Die treibende Kraft hinter Edge Computing und die Vorteile der Technologie

White Paper 226

Version 0

von Steven Carlini

Zusammenfassung

Die Internetnutzung entwickelt sich zunehmend hin zu bandbreitenintensiven Inhalten und vielen verbundenen Geräten. Gleichzeitig nutzen immer mehr mobile Kommunikations- und Datennetze eine Cloud-Computing-Architektur. Um diesen neuen Anforderungen heute und in Zukunft gerecht zu werden, werden Rechenleistung und Speicher an den Netzwerkrand verschoben, um die Datenübertragung zu beschleunigen und die Verfügbarkeit zu optimieren. Edge Computing bringt bandbreitenintensive Anwendungen, die möglichst geringe Latenzen bieten müssen, näher an den Benutzer bzw. die Datenquelle. Dieses White Paper erläutert die treibende Kraft hinter Edge Computing genauer und erforscht die verschiedenen verfügbaren Arten von Edge Computing.

APC by Schneider Electric-White Papers sind nun Teil der vom **Data Center Science Center** von Schneider Electric geschaffenen White Paper-Bibliothek.
DCSC@Schneider-Electric.com

Schneider Electric

Definition von Edge Computing

Beim Edge Computing werden Funktionen zur Datenerfassung und -kontrolle, die Speicherung bandbreitenintensiver Inhalte sowie Anwendungen näher an den Endbenutzer gebracht. Sie werden an einem logischen Endpunkt des entsprechenden Netzwerks (Internet oder privat) positioniert, wo sie einen Teil einer größeren Cloud-Computing-Architektur bilden.

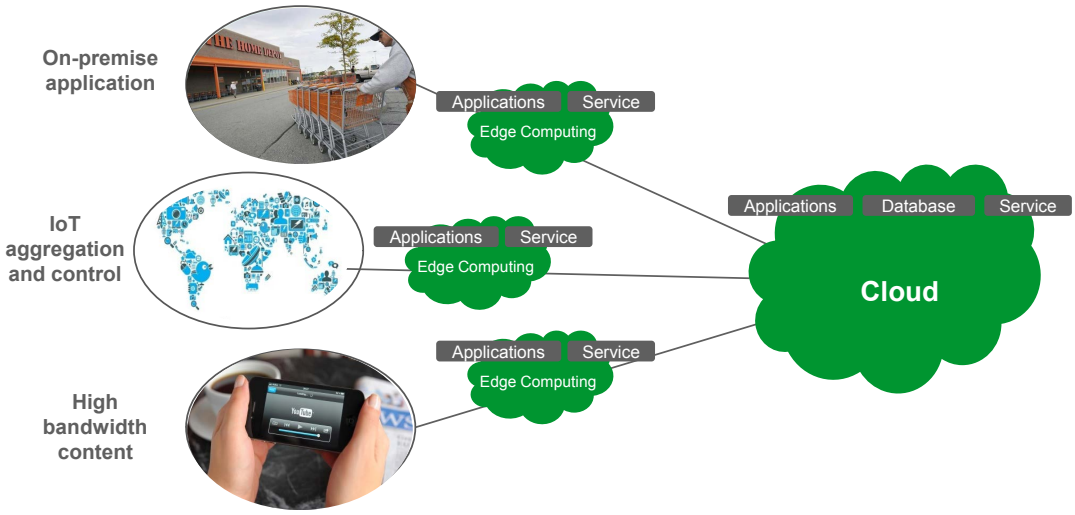


Abbildung 1

Grundlegendes Diagramm zu Cloud Computing mit Edge-Geräten

In diesem White Paper stellen wir Ihnen drei Anwendungsbereiche vor, in denen Edge Computing primär eingesetzt wird.

1. Als Tool, um Informationen aus lokalen Geräten zentral zu erfassen und zu kontrollieren.
2. Zur lokalen Speicherung und Bereitstellung bandbreitenintensiver Inhalte im Rahmen eines Content Distribution Networks (CDN).
3. Als lokales Anwendungs- und Verarbeitungstool zur Replikation von Cloud-Services und zur Isolation des Datacenters von der Public Cloud.

