Cisco Global Cloud Index: Forecast and Methodology, 2015–2020

What You Will Learn

The Cisco® Global Cloud Index (GCI) is an ongoing effort to forecast the growth of global data center and cloud-based IP traffic. The forecast includes trends associated with data center virtualization and cloud computing. This document presents the details of the study and the methodology behind it.

Forecast Overview

Hyperscale Data Centers

• Hyperscale data centers will grow from 259 in number at the end of 2015 to 485 by 2020. They will represent 47 percent of all installed data center servers by 2020.

• Traffic within hyperscale data centers will quintuple by 2020. Hyperscale data centers already account for 34 percent of total traffic within all data centers and will account for 53 percent by 2020.

Global Data Center Traffic

• Annual global data center IP traffic will reach 15.3 zettabytes (ZB) (1.3 ZB per month) by the end of 2020, up from 4.7 ZB per year (390 exabytes [EB] per month) in 2015.

• Global data center IP traffic will grow 3-fold over the next 5 years. Overall, data center IP traffic will grow at a compound annual growth rate (CAGR) of 27 percent from 2015 to 2020.

Data Center Virtualization and Cloud Computing Growth

• By 2020, 92 percent of workloads will be processed by cloud data centers; 8 percent will be processed by traditional data centers.

• Overall data center workloads will more than double (2.6-fold) from 2015 to 2020; however, cloud workloads will more than triple (3.2-fold) over the same period.

• The workload density (that is, workloads per physical server) for cloud data centers was 7.3 in 2015 and will grow to 11.9 by 2020. Comparatively, for traditional data centers, workload density was 2.2 in 2015 and will grow to 3.5 by 2020.

Public vs. Private Cloud

• By 2020, 68 percent of the cloud workloads will be in public cloud data centers, up from 49 percent in 2015 (CAGR of 35 percent from 2015 to 2020).

• By 2020, 32 percent of the cloud workloads will be in private cloud data centers, down from 51 percent in 2015 (CAGR of 15 percent from 2015 to 2020).

Global Cloud Traffic

• Annual global cloud IP traffic will reach 14.1 ZB (1.2 ZB per month) by the end of 2020, up from 3.9 ZB per year (321 EB per month) in 2015.

• Global cloud IP traffic will almost quadruple (3.7-fold) over the next 5 years. Overall, cloud IP traffic will grow at a CAGR of 30 percent from 2015 to 2020.

• Global cloud IP traffic will account for more than 92 percent of total data center traffic by 2020.

Cloud Service Delivery Models

• By 2020, 74 percent of the total cloud workloads will be software-as-a-service (SaaS) workloads, up from 65 percent in 2015.

• By 2020, 17 percent of the total cloud workloads will be infrastructure-as-a-service (IaaS) workloads, down from 26 percent in 2015.

• By 2020, 8 percent of the total cloud workloads will be platform-as-a-service (PaaS) workloads, down from 9 percent in 2015.

Workloads by Application

• By 2020, enterprise workloads will account for 72 percent of total data center workloads, down from 79 percent in 2015.

• By 2020, consumer workloads will account for 28 percent of total data center workloads, up from 21 percent in 2015.

• Within the enterprise segment, compute (29 percent of enterprise workloads by 2020) and collaboration (24 percent of enterprise workloads by 2020) are the two main contributors to workload totals.

• Within the consumer segment, video streaming (34 percent of consumer workloads by 2020) and social networking (24 percent of consumer workloads by 2020) are the two main contributors to workload totals.

• Within the enterprise segment, database/analytics and IoT will be the fastest growing applications, with 22 percent CAGR from 2015 to 2020, or 2.7-fold growth.

- Within the consumer segment, social networking (33% CAGR from 2015 to 2020) and video streaming (32 percent CAGR from 2015 to 2020) will be the fastest growing applications.

