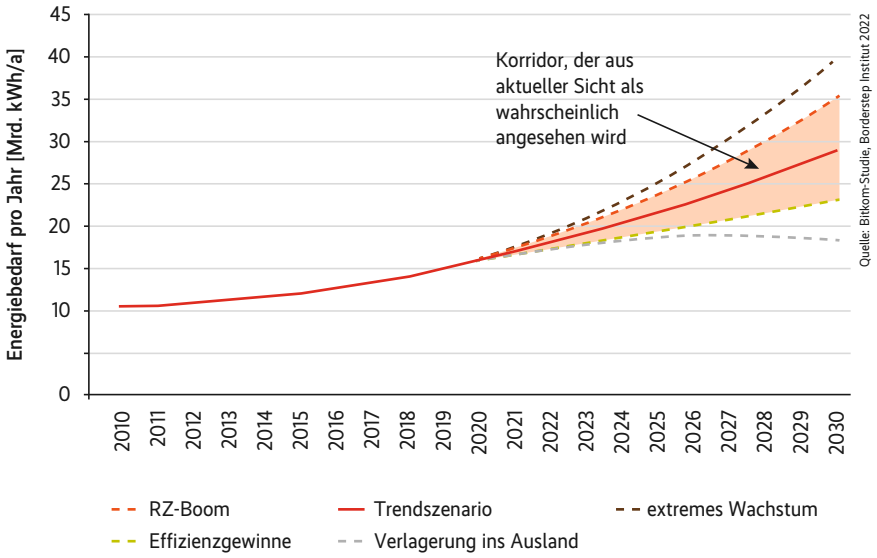
Dass sich der derzeitige Stromverbrauch deutscher Rechenzentren in den nächsten 10 Jahren verdoppelt, liegt durchaus im Bereich des Möglichen. Alle Prognosen prophezeien ein Wachstum, das sich allerdings je nach Modell stark unterscheidet (siehe Abbildung 1). Kein Wunder also, dass die neue Bundesregierung Themen wie Energieeffizienz, Nachhaltigkeit und Klimaneutralität der digitalen Infrastrukturen in ihren Koalitionsvertrag aufgenommen hat.

Zumindest bei der Energieeffizienz und dem daraus resultierenden Stromverbrauch gibt es eine hohe intrinsische Motivation der Rechenzentrumsbetreiber, wettbewerbsfähig zu sein und zu bleiben – allein schon aus Kostengründen, befeuert von den steigenden Energiepreisen. Zudem stieg in den letzten Jahren der Druck der Kunden, aber auch der Politik, die Energieeffizienz und Nachhaltigkeit zu erhöhen und messbare Ziele zu erreichen. Was wäre da einfacher, als eine anerkannte Zertifizierung als Nachweis zu nehmen. Allerdings war dies in den vergangenen Jahren nicht ganz so einfach. Die Geschichte der verschiedenen Ökozertifikate für Rechenzentren ist nämlich bei Weitem nicht so alt wie die der Rechenzentren (siehe Kasten „Die Rechenzentrumsbranche“). Man hat den Eindruck, dass viele der rechenzentrumsspezifischen Zertifizierungen und Normen erst in den letzten 15 Jahren entstanden sind, und zwar unter dem Eindruck des steigenden Energieverbrauchs.

Building Research Establishment Environmental Assessment Methodology

Zertifikate für Gebäude und ganze Unternehmen sind bereits älter. Aus dem Jahr 1990 stammt die BREEAM (Building Research Establishment Environmental Assessment Methodology).

In allen Szenarien steigt der künftige Energiebedarf deutscher Rechenzentren. Nur im Fall, dass RZ-Ressourcen massiv ins Ausland verlagert werden, würde sich die Entwicklung zumindest in Deutschland in etwa 5 Jahren leicht umkehren (Abb. 1).



BREEAM ist ein Bewertungssystem für ökologische und soziokulturelle Aspekte der Nachhaltigkeit von Gebäuden und stammt ursprünglich aus Großbritannien. Zertifizieren lassen sich neue und ältere Gebäude, aber auch Gebäude, die renoviert werden.

Das System wurde vom britischen Forschungsinstitut BRE (Building Research Establishment) entwickelt und ist vor allem in Großbritannien, aber auch im angloamerikanischen Raum verbreitet. Mittlerweile hat es über 600 000 Zertifikate verliehen und mehr als 2 Millionen Gebäude in 93 Ländern registriert. Da BREEAM für jedwede Art von Gebäuden gilt, fehlt ihm die Tiefe, den spezifischen Betrieb von Rechenzentren angemessen zu betrachten.

Leadership in Energy and Environmental Design

Das Gleiche gilt mehr oder weniger für LEED (Leadership in Energy and Environmental Design), das das U.S. Green Building Council 1998 entwickelt hat. LEED ist ein System zum Klassifizieren ökologischer Gebäude und definiert eine Reihe von Standards für umweltfreundliches, ressourcenschonendes und nachhaltiges Bauen. Auch bei LEED fehlt die Detailtiefe für die Rechenzentrumsbranche.

Dennoch sind beide Zertifikate im angloamerikanischen Raum anerkannt und werden dort gern von Rechenzentrumsbetreibern angestrebt. Die Verbreitung im deutschsprachigen Raum ist dagegen gering. Das kann sich allerdings in den nächsten Jahren ändern, da das Thema Nachhaltigkeit immer umfassender betrachtet wird und sich nicht mehr nur auf den Stromverbrauch und die PUE (Power Usage Effectiveness) beschränkt.

iX-TRACT

- Der Stromverbrauch deutscher Rechenzentren steigt stetig an. Ökozertifizierungen sollen helfen, den Energiehunger zu drosseln und die Umweltverträglichkeit zu erhöhen.
- Neben Zertifizierungen für Gebäude und Betriebe gibt es auch solche, die den spezifischen Betrieb von Rechenzentren abbilden.
- Die EMAS III, die ISO 14001 und ISO 50001 spezifizieren einen kontinuierlichen Verbesserungsprozess und die EN 50600/ISO 30134-x arbeitet mit einem Reifegradmodell. Anders als sie stellt der Blaue Engel für Rechenzentren harte Ausschlusskriterien auf.

Das bedeutet, dass das Gebäude selbst in Zukunft genauso viel Beachtung bekommen könnte wie der RZ-Betrieb. Deshalb ist es ratsam, sich vorab mit den Anforderungen von BREEAM und/oder LEED auseinanderzusetzen und sie in das Design und die Architektur neuer Rechenzentren einfließen zu lassen. Denn der CO_2-Fußabdruck beim Bau eines Rechenzentrumsgebäudes ist nicht zu vernachlässigen und bietet einiges an Verbesserungspotenzial.

EMAS Regulation 1836/93

1993 stellte die EU-Kommission die EMAS (Eco-Management and Audit Scheme) Regulation 1836/93 vor und öffnete sie im April 1995 für eine freiwillige Teilnahme. Die EMAS durchlief seitdem mehrere Änderungszyklen und ist seit Januar 2010 in der Version EMAS III gültig. Die EMAS wirbt damit, dass teilnehmende Organisationen und Unternehmen ihre Umweltauswirkungen systematisch erfassen, kontinuierlich die Umweltleistung verbessern und damit die Effizienz steigern und Kosten senken können (siehe Abbildung 2).

Die EU-Verordnung wurde bereits 1996 in die ISO 14001 überführt und hat seitdem eine globale Reichweite. Der kleine Bruder der ISO 14001, die ISO 50001, erschien allerdings erst 2011. Wie BREEAM und LEED sind auch die EMAS-Verordnung und beide ISO-Zertifizierungen nicht speziell für Rechenzentren konzipiert, aber dafür EU- respektive weltweit anerkannt und geschätzt.

ISO 14001 und ISO 50001

Die ISO 14001 ist ein umfassendes Umweltmanagementsystem, während sich die ISO 50001 auf das Energiemanagement beschränkt. Die ISO 50001 bezieht also jegliche Form des Energiebedarfs mit ein und berücksichtigt den Verbrauch von Strom, Gas, Kraftstoffen et cetera. Die ISO 14001 ist breiter aufgestellt und betrachtet das komplette Umweltmanagement, also zusätzlich zum Energieverbrauch den Wasserverbrauch, das Abfallmanagement, Hilfs- und Betriebsstoffe und weitere Faktoren.

Beide Managementsysteme durchlaufen einen kontinuierlichen Verbesserungsprozess (KVP), der anhand des PDCA-Zyklus (Plan – Do – Check – Act) organisiert und abgebildet wird.

Sowohl KVP als auch PDCA-Zyklus sind aus der ISO 9001 hinlänglich bekannt und bringen verschiedene Vorteile, aber auch Pflichten mit sich. Zum einen wird den Betreibern die Pflicht auferlegt, sich kontinuierlich zu verbessern. Die Verbesserungspotenziale und Ziele kann sich das Unternehmen – mehr oder weniger – selbst aussuchen, muss diese dann aber auch dokumentieren und optimalerweise erreichen. Gerade Betreibern älterer Rechenzentren bietet dies die Möglichkeit, sich sukzessive zu verbessern, ohne gleich bei der Erstzertifizierung mit rigiden oder zu hohen Schwellenwerten konfrontiert zu werden.

Eine weitere Motivation für die Zertifizierung nach ISO 50001 ist die seit 2015 für alle Nicht-KMU geltende Verpflichtung, regelmäßig ein Energieaudit nach DIN EN 16247-1 durchzuführen und nachzuweisen. Die Verpflichtung ist im Energiedienstleistungsgesetz (EDL-G) verankert und betrifft auch KMU, sofern sie einem Konzern angehören, der kein KMU ist. Davon sind in Deutschland immerhin rund 50 000 Unternehmen betroffen.

Die Zertifizierung nach ISO 50001 ist dem gleichgestellt und daher für viele Rechenzentren ein guter Grund, sich mit diesem Zertifikat auszeichnen zu lassen. Der hohe Verbreitungsgrad von EMAS respektive ISO 14001 und ISO 50001 sowie die konzeptionelle Anlehnung an die ISO 9001 ist sicherlich für viele Unternehmen eine große Motivation, eines oder mehrere dieser Zertifikate anzustreben.

Als Kritikpunkt bleibt allerdings, dass beide Managementsysteme keine quantitative Differenzierung durchführen. Sie betrachten beispielsweise den Stromverbrauch eines Rechenzentrums genauso wie den Kraftstoffverbrauch einiger weniger Firmenwagen. Hier wäre eine Fokussierung auf die wesentlichen Einflussfaktoren im Bereich der energetischen Leistung oder auf den Umwelteinfluss hilfreich.

Die Rechenzentrumsbranche

Rechenzentren, Energieeffizienz, Nachhaltigkeit, Klimaneutralität und die entsprechenden Zertifikate sind nichts Neues. Bereits 1961 wurde das Deutsche Rechenzentrum in Darmstadt gegründet und stand der deutschen Forschung zur Verfügung. Die deutschen Steuerberater gründeten 1966 die DATEV eG und 3 Jahre später wurde deren erstes Rechenzentrum in Betrieb genommen. Damals übrigens mit einem der ersten Cloud-Angebote weltweit.

Mit der dritten industriellen Revolution in den 1970er-Jahren wurden Rechenzentren auch in deutschen Unternehmen und Behörden etabliert und entwickelten sich dank des mooreschen Gesetzes mit rasanter Geschwindigkeit. Aber erst das Web in den 1990er-Jahren, die Liberalisierung der Telekommunikation in den Jahren 1996 und 1998, die Digitalisierung in den 2000ern und die vierte industrielle Revolution haben zur aktuellen Entwicklung der Rechenzentren weltweit geführt und bescheren insbesondere der jungen Colocation- und der noch jüngeren Cloud-Branche ungeahnte Wachstumsmöglichkeiten.

Die Geburtsstunde der Colocation-Branche fällt zusammen mit dem zweiten Schritt der Liberalisierung der Telekommunikation und dem Dotcom-Boom im Jahre 1998. Die Cloud-Branche, insbesondere Infrastructure as a Service, war dann fast ein Jahrzehnt später dran; der IaaS-Weltmarktführer AWS wurde überhaupt erst 2006 gegründet. Nicht zuletzt die Verbreitung von Smartphones, beginnend mit der Markteinführung des iPhone 1 im Jahr 2007, hat zur großen Nachfrage an Rechenzentrumskapazitäten beigetragen.

Dies spiegelt sich auch im Stromverbrauch der Rechenzentren in Deutschland wider. Von 2010 bis 2020 ist der Stromverbrauch der deutschen Rechenzentren von 10,5 auf 16 TWh um gut 50 Prozent gestiegen. Der prozentuale Anteil der IT – also Server, Speicher und Netz – stieg in dem Zeitraum sogar auf über 70 Prozent.

Heute unterscheidet man vier Klassen von Rechenzentren. Unter den Oberbegriff der Unternehmensrechenzentren fallen die RZs von Firmen, aber auch die von Behörden, Universitäten und Forschungseinrichtungen. Daneben existieren seit Längerem die Colocation-Rechenzentren, deren Betreiber sich im Wesentlichen um Strom, Kühlung und physische Sicherheit kümmern. Ihre Kunden richten dort die eigene Hard- und Software ein. Die dritte Klasse von Rechenzentren umfasst die Hosting- oder Cloud-RZs, deren Betreiber unterschiedlichste IT-Dienstleistungen anbieten. Noch ganz am Anfang stehen die Edge-Rechenzentren. Ihre Zahl wird aber in den nächsten Jahren sprunghaft zunehmen, zusammen mit der Verbreitung von 5G, Smart Home, Smart City und Industrie 4.0.

EU Code of Conduct for Energy Efficiency in Data Centres

Dem steigenden Energieverbrauch der Rechenzentren hat auch die EU-Kommission Rechnung getragen und 2008 den EU Code of Conduct for Energy Efficiency in Data Centres ins Leben gerufen. Der Code of Conduct ist ausdrücklich freiwillig und basiert im Wesentlichen auf Best-Practice-Ansätzen und den dazugehörigen Energieeffizienzzielen. Für den EU Code of Conduct gibt es keine Zertifizierung im üblichen Sinn, wie man sie von der ISO 50001 oder der ISO 14001 kennt. Man kann sich aber sehr wohl von unabhängiger Stelle die Konformität mit dem EU Code of Conduct bescheinigen lassen und damit werben.

Noch genießt der EU Code of Conduct bei vielen deutschen Unternehmen keinen signifikanten Stellenwert, deshalb wird die Konformität damit selten angestrebt. Allerdings fragen mehr und mehr Unternehmen aus dem Ausland danach. Daher sollte man den EU Code of Conduct als Rechenzentrumsdienstleister nicht vernachlässigen oder eine international anerkannte Alternative anstreben.

Blauer Engel DE-UZ 161 und DE-UZ 214

Ein weiteres Zertifikat, das sich ausdrücklich an Rechenzentren richtet, ist der Blaue Engel DE-UZ 161. In Deutschland seit vielen Jahrzehnten bekannt für Konsumgüter, gibt es den Blauen Engel seit 2011 auch für Rechenzentren. Wie der EU Code of Conduct for Energy Efficiency in Data Centres wurde der Blaue Engel ausschließlich für Rechenzentren konzipiert.

In der Version DE-UZ 161 richtet er sich allerdings ausschließlich an Unternehmen, die nicht nur das Rechenzentrum selbst betreiben, sondern auch die dort installierte Hard- und Software. Das machte es insbesondere für die Colocation-Branche unmöglich, ihn zu erreichen. Anfänglich schrieb der Blaue Engel einen hohen Virtualisierungsgrad der Server vor; mittlerweile ist man auf eine CPU-Auslastung von mindestens 20 Prozent umgeschwenkt. Damit will man die Effizienz vorgeben, aber nicht unbedingt den Weg.

Beides ist sicherlich sinnvoll, aber für Betreiber von Colocation-Rechenzentren weder steuerbar noch nachweisbar. Um auch der stark wachsenden Colocation-Branche den Zugang zum Blauen Engel zu ermöglichen, hat das Umweltbundesamt nachgelegt. Seit 2020 gibt es den Blauen Engel auch in der schlankeren Version DE-UZ 214 für Colocation-Rechenzentren.

Die Anforderungen, um den Blauen Engel zu bekommen, sind hoch – das gilt für beide Versionen. Beispielsweise ist bereits bei Antragstellung ein Energieeffizienzbericht vorzulegen, der die zu prüfenden Kriterien beziehungsweise Anforderungen plausibel dokumentiert. Des Weiteren ist ein Energiemanagementsystem vorgeschrieben, das angelehnt ist an die ISO 50001 oder EMAS III. Wichtige Kriterien für den Blauen Engel DE-UZ 161 sind unter anderem die EUE (Energy Usage Effectiveness), der Anteil an erneuerbarem Strom, die Auslastung wesentlicher IT-Komponenten wie CPU, RAM und Storage, die JAZ (Jahresarbeitszahl) des Kühlsystems und der Wirkungsgrad der USV (unterbrechungsfreien Stromversorgung), um nur einige zu nennen. In der Version DE-UZ 214 fallen die Kriterien für die IT komplett weg. Alle anderen Vorgaben sind aber mehr oder weniger identisch, außer dass der EUE durch den PUE ersetzt wird.

Ausschlusskriterien statt kontinuierliche Verbesserungsprozesse

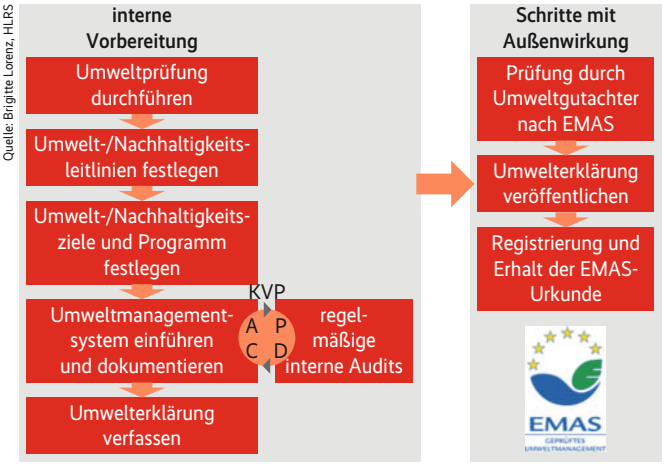
Ein zentrales Kriterium des Blauen Engels in beiden Versionen ist das verwendete Kältemittel: Alle Rechenzentren, die nach dem 01.01.2013 in Betrieb gegangen sind, dürfen ausschließlich halogenfreie Kältemittel einsetzen. Gerade an diesem Kriterium scheiden sich oft die Geister. Die Branche ist mehrheitlich der Meinung, dass es für große Rechenzentren lange keine sinnvollen Alternativen für halogenierte Kältemittel gab, wohingegen das Umweltbundesamt beim Blauen Engel auf halogenfreien Kältemitteln als K.-o.-Kriterium besteht. Das mag einer der Gründe dafür sein, dass sich der Blaue Engel in deutschen Rechenzentren noch nicht durchsetzen konnte. Mittlerweile setzen aber immer mehr Rechenzentrumsbetreiber auf adiabatische Kühlung, wodurch dieser Konflikt – zumindest für diese – obsolet wird.

Aber nicht nur beim Kältemittel stellt der Blaue Engel in beiden Versionen harte Bedingungen an alle, die dieses Zertifikat anstreben. Diese Ausschlusskriterien ziehen sich wie ein roter Faden durch die beiden Anforderungskataloge, ob bei der CPU-Auslastung, der EUE, der PUE oder der JAZ. Weder der kontinuierliche Verbesserungsprozess einer ISO 14001 oder ISO 50001 noch ein Reifegradmodell wie das der EN 50600 genügt dem Anspruch des Blauen Engels. Beides würde die anfängliche Hürde zum Erreichen des Blauen Engels etwas heruntersetzen und damit mehr Unternehmen und Dienstleister motivieren, dieses Zertifikat zu erreichen.

Nicht nachvollziehbar ist allerdings die immer noch sehr geringe Verbreitung des Blauen Engels bei den Rechenzentren der öffentlichen Hand. Gerade dort wäre eine wesentlich höhere Akzeptanz zu erwarten. Derzeit plant das Umweltbundesamt eine Zusammenführung der beiden Umweltzeichen DE-UZ 161 und 214. Diese Zusammenführung birgt die Chance, auch unterschiedliche Stufen beziehungsweise Reifegrade einzuführen und damit den Blauen Engel für Rechenzentren etwas attraktiver zu machen.

EN 50600-4-x

Bei europäischen Rechenzentren nimmt derzeit die EN 50600 eine führende Rolle ein. Sie wurde von, mit und für die Rechenzentrumsbranche entwickelt und gibt detaillierte Vorgaben zum Erreichen der vier Verfügbarkeitsklassen, analog den vier Tier-Standards des Uptime Institute.



Die EMAS-Zertifizierung zielt auf die kontinuierliche Verbesserung der Umweltleistung (Abb. 2).

Bei den Themen Energieeffizienz und Nachhaltigkeit steht diese Norm noch am Anfang. Mit der EN 50600-4-x wurde 2016 aber ein großer und wichtiger Schritt in diese Richtung gemacht. Dieser Teil der EN 50600 beschäftigt sich ausschließlich mit einer klaren Definition der relevanten KPIs (Key Performance Indicators) für Rechenzentren (siehe Abbildung 3). Er behandelt sehr detailliert KPIs für IT- und Netzwerkkomponenten, Stromverteilungs-, Monitoring- und jegliche Sicherheitsinfrastruktur. Damit schließt die EN 50600-4-x eine Lücke, die lange Jahre immer wieder unterschiedlichste Interpretationen hervorrief.

Die EN 50660-4-x ist zudem für alle möglichen Arten, Größen und Verfügbarkeitsklassen von Rechenzentren anwendbar. Durch diese Vielseitigkeit ist sie universell einsetzbar. Beispielsweise bietet sie eindeutige Definitionen für die PUE, den REF (Renewable Energy Factor), die $ITEE_{SV}$ (IT Equipment Energy Efficicency for servers), die $ITEU_{SV}$ (IT Equipment Utilization for servers), den ERF (Energy Reuse Factor), die CER (Cooling Efficiency Ratio), die CUE (Carbon Usage Effectiveness) und die WUE (Water Usage Effectiveness). Darüber hinaus bietet die EN 50600-4-x auch Umrechnungsfaktoren für Energie, die entweder aus Diesel, Gas, Wasserstoff oder Bioethanol gewonnen wird. Daher ist die EN 50600-4-x die notwendige, aber auch hinreichende Grundlage für die folgende EN 50600-5-1.

EN 50600-5-1, CLC/TR 50600-99-1 und CLC/TR 50600-99-2

Denn erst die EN 50600-5-1 stellt zusammen mit der CLC/TR 50600-99-1 und der CLC/TR 50600-99-2 eindeutige Anforderungen an die Reduktion der Umweltauswirkungen, das heißt an das Energiemanagement und die Umweltverträglichkeit der Rechenzentren. Sie lassen sich wiederum im bekannten Reifegradmodell sukzessive erreichen. Unter die Lupe genommen werden die Auslegung, die Beschaffung, der Betrieb und das Ende der Lebensdauer eines Rechenzentrums, also der komplette Lebenszyklus des Rechenzentrums und seiner Infrastruktur.

Analog zu den Normen EN 50600-1, EN 50600-2-x, EN 50600-3-1 und EN 50600-4-x behandelt auch die EN 50600-5-1 alle Komponenten in einem Rechenzentrum. Sie betrachtet also sowohl die Strom- und Kühlungs- als auch die IT-Infrastruktur, außerdem die Sicherheits- und Monitoringinfrastruktur.

Normen, Zertifikate und Empfehlungen

Norm	Veröffentlichung	Schwerpunkt
BREEAM	1990	Bewertungssystem für ökologische und soziokulturelle Aspekte der Nachhaltigkeit von Gebäuden
LEED	1998	Managementsystem für umweltfreundliches, ressourcenschonendes und nachhaltiges Bauen
EMAS-Verordnung	1995	Managementsystem aus Umweltmanagement und Umweltbetriebsprüfung
ISO 14001	1996	Managementsystem für Umwelt-, Energie- und Ressourceneffizienz von Unternehmen
ISO 50001	2011	Managementsystem für Energieeffizienz von Unternehmen
Code of Conduct	2008	Best Practices für energieeffizienten Rechenzentrumsbetrieb (IT und TGA*)
Blauer Engel – UZ 161	2011	Energie- und Ressourceneffizienz von Rechenzentren im Eigenbetrieb (IT und TGA*)
Blauer Engel – UZ 214	2020	Energie- und Ressourceneffizienz von Colocation-Rechenzentren (nur TGA*)
DIN EN 50600-4-x ISO/IEC 30134-x	2016	Definition KPIs für Rechenzentren (IT und TGA*)
DIN EN 50600-5-1	2021	Reifegradmodell für Energie- und Ressourceneffizienz im Rechenzentrum (IT und TGA*)
TSE.Standard 2.0	2021	Managementsystem für Umwelt-, Energie- und Ressourceneffizienz von Rechenzentren

*TGA = technische Gebäudeausrüstung

Aus Umweltsicht bleiben daher keine Wünsche offen. Allerdings existiert die EN 50600-5-1 erst seit Dezember 2021, also gerade mal ein halbes Jahr. Es bleibt abzuwarten, wie stark sie sich verbreiten wird.

ISO/IEC 22237 und ISO 30134

Anlass zur Skepsis gibt die Tatsache, dass die EN 50600 schrittweise von der ISO (International Organization for Standardization) übernommen wird und damit Einzug ins ISO-Regelwerk hält. So wurde beziehungsweise wird aus der EN 50600-1 bis -3-1 die ISO/IEC 22237-1 bis ISO/IEC 22237-7. Die für die Ökozertifizierung relevanten Normen EN 50600-4-x und EN 50600-5-1 gehen über in die ISO 30134-x.

Damit ist die in Europa anerkannte EN 50600 auch auf dem globalen Parkett fest etabliert. Es bleibt also abzuwarten, ob sich die Rechenzentren in Deutschland beziehungsweise in Europa eher für die EN 50600 oder die ISO 22237 respektive ISO 30134 entscheiden werden. Inhaltlich gibt es keinen Unterschied.

Allerdings bildet die EN 50600 nur die Grundlage für die Zertifizierung. Man benötigt einen qualifizierten Zertifizierer, der anhand seines Prüfkatalogs die Konformität des Rechenzentrums mit den Anforderungen der EN 50600-5-1 bestätigt. Hier hat man die Auswahl aus zahlreichen Auditoren, unter anderem den verschiedenen TÜVs.

TSE.Standard 2.0

Eine Zertifizierung nach den Normen EN 50600-5-1, CLC/TR 50600-99-1 und -99-2 bietet zum Beispiel der TÜViT mit seinem Prüfkatalog TSE.Standard 2.0 (Trusted Site Energy Efficiency) an. Hier kann sich der Rechenzentrumsbetreiber von einer unabhängigen Prüfstelle bestätigen lassen, dass er die Anforderungen für das Reifegradmodell der EN 50600-5-1 erfüllt. Eine Übersicht über die unterschiedlichen Normen, Zertifikate und Empfehlungen samt Jahr der Erstveröffentlichung und jeweiligem Schwerpunkt gibt die Tabelle „Normen, Zertifikate und Empfehlungen“.

Fazit

Mittlerweile gibt es eine ausreichende Auswahl unterschiedlichster Ökozertifikate, aus denen sich RZ-Betreiber das für sie sinnvollste aussuchen können. Der geringste Aufwand ist vielleicht bei der ISO 50001 zu erwarten, sodass sie bei manchen Rechenzentrumsbetreibern erste Wahl sein wird. Mit etwas mehr Aufwand kann man die IS0 14001 bekommen, ohne gleich bei der Erstzertifizierung an feste Vorgaben bei den KPIs gebunden zu sein. Beiden gemeinsam ist der kontinuierliche Verbesserungsprozess, aufgrund dessen sich der Betreiber überlegen muss, wie er jedes Jahr seinen Betrieb und seine Dienstleistung effizienter und nachhaltiger anbieten kann.

Ein weiterer Vorteil beider Zertifikate ist die globale Verbreitung und damit Anerkennung. Insbesondere viele Colocation-Rechenzentren haben einen Anteil ausländischer Kunden von mehr als 50 Prozent. Da ist ein global gültiges und anerkanntes Zertifikat unabdingbar.

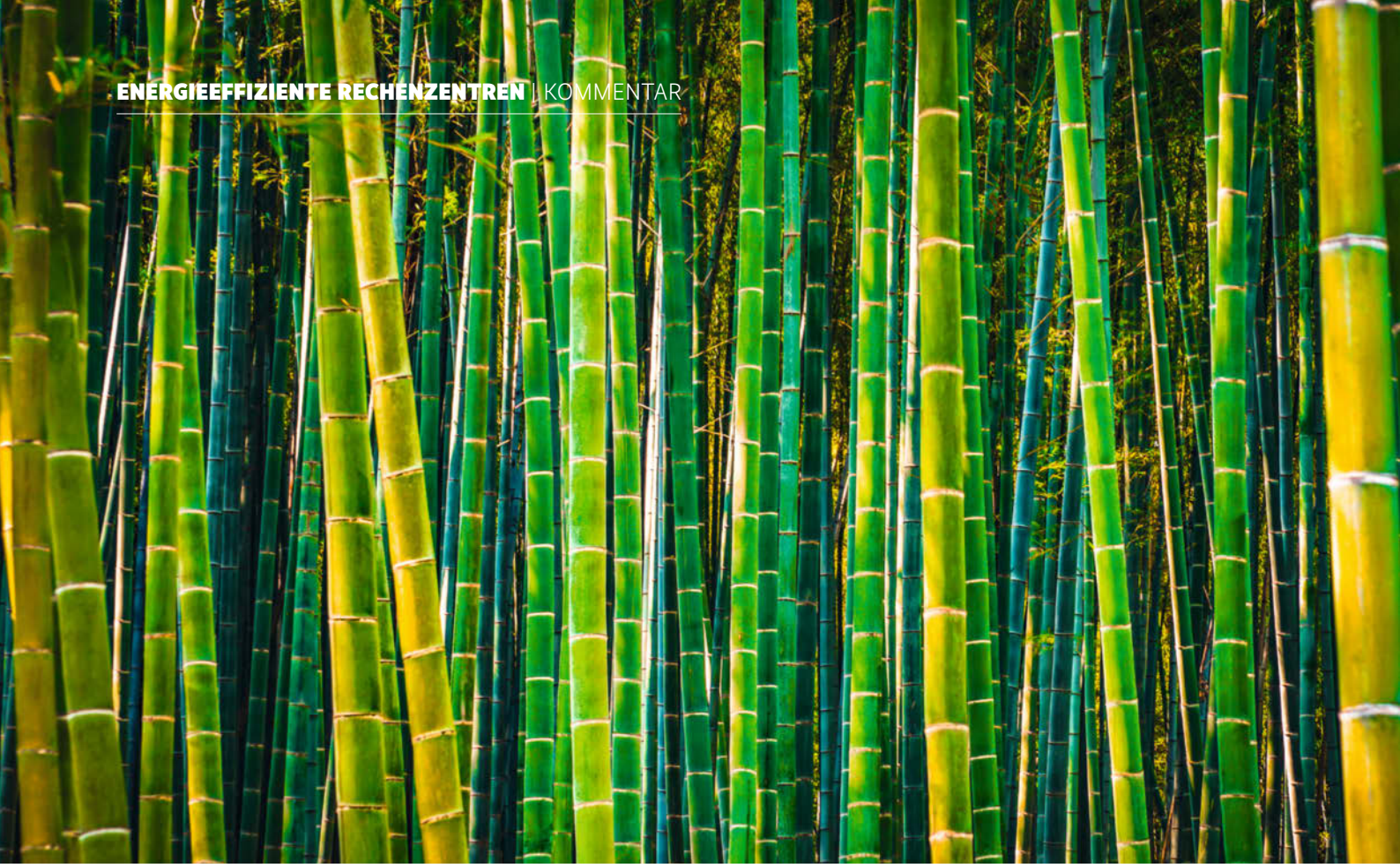
Genau das wird zukünftig der größte Hemmschuh für die Verbreitung des Blauen Engels sein. Als Symbol für besonders umweltschonende Produkte und Dienstleistungen genießt er zwar im deutschsprachigen Raum einen sehr guten Ruf und ist der breiten Öffentlichkeit wohlbekannt. Jenseits dessen hat er allerdings keinerlei Bedeutung und ist daher für viele Rechenzentrumsbetreiber keine Alternative.

Vermutlich werden mehr und mehr Betreiber auf die EN 50600 respektive die ISO 30134-x setzen. Viele haben sich bereits heute die Konformität mit der EN 50600-1 bis EN 50600-3-1 oder ISO/IEC 22237-x bestätigen lassen, da immer mehr Kunden aus den verschiedensten Branchen dies erwarten. Da ist es nur logisch, dass der Betreiber in der Logik des Reifegradmodells bleibt und seine Konformität um die EN 50600-4-x/ISO 30134-x und EN 50600-5-1 ergänzt.

Bei der Diskussion um das richtige Ökozertifikat darf man nicht vergessen, dass ein Rechenzentrumsbetreiber noch eine Reihe weiterer Zertifikate benötigt, um attraktiv für seine Kunden zu sein. Je nach Zielgruppe gehören dazu zum Beispiel ISO/IEC 27001, ISEA3402, PCI-DSS, ISO/IEC 50001 oder ISO/IEC 14001, EN 50600, aber auch die B3S für diejenigen, die zur kritischen Infrastruktur der Bundesrepublik Deutschland gehören. All diese Zertifikate, Nachweise und Audits kosten viel Zeit, Geld und Ressourcen. Daher wird der verantwortliche Geschäftsführer oder Vorstand sich sehr genau überlegen, welche Ökozertifikate er sich zusätzlich leisten kann und will. (sun@ix.de)

Dr. Béla Waldhauser

ist CEO der Telehouse Deutschland GmbH, Leiter der Kompetenzgruppe Datacenter im eco – Verband der Internetwirtschaft e. V. und aktives Mitglied der Policy Group der EUDCA.



Welche Bezugspunkte braucht eine Umweltkennzahl?

Maß-Stäbe setzen

Hubert Sieverding

Mit dem Hinzufügen unterschiedlichster Messgrößen kann man viel kaschieren und schönen, so auch die Energieeffizienz von Servern. Doch welche Bezugswerte haben solche Prospektaussagen und was sagen sie aus? Ein Kommentar.

Mit steigenden Energiekosten rückt die Energieeffizienz der IT-Hardware zunehmend in den Fokus. Doch wann ist ein Server, wann ein Rechenzentrum energieeffizient? Und wie misst man das? Helfen soll ein Blick über den Zaun, auf eine Branche, die einen ähnlich hohen CO_2-Fußabdruck zu verantworten hat: die Automobilindustrie. Sie hat sich längst auf ein weltweit einheitliches Messverfahren zum Ermitteln der Effizienz von Pkws geeinigt: die Worldwide Harmonised Light-Duty Vehicles Test Procedure (WLTP).

Die WLTP legt fest, wie der Energieverbrauch zu messen ist. Zu erbringen ist eine Dienstleistung, nämlich eine bestimmte Last – die Personen im Fahrzeug – unter strikter Vorgabe der Beschleunigung, Höchstgeschwindigkeit und Umgebung – Stadt, Landstraße, Autobahn – von A nach B zu befördern. Größe, Gewicht und Motorisierung des Fahrzeugs spielen dabei keine Rolle. Heraus kommt eine Zahl: CO_2/km oder kWh/km. Der Gesetzgeber nutzt den so ermittelten Flottendurchschnitt eines Herstellers als Vorgabe für Restriktionen.

Zusätzlich gibt es nach dem Vorbild der Elektroindustrie ein Energielabel, das Fahrzeuge in die Energieeffizienzklassen A bis G einteilt. Wer nun aber glaubt, hier würde der Energieverbrauch eines Fahrzeugs klassifiziert, der irrt. Ein Audi Q7 und ein VW up! gehören derselben Effizienzklasse B an, weil auf Druck der Automobillobby das Fahrzeuggewicht mit einfloss.

iX-TRACT

- Effizienzkennzahlen, die die Performance durch die Energieaufnahme teilen, setzen die Performance in den Mittelpunkt und nicht die Energieeffizienz. Um den Fokus auf die Effizienz zu legen, müsste man die Werte tauschen.
- Bemühungen, eine allgemeingültige Kennzahl für die Energieeffizienz zu finden, erliegen gern der Versuchung, auf bestehende, aber nicht vergleichbare Messergebnisse zurückzugreifen.
- Vor allem die gern referenzierten und oft von Herstellern selbst erzeugten Benchmarkergebnisse auf spec.org haben wenig mit dem gemein, was produktive Systeme im RZ zu leisten imstande sind.
- Zu wenig Berücksichtigung findet zudem die ungenügende Auslastung von CPUs und Netzteilen.

Logisch nicht nachvollziehbar – aber gut für den Werbeprospekt der Hersteller.

Übertragen auf die IT hieße dies für Server: Es gäbe einen weltweit einheitlichen Benchmark, der genau festlegt, wie der Energieverbrauch für eine IT-Dienstleistung zu messen ist, unabhängig von der Leistung des Systems. Er würde messen, wie viel Energie ein System für eine vorgegebene Berechnung oder das Bewegen einer bestimmten Datenmenge durchs Netz benötigen würde. Beim Datentransport wäre sogar vorgegeben, mit welchem Durchsatz er zu erfolgen hätte. Heraus käme eine Zahl: Energie/normierte IT-Leistung.

Mit dem produktiven Betrieb hätte dies so wenig zu tun wie der tatsächliche Verbrauch eines Pkw mit den Prospektaussagen. Doch der Gesetzgeber würde den Durchschnittswert für die Besteuerung von Rechenzentren heranziehen. Lediglich in den Prospekten der Server gäbe es zusätzlich ein Energielabel A bis G. Dort allerdings würde die CPU-Performance mit einbezogen und ein Supercomputer mit eigenem Kraftwerk wäre derselben Effizienzklasse B zugeordnet wie der Ein-Sockel-Rackserver mit Standard-CPU. Eine Dystopie?

Performance per Watt

Supercomputer werden seit Jahren nicht nur anhand ihrer Performance verglichen, sondern auch hinsichtlich ihres Stromverbrauchs: Performance pro Watt lautet das Schlagwort. Dies ist insofern einfach, als die in GFlops (Giga Floating Points per Second) ausgegebenen Ergebnisse des Linpack-Benchmarks durch den abgelesenen Stromverbrauch der Rechenknoten dividiert werden: GFlops/W. Storage, Stromversorgung und Kühlung etwa spielen dabei keine Rolle.

IT-typisch ist auch die Betonung dieser Messgröße: Im Zähler steht die Leistung, nicht der Energieverbrauch. Die IT-Industrie sucht und erfindet gern performanceabhängige Kennzahlen. In Anbetracht vollgestellter RZ-Hallen brachte Sun Microsystems vor Jahren sogar die Stellfläche als zusätzliche Dimension mit ein und definierte:

$$SWap = \frac{Performance}{Space \times Power}$$

Vergleichbar mit dem Flottendurchschnitt bei Autos definiert die Norm ISO/IEC 30134-5 „Information technology – Data centres – Key performance indicators“ zwei KPIs (Key Performance Indicators) für Rechenzentren. Der erste, $ITEE_{SV}$ (IT Equipment Energy Efficiency for servers), ergibt sich, wenn man die Summe aller SMPE- durch die Summe aller SMPO-Werte eines Rechenzentrums teilt:

$$ITEE_{SV} = \frac{\sum_{i=1}^{N} SMPE_{i,\,max}}{\sum_{i=1}^{N} SMPO_{i,\,max}}$$

Dabei ist N die Anzahl aller Server im RZ, $SMPE_{i,\,max}$ die maximale Performance des Servers i in Ops (Operations per Second), vom Hersteller gemessen, und $SMPO_{i,\,max}$ die maximale elektrische Leistungsaufnahme des Servers i in W_{el}, ebenfalls gemessen vom Hersteller. Wie genau der Benchmark auszusehen hat, wird nicht definiert. Ergo: ein Flottendurchschnitt nach dem Gusto der Hersteller. Auch interessant: Im Fokus steht die maximale Performance der CPU, und die steht im Zähler. Der zweite KPI, $ITEU_{SV}$ (IT Equipment Utilization for servers), hingegen beschreibt die durchschnittliche Auslastung der CPUs aller Server eines Rechenzentrums in einem Jahr. Er soll der RZ-Betreiber per Monitoring ermitteln.

Beide CPU-lastigen Indizes zusammen eignen sich für eine Effizienzaussage eines RZ, so die Norm. Beispiel: Ein RZ bietet 34 Ops/Watt und ist durchschnittlich zu 68 Prozent ausgelastet. Übertragen auf die Automobilindustrie: Die Fahrzeugflotte eines Anbieters liefert 1,3 PS/g CO_2 und ist zu 68 Prozent ausgelastet. Was aber sagt dieser KPI über den Verbrauch aus? Laut Norm sollen $ITEE_{SV}$ und $IREU_{SV}$ auch bei Entscheidungen zur Neubeschaffung und zur besseren Auslastung herangezogen werden. Nachfolgeserver sollten dabei einen höheren $ITEE_{SV}$ aufweisen.

Mögliche und beliebte Benchmarks

Als Benchmark zum Ermitteln des $ITEE_{SV}$ kommt unter anderem der SPECpower_ssj2008 infrage. Er ermittelt die Rechenleistung in ssj_ops (Server-Side Java Operations per Second) und die durchschnittliche Leistungsaufnahme mithilfe einer Reihe von Java-Programmen und eines externen Strommessgeräts, das die Benchmarksoftware periodisch ausliest. Dabei erhöht er die Last des Systems in 10-Prozent-Schritten von 0 bis 100 Prozent. Das Ergebnis gibt er aus in ssj_ops/Watt – für zehn Laststufen und für Active Idle.

Als Nachfolger des SPECpower_ssj2008 sieht die SPEC gern den SPECsert. Mit ihm steht ein Benchmark in den Startlöchern, der sich nicht auf den Durchsatz der Java-VM beschränkt, sondern auch Memory- und IO-intensive Tests umfasst und so ein weites Anwendungsspektrum abdeckt. Leider finden sich nur wenige Ergebnisse von Messläufen bei der spec.org, weshalb dieser Test kaum Beachtung findet.