Bevor wir jedoch die Anwendungen und Lösungen besprechen, definieren wir zunächst, wie Netzwerke und das Internet funktionieren.

Funktionsweise des Internets

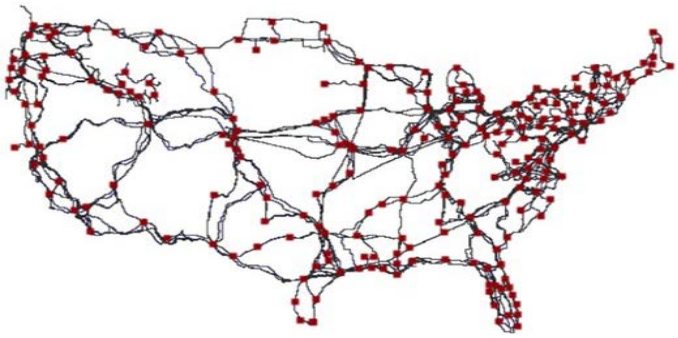
„Horizontale“ Datenübertragung

Quelldaten werden in Pakete umgewandelt, die dann mithilfe des Netzwerkprotokolls „IP“ (Internet Protocol) über ein Netzwerk übertragen werden. Das Routing im Internet wird von einem anderen Protokoll übernommen: dem sogenannten Border Gateway Protocol (BGP). Das Internet wurde so entwickelt, dass es große Ausfälle abfangen kann, in dem einfach eine Route um Problembereiche herum gefunden wird. BGP berücksichtigt beim Daten-Routing nicht den Zeitfaktor. Es untersucht lediglich die Anzahl von Hops zwischen zwei Netzwerken, die miteinander kommunizieren wollen. Diese Hops sind jedoch möglicherweise bereits überlastet oder beim Routing wird eine physisch längere Route mit weniger Hops anstelle einer sehr kurzen Route mit vielen Hops ausgewählt. **Abbildung 2** zeigt eine Karte der Langstrecken-Hops in den USA[1] Zwar bietet das BGP eine hohe Zuverlässigkeit und stellt eine grundlegende Technologie für das Internet dar, jedoch ist es bezüglich Latenzen (Verzögerungen, Jitter und Ruckeln) und Performance nicht die optimale Lösung.

