WHITE PAPER

cadence

Maximieren Sie Ihren ROI durch die Rückgewinnung verlorener Kapazitäten im Rechenzentrum

Effektivität langfristiger Lösungen für das Rechenzentrumsmanagement, Maximierung der Effizienz durch Kapazitätsrückgewinnung

Von Dave King, Cadence

Da die Bedeutung der Leistung, Effizienz und Nachhaltigkeit von Rechenzentren zunimmt, müssen Eigentümer und Betreiber bei der Verwaltung ihrer Rechenzentren und IT-Implementierungen über „Schnelllösungen" hinausgehen. Diese temporären Lösungen sind weder effektiv noch nachhaltig. Um bedeutende Effizienzsteigerungen zu erzielen, sind langfristige Lösungen und ein effektives Management über einen längeren Zeitraum erforderlich. Die Kapazitätsplanung ist ein solcher Ansatz, bei dem Vorhersageinstrumente eingesetzt werden, um verlorene Kapazitäten zurückzugewinnen und die Effizienz von Rechenzentren zu verbessern. Dieses Whitepaper befasst sich mit dem Problem verlorener Kapazitäten und seinen finanziellen Auswirkungen und bietet gleichzeitig eine Lösung durch effektives Rechenzentrumsmanagement und prädiktive Analysen.

Inhalt

Einleitung

Obwohl eine langfristige Planung von entscheidender Bedeutung ist, arbeitet die Rechenzentrumsbranche oft unter Termindruck und priorisiert häufig kurzfristige Lösungen, um der wachsenden Nachfrage nach neuen IT-Implementierungen gerecht zu werden. Kurzfristige Lösungen können zu einer unausgewogenen Auslastung oder ineffizienten Nutzung von Kapazitäten führen, was die finanziellen Erträge verringert und sich negativ auf die ökologische Nachhaltigkeit des Rechenzentrums auswirkt. Laut Gartner nutzen bestehende Rechenzentren in der Regel nur 50 bis 60 % ihrer Rack-Kapazität[1] , was bedeutet, dass ein erheblicher Teil der Kapazitäten ungenutzt bleibt. Angesichts der enormen Investitionen in den Bau und Betrieb eines Rechenzentrums sowie der Herausforderungen beim Bau neuer Einrichtungen ist dieser Wert unglaublich niedrig, wodurch ein erheblicher Teil der Investitionen verschwendet wird. Darüber hinaus prognostizieren Quellen einen Mangel an Ressourcen für den Bau neuer Rechenzentren, die voraussichtlich nicht ausreichen werden, um die zukünftige Nachfrage zu decken[2] . Das bedeutet, dass bestehende Rechenzentren gezwungen sein könnten, die Auslastung der vorhandenen Flächen zu maximieren, bevor neue gebaut werden.

Die Aufrechterhaltung der Ausfallsicherheit von Rechenzentren und die Berücksichtigung komplexer Faktoren wie Luftstrom können jedoch die Umsetzung langfristiger Lösungen ohne ein Management-Tool für Rechenzentren riskant machen. Um von den angebotenen Lösungen für mehr Effizienz und Nachhaltigkeit zu profitieren, müssen Eigentümer und Betreiber deren Auswirkungen sorgfältig und wissenschaftlich bewerten, bevor sie diese fest in das physische Rechenzentrum integrieren.

Bei der Bewertung der Kapazität eines Rechenzentrums werden mehrere Kennzahlen wie verfügbarer Platz, Strom und Netzwerkanschlüsse berücksichtigt. Die Kühlung von Rechenzentren ist jedoch die Kennzahl, welche am schwierigsten zu quantifizieren und genau vorherzusagen ist. IT-Geräte können nicht eingesetzt werden, wenn eine dieser Kennzahlen ihre Grenze erreicht. Ein ineffizienter Einsatz von IT-Geräten kann die verfügbare Kapazität für zukünftige Einsätze verringern und im Laufe der Zeit zu einer Fragmentierung der Kapazitäten führen.

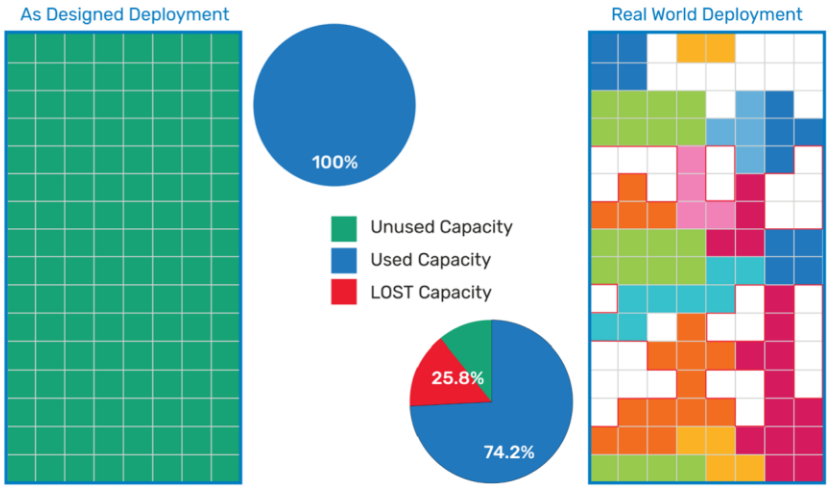

Abbildung 1: Bildliches Beispiel für die Fragmentierung eines Rechenzentrums anhand von Blöcken

Eine langfristige Planung in Verbindung mit einer effektiven Organisation und Analyse der Bereitstellung von Beginn des Lebenszyklus eines operativen Rechenzentrums an kann jedoch unerwartete Einschränkungen bei der zukünftigen Bereitstellung verhindern und somit eine effizientere Nutzung der Kapazität gewährleisten. Eine neue Generation von Rechenzentren verfolgt einen zukunftsorientierten Ansatz und setzt Management-Tools ein, die auf der numerischen Strömungsmechanik (CFD) basieren. Diese helfen dabei, die komplexen Systeme von Rechenzentren in ein verständliches Format zu übersetzen. Alle vorgeschlagenen zukünftigen Änderungen können analysiert werden, einschließlich Kosten-Nutzen-Analysen, Risikobewertungen und Umweltleistungsbewertungen.