Der Benchmark SPEC OMP 2012 nutzt die OpenMP-Bibliothek zum Messen des Multiprocessings einer leistungshungrigen Applikation mit optionaler Energiemessung. Auch SPEC virt_sc 2013 misst optional den Energiebedarf. Im Fokus steht aber die Performance von Virtualisierungsplattformen. Auch der SPECcpu2017, den die iX einsetzt, erlaubt es, während der Performancemessung den Energiebedarf mit einem externen Leistungsmessgerät zu ermitteln.

Die grundsätzliche Kritik an den SPEC-Benchmarks lautet, dass sie lizenzpflichtig sind, dass vor allem die Hersteller selbst die Tests durchführen und dabei tricksen. Beim SPECpower greifen die Hersteller zu Serverkonfigurationen und -optimierungen, die niemand in einem produktiven Rechenzentrum verwenden würde, geschweige denn, dass eine produktive Anwendung darauf überhaupt laufen würde.

Denn die Leistung der CPU hängt nicht nur von Takt und DRAM ab. Moderne Prozessoren steuern außerdem die PCIe-Lanes, und das zulasten der Performance und der Energieeffizienz. Gern verzichten die Hersteller bei optimierten Tests auch auf 10GE-NICs, ein zweites Netzteil und reduzieren den Storage oder die Zahl der Lüfter. Das rückt den gemessenen Server in ein besseres Licht – macht ihn performanter und energieeffizienter. Dieses Verfahren erinnert stark an die Prüfstandmessungen der Automobilindustrie.

Benchmarkergebnisse ohne Benchmark

Als Beispiel für den Missbrauch von Performanceangaben seien die SPECcpu2006-Werte für AMDs CPU Epyc genannt. Sie standen bereits fest, bevor das erste Exemplar überhaupt aus dem Wafer geschnitten wurde. Man hatte sie hochgerechnet

AMD Epyc 7702 in drei Systemen, gemessen mit SPECcpu2017 INTrate					
Auslastung	**12,5 %**	**25,0 %**	**50,0 %**	**81,3 %**	**100,0 %**
SPECcpu2017 INTrate					
System #1	65,50	119,00	184,00	202,00	214,00
System #2	66,00	118,00	188,00	205,00	217,00
System #3	69,30	123,00	189,00	207,00	218,00
Total Power (W)					
System #1	294,52 W	336,75 W	357,88 W	370,24 W	374,17 W
System #2	189,75 W	236,31 W	269,42 W	279,00 W	283,14 W
System #3	167,88 W	209,55 W	233,72 W	244,92 W	251,70 W
Verbrauch der CPU (W)					
System #1	118,32 W	162,53 W	179,94 W	186,53 W	188,40 W
System #2	114,82 W	156,95 W	180,64 W	184,29 W	185,81 W
System #3	111,22 W	145,96 W	162,44 W	166,41 W	169,67 W
Gesamtverbrauch (W)/SPEC					
System #1	4,50 W/SPEC	2,83 W/SPEC	1,95 W/SPEC	1,83 W/SPEC	1,75 W/SPEC
System #2	2,88 W/SPEC	2,00 W/SPEC	1,43 W/SPEC	1,36 W/SPEC	1,30 W/SPEC
System #3	2,42 W/SPEC	1,70 W/SPEC	1,24 W/SPEC	1,18 W/SPEC	1,15 W/SPEC
Verbrauch der CPU (W)/SPEC					
System #1	1,81 W/SPEC	1,37 W/SPEC	0,98 W/SPEC	0,92 W/SPEC	0,88 W/SPEC
System #2	1,74 W/SPEC	1,33 W/SPEC	0,96 W/SPEC	0,90 W/SPEC	0,86 W/SPEC
System #3	1,60 W/SPEC	1,19 W/SPEC	0,86 W/SPEC	0,80 W/SPEC	0,78 W/SPEC

und fairerweise mit „estimated“ markiert. Jedoch ist der Ansatz, von der offiziell dokumentierten Leistungsfähigkeit einer CPU auf die Effizienz eines Servers zu schließen, viel zu kurzsichtig. Messungen von c’t und iX zeigen seit Jahrzehnten, dass die im Labor geschönten SPECcpu-Werte erheblich von der im RZ-Betrieb erreichbaren Leistung abweichen.

Zur Erklärung: Für den SPECcpu2017 erhalten die OEMs vom CPU-Hersteller einen vorkompilierten Benchmark, den sie dann – nach strikten Vorgaben – auf der Zielumgebung ausführen. Zwar ist in den Begleitpapieren exakt angegeben, wie die EXE-Datei erzeugt wurde, doch lassen sich die Kommandos fast nie nachvollziehen. Dazu werden spezielle Compiler mit exotischen Optionen verwendet, Prozesse an Kerne gebunden und alle überflüssigen Dienste deaktiviert.

Kurzum: Die Ergebnisse sind praxisfern und haben mit dem Serienprodukt so wenig zu tun wie Rennsportautos mit ihren gleichnamigen Serienschwestern. Das zeigen beispielsweise die Ergebnisse von AMDs Epyc 7551P in Dells PowerEdge R7415. Bei spec.org wird das System mit 134 CPU2017 INTrates ausgewiesen. Im iX-Labor erreichte es gerade einmal 95 INTrates, also 70 Prozent, beim Einsatz des generischen GNU-Compilers gcc mit der Option -O3. 120 INTrates – 89 Prozent – erreichte das System mit AMDs auf Epyc optimierter Compilersuite [1].

Ein erster Schritt, die Umweltverträglichkeit stärker in den Mittelpunkt zu rücken, bestünde darin, von der Performance im Zähler einer Kennzahl abzurücken. Der Energieverbrauch pro erbrachter CPU-Leistung ist dabei mehr als der reine Kehrwert. Es stellt den Energiehunger der CPUs in den Fokus, nicht die maximale Performance.

Auch ohne externes Leistungsmessgerät lässt sich während eines SPECcpu2017-Laufs die Leistungsaufnahme des CPU-Package einfach periodisch abfragen. Moderne CPUs bieten dazu die RAPL-Schnittstelle. Liefern kann die Werte dann entweder der BMC (Board Management Controller) über die IPMI-Schnittstelle oder das Betriebssystem, abgerufen etwa mit turbostat. Das funktioniert auch in der Cloud, sofern man eine komplette Bare-Metal-Instanz mietet.

SSJ

Das Borderstep Institute führt in seinem Bericht „Energiebedarf der Rechenzentren steigt trotz Corona weiter an“ die gestiegene Energieeffizienz der Server als Gegentrend gegen den steigenden Energiehunger der Rechenzentren an und verweist dabei auf die Ergebnistabelle des SPECpower_ssj 2008 (siehe ix.de/zdp9). Die Argumentationskette lautet: Der steigende Energiebedarf der Rechenzentren hat seine Ursache im erhöhten Energiebedarf der Server, der wiederum Folge der erhöhten Nachfrage nach Rechenzentrumsleistung im Zuge der Digitalisierung und der Pandemie ist. Die durchschnittliche Leistungsaufnahme eines Standardservers sei von 153 Watt im Jahr 2010 auf 228 Watt 2020 gestiegen, die Zahl der SPECpower-Rechenoperationen pro Watt hätte sich gleichzeitig aber verfünfzehnfacht.

Nachweisen lässt sich dieser Zusammenhang jedoch nicht. Die Autoren der Studie gehen einfach davon aus, dass neben den Rechenressourcen auch die Infrastruktur über die Jahre erneuert wurde. Trends wie stromfressende GPUs werden dabei ausgeblendet. Um diesen Zusammenhang herstellen zu können, schlugen die Autoren des Borderstep-Arbeitspapiers „KPI für IT-Leistung und Energieeffizienz von Rechenzentren“ bereits 2016 im Zuge der Normdiskussion eine neue Kennzahl SSJ auf Basis des SPECpower_ssj2008 vor (siehe ix.de/zdp9).

SSJ ist ein abstraktes Maß für die IT-Leistung, die ein Server erbracht hätte, wenn er statt der echten Applikation mit SPECpower_ssj2008 belastet worden wäre. Der Grundgedanke: Für jedes im RZ vorhandene System schlägt man den SPECpower_ssj2008-Wert bei voller Auslastung nach und rechnet ihn auf die durchschnittliche Auslastung der Server – nicht der CPU – herunter. Für datenbank- oder VM-lastige Applikationen seien zusätzliche KPIs zu entwickeln. Vorteil dieses Vorschlags: Ohne die Performance messen zu müssen, kommt man bei Kenntnis des Auslastungsgrads zu einer Effizienzaussage.

KPI4DCE

Auch das Umweltbundesamt schlägt mit KPI4DCE eine Kennzahl vor, die sich ohne Messung ermitteln lässt (siehe ix.de/zdp9). Dazu zieht es unter anderen die Ergebnisse des SPECcpu2006 heran, also die Vorgängerversion des SPECcpu2017, den die iX zur Bewertung der CPU-Leistung nutzt (siehe Artikel „Ein eigenes Ökotop“ ab Seite 84). Das Umweltbundesamt (UBA) bezeichnet KPI4DCE als richtungssicheres Kennzahlensystem zur umfassenden Beurteilung der Energie- und Ressour-

ceneffizienz von Rechenzentren, das erstmalig alle Teilbereiche eines RZ berücksichtigt und die Leistungsfähigkeit der IT miteinbezieht (siehe Artikel „Weiter gefasst“ ab Seite 78).

Die IT-Leistung eines Rechenzentrums setzt sich demnach aus dem Nutzen der Rechenkapazität, des Speicherplatzes und der WAN-Nutzung zusammen. Die Rechenkapazität ermittelt man, indem man die veröffentlichten SPECcpu2006-Ergebnisse der verwendeten CPUs zurate zieht. RZ-Betreiber müssen dazu Anzahl und Typ der installierten CPUs in eine umfangreiche Excel-Tabelle eintragen und bekommen das SPECcpu2006-Ergebnis ihrer CPUs, die auf ganz anderen Systemen gemessen wurden, ergänzt.

Das UBA geht also davon aus, dass man von der Performance der CPU auf die Gesamteffizienz schließen kann, was jeder Benchmark-erfahrene Administrator widerlegen kann (siehe Tabelle „AMD Epyc 7702 in drei Systemen“). Zudem liegen für neuere CPUs keine Ergebnisse mehr vor, da die Hersteller sie mit dem neueren Benchmark SPECcpu2017 vermessen. Beide Benchmarks verwenden aber nicht nur unterschiedliche Tests, sondern auch unterschiedliche Referenzsysteme, sodass die Ergebnisse der 2006er- und der 2017er-Versionen in keiner Weise vergleichbar sind.

Allen Effizienz-KPIs gemein ist, dass sie eine weitere Vergleichbarkeit der Rechenzentren erlauben, die sich heute vor allem durch ihre Nähe zu einem zentralen Internetknoten und der PUE (Power Usage Effectiveness) unterscheiden. Denn die PUE, die das Verhältnis des Gesamtenergiebedarfs zum Bedarf des IT-Equipments beschreibt, nimmt nur die Effizienz der Infrastruktur in den Blick und sagt nichts über die der IT-Komponenten aus, zumal Erstere primär von der Kühlung und damit von der geografischen Lage abhängt.

Idle ist Gift

Eine IT-Applikation nimmt sich die Ressourcen, die sie bekommen kann, und verwendet dazu alle IT-Komponenten, die sie benötigt. Je mehr Leistung das System ihr bietet, desto schneller ist sie fertig und lässt das System im Idle-Zustand zurück, einem Zustand, der für die Effizienz Gift ist, wie die rote Ampel für den Spritverbrauch. Im Idle-Zustand sinkt der absolute Energiebedarf der Komponenten, der relative Anteil der Netzteile steigt hingegen.

Ein Einkaufskonfigurator schlägt die Netzteile in Abhängigkeit von der CPU-Last und den sonstigen Komponenten vor. Es gilt: Je mehr die Maschine leisten muss, desto größer ist das Netzteil dimensioniert. Oft aber fallen, wie beim Motor des Autos, die Netzteile zu groß aus. Überdimensionierte Netzteile haben jedoch einen schlechteren Wirkungsgrad.

Netzteile, die der 80-Gold-Platinum-Spezifikation entsprechen, zeigen bei einer Auslastung ab 20 Prozent einen Wirkungsgrad von mehr als 90 Prozent. Das Optimum liegt mit 94 Prozent bei 50 Prozent Last. Unterhalb von 20 Prozent Last ist die Effizienz schlecht, pendelt oftmals bei 80 Prozent. Gerade das ist der Bereich des Leerlaufs, insbesondere dann, wenn beide Netzteile 2 + 0 geschaltet sind. Dadurch werden aus zweimal 750 Watt 1500 Watt, und 150 Watt Leistungsaufnahme des Servers im Leerlauf gipfeln in 30 Watt Verlustleistung der Netzteile.

Zudem geht die Definition des nominellen Wirkungsgrads davon aus, dass die Netzteile fremdbelüftet werden. Der Hunger der sehr kleinen Netzteillüfter kommt also noch obendrauf. Einfach errechnen kann man den tatsächlichen Wirkungsgrad eines Netzteils, wenn man die zwei IPMI-Messwerte PSUout und PSUin ins Verhältnis setzt.

Messen statt Referenzieren

Denn ineffiziente Systeme aufspüren kann man nur durch Messen. Wozu verfügt jeder Server, jedes Storage, ja sogar jede RZ-Steckdosenleiste über einen BMC, wenn nicht zum Überwachen, Messen und letztlich zum Fakturieren der laufenden Kosten? Monitoringsysteme ermitteln den Stromverbrauch der CPUs und der Netzteile zu jedem Zeitpunkt.

Ganz gleich, mit welchem Benchmark man den Energieverbrauch unter Volllast misst, übrig bleiben kWh/Server. Da jede Applikation vor ihrer Produktivsetzung einen Lasttest durchläuft, ist es ein Leichtes, dabei auch den Energieverbrauch zu messen und ihn bei zukünftigen Releases oder Serverkäufen zu berücksichten. Im Zweifelsfall wirft das Kommando

```
perf stat -e power/energy-cores/,power/↵
                              energy-pkg/,power/energy-ram/
```

nach Ende oder Abbruch die benötigte Energie in Joule aus.

Ohne Zugang zum Betriebssystem gibt es mithilfe des BMC zwei einfache Möglichkeiten, die Energieeffizienz von Servern zu vergleichen. Erstens: Watt / (Gesamtzahl der Kerne × Auslastung). Dabei teilt man den durchschnittlichen Energiebedarf durch das Produkt aus der Zahl der Rechenkerne und der durchschnittlichen Auslastung. Aufgrund physikalischer Grenzen erhöht sich die Leistung pro Kern seit einigen Jahren kaum noch, die Leistungssteigerung findet durch Erhöhung der Anzahl Kerne statt. Gleichzeitig sinkt der Energiebedarf pro Kern seit Jahren.

Zweitens: PSUout/PSUin. Während sich eine schlecht programmierte Applikation durch immer höhere CPU-Performance kaschieren lässt, ändert sich die Qualität der Netzteile seit Jahrzehnten kaum. Dabei liegt in ihnen nachweislich Potenzial. Das zu belegen reichen auf RZ-Ebene zwei Zahlen und eine Aussage: Bei einem RZ-Nettoenergiebedarf von 175 MWh/a führen die Servernetzteile 19 MWh pro Jahr an Verlustleistung als Wärme ab. Diesen Satz sucht man in Unterlagen vergebens. Diese Verschwendung aber ist es wert, beachtet zu werden, vergleichbar mit der Verbrauchsanzeige eines Autos an der roten Ampel: l/h oder kWh/h.

Fazit

Der Weg zur performanceunabhängigen Kennzahl erfordert ein Umdenken, das damit beginnt, den Energieverbrauch in den Zähler zu setzen. Es fällt schwer, bei der Betrachtung der Energieeffizienz vollständig auf die Performance zu verzichten. Doch der BMC liefert ausreichend Messdaten, um die Effizienzinnovation ohne Performance-Messung beurteilen zu können. (sun@ix.de)

Quellen

[1] Hubert Sieverding; Underdog; Dells PowerEdge R7415 mit AMDs Epyc; iX 9/2018, S. 70
[2] Berichte des Borderstep Institute und des Umweltbundesamts: ix.de/zdp9

Hubert Sieverding
arbeitet nach langjähriger Tätigkeit in der Automobilbranche als freier Autor.

Umweltgerechte RZ-Infrastruktur

Unproduktiv und dabei äußerst energiehungrig ist die Infrastruktur der Rechenzentren. Und doch ist sie unerlässlich für den geregelten und zuverlässigen Betrieb der IT. Die Frage lautet also: Wie kann sie genügsamer und umweltgerechter werden, ohne die Verfügbarkeit der IT zu beeinträchtigen? Ansätze dazu gibt es viele.

Stromversorgung – Zukunftsfähiges Energiemanagement im Rechenzentrum 110

Stromausfall – Alternativen zur unterbrechungsfreien Stromversorgung mit Blei-Gel-Akkus 118

HFKW-Alternativen – Natürliche Kältemittel für die RZ-Klimatisierung 122

Supercomputer – Direkte Warmwasserkühlung von HPC-Chips 128

Racks und Reihen – Systemnahe Wärmeübertragung für luftgekühlte Server 134

Abwärme – Flüssigkeitskühlungen für Server 140

Projekte – Abregelungen und Abwärme: Verpuffte Energie nutzen 146

Zukunftsfähiges Energiemanagement im Rechenzentrum

Kraftquelle

Ariane Rüdiger

Zur Nachhaltigkeit gehört auch ein effizienterer Umgang mit Energie. Wie das gelingen könnte, dafür gibt es eine Fülle von Ansätzen, aber noch keinen Königsweg.

Gegen Stromausfälle kennt das RZ-Management umfangreiche Gegenmaßnahmen, allen voran die unterbrechungsfreie Stromversorgung (USV) samt den dazugehörigen Batteriebänken, Dieselgeneratoren und anderen Formen des Energieersatzes. Sie sollen einen Ausfall der Netzressourcen bis zu einigen Tagen oder gar Wochen zuverlässig überbrücken.

Zwar hat der Schutz vor Stromausfällen die höchste Priorität, doch reicht das alleine nicht. Rechenzentren müssen heute in eine ganz andere Rolle hineinwachsen: in die eines aktiven Teilnehmers am Energiemarkt, der weitestgehend mit regenerativen Energien handelt. Das erfordert neben technischen Neuerungen vor allem ein Umdenken der durchaus und aus guten Gründen eher sicherheitsorientiert agierenden RZ-Facility-Manager.

Die Notwendigkeit verdeutlichen ein paar nüchterne Zahlen. Nach dem Ergebnisbericht des Projekts DC-Heat, verfasst vom Borderstep-Institut, hat sich der Energiebedarf der deutschen Rechenzentren zwischen 2010 und 2020 um 60 Prozent auf 16 TWh erhöht. Die Gesamtstromnutzung lag 2019 in Deutschland laut Destatis bei 534 TWh (alle Quellen siehe ix.de/z94w). Mithin liegt der Anteil der Rechenzentren in Deutschland bereits bei drei Prozent. In Frankfurt, einer der vier großen europäischen RZ-Metropolen, lassen die Rechenzentren den restlichen Stromabnehmern kaum noch Kapazitäten übrig (siehe Abbildung 1).

Der Stromverbrauch der Rechenzentren muss also dringend gesenkt werden. Nur so nämlich lässt sich die notwendige Dekarbonisierung von Wirtschaft und Gesellschaft erreichen. Dabei auf Kompensationseffekte in anderen Branchen und bei den Anwendungen der IT zu warten, war bisher nicht zielführend. Das zeigt die aktuelle, vom Borderstep-Institut koordinierte Studie CliDiTrans (Klimaschutzpotenziale der digitalen Transformation). Das Ergebnis: Die ohnehin eher marginalen positiven Klimawirkungen der digitalen Transformation hat die Mehrnutzung wieder aufgefressen. Das heißt: Man muss erstens regulieren und zweitens die Digitalisierung selbst grün gestalten, da sie nicht nur erhebliche Materie- und Energiemengen verschlingt, sondern auch recht ineffizient ist (alle Studien siehe ix.de/z94w).

Gefragt: ganzheitliches Energiemanagement

In den Netzen fließt zukünftig vor allem regenerativer Strom, dessen Erzeugung aufgrund der Wetterbedingungen und Tageszeiten fluktuiert. Deshalb müssen sich alle Netzteilnehmer, auch Rechenzentren, auf eine aktivere Rolle einstellen, nämlich

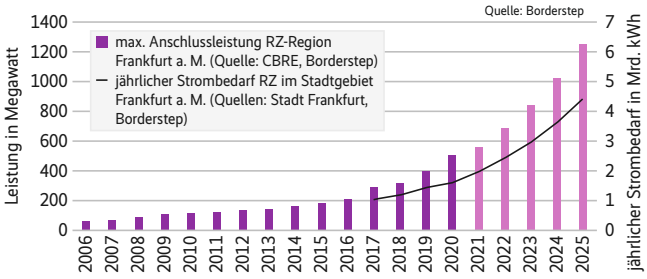
iX-TRACT

- Elektrische Energie sollte in allen Phasen ihres Lebenszyklus – Erzeugung, Nutzung, Umwandlung in Wärme – als wertvolle Ressource betrachtet werden.
- Ein nachhaltiges und ganzheitliches Energiemanagement ist entscheidend für die Zukunft der Rechenzentrumsbranche.
- Dabei müssen Rechenzentren am Energiemarkt eine aktivere Rolle einnehmen. Als Prosumenten können sie sich am Netzausgleich beteiligen.
- Voraussichtlich werden die Energie- und die RZ-Branche weiter zusammenrücken. Beide stellen kritische Infrastrukturen bereit und können von den Ressourcen des jeweils anderen profitieren.

die von Prosumenten, die ihre Potenziale nutzen, um selbst zu einem zuverlässigen, frequenz- und spannungsstabilen Stromnetz beizutragen und davon günstigenfalls zu profitieren.

Gefragt ist ein ganzheitliches aktives Energiemanagement entlang folgender Grundfragen: Wie viel Energie welcher Qualität fließt ins Rechenzentrum? Was geschieht dort mit der Energie? Wie viel Energie fließt zu welchem Zweck aus dem Rechenzentrum heraus und wie wird sie weiter genutzt, um nicht sinnlos zu verpuffen? Am Ende steht auch die Frage, wie sich aus diesen Gegebenheiten ein funktionierendes Geschäftsmodell stricken lässt.

Das Ziel der Rechenzentren muss sein, so wenig Strom wie möglich aus externen Quellen zu beziehen, aber genug, um das Rechenzentrum in dem mit den Kunden vereinbarten Umfang am Laufen zu halten. Dazu gehört eventuell auch, selbst Energie zu erzeugen. Beim Betrieb sollte keine Energie verschwendet werden. Möglicherweise kann der gesteuerte Ressourcenverbrauch sogar zur Stabilisierung des Stromnetzes beitragen. Und schließlich sollte die als Ressource zu betrachtende Abwärme nicht sinnlos verpuffen.



Der Strombedarf der Rechenzentren droht in Frankfurt am Main alle anderen Stromabnehmer an den Rand zu drängen (Abb. 1).

Dass das funktionieren kann, belegt Microsoft: Dessen RZs sollen schon bald nicht nur kein CO_2 mehr ausstoßen, sondern bis 2030 sogar einen absoluten Beitrag zur Einsparung von

Die Emissions-Scopes nach dem Greenhouse Gas Protocol

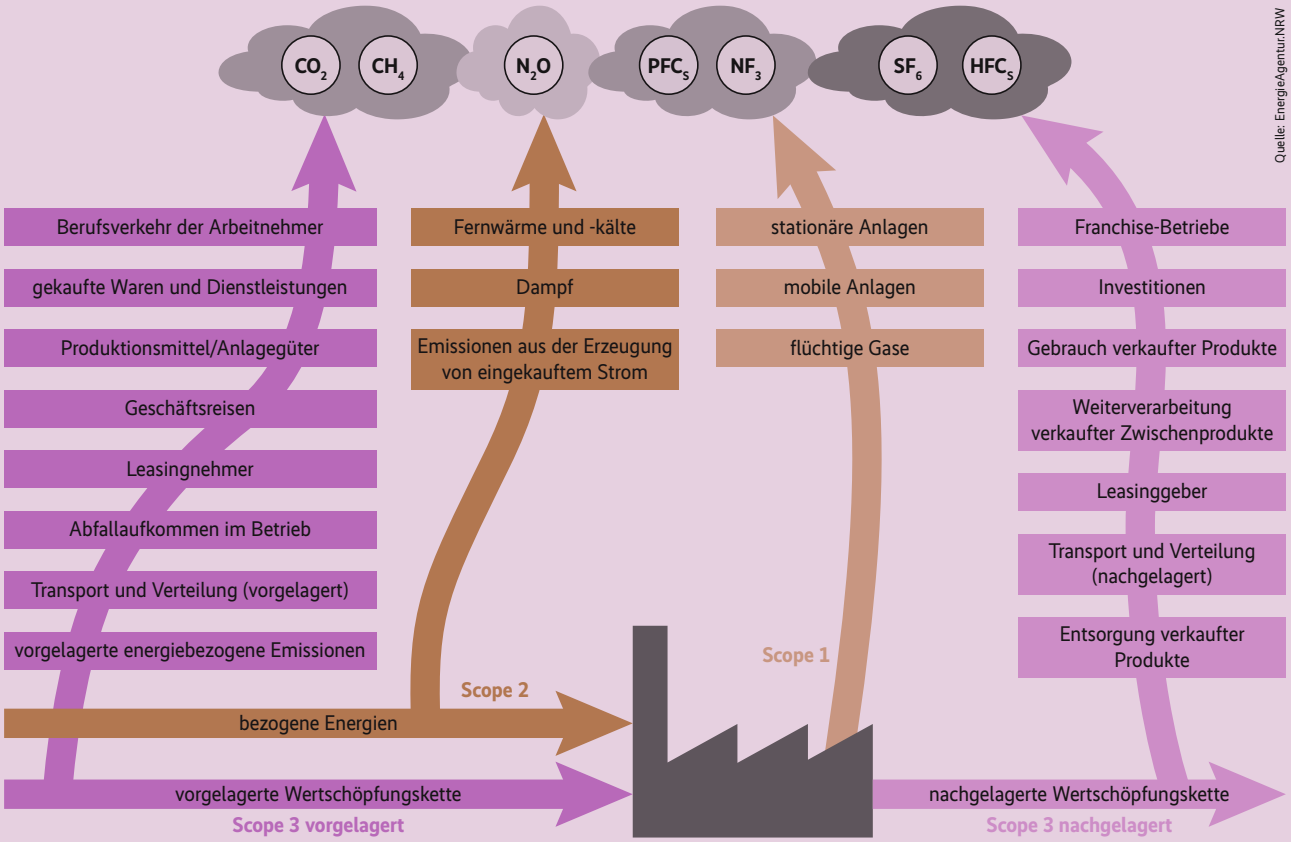
Das GHG-Protocol, einer der meistgenutzten internationalen Standards zur Berechnung unternehmensbezogener Treibhausgasemissionen, beschreibt im Kapitel 4 „Setting Operational Boundaries“ die Emissions-Scopes 1 bis 3 (siehe Abbildung 2):

Scope-1-Emissionen entspringen Quellen, die sich direkt im Besitz oder Geltungsbereich des Unternehmens befinden, etwa aus dem eigenen Heizkessel oder Fuhrpark.

Scope-2-Emissionen stammen aus eingekauften Energien wie Strom, Wärme, Kühlung. Selbst erzeugter Strom fällt nicht unter Scope 2, dafür wird der eingesetzte Brennstoff unter Scope 1 als direkte Emission bilanziert.

Scope-3-Emissionen resultieren aus Tätigkeiten, die nicht direkt zum Unternehmen gehören, etwa aus Geschäftsreisen oder dem Abfallmanagement.

Die Berichterstattung der THG-Emissionen in den Scopes 1 und 2 ist verpflichtend, da ihre Emissionsquellen so definiert sind, dass zwei Unternehmen nicht dieselbe Emission berichten. Bei Scope-3-Emissionen kann es zu Überschneidungen kommen, ihre Bilanzierung ist optional.



Die Emissionskategorien oder Scopes beschreiben die Verantwortlichen der Treibhausgasemissionen – ob diese selbst erzeugt, durch direkte Nutzung oder die Inanspruchnahme von Dienstleistungen verursacht wurden (Abb. 2).

Quelle: Westfalenwind

Im Fuß des Enercon-Windgenerators der Westfalenwind befindet sich das Windcores-Rechenzentrum (Abb. 3).

Treibhausgasen leisten, also Carbon-negative werden und damit ehemalige Emissionen des Unternehmens kompensieren (alle Projekte siehe ix.de/z94w).

Grüne Zertifikate gehören schon beinahe zum guten Ton bei Hostern und Cloud-Anbietern. Schließlich legen deren Kunden Wert darauf, ihre eigene Kohlendioxidbilanz nicht durch schwarze Flecken aus Scope 3, sprich in der Lieferkette, zu verunzieren (siehe Kasten „Die Emissions-Scopes nach dem Greenhouse Gas Protocol“).

Direkt zur Stromquelle

Alternativ ließen sich die eigenen Rechenkapazitäten dahin verlegen, wo grüne Energie reichlich vorhanden und kostengünstig ist, vorzugsweise in nordische Länder wie Schweden, Norwegen oder Island, deren Klima über viele Monate hinweg eine freie Kühlung erlaubt. Viele Hyperscaler wie Facebook und Google sind bereits an solchen Standorten vertreten. Dagegen spricht, dass man seine RZ-Ressourcen nicht ausschließlich im Ausland haben möchte. Attraktiver sind da die großen RZ-Hotspot-Regionen Europas FLAP: Frankfurt, London, Amsterdam und Paris.

Eine Alternative dazu sind Power Purchase Agreements (PPAs), also oft langfristige Stromlieferverträge zwischen einem Abnehmer und einem Erzeuger erneuerbarer Energien. Solche Stromkaufvereinbarungen sind meist eine Sache der Großen und vor allem in den USA gängig. Zudem können in der Regel nur große Stromlieferanten die Anforderungen großer Rechenzentren erfüllen. In Deutschland ist das PPA daher noch nicht so weit verbreitet, wird aber beliebter.

Theoretisch ließe sich der Strom auch selbst erzeugen. Auch dafür sind in erster Linie die Hyperscaler bekannt. 2019 etwa kündigte Google an, zwei Milliarden Dollar in 18 Erneuerbare-Energien-Anlagen auf drei Kontinenten zu investieren. Auch in Deutschland gibt es Beispiele dieser Art. Das bekannteste ist wohl Windcores (siehe Artikel „Aufgefangen“ ab Seite 146). Der Rechenzentrumszweig von Westfalenwind baut RZs in den recht großen hohlen Fuß der Enercon-Windgeneratoren, die den Strom liefern (siehe Abbildung 3). Beispielsweise stehen dort die Server des TV-Streaminganbieters Zattoo. Der Vorteil: Das Platzieren der Systeme direkt beim Erzeuger entlastet das Netz und reduziert die Menge des abgeregelten Stroms. Nur bei Windstille ziehen sie Energie aus dem Netz oder nutzen ihre Notressourcen.

Zwei RZs betreibt Windcloud auf dem Greentec-Campus im küstennahen Enge-Sande und in Bramstedt. Beide sind in aufgelassenen Bundeswehrbunkern untergebracht und ziehen ihre Energie aus den nahegelegenen Windparks. Bis zu 60 kW Rechenleistung haben in den Standorten Platz. Das ist nicht sehr viel, innovativ ist aber das Gesamtkonzept des Betreibers vor allem bei der Abwärmenutzung. Ein RZ-Großprojekt mit einer Gesamtleistung von 180 MW ist bei Hanau geplant. Es beteiligt sich an einem geplanten benachbarten Fotovoltaikpark, von dem es dann seine Energie bezieht.

Mit Brennstoffzellen in die Zukunft

Weiter in die Zukunft weist ein Projekt des schwedischen Forschungsinstituts RISE (Research Institute of Sweden). In Lulea versucht RISE im Rahmen von ICE Edge erstmalig, Festoxidbrennstoffzellen als primäre Energiequelle für Rechenzentren zu nutzen und mit Abwärmerückgewinnung zu kombinieren (siehe Abbildung 4). ICE Edge ist ein Teilprojekt des EU-Vorhabens Wedistrict, das wiederum zu den Horizon-2020-Projekten gehört.

Die Energie der SOFCs (Solid Oxide Fuel Cells) stammt aus Biogas, das sich bis zu 35 Prozent mit Wasserstoff anreichern lässt und dessen Tank sich auf dem RZ-Gelände befindet. Die auf 2,7 kW/m² ausgelegten Zellen sind eigentlich für RZ-Anforderungen zu schwach. Brennstoffzellenhersteller SolidPower möchte aber mit diesen Aggregaten den RZ-Markt erobern und strebt deshalb auch mithilfe des Projekts eine Leistungssteigerung auf 10 kW/m² pro Zelle an. In dem Rechenzentrum arbeiten neun Zellen; die Leistung von dreien wurde jeweils gebündelt. Die IT-Systeme selbst sind immersionsgekühlt (siehe Artikel „Ohne Umwege“ ab Seite 140). Ihre Abwärme trägt zur Temperierung der Brennstoffzellen bei, indem sie die Container mit den Brennstoffzellen erwärmt.

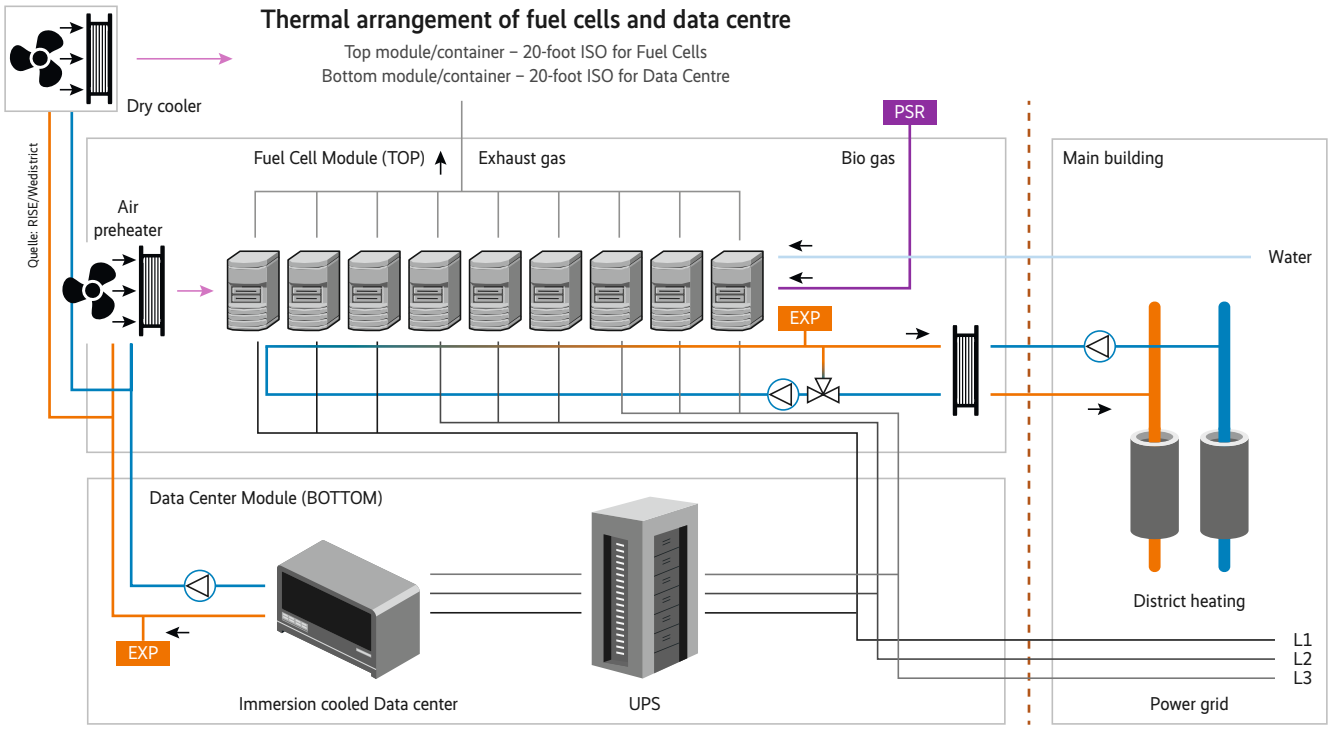
Ein bereits kommerziell genutztes Brennstoffzellen-RZ-Projekt entsteht im britischen Manchester: Dort baut Tele-Data ein RZ, das 1,2-MW-Festoxidbrennstoffzellen von Bloom als zweite Stromquelle integriert. Das Rechenzentrum nutzt sie im Rahmen eines PPA und kann später auf reinen Wasserstoff als Energieträger umsteigen.

Eigene Kraftwerksressourcen fürs RZ nutzen

Gas als Energieträger kann auch dann eine Rolle spielen, wenn die primäre Energiequelle des Rechenzentrums ein stromgeführtes Blockheizkraftwerk (BHKW) ist. Zudem ist Gas als Überbrückung gedacht, wenn Diesel aus Umweltgründen nicht mehr verwendet werden soll. Immerhin gibt es in Deutschland vielerorts ein funktionierendes Ferngasnetz. Zudem deutet vieles darauf hin, dass klassisches Erd- oder Biogas etwa durch Wasserstoff ersetzt werden wird.

Doch haben Wasserstoff und andere gasförmige Alternativen eine etwas geringere Energiedichte als Erdgas. Deshalb muss bei der Umstellung auf Wasserstoff entweder das Rechenzentrum oder das Kraftwerk redimensioniert werden. Grundsätzlich ist ein BHKW beispielsweise bei kleineren Rechenzentren auf Fabrikgeländen denkbar – meist wohl parallel zu einer Versorgung aus dem öffentlichen Netz oder redundant ausgelegt.

Inzwischen entdecken Energieversorger den Rechenzentrumsmarkt als neue Erlösquelle. Ein Beispiel dafür sind die Kraftwerke Mainz-Wiesbaden, die eine ganze Reihe von Wind- und Solaranlagen selbst betreiben und die gewonnene Energie ins Netz einspeisen. Nun plant das Unternehmen den Bau eines Rechenzentrums an einem seiner Standorte, das im Endausbau mehr als 50 MW liefern soll. Es wird bilanziell mit dem regene-



Beim Test-RZ am RISE fließt die Abwärme der Brennstoffzellen ins Fernwärmenetz, die Abwärme der immersionsgekühlten IT erwärmt die Luft um die Brennstoffzellen (Abb. 4).

rativen Strom aus eigener Erzeugung gefüttert statt wie meist üblich mit Zertifikaten gehandelt.

Das wirft aufseiten der Energieerzeugung die Frage auf: Warum nicht gleich mit Gleichstrom ins RZ gehen? Das würde viele verlustreiche Wandlungen einsparen. Denn digitale Systeme arbeiten mit Gleichstrom, nicht mit konventionellem Wechselstrom. Solarstrom könnte in quasi naturnaher Form eingespeist werden, auch zu Batterien passt ein Gleichstromabnehmer hervorragend.

Gleichstrom mit Supraleitung kombiniert

Versuche, mit Gleichstrom betriebene RZs zu etablieren, gab es bereits. Beispielsweise errichtete 2012 die Green-Gruppe ein Gleichstrom-RZ mit 1 MW Leistung auf dem Datacenter-Campus Zürich-West. Von den einschlägigen Anbietern von Stromversorgungsequipment ist zu vernehmen, dass die Durchsetzung von Gleichstrom als Versorgungsenergie sicher auf längere Sicht sinnvoll, momentan aber wenig praktikabel sei.

Die Gründe dafür sind vielfältig, von der unvollständigen Produktpalette für die Infrastruktur über die fehlende Erfahrung aufseiten der Anwender und Gerätehersteller bis zur Gewohnheit. Zudem erfordern erhöhte Sicherheitsanforderungen etwa an Schaltern ein komplexeres Design der Stromkomponenten, da beim Schalten unter Spannung die Gefahr von Lichtbögen besteht. Darüber hinaus sind die digitalen Systeme auf der Stromzugangsseite anzupassen, da die integrierten Konverter entfallen. Tatsächlich kannibalisieren die Hersteller ihr gut laufendes Wechselstromkomponentengeschäft, wenn Anwender verstärkt auf Gleichstrom setzen. Hier mangelt es also auf beiden Seiten an Interesse.

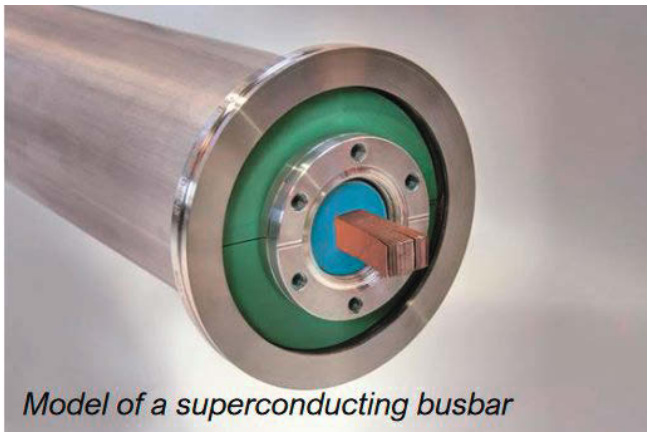
Eine Alternative hat der Supraleitungspionier Vision Electric Superconductors aus Kaiserslautern auf dem OCP-Summit (Open Compute Project) 2021 vorgestellt: ein mit Supraleitungsgleichstrom versorgtes Rechenzentrum. Basis des EOS-RZ-Designs (ECO Open Superconducting) ist eine hochkompakte, supraleitende Stromschiene, die samt den nötigen Aggregaten und Verbindungsmodulen zur externen Energieversorgung mit flüssigem Stickstoff gekühlt wird (siehe Abbildung 5). Die Schienen leiten ab –210 °C verlustfrei und haben eine sehr hohe Energiedichte von 62 A/mm² – etwa 1000-mal mehr als übliche Kupferleitungen.

Die Technik verspricht aufgrund der Material- und Produkteinsparung bei der Strominfrastruktur 80 Prozent weniger Scope-3-CO_2 im Rechenzentrum, 30 Prozent höhere Rackdichte und 30 Prozent Energieeinsparung durch geringere Leitungsverluste. Zudem ist flüssiger Stickstoff ein verbreitetes und günstiges Abfallprodukt industrieller Prozesse. Grundsätzlich wäre auch flüssiger Wasserstoff als Kühlmedium denkbar. Ein 80-MW-Rechenzentrum könnte pro Jahr mit dieser Technik rund vier Millionen Euro sparen.