[1] http://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p565.pdf

Abbildung 2

Karte verschiedener Netzwerk-Hops in den USA

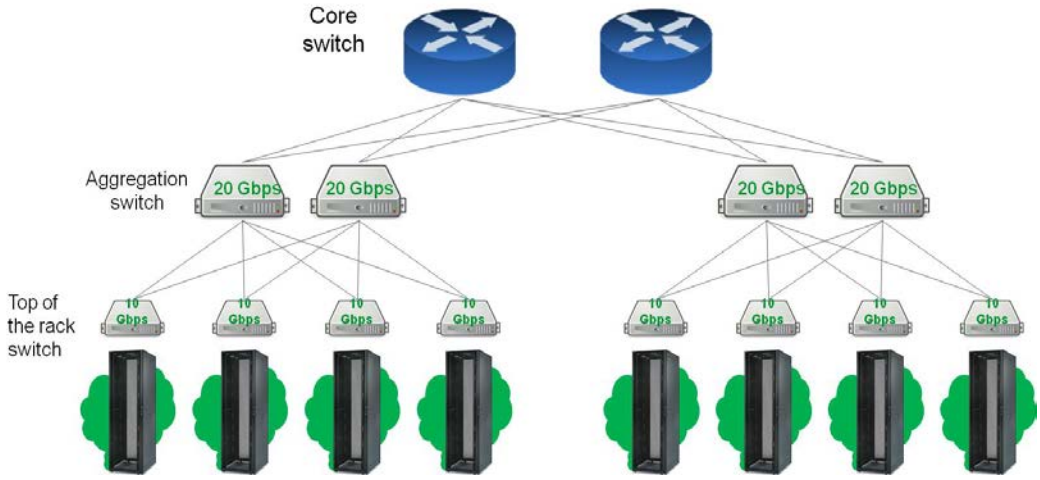


„Vertikale“ Datenübertragung

Wie in **Abbildung 3** veranschaulicht, starten Daten, die aus dem Inneren eines typischen Cloud-Datacenter-Netzwerks nach außen übertragen werden, bei einer physischen Serverschnittstelle und durchlaufen dann ToR- (Top of Rack) oder EoR-Switches (End of Rack). Von jedem ToR-Switch gehen Daten an einen Aggregations-Switch und die Aggregations-Switches leiten Daten durch einen Core-Switch, der den primären Eingabe- und Ausgabepunkt des Datacenters bildet. Jeder dieser Switches überträgt Daten und wird als Netzwerk-Hop angesehen – mitsamt entsprechender Verlangsamung der Daten und möglicher Netzwerküberlastung. Wenn eine Überlastung auf einer Netzwerkebene vorliegt (wenn also nicht genügend Bandbreite für Spitzenauslastung bereitsteht), kann es während dieser Zeiträume hoher Belastung zu weiteren Verzögerungen kommen.

Abbildung 3

Datacenter-Netzwerk



Anwendung Nr. 1: Verteilung bandbreitenintensiver Inhalte

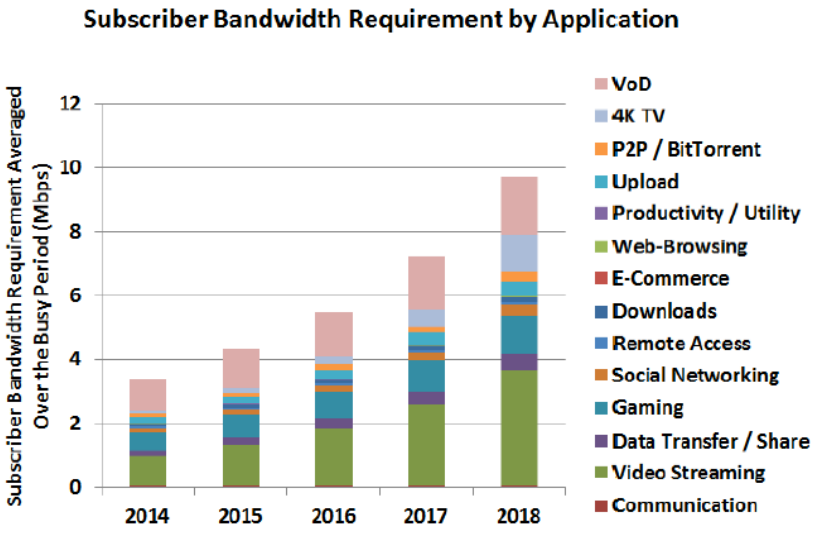
Latenz beschreibt die Zeit zwischen dem Moment, in dem ein Datenpaket übermittelt wird, und dem Moment, in dem es beim Ziel ankommt (One-Way) und wieder zurückkehrt (Roundtrip). Selbst wenn die meisten Daten nur in die eine Richtung übertragen werden, ist dieser Wert nahezu unmöglich zu messen. Deshalb ist die Roundtrip-Zeit von einem einzigen Punkt aus der am häufigsten genutzte Messwert für Latenzen. Roundtrip-Latenzen von weniger als 100 Millisekunden sind üblich – das Ziel sind meist weniger als 25 Millisekunden.

