

Überdimensionierte Datencenter- und Netzwerkraum- Infrastrukturen: Kostenvermeidung

White Paper Nr.37



Zusammenfassung

Die räumliche und strombezogene Infrastruktur von Datacentern und Netzwerkräumen ist in der Regel um mehr als 100 % überdimensioniert. Dieses Weißbuch legt Statistiken zur Überdimensionierung vor. Die damit verbundenen Kosten werden beziffert. Darüber hinaus werden die Hauptursachen für diese Überdimensionierungen dargelegt. Schließlich wird eine Architektur und Methode zur Vermeidung dieses Problems beschrieben.

Einleitung

In diesem Weißbuch wird dargelegt, dass die Überdimensionierung den größten vermeidbaren Kostenfaktor in Bezug auf übliche Datacenter- und Netzwerkraum-Infrastrukturen ausmacht. Die Auslastung der räumlichen und strombezogenen Infrastruktur von Datacentern und Netzwerkräumen beträgt in der Regel weit unter 50 %. Durch die ungenutzte Kapazität von Datacentern und Netzwerkräumen entstehen Kapitalkosten und in der Folge auch Betriebs- und Wartungskosten, die vermieden werden können.

Dieses Weißbuch ist in drei Teile aufgeteilt. Zunächst werden die Fakten und statistischen Erhebungen in Bezug auf die Überdimensionierung beschrieben. Als Nächstes werden die Gründe hierfür dargelegt. Schließlich wird eine Architektur und Methode zur Vermeidung der dadurch verursachten Kosten beschrieben.

Fakten und Statistiken

Auf dem IT-Sektor und im Gebäudemanagement lässt sich immer wieder beobachten, dass verfügbare Grundflächen, die Kapazitäten sowie andere infrastrukturelle Möglichkeiten in den Datacentern nicht vollständig ausgenutzt werden. Damit dieses Phänomen in Zahlen ausgedrückt werden kann, müssen zunächst die in diesem Zusammenhang verwendeten Termini definiert werden.

Definitionen

Für dieses Dokument werden die folgenden Termini nachstehend definiert:

Terminus	Definition
Anlagen-nutzungszeitraum	Der gesamte geplante Nutzungszeitraum des Datacenters. In der Regel 6 bis 15 Jahre. Generell wird von 10 Jahren ausgegangen.
Raumkapazität	Die Maximallast für den Raum. Bereits bei der anfänglichen Einrichtung können alle oder ein Teil der Strom- und Kühlanlagen installiert werden, die für diese Kapazität erforderlich sind.
Installierte Kapazität	Die Lastfähigkeit der installierten Strom- und Kühlanlagen. Entspricht der Raumkapazität bzw. liegt darunter.
Erwartete Last	Der geschätzte Strombedarf zum Zeitpunkt der Einrichtung des Systems und über seine Nutzungsdauer. Erhöht sich in der Regel im Laufe der Zeit.
Istlast	Der tatsächliche Strombedarf zum Zeitpunkt der Einrichtung des Systems und über seine Nutzungsdauer. Erhöht sich in der Regel im Laufe der Zeit.

Ausgangspunkte zur Modellbildung

Zur Erhebung und Analyse von Daten, die im Zusammenhang mit der Überdimensionierung stehen, haben wir von APC eine Einschätzung der Situation bei verschiedenen Benutzern vorgenommen und anhand der ermittelten Daten ein vereinfachtes Modell entwickelt, das zur Beschreibung der infrastrukturellen Kapazitätenplanung für Datacenter herangezogen werden kann. Das Modell stützt sich auf folgende Annahmen:

- ? Der Nutzungszeitraum des Datacenters ist auf 10 Jahre ausgelegt
- ? Für die Datacenterplanung wurde eine maximale Anlagenkapazität und ein geschätzter Anfangsstrombedarf ermittelt
- ? Bei der typischen Nutzungsdauer eines Datacenters steigt die erwartete Last linear vom Startwert der geschätzten Last bei der anfänglichen Einrichtung an und erreicht ihren Höchstwert in der Hälfte der erwarteten Nutzungsdauer.