[1] „Ihr Rechenzentrum ist veraltet. Was nun? – Gartner.“ 03. Mai 2021, https://www.gartner.com/smarterwithgartner/your-data-center-is-old-now-what

[2] „Europa könnte Schwierigkeiten haben, genügend neue Rechenzentren zu bauen – Bericht.“ 25. Juli 2023, https://www.theregister.com/2023/07/25/aggreko_datacenter_demand_europe/

Dieses Dokument schlägt einen Rahmen zur Ermittlung von Kapazitätsverlusten in Rechenzentren vor und skizziert die negativen finanziellen und ökologischen Auswirkungen von Kapazitätsverlusten in Rechenzentren. Im Gegenzug wird in diesem Dokument die Umsetzung einer langfristigen Kapazitätsplanungsstrategie vorgestellt, die Cadence Reality DC nutzt, eine CFD-basierte Lösung für das Rechenzentrumsmanagement unter Verwendung der Digital-Twin-Technologie. Vertriebspartner in Deutschland ist die ALPHA-Numerics GmbH. Anhand einer Fallstudie wird anhand eines realen Projekts ermittelt, welcher potenzielle Return on Investment (ROI) durch die effektive Kapazitätsauslastung mit Cadence Reality DC erzielt werden kann.

Erklärung der Lücke in der Rechenzentrumskapazität

Während der Entwurfsphase eines Rechenzentrums legen Ingenieure die Auslegungskapazität auf der Grundlage der in Auftrag gegebenen Anlagenleistung fest. Diese Leistung ergibt sich aus den Anforderungen des Colocation-Anbieters, des Hyperscale-Anbieters oder des Unternehmenskunden. Die Auslegungskapazität ist in der Regel ein einzelner Kilowattwert, der auf der erforderlichen Leistung des Systems und der Leistung zur effizienten Kühlung des Rechenzentrums basiert, mit einer zusätzlich vereinbarten Redundanz. Eine weitere Variable, die „verfügbare Kapazität", hängt von vier Ressourcenkomponenten ab: Platz, Strom, Netzwerk und Kühlung.

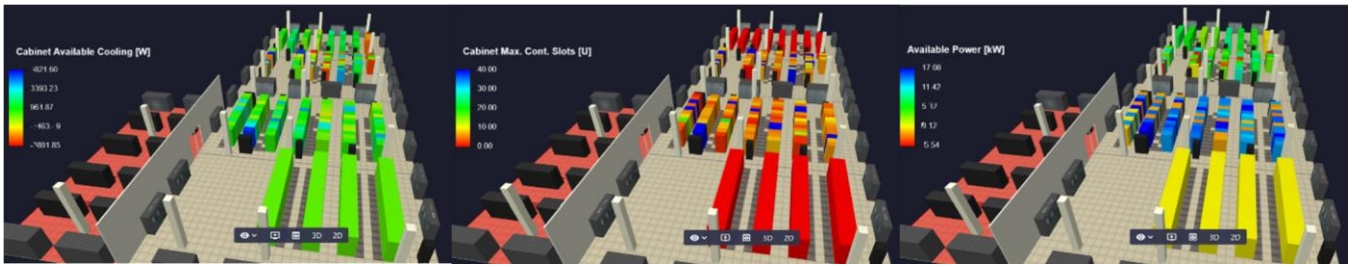

Abbildung 2: Bild und relative Ergebnisse der Kapazitätsverfügbarkeit für ein in Betrieb befindliches Rechenzentrum unter Berücksichtigung von Kühlung, Platz und Strom, erstellt mit Cadence Reality DC Digital Twin

Der Rechenzentrumsdesigner verteilt dann die Leistung auf eine Reihe von Schränken mit einem vorgegebenen Layout und stellt gleichzeitig sicher, dass die einzelnen Schränke eine akzeptable Leistungsdichte aufweisen. Das Kühlsystem wird dann so dimensioniert, dass die erwartete Wärme abgeführt wird und die gewünschte Ausfallsicherheit gewährleistet ist. Schließlich werden das Stromversorgungssystem und das Netzwerk angeschlossen, um eine verteilte Stromversorgung mit wirksamen Ausfallstrategien zu gewährleisten. In der Regel wird dabei von einer fiktiven Leistung pro Rack mit begrenzten Abweichungen ausgegangen, mit Ausnahme vielleicht von erwarteten Zonen mit geringer oder hoher Dichte oder Abweichungen aufgrund unterschiedlicher Arten von Bereichen, wie z. B. Server, Speicher und Netzwerk.

Obwohl dieser konzeptionelle Entwurf erforderlich ist, um das Konzept zu testen und zu validieren, wissen wir jedoch, dass der konzeptionelle Entwurf für die meisten Rechenzentren niemals Realität werden wird. Der Grund dafür ist, dass die tatsächlich einzusetzende IT in der Regel zum Zeitpunkt der Planung noch nicht bekannt ist und dass das Rechenzentrum zwar in seiner Planung festgelegt ist, im Laufe seines Lebenszyklus jedoch viele kleine und große Veränderungen durchläuft, insbesondere durch ständig wechselnde IT-Implementierungen und Änderungen im Schranklayout. Im Durchschnitt wird die IT während des 15- bis 20-jährigen Lebenszyklus eines Rechenzentrums alle drei bis vier Jahre erneuert.[3] Infolgedessen werden die ursprünglichen Designannahmen nach und nach hinfällig, was im Laufe der Zeit zu erheblichen Unterschieden in der Stromverteilung, der Netzwerkverfügbarkeit, der Kühlung und dem Platzbedarf für Rack-U-Slots führt. Das Uptime Institute liefert Beispiele für die Erneuerung von Rechenzentren und die Auswirkungen auf die Stromversorgung und Kühlung von Rechenzentren. Außerdem werden die Trends beim steigenden Stromverbrauch neuerer Server aufgezeigt und wie sich dies negativ auf die Kapazität eines betriebsbereiten Rechenzentrums auswirken kann.[4]

Wenn in einem operativen Rechenzentrum eine der vier Schlüsselkomponenten vollständig ausgeschöpft ist, ist der Einsatz neuer IT-Geräte nicht mehr möglich. Dies kann sogar schon vor der vollständigen Ausschöpfung der Ressourcen eintreten, da die Ressourcen für einen bestimmten Standort nicht übereinstimmen. Daher sinkt die verfügbare Kapazität in jedem Rack praktisch auf null. Bei allen Ressourcenkomponenten kann die Verfügbarkeit für den Einsatz visuell überprüft werden, mit

[3] „Die Lebensgeschichte eines Rechenzentrums – DCD – DatacenterDynamics." 21. Juli 2017, https://www.datacenterdynamics.com/en/analysis/the-data-center-life-story/

[4] „Uptime Institute Data Center Capacity Trends Survey 2022
https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/CapacityTrendsSurvey.Report.02032023.pdf

Ausnahme der Kühlung, die nicht sichtbar ist. Einige Betreiber ignorieren absichtlich oder unabsichtlich die Verfügbarkeit der Kühlung, aber diese Praxis beeinträchtigt wahrscheinlich die Ausfallsicherheit und erhöht das Risiko einer Überhitzung der IT-Geräte im Laufe der Zeit. Wenn Sie nicht überprüfen, ob alle für den von Ihnen gewählten Standort verfügbaren Ressourcen ausfallsicher verfügbar sind, kann dies im Falle eines Ausfalls der Anlageninfrastruktur zu einem unsichtbaren Unfall führen.