Bandbreite bezieht sich auf die Geschwindigkeit der Datenübertragung im Netzwerk. Höchstgeschwindigkeiten von Netzwerkgeräten werden von ihren Herstellern angegeben. Die tatsächlich im jeweiligen Netzwerk erreichte Geschwindigkeit liegt jedoch immer niedriger als dieser Spitzenwert. Übermäßige Latenzen schaffen Datenstaus, durch die nicht die gesamte Netzwerkkapazität genutzt werden kann. Die Auswirkungen von Latenzen auf die Netzwerkbandbreite können temporär auftreten und den Verkehr wie eine Ampel nur wenige Sekunden behindern oder sie sind von Dauer wie eine nur einspurig befahrbare Brücke. Die wahrscheinlichste Ursache für Netzwerküberlastung sind bandbreitenintensive Videoinhalte. Wie in **Abbildung 4** veranschaulicht, zählen VoD, 4K-Fernsehen und Video-Streaming zu den am schnellsten wachsenden bandbreitenintensiven Anwendungen[2].

[2] ACG Research, The value of content at the edge, 2015, S. 4

Abbildung 4

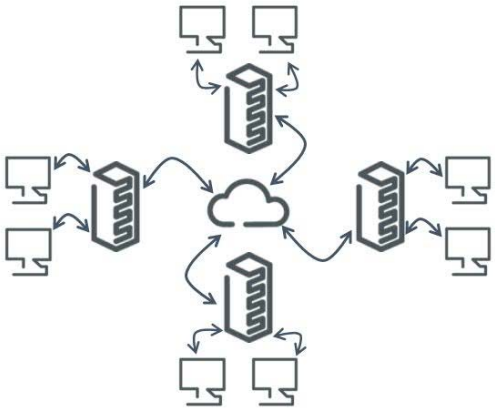
Zunahme bandbreitenintensiver Anwendungen



Um die Netzwerkbelastung zu minimieren und so das Streamen bandbreitenintensiver Inhalte heute und in Zukunft zu optimieren, nutzen Anbieter ein System aus Computern mit Internetverbindung, das Inhalte näher am Benutzer zwischenspeichert. So können die Inhalte schnell an verschiedene Benutzer übertragen werden, indem Inhalte auf verschiedenen Servern dupliziert und den Benutzern je nach Nähe bereitgestellt werden. Diese Computer, die Inhalte zwischenspeichern, stellen ein Beispiel für Edge Computing dar (**Abbildung 5**).

Abbildung 5

Einfaches CDN-Diagramm (Content Distribution Network)



Anwendung Nr. 2: Edge Computing als IoT-Aggregations- und Kontrollpunkt

Die Technologien, die alles „smart“ vernetzen – Städte, Landwirtschaft, Fahrzeuge, Gesundheit usw. – werden in Zukunft die Implementierung einer Vielzahl von IoT-Sensoren (Internet of Things) erfordern. Ein IoT-Sensor wird als Nicht-Computer-Knoten (oder -Objekt) mit einer IP-Adresse definiert, der mit dem Internet verbunden ist.

Durch die stetig sinkenden Preise für entsprechende Sensoren wird die Anzahl verbundener IoT-Geräte weiter rapide ansteigen. Cisco schätzt, dass bis 2020 50 Milliarden IoT-Geräte mit dem Internet verbunden sein werden[3]. IoT kann Vorgänge folgendermaßen automatisieren:

- Durch automatische Erfassung von Informationen zu physischen Assets (Maschinen, Geräten, Anlagen, Fahrzeugen) zur Überwachung des Status oder Verhaltens
- Durch Nutzung dieser Informationen, um Transparenz zu schaffen, die Kontrolle zu gewährleisten und so Prozesse und Ressourcennutzung zu optimieren

[3] Dave Evans, The Internet of Things: How the Next Evolution of the Internet Is Changing Everything, Cisco Internet Business Solutions Group, S. 3

M2M (Machine to Machine) bezieht sich auf Technologien, die es kabellosen und -gebundenen Systemen ermöglichen, mit anderen Geräten desselben Typs zu kommunizieren. M2M bildet einen essenziellen Bestandteil des Internet of Things und birgt dank der vielseitigen Einsatzmöglichkeiten in Smart Citys verschiedene Vorteile für Industrie und andere Branchen.