Anschluss auf Mittelspannungs- oder Niederspannungsebene?

Heute ist es üblich, die Rechenzentrumsräume auf Niederspannungsebene bei 230 V mit USVs abzusichern, die auf den Systemen die benötigte Spannung zur Verfügung stellen. Denkbar – und zwar gerade dann, wenn der Stromanbieter auch der RZ-Betreiber ist – ist auch ein übergreifender Anschluss auf der

Quelle: Vision Electric Superconductors



Die supraleitende Stromschiene ist von einer stickstoffgekühlten Kammer ummantelt (Abb. 5).

Flexible Energiesteuerung im RZ

Methode	UPS	Zeitverlagerung	Ortsverlagerung	Backup-Generator	Backup-Batterie
Zielapplikation	Netzdienstleistung, z. B. Frequenzregulierung	Energiemarkt	Energiemarkt	Energiemarkt	Energiemarkt und Systemdienstleistungen
Verbreitung	sehr verbreitet	Lastmanagement ja, Lastverschiebung nein	Lastmanagement ja, Lastverschiebung begrenzt	sehr verbreitet, meist Diesel	Pilotstadium
Stromkapazität	entspricht RZ-Kapazität	30–50 %	30–50 %	entspricht RZ-Kapazität	entspricht RZ-Kapazität
zeitliche Reichweite	einige Minuten	Stunden (terminierte Fertigstellung)	Stunden, hängt von Rechenleistung des RZ ab	2–8 Stunden, abhängig von Brennstoffvorräten	bis zu 50 %
Implementierungs-bereitschaft	hoch, abhängig vom Batterietyp	niedrig bis mittel, eher bei Hyperscalern	nur bei Betreibern mehrerer Rechenzentren	wahrscheinlich bei CO_2-armen Brennstoffen, Zuverlässigkeitsbedenken	Zusatzeinnahmen entscheiden über Wirtschaftlichkeit

Mittelspannungsebene von 110 kV. Dann lässt sich dort eine passend dimensionierte Überbrückungsressource zu den Niederspannungsbereichen implementieren. Das wiederum bedeutet, dass ihr Strombedarf im Verhältnis zur nächsten Ebene gemittelt wird.

USVs funktionieren in einem solchen Design auf der übergreifenden Ebene. Dieses spart eine Menge Infrastruktur auf der Niederspannungsebene und erlaubt, gekoppelt mit entsprechenden Softwaremechanismen auf der IT-Ebene, eine flexiblere Lastverteilung auf die einzelnen RZ-Bereiche. Das hat unter Umständen Vorteile, wenn der Rechenzentrumsbetreiber ein Einkommen aus der Bereitstellung von Netzdienstleistungen generieren will, denn die USVs auf der Mittelspannungsebene werden naturgemäß größer sein als die einzelnen USVs auf der Niederspannungsebene.

Insbesonders in Hyperscale-RZs mit Rackdichten von bis zu über 20 kW sind weitere rackinterne Sekundärstromquellen und USVs im Gespräch. Das Open Compute Project hat ein skalierbares Power Shelf entwickelt, dessen Lithiumbatterien bis zu 15 kW für 300 Sekunden liefern können. Zudem sind neue Batterietechniken wie Nickel-Zinn in der Entwicklung, die weniger Gewicht mit größerer Leistungsdichte, längerer Lebensdauer und mehr Nachhaltigkeit verbinden.

Ist der Strom erst einmal im Rechenzentrum, sollte man ihn möglichst sinnvoll nutzen, das heißt, ihn möglichst wenig an unproduktive Komponenten zu verschwenden, ohne die Zuverlässigkeit zu beeinträchtigen. Bei der Kühlung, der energiehungrigsten RZ-Komponente, ist man hier schon recht weit gediehen. Beim Strom sind zunächst die parasitären Ströme zu erfassen und so weit wie möglich einzuschränken. Um Erfolge mess- und vergleichbar zu machen, benötigt man geeignete Parameter.

Quelle: Vision Electric Superconductors

Ein RZ mit supraleitender Gleichstromversorgung kommt mit wesentlich weniger Komponenten aus und ist einfacher aufgebaut (Abb. 6).

Mit der HUE (Hardware Utilization Effectiveness) haben Jayati Athavale und Anthony Chan von Meta einen solchen auf dem OCP Summit 2021 vorgeschlagen. HUE ist der Quotient von Gesamtstrombedarf der IT geteilt durch den Strom, der tatsächlich für die Datenverarbeitung benötigt wird. Dieser berechnet sich, indem man vom Gesamtstrom für die IT sämtliche Quellen parasitärer Ströme abzieht. Zu ihnen gehören Idle-Stromverbrauch, Klimakomponenten wie Lüfter oder Pumpen direkt auf oder in den IT-Systemen, Netzteilverluste und Leckströme (siehe Artikel „Ein eigenes Ökotop“ ab Seite 84). Anwenden lässt sich die HUE auf Rechner, Chips, Racks oder ganze Rechenzentrumsräume.

Das Rechenzentrum als Microgrid

Solche Effizienzmessungen sind allein aber nicht ausreichend. Vielmehr ist das Rechenzentrum in ein integriertes energetisches System umzugestalten, das mehrere verteilte Energieressourcen und -verbraucher verwaltet. Ein solches Microgrid verfügt über seine eigene Steuerungssoftware, kann Energie aus dem übergeordneten Stromnetz beziehen, einspeisen oder eine gewisse Zeit autonom arbeiten. Zu einem solchen Microgrid können diverse regenerative oder nicht regenerative Energiequellen, Energiespeicher, Prognosesoftware, Kommunikationstechnik, Software für die Echtzeitsteuerung und die Angebots- beziehungsweise Nachfragesteuerung gehören. Energieversorger vergüten Beiträge zur Netzstabilisierung. Die können auch von Ressourcen stammen, die der Stromabsicherung dienen, etwa beim notwendigen Diesellauf. Beispiele für denkbare Mechanismen und die dafür einsetzbaren Ressourcen zeigt die Tabelle „Flexible Energiesteuerung im RZ“.

Wie wichtig die in den Rechenzentren vorhandenen Ressourcen für Strominfrastrukturen sein könnten, die vorwiegend auf erneuerbaren Energien fußen, zeigen weitere Daten aus der Studie von Bloomberg NEF, Eaton und Statkraft (siehe Abbildung 7). Danach stellen die Rechenzentren in den Niederlanden, England, Deutschland, Irland und Großbritannien im Jahr 2030 perspektivisch zwar 16,8 GW oder etwa neun Prozent der Spitzenlast – heute sind es 3,8 GW oder etwa zwei Prozent –, gleichzeitig bildet ein Teil dieser Last aber auch eine große Flexibilitätsreserve.

Partnerschaft zwischen Energie- und RZ-Branche

Eine Studie der SDIA (Sustainable Digital Infrastructure Alliance) prognostiziert deshalb ein Zusammenwachsen von Ener-

gie- und Rechenzentrumsbranche. Beide hätten im Grunde dieselben Interessen: Sie stellen eine permanent benötigte und für das Funktionieren der gesamten Gesellschaft kritische Ressource bereit und arbeiten daher mit höchsten Verfügbarkeits- und Sicherheitsanforderungen. Rechenzentrumsbetreiber sind auf günstigen und permanent verfügbaren Strom angewiesen, Elektrizitätsanbieter auf Ausgleichsmöglichkeiten, wenn Erzeugung und Abnahme fluktuieren. Zudem verfügen sie oft über große Grundstücke und Anlagen, auf denen sich Rechenzentren unterbringen lassen.

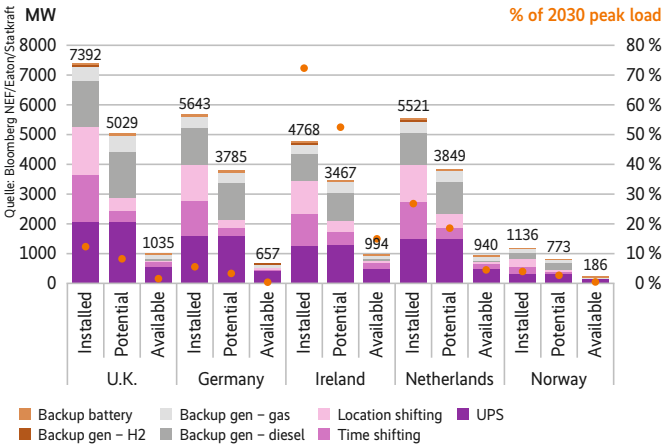
Dass solche Synergien bereits gesehen und umgesetzt werden, belegen die oben erwähnten Beispiele wie das der Kraftwerke Mainz-Wiesbaden (KMW). Das dort entstehende Rechenzentrum gehört dem Energielieferanten, der seine eigene regenerative Energieerzeugung für den RZ-Betrieb einsetzt, gleichzeitig aber vorhandene Energieerzeugungsressourcen auf seinem Gelände so als Backup verwendet, dass Dieselaggregate und USVs unnötig werden, was die Baukosten und den Kohlendioxid-Fußabdruck verringert. Dazu benötigt man einen RZ-Partner, der weiß, was IT-Service- und Kollokationskunden wollen. Dieses Know-how will KMW mit dem eigenen spezifischen Energiewissen bündeln – eine recht interessante Kombination.

Damit Vor-Ort-Ressourcen ihre neue Aufgabe, gleichzeitig Notstromversorgung fürs RZ und Flexibilitätsreserve fürs übergeordnete Stromnetz zu sein, optimal erfüllen können, muss die Notstromversorgung größer als bisher dimensioniert werden. Hierzu hat Hitachi Energy Kalkulationen angestellt. Danach wäre es für eine Überbrückungszeit von drei Tagen finanziell profitabel, Solarstromerzeugung und Batteriespeicher zu kombinieren. Allerdings wäre die dafür nötige Solarfläche sehr groß.

Eher realisierbar erscheint nach diesen Berechnungen heute eine Kombination aus Fotovoltaik, Batterie und Diesel. Die Kosten hängen weitgehend davon ab, wie groß Batteriespeicher und Solaranlage dimensioniert sind: Je größer die Batterieanlage und je kleiner die Fotovoltaikfläche, desto höher die Gesamtinvestition. Allerdings könnte sich dies mit der Verfügbarkeit günstigerer Batterietypen ändern. Ein völliger Verzicht auf Batterien zugunsten von Brennstoffzellen ist derzeit nicht wirtschaftlich. Dazu müssen Brennstoffzellen als Energieträger erst erheblich billiger werden.

Elektrische Energie als Wärme recyceln

Bleibt am Ende das Energierecycling: Nahezu 100 Prozent des aufgenommenen Stroms wandeln elektrische Systeme in Wär-



Die Elektrizitätsflexibilitätsreserven in Rechenzentren setzen sich in den fünf europäischen Ländern im Jahr 2030 aus unterschiedlichen Quellen zusammen (Abb. 7).

me um, die ein RZ mit weiterem, meist hohem Energieaufwand abführen muss. Oft wird die Abwärme in die Luft entlassen. Auch Abwärme im Wasser zu deponieren, ist üblich. Das geplante Rechenzentrum der KMW beispielsweise soll Abwärme zumindest teil- oder zeitweise im Rhein verklappen. Gleichzeitig werden einige Kraftwerke anderer Betreiber am Fluss stillgelegt, die ebenfalls Rheinwasser zum Kühlen nutzten, sodass daraus keine zusätzliche Erwärmung folgen dürfte. Heute weiß man, dass schon eine Erwärmung des Flusswassers um 1 °C das komplette Ökosystem verändert.

Schon 2012 setzte der Schweizer Betreiber Deep Green Datacenter darauf, sein RZ mit Wasser aus dem angrenzenden Bodensee zu kühlen. Allerdings ging das Unternehmen bereits 2013 pleite, aus den ehrgeizigen Bauplänen wurde nichts. Projekt Nautilus, ein auf dem OCP Summit 2021 vorgestelltes Start-up, das RZs auf ausgedienten Schiffen unterbringen will, setzt ebenfalls auf Seen, Flüsse oder Meerwasser zum Kühlen. Und auch Microsofts experimentelle Unterwasserrechenzentren des Projekts Natick kühlen mit Wasser.

Besonders weit gedacht ist das alles nicht, aus drei Gründen: Erstens sind die Flüsse, Seen und Meere ohnehin einem Erwärmungsstress durch den Klimawandel ausgesetzt. Zweitens lässt sich die Ressource Wärme viel besser zum Heizen oder Kühlen nutzen. Drittens muss diese unfassbare Verschwendung jedem effizienzbewussten Ingenieur die Schamesröte ins Gesicht treiben.

Welche Art der Abwärmenutzung sinnvoll ist, hängt von der Temperatur der Abwärme ab. Bei klassischen luftgekühlten Systemen liegt sie meist unter 30 °C – für Wärmenetze zu wenig. Doch gibt es andere Möglichkeiten. Windcloud etwa erwärmt damit in Gewächshäusern gezüchtete Spirulina-Algen,

Quelle: Quarnot

Die Abwärme der Quarnot-Knoten fließt direkt in die Wärmeversorgung (Abb. 8).

ein margenträchtiges Produkt, eingesetzt in der gesunden Ernährung und zum Düngen. Auch Kräuter oder empfindliche Gemüse können in solchen Gewächshäusern gedeihen.

Green Mountain beheizt mit der Abwärme seiner Rechenzentren zwei Tierfarmen. In der an Land befindlichen Hummerfarm brauchen die teuren Krebstiere bei 20 °C warmem Wasser nur ein bis zwei statt fünf Jahre bis zur Schlachtreife. In der Forellenfarm, deren Wassertemperatur bei 12 bis 15 °C liegt, werden jährlich bis zu 9000 Tonnen Forellen abgefischt.

Einfacher wird die Nutzung von aus der Luftkühlung stammender Abwärme, wenn Häuser Niedrigtemperaturheizungen verwenden, doch sind die bislang selten. Als Glücksfall muss man es sehen, wenn wie beim Projekt in Hanau direkt neben einem geplanten Rechenzentrum ein Fernwärmekraftwerk entsteht, in das die RZ-Abwärme direkt eingespeist wird.

Mehr als lauwarme Luft

Erhöhen kann man die Abwärmetemperatur bei Bedarf mit Wärmepumpen. Allerdings verbrauchen sie ihrerseits Energie. Da in Deutschland die Regulierung der Abwärmenutzung aus Rechenzentren gewaltige Lücken aufweist, sind bisher weder Betreiber noch Wärmeverbraucher sonderlich motiviert, die Technik voranzutreiben.

Adsorptionskälteanlagen können die Abwärme in Kühlenergie umwandeln, wenn auch nicht gänzlich ohne Materie- und Energieaufwand. Auch solche Konzepte haben es in Deutschland schwer. Das zeigt sich daran, dass der Adsorptionskälte-Partner des DC-Heat-Projekts wegen wirtschaftlicher Schieflage bereits aus dem Vorhaben ausscheiden musste.

Neben Finanzierungsregeln für die Anschlussstücke zum nächsten Fern- oder Nahwärmenetz fehlen vielerorts Vorgaben, die die Planung von Abwärmenutzung für Rechenzentren und die Abnahme von Fernwärme für Haushalte und Betriebe verbindlich regeln. Deshalb bilden sich nur mühsam die kooperativen Strukturen, die zur Durchsetzung solcher Vorhaben nötig sind. Auch das verbissene Festhalten am Gewohnten behindert Abwärmenutzungskonzepte, die Rechenzentren einbeziehen. Effizient wird Abwärmenutzung nämlich erst mit der wesentlich sparsameren und ertragreicheren Flüssigkühlung, die es inzwischen in mehreren Varianten gibt (siehe Artikel „Ohne Umwege“ ab Seite 140).

Cloud&Heat konnte sein Konzept von in Wohngebäuden verteilten Kleinrechenzentren mit Abwärmenutzung durch die betreffenden Gebäude lange nicht in gewünschtem Umfang umsetzen. Heute gibt es immerhin in Frankfurt mit dem grundsanierten ehemaligen Gebäude der EZB ein Vorzeigeprojekt. Die neueste Idee der Firma: In den Vereinigten Arabischen Emiraten speisen Cloud&Heat-Anlagen ihre Abwärme in die Entsalzungsanlagen ein. Rechenzentren von 40 MW Leistung würden ausreichen, um das gesamte Flaschenwasser für das Land zu entsalzen.

Das französische Start-up Quarnot verteilt die komplexen, leistungshungrigen Rechen-Tasks seiner Kunden auf alle von ihm betriebenen Rechenknoten mit 12 bis 24 AMD-Ryzen-Prozessoren oder mit OCP-Leopard- oder -Capri-Servern mit Intel-CPUs (siehe Artikel „Ein eigenes Ökotop“ ab Seite 84). Außerdem verwendet die Firma für die Knoten gebrauchte OCP-Hardware und eine spezielle Software, die die Knoten in das Wärmesystem des jeweiligen Gebäudes einbindet. Die Wasserkühlung der Knoten erlaubt die Nutzung von bis zu 96 Prozent der erzeugten Abwärme. Bei Bezug und Einbau der OCP-Komponenten half IT Renews (siehe Artikel „Ohne Ende“ ab Seite 30). Neben den Rechenknoten bietet Quarnot auch einen Cryptoheater (QC-1) an, der Blockchains oder Kryptowährungen berechnet. Zudem nutzt die Firma die Immersionskühlung, die die Abwärmenutzung erheblich erleichtert (siehe Abbildung 8 und Artikel „Ohne Umwege“ ab Seite 140).

Projekte weisen die Richtung

Auch in Deutschland bewegt sich langsam etwas bei der RZ-Abwärmenutzung (siehe Artikel „Aufgefangen“ ab Seite 146). Seit 2020 läuft das Horizon-2020-Projekt ReUseHeat, das sich nur teilweise mit der Abwärmenutzung von Rechenzentren beschäftigt. In Braunschweig, wo die Veolia-Tochter BS|Energy rund 45 Prozent der Stadt mit Wärme versorgt, soll ein Nahwärmenetz der vierten Generation entstehen, das auch mit der Niedertemperaturabwärme eines Rechenzentrums zurechtkommt, das in der Nachbarschaft als Wärmequelle entsteht und 400 neue Häuser beheizen soll.

Das Projekt Byte2Heat, das unter Federführung des Instituts für Energiewirtschaft und Rationelle Energiegewinnung der Universität Stuttgart von 2021 bis 2024 läuft, soll Hindernisse der Abwärmenutzung aufspüren und beseitigen. Geplant ist zunächst eine Bedarfsanalyse bei RZ-Betreibern und potenziellen Nutzern. Anschließend sollen Werkzeuge wie ein Wirtschaftlichkeitsrechner, Vertragsbaukästen und Finanzierungsmodelle sowie eine Webplattform entstehen.

Gleichzeitig entsteht im selben Institut im Auftrag des Umweltbundesamtes ein Rechenzentrumskataster mit detaillierten Angaben etwa zur energetischen Situation von Rechenzentren und ihren Leistungen. Vorläufig geben Interessierte dort freiwillig ihre Daten ein. Später soll es das Kataster erlauben, bei der Bauplanung von Rechenzentren mögliche Abwärmenutzer einzubeziehen. Es könnte zum Vorbild eines gesamteuropäischen RZ-Katasters werden. Eines Tages nämlich sollen Energieverbrauch und Leistungen von Rechenzentren vergleichbar sein. Damit könnte die Energieeffizienz von Rechenzentren zum Einkaufskriterium CO_2-bewusster IT- und Multi-Cloud-Manager werden. (sun@ix.de)

Quellen

Alle Studien, Projekte und Quellen siehe ix.de/z94w

Ariane Rüdiger

ist freie IT-Journalistin.

software quality lab
Innovation Meets Quality

powered by iX

SEMINARE FÜR MEHR SOFTWARE QUALITÄT

AGILE METHODEN

Werden Sie schneller und flexibler durch agiles Vorgehen!

Requirements Engineering für die agile Softwareentwicklung	**30.06.2022 - 01.07.2022**	**Online**

PROGRAMMIERUNG & CODE

Sichern Sie nachhaltig das technische und wirtschaftliche Überleben Ihres Softwaresystems!

Code Review praktisch betrachtet	**01.07.2022**	**Online**
Testgetriebene Softwareentwicklung (TDD)	**01.08.2022 - 03.08.2022**	**Online**

REQUIREMENTS ENGINEERING

Gute Requirements sind der Grundstein für ein erfolgreiches Projekt!

IREB Certified Professional for Requirements Engineering Advanced Level (CPRE-AL): Requirements Management	**18.07.2022 - 21.07.2022**	**Online**
IREB Certified Professional for Requirements Engineering Advanced Level (CPRE-AL): Requirements Modeling	**04.07.2022 - 06.07.2022**	**Online**
IREB Certified Professional for Requirements Engineering Advanced Level (CPRE-AL): Requirements Elicitation	**11.07.2022 - 13.07.2022**	**Online**

TESTEN

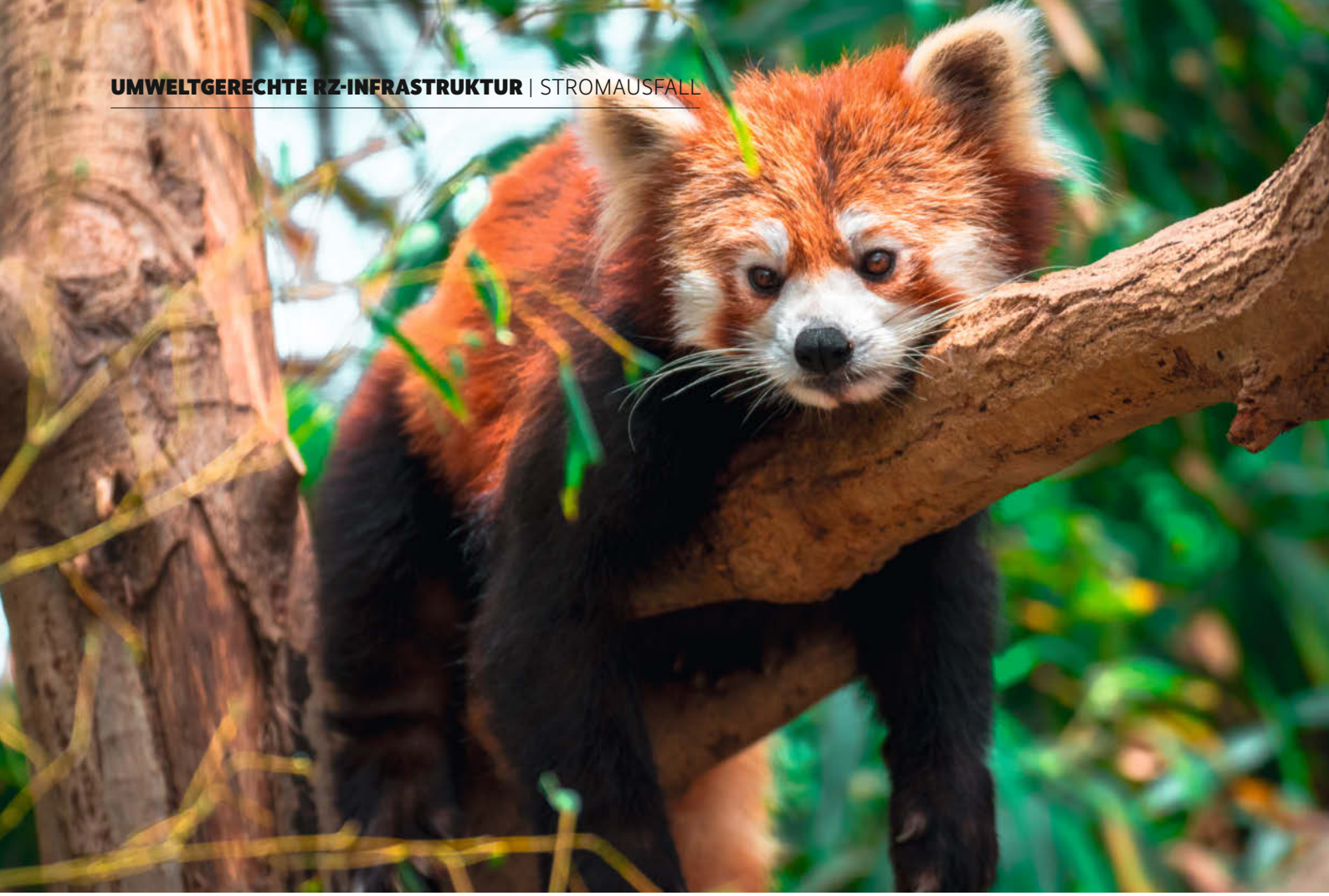
Durch erfahrene Test- und QS-Experten unterstützen wir Sie bei der Durchführung Ihrer Qualitätssicherungs- und Test-Aktivitäten.

ISTQB Certified Tester - Advanced Level: Test Analyst (CTAL-TA)	**27.06.2022 - 30.06.2022**	**Online**

E-ACADEMY

Weiterbildungen bequem vom Büro oder vom Homeoffice aus!

Infos unter https://www.software-quality-lab.com/leistungen/e-academy/



Alternativen zur unterbrechungsfreien Stromversorgung mit Blei-Gel-Akkus

Voller Energie

Bernd Schöne

Blei-Gel-Akku-USVs gelten in Rechenzentren derzeit als alternativlos. Dabei könnten ihnen solche mit Supercaps und Schwungrad sehr wohl Konkurrenz machen.

iX-TRACT

- Unterbrechungsfreie Stromversorgungen überbrücken die Umschaltzeit zwischen Netz- und Notstrom.
- USVs müssen dazu zuverlässig sein und über ausreichende Kapazität verfügen.
- Mit Supercap- und Schwungrad-USVs stehen zwei zuverlässige und nachhaltige, da batteriefreie Alternativen zu den gängigen Modellen mit Blei-Gel-Akkus bereit.
- Derzeit sind ihre Anschaffungskosten hoch, aber die Wartungs- und Erneuerungskosten niedrig.
- Beide Techniken werden im Rahmen der Energie- und Mobilitätswende inzwischen für unterschiedliche Bereiche weiterentwickelt.

Systeme zur unterbrechungsfreien Stromversorgung sind so etwas wie die Mauerblümchen im Rechenzentrum. Jedes RZ hat eins, aber kaum jemand redet über die komplexe und nicht ganz billige Technik. Was eine USV tut, ist längst nicht jedem in der IT klar. Sie gehört schließlich mehr zur Gebäudetechnik, wie die Kühlanlage. Das liegt auch mit daran, dass eine USV nicht in Erscheinung tritt, wenn alles rundläuft.

Nur bei einer Unterbrechung der primären Stromversorgung tritt sie in Aktion. Das aber passiert weit häufiger, als die meisten glauben. Ihre Arbeit beginnt nicht erst bei einem Blackout, sondern bereits bei kleinen, oft nur Zehntelsekunden langen Aussetzern, bei kleinen Abweichungen vom Nennwert der Spannung oder der Frequenz. Eine USV korrigiert diese Störungen oft Hunderte Mal pro Jahr und schützt damit die sensiblen Netzteile der IT-Systeme.

Bei einem längeren Blackout übernimmt die USV die Rolle des Lückenfüllers. Jedes mittlere und größere Rechenzentrum unterhält Notstromaggregate, meist spezielle Dieselaggregate, die schnell starten können, um so zügig die gesamte Stromver-

sorgung des Rechenzentrums zu übernehmen. Sie werden ständig vorgeheizt, um den Startprozess auf die Vorgaben der DIN 6280-13 zu beschleunigen. Nach ihr hat ein Ersatzstromaggregat eine Umschaltzeit von höchstens 15 Sekunden, bis der Diesel also seinen Strom einspeisen kann.

Zudem schreiben Hersteller und Betreiber monatliche Testläufe der Diesel vor. Das erzeugt Abgase und Lärm und belastet etwaige Anwohner. Technisch gibt es aber kaum Alternativen, denn RZs sollen während des Blackouts durchlaufen können. Neben Dieselgeneratoren mit dem passenden Kraftstoffvorrat sind nur Blockheizkraftwerke in der Lage, die Stromversorgung über Stunden oder gar Tage zu übernehmen. Sie lassen sich mit unterschiedlichen Gasarten wie Erd- oder Biogas betreiben, produzieren neben dem Strom aber auch Wärme, was für das RZ aber ein passendes Wärmenutzungskonzept voraussetzen würde, etwa mit Adsorptionskühlung (siehe Artikel „Vorangeschritten“ ab Seite 128). Das und viele andere Argumente bringen RZ-Betreiber an, um am Diesel festhalten zu können.

Batterien und Brennstoffzellen: Nachschub für den Giftmüll

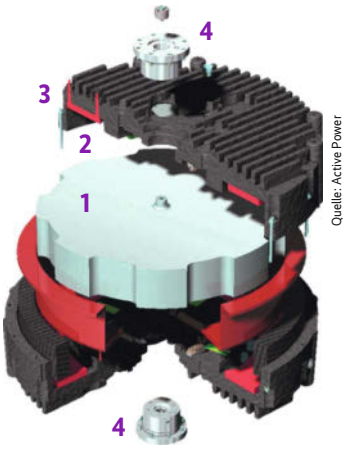
Die Zwischenzeit, bis das Notstromaggregat Energie liefert, überbrücken meist Batteriesysteme. Sie übernehmen die Stromversorgung bei einem Spannungsabfall binnen Millisekunden. Viele RZ-Betreiber vertrauen dem Schnellstart der Diesel nicht und verlangen eine Überbrückungszeit von ein bis zwei Minuten für einen zweiten Startversuch. Ein Blick in die Anschlusswerte eines modernen Rechenzentrums macht deutlich, welche Last bei einem Ausfall auf die USV zukommt.

Den Standard bilden bis heute Bleibatterien, wie sie auch Lkw verwenden. Sie liefern die Energie sofort, ohne Anlaufverzögerung, sind hochbelastbar und preiswert. Für die nötige Kapazität werden meist in einem separaten Raum ganze Regale mit Bleibatterien zusammengeschaltet. Sie halten aber selten länger als sechs Jahre. Danach sind sie fachgerecht zu recyceln, damit das giftige Blei nicht in der Umwelt landet. Wartungsintensive Testläufe sind nötig, um schwächelnde Batterien rechtzeitig zu erkennen. Batterieversagen, so Betreiber von Rechenzentren, ist einer der häufigsten Gründe dafür, dass eine unterbrechungsfreie Stromversorgung nicht gewährleistet werden kann.

Die Hersteller suchen seit Längerem nach Alternativen zu der teuren und nicht sehr zuverlässigen Technik. Alternativen bilden Schwungmassenspeicher, Brennstoffzellen, Lithium-Ionen-Batterien sowie Kondensator-Arrays. Die aus dem Automobilbau bekannten Brennstoffzellen haben sich bislang nicht im Markt etablieren können. Technisch ist eine Brennstoffzelle eine galvanische Zelle, die mithilfe eines Brennstoffs und eines Oxidationsmittels elektrische Energie erzeugt. Sie existieren in zahlreichen Varianten.

Der prinzipielle Aufbau aller Zellen ist gleich: Zwei Elektroden sind durch einen Elektrolyten getrennt, der den Ionenaustausch ermöglicht. Mechanische Komponenten gibt es keine, dafür sehr dünne, semipermeable Trennwände, die praktisch nicht zu warten sind. Im Falle einer Störung muss die Brennstoffzelle meist ausgetauscht werden, zumindest aber in eine Werkstatt. Brennstoffzellen haben zudem den Nachteil, dass sie zwar über eine hohe Energiedichte verfügen, aber nur über eine geringe Leistungsdichte. Leistung aber ist gefordert, wenn ein Puffersystem ein ganzes Rechenzentrum mit Strom versorgen soll. Da nützt es wenig, wenn die Stromquelle auch länger als die geforderten 30 Sekunden durchhalten würde.

An Bedeutung gewinnen langsam Lithium-Ionen-Batterien, doch sind sie deutlich teurer als Bleiakkumulatoren, die verwendeten Rohstoffe sind knapp und viele Fragen zum Recycling noch nicht beantwortet. Nach einigen Tausend Ladezyklen verwandeln sie sich in Sondermüll. Für sie spricht nur die längere Lebensdauer. Viele Anbieter von USVs lassen ihren Kunden die Wahl zwischen den etablierten Bleibatterien und Lithium-Ionen-Akkus.



Quelle: Active Power

1 multifunktionaler Rotor

2 Vakuum im Gehäuse zur Reduktion der Luftreibung

3 Feldspule zur Reduktion der Gewichtskraft auf die Lager und Minimierung der Reibungsverluste

4 austauschbare Lagerkassetten

Vakuum und frei schwebende Magnetlager minimieren die Verluste und erhöhen die Lebensdauer (Abb. 1).

Supercaps: empfindlich gegen hohe Ladespannungen und Temperaturen

Eine weitere Alternative sind die Super- oder Goldkondensatoren, auch Supercaps genannt. Sie haben bereits einen gewissen Marktanteil erobert. Auch sie sind teurer als Bleibatterien, dafür ohne Giftstoffe, zudem recht langlebig und wartungsarm. Vor allem neigen sie nicht wie chemische Akkus dazu, sich selbst zu zersetzen und ätzende Stoffe im Rechenzentrum freizusetzen. Kondensatoren sind neben Widerständen die ältesten Bauelemente der Elektronik. Es gab sie schon vor der ersten Röhre.

Supercaps dagegen sind erst seit gut 40 Jahren erhältlich. Technisch spricht man von Doppelschichtkondensatoren. Sie speichern sehr viel mehr Energie als ihre Vorläufer. Als die Firma Panasonic 1978 die ersten auf den Markt brachte, revolutionierte das die Elektronik, denn nun standen auf einmal Energiespeicher zur Verfügung, die eine Batterie ersetzen konnten, ohne eine zu sein. Goldkondensatoren halten rund zehn Jahre und damit länger als Akkus. Man kann sie nicht tiefentladen und sie zeigen keinen Memory-Effekt. Sie benötigen zum Laden nur eine Gleichspannung und keine Ladeelektronik, da sie nicht überladen werden können. Sie enthalten weder seltene

Quelle: Piller

Die Schwungrad-USV Powerbridge, hier im Querschnitt, liefert eine Leistung im MW-Bereich (Abb. 2).

Schwungräder mit Lagerung durch Magnete oder Spindellager

Typ	Magnetlager	Spindellager
Kosten	erheblich	moderat
Platzbedarf	erheblich	gering
Lagerverluste	nahezu null	0,5 bis 2 kW
Wartung	nicht erforderlich	jährlicher Ölwechsel (bei ölgeschmierten Lagern); Lagerwechsel nach 1 bis 3 Jahren bei fettgeschmierten Lagern
Steifigkeit	nur für den stationären Einsatz geeignet	sehr gut, auch für Stöße in Fahrzeugen geeignet
Zubehör	aufwendige Elektronik, Sensoren und ein Notlauflager	Ölkreislauf mit Filter und Pumpe
Betriebssicherheit	Bei Steuerungsausfall ist ein Totalverlust des Schwungrads wahrscheinlich.	sehr sicher

Rohstoffe noch gesundheitsschädliche Substanzen und auch kein Gold, auch wenn der Markenname anderes suggeriert. Wichtigster Bestandteil ist Aktivkohle. Trotz anfänglich gepfefferter Preise etablierten sie sich schnell als Pufferbatterien für RAM-Speichersysteme. Auch in elektrischen Zahnbürsten oder Rücklichtern von Fahrrädern fanden sie schnell Einzug. Heute findet man sie in Sensoren und IoT-Devices.

Supercaps haben allerdings auch Nachteile. Sie tolerieren keine hohe Ladespannung, schon bei maximal 5 Volt wird es gefährlich. Die meisten Produkte sind auf eine Nennspannung von 2,7 Volt begrenzt, die Ströme dürfen 1 Ampere pro Kondensator nicht übersteigen. Sie vertragen auch keine zu hohen Temperaturen, denn sonst sinkt die Lebenserwartung rapide. Sie verfügen über einen recht hohen Innenwiderstand von 30 Ohm, was ihre Einsatzmöglichkeiten limitiert.

Allerdings wurde in den letzten Jahrzehnten auch viel geforscht. Neue Doppelschichtkondensatoren kamen auf den Markt, mit geringerem Innenwiderstand und damit besser geeignet für Anwendungen wie USVs. Diese Neulinge sind aber noch recht teuer, was einer der Gründe sein kann, warum sie sich in Rechenzentren kaum durchsetzen. In der Steuerungstechnik, wo recht geringe Ströme fließen und die erforderliche Leistung überschaubar ist, haben sie sich aber bereits gut etabliert. Derzeit werden erhebliche Anstrengungen unternommen, um sie zum Speichern der Bremsenergie von Pkw, Lkw und Bahnen verstärkt nutzen zu können. Von diesen neuen Produkten, einer steigenden Nachfrage und sinkenden Preisen könnten auch die Rechenzentren profitieren.

Eine Sonderform stellen die Lithium-Ionen-Kondensatoren dar, eine Mischform aus Akku und Kondensator, bei dem die Kohlenstoffanode mit Lithium dotiert wurde. Damit ist dieser Kondensator nicht so nachhaltig wie klassische Supercaps, dafür aber leistungsfähiger und er soll eine Lebenserwartung von über 15 Jahren haben. Die Firma Socomec nutzt diesen Kondensatortyp für ihre USVs als Puffer.

Schwungräder: Energie mechanisch in einer rotierenden Masse speichern

Schwungmassenspeicher oder rotationskinetische Speicher (RKS) können elektrische Energie ganz ohne Chemie speichern. Energieträger ist eine rotierende Masse. Schon Carl Benz vertraute bei seinem Motordreirad anno 1886 auf eine Schwungmasse. In Forschungseinrichtungen verwendet man sie, um kurzfristig große Energiemengen freisetzen zu können, ohne die Energieversorgung der umliegenden Gemeinden zu gefährden. Beispielsweise verfügt der Forschungscampus in Garching bei München über eine Schwungmassenanlage; Nutzer ist das dortige Max-Planck-Institut für Plasmaphysik. Das Schwungrad wird innerhalb von 20 Minuten beschleunigt. Danach kann es 10 Sekunden lang eine Leistung von 150 MW respektive 580 MVA für das Fusionsexperiment ASDEX Upgrade abgeben.

Die meisten Schwungradspeicher arbeiten elektrisch, indem sie den Rotor mit einem Elektromotor beschleunigen und durch einen elektrischen Generator abbremsen. Der Vorteil: Schwungräder funktionieren über viele Jahre und Hunderttausende von Ladezyklen, ohne zu altern. Klassische Akkus, auch die neuen Lithium-Ionen-Batterien, altern schnell, und eine häufige Nutzung beschleunigt diesen Prozess. Sie büßen dann schnell an Kapazität ein. Zudem dürfen sie weder zu warm noch zu kalt gelagert werden. Schwungmassenspeicher zeigen dieses Verhalten nicht, sie kommen ohne Klimatisierung aus, außer in heißen Umgebungen von über 40 °C.

Batterien benötigen zudem viele Stunden, um nach einer Entladung wieder betriebsbereit zu sein. Schwungräder sind binnen Minuten voll arbeitsfähig. Bleibatterien skalieren anstandslos und erfordern nur geringe Anfangsinvestitionen. Schwungradspeicher kosten etwa das Dreifache. Die Batterien halten allerdings kaum länger als sechs Jahre, die Schwungmassen laut Hersteller 25 Jahre. Nach der Hälfte der Zeit ist ein Austausch der Lager einzukalkulieren.

Batterien müssen zudem vor Tiefentladungen geschützt werden, Schwungmassen nicht. Der Speicher ist leer, wenn die Schwungmasse steht. Ein Nachteil sind die durch Luftreibung und die Lager verursachten Energieverluste, die zwischen 3 und 20 Prozent pro Stunde betragen können. Die Anbieter minimieren die Verluste, indem sie die Schwungmasse in einem Vakuum unterbringen und zudem die Lagerung durch Magneten entlasten (siehe Abbildung 1). Beide Maßnahmen erhöhen auch die Lebensdauer. Durch das Vakuum entfallen Varianten wie Luft- oder Wasserlager, neben der frei schwebenden Lagerung kommen nur klassische Spindellager infrage (siehe Tabelle „Schwungräder mit Lagerung durch Magnete oder Spindellager“).

Kinetik mit Hightech

Schwungmassenspeicher sind mechanische Präzisionsgeräte, die genau justiert und ausgewuchtet sein müssen. Um die nötigen Energiemengen speichern zu können, bewegen sie tonnenschwere Gewichte. Die Gesetze der Physik sind eindeutig: Wer doppelt so viel Energie speichern will, muss eine doppelt so große Masse in Bewegung setzen. Verdoppelt man hingegen die Geschwindigkeit, vervierfacht sich die in der Drehbewegung gespeicherte Energie. Dafür werden die Lager des Systems stärker belastet.

Damit die USV nicht zu groß und zu schwer wird, müssen die Ingenieure einen Kompromiss zwischen Größe und Gewicht der Schwungmasse und der Rotationsgeschwindigkeit finden. Schwungräder mit bis zu 10 000 U/min gelten als Langsamläufer, die sehr große Gewichte benötigen. Die Schnellläufer arbeiten mit 30 000 U/min oder mehr. Sie kommen mit geringeren Gewichten aus und sind daher kleiner. Da Lager und Schwungmasse brechen können, ist in beiden Fällen ein stabiles Gehäuse nötig. Das alles macht Schwungmassenspeicher recht teuer und schwer. Ein belastbarer Boden im Rechenzentrum ist Voraussetzung.

Schon eine langsam rotierende Schwungmasse mit einem geschmiedeten Stahlrotor bringt es bei 7700 U/min auf 939 km/h Spitzengeschwindigkeit und auf eine gespeicherte Energie von

10,5 MJ. Vor allem bei schnell drehenden Systemen mit 20 000 bis über 50 000 Umdrehungen pro Minute fertigen Hersteller die Rotoren aus kohlenstofffaserverstärkten Kunststoffen (CFK) oder Kevlar statt aus Stahl. Es ist leichter, die mechanische Energie aufgrund der hohen Drehzahl aber vergleichbar. Die gespeicherte kinetische Energie eines Schwungrads ist proportional zur Masse seines Rotors, zum Quadrat seines Radius und zum Quadrat seiner Rotationsgeschwindigkeit (U/min).