Das oben definierte Modell führt zum in Abbildung 1 dargestellten Planungsmodell. Dieses kann als repräsentatives Modell zur Systemplanung angesehen werden.

Error! No topic specified.

Abbildung 1 – Anlagenkapazität von Datacentern und Bedarf über den Nutzungszeitraum

Die Abbildung zeigt einen typischen Planungszyklus. Die installierte Kapazität der vorhandenen Strom- und Kühlanlagen entspricht der Raumkapazität. Das heißt, dass das System von Anfang an vollständig ausgebaut ist. Laut Planung beträgt die anfängliche erwartete Last des Datacenters oder Netzwerkraums 30 %; diese steigt im Laufe der Zeit auf einen Maximalwert an. Die anfängliche Istlast liegt jedoch unter dem Wert für die anfängliche erwartete Last und steigt auf einen Höchstwert an, der weit unter der installierten Kapazität liegt. (Die Nennkapazität der eigentlichen installierten Anlage kann dabei aufgrund von Redundanz oder vom Benutzer gewünschten Wertebereichen größer als die installierte Kapazität sein.)

Daten aus vorhandenen Installationen

APC hat Daten von einer Vielzahl an Kunden zusammengetragen, um das tatsächliche Ausmaß der Überdimensionierung bei bereits bestehenden Installationen zu ermitteln. Diese Daten wurden anhand von Analysen bestehender Installationen sowie durch Befragungen unserer Kunden gewonnen. Dabei zeigte sich, dass die erwartete Anfangslast in der Regel 30 % der erwarteten Maximallast beträgt. Außerdem stellte sich

heraus, dass die anfängliche Istlast in der Regel 30 % der anfänglichen erwarteten Last beträgt; die maximale Istlast liegt zudem in der Regel bei circa 30 % der installierten Kapazität. Diese Daten werden in Abbildung 1 zusammengefasst. Durchschnittliche Datacenter werden damit letztendlich um das Dreifache größer als erforderlich ausgeführt. Zum Zeitpunkt der Einrichtung sind viele Datacenter häufig sogar um das Zehnfache überdimensioniert.

Durch Überdimensionierung bedingte Zusatzkosten

Die durch die Überdimensionierung bedingten Kosten über die Nutzungsdauer können in zwei Kategorien unterteilt werden: Kapitalkosten und Betriebskosten.

Die kapitalgebundenen Zusatzkosten werden durch den schattierten Bereich in Abbildung 1 reflektiert. Bei diesem schattierten Bereich handelt es sich um den Teil der Systemkapazität, die in einer durchschnittlichen Installation ungenutzt bleibt. Die überschüssige Kapazität lässt sich direkt in überschüssige Kapitalkosten umwandeln. Zu den zusätzlichen Kapitalkosten zählen die Kosten für überschüssige Strom- und Kühlanlagen sowie die zusätzlichen Anlagen- und Installationsaufwendungen, darunter auch Kabel- und Leitungssysteme.

Für die Strom- und Kühlsysteme in einem herkömmlichen 100 kW-Datacenter fallen im Schnitt Kapitalkosten von 500.000 € an; das ergibt 5 € pro Watt. Diese Analyse verdeutlicht, dass 70 % bzw. 350.000 € dieser Investitionskosten verschwendet werden. In den ersten Jahren ist diese Verschwendung sogar noch größer. Wenn die Zeitkosten in die Rechnung einbezogen werden, beträgt der übliche durch Überdimensionierung bedingte Verlust nahezu 100 % der Gesamtkapitalkosten. Das heißt, dass sich alleine durch die Zinsen des Ursprungskapitals die tatsächlichen Kapazitätsanforderungen bezahlen ließen.