Das Industrial Internet of things (IIoT), das die Nutzung der Sensordaten, die Kontrolle der M2M-Kommunikation sowie Automatisierungstechnologien umfasst, sorgt für große Mengen an Daten und Netzwerkverkehr. Unternehmenseigene industrielle IT-Systeme und Netzwerktechnologien werden zu kommerziellen Standard-IT-Systemen migriert, die über IP-Netzwerke (Internet Protocol) kommunizieren.

Die Öl- und Gasgewinnung ist ein gutes Beispiel für diese IIoT-Anwendung. Fliegende Drohnen, die aus der Luft Daten erfassen, um die Einsatzorte bei der Ölexploration zu untersuchen, generieren große Datenmengen in Form von HD-Videos. An diesen Einsatzorten gestaltet sich die Koordination der Flotten aus riesigen Lastwagen, Kränen und Baggern schwierig. Bei älteren Methoden zur Verkehrsregelung kamen unbemannte Helikopter für die Videoüberwachung zum Einsatz. Selbststeuernde Drohnen können Einsatzorte 24 Stunden täglich aufnehmen und den Verantwortlichen so immer eine topaktuelle Übersicht des Ressourceneinsatzes bieten. Durch den Einsatz von Edge Computing können die Drohnen Daten in Echtzeit übermitteln und schnell Anweisungen entgegennehmen.

Abbildung 6

Öl- und Gasgewinnung: Drohnen sammeln Unmengen von Daten auf Ölfeldern und nutzen Edge Computing, um Datenübertragung und Steuerbefehle in Echtzeit zu ermöglichen.

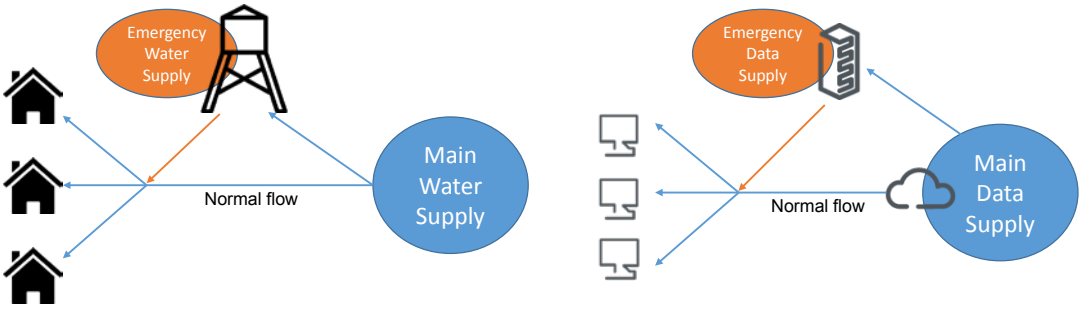
Anwendung Nr. 3: Lokale Anwendungen

Die Gewährleistung bzw. die Optimierung der Verfügbarkeit von IT und Netzwerken hat fast immer höchste Priorität. Cloud Computing nutzt seit jeher eine zentralisierte Architektur. Edge Computing sorgt hierbei für eine weiter verteilte Cloud-Computing-Architektur. Der Hauptvorteil besteht darin, dass jegliche Unterbrechungen auf einen einzelnen Punkt im Netzwerk beschränkt sind und nicht das gesamte Netzwerk betreffen. Ein DDoS-Angriff (Distributed Denial of Service) oder ein langfristiger Ausfall beschränkt sich nur auf das Edge-Computing-Gerät und dessen lokale Anwendungen – und wirkt sich nicht auf alle Anwendungen im zentralen Cloud-Datacenter aus.

Unternehmen, die ihre Systeme zu externen Cloud-Computing-Architekturen migriert haben, können so die Vorteile des Edge Computing nutzen, um Redundanz und Verfügbarkeit zu steigern. Geschäftskritische Anwendungen bzw. Anwendungen, die für den Betrieb wichtiger Funktionen des Unternehmens erforderlich sind, können lokal dupliziert werden. Stellen Sie sich eine kleine Stadt vor, die als Hauptquelle eine sehr große gemeinsame Wasserversorgung nutzt (siehe **Abbildung 7**). Sollte diese Wasserversorgung durch einen Ausfall der Hauptversorgung oder des Verteilungsnetzwerks unterbrochen werden, gibt es noch einen Notfalltank in der Stadt.