Diese Kapazitätsherausforderung hat einen Dominoeffekt zur Folge. Jede ineffiziente Bereitstellung führt zu einer Fragmentierung und schränkt somit die Verfügbarkeit für die nächste Bereitstellung ein. Die einzelnen Arten der Ressourcenfragmentierung in Bezug auf Platz, Strom, Kühlung und Netzwerk werden später in diesem Dokument erläutert.

Die daraus resultierende wachsende Lücke zwischen der geplanten verfügbaren Kapazität und der Betriebskapazität wird als „verlorene“ Kapazität quantifiziert. Verlorene Kapazität wirkt sich sowohl finanziell als auch ökologisch auf Rechenzentrumsunternehmen aus. Zwar erreichen fast alle Rechenzentren nicht ihre maximale geplante Kapazität, doch ist das Ausmaß dieses Versagens erheblich.

Beispiel: Betrachten wir ein Rechenzentrum, das für 15 Jahre für 1 MW an Geräten ausgelegt ist. Nach nur 6 Jahren hat das Rechenzentrum jedoch 50 % ungenutzte Kapazität und kann keine weiteren Geräte sicher unterbringen. Das bedeutet, dass das Unternehmen trotz der Bezahlung von 1.000 kW nur 500 kW „ ” erreichen kann. Es besteht ein zusätzlicher Bedarf von 500 kW, der nicht gedeckt werden kann. Infolgedessen muss das Unternehmen in den Bau eines neuen Rechenzentrums investieren, um diesen Bedarf zu decken, was zu erheblichen Investitionsausgaben führt. Im Wesentlichen zahlt das Unternehmen zweimal für denselben Bedarf. Diese Situation verdeutlicht die tatsächlichen Kosten des Kapazitätsverlusts in einem Rechenzentrum.

Kapazitätsplanungs- und Prognosetools basieren auf der Annahme, dass die gesamte ungenutzte Kapazität verfügbar ist. Dies ist jedoch nie der Fall. Schauen wir uns an, wie Fläche, Strom, Kühlung und Netzwerk fragmentiert werden und somit für zukünftige Bereitstellungen verloren gehen können.

Platz

Bei der Beurteilung, ob Geräte in einer Einrichtung auf der Grundlage des verfügbaren Platzes eingesetzt werden können, reicht es nicht aus, sich ausschließlich auf den gesamten U-Slot-Platz zu verlassen; auch die Größe des zusammenhängenden U-Platzes ist ein entscheidender Faktor, der berücksichtigt werden muss. Wenn beispielsweise im gesamten Rechenzentrum nur wenige 1U-Slots verfügbar sind und der erforderliche Platzbedarf für die IT 2U beträgt, kann der Einsatz nicht durchgeführt werden. Dieses extreme Beispiel verdeutlicht, wie wichtig es ist, zusammenhängenden Platz gegenüber dem Gesamtplatz zu priorisieren.

In der Regel nimmt die Menge an zusammenhängendem U-Platz im Laufe des Betriebszyklus aufgrund der „Fragmentierung” des U-Slot-Platzes durch die Bereitstellung und Außerbetriebnahme von IT-Geräten ab.

Beispiel: Nächste Woche muss ein neuer 4U-Server installiert werden, und der aktuelle Zustand des Schranks lässt darauf schließen, dass Steckplatz 28 der erste verfügbare U-Steckplatz ist. Allerdings soll einer der 4U-Server am unteren Ende des Schranks morgen außer Betrieb genommen werden. Zum Zeitpunkt der Installation ist der tatsächlich erste verfügbare U-Steckplatz Steckplatz 10. Der Installateur weiß dies nicht, und der Plan sieht vor, die IT in Steckplatz 28 zu installieren, was zu einer Fragmentierung der verbleibenden verfügbaren U-Steckplätze führt, die nun in zwei Abschnitte statt in einen zusammenhängenden Block unterteilt sind (siehe Abbildung 3).

Rack-Platz, der zum Zeitpunkt der Planung einer Bereitstellung als verfügbar angesehen wird
Planung einer Bereitstellung (grün markiert)
Rack-Platz mit geplanter Stilllegung (rot markiert)
Rack-Platz nach der Bereitstellung, was zu einer Fragmentierung des verfügbaren Platzes

Abbildung 3: Beispiel für eine Fragmentierung im durchgehenden U-Slot-Platz eines Schranks

Bereitstellungspläne werden in der Regel kurzfristig erstellt, wobei nur der aktuelle Zustand jedes Racks berücksichtigt wird, ohne zu bedenken, wie sich dies auf den verfügbaren Platz für zukünftige Bereitstellungen auswirkt. Oft sehen wir Schränke, die scheinbar willkürlich gefüllt sind, aber dies ist nur das Ergebnis langjähriger Bereitstellungs-/Stilllegungszyklen.

Das gleiche Problem sehen wir auch im Rechenzentrum, insbesondere in Colocation-Rechenzentren. Wenn Colocation-Anbieter die Fläche im Rechenzentrum für Kunden aufteilen und zuweisen, geht leicht Kapazität verloren. Wenn Kunden wegziehen und neue Kunden hinzukommen, wird der Platz immer weiter aufgeteilt (in der Regel in Käfige), bis für jede weitere Bereitstellung immer weniger nutzbarer Platz übrig bleibt. Colocation-Anbieter stehen dann vor der Aufgabe, bestehende Kunden umzusiedeln, um verlorene Kapazität zurückzugewinnen.

Strom

Bei der Planung eines Rechenzentrums wird die Gesamtleistung auf verschiedene Weise aufgeteilt, unter anderem über angeschlossene Fernstromverteiler, die den Strom an die Racks oder Sammelschienen mit Abzweigungen verteilen. Einige Designs erfordern möglicherweise eine höhere Stromverfügbarkeit in bestimmten Bereichen, z. B. in Zonen mit hoher Dichte oder Netzwerkzonen. Die Stromverteilung erfolgt entsprechend den spezifischen Bereichen des Rechenzentrums und den Arten der stationären Geräte, z. B. Netzwerk, Speicher und Server. Die Stromverteilung im Raum ist zwar nicht gleichmäßig über das gesamte Raumdesign verteilt, geht jedoch auf Schrankebene von einer gewissen Gleichmäßigkeit aus.

In operativen Rechenzentren weisen die neu installierten IT-Geräte jedoch unterschiedliche Leistungsdichten und Größen auf, und das Verhältnis zwischen dem Platzbedarf und der Leistungsaufnahme ist je nach Technologie unterschiedlich. Beispielsweise verwenden rackmontierte Speichersysteme eine große Anzahl von Festplatteneinschüben mit geringer Leistung und können einen gesamten Schrank füllen, ohne die Auslegungsleistung zu erreichen. Im Gegensatz dazu können Server und Blades die gesamte Leistung eines Schranks in weniger als der Hälfte des verfügbaren U-Raums verbrauchen.