In den zusätzlichen Kosten über die Nutzungsdauer, die durch Überdimensionierung entstehen, sind auch die Ausgaben zum Betrieb der Einrichtung enthalten. Diese Kosten umfassen Wartungsverträge, Verbrauchsmaterial und Strom. Die Wartungskosten liegen im Regelfall ein wenig unter den Kapitalkosten über die Nutzungsdauer eines Datacenters oder Netzwerkraums, wenn die Geräte gemäß den Anweisungen des Herstellers gewartet werden. Da aufgrund der Überdimensionierung auch die nur teilweise genutzten Geräte gewartet werden müssen, ist ein Großteil der Wartungskosten überflüssig. In unserem Beispiel des 100 kW-Datacenters belaufen sich diese überflüssigen Kosten im Laufe der Nutzungsdauer des Systems auf 250.000 €.

Zudem fallen überschüssige Energiekosten an, wenn die Datacenter oder Netzwerkräume zu großzügig dimensioniert werden. Die durch Standby-Kosten bedingte Verlustrate eines Stromversorgungssystems für ein Datacenter oder einen Netzwerkraum beträgt im Schnitt 5 % des Strombedarfs. Wenn die Kosten für die Kühlungssysteme in die Rechnung einbezogen werden, erhöht sich diese Rate auf 10 %. Bei einem 100 kW-Datacenter mit durchschnittlicher Überdimensionierung akkumuliert sich die verschwendete elektrische Energie

über die Systemnutzungsdauer von 10 Jahren auf einen Gesamtwert von 600.000 kWh, was einer Summe von 55.000 € entspricht.

Die überschüssigen Gesamtkosten über die gesamte Nutzungsdauer eines Datacenters oder Netzwerkraums betragen im Schnitt 70 % der Infrastrukturkosten für die Strom- und Kühlanlagen. Diese Kosten könnten theoretisch aufgefangen werden, wenn die Infrastruktur an die tatsächlichen Anforderungen angepasst würde.

Für viele Unternehmen führt die Verschwendung von Kapital und Geld zu verlorenen Alternativkosten, die um ein Vielfaches über den Baraufwendungen liegen können. So haben beispielsweise Internethost-Unternehmen durch Kapitalbindung in einer ungenutzten Installation keine Möglichkeit, Installationen in anderen Anlagen vorzunehmen.

Wodurch entsteht Überdimensionierung?

Die Daten zeigen, dass in bestehenden Installationen eine hohe und recht variable Überdimensionierung der Infrastruktur von Datacentern und Netzwerkräumen anzutreffen ist. In diesem Zusammenhang stellt sich automatisch die Frage, ob diese Überdimensionierung geplant und erwartet wird, ob sie vielmehr durch fehlerhafte Planung bedingt ist, oder ob es für dieses Phänomen handfeste Gründe gibt.

Geplante Überdimensionierung

Befragungen der Verantwortlichen üblicher Installationen haben ergeben, dass Datacenter so geplant werden, dass sie den künftig erwarteten maximalen Anforderungen entsprechen. Die Raumkapazität und die installierte Kapazität werden dabei absichtlich etwas höher als die erwartete Maximallast angesetzt. Viele Kunden wenden ein Standardverfahren an, in dem sie das Stromversorgungssystem niedriger kalkulieren und nur einen Bruchteil der vollen Kapazität ausnutzen, beispielsweise 80 %; diese Vorgehensweise ist durch die Annahme bedingt, dass die Zuverlässigkeit des Systems auf ein Maximum erhöht wird, wenn es nicht vollständig ausgelastet ist.

Die Praxis, die installierte Kapazität bei Datacentern größer als die erwartete Maximallast anzusetzen, wird in Abbildung 1 wiedergegeben. Hierbei handelt es sich um eine absichtliche Form von Überdimensionierung. Bei dieser Art von Überdimensionierung handelt es sich zwar um eine Art Unterauslastung, sie ist aber nicht der Hauptfaktor bei den Gesamtsatzkosten.