Abbildung 7

Städtisches Wasserversorgungssystem zur Veranschaulichung von Edge Computing

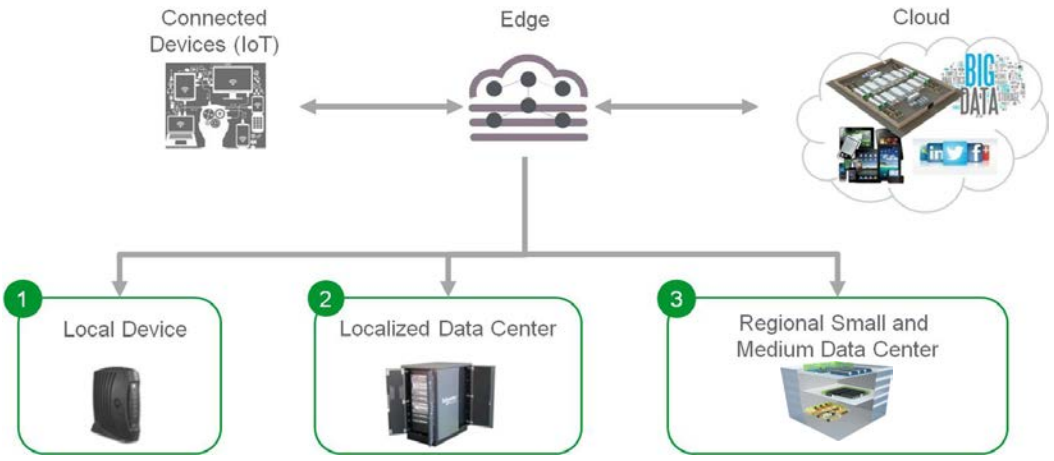


Arten von Edge Computing

Im Allgemeinen gibt es drei Arten von Edge Computing (siehe **Abbildung 8**).

Abbildung 8

Arten von Edge Computing



Lokale Geräte:
Geräte, die einen bestimmten definierten Zweck erfüllen sollen. Die Implementierung erfolgt umgehend und sie eignen sich für Heim- oder Kleinbüros. Beispiele hierfür ist der Betrieb des Sicherheitssystems eines Gebäudes (Intel SOC-Appliance) oder die Speicherung lokaler Videoinhalte auf einem digitalen Videorekorder. Ein weiteres Beispiel sind Cloud-Speicher-Gateways. Hierbei handelt es sich um lokale Geräte, die für gewöhnlich als Netzwerk-Appliance oder -server eingesetzt werden, die Cloud-Speicher-APIs, wie z. B. SOAP oder REST, übersetzen. Cloud-Speicher-Gateways ermöglichen es Benutzern, Cloud-Speicher in Anwendungen zu integrieren, ohne die Anwendungen selbst in die Cloud zu verschieben.

Lokalisierte Datacenter (1 bis 10 Racks):
Diese Datacenter bieten erhebliche Verarbeitungs- und Speicherkapazitäten und lassen sich schnell in bestehende Umgebungen implementieren. Diese Datacenter sind oft als vorkonfigurierte Systeme (Configure to Order, CTO) verfügbar, die vorgefertigt und dann vor Ort eingebaut werden (siehe links: **Abbildung 9**). Eine andere Form lokalisierter Datacenter sind vorgefertigte Micro-Datacenter, die in einer Fabrik montiert und dann vor Ort abgeliefert werden (siehe rechts: **Abbildung 9**). Diese Ein-Gehäuse-Systeme können in robusten Gehäusen – samt Feuchte-, Rost-, Brandschutz usw. – oder in normalen IT-Gehäusen für Büros verbaut werden. Die Ein-Rack-Versionen können vorhandene Infrastrukturen, Kühlsysteme und Energieversorgung nutzen und reduzieren so die Investitionskosten für neue spezielle Standorte. Für die Installation ist die Auswahl eines Standorts nahe an der Energieversorgung und Glasfaserleitung des Gebäudes erforderlich. Die Multi-Rack-Versionen bieten dank ihrer Größe mehr Leistung und Flexibilität, erfordern jedoch mehr Planungs- und Installationsaufwand und benötigen ein eigenes Kühlsystem. Diese 1-bis-10-Rack-Systeme eignen sich für eine Vielzahl von Szenarien, in denen geringe Latenz, hohe Bandbreite und/oder gesteigerte Sicherheit oder Verfügbarkeit erforderlich sind.