Data Center Storage

- By 2020, data center storage installed capacity will grow to 1.8 ZB, up from 382 EB in 2015, nearly a 5-fold growth.
- By 2020, the total global installed data storage capacity in cloud data centers will account for 88 percent share of total DC storage capacity, up from 64.9 percent in 2015.

Data in Data Centers, Big Data, and IoE

- Globally, the data stored in data centers will quintuple by 2020 to reach 915 EB by 2020, up 5.3-fold (a CAGR of 40%) from 171 EB in 2015.
- Big data will reach 247 EB by 2020, up almost 10-fold from 25 EB in 2015. Big data alone will represent 27 percent of data stored in data centers by 2020, up from 15 percent in 2015.
- The amount of data stored on devices will be 5 times higher than data stored in data centers, at 5.3 ZB by 2020.
- Driven by the Internet of Things, the total amount of data created (and not necessarily stored) by any device will reach 600 ZB per year by 2020, up from 145 ZB per year in 2015. Data created is two orders of magnitude higher than data stored.

Consumer Cloud Storage

- By 2020, 59 percent (2.3 billion) of the consumer Internet population will use personal cloud storage, up from 47 percent (1.3 billion users) in 2015.
- Globally, consumer cloud storage traffic per user will be 1.7 gigabytes per month by 2020, compared to 513 megabytes per month in 2015.

Multiple-Device and -Connection Ownership

- North America (7.3), followed by Western Europe (5.5), had the highest average number of devices or connections per user in 2015, followed by Middle East and Africa (5.4), Latin America (4.7) Central and Eastern Europe (4.5), and Asia Pacific (4.5).
- By 2020, North America (13.6), followed by Western Europe (9.9), will have the highest average number of devices or connections per user, followed by Central and Eastern Europe (6.2), Latin America (5.2), Middle East and Africa (5.0), and Asia Pacific (5.0).

Regional Cloud Readiness

Network Speeds and Latency

- Asia Pacific leads all regions with an average fixed download speed of 33.9 Mbps. North America follows with an average fixed download speed of 32.9 Mbps. Central and Eastern Europe and Asia Pacific also lead all regions in average fixed upload speeds with 19.3 Mbps and 19.0 Mbps, respectively.
- Asia Pacific leads all regions in average fixed network latency with 26 ms, followed by Central and Eastern Europe with 30 ms.
- Asia Pacific leads all regions with an average mobile download speed of 18.5 Mbps. Western Europe follows with an average mobile download speed of 18.2 Mbps. North America and Asia Pacific lead all regions in average mobile upload speeds with 9.9 Mbps and 8.9 Mbps, respectively.
- Western Europe and Asia Pacific lead all regions in average mobile network latency with 57 ms and 73 ms, respectively.

Top Seven Data Center and Cloud Networking Trends

Over the last few years, the telecommunications industry has seen cloud adoption evolve from an emerging technology to an established networking solution that is gaining widespread acceptance and deployment. Enterprise and government organizations are moving from test environments to placing more of their mission-critical workloads into the cloud. For consumers, cloud services offer ubiquitous access to content and services, on multiple devices, delivered to almost anywhere network users are located.

The following sections identify seven important trends in data center and cloud networking that are accelerating traffic growth, changing the end-user experience, and placing new requirements and demands on data center and cloud-based infrastructures.

1. Growth of Global Data Center Relevance and Traffic
   - Growth of Hyperscale Data Centers
   - Global Data Center IP Traffic: Three-Fold Increase by 2020
   - Data Center Traffic Destinations: Most Traffic Remains Within the Data Center
   - Global Data Center and Cloud IP Traffic Growth
   - SDN/NFV Architecture Effects: Wild Card
2. Continued Global Data Center Virtualization
   - Public vs. Private Cloud

Trend 1: Growth of Global Data Center Relevance and Traffic

From server closets to large hyperscale deployments, data centers are at the crux of delivering IT services and providing storage, communications, and networking to the growing number of networked devices, users, and business processes in general. The growing importance of data analytics—the result of big data coming from ubiquitously networked end-user devices and IoE alike—has added to the value and growth of data centers. They touch nearly every aspect of an enterprise, whether internal/employee-related data, communication or processes, or partner- and customer-facing information and services. The efficient and effective use of data center technology such as virtualization, new software-based architectures, and management tools and use of public vs. private resources and so on can all add to the agility, success, and competitive differentiation of a business.