Planungsprozess und Mängel

In den Planungsprozess für Datacenter und Netzwerkräume fließt eine Reihe von Annahmen in Bezug auf künftige Anforderungen ein. Folgende Überlegungen werden getroffen:

- ? Die Kosten, die durch unzureichende Kapazität im Datacenter oder Netzwerkraum entstehen, sind sehr hoch und müssen vermieden werden.

- ? Eine Vergrößerung der Kapazität im Laufe der Nutzungsdauer ist sehr kostenintensiv.
- ? Die Arbeit, die mit der Kapazitätsvergrößerung während der Nutzungsdauer verbunden ist, kann Ausfallzeiten verursachen, die ein großes und unzumutbares Risiko darstellen.
- ? Alle Konstruktions- und Planungsarbeiten für die maximale Kapazität müssen im Vorfeld durchgeführt werden.
- ? Die benötigte Last des Datacenters oder Netzwerkraums erhöht sich im Laufe der Zeit, lässt sich jedoch nicht zuverlässig vorkalkulieren.

Das Ergebnis dieser Annahmen ist, dass Datacenter bzw. Netzwerkräume so geplant, konstruiert und ausgebaut werden, dass sie einen unbekanntem Bedarf abdecken. Zudem wird die Kapazität von Datacentern oder Netzwerkräumen so eingeplant, dass sie in realistischen Wachstumsszenarien traditionell eher höher tendiert.

Begründete Ursachen für die Überdimensionierung

Der Planungsprozess führt zu Plänen, die im Allgemeinen eine ziemlich schlechte Auslastung liefern, wie die Ergebnisse zeigen. Aus betriebswirtschaftlicher Sicht muss dies als Fehlschlag betrachtet werden. Allerdings führt die oben beschriebene Analyse des Planungsprozesses nicht zu schwer wiegenden Mängeln. Dieser scheinbare Widerspruch kann durch eine genauere Analyse der Daten und der Prozesseinschränkungen beigelegt werden. Abbildung 2 zeigt die Verteilung der maximalen Auslastung vorhandener Installationen, d. h. die maximale Istlast geteilt durch die maximale installierte Kapazität. Die Datenanalyse führt zu folgenden Erkenntnissen:

- ? Der erwartete Prozentsatz für die tatsächliche Auslastung beträgt circa 30 %
- ? Der erwartete Wert für zusätzliche bzw. überflüssige Kapazität beträgt 70 %
- ? Der Istwert der Auslastung weist hohe Schwankungen auf. Dadurch ist eine Vorhersage während der Entwurfsphase kaum möglich.
- ? Wenn die installierte Kapazität statt auf die gängigen Werte generell auf den erwarteten Wert von 30 % gesetzt würde, könnten 50 % der Systeme im Laufe der Nutzungsdauer nicht den Wert für die benötigte Last erreichen.
- ? Das derzeitige Dimensionierungsverfahren ist ein Kompromiss, bei dem sich überdimensionierte Systeme gegen die hohen Schwankungen der maximalen Istlast absichern, indem die Wahrscheinlichkeit verringert wird, dass die erforderliche Last während der Nutzungsdauer nicht erreicht werden kann.