Diese Diskrepanz zwischen verschiedenen IT-Geräten führt zu mehreren Problemen. Schränke, die viele Geräte mit geringem Stromverbrauch enthalten, haben keinen Platz mehr, bevor sie keine Leistung mehr haben, sodass die verbleibende Leistung nicht verfügbar ist. Umgekehrt haben Schränke mit dichteren Verarbeitungsgeräten keine Leistung mehr, bevor der Platz im Schrank aufgebraucht ist.

Schrank	1	2	3	4	5
Gesamt verfügbarer U-Slot-Speicherplatz	17U	19U	30U	21U	21U
Der größte zusammenhängende Block von U-Steckplätze	8U	8U	26U	8U	10U
Verfügbare Leistung	-1,2 kW	1,7 kW	0 kW	2 kW	1,7 kW

Verfügbar für den Einsatz von 11 U 1,4 kW	Nicht verfügbar	Nicht verfügbar	Nicht verfügbar	Nicht verfügbar	Nicht verfügbar

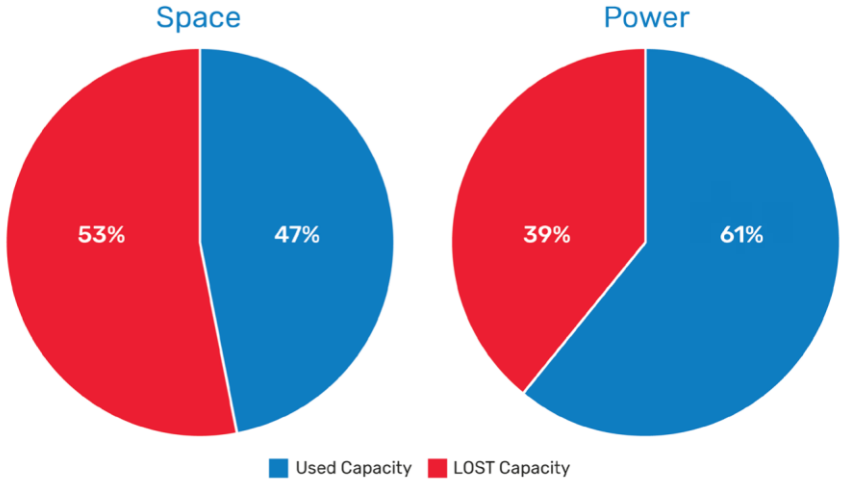

Abbildung 4: Beispiel für Raum- und Leistungsfragmentierung

Beispiel: Abbildung 4 zeigt fünf Schränke mit unterschiedlichen Lasten.

Es besteht die Anforderung, einen neuen HPC-Cluster in diesen Schränken zu installieren, der 11 HE Rackplatz und 1,4 kW Strom benötigt. Aufgrund der Art und Weise, wie die Clusterkomponenten kommunizieren, müssen sie in einem zusammenhängenden Blockplatz untergebracht werden.

Aufgrund dieser Einschränkungen kann keines der Schränke die neuen Geräte aufnehmen, sowohl aus Platz- als auch aus Stromgründen. Daher ist die verbleibende Kapazität in diesem Szenario nicht verfügbar und geht verloren.

Darüber hinaus kann der Phasenausgleich zu einer Fragmentierung der verfügbaren Kapazität führen. Eines der Hauptziele des Facility Managers ist es, die Ausfallsicherheit des Rechenzentrums zu gewährleisten und einem ausgewogenen Phasenausgleich Vorrang vor der Kapazität einzuräumen.

Beispiel: Eine 90-kW-PDU verfügt über 30 kW auf jeder der drei Phasen. Wenn jedoch die Schränke auf Phase eins nur mit 10 kW Leistung belegt sind, müssen die verbleibenden 20 kW aus Phase eins auf andere Schränke verteilt werden, um die Auslegungskapazität aufrechtzuerhalten, da sonst die 20 kW Kapazität verloren gehen. Der Facility Manager wird versuchen, dieses Ungleichgewicht so weit wie möglich zu vermeiden, oft auf Kosten der nutzbaren Kapazität.

Netzwerk

Bei der Konzeption des Datennetzwerks werden viele Aspekte berücksichtigt, darunter Funktionsbereiche wie der Haupt- oder horizontale Verteilungsbereich (MDA oder HDA), die Netzwerktopologie, strukturierte oder unstrukturierte Verkabelung, Kabeltypen und Patching, um nur einige zu nennen. Diese Konzeption wird schließlich in Netzwerkausrüstung und Patching umgesetzt sowie in die Zuweisung der verfügbaren Ports für bestimmte Netzwerke oder zukünftige Erweiterungen. Die installierte Hardware hat direkten Einfluss darauf, wie viele Ports in der Betriebsphase belegt sind.

Beispiel: Rackmount-Servergeräte können bis zu zehn Kupferports für eine Rechenleistung von etwa 500 W nutzen, während ein Blade-Chassis mit der gleichen Anzahl von Ports eine Rechenleistung von bis zu 5 kW bereitstellen kann. Wenn ein Rechenzentrum mit einer Portdichte von 48 Netzwerkports und 5 kW Leistung pro Schrank ausgestattet ist, würde der Einsatz eines Blade-Chassis die gesamte Leistung nutzen, aber 80 % der Ports ungenutzt lassen, und die Rackmount-Server könnten alle Ports nutzen, aber nur 50 % der Leistung.

Kühlung

Wie bereits erwähnt, ist die Kühlung die anspruchsvollste und komplexeste Kapazitätskennzahl eines Rechenzentrums. Erstens ist der Luftstrom unsichtbar, was seine Steuerung erschwert. Zweitens ist es ohne spezielle physikalisch basierte Simulationssoftware für Rechenzentren fast unmöglich zu verstehen, wo die Kühlung verteilt ist, warum Hotspots auftreten und wie sich die Luftzufuhr verbessern lässt. Selbst bei einem „stationären" Rechenzentrumsdesign ist die Berechnung der Kühlung ohne die Hilfe speziell entwickelter Tools schwierig.