> Die in einem Schwungrad mit dem Trägheitsmoment J gespeicherte Energiemenge ist $E_{rot} = 1/2 \times J \times \omega^2$, wobei ω die Winkelgeschwindigkeit ist, also 2 π mal Drehzahl. Die in einem Schwungrad gespeicherte Energie ergibt sich als Produkt vom Quadrat der Winkelgeschwindigkeit und dem Massenträgheitsmoment geteilt durch zwei. Das Massenträgheitsmoment errechnet sich aus der Masse und dem Quadrat des Radius.

Die Eiskunstläuferin zeigt die Zusammenhänge am elegantesten. Sie holt mit ausgebreiteten Armen Schwung, dreht sich und zieht dann die Arme an sich. Sie wird immer schneller. Dahinter steckt das Gesetz der Drehimpulserhaltung. Zwar nimmt die Reibung der Kufen der Eiskunstläuferin etwas Bewegungsenergie, viel stärker wirkt aber das Gesetz von sich drehender Masse und dem Abstand der Masse zum Mittelpunkt der Drehbewegung. Neben der Ausgangsgeschwindigkeit gehen nur diese Parameter in die Berechnung des Drehimpulses ein. Verlagert sie die Arme zum Körper, erzwingt der Erhaltungssatz eine höhere Geschwindigkeit. Ohne Reibung würde sie sich ewig drehen. Aber bis die Energie komplett verbraucht ist, dauert es lange genug, um etliche Kunstfiguren zu generieren.

Die Zukunft der Schwungräder

Der Bonner Sicherheitsberater und Rechenzentrumsplaner Rainer von zur Mühlen hat die Planung zahlreicher Rechenzentren für Großbanken und Behörden geleitet und schätzt die Situation so ein: „Dynamische USVs mit kinetischem Energiespeicher sind in der Erstanschaffung zumeist etwas teurer als Bleibatterien, dafür aber sicherer. Wir haben sie vor allem in Hochverfügbarkeitsrechenzentren gerne eingesetzt, weil sie weniger Ausfallrisiken bergen. Bei Anlagen mit Bleibatterien kommt es vor, dass ein Totalausfall der USV schon eintritt, wenn eine einzelne Batterie versagt. Da der Ausfall zumeist eine ganze Serie der Batterien mitreißt, kommt es vor, dass Ersatzbeschaffungen der Batterien sich über Wochen oder gar Monate hinziehen können. Batterieversagen passiert häufiger, als man denkt. Deswegen werden die Blöcke meistens weit vor dem statistischen Lebensdauerende ersetzt. Um Versagen abzuwenden, sollte man auch die Zellen einzeln und umfassend messtechnisch überwachen. Was verflixt teuer ist.“

Auch Schwungräder könnten von massiven Forschungsanstrengungen für die Energie- und Verkehrswende profitieren. Geplant ist etwa, Schwungmassen in Ladestationen von Elektroautos zu integrieren. Besitzer von Elektroautos wollen den Ladevorgang aus nachvollziehbaren Gründen so schnell wie möglich hinter sich bringen. Dazu müssen enorme Energiemengen binnen weniger Minuten in die Batterie gelangen. Außerhalb von Ballungszentren kann die Stromversorgung die hohen Ströme nicht zur Verfügung stellen. Hier sollen Schwungräder als Puffer einspringen.

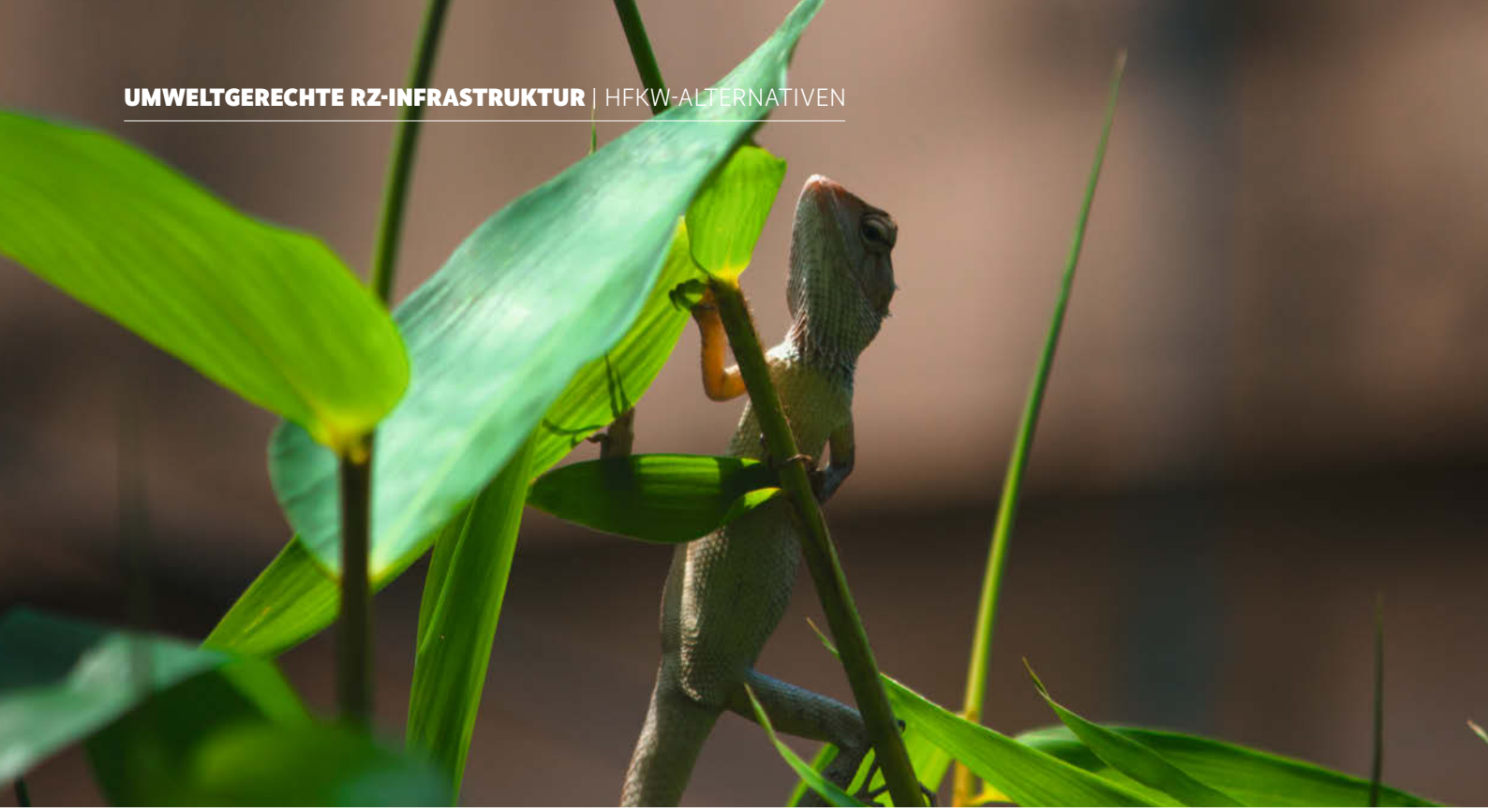
Die Adaptive Balancing Power GmbH stellte auf der E-Mobility-Messe Power2Drive 2021 in München den Amperage HPC-Booster vor. Er soll mit allen gängigen Schnellladesäulen kompatibel sein und bis zu 350 kW Energie speichern. Auch die Betreiber von Windparks haben die Schwungmassen entdeckt. Eventuell werden diese demnächst in die Parks integriert, um bei Flaute die Stromversorgung zu übernehmen.

Vielleicht schrauben die neuen Anwendungsgebiete die Stückzahlen nach oben und reduzieren so die Kosten. Die deutsche Firma Piller liefert unter dem Namen Powerbridge und seit der Übernahme von Active Power unter dem Namen CleanSource verschiedene Schwungmassenspeicher mit einer Leistung im MW-Bereich (siehe Abbildung 2).

Vor allem könnten neue Anstrengungen in der Forschung helfen, die Nachteile der Schwungmassen zu minimieren. Derzeit sind sie teuer und schwer: USVs mit Schwungmassen erreichen Energiedichten von nur 10 Wh/kg, brauchen also 100 kg pro gespeicherter Kilowattstunde, während Lithium-Ionen-Batterien rund 180 Wh/kg bieten. Mit leichteren Schwungrädern aus Kunststoff, höheren Drehzahlen und besseren Lagern besteht hier noch Aufholpotenzial. Bis zu 100 000 U/min sollten mit Schwungmassen aus Kunststoff möglich sein. (sun@ix.de)

Bernd Schöne

ist freier Journalist.



Natürliche Kältemittel für die RZ-Klimatisierung

Grün getarnt

Dr. Daniel de Graaf

Viele Alternativen zu den umwelt- und klimaschädlichen Kältemitteln, die die EU vom Markt haben will, sind eher Mogelpackungen. Dabei geht es auch mit natürlichen Kältemitteln.

Elektrische Energie, die Server aufnehmen und in Wärme umwandeln, muss vollständig abgeführt werden, will man keinen Systemausfall durch Überhitzung riskieren. Deshalb kommt der Klimatisierung im Rechenzentrumsbetrieb eine ebenso wichtige Rolle zu wie der Stromversorgung. Hier herrscht noch immer die Luftkühlung vor. Sie führt einen Luftstrom im Kreis, der durch die Kältemaschine abgekühlt und als Zuluft über einen Doppelboden in das Rechenzentrum eingebracht wird. Die Lüfter in den Servern saugen sie an und kühlen damit die sich erwärmenden Komponenten. Die erwärmte Luft steigt nach oben, wo Ventilatoren sie über den Wärmeübertrager des Klimasystems führen, dort wird sie erneut abgekühlt, damit sie als Zuluft wieder zur Verfügung steht.

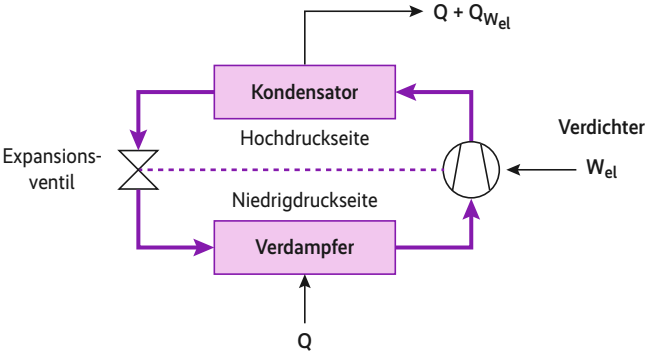
Die für die Luftkühlung am häufigsten verwendeten Kältemaschinen greifen ihrerseits auf das Kernprinzip der Kältetechnik zurück, den Kältekreislauf. Jeder Kühlschrank und jedes RZ-Kühlsystem mit Luftkühlung macht es sich zunutze. Dabei entzieht eine bei niedriger Temperatur siedende Flüssigkeit, das Kältemittel, einem Medium, in diesem Fall der Luft des Rechenzentrums, durch Verdampfen die Wärme (siehe Abbildung 1). Das jetzt gasförmige Kältemittel wird im Kompressor verdichtet, das heißt auf höheren Druck und höhere Temperatur gebracht. Im Verflüssiger oder Kondensator wird die Wärme an die Umgebung abgegeben, wodurch sich das Kältemittel wieder verflüssigt. Über ein Expansionsventil wird das Kältemittel anschließend wieder auf niedrigeren Druck entspannt.

Obwohl man umgangssprachlich gern die Begriffe Kältemaschine oder Kälteerzeuger verwendet, erzeugt dieser Prozess keine Kälte, sondern verschiebt Wärme (Q) gegen ein Temperaturgefälle – zum Beispiel aus dem kühlen Rechenzentrum hinaus ins sommerlich heiße Freie. Den Kompressor oder Verdichter treibt ein Elektromotor an, der einen erheblichen Strombedarf hat.

Eine Riesenbelastung für das Klima

In fast jedem luftgekühlten Rechenzentrum kommen Kältemaschinen und damit Kältemittel zum Einsatz. Bei kleineren Kälteleistungen, die für Serverräume und kleinere Rechenzentren ausreichen, arbeiten sie mit Direktverdampfung, das heißt, das Kältemittel eines Split-Gerätes steht direkt mit der zu kühlenden Raumluft über den Verdampfer in Kontakt. Bei einem mittleren und großen Kältebedarf setzt man dagegen auf Flüssigkeitskühler, deren Verdampfer mit einem Kälteträger – Wasser oder Sole – beaufschlagt werden, sodass die Serverabwärme indirekt über eine Flüssigkeit an das verdampfende Kältemittel abgegeben wird.

In der Regel kommen teilfluorierte Kohlenwasserstoffe (HFKW) als Kältemittel zum Einsatz. Sie haben die früher verwendeten und mittlerweile verbotenen Fluorchlorkohlenwasserstoffe (FCKW) ersetzt und sind nicht ozonschichtschädi-



Im Kältekreislauf wird die Summe aus aufgenommener Wärme (Q) und aufgewendeter elektrischer Arbeit (W_{el}) über den Kondensator an die Umgebung abgegeben. Spontan, also ohne Energieaufwand für den Antrieb des Kompressors, würde dieser Prozess nicht ablaufen (Abb. 1).

gend, besitzen jedoch wie FCKW eine erhebliche Klimawirkung (siehe Tabelle „Treibhauspotenzial ausgewählter Kältemittel“).

Die Klimawirkung eines Treibhausgases, das Treibhauspotenzial, gibt das GWP (Global Warming Potential) an, die Referenzgröße ist Kohlendioxid (CO_2) mit einem GWP von 1. Ein häufig verwendeter HFKW in Kältemaschinen, der auch in Rechenzentren eingesetzt wird, ist R134a mit einem GWP von 1430. 1 kg R134a bewirkt den gleichen Treibhauseffekt wie 1,43 t CO_2, man nennt das auch die äquivalente Menge CO_2 und gibt als Einheit CO_2-Äquivalent, CO_2-eq oder CO_2e an. Aufgrund ihres hohen Treibhauspotenzials sind HFKW sehr schädlich für das Klima, selbst wenn davon eher kleine Mengen im Vergleich zu anderen Treibhausgasen wie CO_2 oder Methan freigesetzt werden.

Die Kältemittelemissionen entstehen beim Befüllen, im Betrieb und bei der Wartung und Entsorgung von Kälteanlagen. Jeder Flüssigkeitskühler verliert im Betrieb im Durchschnitt jährlich 3 Prozent der Kältemittelfüllmenge, Havarien eingerechnet. Dazu kommen die Entsorgungsemissionen von 18,9 Prozent der verbliebenen Füllmenge. Unterm Strich verursacht das Klimasystem des Rechenzentrums daher nicht nur eine signifikante Klimaschädigung durch indirekte Emissionen aufgrund seines Energiebedarfs, sondern auch durch direkte Emissionen, geschuldet dem Verlust des Kältemittels.

Treibhauspotenzial ausgewählter Kältemittel (Einzelstoffe und Gemische)

Kältemittel	Chemische Bezeichnung / Zusammensetzung	GWP$_{100}$* [1]
R32	Difluromethan	675
R134a	Tetrafluorethan	1430
R125	Pentafluorethan	3500
R227ea	Heptafluorpropan	3220
R1234yf	2,3,3,3-Tetrafluorpropen	4 [2]
R1234ze(E)	Trans-1,3,3,3-Tetrafluorprop-1-en	7 [2]
R407C	23 % R32, 25 % R125, 52 % R134	1774
R410A	50 % R32, 50 % R125	2088
R513A	44 % R134a, 56 % R123yf	631
R515B	8,9 % R227ea, 91,1 % R1234ze	293
R290	Propan	3 [2]
R600a	Isobutan	3 [2]
R717	Ammoniak	0
R718	Wasser	0
R744	Kohlendioxid	1

* Treibhauspotenzial, bezogen auf 100 Jahre

iX-TRACT

- Die oft als Kältemittel eingesetzten teilfluorierten Kohlenwasserstoffe (HFKW) schädigen zwar nicht die Ozonschicht, dafür aber mit ihrem hohen Treibhauspotenzial das Klima.
- Die EU verknappt die verfügbare HFKW-Menge absichtlich, damit Kunden auf klimafreundlichere Alternativen umsteigen.
- Die von der Kältemittelindustrie eingeführten ungesättigten HFKW (uHFKW) sind zwar weniger klimaschädlich, dafür aber im Brandfall hochgefährlich. Deshalb setzt die Industrie auf nicht brennbare Mischungen, die aber keine echten Alternativen sind.
- Bleiben die natürlichen, halogenfreien Kältemittel, die auch der Blaue Engel für Rechenzentren vorschreibt.

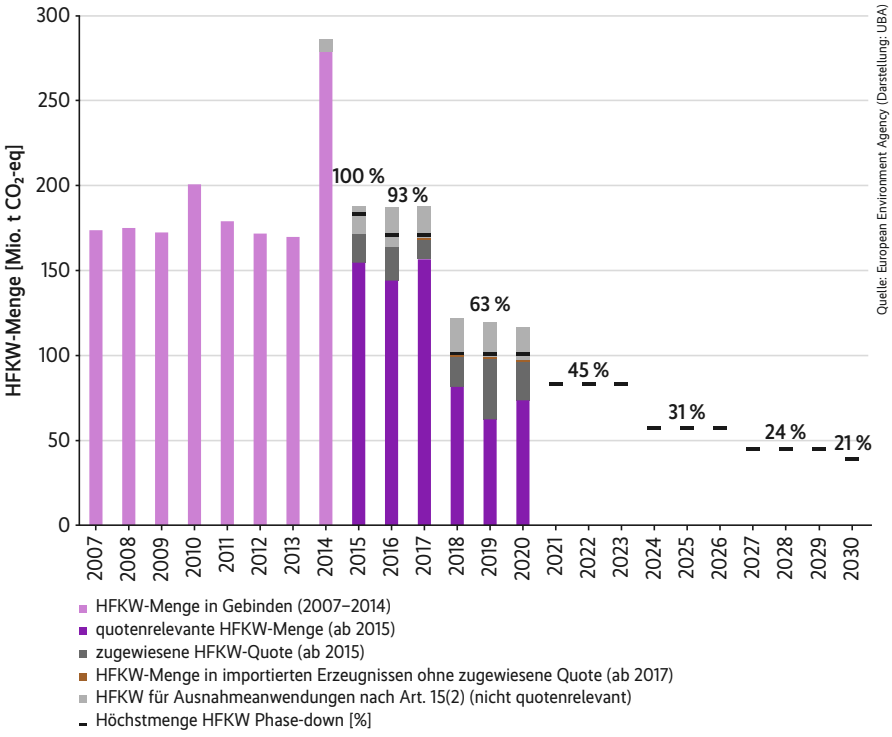
Der lange Weg der Reduktion

Insgesamt waren die HFKW in der EU und dem Vereinigten Königreich im Jahr 2019 für 92,5 Millionen Tonnen CO_2-Äquivalente verantwortlich, das sind 2,3 Prozent des Gesamttreibhausgasausstoßes dieser Länder. Um Verbrauch und Emissionen von HFKW deutlich zu senken, hat die EU die bereits bestehende Regelung der F-Gase (fluorierte Treibhausgase), die F-Gas-Verordnung Nr. 517/2014, geändert und ein neues Element aufgenommen, genannt das HFKW Phase-down [2]. Nach diesem Schema wird die in die EU jährlich eingeführte Menge an HFKW bereits seit 2015 schrittweise bis ins Jahr 2030 auf 21 Prozent der Ausgangsmenge abgesenkt. Als Ausgangsmenge gilt die Durchschnittsmenge der Jahre 2009 bis 2012: 183,1 Millionen Tonnen CO_2-Äquivalent, sie soll um 79 Prozent reduziert werden auf dann 38,5 Millionen Tonnen (siehe Abbildung 2).

Veranschlagt wird hierfür jedoch nicht die Masse der HFKW, sondern Tonnen CO_2-Äquivalent, also die Masse HFKW mal GWP-Wert des Kältemittels. Damit verknappt die EU das Angebot und schafft einen Anreiz, insbesondere Kältemittel mit sehr hohem Treibhauspotenzial durch klimafreundlichere Alternativen zu ersetzen. Jeder Inverkehrbringer von HFKW muss die im vergangenen Jahr auf den Markt gebrachte HFKW-Menge bei der EU melden und bekommt dafür eine Quote für das Folgejahr zugeteilt. Heute ist die Vermarktungsmenge bereits auf 55 Prozent reduziert.

Die beabsichtigte Verknappung des Angebots an HFKW-Kältemitteln konnte man nicht nur an den vereinzelt auftretenden Lieferengpässen, einer bisher nicht dagewesenen Zahl an HFKW-Diebstählen und Fällen illegaler Einfuhr im Maßstab mehrerer Tonnen erkennen, sondern vor allem an den steigenden Preisen. Im Vergleich zur Zeit vor dem Phase-down mussten Kunden kräftige Aufschläge zahlen, die im zweiten Quartal 2018 in Preissteigerungen von 700 Prozent und mehr gipfelten. Mittlerweile haben sich die Preise bei 250 bis 300 Prozent im Vergleich zu 2014 stabilisiert (siehe Abbildung 3).

Das schrumpfende HFKW-Angebot sollte Betreiber von Rechenzentren spätestens jetzt aufhorchen lassen, denn damit sind nicht nur steigende Preise verbunden, sondern auch eine Gefährdung ihrer Verfügbarkeit. Eine Havarie der Kälteanlage mit großen Kältemittelverlusten kann im ungünstigen Fall auf-



Das in der F-Gas-Verordnung vorgeschriebene HFKW Phase-down nimmt als 100 Prozent die HFKW-Ausgangsmenge von 183,1 Mio. t CO_2-eq an, die 2015 auf dem europäischen Markt in Verkehr gebracht werden durfte. Sie wird bis 2030 schrittweise auf 38,5 Mio. t gesenkt (Abb. 2).

grund der HFKW-Knappheit nicht zeitnah behoben werden und zu einem wochenlang andauernden RZ-Ausfall führen.

Neue fluorierte Stoffe: Ausweg oder Sackgasse?

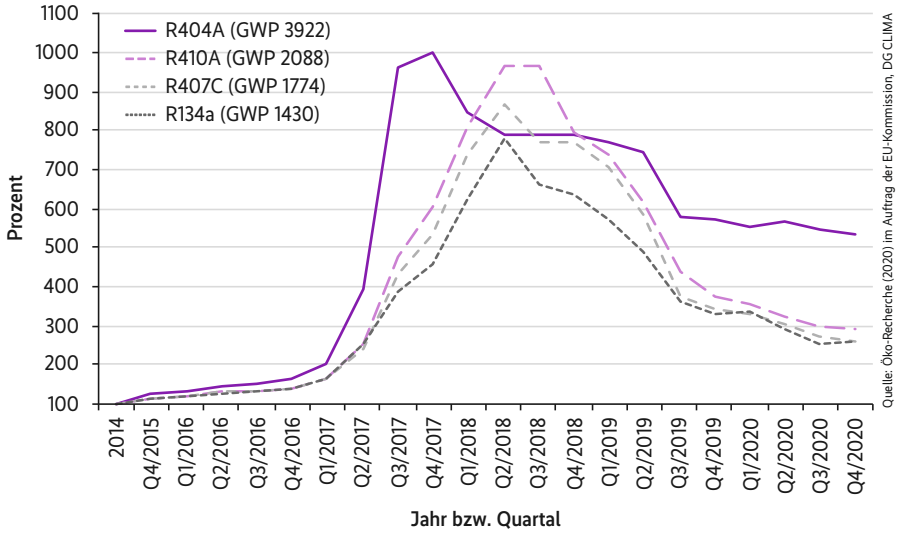
Die Zeichen, die darauf hindeuten, dass es mit den bisherigen, stark klimaschädlichen Kältemitteln in der RZ-Klimatisierung nicht weitergehen kann, sind zahlreich und unübersehbar. Die Frage ist daher, welche alternativen Stoffe als Ersatz taugen und dabei nachhaltig sind, also sowohl energieeffizient als auch klima- und umweltfreundlich. Diese Frage haben die Kältemittelhersteller bereits wenig überraschend beantwortet: fluorierte beziehungsweise halogenierte Stoffe! Diese werden seit einigen Jahren, flankiert durch gut budgetierte Werbekampagnen, als klimafreundlich, teilweise sogar als nachhaltig vermarktet. Illustriert wird dies gern mit grünen Symbolen, etwa einem stilisierten grünen Blatt. Doch bleibt die Frage, ob diese neuen Kältemittel tatsächlich das halten, was ihre Hersteller versprechen.

Um die Klimaschädlichkeit zu reduzieren, haben die Hersteller der HFKW-Kältemittelchemie eine Kohlenstoff-Kohlenstoff-Doppelbindung hinzugefügt und damit die Reaktivität der neuen Stoffe deutlich erhöht. Die als Hydrofluorolefine (HFO) vermarkteten ungesättigten HFKW (uHFKW) zersetzen sich dadurch in der Atmosphäre sehr schnell, was das Treibhauspotenzial senkt. Die bekanntesten und verbreitetsten Vertreter sind R1234yf und R1234ze(E), die GWP-Werte von 4 respektive 7 aufweisen. Die höhere Reaktivität führt dazu, dass uHFKW brennbar sind, wobei im Brandfall und an heißen Oberflächen Verbrennungsprodukte wie giftiger und stark ätzender Fluorwasserstoff (HF) entstehen. Deshalb wurden zahlreiche Mischungen mit herkömmlichen HFKW wie R134a erzeugt, um ein nicht brennbares Produkt mit niedrigerem GWP zu erhalten. Dazu zählen die Kältemittel R513A mit 56 Prozent R1234yf und 44 Prozent R134a oder R515B, bestehend aus 91,1 Prozent R1234ze(E) und 8,9 Prozent R227ea. Diese Mischungen haben aber doch noch recht hohe GWP-Werte von 631 beziehungsweise 293.

Viele neue Nachteile im Gepäck

Mit den fluorierten Stoffen als vermeintlicher Alternative bleibt die Abhängigkeit vom knappen Rohstoff Flussspat (CaF_2) erhalten, der als Fluorlieferant dient und nur an sehr wenigen Orten der Welt vorkommt. Flussspat ist einer von dreißig Rohstoffen, die die EU als kritisch einstuft, da deren hohe wirtschaftliche Bedeutung mit einem großen Versorgungsrisiko einhergeht (siehe Artikel „Nicht in den Himmel“ ab Seite 22).

In größeren Flüssigkeitskühlern mit mehr als 250 kW Nennkälteleistung, die auch in Rechenzentren zum Einsatz kommen, wird vermehrt R1234ze(E) als R134a-Ersatz verwendet. In ihrer Energieeffizienz sind die beiden Stoffe vergleichbar. Allerdings ist die Kälteleistung von R1234ze(E) um gut 20 Prozent gerin-



Die Darstellung der Großhandelspreise gängiger HFKW-Kältemittel beruht auf den durchschnittlichen Preisen von drei großen Gashandelsunternehmen, normiert auf das Referenzjahr 2014 (Abb. 3).

ger, was zur Folge hat, dass die Maschinen 20 Prozent größer ausfallen müssen, um die gleiche Leistung wie eine R134a-Maschine zu erzielen. Entsprechend höher ist der Materialbedarf.

Ressourceneffizienz sieht aber anders aus. Und sollte eine Neuentwicklung nicht deutlich energieeffizienter sein als die Technik, die sie ersetzt? Sogar als gänzlich unbrauchbare R134a-Alternative müsste R1234ze(E) angesehen werden, falls sich die Erkenntnisse der Universität New South Wales in Sydney, Australien, bestätigen. Die Wissenschaftler wiesen nach, dass R1234ze(E) teilweise zu Trifluormethan (HFKW-23) zerfällt [3]. Der Stoff, der unter der Bezeichnung R23 auch als Kältemittel in der Tiefstkühlung verwendet wird, hat ein sehr hohes GWP von 14 800. Das Zerfallsprodukt eingerechnet, würde nach diesem Befund R1234ze(E) als vorgeblich klimafreundlicher R134a-Ersatz das Klima mit einem GWP von 1000 ähnlich schwer schädigen wie R134a selbst.

Ammoniak und Wasser im Vergleich

Kältemittel	Ammoniak (NH_3)	Wasser (H_2O)
Normalsiedepunkt (p = 1 bar)	–33,3 °C	100 °C
Siededruck (10 °C)	6,2 bar (abs)	0,012 bar (abs)
Kondensationsdruck (40 °C)	15,6 bar (abs)	0,074 bar (abs)
Dampfdichte (10 °C)	4,876 kg/m³	0,0094 kg/m³
Volumetrische Kälteleistung	5966 kJ/m³	23 kJ/m³
Brennbarkeit	15–34 Vol.-%	nicht brennbar
Toxizität	sehr toxisch	nicht toxisch
Materialverträglichkeit	Ammoniak-Wasser-Gemisch ungeeignet für Buntmetalle und Aluminium; geeignet für Stahl (in Kombination mit Korrosionsinhibitor), Edelstahl, Teflon	bei Zugabe von Lithiumbromid zur Reduktion des Dampfdrucks in Verbindung mit Luft: hohe Korrosionswirkung
Ozonabbaupotenzial (ODP)	0	0
Treibhauspotenzial (GWP)	0	0

Nachhaltige Alternativen mit natürlichen Kältemitteln

Glücklicherweise gibt es Alternativen, mit denen sich die RZ-Klimatisierung nicht nur klimafreundlich, sondern auch nachhaltig gestalten lässt: halogenfreie, natürliche Kältemittel. Verbindungen wie Kohlenwasserstoffe, Ammoniak (R717) und Kohlendioxid (R744) werden als natürlich bezeichnet, da sie in signifikanten Mengen in der Natur zu finden sind. Sie enthalten kein Chlor oder Fluor, sind gut erforscht und kommen teilweise seit über 100 Jahren in Kältemaschinen zum Einsatz. Ähnlich böse Überraschungen für die Umwelt wie bei den uHFKW durch unbekannte Abbauwege lassen sich hier ausschließen.

Außerdem sind natürliche Kältemittel aufgrund ihrer thermodynamischen Eigenschaften wie der volumetrischen Kälteleistung und der niedrigen Siedetemperatur den HFKW- und uHFKW-Kältemitteln überlegen. Insbesondere Ammoniak zeigt bei der spezifischen Wärmekapazität, Wärmeleitfähigkeit und spezifischen Verdampfungsenthalpie Spitzenwerte, die eine herausragend energieeffiziente Anlagentechnik ermöglichen.

Das war bereits Carl Linde bewusst, der 1876 die erste Ammoniakkältemaschine für eine Brauerei entwarf. Dass er sich mit Ammoniak sogar für einen besonders klimafreundlichen Stoff ganz ohne Treibhauspotenzial entschieden hatte, konnte der Ingenieur damals noch nicht wissen. Auch alle anderen natürlichen Kältemittel weisen sehr niedrige GWP-Werte von maximal 6 auf. Allerdings: Ammoniak ist entflammbar, toxisch und verträgt sich nicht mit Buntmetallen und Aluminium (siehe Tabelle „Ammoniak und Wasser im Vergleich“). Das gilt es beim Anlagendesign und der Installation zu berücksichtigen.

Für Rechenzentren kommen für eine HFKW-freie Klimatisierung im Wesentlichen Kältemaschinen mit den Kältemitteln Ammoniak, Wasser (R718) oder dem Kohlenwasserstoff Propan (R290) infrage. Flüssigkeitskühler mit Propan sind in einem Leistungsbereich von 5 kW bis 1,2 MW erhältlich. Ammoniak-

Quelle: STACKIT

STACKIT kühlt sein Colocation-RZ mit zwei 1,8-MW-Ammoniak-Kältemaschinen (Abb. 4).

aggregate sind serienmäßig ab 100 kW bis über 6 MW verfügbar. Damit lässt sich der Kältebedarf eines Serverraums genauso wie der eines großen Colocation-Rechenzentrums decken. Wie bei den uHFKW ist die Brennbarkeit von Propan und Ammoniak bei der Anlagenplanung und dem Betrieb normenkonform zu berücksichtigen. Gleiches gilt für die Toxizität von Ammoniak.

Kältemittelleckagen lassen sich auch bei einem auf Dichtheit ausgelegten Anlagendesign nie gänzlich ausschließen. Durch Außenaufstellung, Belüftung des Maschinenraums und weitere Maßnahmen zum Vermeiden zündfähiger Atmosphären ist die Sicherheit im Betrieb jedoch gewährleistet. Zahlreiche Beispiele kleiner und großer Rechenzentren belegen, dass sich Anlagen mit natürlichen Kältemitteln für die IT-Klimatisierung eignen. Der Colocation-Anbieter STACKIT etwa kühlt mehrere seiner Rechenzentren mit Ammoniak-Flüssigkeitskühlern, die Nennkälteleistungen im Megawattbereich aufweisen (siehe Abbildung 4). Nur wenige Kilowatt umfassen hingegen die R290-Maschinen für die Serverraumkühlung in Logistikzentren eines großen Lebensmitteleinzelhandelsunternehmens (siehe Abbildung 5).

Serverkühlung auch ohne Kältemaschine

Alle Kältemittelsorgen vollständig los ist, wer auf maschinelle Kühlung verzichtet. Schöpft man den zulässigen Temperaturbereich für den ausfallsicheren Serverbetrieb gemäß ASHRAE-Richtlinie aus, ist eine Zulufttemperatur von 32 °C im Rechenzentrum tolerierbar [4]. Dieses Temperaturniveau ist mit adiabater Verdunstungskühlung zu jedem Zeitpunkt im Jahr, also auch bei extremer Hitze von über 40 °C, gewährleistet, da die Verdunstung von Wasser eine Abkühlung um 10 K und mehr gegenüber der Außenluft mit sich bringt. Eine Kältemaschine erübrigt sich damit.

Quelle: Frigoteam Handels GmbH

Diese zwei R290-Flüssigkeitskühler sind in Außenaufstellung mit jeweils 12 kW Nennkälteleistung installiert (Abb. 5).

Gleiches gilt für die Flüssigkühlung, die die Wärme direkt von den sich erwärmenden Komponenten abführt (siehe Artikel „Ohne Umwege“ ab Seite 140). Je nach System betragen die Vorlauftemperaturen 45 bis 50 °C, die des Rücklaufs zwischen 50 und 55 °C. Die Rücklauftemperaturen sind damit hoch genug, um zum Heizen oder zur Warmwasserbereitung in der unmittelbaren Umgebung genutzt zu werden (siehe Artikel „Aufgefangen“ ab Seite 146). Für die Einspeisung in ein Fernwärmenetz mit Temperaturniveaus von 70 °C und mehr sind allerdings Wärmepumpen nötig.

Der anlagentechnische und organisatorische Aufwand, um RZ-Abwärme auch über die Grundstücksgrenzen hinaus zu nutzen, wird bisher eher gemieden. Mittelfristig wird das jedoch zur Regel werden, sollte die neue Bundesregierung ihre eigenen Ambitionen ernst nehmen. Im Koalitionsvertrag ist zu lesen: „Wir werden Rechenzentren in Deutschland auf ökologische Nachhaltigkeit und Klimaschutz ausrichten, u. a. durch Nutzung der Abwärme.“ Damit ist klar, dass sich in deutschen Rechenzentren in Sachen Nachhaltigkeit insgesamt noch deutlich mehr als bisher tun wird. Da der Koalitionsvertrag den Blauen Engel für Rechenzentren der öffentlichen Hand zum Standard erhoben hat, wird es auch für die HFKW-Kälteanlagen in Colocation-Rechenzentren zusehends enger, denn er schreibt halogenfreie Kältemittel als verpflichtendes Vergabekriterium fest (siehe Artikel „Dickicht der anderen Art“ ab Seite 100).

Fazit

Der Einsatz von HFKW-Kältemitteln in Kälte- und Klimaanwendungen wird massiv zurückgedrängt, auch vor der Rechenzentrumsklimatisierung macht diese Entwicklung nicht halt. Für einen zukunftssicheren Betrieb empfehlen sich Kälteanlagen mit natürlichen Kältemitteln, die in allen Größen zur Verfügung stehen und sich in der Praxis bewährt haben. (sun@ix.de)

Quellen

[1] Intergovernmental Panel on Climate Change (IPCC); Climate Change 2007 – The Physical Science Basis; Contribution of Working Group I to the fourth Assessment Report of the IPCC; Cambridge and New York 2007

[2] EU-Verordnung Nr. 517/214 des Europäischen Parlaments und des Rates vom 16. April 2014 über fluorierte Treibhausgase (F-Gas-Verordnung)

[3] Jyoti S. Campbell, Scott H. Kable, Christopher S. Hansen; Photodissociation of CF_3CHO provides a new source of CHF_3(HFC-23) in the atmosphere: implications for new refrigerants; 2021 (Vorveröffentlichung)

[4] ASHRAE (American Society of Heating, Refrigeration and Air-Conditioning Engineers); Thermal Guidelines for Data Processing Environment; 4th Edition; Atlanta 2015

Dr. Daniel de Graaf

ist wissenschaftlicher Mitarbeiter im Umweltbundesamt und dort zuständig für Industriekälte und stationäre Klimatisierung.

MAGAZIN FÜR PROFESSIONELLE INFORMATIONSTECHNIK

iX Special 2022 – Green IT

Postfach 61 04 07, 30604 Hannover; Karl-Wiechert-Allee 10, 30625 Hannover

Redaktion: Telefon: 0511 5352-387, Fax: 0511 5352-361, E-Mail: post@ix.de
Abonnements: Telefon: 0541 80009-120, Fax: 0541 80009-122, E-Mail: leserservice@heise.de

Herausgeber: Christian Heise, Ansgar Heise

Chefredakteur: Dr. Oliver Diedrich (odi@ix.de) -616

Konzeption und redaktionelle Leitung: Susanne Nolte (sun@ix.de) -689

Redaktionelle Mitarbeit: Nicole Bechtel (nb@ix.de) -378, Moritz Förster (fo@ix.de) -374, André von Raison (avr@ix.de) -377, Ute Roos (ur@ix.de) -535, Bert Ungerer (un@ix.de) -368, Jonas Volkert (jvo@ix.de) -286, Ulrich Wolf (ulw@ix.de) -379

Ständiger Mitarbeiter: Jürgen Seeger (js@ix.de)

Autoren dieser Ausgabe: Dr. Ludger Ackermann, Fridtjof Chwoyka, Andreas Fertig, Dr. Daniel de Graaf, Jens Gröger, Achim Guldner, Tobias Haar, Dr. Dirk Harryvan, Dr. Eva Kern, Marina Köhn, Dr. Sandro Kreten, Prof. Dr. Volker Lindenstruth, Martin Lippert, Rudolf Meier, Prof. Dr. Stefan Naumann, Susanne Nolte, Frank Pientka, Ariane Rüdiger, Dr. Alexander Schatten, Bernd Schöne, Jürgen Seeger, Hubert Sieverding, Detlef Thoms, Dr. Béla Waldhauser

Redaktionsassistenz: Carmen Lehmann (cle@ix.de) -387, Michael Mentzel (mm@ix.de) -153

Abbildungen © Adobe Stock

Titel: Idee: Susanne Nolte; Gestaltung: Lisa Hemmerling, Adobe Stock

Aufmacher: Idee: Susanne Nolte, Gestaltung: Lisa Hemmerling
Bildmotive: Aamon (S. 134), alan1951 (S. 128), Alex Bramwell (S. 67), Antony (S. 122), black-rabbit3 (S. 140), charles (S. 118), dangdumrong (S. 12), enjoynz (S. 48, 62), euruspluvia (S. 40), EwaStudio (S. 100), eyetronic (S. 8), geoffrey (S. 96), kanzilyou (S. 84, 105), languste15 (S. 36), Li Ding (S. 5, 71, 109), LinZay (S. 30), Maksym (S. 58), Manuel (S. 22), Min Cheol Kim (S. 92, 146), Nedilko (S. 110), photocech (S. 16), Pietro D'Antonio (S. 3), Roman (S. 7), Stefan Körber (S. 54), Subbotina Anna (Titelmotiv, S. 4), Suphatthra China (S. 72), Unclesam (S. 4, 39), Wildspaces (S. 78), wusuowei (S. 151)

Layout und Satz: Beatrix Dedek, Madlen Grunert, Lisa Hemmerling, Cathrin Kapell, Kirsten Last, Steffi Martens, Marei Stade, Matthias Timm, Heise Medienwerk, Rostock

Chefin vom Dienst: Barbara Gückel

Korrektorat: Barbara Gückel; Martina Lübke, Heise Medienwerk, Rostock

Hergestellt und produziert mit Xpublisher: www.xpublisher.com

Xpublisher-Technik: Melanie Becker, Kevin Harte, Thomas Kaltschmidt, Pascal Wissner

Verlag und Anzeigenverwaltung:
Heise Medien GmbH & Co. KG, Postfach 61 04 07, 30604 Hannover; Karl-Wiechert-Allee 10, 30625 Hannover; Telefon: 0511 5352-395, Fax: 0511 5352-129

Geschäftsführung: Ansgar Heise, Beate Gerold

Mitglieder der Geschäftsleitung: Jörg Mühle, Falko Ossmann

Anzeigenleitung: Michael Hanke -167, E-Mail: michael.hanke@heise.de, www.heise.de/mediadaten/ix

Anzeigenpreise: Es gilt die Anzeigenpreisliste Nr. 33 vom 1. Januar 2022.

Leiter Vertrieb und Marketing: André Lux -299

Werbeleitung: Julia Conrades -156

Druck: Dierichs Druck + Media GmbH & Co. KG, Frankfurter Straße 168, 34121 Kassel

Sonderdruck-Service: Julia Conrades -156

Verantwortlich: Textteil: Dr. Oliver Diedrich; Anzeigenteil: Michael Hanke

Abo-Service: Heise Medien GmbH & Co. KG, Leserservice, Postfach 24 69, 49014 Osnabrück, Telefon: 0541 80009-120, Fax: 0541 80009-122, E-Mail: leserservice@heise.de

Vertrieb Einzelverkauf (auch für Österreich, Luxemburg und Schweiz):
DMV DER MEDIENVERTRIEB GmbH & Co. KG, Meßberg 1, 20086 Hamburg, Telefon: +49 40 3019-1800, Fax: +49 40 3019-1815, info@dermedienvertrieb.de, Internet: dermedienvertrieb.de

iX Special 2022 – Green IT: Einzelpreis 14,90 €, Österreich 16,40 €, Schweiz 27.90 CHF, Luxemburg 17,10 €

Printed in Germany

ISSN 0935-9680

Die Inserenten*

Redaktioneller Teil

Alkmene Verlags- und Mediengesellschaft mbH	Frankfurt am Main	19
AMD International Sales & Service Ltd.	USA-Sunnyvale	28, 29
B1 Systems GmbH	Vohburg	155
Cordaware GmbH	Pfaffenhofen	2
KfW Bankengruppe	Frankfurt am Main	11
M-net Telekommunikations GmbH	München	21
Techbuyer GmbH	Goch	95
Thomas Krenn AG	Freyung	156
zlb Zentral- u. Landesbibliothek Berlin	Berlin	47

Veranstaltungen

enterJS	iX, heise Developer, dpunkt.verlag	6
iX Workshops	iX, heise Events	15
StackFuel	StackFuel	25
PSW	iX	51
qSkills	qSkills, iX	57
unique code	iX	70
Mastering Kubernetes	heise Academy	75
Maker Faire	Make:	77
storage2day	iX, dpunkt.verlag	83
SQL-Webinar-Serie	heise Academy	91
Software Quality Lab	Software Quality, iX	117

Ein Teil dieser Ausgabe enthält Beilagen der Heise Gruppe GmbH & Co. KG, Hannover.