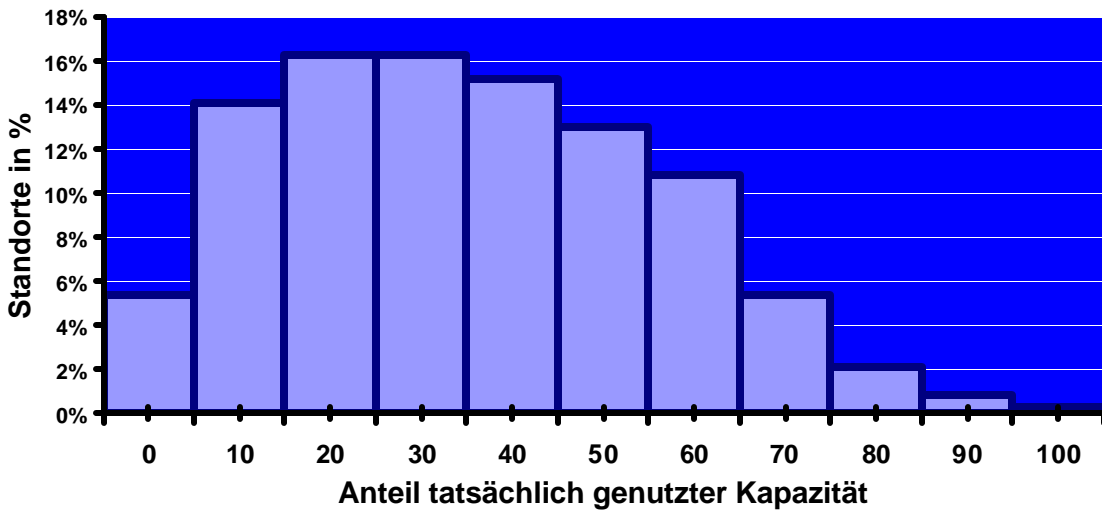


Abbildung 2 – Maximaler Auslastungsanteil von Datacentern

Die Analyse führt zu dem überraschenden Ergebnis, dass durch die Entwurfsbeschränkungen und die unvorhersehbaren künftigen Kapazitätsanforderungen die derzeitige Planungsmethode für Datacenter und Netzwerkräume durchaus logisch ist. Wenn die Unterdimensionierung von Datacentern und Netzwerkräumen für die Unternehmen zu hohen Kosten führt, ist eine deutliche Überdimensionierung die beste Möglichkeit zur Reduzierung der erwarteten Systemgesamtkosten.

Architektur und Methode zur Vermeidung der Überdimensionierung

Da sich künftige Bedarfe während des Planungsprozesses äußerst schwer vorhersagen lassen, ist die Planung von Datacentern und Netzwerkräumen eine große Herausforderung. Vor diesem Hintergrund muss versucht werden, eine Datacenter- und Netzwerkraum-Infrastruktur zu erstellen, die dem unvorhersehbaren Bedarf Rechnung trägt.

Einschränkungen der Anpassungsfähigkeit

Nach der Betrachtung des Ausmaßes dieser Überdimensionierung stellt sich automatisch folgende Frage: Warum wird die Infrastruktur von Datacentern und Netzwerkräumen vorab ausgebaut und nicht so, dass sie mit den tatsächlichen Lastanforderungen dynamisch Schritt hält?

Viele Datacenter werden in der Tat schrittweise vergrößert. So werden beispielsweise Geräteschränke häufig nach und nach installiert. Auch die Installation der Stromverteilung erfolgt in vielen Fällen in mehreren Phasen. In

einigen Fällen kann ein redundantes USV-Modul schrittweise installiert werden. Durch diese Maßnahmen können Kosten über die gesamte Nutzungsdauer eingespart werden. Allerdings ist die nachträgliche Installation dieser Geräte um einiges kostenintensiver als die anfängliche Installation, so dass sich die meisten Planungsbeauftragten für eine direkte Komplettinstallation entscheiden. Aus diesem Grund wird in der Praxis nur ein Bruchteil der möglichen Kosteneinsparungen erzielt.

Methode und Ansatz zum Erstellen einer anpassbaren Infrastruktur

Es wäre ideal, eine Methode und Architektur zu verwenden, die sich kontinuierlich an die veränderten Anforderungen anpasst. Eine derartige Methode und Architektur hätte folgende Merkmale:

- ? Die einmalige Konstruktion wäre stark eingeschränkt oder würde entfallen
- ? Die strombezogene Infrastruktur würde aus vorgefertigten Modulen bestehen
- ? Die Komponenten passen durch normale Türen und in herkömmliche Personenfahrstühle und ließen sich einfach und ohne Arbeiten an Strom führenden Leitungen anschließen
- ? Spezielle Vorbereitungsarbeiten an den Räumlichkeiten (z. B. Einziehen von Zwischenböden) würden entfallen
- ? Das System könnte ohne Modifikationen in N-, N+1- oder 2N-Konfiguration arbeiten
- ? Installationsarbeiten wie Verkabeln, Bohren und Schneiden würden entfallen
- ? Zur Erhöhung der Kapazität wären keine speziellen Genehmigungs- oder Überwachungsverfahren erforderlich
- ? Die Gerätekosten für das modulare Stromsystem wären dieselben wie für herkömmliche zentralisierte Systeme oder lägen darunter
- ? Die Wartungskosten für das modulare Stromsystem wären dieselben wie für herkömmliche zentralisierte Systeme oder lägen darunter.

Praktische und erreichbare Anpassungsebenen

Wenn ein anpassbares System für die räumliche Infrastruktur verwendet wird, kann die durch Überdimensionierung verursachte Verschwendung, die in Abbildung 1 dargestellt ist, deutlich reduziert werden. Diese Einsparungen werden in Abbildung 3 unten gezeigt. Die installierte Kapazität ist dabei anfangs nicht auf die Raumkapazität ausgelegt, und die installierte Kapazität wird zur Anpassung an die Istlast geändert.

Error! No topic specified.

Abbildung 3 – Anlagenkapazität von Datacentern und Bedarf über den Nutzungszeitraum

Ein Beispiel für ein anpassbares System, das die oben dargestellten Anforderungen erfüllt, ist die APC InfraStruXure-Architektur. An dieser Stelle wird dieses System nicht in aller Ausführlichkeit beschrieben. In der

? 2003 American Power Conversion. Alle Rechte vorbehalten. Ohne die ausdrückliche schriftliche Erlaubnis des Urheberrechtsinhabers darf kein Teil dieser Veröffentlichung für irgendwelche Zwecke verwendet, vervielfältigt oder in einem Datenempfangssystem gespeichert oder darin eingesehen werden, unabhängig davon, auf welche Weise und mit welchen Mitteln dies geschieht. www.apc.com

Version 2002-4

InfraStruXure-Architektur können über 70 % des Stromversorgungssystems so bereitgestellt werden, dass die erhöhten Anforderungen des Datacenters oder Netzwerkraums genau verfolgt werden. In der Praxis wird anfänglich nur die Niederspannungs-Hauptverteilung (NSHV) installiert, die so dimensioniert ist, dass sie der maximalen Raumkapazität entspricht. USV, Batteriesystem, Stromverteiler, Bypass-Schaltgeräte und Rackverkabelung werden im Laufe der Zeit gemäß der variablen Last installiert.

In diesem Zusammenhang soll darauf hingewiesen werden, dass sich diese Ausführungen auf die Strom- und Kühlanlagen konzentrieren, die einen wesentlichen Kostenfaktor für Datacenter und Netzwerkräume ausmachen. Die gleiche Analyse kann und muss erweitert werden, um den physischen Platzbedarf, Brandschutzanforderungen und Sicherheitsbestimmungen zu ermitteln.

Ergebnisse

Datacenter und Netzwerkräume werden nach gängiger Praxis dreimal größer ausgeführt, als es die benötigte Kapazität erfordert. Durch die Überdimensionierung entstehen übermäßige Kapital- und Wartungskosten, die einen Großteil der Gesamtkosten über die Nutzungsdauer ausmachen. Die meisten dieser zusätzlichen Kosten können aufgefangen werden, indem eine Methode und Architektur angewendet wird, die sich auf kostenwirksame Weise an die veränderlichen Anforderungen anpassen und gleichzeitig hohe Verfügbarkeit bieten.

Quellenangaben

Mitchell-Jackson, J.D., Koomey, J.G., Nordman, B., Blazek, M., „Data Center Power Requirements: Measurements From Silicon Valley“, 16. Mai 2001. Akademische Abschlussarbeit, Energy and Resources Group, University of California. Berkeley, Kalifornien. Abrufbar unter <http://enduse.lbl.gov/Projects/InfoTech.htm>