Aufgrund der Gleichmäßigkeit von Raum und Leistung im Rechenzentrumsdesign zielt das Kühlungsdesign auf einen ähnlichen Grad an Gleichmäßigkeit ab, wobei der Luftstrom gleichmäßig auf jedes IT-Gerät verteilt wird, um der Verteilung der anderen Ressourcen zu entsprechen. Strategien zum Wärmemanagement sind darauf ausgelegt, die Kühlungsversorgung an den Kühlungsbedarf der IT anzupassen, um sicherzustellen, dass alle IT-Geräte relativ gleiche Temperaturen (oft als Einlasstemperatur gemessen) aufweisen, die für den Betrieb und die effektive Ableitung der Abwärme erforderlich sind. Im Wesentlichen ist das Wärmemanagementsystem darauf ausgelegt, die IT-Geräte auf die für den Betrieb erforderlichen Temperaturen zu kühlen. All dies basiert auf einer angenommenen Verlustleistung pro Rack und, was wichtig ist, einem angenommenen Luftstrom pro kW.

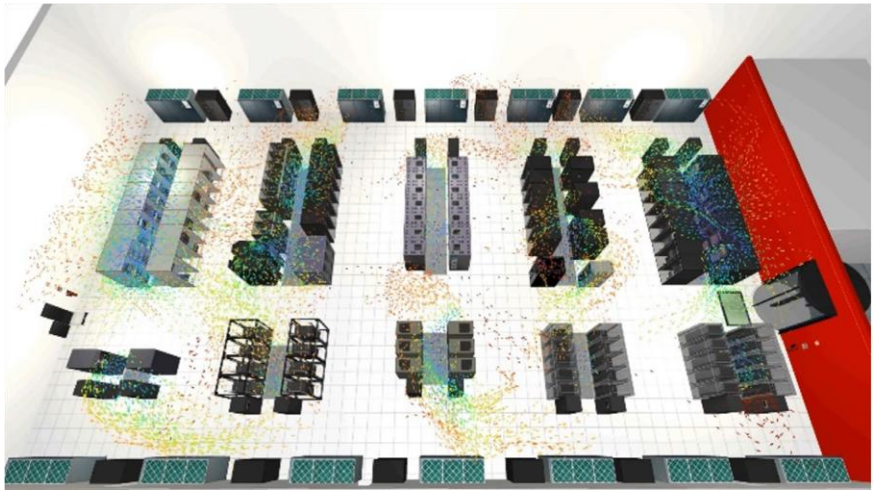

Abbildung 5: Cadence Reality DC Digital Twin mit ungleichmäßiger Luftstromverteilung, dargestellt mit Stromlinienplots des Luftstroms.

Die Bewertung der Kühlleistung eines in Betrieb befindlichen Rechenzentrums ist aufgrund seiner Dynamik erheblich schwieriger. Die Gleichmäßigkeit der Kühlung wird durch minimale Änderungen in der Betriebsphase bis hinunter auf Rack-Ebene drastisch beeinflusst, was zu Änderungen im Luftstrom und in der Kühlung des gesamten Rechenzentrums führt. Es gibt viele Faktoren, die zu einer Fragmentierung der Kühlgleichmäßigkeit im gesamten Rechenzentrumsbetrieb führen, darunter: ← Der Einsatz unterschiedlicher IT-Geräte: Jedes IT-Gerät benötigt unterschiedliche Luftmengen, die in verschiedene Richtungen strömen müssen, um die Chips im Inneren zu kühlen. Diese Unterschiede bestehen nicht nur zwischen verschiedenen Arten von IT-Geräten, sondern auch zwischen Herstellern und sogar zwischen verschiedenen Generationen desselben Modells. ← Standort der Geräte: Die Fähigkeit des Kühlsystems, die richtige Luftmenge an die Einlässe aller IT-Geräte in der Anlage zu liefern, hängt davon ab, wo sich diese befinden. ← Luftstromverteilung auf bestimmte Bereiche oder geschlossene Reihen: Dies kann zu einer Diskrepanz zwischen dem Kühlbedarf und der bereitgestellten Kühlleistung führen. In diesem Szenario kommt es in Bereichen mit weniger Luft als erforderlich wahrscheinlich zu einer Überhitzung, während Bereiche mit übermäßiger Kühlung ineffizient arbeiten. ← Stromverbrauch: Der Stromverbrauch variiert im Laufe des Tages je nach Nutzung der Anwendungen durch die Benutzer, was zu einem Anstieg oder Rückgang des Kühlbedarfs führen kann. ← Schlechtes Leckagemanagement: Kabeldurchführungen, fehlende Blindplatten, schlechte Einhausungskonstruktion oder schlechte Schrankauswahl ermöglichen es der Kühlung, die IT zu umgehen und zum Kühlsystem zurückzukehren. ← Kühlungssteuerung: Ob es sich nun um Steuerungsschemata für Kühlaggregate handelt, die die Wirksamkeit anderer Kühlaggregate zunichte machen, um schlechte Sensorpositionen oder um ineffektive Steuerungslogik – schlechte Steuerungsstrategien können die verfügbare Kühlung beeinträchtigen.

Abbildung 6: Abluftströmungsmuster auf Schrankebene für einen Konstruktionsschrank (rechts) mit derselben Ausrüstung und einer gleichmäßig verteilten Last und einen Betriebsschrank (links) mit unterschiedlichen IT- und Leistungsdichten

Ein weiterer Aspekt bei der Gestaltung und Betriebskapazität von Rechenzentren, der für den Facility Manager oberste Priorität hat, ist die Ausfallsicherheit, d. h. die Gewährleistung, dass das Rechenzentrum einem Ausfallszenario sicher standhalten kann. Strom- und Datennetzwerke sind in ihrer Gestaltung und ihrem Betrieb auf einem Redundanzniveau recht unkompliziert, wobei viele eine standardmäßige 2N-Architektur für Ausfälle verwenden. Es ist jedoch erheblich schwieriger zu verstehen, wie eine Anlage auf einen Ausfall der Kühlung oder einen Wartungsplan reagieren wird, wenn man keine Simulation einsetzt, um die Reaktion der Anlage vorherzusagen.

Das Problem wird durch die mangelnde Verantwortlichkeit sowohl der Facility Manager als auch der IT-Abteilungen für die Regulierung des Luftstroms und der Kühlung im Rechenzentrum noch verschärft. In vielen Unternehmen sorgt der Facility Manager dafür, dass das Kühlsystem ausreichend ist, hat aber keine Kontrolle über den IT-Einsatz. Ebenso haben die IT-Teams die Kontrolle über den IT-Einsatz, aber das Kühlsystem liegt außerhalb ihres Zuständigkeitsbereichs. Aufgrund dieses Widerspruchs kann das Luftstrommanagement unbeabsichtigt vernachlässigt werden, ohne dass jemand die Verantwortung dafür übernimmt. Die ultimative Folge ist eine unzureichende Planung der Bereitstellungen, was zu Hotspots und Kapazitätsverlusten führt.