Wir bitten unsere Leser um freundliche Beachtung.

*Die hier abgedruckten Seitenzahlen sind nicht verbindlich. Redaktionelle Gründe können Änderungen erforderlich machen.

Direkte Warmwasserkühlung von HPC-Chips

Vorangeschritten

Bernd Schöne

Warmes Wasser kühlt Prozessoren besser als kalte Luft – und betreibt Freikühler und Adsorptionskältemaschinen, doch meist nur bei Supercomputern.

Die Wasserkühlung großer Rechner ist nicht neu. Während sich IBMs Mainframes gerade von ihr verabschieden, gehen Supercomputer den umgekehrten Weg. Auch die letzten Hersteller stellen fest, dass eine Luftkühlung die von ihnen erzeugte Hitze nicht mehr bewältigen kann.

Zum Kühlen nutzt man vor allem mechanisch betriebene Kältemaschinen nach der Kompressionsmethode. Sie verdichten ein gasförmiges Kältemittel, das sich dadurch erwärmt. Im Verflüssiger gibt es die Wärme an die Umgebung ab, wodurch das Kältemittel kondensiert. Während der Druckabsenkung in einer geeigneten Vorrichtung verdampft es wieder, entzieht seiner Umgebung Wärme und produziert so Verdunstungskälte.

Nachdem 1995 erst FCKW-haltige Kältemittel und später deren Nachfolger FKW aufgrund ihrer ozonschädigenden Wirkung verboten wurden, unterwirft die EU die heute verwendeten HFKW-Kältemittel einem absichtlich herbeigeführten Verknappungsprozess, der die Betriebssicherheit von Kühlanlagen mit Kompressorkühlung gefährdet. Die EU will den Bestand von HFKW-Kältemittel wie R134a, R407C und R410A aufgrund ihres hohen Treibhausgaspotenzials bis 2030 auf 21 Prozent der Menge von 2015 drücken (siehe Artikel „Grün getarnt“ ab Seite 122).

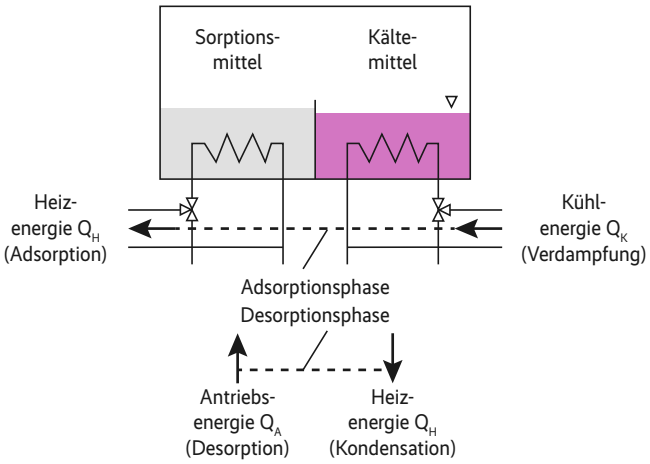
Zu den heutigen Großrechnern, die ohne Kompressionskältemaschinen auskommen, gehört der 3 PFlops schnelle SuperMUC am LRZ (Leibniz-Rechenzentrum) in Garching bei München. Er wird über Freikühler auf dem Dach gekühlt. Die noch benötigte Restkälte erzeugen Adsorptionskältemaschinen. Sie arbeiten ohne umweltschädliche Kältemittel, da sie ausschließlich Wasser und das Prinzip der Sorption nutzen, benannt nach dem lateinischen „sorbere“, „schlürfen“. Adsorption bezeichnet die Anreicherung von Stoffen, etwa Gasen oder Flüssigkeiten, an der Oberfläche eines Festkörpers. In Adsorptionsprozessen wird Wasserdampf vom Sorptionsmaterial wie Silikagel oder Zeolith angesaugt, wodurch Wasser kondensiert, bei der Desorption wieder verdampft und dadurch die Umgebung kühlt (siehe Abbildung 1).

Nicht zu verwechseln sind Adsorptionskältemaschinen mit Absorptionskältemaschinen. Sie werden als geräuschlos arbeitende Alternative angeboten und sind zum Beispiel als Minibar in Hotelzimmern zu finden. In ihnen wird der Kompressor des normalen Kühlschranks durch einen Lösungskreislauf ersetzt. Bei der Absorption dringen Gase in eine Flüssigkeit ein, während bei der Adsorption alle Prozesse an der Oberfläche eines Festkörpers ablaufen.

iX-TRACT

- Kompressionskühlung ist energieaufwendig und mit umweltschädlichen Kältemitteln verbunden.
- Umweltfreundlicher sind Adsorptionskältemaschinen. Sie arbeiten mit weniger Energieaufwand und mit Wasser, das muss aber 60 °C warm sein.
- Warmwasser- und Adsorptionskühlung findet sich vor allem im HPC-Bereich, wo der hohen Energiedichte der High-End-Prozessoren mit Luftkühlung nicht mehr beizukommen ist.

Adsorptionskältemaschinen können Wärmeenergie in Nutzkälte verwandeln, ohne die Umwelt mit schädlichen Kühlmitteln zu belasten. Das funktioniert am LRZ auch deshalb, weil das Kühlwasser des SuperMUC relativ hohe Temperaturen ausweisen darf. Möglich wurde das durch intensive Forschungsarbeit von IBM Schweiz, getrieben vom wachsenden Wunsch nach grünen Superrechnern. Heute entfallen bis zu 50 Prozent des Energieverbrauchs und der CO_2-Emissionen eines durchschnittlichen luftgekühlten Rechenzentrums auf die Kühlsysteme. Schweizer Wissenschaftler griffen daher die Idee der direkten Wasserkühlung aus Mainframe-Tagen auf, denn Wasser leitet Wärme rund 4000-mal besser als Luft. Vor allem Hochleistungsrechner sind bereits jetzt kaum noch anders zu kühlen.



Im Adsorptionsprozess kondensiert Wasserdampf am Sorptionsmaterial und verdampft bei der Desorption wieder (Abb. 1).

Eine Domäne des HPC-Segments

Außerhalb der HCP-Gemeinde investiert allerdings kaum jemand in diese Technik. In vielen Rechenzentren fehlen die entsprechenden Voraussetzungen oder schlicht der Platz, die neue Infrastruktur zu installieren. Die aktuelle Studie „Rechenzentren in Deutschland“ des Bitkom aus dem Jahr 2022 zeigt, dass für die Betreiber die Zuverlässigkeit die höchste Priorität hat, der Energieverbrauch aber ein wichtiges Thema ist. Denn der jährliche Energiebedarf der Rechenzentren und kleineren IT-Installationen ist in Deutschland laut Bitkom von 2010 bis 2020 von 10,5 auf 16 TWh gestiegen (siehe Abbildung 2).

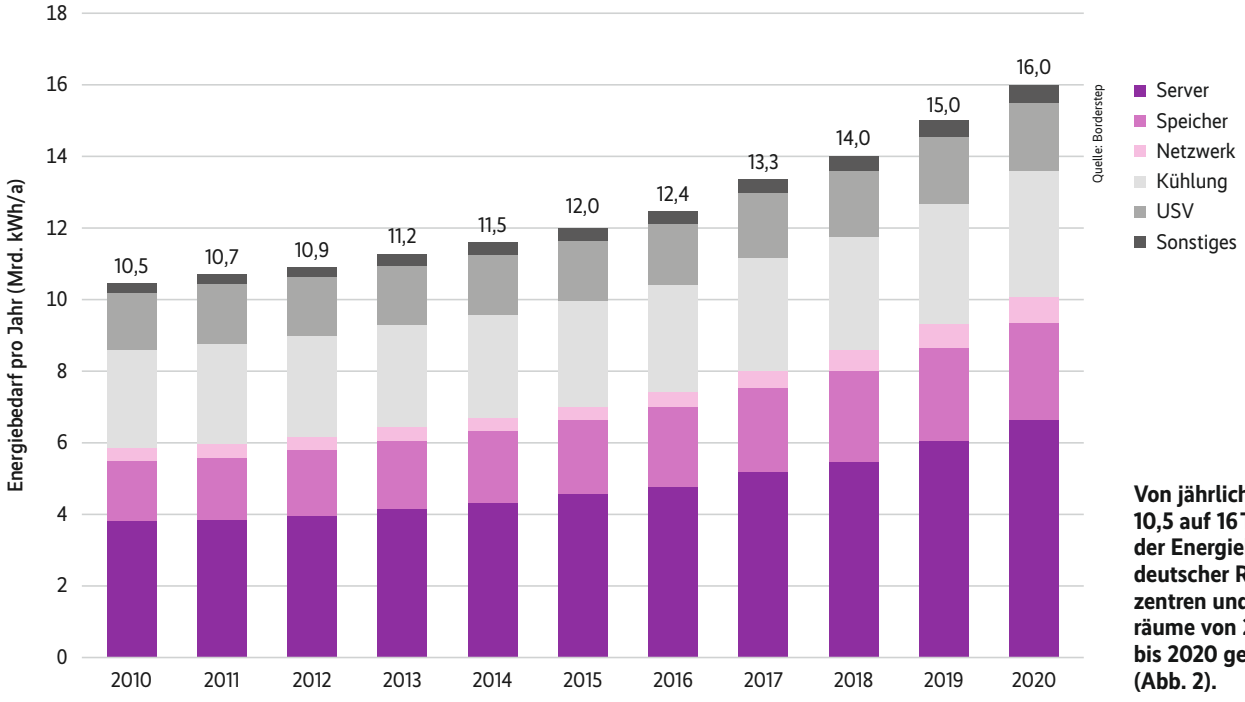
Die frei werdende Energie wieder zu nutzen, etwa um sie an das Fernwärmenetz abzugeben, erscheint der Mehrheit aber nicht möglich, weil Abnehmer und Regularien fehlen. Viele halten eine Folgenutzung der teuer erzeugten Wärme erst nach Umstellung auf eine Warmwasserkühlung für sinnvoll. Dazu wird im Rechenzentrum ein geschlossener Wasserkreislauf mit Wärmetauschern installiert.

„Über Warmwasserkühlung wird viel geredet, aber sie wird nur im HPC-Bereich wirklich eingesetzt“, zu diesem Schluss kommt Ralph Hintemann, Autor der Bitkom-Studie. „Für die Betreiber von kommerziellen Rechenzentren bedeutet diese Kühlmethode einen Mehraufwand, der vom Kunden bisher nicht honoriert wird.“ Generell ist die Wissensbasis in diesem Themengebiet schlecht. „Es gibt kaum Studien über Rechenzentren, und die beteiligten Firmen halten sich aus wettbewerbsstrategischen Gründen mit Informationen zurück“, so Hintemann. „Technische Details, Preise und Vertragskonditionen regelt die Industrie lieber intern.“

Über die absolute Grenze der thermischen Leidensfähigkeit der Chips ist wenig bekannt (siehe auch Kasten „Die Angst vor der Wärme“). Die Industrie verweist auf die Datenblätter, die die maximal garantierte Gehäusetemperatur angeben. „Moderne Server kommen mit Wassertemperaturen von 60 °C und mehr durchaus klar“, so Ralph Hintemann, „spezielle Chips können sogar bei höhere Temperaturen ohne Störungen betrieben werden.“ Temperaturen oberhalb von 70 °C gelten aber als kritisch, Ausfälle und Fehlkalkulationen können die Folge sein.

Wasser bis zum Chipgehäuse

Das LRZ der Bayerischen Akademie der Wissenschaften gehörte zu den ersten Rechenzentren, die ihren Supercomputer mit Warmwasserkühlung ausstatteten, und wurde dafür 2012 mit



Von jährlich 10,5 auf 16 TWh ist der Energiebedarf deutscher Rechenzentren und Serverräume von 2010 bis 2020 gestiegen (Abb. 2).

dem Deutschen Rechenzentrumspreis in der Kategorie „Energie- und Ressourceneffiziente Rechenzentren“ ausgezeichnet. Die Technik dafür entwickelte IBM Research Zürich zusammen mit der ETH Zürich, der ETH Lausanne und dem Schweizer Kompetenzzentrum für Energie und Mobilität (CCEM). Sie senkten die Stromrechnung des LRZ jährlich um rund eine Million Euro, weil sie den Gesamtenergieverbrauch des Rechners um 40 Prozent reduzierten.

In dem mit 150 000 Rechenkernen ausgestatteten SuperMUC wird über 40 °C heißes Wasser direkt auf die Prozessorgehäuse aufgebracht. Dadurch leitet diese Direct Liquid Hot Water Cooling oder High Temperature Direct Liquid Cooling (HT-DLC) genannte Technik die Wärme direkt von der Wärmequelle ab – bis zu 3 kW/cm³ (siehe Abbildung 5). Bei den Betriebstemperaturen tastete sich das LRZ vorsichtig nach oben. Beim SuperMUC hat allein der Wegfall der CPU-Lüfter zu einer Energieersparnis von 5 Prozent geführt, weitere 4 Prozent gehen auf die geringere Betriebstemperatur der Prozessoren durch bessere Kühlung zurück. Die freie Kühlung des LRZ spart weitere 15 Prozent ein.

Die Angst vor der Wärme

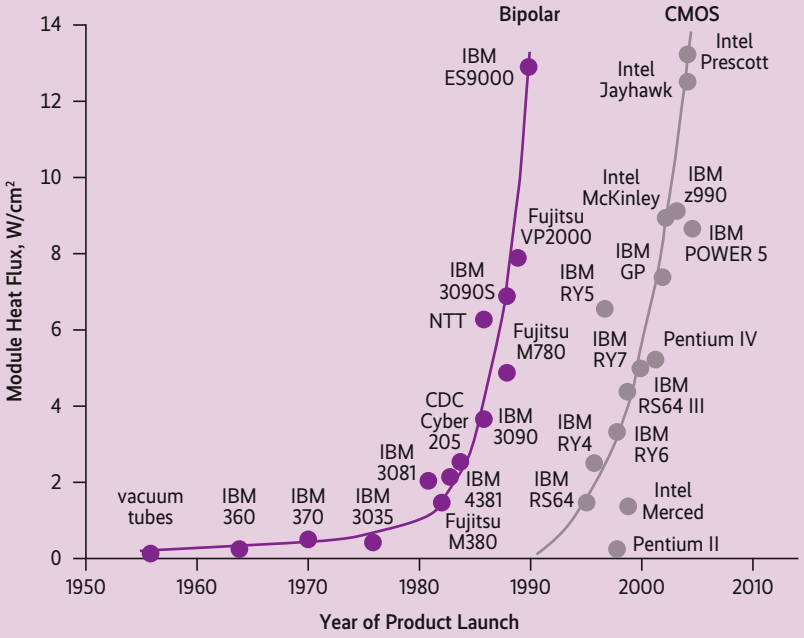
Wer in den 70er-Jahren ein Rechenzentrum besuchte, musste sich warm anziehen, selbst im Hochsommer. Temperaturen unter 18 °C waren keine Seltenheit. Niemand wollte den Wärmetod der sündhaft teuren Mainframes riskieren. Da der Computer teuer und Energie billig war, störte sich niemand an den horrenden Energiemengen, die das kostete. Üppig und reichlich strömte kühle Luft durch die Doppelböden der großzügig zugeschnittenen Rechenzentren.

Rechenzentren waren Teil der Selbstdarstellung, oft verglast und dadurch dem Sonnenlicht ausgesetzt. Die Verehrung für „unseren“ Computer war in dieser Zeit auch viel zu groß. Sparkassenvorstände und Stadtväter waren stolz darauf, eine IBM360 zu besitzen, auch wenn sie nur gepachtet war. Zu den Mietbedingungen zählten ausreichend gekühlte Räume. Die bipolaren Chips im Innern vertrugen keine große Hitze, erzeugten aber viel davon (siehe Abbildung 3).

Als Vorgabe verwendeten viele RZ-Betreiber die Anregungen des Herstellers IBM und des 1894 gegründeten US-Branchenverbandes ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers). Sie empfahlen Lufttemperaturen von 15 bis 32 °C, was in der Praxis auf 18 °C hinauslief, mit einer Reserve nach unten. 18 °C galt als idealer Wert, um Schäden von den empfindlichen bipolaren Chips fernzuhalten. Diese wurden in der zentralen Recheneinheit direkt mit kaltem Wasser temperiert.

Eine beeindruckende Grafik von ASHRAE verdeutlichte den Zusammenhang zwischen Ausfallrate und Betriebstemperatur. Alles über 20 °C Umgebungstemperatur galt für die kostbaren Systeme als zu gefährlich. Um die Luft auf die idealen 18 °C zu kühlen, mussten Kältemaschinen Wasser auf Temperaturen von unter 14 °C temperieren – vor allem im Sommer ein enorm aufwendiger Prozess, der viel Strom kostete. Denn aus den Gesetzen der Thermodynamik folgt: Der Strombedarf wächst proportional zur Temperaturdifferenz. Je kälter das Wasser, desto teurer. Nicht berücksichtigt wurde damals, dass Computer oft nach fünf bis sechs Jahren ausgetauscht wurden, also lange bevor die erhöhten Ausfallraten wirklich ins Gewicht fielen.

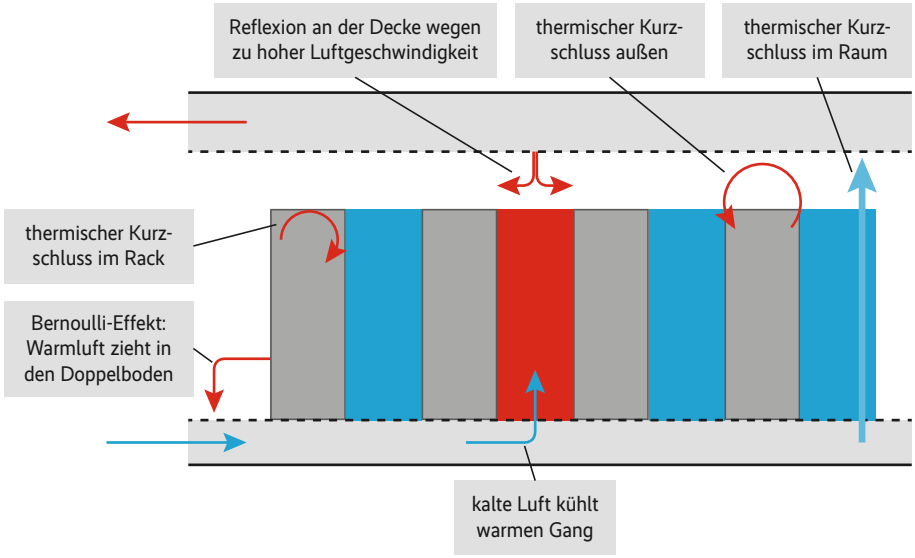
Über die Wohlfühltemperatur von Chips ist viel geforscht worden. Denn inzwischen ist eine neue Chiptechnik in die Superrechner eingezogen, die sich anstandslos bei Betriebstemperaturen von 45 °C und mehr betreiben lässt. Im Jahr 2011 legte die ASHRAE eine Studie mit dem Namen „Thermal Guidelines for Data Processing Environments – Expanded Data Center Classes and Usage Guidance“ vor. Anhand umfangreicher Analysen von Herstellerempfehlungen sieht man 27 °C als tolerabel an. Fast alle Hersteller geben neue Komponenten sogar für höhere Temperaturen frei, wodurch theoretisch auch heute schon 38 bis 40 °C möglich sein sollten. Höhere Temperaturen bedingen allerdings auch größere Leitungsverluste und mehr Rechenfehler und Serverausfälle.



Bipolare Chips durchliefen den gleichen Anstieg der Wärmedichte wie nach ihnen die CMOS-Chips (Abb. 3).

Dennoch wurden Rechenzentren ohne technische Notwendigkeit über Jahre weiter auf arktische Temperaturen gekühlt, weil Verträge zwischen Rechenzentrumsbetreibern und Nutzern oder Leasingverträge dies vorschrieben, auch unter Hinweis auf technische Vorgaben und Empfehlungen wie die des US-Verbandes ASHRAE. Etliche Nutzer hegten auch die nicht ganz unbegründete Furcht vor schlecht belüfteten Stellen in ihren Rechenzentren. Wer mit Luft kühlt, sollte darauf achten, dass nicht Kabelbäume, Schmutz oder thermische Kurzschlüsse den freien Strom der kühlen Luft behindern (siehe Abbildung 4).

Auch birgt der seit Mainframe-Tagen im Rechenzentrum so beliebte Doppelboden Risiken, wenn dort im Laufe der Zeit immer mehr Datenleitungen oder andere Installationen eingezogen werden, die dann den Luftstrom behindern. Vorsichtige Administratoren inspizierten nach Umbauten ihr Rechenzentrum mit einer Infrarotbrille, um unzulässig erwärmte Apparate ausfindig zu machen – und das nicht ohne Grund. Denn: Staut sich die Wärme, droht ein Schwelbrand oder zumindest die Zwangsabschaltung der Rechner.



So schematisch wie in der Theorie funktioniert der Luftstrom im Rechenzentrum nicht. Vor allem thermische Kurzschlüsse lauern an vielen Stellen (Abb. 4).

Warmes Wasser kühlt wesentlich besser als kalte Luft, die ein guter thermischer Isolator, also ein extrem schlechter Wärmeleiter ist. Auch ist ihr Vermögen, Wärme aufzunehmen, ausgesprochen schlecht. Wasser besitzt eine viermal höhere spezifische Wärmekapazität und eine zehnmal höhere Wärmeleitfähigkeit als Luft. Selbst bei Vorlauftemperaturen von 50 °C kühlt Wasser die CPU-Oberfläche zuverlässig unter die erlaubten Grenztemperaturen. Sie bleiben kühler als bei Luftkühlung mit auf 20 °C temperierter Luft, erklärt Michael Ott, Senior Reseacher und Green-IT-Spezialist der Future-Computing-Gruppe am LRZ, der nicht mit höheren Ausfallraten rechnet. „Nach zehn Jahren Betrieb solcher Maschinen stellen wir tatsächlich fest, dass die Ausfallraten niedriger sind als bei luftgekühlten Maschinen“, so Ott.

Moderat temperiert und wiederverwendet

Neben dem Leibniz-Rechenzentrum gehört das Rechenzentrum des KIT (Karlsruher Institut für Technologie) zu den Vorreitern des HT-DLC, allerdings bei moderateren Temperaturen, mit freier Kühlung und Abwärmenutzung. „Mit der Warmwasserkühlung erhöhen wir die Energieeffizienz des Gesamtsystems“, erläutert Simon Raffeiner, Betriebsleiter HPC am Nationalen Hochleistungsrechenzentrum in Karlsruhe. „Wir versorgen den Rechner mit 40 °C warmem Wasser. Nachdem die Komponenten im System gekühlt wurden, liegt die Wassertemperatur bei mehr als 45 °C. Da die Außentemperatur in Karlsruhe selten mehr als 40 °C beträgt, können wir das warme Wasser vereinfacht gesagt direkt zu den Rückkühlern auf dem Dach pumpen, es kühlt sich dort auf 40 °C ab und kann ohne Umwege wieder zurück in den Rechner geleitet werden.“

„Statt großer Kältemaschinen, deren Stromverbrauch bis zu 50 Prozent der zu kühlenden Leistung ausmachen kann, müssen wir nur einige kleine Pumpen betreiben. Auch benötigen wir im Gegensatz zu vielen anderen Rechenzentren kein zusätzliches Wasser für eine Verdunstungskühlung. Darüber hinaus eignet sich so warmes Wasser für die Heizung des benachbarten Bürogebäudes. Das haben wir auch so umgesetzt. Eine Ausweitung der Wärmenachnutzung auf weitere Gebäude, möglicherweise auch eine Einspeisung in das Fernwärmenetz des Campus, ist angedacht. Fernwärme gestaltet sich allerdings aufgrund des Temperaturniveaus und der stark schwankenden Abnahmemengen im Jahresverlauf als herausfordernd.“

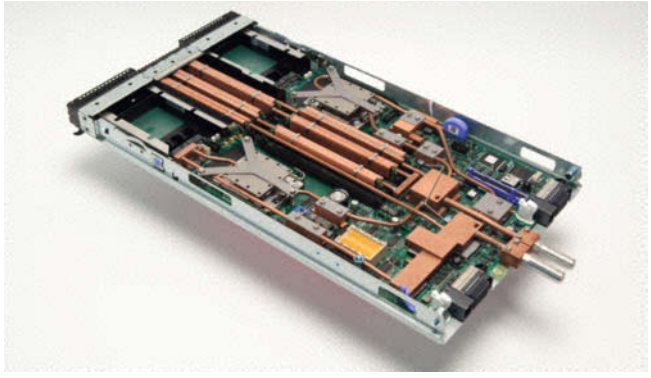
Auch das Forschungszentrum Jülich nutzt die Wärme des dortigen Superrechners vom Typ Bull Sequana X1000. Der JUWELS genannte Computer soll im Herbst 2022 an die Gebäudeheizung gekoppelt werden. Eine Wärmepumpe erhöht die Temperatur des 35 °C warmen Kühlwassers des von der französischen IT-Firma Atos konstruierten modularen Rechners auf 70 °C. Seit 2021 installiert die Universität Paderborn ein ähnliches System. Jülich betreibt mit dem Fujitsu-Rechner QPACE3 seit 2016 einen zweiten warmwassergekühlten Rechnercluster. Die Vorlauftemperatur von bis zu 40 °C ermöglicht ganzjährig eine freie Kühlung. Bei niedrigen Außentemperaturen wird eine niedrigere Vorlauftemperatur eingestellt, da sich das positiv auf den Strombedarf des Rechners auswirkt.

Neben dem SuperMUC betrieb das LRZ einen kleinen Cluster mit AMD-Prozessoren, der ebenfalls mit warmem Wasser gekühlt wurde. Die Vorlauftemperaturen lagen bei 25 bis 45 °C; der Rücklauf wies bereits Temperaturen von 55 bis 60 °C auf. Diese hohen Temperaturen ermöglichten es, die Wärme einer Adsorptionskältemaschine zuzuführen. Diese gewann daraus 11 kW als Kälteleistung für konventionell gekühlte IT-Systeme.

Eine Version ohne Wärmepumpen

2016 ging in Garching der Rechencluster CooLMUC-2 in Betrieb, gefertigt diesmal von Lenovo. IBM Research Zürich und Fahrenheit lieferten die Kühlung. Dank der Wassertemperatur von 60 °C kommt CooLMUC-2 ohne Wärmepumpe aus, er produziert die von den Adsorptionskältemaschinen verlangte Wassertemperatur direkt. Denn diese Maschinen benötigen im Vergleich zu üblichen Kompressionskältemaschinen zwar nur halb so viel Energie, erfordern aber 60 °C warmes Wasser. Die erzeugte Kälte dient der Kühlung der luftgekühlten Systeme im LRZ. Inzwischen wurde CooLMUC-3 installiert, das erste Sys-

Quelle: IBM



Das durch die Kupferleitungen geführte Wasser verdampft im Arteriensystem auf dem Chipgehäuse und kondensiert beim Rücktransport (Abb. 5).

tem, bei dem nicht nur die CPUs, sondern alle Komponenten mit warmem Wasser gekühlt werden, also auch die Stromversorgung und die Netzwerkkomponenten – ein wichtiges Detail, das oft unterschlagen wird.

Auch der SuperMUC Next Generation (SuperMUC-NG) wird über Warmwasserkühlung gekühlt. „Eine Einspeisung in das Garchinger Fernwärmenetz ist nicht möglich", erläutert Michael Ott, „die Stromrechnung für die nötigen Wärmepumpen wäre zu hoch. Das von den Rechnern zurückströmende Wasser hat Temperaturen von etwa 55 Grad, und das ist mindestens 5 Grad zu wenig für ein Fernwärmenetz." Um die Wärme trotzdem sinnvoll zu nutzen, wird sie zur Hälfte in Adsorptionskältemaschinen in Kälte verwandelt, der Rest landet wie später die andere Abwärme auf dem Dach des LRZ-Gebäudes, wo die Rückkühler sie an die Umgebung abgeben. Insgesamt gibt der SuperMUC-NG 3400 kW thermische Energie an die Umgebung ab.

„Wir haben mit der Warmwasserkühlung seit der Einführung dieser Technik im Jahr 2011 gute Erfahrungen gemacht", sagt Ott, „ein Drittel der Betriebskosten vom SuperMUC-NG sind Stromkosten, und die konnten wir so reduzieren. Den oft beschworenen Wasserschaden durch plötzlich undicht werdende Verschraubungen hat es bislang nicht gegeben, aber das LRZ konnte Jahr für Jahr ein Drittel der Stromkosten einsparen." Ein Umstieg auf Luftkühlung kommt nicht infrage, schon allein aufgrund der weiter steigenden Energiedichte innerhalb der Prozessoren.

Der Trend zur steigenden Wärmedichte bei Chips

Der nächste SuperMUC ist bereits geplant, diesmal mit Intels Ponte Vecchio. Die HPC-GPU gibt es für Luft- und Wasserkühlung, doch nur letztere Variante liefert die volle Rechenleistung. Die maximale Oberflächentemperatur gibt Intel mit 81 °C an, die Anschlussleistung der wassergekühlten mit 600 Watt, die luftgekühlte Variante ist nur bis 450 Watt belastbar. Teilt man die Anschlussleistung durch die Oberfläche des Chips, ergibt sich eine Wärmedichte, die rund zehnmal höher ist als die einer Herdplatte.

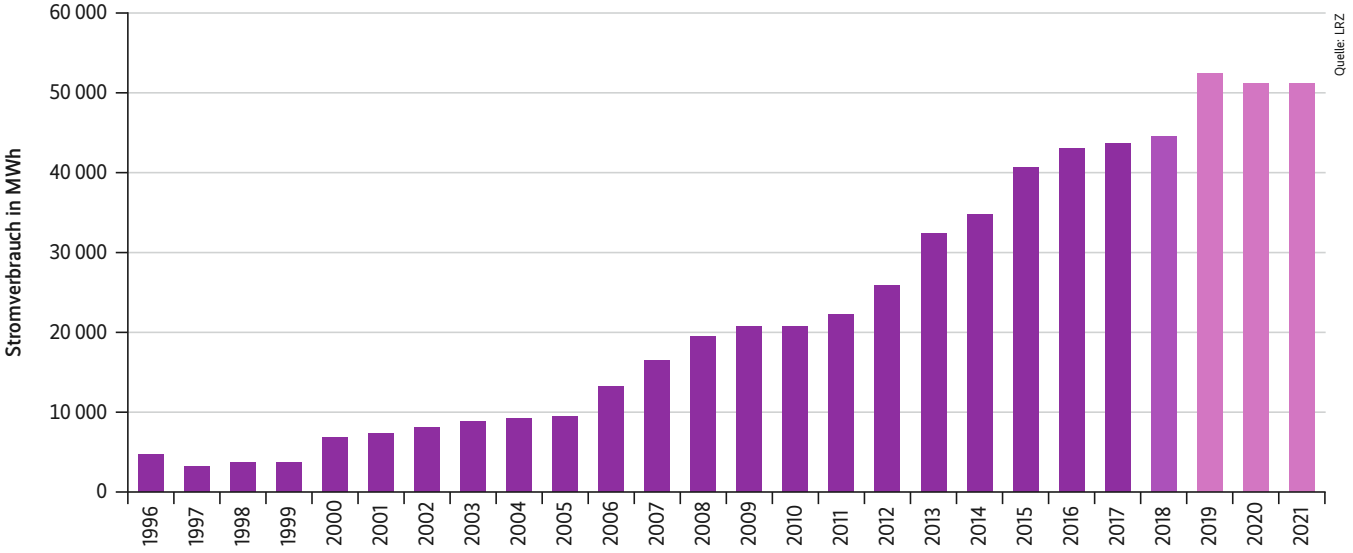
Dieser Trend wird sich fortsetzen, denn die Zahl der Rechenkerne pro Chip steigt weiter. Die nächste Generation der High-End-CPUs und -GPUs wird einen Strombedarf von 1000 Watt aufweisen. „Der Warmwasserkühlung gehört bei HPC-Rechenzentren die Zukunft, die Leistungsdichten an den Prozessoren steigen weiter. Ohne Wasserkühlung sind diese Systeme nicht zu betreiben, denn HPC-Rechner laufen fast rund um die Uhr unter Volllast", so Michael Ott.

Auch die neuen LRZ-Rechner erhalten Adsorptionskältemaschinen. Kälte ist ein gefragtes Gut im Rechenzentrum, denn längst nicht alle Komponenten in einem Rechenzentrum lassen sich direkt an ein wassergespeistes Kühlsystem koppeln, etwa Switches und Speichersysteme – vor allem bei Festplatten steigt die Ausfallwahrscheinlichkeit bei hohen Betriebstemperaturen. Festplatten lassen sich alternativ nur mit einer zweiphasigen Immersionskühlung betreiben, die aber aufwendig zu implementieren ist (siehe Artikel „Aufgefangen" ab Seite 146).

Für das derzeitige Fernwärmenetz der dritten Generation wären Temperaturen von 80 °C optimal. Das überschreitet aber die maximal zulässige Oberflächentemperatur der HPC-Prozessoren, bei etwas niedrigeren Temperaturen steigt bereits die Wahrscheinlichkeit, dass durch thermische Prozesse Berechnungen zum falschen Ergebnis führen. „Insbesondere bei den CPUs ist die Zahl der Modelle, für die die Hersteller so hohe Temperaturen zulassen, im HPC-Bereich nur noch sehr gering. Bei SSDs lässt die Leistung jenseits der 50 °C dann auch sehr schnell nach", erläutert Simon Raffeiner vom KIT. Bei 85 °C Oberflächentemperatur ist die Belastungsgrenze fast aller Chips erreicht.

Chips werden künftig wieder wärmeempfindlicher

Die Prozessoren der nächsten Generation werden noch höher integriert sein als die jetzigen und dadurch vermutlich empfindlicher auf hohe Temperaturen reagieren. Haben sich die Ingenieure in den letzten Jahren in höhere Temperaturbereiche vorgetastet, dürfte es in den nächsten Jahren wieder nach unten gehen. Beispielsweise soll 2023 in den USA der Superrechner Frontier mit einer Rechenleistung von 1,5 EFlops in Betrieb gehen. Herzstück ist eine spezielle Version der CPU Epyc 7A53 und der GPU Instinct MI250X mit 560 Watt pro Chip von AMD. Höher als 32 °C soll die Wassertemperatur dort nicht sein, was sich weder für die Gebäudeklimatisierung noch für die Adsorptionskühlung eignet.



Der Energiebedarf des vorbildlichen Leibniz-Rechenzentrums hat sich zwischen 2005 und 2020 mehr als verfünffacht (Abb. 6).

„Im HPC-Bereich hat die Adsorptionskühlung eigentlich aufgrund der durch die höheren Leistungsdichten sinkenden Temperaturen keine Zukunft“, erläutert Michael Ott, „zudem erhöht der Einsatz von Adsorptionskältemaschinen die Systemkomplexität deutlich. Wir hatten am Anfang auch viel mit Kinderkrankheiten zu kämpfen.“ Der nächste SuperMUC in Garching wird kaum noch Kaltwasser benötigen. Fast die gesamte erzeugte Wärme wird dann über die Rückkühler auf dem Dach an die Umgebungsluft abgegeben. „Für unseren nächsten Hochleistungsrechner streben wir mehr als 98 Prozent Kühlung durch Warmwasserkühlung an. Dann ist der Bedarf an Kaltwasser letztendlich so gering, dass es vollkommen egal ist, ob das mit 5 oder 30 Prozent Energieaufschlag erzeugt wird“, so Michael Ott.

Während die Warmwasserkühlung bei HPC also schon fast Standard ist, warten andere Rechenzentren noch ab. „Sobald große Cloud-Anbieter verstärkt Warmwasserkühlung verwenden, dürften auch die Mittelständler und die Betreiber der kleineren Rechenzentren nachziehen“, hofft Ralph Hintemann, „dann wird auch die Modellpalette vielfältiger.“

Die Kühlung ist allerdings nicht der einzige große Energieverbraucher. Da sind zum einen die Recheneinheiten selbst, deren Energieeffizienz in GFlops/W gemessen wird, und zum anderen die Herstellungsprozesse des Supercomputers. Da die Stromkosten etwa den Hardwarekosten entsprechen, steht nach etwa fünf bis sieben Jahren ein neuer Rechner an, der bei vergleichbaren Stromkosten viel mehr leistet. Der 2020 eingeweihte und ebenfalls warmwassergekühlte Stuttgarter Supercomputer Hawk beispielsweise hat 44 Millionen Euro gekostet und zieht jedes Jahr Strom im Wert von etwa 4 Millionen Euro. Eine Zweitnutzung ist bei Supercomputern nicht vorgesehen und wird es auch bei Hawk nicht geben.

Fazit

Während man in Jülich und Garching gleich mehrere warmwassergekühlte Rechner nutzt und über die Jahre durchweg positive Erfahrungen gemacht hat, sieht das bei kommerziellen Rechenzentren ganz anders aus. Die höheren Investitionskosten rechnen sich – noch – nicht. Die Betreiber scheuen die technisch möglichen hohen Temperaturen und verschenken damit die Chance, die Abwärme sinnvoll zu nutzen. Den Grund erfährt man hinter vorgehaltener Hand: die Gewährleistung. Niemand möchte im Fall eines möglichen Schadens als Verursacher gelten. Die gerichtlichen Gutachter könnten die bekannten Tabellen der ASHRAE zücken und auf die dort empfohlene optimale Kühltemperatur verweisen. Oft sind diese Zahlen bereits in den entsprechenden Verträgen fixiert. Wenn dann noch individuelle Sicherheitsmargen beschlossen werden, rattern die Kompressionskühler fast so ungebremst wie in den 60er-Jahren. Die Stromkosten zahlt der Kunde des Rechenzentrums – und über Umlagen indirekt wir alle.

Im Juli 2021 beschloss die EU, den CO_2-Ausstoß von Rechenzentren bis 2030 um mindestens 55 Prozent unter den Wert von 1990 zu drücken, wie immer das auch gelingen mag angesichts der explodierenden Anzahl der RZs und ihres nicht minder schnell fortschreitenden Stromhungers. Selbst das vorbildliche LRZ in Garching benötigt jedes Jahr mehr Strom. Sein Energiebedarf stieg von 10 GWh im Jahr 2005 trotz durchdachter Kühltechnik bis 2019 auf über 40 GWh (siehe Abbildung 6). Das entspricht dem Bedarf einer Stadt mit 30 000 Einwohnern – Tendenz steigend.

Zum Massenphänomen wird die Warmwasserkühlung so nicht. Firmen, die sich hier engagieren, kämpfen um jeden Kunden. Nachträglich die Voraussetzungen für eine Warmwasserkühlung zu schaffen, ist höchst unwirtschaftlich. Innovative Wege werden nur dort beschritten, wo es keine Alternative gibt, weil die enormen Wärmemengen sonst nicht vom Chip wegzubekommen sind. Das könnte sich rächen, denn bei einem möglichen Stromengpass zählt jede Kilowattstunde, verbunden mit der Frage: Ist dieses Rechenzentrum systemrelevant? (sun@ix.de)

Bernd Schöne

ist freier Journalist.

Systemnahe Wärmeübertragung für luftgekühlte Server

Nah dran

Susanne Nolte, Hubert Sieverding

Auch in Colocations und überall, wo die Luftkühlung Usus ist, lohnt es sich, die Klimatisierung möglichst nah an die Server zu bringen – vor allem, wenn man die Abwärme nutzen will.

X-TRACT

- Für den optimalen Abwärmetransport zwischen luftgekühlten Servern und der Kühlanlage bringt man den flüssigkeitsgeführten Kühlkreis möglichst nah an die Systeme heran.
- Effizienter als eine Kaltgangeinhausung arbeiten Warmgangeinhausungen und Reihenkühlungen, da sie das Warmluftvolumen deutlich reduzieren. Beide setzen aber einen abgeschotteten Warmgang voraus.
- Andere Techniken wie Seiten- und Rückwandkühler reduzieren das Warmluftvolumen auf den hinteren Bereich des Racks.
- Bei allen Systemen erübrigt sich eine Raumkühlung. Einige Systeme erlauben sogar das Weiterverwenden vorhandener Racks. Anpassungen sind aber in jedem Fall nötig.

Rechenzentren, die die Kosten ihrer Klimatisierung senken wollen, nutzen oft eine Kaltgangeinhausung. Sie ist eine Weiterentwicklung der Rechnerräume der ersten Generation: Dort wurde über einen Doppelboden eiskalte Luft eingeblasen. Rechner und Laufwerke standen frei im Raum und das Personal trug ganzjährig Wollpullover.

Seit die Energiedichte zunimmt und sich 19″-Racks durchgesetzt haben, gilt dieses Prinzip als ineffizient. Besser ist es, die Vermischung von Warm- und Kaltluft zu verhindern. Deshalb stehen die Racks in Reihen, Vorderseite zu Vorderseite mit einem Kaltgang dazwischen und Rückseite zu Rückseite mit einem Warmgang dazwischen. Die Systeme saugen die Kaltluft auf der Vorderseite ein und geben die Warmluft auf der Rückseite ab.