Abbildung 9

Beispiel für ein vorkonfiguriertes (links) und ein vorgefertigtes Micro-Datacenter (rechts)

Regionales Datacenter:
Datacenter, die über mehr als 10 Racks verfügen und sich näher am Benutzer und der Datenquelle befinden als zentrale Cloud-Datacenter, werden als „regionale Datacenter“ bezeichnet. Aufgrund ihrer Skalierung verfügen sie über mehr Verarbeitungs- und Speicherkapazität als lokalisierte Datacenter mit 1 bis 10 Racks. Selbst wenn sie vorgefertigt sind, dauert der Aufbau aufgrund von wahrscheinlich erforderlichen Bauarbeiten und Zulassungen sowie von lokalen Compliance-Problemen länger als bei lokalisierten Datacentern. Darüber hinaus benötigen sie eine eigene Energieversorgung sowie eigene Kühlsysteme. Latenzen sind von der physischen Nähe zu Benutzern und Daten sowie von der Anzahl der Hops zwischen ihnen abhängig.

Fazit

Edge Computing kann Latenzprobleme lösen und es Unternehmen ermöglichen, ihre Chancen mithilfe einer Cloud-Computing-Architektur besser zu nutzen. Aus bandbreitenintensivem Video-Streaming generierte Workloads verursachen Netzwerküberlastung und Latenzen. Edge-Datacenter bringen bandbreitenintensive Inhalte näher an den Benutzer und Anwendungen, die möglichst geringe Latenzen bieten müssen, näher an die Daten. Die Rechenleistung und Speicherkapazitäten werden direkt am Netzwerkrand integriert, um Übertragungszeiten und Verfügbarkeit zu optimieren. Die verschiedenen Arten von Edge Computing beinhalten lokale Geräte, lokalisierte Datacenter und regionale Datacenter. Die Art, die die Bereitstellungsgeschwindigkeit und die Kapazität bietet, die für künftige IoT-Anwendungen erforderlich sein werden, sind die lokalisierten 1-bis-10-Rack-Versionen. Diese können schnell und einfach entworfen und bereitgestellt werden: in vorkonfigurierten oder vorgefertigten Varianten.

Über den Autor

Steven Carlini ist Director of Marketing for Data Center Solutions bei Schneider Electric. Im Laufe seiner Karriere arbeitete er an einigen der innovativsten Lösungen, die die Datacenter-Landschaft und -Architektur verändert haben. Er hat einen BSEE-Abschluss von der University of Oklahoma und einen MBA in International Business von der University of Houston. Er ist anerkannter Experte auf dem Gebiet und häufig Podiumsgast und Sprecher bei Branchen-Events.

Ressourcen

Cost Advantages of Using Single-Rack Micro Data Centers
White Paper 223

Practical Options for Deploying Small Server Rooms and Micro Data Centers
White Paper 174

Alle White Paper anzeigen
whitepapers.apc.com

Alle TradeOff Tools™
tools.apc.com

Kontaktieren Sie uns

Rückmeldungen und Anmerkungen zum Inhalt dieses White Paper:

Data Center Science Center
dcsc@schneider-electric.com

Falls Sie Kunde sind und Fragen zu Ihrem spezifischen Datacenter-Projekt haben:

Wenden Sie sich an Ihre Schneider Electric-Vertretung unter
www.apc.com/support/contact/index.cfm