Kombinierte Kapazität in Netzwerk, Stromversorgung, Platz und Kühlung

Es reicht nicht aus, die vier Schlüsselkennzahlen des Rechenzentrums einzeln zu betrachten, da sie sich alle gegenseitig beeinflussen und das Problem dadurch weitaus komplexer machen. Beispielsweise können Netzwerkports oder die Stromverfügbarkeit eine Bereitstellung in einem Bereich erzwingen, der sich negativ auf die Leistung des Kühlsystems auswirken kann. Daher ist es von entscheidender Bedeutung, die verschiedenen Teile des Rechenzentrums-Ökosystems sorgfältig zu bewerten und auszugleichen, um eine langfristige Kosten-Nutzen-Analyse für jede Bereitstellung zu optimieren und zu entschlüsseln.

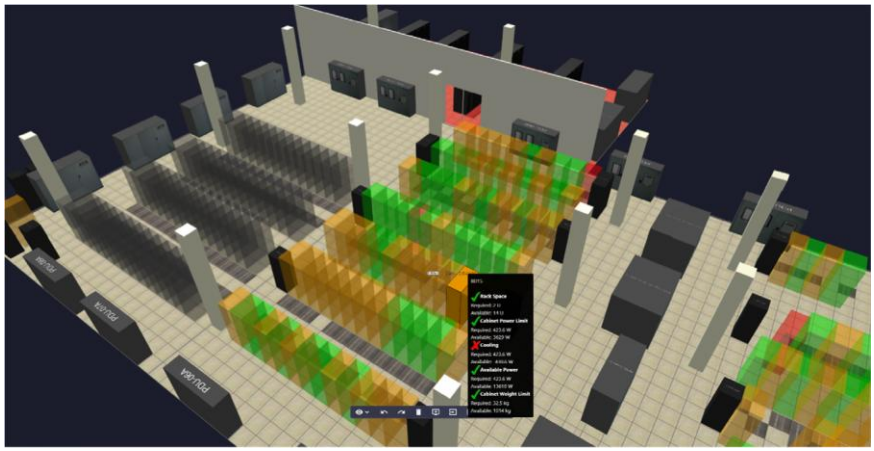

Abbildung 7: Cadence Reality DC Digital Twin zeigt verfügbaren Rack-Platz und Stromversorgung für IT-Bereitstellung, jedoch unzureichende Kühlung.

Die Betriebskosten für nicht ausgelastete Kapazitäten

Unternehmen geben neue Rechenzentren in Auftrag, um den gestiegenen Bedarf an Rechenleistung zu decken. Die IT-Abteilung teilt dem Unternehmen die erforderliche Rechenleistung mit, die zur Deckung des Bedarfs benötigt wird. Anschließend wird ein Rechenzentrum mit den entsprechenden Konstruktionsparametern und maximalen Kapazitäten in Auftrag gegeben, die diesen Leistungsbedarf erfüllen, wie bereits weiter oben in diesem Dokument ausführlicher erläutert. Ein Rechenzentrum verursacht viele Kosten, die sich jedoch in zwei Gruppen einteilen lassen: Investitionsausgaben (CAPex) und Betriebskosten (OPex).

CAPex bezieht sich auf die Kosten für Sachanlagen, die unverzichtbar und dauerhaft sind. OPex bezieht sich auf die geschätzten Kosten, die während des Betriebszyklus des Rechenzentrums anfallen und nicht feststehen. Diese Kosten lassen sich wiederum in weitere Gruppen unterteilen, darunter: ← Feste Betriebskosten – Wartungskosten, Personalkosten und Steuern

- ← Lastabhängige Betriebsausgaben – Strom- und Kühlungskosten für die installierte Last
- ← Kosten für den Lagerumschlag, z. B. IT-Ausrüstung

In der Inbetriebnahmephase werden die OPex häufig niedriger eingeschätzt als die tatsächlichen OPex-Kosten, die dem Rechenzentrum während seines gesamten Lebenszyklus entstehen.[5] Dieser Anstieg der prognostizierten OPex kann durch unvorhergesehene Erhöhungen des Anwendungs- und Strombedarfs verursacht werden, wodurch die Kapazität vorzeitig erschöpft wird. Darüber hinaus führt eine ineffiziente Nutzung des Rechenzentrums dazu, dass die Kapazität schneller als erwartet ausgeschöpft ist, was Unternehmen dazu zwingt, zusätzliche Kapazität durch neue Rechenzentren oder die Cloud zu erwerben. Dies wird im folgenden Abschnitt anhand von Beispielen näher erläutert, die anhand einer anonymisierten Fallstudie aus der Praxis zeigen, wie ungenutzte Kapazitäten zu verschwendeten CAPex-Kosten und erhöhten OPex-Kosten führen können.

Fallstudie zur Kapazitätsauslastung

Die folgende Fallstudie untersucht ein hypothetisches Rechenzentrum mit einer Leistung von 1,3 MW und einer erwarteten Lebensdauer von 15 Jahren. Die Investitionskosten (CAPex) dieses Falls werden auf 50 Millionen US-Dollar geschätzt, berechnet mit dem vom Uptime Institute angebotenen Rechner für die Gesamtbetriebskosten (TCO).[6] Dieser Rechner verwendet die statische Annuitätenmethode. Parameter wie die Auslegungskapazität (1,3 MW), das Kühlsystem, die USV-Architektur und der Redundanzgrad werden eingegeben, um eine Kostenübersicht zu erhalten.

Die Kosten für die IT-Ausrüstung können bei den Berechnungen vernachlässigt werden, da sich die Verarbeitungsanforderungen des Unternehmens nicht ändern, wenn das Rechenzentrum seine Kapazität nicht erreicht. In

[5] „Verwendung eines Gesamtbetriebskostenmodells (TCO) für Ihr Rechenzentrum." 1. Oktober 2013, https://www.datacenterknowledge.com/archives/2013/10/01/using-a-total-cost-of-ownership-tco-model-for-your-data-center/

[6] „Ein einfaches Modell zur Ermittlung der tatsächlichen Gesamtbetriebskosten für Daten ..." https://datacenters.lbl.gov/sites/default/files/%28TUI3011B%29SimpleModelDeterminingTrueTCO.pdf

dieser Fallstudie wird davon ausgegangen, dass die verbleibende IT-Ausrüstung, die aufgrund der nicht verfügbaren Kapazität nicht eingesetzt werden kann, in einer anderen Einrichtung eingesetzt wird.

Die Ausgaben werden jährlich für die ROI-Berechnungen erfasst. Da es sich bei den CAPex um Fixkosten handelt, können sie gleichmäßig auf die erwartete Lebensdauer des Rechenzentrums von 15 Jahren verteilt werden, was einer jährlichen Rate von 6,3 Mio. USD entspricht.

Die festen Betriebskosten können im Gegensatz zu den CAPex im Laufe der Zeit variieren, obwohl sie im Allgemeinen außerhalb der Kontrolle des Eigentümers/Betreibers des Rechenzentrums liegen. In diesem Fall werden die festen Betriebskosten auf 3,1 Mio. USD pro Jahr berechnet.