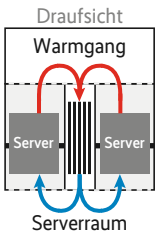
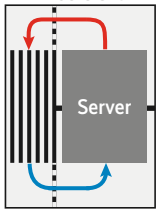
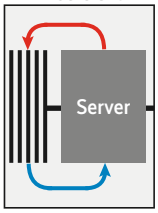
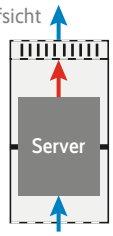
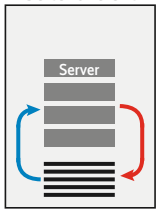
Kaltgangeinhausung

Durch das Einhausen schottet man den Kaltgang ab, sowohl nach oben als auch an den Enden. In ihn wird durch den Doppelboden Kaltluft eingeblasen. Die von den Systemen erwärmte Luft steigt aus deren Rückseite in den Serverraum außerhalb der Einhausung nach oben und sammelt sich unter der Decke. Von dort aus wird sie abgesaugt, gekühlt und anschließend über einen Doppelboden wieder den Rechnern zugeführt. Das Konzept der Kaltgangeinhausung verlangt einen Doppelboden und eine ausreichende Raumhöhe. Diesen baulichen Nachteilen steht gegenüber, dass sich alle Schränke komplett für IT-Equipment nutzen lassen und die Redundanz der Klimatisierung einfach zu implementieren ist.

Doch hält sich die Effizienz dieses Prinzips in Grenzen: Erstens erwärmt sich die in den Doppelboden eingeblasene Kaltluft bereits auf dem Weg zu den Servern um 2 bis 3 °C. Zweitens ist der Weg der Warmluft bis zum RZ-eigenen Kühlsystem lang und nutzt ein äußerst ungeeignetes Medium, die Luft. Drittens handelt es sich bei dieser Luft um die Raumluft der Serverräume. Da darin zumindest ab und zu Menschen arbeiten, ist der Temperaturbereich dieses Mediums auf 18 bis 25 °C begrenzt, was seine Eignung weiter herabsetzt. Vor allem aber erfordern derart geringe Temperaturen eine Kältemaschine und Chemikalien mit niedrigem Siedepunkt (siehe Artikel „Grün getarnt“ ab Seite 122).

Sinnvoller wäre es, die Wärme nah an der Quelle zu entnehmen, also den flüssigkeitsführenden Kühlkreislauf so nah wie möglich an die Server heranzubringen und dabei die Raumluft

Funktionsprinzipien der Kühltechniken

Kühltechnik	Reihenkühler	Seitenkühler	Schrankkühler	Rückwandkühler	Rackeinschubkühler
Funktionsprinzip					
typische Kühlleistung	bis 70 kW	bis 80 kW	bis 3 kW	bis 50 kW	bis 7 kW
Positionierung	zwischen den Racks	neben das Rack oder zwischen zwei Racks	im abweichend breiten Rack seitlich neben den Servern	Rückseite im Rack	in den unteren Höheneinheiten
Weg der Warmluft	vom Rack über den Warmgang zum Kühler	von der Serverrückseite im Rack seitlich zum Kühler	von der Serverrückseite im Rack seitlich zum Kühler	von der Serverrückseite direkt in die Kühlertür geblasen oder durch Ventilatoren gezogen	von der Serverrückseite im Rack nach unten zum Kühler
Weg der Kaltluft	vom Kühler durch den Serverraum zu den Racks	vom Kühler seitlich ins Rack zur Vorderseite der Server	vom Kühler seitlich ins Rack zur Vorderseite der Server	normale Raumluft	vom Kühler im Rack nach oben zur Vorderseite der Server
Anforderungen an Racks	Racks müssen vorn und hinten offen sein (evtl. Gittertüren)	betreffende Seitenteile werden ersetzt, vorn und hinten geschlossen	nur mit geeigneten Racks umsetzbar, vorn und hinten geschlossen	Rücktür wird durch Kühltür ersetzt, vorn offen	Standardracks, vorn und hinten geschlossen
Luft-Wasser-Wärmetauscher	außerhalb des Racks	außerhalb des Racks	im Rack	im Rack, Wasserzufuhr außerhalb	im Rack
Merkmale	zusätzlicher seitlicher Platzbedarf	zusätzlicher seitlicher Platzbedarf	zusätzlicher Platzbedarf notwendig	Tiefe des Racks vergrößert sich	Platz für Server verringert sich
Skalierbarkeit	frei skalierbar, mehrere Racks und mehrere Kühler teilen sich einen Warmgang	2 Racks pro Kühler, 1 Rack pro Kühler oder 2 Kühler pro Rack	1 Kühler pro Rack	1 Kühler pro Rack	1 Wärmetauscher pro Rack
Redundanz	frei dimensionierbar, mehrere Racks und mehrere Kühler teilen teilen sich einen Warmgang	2 Kühler pro Rack, je nach Hersteller mehrere Kühlmodule pro Kühlschrank, doppelte Wasserführung möglich	–	je nach Hersteller möglich	–

außen vor zu lassen. Das hat zwei große Vorteile: Zum einen reduziert man das involvierte Warmluftvolumen und damit den Energieaufwand drastisch, den man zum Abführen der Wärme aufbringen muss, da sich etwa Wasser wesentlich besser dafür eignet (siehe Artikel „Ohne Umwege“ ab Seite 140). Zum anderen erlaubt der flüssigkeitsführende Kühlkreislauf wesentlich höhere Temperaturen, je nachdem, wie weit man das Warmluftvolumen zwischen ihm und den Servern verringern kann.

Kürzere Wege zum Wasser

Möglich wäre es beispielsweise, den Warmgang vom Serverraum zu isolieren und direkt mit dem Kühlkreislauf zu koppeln. Alternativ kann man den Kreislauf durch die Racks oder die Rackreihen führen. Alle flüssigkeitsführenden Komponenten sind üblicherweise in Kondensatwannen untergebracht. In allen Varianten kann man auf hohe Räume verzichten und spart den Doppelboden ein. Ist ein Doppelboden im Gebäude integriert, lässt er sich weiter für die Verkabelung nutzen, ohne dass sie mit der Kaltluftzuführung kollidiert. Die Wasserrohre lassen sich von unten oder oben an die Kühler heranführen. Am elegantesten ist es aber, die Rohre in Wannen unterwärts durchs Rechenzentrum und die Kabel in offenen Kabelwannen oberhalb der Racks zu führen.

Da die Server die kalte Luft vorn einziehen, sind die Wärmetauscher oder Wärmeübertrager unten, hinten, an der Seite oder oben angebracht. Topsysteme finden sich aufgrund ihrer beschränkten Leistung eher in Netzwerk- und Telko- als in Serverracks und bleiben unberücksichtigt. Die Tabelle „Funktionsprinzipien“ stellt die unterschiedlichen Methoden zur systemnahen Wärmeübertragung luftgekühlter Server gegenüber. Das Prinzip des Rackeinschubkühlers ist von der Seite, die anderen sind von oben dargestellt.

Warmgangeinhausung

Der Warmgang zwischen den Rückseiten zweier Rackreihen ist vom Rest des Serverraums abgeschottet. In ihm sammelt sich die Abwärme der Systeme. Zwischen den Serverracks aufgestellte Reihenkühler ziehen die Wärme daraus ab und übertragen sie auf Wasser oder eine Kühlflüssigkeit. Auf ihrer Rückseite geben sie die von der Abwärme befreite Luft in den Raum und damit an die Frontseite der Racks ab. Die Kühlung funktioniert nach dem Split-Prinzip, das heißt, das erwärmte Wasser wird durch Rohre aus dem Raum geführt und dort heruntergekühlt.

Warmgangeinhausungen benötigen Platz für die Reihenkühler, was die Packungsdichte einer Rackreihe verringert. Diesen Nachteil kann man durch höhere Racks teilweise ausgleichen, andererseits ist heute eher die Wärmedichte des IT-Equipments als der Platzbedarf der begrenzende Faktor. Da eine Warmgangeinhausung üblicherweise mit mehreren Reihenkühlern arbeitet, ist die Redundanz automatisch gegeben und eine Auslegung n + 1 oder n + x jederzeit möglich. Regeln lässt sich das System über Schwellenwerte, teilweise auch zonenabhängig.

Wie bei allen Trennungen von Warm- und Kaltluftbereichen ist sicherzustellen, dass leere Servereinschübe und die Seiten der Racks mit Blenden abgedichtet werden. Üblicherweise kom-

men die Racks ohne Front- und Rücktüren aus. In Colocations mit erhöhten Sicherheitsanforderungen dürften Warmgangeinhausungen für die meisten Cages oder Einzelräume überdimensioniert sein. Allerdings lässt sich der Warmgang statt doppelreihig auch einreihig als Reihenkühlung aufbauen, wie das Beispiel WOBCOM zeigt (siehe Kasten). In Mehrkundenräumen muss das Sicherheitskonzept mit den Eigenheiten der Warmgangeinhausung zusammenpassen.

Reihenkühler

Ebenfalls zwischen den Racks in die Reihe integriert sind Reihenkühler, auch Rack Cooler genannt. Mit ihnen lassen sich ein- oder zweireihige Warmgangeinhausungen aufbauen. Es gibt sie in unterschiedlichsten Ausführungen und Leistungsstufen von vielen Herstellern. Eine vollständige Übersicht würde den Umfang sprengen. Eine Auswahl: Climaveneta Side Cooler von Mitsubishi haben eine Bruttokälteleistung von 16 bis 47 kW. Auch verfügbar sind Varianten mit redundantem Wasser- oder Flüssigkeitskreislauf. Je nach Kälteleistung sind die Geräte 30 oder 60 cm breit und 100 oder 120 cm tief. Es bestehen keine Abhängigkeiten zu Racks bestimmter Anbieter.

Rittal bietet seine Cooler passend zu den eigenen Racks an. Zwar existieren keine technischen Abhängigkeiten, doch sind sie in Design und Höhe angepasst. Die Reihenkühler von Schneider Electric skalieren bei 30 cm Breite und 1095 mm Tiefe von 40 bis 60 kW. Leistungsfähigere Systeme sind 60 cm breit. Auch hier gibt es Ausführungen für Wasser oder Kältemittel.

Mit den Reihenkühlern von Stulz sind vorhandene Racks nachrüstbar. Die Kühlleistung der Kaltwasserausführung reicht von 11 bis 58 kW. Das Besondere: Die Kühler in der Breite 30, 40 oder 60 cm und der Bautiefe 100 oder 120 cm stehen 20 cm vor und blasen die kalte Luft seitlich aus, sodass sie ohne Umwege zu den Racks strömen kann. Die drei unten, mittig und oben angeordneten Klimazonen werden separat geregelt.

Seitenkühler

Seitenkühler kommen ohne Umweg über den Warmgang aus. Bei dieser Bauform ersetzt man ein Seitenblech des Racks durch ein Luftführungsblech, dadurch bilden Rack und Kühler eine Einheit. Der größte Unterschied zur Warmgangeinhausung respektive Reihenkühlung besteht darin, dass die Racks zu allen anderen Seiten geschlossen sein müssen.

Der Seitenkühler saugt die bis zu 46 °C warme Luft aus dem Bereich zwischen Serverrückseite und Rücktür ab und bläst die auf 25 bis 30 °C abgekühlte Luft seitlich zwischen Fronttür und Serverfront zurück. Beide Bereiche müssen auch hier durch Blenden voneinander getrennt sein. Da bei dieser Technik die bewegte Luftmenge deutlich geringer ist als bei den mit einem Warmgang arbeitenden Varianten, ist sie erheblich effizienter.

Zwei Racks können sich eine Einheit teilen, genauso können zwei Seitenkühler ein Rack bedienen. Da Rack und Kühler zusammenarbeiten müssen, sind es oft die Rackhersteller, die passende Seitenkühler offerieren. Eine Redundanz lässt sich auch herstellen, wenn die Seitenkühler modular aufgebaut sind und aus mehreren Segmenten bestehen. Damit lässt sich die Kühlung auch feiner abstufen. Technisch unterscheiden sich Seiten- und Reihenkühler nicht, nur darin, ob sie die Luft frontal oder seitlich einziehen und ausblasen. Verantwortlich dafür ist die Anordnung von Wärmeübertragern, Ventilatoren und Luftleitblechen. Wärmeübertrager und Ventilatoren sind häufig redundant ausgelegt. Zudem bieten alle Kühlsysteme eine Netzwerkanbindung, ein Monitoring und eine regelbasierte Steuerung.

Beispiel WOBCOM

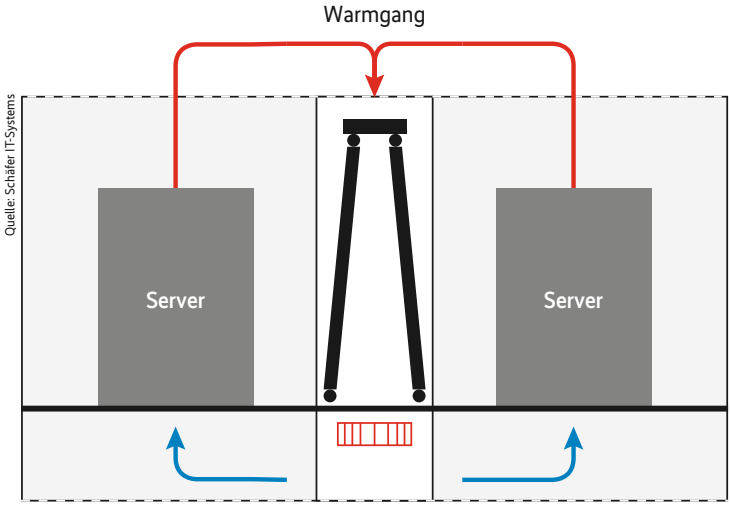
Die besonders für kleinere Rechenzentren vorteilhafte Warmgangeinhausung findet sich beispielsweise im Wolfsburger Nordkopfcenter, gleich neben dem Bahnhof und nur durch den Mittellandkanal von der Autostadt getrennt. Den Kernbereich des keilförmig geschnittenen achtstöckigen Bürogebäudes des Energieversorgers LSW aus dem Jahr 2017 nutzt die WOBCOM, ebenfalls eine Tochter der Stadtwerke Wolfsburg, als Rechenzentrum. Die ungewöhnliche Lage für einen Rechenzentrumsneubau begründet das umfangreiche Glasfasernetz, das nahezu alle in Wolfsburg ansässigen KMU so von zentraler Stelle erschließt.

Ungewöhnlich sind auch die Räumlichkeiten, in denen die IT untergebracht ist, denn sie erstrecken sich über alle acht Etagen, von denen bisher fünf genutzt werden. Bedingt durch die Bauform des Gebäudes, das sich dem Verlauf der Straßen anpassen muss, ist der Kernbereich nicht rechtwinklig, sondern erinnert an ein angeschnittenes Quadrat. In jeder Etage finden 15 Racks mit 48 Höheneinheiten Platz, die über Abschottungen hermetisch mit dem Warmgang abschließen. Zwei vor den Wänden installierte einreihige Warmgangeinhausungen kapseln die auf maximal 27,5 Grad erwärmte Luft ein, die von Reihenkühlern auf 22 Grad abgekühlt ihren Kreislauf durch die Racks erneut startet (siehe Abbildung 1).

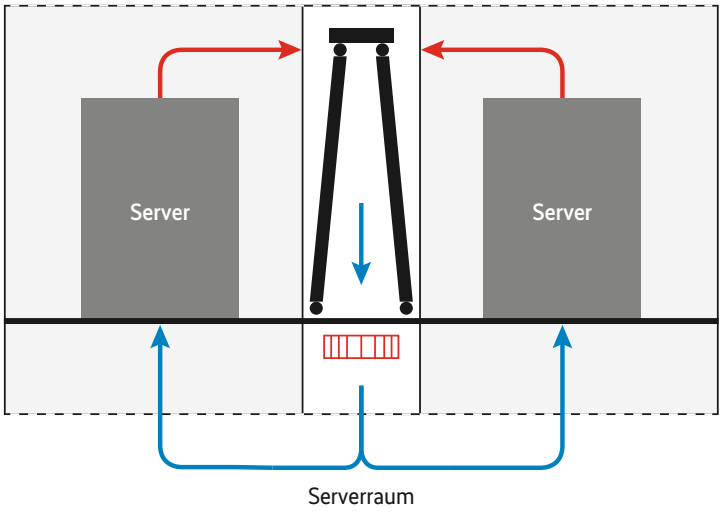
Die Bautiefe der Climaveneta Rack Cooler bestimmt auch die ungewöhnliche Racktiefe von 120 cm. Nicht alle verwendeten IT-Komponenten sind ideal für derart tiefe Racks, wie Giovanni Coppa von der WOBCOM erläutert. Beispielsweise würde die kurze Bauform der verwendeten Router zu einem thermischen Kurzschluss führen, was aus Blech gefertigte Lufttunnel verhindern.

Da die Racks in allen RZ-Etagen exakt gleich angeordnet sind, strömt das Wasser über alle Etagen vom Dach bis zum Keller. Im Kellergeschoss ist die Niedrigtemperaturheizung und die Klimatisierung untergebracht. Die Rückkühlung ist auf dem Dach installiert.

Racks und Kühler bilden eine Reihe. Passend zu den Kühlern wurden Racks mit 120 cm Tiefe und 48 HE gewählt (Abb. 1).



Bei Schäfers Hybridmodell mit geschlossener Front geht die Luft durch den Warmgang, aber nicht durch den Serverraum (Abb. 2).



Bei dem Hybridmodell mit geschlossener Rückseite geht die Luft durch den Serverraum, aber nicht durch einen Warmgang (Abb. 3).

Die 1200 mm tiefen Climaveneta Coolside CW von Mitsubishi liefern eine Kühlleistung von 8,81 bis 68,4 kW, Rittals Liquid Cooling Package LCP Rack CW 30 bis 53 kW – fürs hauseigene Rack. Die Economy-Versionen der SideCooler von Schäfer IT-Systems schaffen mit zwei Wärmeübertragern 10 bis 30 kW. Sie sind wie alle SideCooler-Versionen 30 cm breit, in unterschiedlichen Höhen und Tiefen erhältlich und zwischen herstellerfremden Racks aufstellbar.

Hybride

Gleich für mehrere Kühlarchitekturen einsetzbar sind die Enterprise- und HPC-Modelle (High-Performance Computing) der SideCooler-Serie. Sie lassen sich als Reihen- oder als Seitenkühler verwenden, außerdem für Hybridarchitekturen. Eine Warmgangeinhausung verlangt die Hybridvariante mit geschlossener Front und offener Rückseite. Die Kühler ziehen die Luft wie ein Reihenkühler aus dem Warmgang, geben sie aber nicht in den Serverraum, sondern direkt seitlich in die Racks rechts und links ab (siehe Abbildung 2).

Dieser Hybrid versucht, mehrere Vorteile der offenen Reihen- und der geschlossenen Seitenkühlung zu verbinden: Wie die Warmgangeinhausung und die Reihenkühlung gleicht diese Variante die unterschiedliche Wärmelast einzelner Racks aus und bietet dem Wärmestrom mehr Platz und weniger Kollisionspotenzial mit der Verkabelung. Nach vorn hin verringert sie zugleich die Geräuschemissionen im Serverraum – ein Vorteil aller Systeme mit geschlossenen Fronten.

Ohne Warmgangeinhausung kommt die Hybridvariante mit geschlossener Rückseite aus. Die Kühler ziehen die Abwärme seitlich direkt aus den Racks, geben die kühle Luft aber in den Serverraum ab (siehe Abbildung 3). Hier sind die Geräuschemissionen zwar höher, die Systeme sind durch die offene Front aber von vorn leichter zugänglich. Schäfers Reihen-, Seiten- oder Hybridkühler fürs Enterprise und HPC schaffen 30, 40 oder 60 kW bei einem Luftstromvolumen von 5000, 6000 oder 8000 m³/h. Als Kühlmedien können Kaltwasser oder Wasser-Glykol-Gemische dienen. Damit die Kühler die Luft je nach verwendeten Front- und Seitenabdeckungen frontal oder seitlich einziehen und ausblasen können, sind die Wärmeübertrager schräg angebracht. Bei den Varianten mit geschlossener Front, bei denen die Kaltluft direkt in die Racks zurückgelangt, kann die Kaltwassertemperatur bis zu 30 °C hoch sein.

Schrankkühler

Bilden Seitenkühler und Rack eine bauliche Einheit, hat man es mit einem Schrankkühler zu tun. Wie bei der Seitenkühlung ist das Rackinnere samt Klimagerät hermetisch von der Umgebungsluft abgeschottet. Die Luftzirkulation im Rack muss durch eine Trennung zwischen Kalt- und Warmluftsegment gewährleistet sein. Zwischen den beiden Segmenten kann sich die Luft nur bewegen, wenn sie durch die Serverchassis und die Wärmetauscher hindurchströmt.

Rittal nutzt ein Rack mit 80 cm Breite, um seitlich hochkant einen Wärmetauscher einzubauen. Auch hier gilt das Split-Prinzip, bei dem die Rückkühlung außerhalb des Serverraums stattfindet. Die maximale Kühlleistung fällt mit 3000 Watt deutlich geringer aus als bei den Seitenkühlern.

Rückwandkühler

Ebenfalls vollständig auf eine Einhausung verzichten können Rückwandkühler, außerdem auf die hermetische Abschottung des Rackinneren nach außen. In diesem Fall sind die Wärmeübertrager, die kalte Luft in den Serverraum entlassen, in die Rücktür integriert. Bei diesem Verfahren bildet das Rack für sich eine Warmlufteinhausung, beschränkt auf den schmalen Bereich zwischen Serverrückseite und Rückwandkühler. Der muss aber vom Rest des Racks abgedichtet sein.

Der Vorteil dieses Prinzips liegt im geringen Platzbedarf, denn die Klimatüren sind nur 10 bis 12 cm tief. Allerdings ist jedes Rack mit einer Tür und Wasseranschlüssen auszustatten. Die Systeme von Schäfer IT-Systems und Rittal arbeiten ohne eigene Ventilatoren und sind entsprechend energieeffizient. Die Lüfterleistung der Systeme muss jedoch einen ausreichend hohen Druck aufbauen.

Rittals System liefert eine Kühlleistung von 10 bis 20 kW bei einer typischen Vorlauftemperatur von 15 °C, das von Schäfer 18 bis 30 kW bei frei wählbaren Vor- und Rücklauftemperaturen. Beide Hersteller verwenden wasserführende Scharniere, sodass

Algorithmen sind die größten Verschwender

Das Darmstädter Rechenzentrum für den Green IT Cube des GSI errang den Blauen Engel für Rechenzentren durch Rückwandkühler und Standardserver. Bei dem Konzept spielt Verdunstungskälte eine entscheidende Rolle. Das Wasser wird zu einem Kühlturm gepumpt, wo ihm die Wärme entzogen wird, indem ein Teil davon verdunstet. Strom benötigt nur die Umwälzpumpe. Verbraucht wird allerdings Wasser. Entwickelt hat das Konzept Volker Lindenstruth, Professor für Hochleistungsrechnerarchitektur an der Goethe-Universität und stellvertretender Direktor des Frankfurt Institute for Advanced Studies.

iX: Herr Prof. Lindenstruth, Sie haben das RZ quasi neu erfunden, Ihr Green Cube in Darmstadt wurde dafür mit dem Blauen Umweltengel ausgezeichnet. Wodurch wurden Sie motiviert? Eigentlich ist das ja nicht die Aufgabe eines Professors für Hochleistungsrechnerarchitekturen.

Prof. Lindenstruth: Weltweit verursachen Superrechner jährlich rund 30 Milliarden Euro Stromkosten, davon allein 12 Milliarden für die Kühlung. Das erschien mir zu hoch. Außerdem musste ich vor mehr als zehn Jahren eine Lösung finden, bei gedeckeltem Budget enorme Rechenleistung zur Verfügung zu stellen. Also habe ich nicht nur den Rechner, sondern auch das Rechenzentrum und die verwendeten Algorithmen optimiert. Wenn Rechner und Rechenzentrum die Ressourcen optimal nutzen, werden sie schon aus diesem Grund „grün".

Ein großes Problem war damals die Luftkühlung. Rund 60 Prozent der teuren Energie der Server wurde zusätzlich für die Kaltluftkühlung benötigt. Luft leitet die Wärme extrem schlecht, um die Computer sicher zu kühlen, sind daher enorm tiefe Temperaturen nötig. Bei dem Konzept meiner Arbeitsgruppe wird mit Wasser gekühlt, das durch Wärmetauscher in den Servertüren geleitet wird. Die Lamellen sind Industriestandard und werden auch im Automobilbau verwendet.

Alle handelsüblichen Server können verwendet werden, dadurch sinken die Hardwarepreise für die Rechner. Im Massenmarkt herrscht ein weit größerer Preisdruck als im Bereich der Spezialkonstruktionen, die oft nur in winzigen Stückzahlen gefertigt werden. Die Wasserkühlung spart zudem Platz. Da der Doppelboden wegfällt, stapeln wir die Racks in Hochregallagern. Die Technik entstammt der Architektur von Parkhäusern. Beim Neubau unseres Rechenzentrums in Frankfurt wurde alles kostengünstig und schnell verschraubt. Schweißen war unnötig.

Prof. Volker Lindenstruth

Das erhitzte Kühlwasser wird außerhalb des Rechenzentrums durch einen Kühlturm geleitet. Kühltürme sind ein enorm effizientes Kühlsystem, weil es Verdunstungskälte nutzt. In Frankfurt können wir das Wasser des Main nutzen, es kostet also nichts. Aber auch wer Brauchwasser zukaufen muss, kommt immer noch sehr günstig zur benötigten Kühlleistung. 1 MW Kühlleistung verbraucht 2 m³/h Wasser.

In Darmstadt wurde ein Kühlturm als Rückkühler installiert. Warum nutzt man in so einem Fall nicht einfach Freikühler?

Wärmetauscher benötigen an vielen Tagen im Sommer große Lüfter, um dem Kühlwasser die Wärme zu entziehen. Verdunstungskühlung funktioniert in unseren Breiten das ganze Jahr. Bei dieser Methode steht die Wasserabgabe einer feuchten Oberfläche mit dem Wasseraufnahmevermögen der umgebenden Luft im Gleichgewicht. Entscheidend ist die Kühlgrenztemperatur, die aber in Deutschland unkritisch ist. Auch im Hochsommer können wir ohne Zusatzaggregate kühlen.

Wieso setzen sich Neuerungen in RZs nur langsam durch?

Der Markt ist sehr konservativ und die Planungszeiten sind lang. Es vergehen fünf bis zehn Jahre, bis ein neues Rechenzentrum betriebsbereit ist. Die für die Planung verantwortlichen Abteilungen haben mit den laufenden Betriebskosten oft nichts zu tun. Der Leidensdruck in diesem Bereich ist entsprechend gering. Die enorm hohen Stromrechnungen galten zudem lange als unvermeidbar.

Manche Entscheider haben zudem schlicht Angst vor Wasser in der Nähe ihrer teuren Computer. Doch auch in einem normalen Rechenzentrum gibt es Wasser, von der Klimaanlage bis zur Feuerlöschanlage. Auch hier kann mal eine Leitung platzen. Auf Wunsch arbeitet unser System mit Unterdruck, dadurch tritt kein Wasser aus. Dazu ist ein zusätzlicher Wassertank nötig. Ein Aufwand, den man sich aber auch sparen kann.

sich die Türen weit öffnen lassen. Mit Racks anderer Hersteller sind die Systeme nicht kompatibel. Da diese Türen ohne Ventilatoren auskommen, verteilen Filter die Luft über die Fläche. Redundanz ist nicht gegeben und aufgrund der geringen Komplexität auch nicht notwendig.

Anders arbeitet die Cooling Door von Mitsubishi. Die in unterschiedlichen Höhen verfügbare Tür mit einer Kühlleistung von 27 bis 39 kW lässt sich in vorhandene Racks einbauen. Die Cooling Door nutzt Ventilatoren und Sensoren, um Hotspots gezielt zu kühlen. Die Flüssigkeit wird über flexible Stahlschläuche geführt und kann redundant in zwei Kreisläufen ausgeführt werden, die Vorlauftemperatur des Wassers darf bis zu 20 °C betragen.

Auch die CyberRack Active Rear Door von Stulz kann die Rücktür eines Racks ersetzen und es so in ein gekühltes Serverrack verwandeln, das dann durch die Serverschrankkühlung keine Wärmelast mehr in den Raum abgibt. Die Bautiefe beträgt 33 cm. Die Wärmetauschtür von Stulz verwendet ebenfalls Ventilatoren und kann das Rack in verschiedene Klimazonen einteilen. Die maximale Kühlleistung variiert zwischen 19 und 32 kW.

Einschubkühler

Ohne Spezialracks kommen Rack Cooler aus. Sie werden als 19″-Einschub in vorhandene Racks eingesetzt, belegen typischerweise die unteren sieben Höheneinheiten und bieten eine Kälteleistung von 3 bis 7 kW.

Fazit

Reihen- oder Seitenkühler benötigen Platz und sind ideal für kleinteilige Rechnerräume und Systeme mit hoher Energiedichte. Warmgangeinhausungen und Reihenkühler skalieren gut –

Schäden durch Wasser hatten wir noch nie. Zudem schützen Sensoren und Ventile die Computer. Viel wichtiger bei der Kühlung mit Wasser ist die Analyse auf Bakterien und vor allem Legionellen. Darauf muss man achten, will man die Umgebung nicht gefährden.

Die Anzahl der Rechenzentren, die unsere Technologie nutzen, wächst ständig. Neben dem Rechenzentrum meiner eigenen Universität, der Goethe Universität Frankfurt, existiert der Green Cube in Darmstadt und das an dieses Rechenzentrum angelehnte Datacenter von Airbus in Ottobrunn bei München. Auch Finanzinvestoren haben ihr Interesse bekundet, doch wird in diesem Bereich über Verträge nicht geredet. Weitere Rechenzentren sind in Planung.

Wie könnte man die IT noch effizienter, also „grüner" gestalten?

Durch bessere Software. Die meisten Pakete enthalten Algorithmen, die für moderne Prozessoren Gift sind. Hier wird viel Leistung verschwendet. Mein Team hat das Open-Source-Paket für Hochenergiephysik, mit dem beim CERN in Genf die Spuren der Teilchenkollisionen ausgewertet werden, komplett neu geschrieben und auf moderne Grafikkartencluster optimiert. Es wurde dadurch um den Faktor 10 000 schneller. Entsprechend weniger Server werden benötigt. Kommerzielle Anbieter sehen wenig Veranlassung, Ähnliches zu tun. Es ist viel lohnender, mehr Lizenzen für noch mehr Server zu verkaufen. Auch die teuren Supercomputer für die Forschung liefern oft nur zu 10 Prozent der theoretisch möglichen Leistung, weil sie mit veralteter und ineffizienter Software betrieben werden.

Hat Ihre Technik Grenzen?

Jede Technik hat Grenzen, aber wir können Wärmelasten bis zu 50 kW pro 19"-Schrank mit unserer Technik sicher kühlen – 70 kW wurden kürzlich demonstriert. Bei höheren Leistungen müsste man zur direkten Wasserkühlung der Prozessoren greifen, aber auch diese Technik ist in unser Konzept integrierbar. Das haben wir schon gezeigt. Mit der von der ETH Zürich entwickelten Warmwasserkühlung lässt sich aber nicht die gesamte Wärme aus dem Server bringen, es besteht zusätzlicher Kühlbedarf. Wenn die Prozessorleistungen über 1000 Watt steigen sollten, werden wir aber wohl auf diese Technik zurückgreifen müssen.

Was benötigt der Planer für ein günstiges und „grünes" Rechenzentrum?

Inhouse-Kompetenz! Wer sich ausschließlich auf die Industrie verlässt, zahlt leicht das Doppelte. Google und Amazon haben das erkannt und bauen ihre Rechenzentren selbst. Wer 1000 Server benötigt, bekommt heute ohne Aufpreis ein maßgeschneidertes Motherboard. Er muss es nur noch integrieren. Zum Beispiel mit Grafikkarten. Diese sind für viele Aufgaben im HPC-Bereich das Mittel der Wahl, wie ein Blick auf die Top500-Liste der Superrechner belegt. Unser jüngstes Projekt war ein Rechner für das CERN-Experiment ALICE in Genf, der ohne Hardwareexpertise auf unserer Seite um den Faktor sieben teurer geworden wäre.

Der SuperMUC in Garching wird mit warmem Wasser gekühlt, Ihre Server im Green Cube in Darmstadt mit kaltem. Was ist der Unterschied?

Im Falle des SuperMUC werden die Kühlkörper der Prozessoren direkt mit warmem Wasser gekühlt. Hierdurch hat das Kühlwasser insgesamt höhere Temperaturen und kann im Anschluss in Adsorptionskältemaschinen benutzt werden, um kaltes Wasser für weitere Kühlung zu verwenden. Diese Kühlsysteme müssen für die spezifische Hardware angepasst werden. Im Green Cube werden einfach handelsübliche Computer eingesetzt, die auf dem Massenmarkt verfügbar sind und damit aus dem Marktsegment mit dem größten Wettbewerb kommen. Hier wird der Server selbst mit Luft gekühlt und die Luft dann mithilfe eines Wärmetauschers in der Rückwand des Racks gekühlt.

Welche Baustellen gibt es noch für „grüne" Rechenzentren?

Wir erleben gerade einen Boom bei Edge-Rechenzentren. Das sind kleine Systeme, die zum Beispiel in einen Container integriert wurden. Man kann sie schnell montieren und demontieren. Sie bringen Rechenleistung an wichtige Netzknoten, etwa für 5G. Hier werden lokale Netzwerke eine bedeutende Rolle beim Internet der Dinge spielen. Meine Arbeitsgruppe hat auch in diesem Bereich ein Patent angemeldet, das zeigt, wie Wasserkühlung auch unter diesen beengten Bedingungen umgesetzt werden kann.

Das Interview führte Bernd Schöne.

auch in der Redundanz. In der Regel lassen sich Racks anderer Hersteller mit Gitter- oder ganz ohne Türen verwenden, so Höhe und Tiefe passen. Einige der Seiten- und Rückwandkühler eignen sich für die Umrüstung vorhandener Racks. Allerdings ist pro Rack ein Rückwandkühler notwendig, Seitenkühler können allein zwei Racks oder zu zweit ein Rack versorgen. Verwirrend können mitunter die Bezeichnungen der Hersteller sein, die zuweilen eine Systematik vermissen lassen.

Allen Techniken gemein ist, dass eine freie Luftbewegung zwischen dem Kaltluftsegment vor und dem Warmluftsegment hinter den Servern durch Blenden zwischen und neben den Servern unterbunden werden muss, da der durch die Serverlüfter aufgebaute Überdruck im Wärmesegment ausschließlich über die Kühler abgebaut werden soll. Teilweise sind die Blenden für die auf systemnahe Kühlungen zugeschnittenen Racks verschiebbar. Sollten die Hersteller der umzurüstenden Racks keine passenden Blenden liefern, kann eventuell der lokale Metallbauer aushelfen.

Auch sind in allen Fällen Wasser- oder Kühlmittelleitungen bis zu den Racks oder den Rackreihen zu verlegen. Dafür können aber alle vorgestellten Varianten – richtig dimensioniert – die Serverwärme zu 100 Prozent abführen, ganz ohne Kühlung der Serverräume. Zum Vergleich: Eine Flüssigkeitsplattenkühlung kann etwa 80 Prozent der Wärme fortschaffen, eine hundertprozentige systemnahe Flüssigkeitskühlung schafft, ganz ohne Ventilatoren, sonst nur die Immersions- oder Tauchkühlung – deren Implementierung aber wesentlich aufwendiger ist (siehe Artikel „Ohne Umwege" ab Seite 140). (sun@ix.de)

Hubert Sieverding

arbeitet nach langjähriger Tätigkeit in der Automobilbranche als freier Autor.

Flüssigkeitskühlungen für Server

Ohne Umwege

Fridtjof Chwoyka

Technisch und ökologisch hat die Luftkühlung längst ausgedient. Doch auf die Flüssigkeitskühlung sind RZ-Server noch nicht gut vorbereitet.

Mit dem stetig steigenden Energiebedarf der Rechenzentren im Allgemeinen steigt auch der Energiebedarf der dortigen Kühlung. Denn jede elektrische Energie, die ein Server als Leistung aufnimmt, wandelt er in Wärme um, die mit hohem technischen und energetischen Aufwand aus den Rechenzentren abgeführt werden muss. Dabei bilden die Prozessoren eines Servers die größte Wärmequelle. Ihre Kühlung dient vor allem dem Schutz vor Überhitzung und Ausfall. In heutigen Servern tummeln sich aber immer mehr Prozessoren, etwa GPUs, KI- und Netzwerkprozessoren. Ihnen gesellen sich unter anderem Spannungswandler und Speicherbausteine als weitere relevante Abwärmequellen hinzu.

Zum Abführen der Abwärme bietet sich zunächst die Umgebungsluft als Kühlmedium an. Luft ist überall verfügbar, die Technik ist einfach und ausgereift. Bis zu einem gewissen Maß der Leistungsaufnahme der Prozessoren genügt deren Oberfläche, den stetigen Wärmestrom an die Umgebung abzuführen. Steigt die Leistungsaufnahme, reicht ihre Oberfläche nicht mehr aus, die Abwärme an die Umgebung abzugeben und gleichzeitig die zulässigen Temperaturen im Chip einzuhalten.

Mit luftdurchströmten Kühlkörpern lässt sich die Oberfläche der Prozessoren vergrößern und die Wärme innerhalb der zulässigen Temperaturen abführen – bei heutigen Servern State of the Art. Die abführbare Wärmemenge hängt dabei maßgeblich von der Fläche des Kühlkörpers, der Temperaturdifferenz zwischen der Umgebung und dem Kühlkörper sowie dem verwendeten Kühlmedium, der Luft, ab.

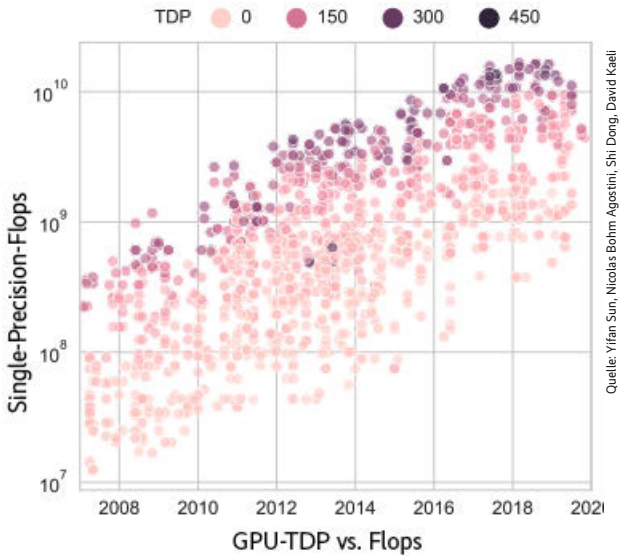
Für die Leistungsfähigkeit der Wärmeübertragung hat die wirksame Fläche des Kühlkörpers einen großen Anteil. Vereinfacht heißt das: Je größer die wirksame Fläche eines Kühlkörpers ist, desto größer der Wärmestrom, der sich auf das Kühlmedium übertragen lässt. Heute begrenzt vor allem der im Servergehäuse verfügbare Platz die Größe des Kühlkörpers. Verbessern lässt sich dessen Leistungsfähigkeit mit einem gezielten Luftstrom, üblicherweise erzeugt von vielen kleinen Ventilatoren, die die Luftmenge in Abhängigkeit von der Prozessortemperatur regeln können (siehe Artikel „Ein eigenes Ökotop“ ab Seite 84). Die Grundlagen der Wärmeübertragung liefert die Gleichung

$$\dot{Q} = \dot{M} c_p \Delta\vartheta$$

Dabei beschreibt $\dot{Q}$ den übertragenen Wärmestrom, $\dot{M}$ den Massenstrom des Kühlmediums, c_p ist die spezifische Wärme-

X-TRACT

- Die Wärmekapazität von Luft ist begrenzt. Auch mit Kühlkörpern und Lüftern gelangt die Luftkühlung bei einer TDP von 300 W an ihr Limit.
- Deutlich mehr Wärme kann die Flüssigkeitskühlung abführen, die sich technisch in zwei Arten unterteilt.
- Die Flüssigkeitsplattenkühlung überträgt die Wärme über die auf den Wärmequellen angebrachten Kühlplatten in den durchs Rack geführten Kühlkreislauf.
- Bei der Immersionskühlung werden die Systeme komplett in nicht leitende Kühlflüssigkeit getaucht.



Zwischen 2008 und 2010 stieg das obere Ende der TDP von GPU von 150 auf 300 Watt (Abb. 1).

kapazität des Kühlmediums und $\Delta\vartheta$ ist die Temperaturdifferenz des Kühlmediums zwischen dem Eintritt in den Kühlkörper und dem Austritt aus dem Kühlkörper. Diese Beschreibung erfasst aber nur den Wärmestrom des Kühlmediums. Eine komplette Beschreibung der Wärmeübertragung von Prozessoren über die Kühlkörper an die Umgebungsluft muss die Geometrie des Kühlkörpers berücksichtigen; darauf soll aufgrund der Komplexität hier verzichtet werden.

Begrenzte Wärmekapazität der Luft

Drei wesentliche Faktoren begrenzen den Wärmestrom, den Kühlkörper abführen können: Erstens lässt sich die Luftmenge $\dot{M}$ nur in gewissen Grenzen anheben, bevor der Verlust durch hohe Strömungswiderstände zu groß und die Kühlung ineffizient wird. Der zweite Faktor ist die Temperatur der Kühlluft. Je kälter sie ist, desto mehr Leistung kann sie abführen. Die Kälte ist aber gerade im Sommer mit hohem energetischen Aufwand zu produzieren.

Der dritte und wesentliche Faktor ist die geringe spezifische Wärmekapazität von Luft. Diese beträgt bei 20 °C und 1 bar Luftdruck 1,007 kJ/kg K. Wasser dagegen hat eine spezifische Wärmekapazität bei 20 °C und 1 bar von 4,181 kJ/kg K. Das heißt: 1 kg Wasser kann einen viermal so großen Wärmestrom transportieren wie 1 kg Luft. Hersteller geben für die Wärmeabgabe ihrer Prozessoren eine TDP (Thermal Design Power) an. An ihr ist die Kühlung auszulegen. Betrachtet man die Entwicklung der TDP performanter GPUs, zeigt sich, dass sie sich von 150 Watt im Jahr 2008 auf über 300 Watt im Jahr 2010 verdoppelt hat. Danach hat sich der Anstieg nicht weiter fortgesetzt und die TDP-Werte für GPUs liegen konstant bei etwa 300 Watt. In Abbildung 1 ist diese Entwicklung anhand der dunkler werdenden Punkte gut zu erkennen.