Die lastabhängigen Betriebskosten sind die einzige echte Variable und hängen von der in der Anlage installierten Last ab. Diese Variable unterliegt ebenfalls in gewissem Maße der Kontrolle des Eigentümers/Betreibers des Rechenzentrums. Eine ineffiziente Nutzung der Kapazität des Rechenzentrums kann die lastabhängigen Betriebskosten erheblich erhöhen. Eine effizientere Nutzung des Rechenzentrums und eine bessere Auslastung der Kapazitäten können diese prognostizierten Kosten senken. Wie bereits in diesem Dokument erläutert, wurde die typische Kapazitätsauslastung (gekennzeichnet als genutzte maximale Last) eines Rechenzentrums für dieses Beispiel auf 50 % gerundet. Bei Anwendung dieser Zahl kostet eine Auslastung von 50 % 0,9 Mio. USD pro Jahr. Im Gegensatz dazu kostet eine Auslastung von 100 % theoretisch 1,4 Mio. USD pro Jahr, was etwas weniger als 50 % ist, da die Berechnung auch die Ineffizienz der Anlage berücksichtigt.

Abbildung 8 vergleicht die geschätzten jährlichen Gesamtkosten für das Beispiel-Rechenzentrum bei einer Auslastung von 50 % und 100 %. Die variablen Kosten sind die lastabhängigen Betriebsausgaben und gleichzeitig die geringsten Kosten. Daher ermöglicht dieser geringe Kostenanstieg (ca. 5 %) dem Eigentümer/Betreiber, 50 % der im Voraus bezahlten Kapazität zurückzugewinnen.

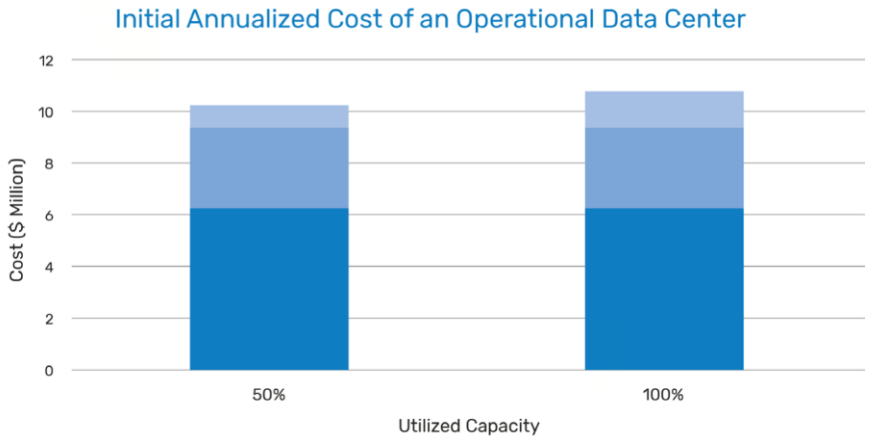

Abbildung 8: Jährliche Kosten eines operativen Rechenzentrums bei einer Auslastung von 50 % und 100 %

Wie bereits erwähnt, bleibt der IT-Bedarf des Unternehmens unverändert. Bei einer Auslastung von nur 50 % sind beispielsweise noch weitere 0,65 MW Rechenleistung erforderlich, um die volle Auslegung von 1,3 MW zu erreichen. Das bedeutet, dass ein weiteres identisches Rechenzentrum gebaut werden muss, um den Geschäftsbedarf zu decken. Es treten dieselben Probleme auf, und auch die zweite Anlage gilt bei einer Auslastung von 50 % als voll ausgelastet. Daher sind die Gesamtkosten für 1,3 MW Rechenleistung doppelt so hoch wie ursprünglich veranschlagt.

Abbildung 9 zeigt die jährlichen Gesamtkosten eines neuen Rechenzentrums, das entsprechend der Auslastung gebaut wurde.

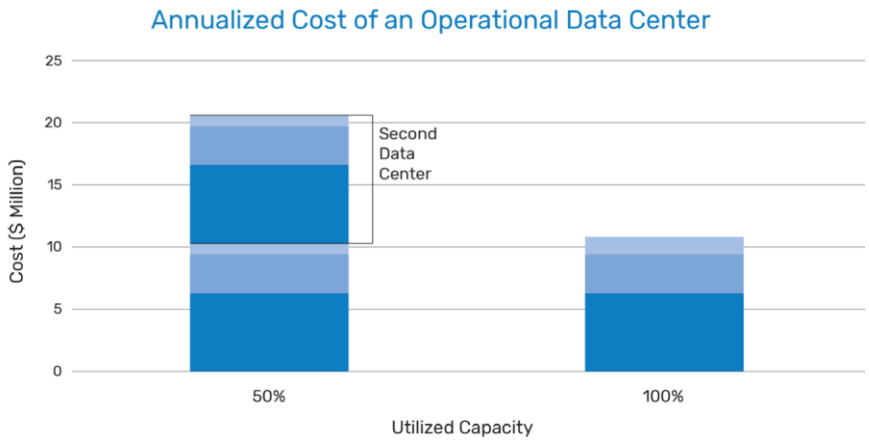

Abbildung 9: A: Jährliche Kosten eines operativen Rechenzentrums, das zu 50 % ausgelastet ist, mit zwei Rechenzentren und einem einzelnen Rechenzentrum, das zu 100 % ausgelastet ist

Dadurch steigen die Kosten pro Kilowatt Rechenleistung ($/kW). Für ein typisches Rechenzentrum mit einer Auslastung von 50 % sind die Kosten in diesem Beispiel fast doppelt so hoch wie für ein vollständig ausgelastetes einzelnes Rechenzentrum (siehe Abbildung 10).

Da jedes Rechenzentrum nicht vollständig ausgelastet ist, zahlt der Eigentümer/Betreiber effektiv zweimal die zusätzlichen Kosten für die Unterauslastung, die deutlich höher sind als die Kosten für eine einzige, vollständig ausgelastete Anlage.