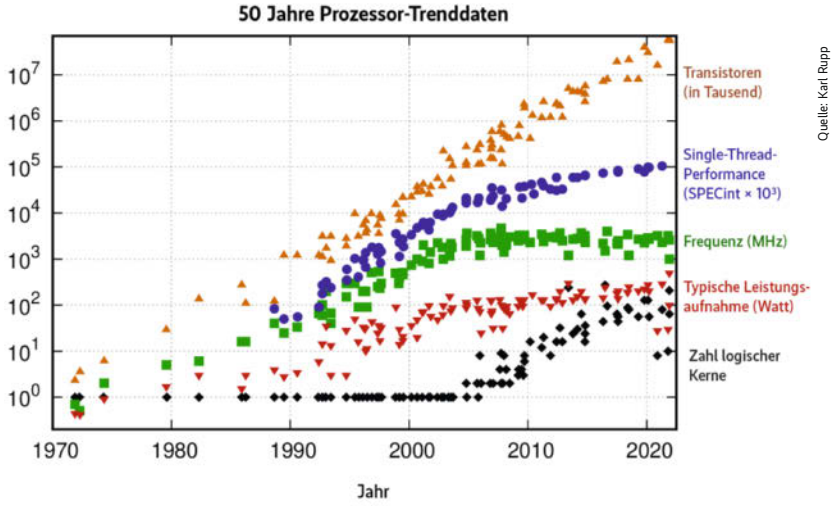
Bei CPUs steigt der Energiehunger seit etwa 2002 nur noch leicht an und nähert sich der 250-Watt-Marke (siehe Abbildung 2). Der Grund, warum die TDP-Werte nicht weiter ansteigen: Mit einer Luftkühlung lässt sich eine TDP von mehr als 300 Watt nicht mehr effizient bewältigen, weil sowohl die notwendigen Kühlkörper als auch die Strömungswiderstände durch die notwendigen Luftmengen zu groß wären. Dennoch deuten die Entwicklungsroadmaps der Prozessorhersteller darauf hin, dass die TDP-Werte in den nächsten Jahren auf über 400 Watt steigen werden.

Mit Platten bestücken oder tauchen

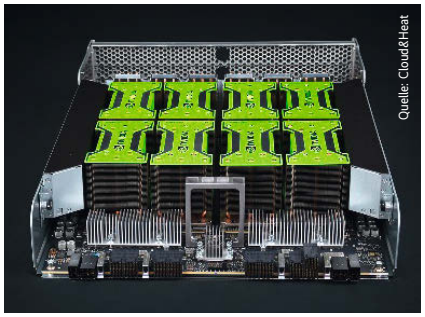
Das würde den Umstieg auf ein Kühlmedium mit einer hohen spezifischen Wärmekapazität wie Wasser erfordern. Technisch gibt es zwei Möglichkeiten, Prozessoren mit Wasser und anderen Flüssigkeiten zu kühlen: Man verwendet entweder Flüssigkeitskühlplatten für das Cold-Plate Liquid oder Direct Liquid Cooling, oder eine Immersions- oder Tauchkühlung, genannt Immersion oder Liquid Submersion Cooling. Bei ersterer Technik platziert man Kühlplatten auf den Wärmequellen im Server, durch die eine Flüssigkeit hindurchströmt. Bei der Immersionskühlung taucht man das gesamte Gerät in die Kühlflüssigkeit ein.

Die Flüssigkeitsplattenkühlung für Server ist nicht neu. Bereits in den 1980er-Jahren setzte unter anderem IBM sie für ihre Großrechner ein. Nach einer kurzen Pause in den 1990er-Jahren kam sie im Jahr 2008 mit IBMs System Power 575 wieder zurück. Die größte Verbreitung hat die Flüssigkeitsplattenkühlung inzwischen beim HPC (High-Performance Computing) erfahren (siehe Artikel „Vorangeschritten“ ab Seite 128). Derzeit breitet sie sich auch auf Bereiche jenseits des HPC aus, zusammen mit der noch neuen Technik der Immersionskühlung, die etwa im Jahr 2019 die Marktreife erlangte. Die einzige Gemeinsamkeit beider Kühltechniken besteht aber darin, dass beide eine Flüssigkeit als Transportmedium verwenden.

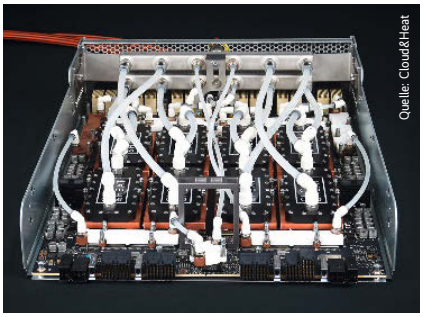
Der Wechsel von Luft auf Wasser bedeutet auch einen Wechsel des Aggregatzustands und damit auf ein Medium mit einer wesentlich höheren Dichte. In dem für Rechenzentren infrage kommenden Temperaturbereich benötigt Wasser weniger Platz als die gleiche Menge Luft. Zum Vergleich: Um einen Wärmestrom von 300 Watt abzuführen, müssen bei einer Luftkühlung 35 m³/h, also etwa der Rauminhalt eines Zweipersonenbüros



In den letzten 20 Jahren hat sich der Energiebedarf 250 Watt angenähert. Im selben Zeitraum stieg die Zahl der Kerne und Transistoren deutlich schneller (Abb. 2).



Die hohen luftdurchströmten Kühlkörper und die Luftleitsysteme benötigen viel Platz (Abb. 3).



Deutlich weniger Platz benötigen dagegen die Flüssigkeitskühlplatten und die Schläuche (Abb. 4).

pro Stunde, bewegt werden, während eine Wasserkühlung nur 0,0125 m³ Wasser pro Stunde, also 12,5 l/h, benötigt.

Das hat Folgen für das Serverdesign. Die Abbildungen 3 und 4 verdeutlichen den Unterschied zwischen einer Luft- und einer Flüssigkeitsplattenkühlung anhand eines dicht gepackten Rechenmoduls: Bei der Variante mit der Luftkühlung sind die großen Kühlkörper für die notwendige Bauhöhe des Chassis verantwortlich (siehe Abbildung 3). Bei dem identischen Rechenmodul in Abbildung 4 ersetzen flache Kühlplatten die luftdurchströmten Kühlkörpertürme. Ohne die Schläuche hat sich die Höhe der eingebauten Komponenten halbiert.

Mehr Platz im Rack

Bei diesem System sind auch weitere Einheiten wie Netzwerkchips und Spannungswandler mit Kühlplatten ausgestattet. Allerdings müssen die Platten für jeden Servertyp individuell entworfen und produziert werden. Die Flüssigkeitsplattenkühlung erfasst etwa 80 Prozent der Abwärme, die anderen 20 Prozent werden per Luft gekühlt. Dies betrifft vor allem die Netzteile und einige weitere Komponenten, die sich derzeit nicht sinnvoll mit Flüssigkeit kühlen lassen.

Durch die Schläuche wird das Kühlwasser zu den Kühlplatten geführt und in ihnen von den Prozessoren erwärmt (siehe Abbildung 5). Das erwärmte Wasser wird auf Rackebene im TCS (Technology Cooling System) gesammelt und noch im Rack über Wärmeübertrager, genannt CDUs (Cooling Distribution Units), an das Gebäudekühlsystem FWS (Facilities Water System) übergeben. Die Abwärme aus dem Gebäudekühlsystem wird über den Rückkühlkreis, das CWS (Condenser Water System), mit einem Rückkühler an die Umgebungsluft oder einen Heizkreislauf abgeführt. Je nach Konzept lässt sich die CDU in einen dedizierten Schrank auslagern, der sich meist zwischen den Racks befindet.

Einen kompletten Kühlkreis für die Flüssigkeitskühlung zeigt Abbildung 6. Angeschlossen sind dort eine interne und eine externe CDU. Ob die Schränke intern mit Platten- oder Tauchkühlung arbeiten, spielt für den rechenzentrumsweiten Kühlkreislauf keine Rolle. Als Kupplung zwischen Server und Rack dienen doppeltdichtende und damit tropffreie Steckverbindungen. Dadurch lassen sich Server oder Komponenten austauschen, ohne dass das Kühlmedium austritt oder der Kühlkreislauf abgeschaltet werden muss. Im Unterschied zum Server mit Luftkühlung sind hier nur zwei Stecker mehr abzuziehen.

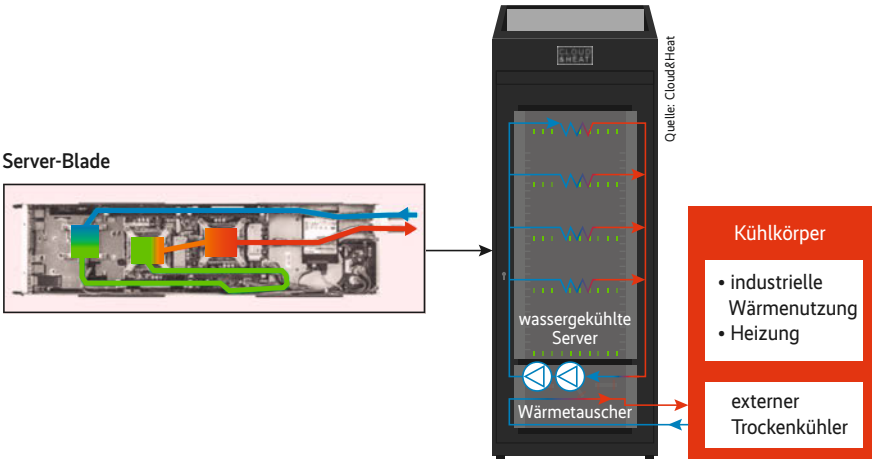
Rundumkühlpaket

Während man das Kühlmedium bei der Flüssigkeitsplattenkühlung in einem geschlossenen Kühlkreislauf und damit vom IT-System fernhält, bringt man es bei der Immersionskühlung absichtlich mit den Systemen in Kontakt. Dazu taucht man die Server ohne zusätzliche Kühlkörper einfach in eine mit dem Kühlmedium gefüllte Wanne. Diese Bauart wird als offener Kühlkreislauf bezeichnet, da das Kühlmedium mit seiner Umgebung in Kontakt kommt (siehe Abbildung 7).

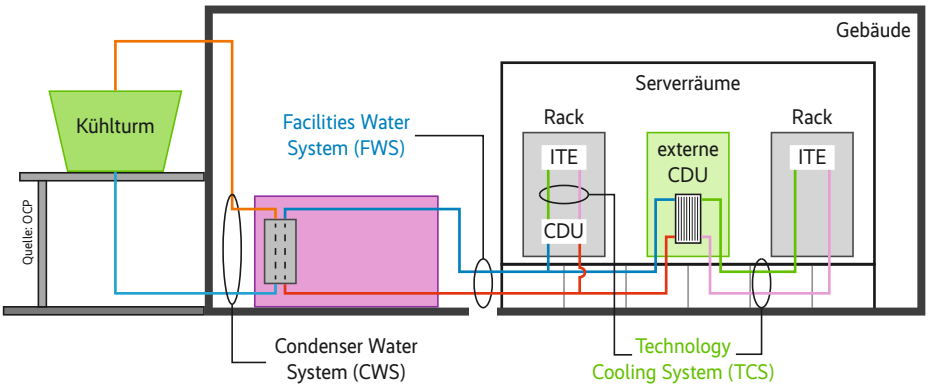
Das funktioniert selbstredend nur mit einem nicht leitenden Kühlmedium. Zu solchen dielektrischen Flüssigkeiten zählen entionisiertes Wasser, Flüssigkeiten aus Kohlenwasserstoffverbindungen, etwa Öle, sowie Fluorkohlenwasserstoffverbindungen. Meist verwenden die Anbieter Kohlenwasserstoffverbindungen. Dafür kommen keine klassischen Racks mehr zum Einsatz, sondern Pods, die eher an Gefriertruhen denn an Serverschränke erinnern (siehe Abbildung 8). Anders als die Platten- kann die Immersionskühlung auf jegliche Art von Kühlkörpern verzichten und so mit jeder Serverhardware funktionieren. Selbst die Netzteile werden bei ihr direkt gekühlt.

Flüssigkeit verlieren oder verdampfen

Zudem entfallen alle beweglichen und damit ausfallgefährdeten Teile im Server, vor allem Lüfter. Anpassungen sind besonders bei deren Steuerung vorzunehmen (siehe Kasten „Standardserver im Pod“). Darüber hinaus wird 100 Prozent der Serverabwärme durch die Kühlflüssigkeit abgeführt. Zusätzlich schottet das Kühlmedium die Server von der Umgebung ab und schützt sie dadurch auch gegen schädliche Umwelteinflüsse wie Staub und Feuer. Kurzschlüsse etwa durch zu hohe Luftfeuchtigkeit und Funkenbildungen sind ebenfalls nicht zu befürchten.



Statt kleiner Lüfter führen Wasserschläuche die Wärme ins Kühlsystem des Rechenzentrums ab, ohne dass die Raumluft der Serverräume und -hallen darin involviert ist (Abb. 5).



Der komplette Kühlkreis im RZ beschränkt sich bei der Flüssigkeitskühlung auf die Racks, die Leitungen und die Wärmeübertrager. In den schematischen Kühlkreislauf ist je ein Rack mit interner und mit externer CDU eingebunden (Abb. 6).

Dadurch, dass die Server in der Kühlflüssigkeit liegen, ist der Austausch eines Servers aufwendiger als bei anderen Kühltechniken. Der Pod selbst lässt sich währenddessen weiterbetreiben, nur der zu tauschende Server muss zuerst in ein spezielles Transportbehältnis gelegt werden, um die Kühlflüssigkeit nicht im Raum zu verteilen. Sie ist für den Menschen bei Hautkontakt ungefährlich, nur bei der Temperatur ist Vorsicht geboten.

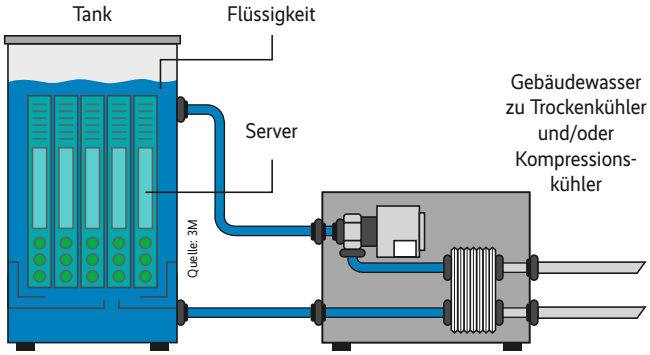
Beide Flüssigkeitskühltechniken arbeiten wie die Luftkühlung mit einer einphasigen Kühlung. Dies bedeutet, der Aggregatzustand des Kühlmediums ändert sich nicht, es findet kein Phasenwechsel statt. Den abtransportierten Wärmestrom bestimmt allein die spezifische Wärmekapazität des Kühlmediums. Mit einem speziellen Kühlmedium lässt sich aber auch eine Zweiphasenkühlung aufbauen.

In diesem Fall bestimmt die Verdampfungsenthalpie den abtransportierten Wärmestrom. Die Verdampfungsenthalpie ΔHv beschreibt die Energie, die notwendig ist, um eine Flüssigkeit in ein Gas umzuwandeln. Außerdem ist die Temperatur am Verdampfungsort auf die Verdampfungstemperatur festgelegt und kann nicht überschritten werden. Diese Art der Kühlung ist aus Sicht der Thermodynamik die leistungsfähigste Art des Wärmetransports. Um einen Wärmestrom von 300 Watt abzuführen, ist lediglich eine Menge von 0,133 g Wasser zu verdampfen. Für die Flüssigkeitsplattenkühlung sind Zweiphasensysteme bereits verfügbar und im Handling mit ihren Einphasen-Geschwistern identisch. Immersionskühlsysteme gibt es ebenfalls zweiphasig. Dabei verdampft die – in der Regel fluorhaltige – Flüssigkeit, das aufsteigende Gas kondensiert an einer Kühlspirale, die wiederum die Flüssigkeit eines Kühlkreislaufs erwärmt (siehe Abbildung 10).

Allerdings stellen hier die Verluste des Kühlmediums, die sich beim Servertausch im offenen Kühlkreislauf nicht verhindern lassen, eine zusätzliche Herausforderung dar. Bei der zweiphasigen Plattenkühlung treten aufgrund des geschlossenen Kreislaufs keine Flüssigkeitsverluste auf. Deshalb ist zu erwarten, dass sich bei der Immersionskühlung die Einphasen-Variante etablieren, bei den Kühlplatten sich aber die Zweiphasen-Version langfristig durchsetzen wird.

Vorteile gegenüber der Luftkühlung

Zukünftig ist eine Serverkühlung aufgrund des weiter steigenden Energiehungers der Prozessoren nur mit einer Flüssigkeitskühlung effizient und nachhaltig möglich. Beide Varianten



Bei der Einphasenimmersionskühlung lässt eine Pumpe die Kühlflüssigkeit durch einen Wärmeübertrager zirkulieren (Abb. 7).

können 100 kW und mehr aus einem Rack mit 42 Höheneinheiten zuverlässig abtransportieren. Ein luftgekühltes Rack schafft maximal 30 kW Leistungsaufnahme. Damit erlauben flüssigkeitsgekühlte Pods und Racks eine höhere Energiedichte und damit eine kompaktere Bauweise. Selbst mit den horizontal aufgestellten Immersionspods lässt sich die Fläche in einem Rechenzentrum besser nutzen: Die Pods müssen nur von einer Seite aus zugänglich sein, Warmgänge entfallen.

Die Temperaturen innerhalb der Kühlkreiskomponenten CWS, FWS und TCS liegen für beide Arten der Flüssigkeitskühlung typischerweise oberhalb von 45 °C, das entspricht der ASHRAE TC9.9 Facility Water Requirements Liquid Cooling Class W5. Zum Vergleich: Die klassische Luftkühlung arbeitet mit einer Kühlwassertemperatur im Gebäudekühlsystem bis 27 °C und damit im Bereich von Class W1 und W2.

Durch die hohen Systemtemperaturen von mehr als 45 °C lassen sich die Systeme in Europa ganzjährig mit freier Kühlung temperieren. Dadurch kann auf eine maschinelle Kälteerzeugung mit Kältemaschinen verzichtet oder die Anlagengröße stark reduziert werden. Auch der Einsatz adiabatischer Rückkühler ist nicht notwendig. Das spart Wasser und die komplizierte Anlagentechnik zur Wasseraufbereitung ein. Insgesamt werden die Systeme zudem auf der Gebäudekälteseite technisch einfacher und damit robuster.

Durch die hohe Systemtemperatur von mehr als 45 °C eignet sich die Flüssigkeitskühlung besser für die Abwärmenutzung. Typischerweise liegt die Systemtemperatur bei beiden Techniken bei 55 °C im Zulauf zum Server und bis zu 75 °C im Rücklauf. Bei den Systemen mit Zweiphasenkühlung liegen die Systemtemperaturen derzeit bei etwa 55 °C, haben aber noch

Einer Wanne oder einer Gefriertruhe ähneln die Pods, die mit Immersionskühlung arbeiten und die klassischen Racks ersetzen (Abb. 8).

Potenzial nach oben. Die Luftkühlung arbeitet mit Zulufttemperaturen von maximal 30 °C.

Noch einige Hürden zu nehmen

Dadurch, dass die Klimaanlage keine großen Luftmengen mehr durch die Räume des Rechenzentrums bewegen muss, vereinfacht sich auch dessen technische Infrastruktur. Die Menge der benötigten Umluftklimageräte reduziert sich im besten Fall auf null, der für den Lufttransport notwendige Platz entfällt. Dadurch, dass die Immersionskühlung auf die Lüfter in den Servern ganz verzichtet, steigt die Energieeffizienz der Server um bis zu 30 Prozent an; bei der Flüssigkeitsplattenkühlung erhöht sie sich um etwa 15 Prozent. Zugleich kann durch die leistungsfähigere Kühlung die Rechenleistung angehoben werden.

Die Nachteile beider Techniken halten sich in Grenzen und sind weniger der Technik selbst als vielmehr der fehlenden Op-

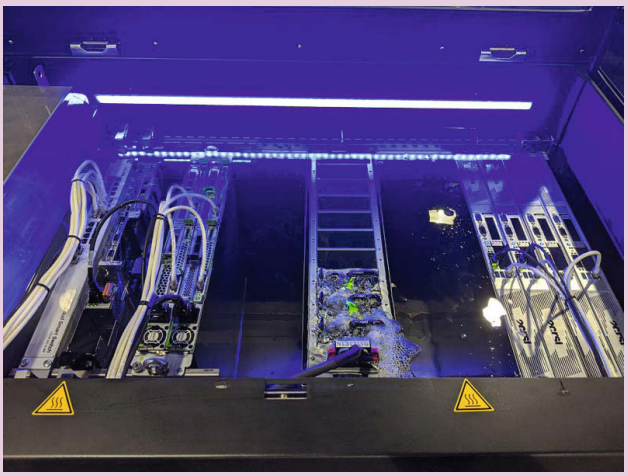
Standardserver im Pod

Ganz ohne Anpassungen lassen sich auf Luftkühlung optimierte Server nicht in einen Pod tauchen (siehe ix.de/zf9t). Zuerst ist die Materialkompatibilität zu der gewählten Flüssigkeit zu prüfen. Vor allem stellen die ein- und die zweiphasigen Kohlen- und Fluorkohlenwasserstoffe unterschiedliche Anforderungen an tauchbare Materialien. Zu berücksichtigen ist auch das thermische Design, da sich das thermische Verhalten der Komponenten in der Flüssigkeit verändern kann. Besondere Aufmerksamkeit verdient das elektronische Design, beispielsweise ist die Signalintegrität zu gewährleisten. Auch müssen BIOS-, Firmware- und IPMI-Funktionen einen dauerhaften Betrieb in der Flüssigkeit erlauben.

Heatpipes für Chips mit geringerer Leistungsaufnahme können entfallen, Lüfter sind in jedem Fall zu deaktivieren oder zu entfernen. Jegliches Airflow-Management muss deaktiviert werden. Das gilt auch für die Lüftererkennung und alle Sicherheitsfunktionen, die das Starten des Servers ohne funktionsfähige Lüfter verhindern. Zudem muss sich das Temperaturmanagement anpassen lassen, da die Temperatur der Flüssigkeit außerhalb der im BIOS eingestellten Schwellenwerte liegen kann.

Die meisten offenen Servernetzteile sind für die Immersionskühlung geeignet. Sie müssen aber vollständig in die Flüssigkeit eingetaucht sein. Integrierte Lüfter müssen deaktiviert, abgeklemmt oder entfernt werden. Etwaige integrierte thermische Abschaltfunktionen, die etwa auf zu niedrige Temperaturen der Flüssigkeit oder das Fehlen der Lüfter reagieren, sind durch Änderungen an der Software oder Hardware anzupassen oder zu deaktivieren.

Fürs Eintauchen geeignet sind alle Halbleiter- und versiegelte Heliumlaufwerke. Zwar bildet auch das Innere anderer Festplatten eine Art Reinraum, der vor allem gegen das Eindringen kleinster Staubpartikel geschützt ist, doch dringt durch den nach dem Abschalten auftretenden Druckabfall Flüssigkeit ein und zerstört sie. In einem zweiphasigen System lassen sich unversiegelte Festplatten im oberen, gasgefluteten Bereich betreiben. Vorteilhaft wirkt sich die Flüssigkeit auf die von schnell drehenden Festplatten ausgehenden Schallwellen aus.



Quelle: ServeTheHome.com

Von oben in den Pod eingetaucht und am Rahmen fixiert werden die Systeme bei der Immersionskühlung. Etwaige mit eingetauchte Lüfter haben dann aber keine Funktion mehr (Abb. 9).

Bei Kupferverbindungen sind in der Regel keine Auswirkungen auf die Signalübertragung durch Leitungsverluste, Impedanz oder Übersprechen zu erwarten. Allerdings können Steckverbinder aufgrund der erhöhten kapazitiven und der verringerten induktiven Impedanzfehlanpassung ausfallen, wenn sie in die Flüssigkeit getaucht werden. Bei Kabeln für hochperformante Verbindungen wie schnelles Ethernet oder PCIe können kleine Änderungen große Auswirkungen auf die Performance haben, wenn etwa Flüssigkeit in die Kabelummantelung gelangt. Deshalb sollten alle Kabel auf ihre langfristige Zuverlässigkeit getestet werden.

Taucht man die Stecker von Lichtwellenleitern in eine dielektrische Flüssigkeit, kann sich diese an die Stelle der Luft zwischen den Ferrulen setzen. Das verändert den Brechungsindex an den optischen Schnittstellen und kann zu Signalreflexionsverlusten führen. Vermeiden lässt sich das durch das Verwenden von Transceivern, wie sie bei SFPs oder QSFPs üblich sind. Beispielsweise kann man die Kabel an den Transceivern befestigen oder zu Dichtmitteln greifen, die das Eindringen von Flüssigkeit in den Luftspalt verhindern. Alternativ kann man Steckverbinder mit Silizium-Photonik verwenden, die keinen Luftspalt haben, oder Port-Extender, um optische Ports aus der Fluidumgebung herauszuverlagern.

Grundlegende Änderungen des mechanischen Designs werden dann notwendig, wenn man Server für die Immersionskühlung optimieren will. Neben der Zugänglichkeit von Ports und Austauschkomponenten stünden vor allem neue Gehäuse- und Rahmenkonstruktionen sowie eine vertikale Positionierung der CPUs auf der To-do-Liste. Insbesondere in einphasigen Systemen ist die Anordnung der Komponenten in der Strömungsrichtung zu optimieren, sodass etwa Hotspots wie CPUs und GPUs stromabwärts platziert sind. Auch in der Firmware gäbe es Anpassungsbedarf, etwa über einen Immersion-Modus.

Zudem stellt die Tauchkühlung Anforderungen an das Gebäude und die Gebäudetechnik, zu denen der Anschluss an die Wärmeübertrager des RZs, die Stromversorgung, die Raumbelüftung, die Statik und die Zugänge zählen, und setzt ein Flüssigkeitsmanagement voraus. Für den sicheren Umgang mit der Tauchkühlung sind die üblichen Prozesse im RZ an die Technik anzupassen. Dazu gehören Maßnahmen zur Eindämmung von Flüssigkeiten, also zur Druckentlastung, zum Vermeiden und Eindämmen von Lecks und zum Umgang von Verschüttungen. Das beinhaltet auch Doppelhüllen- oder Leckwannen, ein Monitoring, eine vollständige Flüssigkeits-, Gesundheits- und Sicherheitsdokumentation und Schulungen des Personals über die Eigenschaften der dielektrischen Flüssigkeit und den Umgang mit Verschüttungen. *Susanne Nolte*

timierung der Systeme geschuldet. Derzeit sind alle Server – von Supercomputern abgesehen – auf Luftkühlung optimiert. Das spiegelt sich in der Anordnung der einzelnen Komponenten auf den Serverboards wider, in der Größe der Servergehäuse und den eingesetzten Steckverbindungen.

Es gibt auch keine Normen oder Industriestandards, die die unterschiedlichen Kühltechniken harmonisieren, sodass die Server unterschiedlicher Hersteller mit dem jeweiligen System funktionieren. Bei der Tauchkühlung fällt das weniger ins Gewicht, bei der Plattenkühlung aber spielt das Serverdesign eine entscheidende Rolle. Auch sind die Anschlüsse an die Wasserverteilung im Rack nicht standardisiert, jeder Anbieter verwendet seine eigenen, die nicht zueinander kompatibel sind. Hier sei kurz auf das Open Compute Project verwiesen, das gerade an einer Standardisierung der beiden Kühlungstechniken arbeitet.

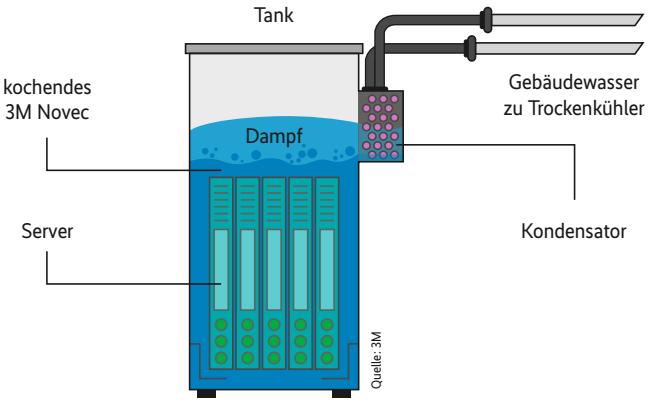
Darüber hinaus sind Bestandsrechenzentren nicht auf die Anforderungen der neuen Kühltechnik vorbereitet. Das gilt vor allem für die Leistungsdichten und die vorhandene Gebäudekühltechnik. Ohne Umbaumaßnahmen ist der Einsatz einer Flüssigkeitskühlung dort nicht effizient möglich.

Auch ist sie nicht für jede IT-Hardware sinnvoll: Bei Speichersystemen mit niedrigen thermischen Lasten und zentralen Netzwerkschränken besteht derzeit keine Notwendigkeit für Flüssigkeitskühlung. Technisch ist es aber ohne Weiteres möglich, diese IT-Hardware in die Flüssigkeitskühlung zu integrieren. Sinnvoll sind beide Kühltechniken bei Hardware mit hoher Leistungsdichte einsetzbar, etwa dort, wo hohe Rechenleistung gefragt ist. Gerade für den Einsatz beim Edge-Computing ist die Immersionskühlung prädestiniert, da sie es erlaubt, auf maschinelle Kälteerzeugung komplett zu verzichten.

Fazit

Der Wechsel von Luft auf Wasser als effizienteres Kühlmedium für Standardserver außerhalb des HPC-Bereiches hat gerade begonnen. Nur mit diesem Wechsel ist es möglich, die zu erwartende steigende Leistungsdichte im Rechenzentrum zu handhaben. Welche der beiden Techniken, Flüssigkeitsplatten- oder Immersionskühlung, die geeignetere für den jeweiligen Einsatzzweck ist, ist lässt sich nicht pauschal sagen. Bei der Plattenkühlung ist die Umrüstung von Bestandsrechenzentren einfacher, Racks und Infrastruktur müssen nur marginal angepasst werden. Mit ihrer höheren Effizienz und den weiteren Vorteilen ist die Immersionskühlung bei Servern mit hoher Leistungsdichte zu bevorzugen.

Komplett auf die Luftkühlung wird sich voraussichtlich nur in Edge-Rechenzentren verzichten lassen. In den anderen Bereichen werden die verschiedenen Kühltechniken wohl koexistieren oder Zwitter wie Schrank- und Reihenkühlungen Einzug halten (siehe Artikel „Nah dran“ ab Seite 134). (sun@ix.de)



Bei der Zweiphasenimmersionskühlung verdampft die Kühlflüssigkeit und kondensiert an der Kühlspirale (Abb. 10).

Quellen

OCP-Empfehlungen: ix.de/zf9t

Fridtjof Chwoyka

ist Senior Data Center Consultant bei der Data Center Excellence GmbH mit dem Schwerpunkt EN 50600, Nachhaltigkeit und Kühltechniken sowie Auditor für den Blauen Engel für Rechenzentren.

Abregelungen und Abwärme: Ungenutzte Energie nutzen

Aufgefangen

Hubert Sieverding

Auch Energie lässt sich recyceln, gerade da, wo sie im großen Maßstab von Strom in Wärme umgewandelt wird: im Rechenzentrum. Andere Projekte nutzen den abgeregelten Strom oder Kühlverfahren ohne Wasser.

Server sind wenig effektive Elektroheizungen, denn gemäß Energieerhaltungssatz wandeln sie den eingebrachten Strom fast vollständig in Wärme um. Nicht erst seit die Strompreise durch die Decke schießen, machen sich RZ-Betreiber Gedanken, wie sie diese Abwärme sinnvoll nutzen können. Einige Projekte stellt dieser Artikel vor. Doch auch wenn eine Nachnutzung der Abwärme gelingt, bleibt doch die Menge des verbrauchten Stroms erschreckend hoch. Das wahre Einsparpotenzial liegt im Code und im Wirkungsgrad der Komponenten.

Anders als früher suchen Rechenzentren nicht mehr die Nähe der Nutzer. Wie jeder Server hat auch ein Datacenter drei Schnittstellen: Stromversorgung, Netz und Abwärme. Rechenzentren entstehen da, wo zumindest eine davon möglichst optimal ist. Sie finden beispielsweise in nordischen Ländern gute Bedingungen: Der Strom kommt aus nachhaltigen Quellen und die Abwärme lässt sich mit wenig Aufwand in die – kühle – Umgebung abgeben. Frankfurt am Main wiederum ist mit Europas größtem Internetknoten DE-CIX ein beliebter Standort, allerdings auch eine der wärmsten Städte Deutschlands.

Windcores

Beginnen soll die Rundreise an der Quelle grünen Stroms, in einer Windkraftanlage (WKA) von Westfalenwind in der Nähe von Paderborn. Inzwischen erreichen die Windräder eine Nabenhöhe von mehr als 160 Metern und an der Basis einen Durchmesser von über 10 Metern. Die Errichtung des Turms ist nur der offensichtliche Teil beim Bau einer WKA. Fast noch umfangreicher ist der Anschluss ans Stromnetz, in der Regel über ein Erdkabel, das oft kilometerweit durch Feld und Wiese zu verlegen ist. Ist die Erde schon mal offen, lässt sich einfach ein Glasfaserstrang mit verbuddeln.

Entstanden ist daraus die Idee, in den Turm ein RZ einzubauen, das direkt von der Stromquelle profitiert: Herausgekommen ist die patentierte Technik mit Namen Windcores. Laut Westfalenwind betrug der Anteil der abgeregelten Energie aus erneuerbaren Quellen im Jahr 2019 5,4 TWh von 244 TWh, davon entfielen 96 Prozent auf die Windkraft. Zum Vergleich: Alle deutschen RZs genehmigten sich im Jahr 2020 etwa 16 TWh, das heißt, mit der regenerativen Energie, die für die Tonne produziert wurde, hätte sich ein Drittel der RZs betreiben lassen.

Abgeregelt werden WKA immer dann, wenn ein Überangebot an Strom vorhanden ist. Die Gründe sind vielfältig: Es ist Wochenende, der Wind bläst heftig oder die Sonne speist die Fotovoltaikanlagen im Land, die ja ebenfalls abgeregelt werden, weil das Kohlekraftwerk unter Dampf steht. Nutzbar wäre die Überschussenergie zur Wasserstofferzeugung – ja, wenn es entsprechende Anlagen gäbe. Leider ein Konjunktiv. Auch könnte man die Energie speichern. Ein weiterer Konjunktiv.

iX-TRACT

- Den Strom, den elektrische Geräte aufnehmen, geben sie als Wärme wieder ab. Das gilt auch für Server. Dort, wo große Kühlanlagen die Wärme abführen, lässt sie sich auch zum Heizen nutzen.
- Bei der Abwärmenutzung sind Flüssigkühlungen im Vorteil, da sie höhere Wassertemperaturen erzeugen.
- Wenn dem öffentlichen Stromnetz die Überlastung droht, wird regenerativer Strom direkt an der Quelle abgeregelt. Ähnlich wie Haushalte mit Fotovoltaikanlage können Rechenzentren im Windradturm ihn direkt an der Quelle nutzen und damit das Netz entlasten.
- Den Einsatz giftiger Kühlflüssigkeiten vermeidet Kyoto Cooling, auch auf Wasser kann es verzichten.

Statt die Propeller auf Durchzug zu schalten, laufen dank Windcores die Server zum Nulltarif und auch sonst ist – solange es windet – der Weg von der regenerativen Energiequelle bis zum Server kurz. Die physische Sicherheit der Anlage ist dank dicker Stahlbetonwände gewährleistet. Stahlplattformen mit mehreren Ebenen schaffen Platz für bis zu 50 Racks und die Abwärme der Server verliert sich auf 150 Meter Turmhöhe (siehe Abbildung 1). Einzig die Administrierenden benötigen für Wartungsarbeiten vor Ort Führerschein und Dienstwagen.

Fernwärme

Überall dort, wo grüner Strom teuer bezahlt werden muss, ist die Abwärmenutzung per Fernwärmeeinspeisung im Gespräch. Die Geschichte der Fernwärme reicht bis in 14. Jahrhundert zurück. Damals wurden heiße Quellen angezapft. Im großen Stil eingeführt wurde die erste Generation um 1880 in den USA. Sie nutzte Wasserdampf und entsprechend gefährlich war ihr Betrieb. In der zweiten Generation strömte Mitte des 20. Jahrhunderts kochend heißes Wasser von den Kraftwerken zu den Haushalten. Die heutige Generation verwendet üblicherweise 70 °C warmes Wasser im Vorlauf und 40 °C im Rücklauf.

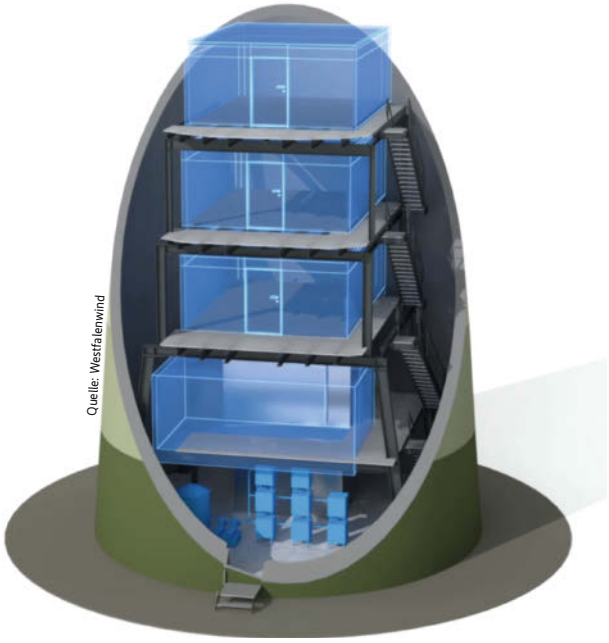
Ein Fernwärmenetz besteht aus mindestens einer Wärmequelle, meist vielen Wärmesenken und einem Verteilnetz in Form isolierter Rohre. Das Wasser fließt im Kreis, warm zur Senke und abgekühlt zur Quelle zurück, wo es erneut erwärmt wird. Heute wird die Wärme vorwiegend in Gaskraftwerken oder Müllverbrennungsanlagen erzeugt. Auch Kohlekraftwerke finden sich noch in vielen Städten. Sie verheizen Kohle, erzeugen Dampf zum Antrieb von Generatoren und nutzen das Fernwärmenetz quasi als Kühlung.

Der Wärmebedarf ist dabei von drei Faktoren abhängig: vom Heizbedarf, vom Warmwasserbedarf und von den Leitungsverlusten. Die Menge Wasser, die durch ein Rohr passt, und damit die transportierbare Energiemenge bei gleicher Temperatur, ist beschränkt. Jedoch wird im Winter mehr Energie benötigt als an heißen Sommertagen, an denen nur der Warmwasserbedarf gedeckt werden muss. Aus diesem Grund wird die Vorlauftemperatur bei niedrigen Außentemperaturen erhöht, womit die transportierbare Wärmemenge steigt.

Auch wenn die verwendeten Leitungen sehr gut isoliert sind, gibt es dennoch geringe Verluste, die wiederum von der Differenztemperatur zur Umgebung abhängen. Als vierte Generation der Fernwärmenetze gilt das Ultra-Niedrigtemperatur-Fernwärmenetz. Es arbeitet mit deutlich niedrigeren Vorlauftemperaturen von unter 50 °C und wird von vielen Quellen gespeist, darunter Geothermie, Solarthermie und RZ-Abwärme.

Eurotheum

Ein Vorzeigeprojekt der RZ-Abwärmenutzung ist das Eurotheum an der Neuen Mainzer Straße in Frankfurt. Das Hochhaus mit 31 Stockwerken wurde 1999 fertiggestellt und ist 110 Meter hoch. Es umfasst Büros, ein Hotel und Apartments auf den oberen sieben Etagen. Der frühere Hauptmieter, die EZB, hinterließ mit ihrem Umzug im Jahr 2015 zwei Rechenzentren im Keller und in der 7. Etage, gemäß damaligem Standard luftgekühlt. Das Gebäude selbst ist an das Fernwärmenetz der Mainova angeschlossen. 2017/18 führte der Dresdener Spezialist Cloud&Heat die Modernisierung hin zu einem redundanten, hochsicheren und hochverfügbaren Rechenzentrum mit dem Ziel der Abwärmenutzung durch. Da es sich nicht um einen RZ-Neubau handelt, existieren für den Energieverbrauch Vergleichszahlen zum vorherigen Ausbau (siehe ix.de/z95h).



Westfalenwind baut Rechenzentren in den Turm von Windkraftanlagen. Bis zu 50 Racks finden auf mehreren Etagen Platz (Abb. 1).

Die genutzte Abwärme stammt aus einer Direktkühlung der wichtigsten IT-Komponenten. Hierbei wird Heißwasser mit einer Temperatur von über 60 °C erzeugt, was der typischen Vorlauftemperatur der Mainova-Fernwärme (70 °C) sehr nahekommt. In Folge reduziert sich der Bedarf an Lüfterleistung und Rückkühlung, was zusammen mit der eingesparten Fernwärme auch zur Reduzierung der Kosten führt.

Bei marktüblichen Servern beträgt der Energiekonsum der Lüfter bis zu 25 Prozent der gesamten Leistungsaufnahme. Nahezu alle Serverhersteller bieten wassergekühlte Systeme an, typischerweise als Blade. Als Beispiel sei die Lenovo SD650 genannt. Cloud&Heat nimmt hingegen auch günstige Standardhardware und baut diese auf Wasserkühlung um. Dabei werden die Kühlkörper der CPUs, GPUs und SSDs durch Wärmetauscher ersetzt und die Wasserschläuche am DRAM und anderen sehr warmen Komponenten vorbeigeführt.

Jedes Rack enthält auf den untersten Einschüben redundante Wärmetauscher – fachlich korrekt Wärmeübertrager. Als typische Rücklauftemperatur nach Kühlung weist Cloud&Heat 50 °C aus. Alle Racks sind per Leitungssystem im Doppelboden in zwei redundante Kühlkreisläufe integriert, die die Wärmeübertrager zur Gebäudeheizung und die Rückkühler für die Sommermonate umfassen. Die zur Wasserzirkulation benötigte Energie ist wesentlich geringer als die Lüfterleistung in konventionellen Systemen.

Auch die Restwärme nutzen

Die per Luftkühlung abgeführte Restwärme von etwa 30 Prozent wurde in der ersten Phase nicht genutzt, sondern an die Umwelt abgegeben. Dazu nutzte Cloud&Heat bis zu einer Außentemperatur von 3 °C die Außenluft. Dies funktioniert im warmen Frankfurt aber nur an wenigen Tagen im Jahr. Im Fall des Eurotheums konnten so 17 Prozent des Restkühlbedarfs gedeckt werden. Bei höheren Temperaturen wurden die als Stromfresser bekannten Kälteanlagen hinzugeschaltet, die dank Wasserkühlung der Rechner deutlich weniger zu tun bekommen (siehe Kasten „Kälteanlagen“).

Abbildung 3 zeigt die Cloud&Heat-Kühlung des Eurotheums im Vergleich zur klassischen Luftkühlung. Zwei Drittel der ein-

gebrachten Energie lassen sich zum Heizen und zur Warmwasserbereitung nutzen. Von 1370 MWh/Jahr sinkt der Bedarf an Hilfsstrom für Lüfter und Kühlung auf 420 MWh/Jahr. Zudem speist das System 933 MWh/Jahr an Fernwärme ein. Die Kostenersparnis beträgt beim Strom- und Fernwärmekostenniveau von 2018 abhängig vom Anteil der freien Kühlung bis zu 255 000 Euro/Jahr und dürfte inzwischen deutlich höher sein. Laut Cloud&Heat reduzieren sich die CO_2-Emissionen um 710 Tonnen/Jahr. In einem neueren Whitepaper hat Cloud&Heat die Einsparpotenziale konkretisiert (siehe ix.de/z95h). Inzwischen wird auch die Abluft der Abwärmenutzung zugeführt (siehe Abbildung 4). Durch Verbesserung der Direktkühlung konnte der wassergekühlte Anteil auf 83 Prozent gesteigert werden. Die restlichen 17 Prozent luftgekühlter Wärme können per Wärmepumpe auf die Vorlauftemperatur von 63 °C angehoben werden, statt sie über einen Rückkühler an die Umgebung abzugeben.

Kälteanlagen

Egal, ob Kühlschrank, Wärmepumpe oder Kaltwasserstrang, das Kernelement einer jeden Kälteanlage ist ein Kältemittelkreislauf und besteht aus sechs Komponenten. Ein Verdichter komprimiert ein gasförmiges Kältemittel, das über einen Kondensator Wärme abgibt. Diese Komponente wird auch Verflüssiger genannt, weil das gasförmige Kältemittel beim Abkühlen seinen Aggregatzustand zu flüssig wechselt. Anschließend fließt es durch eine Rohrleitung zu einem Drosselorgan. Die Expansion infolge des Druckabfalls führt zu einer Temperaturabsenkung und in einem nachgelagerten Verdampfer zur Wärmeaufnahme.

Grundsätzlich gilt der zweite Satz der Thermodynamik: Wärme fließt von selbst nur von einem Körper höherer zu einem Körper niedrigerer Temperatur. Damit also Wärme abgeführt werden kann, muss eine Temperaturdifferenz vorliegen. Rechenzentren, die luftgekühlte Komponenten einsetzen, arbeiten typischerweise mit circa 30 °C im Warmgang und 20 °C im Kaltgang. Die Temperatur des Kältemittels im Verdampfer muss also deutlich unterhalb von 20 °C liegen und nach der Kompression die Außentemperatur deutlich übersteigen, damit die Wärme im Verflüssiger abgeführt werden kann.

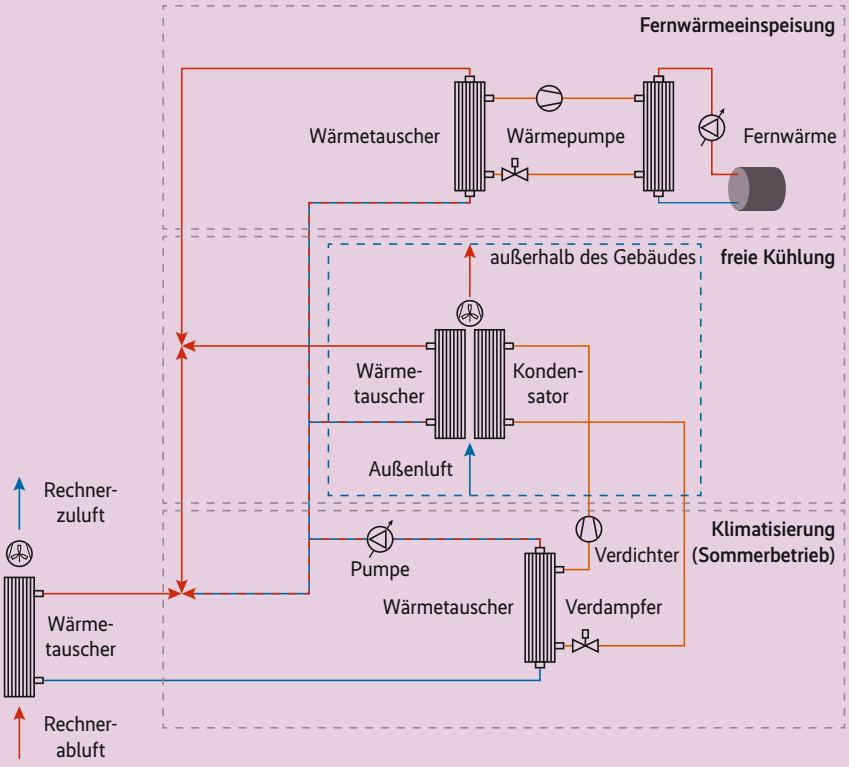
Häufig wird als Alternative zu Kälteanlagen die freie Kühlung genannt. Bei der direkten freien Kühlung wird Außenluft direkt in den Kaltgang geleitet. Dies ist nicht unproblematisch, denn Luft ist mal zu kalt, mal zu warm, mal zu feucht oder zu trocken und vor allem auch mit Schadstoffen belastet. Bei der indirekten freien Kühlung unterscheidet man mehrere Verfahren. Die einstufige Freiluftkühlung vermeidet die oben genannten Nachteile durch Einsatz eines Luft-Luft-Wärmetauschers. Neben dem Rotationsverfahren, dem unten beschriebenen Kyoto Cooling, gibt es Plattenwärmetauscher. Auch hier muss die Wärmeabfuhr mit Ventilatoren unterstützt werden.

Bei der zweistufigen indirekten freien Kühlung wird die Rechnerabwärme mittels Wärmetauscher auf einen Wasserkreislauf übertragen und funktioniert wie die Kühlung eines Verbrennungsmotors im Fahrzeug. Die Abkühlung übernimmt ein Trockenkühler außerhalb des Gebäudes. Dort findet erneut eine Wärmeübertragung statt, diesmal vom glykolhaltigen Wasser auf die kältere Außenluft. Die Wärmeabfuhr wird durch Ventilatoren verbessert. Die zweistufige indirekte freie Kühlung benötigt zusätzlich Energie für die Wasserpumpen.

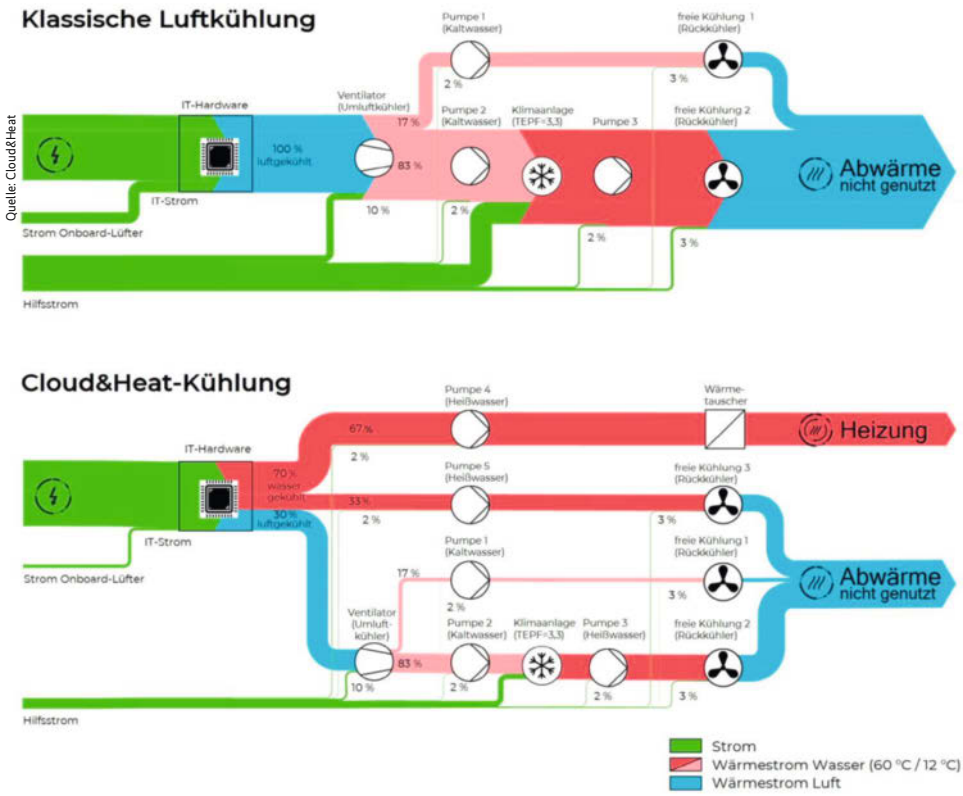
Ist die Temperaturdifferenz zwischen der Flüssigkeit und der Außenluft zu gering, wird eine Kälteanlage hinzugeschaltet. Die Wärme wird dabei per Wärmetauscher einem Verdampfer zugeführt. Der Verdichtung durch Kompressoren folgt die Kondensation. Die Kühler besitzen dafür zwei Kühlmittelkreisläufe. Durch einen fließt an kalten Tagen das von der Abluft erwärmte Wasser direkt. An heißen Tagen wirkt ein zweiter Kreislauf als Verdampfer. Die Kompression verbraucht nicht nur weiteren Strom, sondern erwärmt das Medium zusätzlich.

Ein Mischbetrieb aus freier Kühlung und Klimatisierung ist möglich. Oftmals teilen sich der Wärmetauscher zur freien Kühlung und der Kondensator eine Kühleinheit, die typischerweise auf dem Dach platziert wird. Die gesamte Anlage wird mehrfach redundant ausgeführt.

Für eine Abwärmenutzung muss zusätzlich eine Wärmepumpe integriert werden. Dabei handelt es sich um eine Kältemaschine, die die per Verdampfer aufgenommene Wärme an die Fernwärmeleitung abgibt. Meist wird Fernwärme an den Tagen benötigt, an deren die freie Kühlung ohne Kälteanlage ausreichen würde. Daher nimmt die Wärmepumpe statt der Trockenkühler die Wärme ab, hebt die Temperatur auf circa 70 °C und fungiert als zusätzlicher, am höchsten priorisierter Kreislauf (siehe Abbildung 2). Der Rest ist Regelungstechnik.



Zur Abwärmenutzung luftgekühlter Rechenzentren wird zusätzlich eine Wärmepumpe installiert, die die Temperatur auf circa 70 °C anhebt (oben rechts). Daneben und darunter der typische Aufbau zur RZ-Klimatisierung: Im Winterbetrieb reicht die freie Kühlung zur Senkung der Ablufttemperatur auf circa 20 °C aus (oben). Steigt die Außentemperatur an, wird eine Kälteanlage (unten rechts) hinzugeschaltet (Abb. 2).



Im Vergleich zur klassischen Luftkühlung werden im Eurotheum-Hochhaus per Direktkühlung 67 Prozent der Abwärme für Heizung und Warmwasser genutzt (Abb. 3).

Cloud&Head bietet RZ-Container zum Anschluss an das Fernwärmenetz an und möchte damit Stadtwerke motivieren, in den RZ-Markt einzusteigen. Noch effizienter wäre die Abwärmenutzung, wenn die CPU-Hersteller den zulässigen Temperaturbereich anheben würden. Dann wäre eine Einspeisung ins Fernwärmenetz ohne zusätzliche Wärmepumpe möglich.

Westville

Das bisher größte Projekt zur Abwärmenutzung eines luftgekühlten Datacenters entsteht zurzeit im Westen des Frankfurter Gallusviertels auf dem ehemaligen Avaya-Gelände. Geplant ist ein Wohnquartier mit 1300 Mietwohnungen, drei Kitas und Nahversorgern, angeschlossen ans Frankfurter Fernwärmenetz der Mainova. Auf der anderen Seite der Kleyerstraße befindet sich das Rechenzentrum der Telehouse und unter der Straße liegen mehrere ungenutzte, aber ausreichend dimensionierte Leerrohre. Zukünftig wird die 30 °C warme Serverabluft mindestens 60 Prozent des Wärmebedarfs Westvilles von insgesamt 4000 MWh/Jahr beisteuern.

Telehouse verwendet in seinem Rechenzentrum luftgekühlte Standardhardware. Per Warmgangeinhausung wird die auf 30 °C erwärmte Rechnerabluft per Wärmeübertrager auf unter 24 °C abgekühlt und zu den Servern zurückgeführt. Je nach Jahreszeit und Bedarf bildet zukünftig ein Teil der so abgeführten Wärme einen Kreislauf mit der 500 Meter entfernten Heizzentrale des Stadtquartiers, quasi ein Minifernwärmenetz niedrigster Temperatur (siehe Abbildung 5).

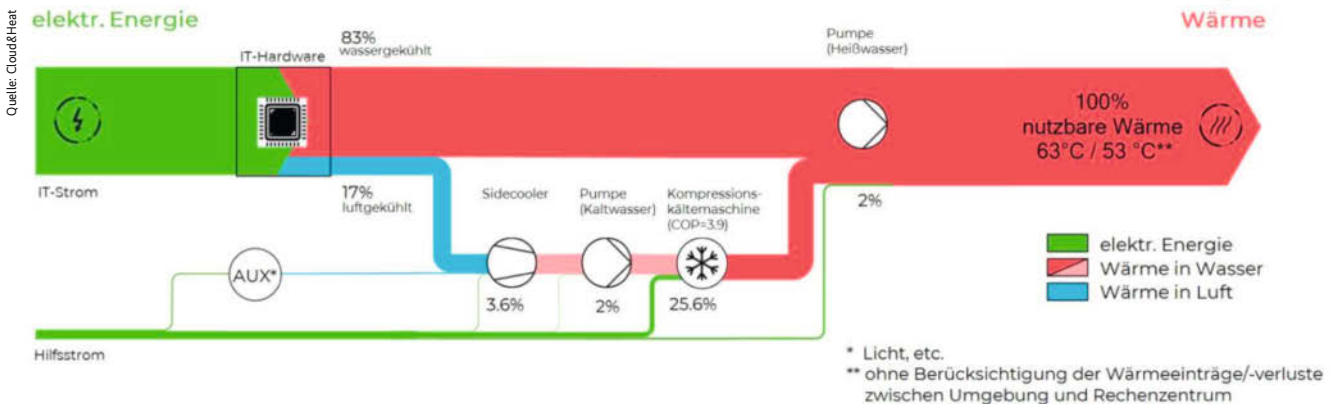
Dort heben zwei Großwärmepumpen mit einer thermischen Leistung von je 320 kW und einem COP (Coefficient of Performance) von 4,5 die Temperatur auf das Niveau der Fernwärme von 70 °C an. Der COP gibt das Verhältnis von Wärmeleistung zur erforderlichen Antriebsenergie an. Im Fall von Westville hat also jede Pumpe eine Leistungsaufnahme von 70 kW.

Telehouse stellt die Energie kostenfrei zur Verfügung. Laut Manuel Gerdsmeyer von Mainova beträgt der Strombedarf aller Rechenzentren in Frankfurt zurzeit circa 1 TWh/Jahr, Tendenz steigend. Hochgerechnet bis 2030 reichte die Abwärme aus, um den Wärmebedarf aller Privathaushalte und Bürogebäude der Stadt zu bedienen, allerdings ist die zurzeit verwendete Fernwärmegeneration dazu ungeeignet.

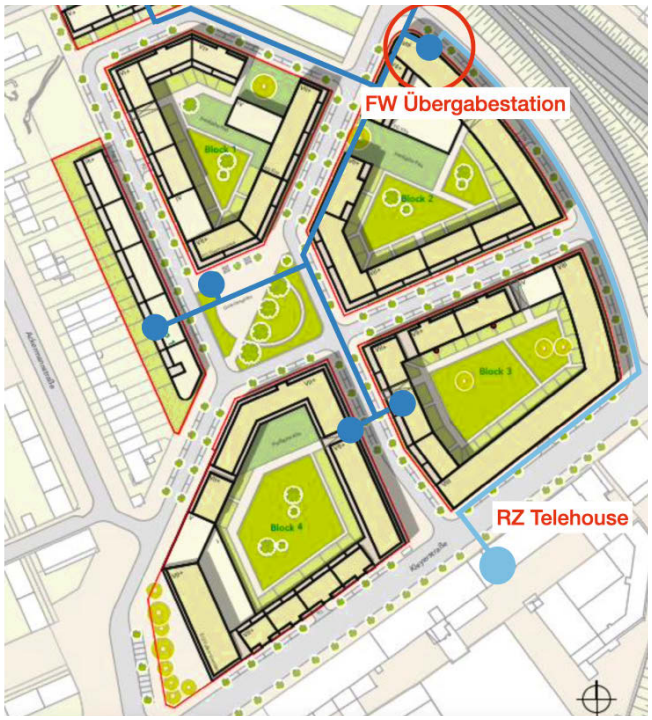
Ein ähnlich geartetes Projekt befindet sich in Hattersheim, einer Ortschaft in der Nähe des Frankfurter Flughafens, im Aufbau. Dort entsteht ein Neubaugebiet, das vom Rechenzentrum Frankfurt4 der NTT Global Data Centers EMEA GmbH mit Wärme versorgt werden wird. Anders als in Westville ist hier der Weg zur Wärmepumpe kurz. Die Fernwärmeübergabestation befindet sich neben der Halle auf dem Rechenzentrumsgelände.

Kyoto Cooling

Wie die vorgestellten Beispiele zeigen, ist die Temperaturdifferenz bei IT-Luftkühlung relativ gering und der Aufwand zur Abwärmenutzung hoch. Schließlich ziehen die nachgeschalteten Wärmepumpen erneut Strom. Daher setzten einige Rechenzentren zur Verbesserung der Power Usage Effectiveness (PUE) auf freie Kühlung. Die Idee allerdings, wie in der Tiermast üblich der Halle per Ventilator Luft zuzuführen, stirbt häufig bereits beim ersten Brandschutzgutachten. Auch die Luft-Luft-Kühlung per



In der zweiten Version hebt Cloud&Heat auch den Restanteil der Abwärme aus der Luftkühlung per Wärmepumpe auf 63 °C an (Abb. 4).



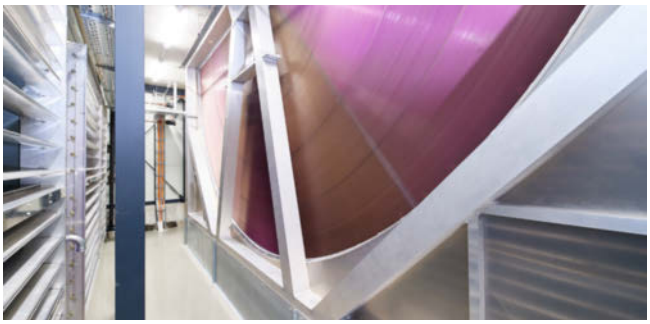
Quelle: Mainova

Im Plan zur Fernwärme- und RZ-Abwärmeversorgung des geplanten Stadtquartiers Westville in Frankfurt sind die Fernwärmeleitungen dunkelblau, die Niedertemperaturleitung vom RZ Telehouse zur etwa 500 Meter entfernten Übergabestation ist hellblau dargestellt (Abb. 5).

Kyoto-Rad ist bei den Hütern der Vorschriften nicht unumstritten. Das laut Noris Network in Nürnberg modernste Rechenzentrum Europas nutzt Kyoto Cooling als energieeffiziente Methode zur freien Kühlung (siehe Abbildung 6). Eine Abwärmenutzung wie in den anderen vorgestellten Projekten findet nicht statt.

Der Grundgedanke des Kyoto-Rads ist denkbar einfach: Eine sechs Meter im Durchmesser fassende Radkonstruktion aus feinsten Alurohren dreht sich mit wenigen Umdrehungen pro Minuten in einem Gehäuse mit vier Kammern, jeweils zwei übereinander und zwei nebeneinander. Zwei Kammern unter- und oberhalb einer Radhälfte teilen mit dem Rechenzentrum jeweils Warm- und Kaltluftbereich. Ventilatoren ziehen die warme Abluft aus dem Deckenbereich der Halle durch die Speichen des Rads, die sich im Laufe einer halben Umdrehung erhitzen und so die Luft kühlen. Diese strömt durch den Kaltgang, zum Beispiel den Doppelboden, erneut zu den Geräten.

Mit Durchschreiten einer Schleuse wird in der anderen Hälfte des Raums das Rad mit Außenluft abgekühlt. Hier drücken Ventilatoren die Außenluft durch die Aluspeichen und kühlen so das Metall (siehe Abbildung 7). Die Fläche des riesigen Rades



Quelle: Noris Network AG

Das Kyoto-Rad überträgt die Wärme zwischen zwei warmen und zwei kalten Kammern (Abb. 6).

und seine sehr große Oberfläche sorgen dafür, dass es nicht zum thermischen Kurzschluss kommt. Nur an sehr heißen Tagen helfen Kühlaggregate nach. Die ganze Anlage ist wasserfrei und für bestehende Räumlichkeiten durch Anflanschen eines Containers nachrüstbar. Die Reduktion des Sauerstoffgehaltes durch bedarfsgerechtes Einbringen von Stickstoff an der Schleuse kann im Brandfall ein Anfachen des Feuers wirksam unterbinden. Minimale Außenluftleckagen lassen sich an der Schleuse aber trotz der sehr langsamen Umdrehung des Rads nicht vollständig vermeiden.

Fazit

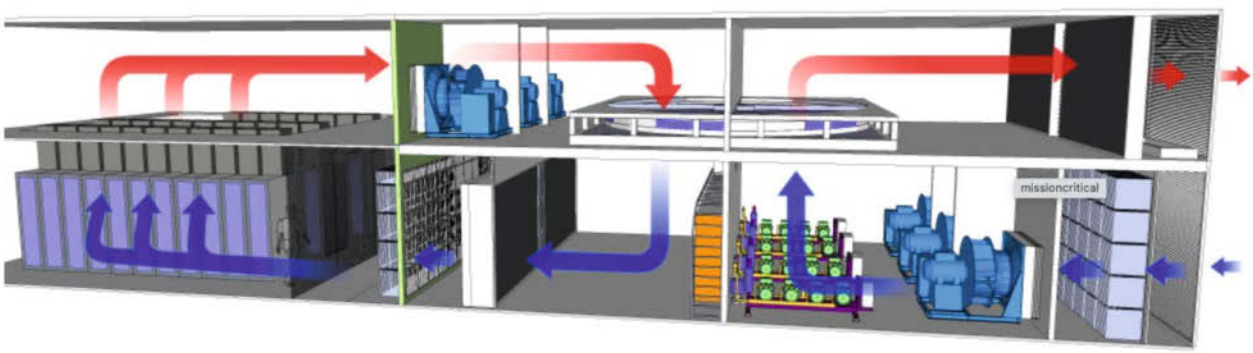
Besonders die Wasserkühlung eignet sich aufgrund der hohen Temperaturen zur Einspeisung in bestehende Fernwärmenetze, was die Gesamtenergiebilanz erheblich verbessert. Die Abwärmenutzung luftgekühlter Server zur Fernwärmeeinspeisung ist möglich, erfordert jedoch zusätzliche Großwärmepumpen mit hohem Energiebedarf. Ideal wäre die Integration von Rechenzentren in Ultra-Niedrigtemperatur-Fernwärmenetze. Diese sind allerdings erst im Aufbau und eignen sich für Neubaugebiete mit Niedrigtemperaturheizungen. Dennoch ist die Abwärmenutzung nur ein Glied der Kreislaufwirtschaft und schützt nicht vor der vordringlichen Pflicht, den Strom zuerst nachhaltig zu produzieren und ihn möglichst sparsam und effizient einzusetzen. (sun@ix.de)

Quellen

Cloud&Heat-Whitepapers siehe ix.de/z95h

Hubert Sieverding

arbeitet nach langjähriger Tätigkeit in der Automobilbranche als freier Autor.



Quelle: KyotoCooling

Völlig ohne Wasser arbeitet Kyoto Cooling, ermöglicht jedoch auch keine Abwärmenutzung (Abb. 7).

Literatur zu nachhaltiger IT

Fit im Kopf

Jürgen Seeger

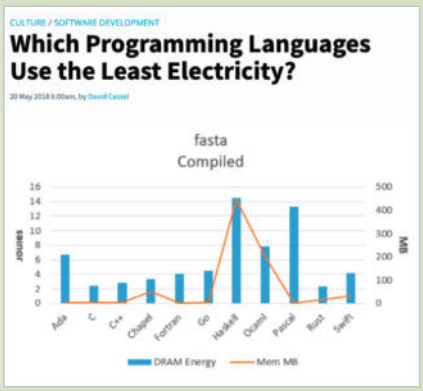
Mehr Nachhaltigkeit der IT wird nicht nur durch gut angepasste Hardware und wohlüberlegte RZ-Infrastruktur erreicht oder durch „grünen" Strom. Eine ebenfalls entscheidende Rolle spielt die Software. Beide Aspekte deckt der Buchmarkt ab.

Leif Geiger et al.
Ressourceneffiziente Programmierung
Bitkom 2021
https://www.bitkom.org/sites/default/files/2021-03/210329_lf_ressourceneffiziente-programmierung.pdf

Es muss nicht unbedingt gleich ein richtiges Buch sein. Für Eilige und nicht so Lesefreudige hat der Bitkom die wesentlichen Aspekte in einem 2021 erschienenen Papier zusammengefasst. Acht Autoren handeln unter dem Titel „Ressourceneffiziente Programmierung" ab, wie – so die Unterzeile – „Ressourcenschonung, Langlebigkeit und Nachhaltigkeit in der Softwareentwicklung berücksichtigt werden können". Von der Architektur über Implementierung und Benchmarking bis zu gesellschaftspolitischen Fragen werden alle relevanten Facetten der Problematik behandelt.

David Cassel
Which Programming Languages Use the Least Electricity
TheNewStack 2018
https://thenewstack.io/which-programming-languages-use-the-least-electricity/

Einen Teilaspekt greift ein anderes Papier auf, das auf The New Stack veröffentlicht wurde und die Ergebnisse einer Untersuchung zur Frage zusammenfasst, welche Programmiersprachen den wenigsten Strom verbrauchen. C-Programmierer haben natürlich schon immer geahnt, dass ihre Sprache zur Rettung der Welt beiträgt, dass aber auch Rust so gut abschneidet, überrascht dann doch ein wenig.

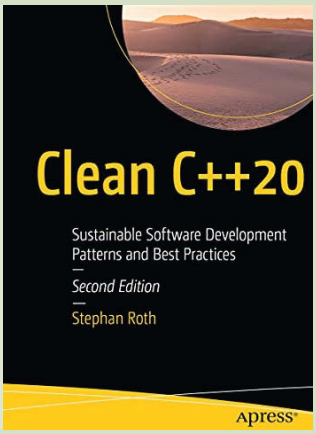


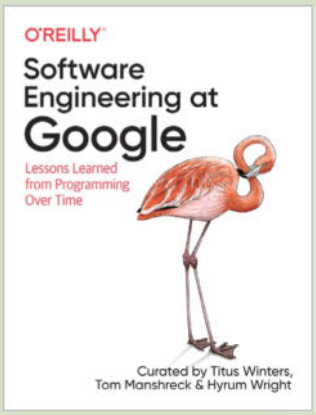
David Thomas, Andrew Hunt
Der pragmatische Programmierer: Ihr Weg zur Meisterschaft
Carl Hanser Verlag 2021 (2., überarbeitete Auflage)
304 Seiten; gebundenes Buch 39,99 €; E-Book 31,99 €

Vor gut 20 Jahren veröffentlichten David Thomas und Andrew Hunt die englische Originalversion ihres Standardwerks „Der pragmatische Programmierer", 2021 ist die 2. Auflage erschienen. Zwar kommt der Begriff „Nachhaltigkeit" dort nicht vor, und wenn es um Ressourcen geht, reden die Autoren nur von der sauberen Freigabe von Speicheranforderungen. Aber weil der gekonnte Umgang mit Algorithmen und Datenstrukturen eine zwingende Voraussetzung für ressourcenschonende Programmierung ist, gehört dieses Buch zur Pflichtlektüre in Sachen nachhaltiger Softwareentwicklung.

Stephan Roth
Clean C++20: Sustainable Software Development Patterns and Best Practices
Apress 2021 (2. Auflage)
508 Seiten; Taschenbuch 34,99 €; E-Book 24,49 €

Bei so viel Nachhaltigkeit drängt sich dem Insider schon längst der Begriff „Clean Code" auf, geprägt von Robert Cecil Martin 2008 in seinem gleichnamigen Buch. Mittlerweile ist Martin bei „Clean Agile" angekommen, und das Paradigma des „sauberen" Codes wurde weithin akzeptiert und ausgebaut. Auf die aktuelle C++-Version bezogen hat das Stephan Roth in „Clean C++20", einer Aktualisierung seines 2017 erschienenen Buchs „Clean C++".



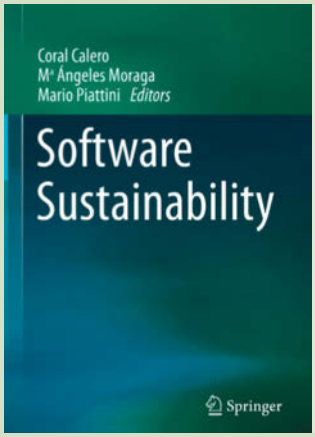


Titus Winters, Tom Manshreck, Hyrum K. Wright
Software Engineering at Google: Lessons Learned from Programming Over Time
O'Reilly UK Ltd. 2020
583 Seiten; Taschenbuch 40,99 €; E-Book 30,74 €

Dass der Begriff nachhaltige Software sich nicht unbedingt auf die Optimierung des Ressourcenverbrauchs zur Laufzeit beziehen muss, erfährt man in „Software Engineering at Google". Titus Winters, Tom Manshreck und Hyrum Wright geben fundierte Einblicke in die Softwareentwicklung in einem weltweit verteilt arbeitenden Team. Sie unterscheiden streng zwischen Programmierung und Softwareengineering und sehen die Art von Software als nachhaltig an, die an veränderte Anforderungen mit minimalen Änderungen angepasst werden kann. Sozusagen Nachhaltigkeit in der Entwicklungsphase, von Praktikern für Praktiker geschrieben.

Carola Lilienthal
Langlebige Software-Architekturen: Technische Schulden analysieren, begrenzen und abbauen
dpunkt (Heise-Tochter) 2019 (3. Auflage)
316 Seiten; Taschenbuch 34,90 €; E-Book 30,99 €; Bundle 39,90 €

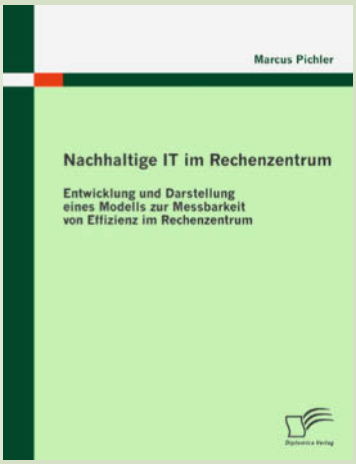
Folgt man Carola Lilienthal, ist die Nachhaltigkeit von Softwaresystemen vor allem eine Frage der Architektur. In „Sustainable Software Architecture" plädiert sie für ein sauberes, auf Langlebigkeit angelegtes Design und arbeitet unter anderem mit dem Begriff der technischen Schulden. Kurz gesagt: Schnell hingeschluderte Lösungen zahlen sich (meist) nicht aus, bei der ersten größeren Änderung wird es teuer. Die dritte Auflage der deutschen Version ist unter dem Titel „Langlebige Software-Architekturen" erhältlich.



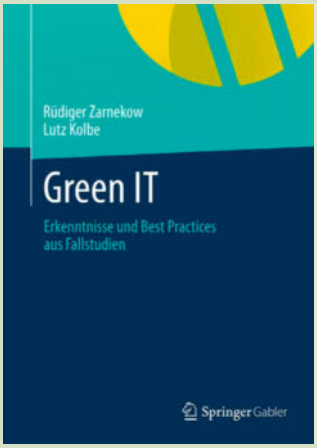
Coral Calero, Ángeles Moraga, Mario Piattini
Software Sustainability
Springer 2021
420 Seiten; gebundenes Buch 123,78 €; E-Book 86,85 €

Einen explizit akademischen Ansatz verfolgt der Sammelband „Software Sustainability". Dort haben die Herausgeber versucht, in 17 Kapiteln umfassend zu klären, was es mit der Nachhaltigkeit von Software auf sich hat, und zwar nicht nur im technischen, sondern auch im sozialen und gesellschaftlichen Sinn. Das fängt bei Architekturentscheidungen an und hört bei detaillierten Messergebnissen zur Berücksichtigung von Nachhaltigkeit nicht auf. Interessant für alle, die theoretisch ganz tief in die Thematik einsteigen wollen.

Marcus Pichler
Nachhaltige IT im Rechenzentrum: Entwicklung und Darstellung eines Modells zur Messbarkeit von Effizienz im Rechenzentrum
Diplomica Verlag 2009
160 Seiten; Taschenbuch 48 €; E-Book 23 €

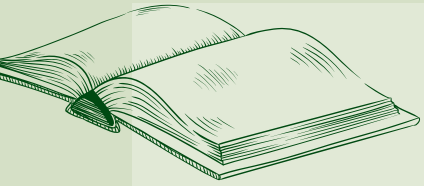


Green IT könnte man auch als Synonym betrachten für „Strom sparen im Rechenzentrum", denn es geht ans Portemonnaie, wenn das RZ nicht grün genug betrieben wird – wenn man nicht gerade wie die Bitcoin-Miner in Billigstromländer ausweicht. Literatur zum grünen IT-Betrieb gehört seit Längerem zum Portfolio der IT-Verlage, das entsprechende „For Dummies" bei Wiley erschien 2009, im selben Jahr wurde auch Marcus Pichlers Untersuchung „Nachhaltige IT im Rechenzentrum" veröffentlicht. Natürlich sind die untersuchten Bauteile heute hoffnungslos veraltet, aber der methodische Ansatz zum Messen der unterschiedlichen Hardwarekonfigurationen ist immer noch interessant.



Rüdiger Zarnekow, Lutz Kolbe
Green IT: Erkenntnisse und Best Practices aus Fallstudien
Springer Gabler 2013
192 Seiten; gebundenes Buch 59,99 €; E-Book 46,99 €

Einen auf den ersten Blick praxisnahen Ansatz verfolgt der Sammelband von Rüdiger Zarnekow und Lutz Kolbe. Sie haben 2013 in „Green IT" Erkenntnisse und Best Practices aus Fallstudien zusammengefasst, die sich in erster Linie um den IT-Betrieb drehen. Leider schränkt nicht nur das lange zurückliegende Erscheinungsjahr den Nutzwert arg ein, sondern auch die unkritische Darstellung der Anwender. Die umfasst zwar ein breites Spektrum von den Hannoverschen Verkehrsbetrieben bis zum Axel-Springer-Verlag, aber Stil und Inhalt der Fallstudien erinnern stark an die unsäglichen „Success Stories", die PR-Agenturen immer wieder bei Fachzeitschriften unterbringen möchten.



Niklas Reisinger
Green-IT-Strategien für den Mittelstand: Nachhaltige Lösungen in der IT und durch IT-Unterstützung
Diplomica Verlag 2014
100 Seiten; Taschenbuch 39,99 €; E-Book 29,99 €

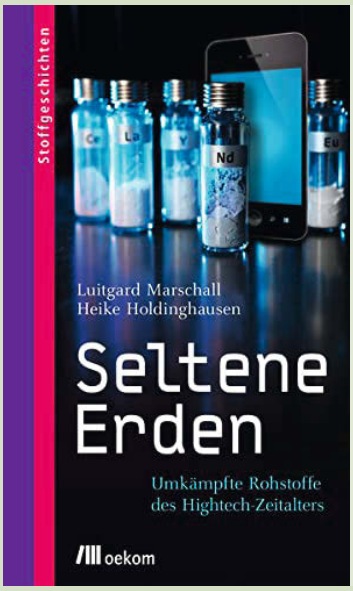
Wer keinen Tipp fürs eigene RZ braucht, sondern erst einmal nur fundiert mitreden können möchte, dem sei Niklas Reisingers 2014 erschienenes „Green-IT-Strategien für den Mittelstand" empfohlen. Reisinger betrachtet nicht nur den IT-Betrieb, sondern den ganzen Lebenszyklus von der Herstellung der Komponenten bis zu ihrer Entsorgung. Zudem erfährt man auch, inwiefern Informationstechnik zum Einsparen von Ressourcen beiträgt. Das Buch ist zwar lange vor der Coronapandemie erschienen, doch das Standardbeispiel durch Videokonferenzen obsolet gewordener Dienstreisen überzeugte auch schon 2014.

Ronald Deckert
Digitalisierung und nachhaltige Entwicklung: Vernetzt Denken, Fühlen und Handeln für unsere Zukunft
Springer Gabler 2020 (2. Auflage)
60 Seiten; Taschenbuch 14,90 €; E-Book 4,99 €

Einen weiten Bogen schlägt Ronald Deckert in der zweiten Auflage von „Digitalisierung und nachhaltige Entwicklung". Wie der Titel andeutet, geht es „ums Ganze", nicht nur um grüne IT, sondern um Nachhaltigkeit im Kontext von Digitalisierung und deren gesellschaftlichen Implikationen. Das tangiert unter anderem die Vernetzung in Natur und Gesellschaft sowie lebenslanges Lernen. Weniger wäre mehr gewesen, aber die Irgendwie-hängt-alles-mit-allem-zusammen-Fraktion wird diesen Rundumschlag wahrscheinlich goutieren.

Michael Jacob
Digitalisierung & Nachhaltigkeit: Eine unternehmerische Perspektive
Springer Vieweg 2019
101 Seiten; gebundenes Buch 54,99 €; E-Book 42,99 €

Eine „unternehmerische Perspektive" auf Digitalisierung und Nachhaltigkeit entwickelt Michael Jacob auf rund 100 Seiten. Ein anderer möglicher Untertitel wäre „Digitalisierung und Nachhaltigkeit für CEO-Dummies" gewesen, was keineswegs despektierlich gemeint ist. Denn Jacob umreißt knapp und prägnant die entsprechenden Problemfelder und gibt Hinweise darauf, wie die Firmenleitung die durchaus widersprüchlichen Anforderungen Digitalisierung und Nachhaltigkeit miteinander versöhnen kann. Im Prinzip mag ich kurze Bücher und zitiere auch gern den Goethe zugeschriebenen Satz „Entschuldige meinen langen Brief, für einen kurzen hatte ich keine Zeit". Aber 54,99 Euro für ein 100-Seiten-Buch sind, milde ausgedrückt, eine recht offensive Preisgestaltung.



Luitgard Marschall, Heike Holdinghausen
Seltene Erden: Umkämpfte Rohstoffe des Hightech-Zeitalters
Oekom 2017
192 Seiten; gebundenes Buch 24 €; E-Book 18,99 €

Ganz weit vorne im Produktionsprozess setzen Luitgard Marschall und Heike Holdinghausen mit der Auswertung der Ergebnisse einer Fraunhofer-Studie an. „Seltene Erden: Umkämpfte Rohstoffe des Hightech-Zeitalters" ist zwar bereits Ende 2017 erschienen, aber die wesentlichen Aussagen zum Recycling gelten – leider – immer noch: Nur ein kaum nennenswerter Prozentsatz dieser begehrten Rohstoffe wird industriell recycelt. Dabei sind die Verfahren dafür im Labor sattsam erprobt, aber noch fehlen die Investoren für kommerzielle Anlagen.

B1 Consulting Managed Service & Support

individuell – umfassend – kundenorientiert

Neue oder bestehende Systemlandschaften stellen hohe Anforderungen an Ihr IT-Personal. Mit einem individuellen Support- und Betriebsvertrag von B1 Systems ergänzen Sie Ihr Team um die Erfahrung und das Wissen unserer über 140 festangestellten Linux- und Open-Source-Experten.

Unsere Kernthemen:

Linux Server & Desktop · Private Cloud (OpenStack & Ceph) · Public Cloud (AWS, Azure, OTC & GCP) · Container (Docker, Kubernetes, Red Hat OpenShift & Rancher) · Monitoring (Icinga, Nagios & ELK) · Patch Management · Automatisierung (Ansible, Salt, Puppet & Chef) · Videokonferenzen

Unser in Deutschland ansässiges Support- und Betriebsteam ist immer für Sie da – mit qualifizierten Reaktionszeiten ab 10 Minuten und Supportzeiten von 8x5 bis 24x7!

B1 SYSTEMS

B1 Systems GmbH - Ihr Linux-Partner

Linux/Open Source Consulting, Training, Managed Service & Support

ROCKOLDING · KÖLN · BERLIN · DRESDEN

www.b1-systems.de · info@b1-systems.de

Teamwork im Osten

Laura & Malte sind für Sie #NäherDran

Ihr schneller Draht zu Thomas-Krenn: Unsere Vertriebs-Buddies Laura Reischl & Malte Rosenberger sind direkte Ansprechpartner für unsere Kunden im Osten Deutschlands. Damit erreichen Sie mit Ihrem Anliegen nicht nur sofort unsere Zentrale in Südostbayern – auf Wunsch besuchen Sie unsere Server Buddies auch vor Ort, um Ihre IT-Projekte zu besprechen!

Buddy-Beratung unter: thomas-krenn.com/ost

Neu: E-Book Windows Server 2022
Jetzt kostenlos downloaden!
thomas-krenn.com/ebook-ws22

THOMAS KRENN®