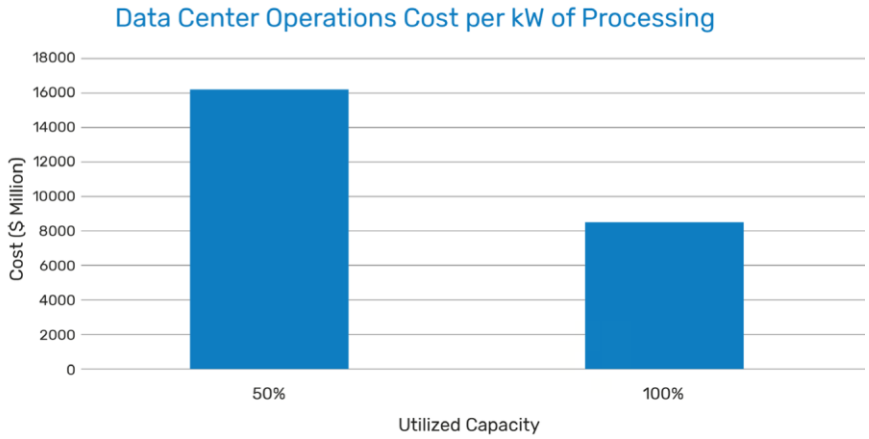

Abbildung 10: Betriebskosten eines Rechenzentrums pro kW Rechenleistung bei einer Auslastung von 50 % und 100

Es ist wichtig zu verstehen, dass kein Rechenzentrum eine Auslastung von 100 % erreichen kann, aber dieses Beispiel veranschaulicht, wie viel die Unterauslastung der Rechenzentrumskapazität tatsächlich kosten kann. Es zeigt auch den ROI, den ein Eigentümer/Betreiber eines Rechenzentrums durch eine bessere Auslastung der Rechenzentrumskapazität erzielen kann.

Wenn ein Rechenzentrum ungenutzte Kapazitäten zurückgewinnt, entfallen die Kosten für Strom, Wasser usw., da es bereits die vorhandene Kühlung nutzt. Die einzigen Kosten, die anfallen, sind die Kosten für die Änderungen, die zur Rückgewinnung der ungenutzten Kapazitäten erforderlich sind, wie z. B. die Kosten für die Installation neuer Geräte, etwaige Mehrkosten für die Stromversorgung der Kühlinfrastruktur und manuelle Arbeit.

So vermeiden Sie Kapazitätsverluste

Ein vollständiges Vermeiden von Kapazitätsverlusten ist für kein Rechenzentrum möglich, aber sie können erheblich minimiert werden. Um Kapazitätsverluste zu vermeiden, muss jede Änderung am Rechenzentrum und jede implementierte Bereitstellung unter Berücksichtigung der langfristigen Auswirkungen auf die Kapazität effektiv bewertet werden. Dies gilt insbesondere für die Optimierung der Bereitstellung und die Analyse der Fragmentierung von Raum, Strom, Netzwerk und Kühlung.

Es gibt eine Lösung für das Kapazitätsproblem. Cadence Reality DC Digital Twin kombiniert moderne Rechenzentrumsmanagementsysteme und simuliert komplexe Betriebsfaktoren wie Kühlung und Luftstrom mithilfe von Computational Fluid Dynamics (CFD). Dieses Tool bewertet die aktuelle Kapazität des Rechenzentrums in Bezug auf Strom, Platz, Netzwerk und Kühlung und liefert gleichzeitig Einblicke in die zukünftige Bereitstellung. Benutzer können dann effektivere Bereitstellungen für die Kapazitätsauslastung wählen und gleichzeitig die Ausfallsicherheit des Rechenzentrums gewährleisten.

Ein gewisser Kapazitätsverlust ist unvermeidbar, aber mit einem Tool wie Cadence Reality DC Digital Twin lassen sich die Folgen jeder Bereitstellung effektiv nachvollziehen. Die Kapazität kann maximiert werden, was sowohl zu finanziellen Einsparungen für das Unternehmen als auch zu effizienteren Rechenzentren für die Umwelt führt. Schließlich kann eine ordnungsgemäße Kosten-Nutzen-Analyse durchgeführt werden, die bei der Prognose der Kosten für den nächsten Rechenzentrumsbau hilft.

Fazit

In einer idealen Welt würden die vier Kapazitätskomponenten des Rechenzentrums gleichzeitig eine Auslastung von 100 % erreichen. Um dies zu erreichen, muss die Anlage nach einem bestimmten Plan entworfen werden, von dem kaum abgewichen werden darf. In den heutigen unternehmenskritischen Anlagen ist dies jedoch selten möglich. Um mit den technologischen Verbesserungen Schritt zu halten, müssen Geschäftspläne kontinuierlich weiterentwickelt werden. Damit diese Weiterentwicklung nicht zu Lasten der Kapazität geht und die Rentabilität der Anlage nicht erheblich beeinträchtigt, muss jede Bereitstellung vor der Umsetzung bewertet werden. Physikalische Simulationstechniken bieten Betreibern von Rechenzentren eine wissenschaftliche Methode, um sich einen genauen Überblick über den aktuellen und zukünftigen Zustand ihrer Anlage zu verschaffen.

Es ist wichtig zu beachten, dass sich dieses Whitepaper auf die Kapazitätsauslastung im Unternehmensbereich konzentriert. Diese Methoden können jedoch in ähnlicher Weise auch für andere Rechenzentrumsbesitzer/-betreiber, wie z. B. Colocation-Anbieter, verwendet werden.

Cadence ist ein führendes Unternehmen im Bereich Elektronikdesign und Computertechnik und nutzt seine Strategie des intelligenten Systemdesigns, um Designkonzepte in die Realität umzusetzen. Zu den Kunden von Cadence zählen die kreativsten und innovativsten Unternehmen der Welt, die außergewöhnliche Elektronikprodukte von Chips über Platinen bis hin zu kompletten Systemen für die dynamischsten Anwendungen liefern.**www.cadence.com**

Für weitere Informationen wenden Sie sich bitte an

ALPHA-Numerics GmbH
Römerstraße 32
56355 Nastätten
Deutschland

info@alpha-numerics.de
+49 6772 969 3430
www.alpha-numerics.gmbh

Als Pionier bei der Einführung des Digital Twin von elektronischen Geräten im Rechenzentrum bietet ALPHA-Numerics leistungsstarke Softwarelösungen für Designer und Betreiber. Diese von dem amerikanischen Unternehmen Cadence Design Systems entwickelte Software basiert auf virtuellen 3D-Modellen, welche die Vorhersage der Kühlluftwege sowie der thermischen Gegebenheiten berechnet (CFD-Technologie) und in einem 3D-Modell darstellen kann.

Digitale Zwillinge von Rechenzentren eröffnen neue Perspektiven, welche über die klassischen Funktionalitäten von DCIM-Software für das tägliche Management von Computer- oder Telekommunikationsräumen hinausgehen. Sie sind wertvolle Verbündete bei der Reduzierung ihres ökologischen Fußabdrucks, eine große Herausforderung angesichts des Klimawandels. Sie ermöglichen es, nachhaltige Lösungen für den Einsatz von KI (Künstliche Intelligenz) oder Edge Computing zu finden. Auf der Ebene der elektronischen Geräte können Designer dank unserer Lösungen innovative Wege für die Kühlung immer kompakterer und hochgradig wärmeabgebender Systeme entwerfen und validieren.