The increased focus on business agility and cost optimization has led to the rise and growth of cloud data centers. Cloud data centers have the five essential characteristics of cloud computing as listed by National Institute of Technology (NIST). These five characteristics are on-demand self-service, broad network access, resource pooling, rapid elasticity or expansion, and measured service. For more details, refer to Appendix E. Cloud adoption enables faster delivery of services and data, increased application performance, and improved operational efficiencies.

Although security and integration with existing IT environments continue to represent concerns for some potential cloud-based applications, a growing range of consumer and business cloud services are currently available. Today's cloud services address varying customer requirements (for example, privacy, mobility, and multiple device access) and

support near-term opportunities as well as long-term strategic priorities for network operators, both public and private.
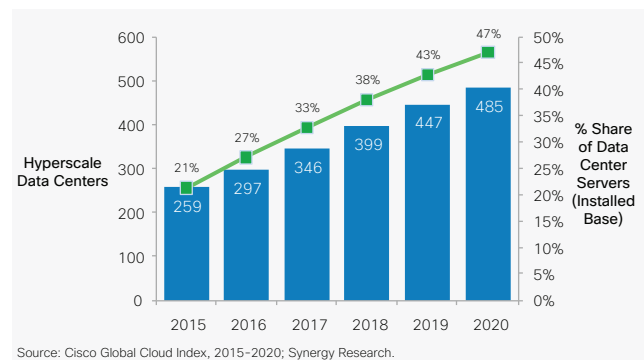
Hyperscale Data Center Growth

The increasing need for data center and cloud resources from both the business and consumer service perspective has led to the development of large-scale public cloud data centers called hyperscale data centers. Hyperscale cloud operators are increasingly dominating the cloud landscape.

To be a hyperscale cloud operator, a company must meet the following criteria defined in terms of annual revenues:

- More than US$1 billion in annual revenue from infrastructure as a service (IaaS), platform as a service (PaaS), or infrastructure hosting services (for example, Amazon/AWS, Rackspace, Google)

- More than US$2 billion in annual revenue from software as a service (SaaS) (for example, Salesforce, ADP, Google)

- More than US$4 billion in annual revenue from Internet, search, and social networking (for example, Facebook, Yahoo, Apple)

- More than US$8 billion in annual revenue from e-commerce/payment processing (for example, Amazon, Alibaba, eBay)

Twenty-four hyperscale operators were identified using the preceding criteria. The data centers operated by these companies are what we consider as hyperscale. The hyperscale operator might own the data center facility, or it might lease it from a colocation/wholesale data center provider.
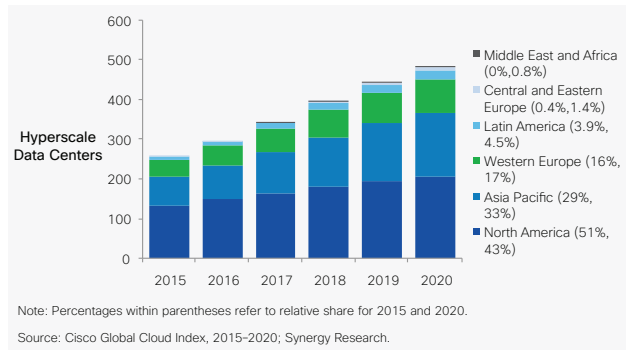
Figure 1. Data Center Growth

Source: Cisco Global Cloud Index, 2015–2020; Synergy Research.

These hyperscale data centers will grow from 259 in number at the end of 2015 to 485 by 2020. They will represent 47 percent of all installed data center servers by 2020. In other words, they will account for 83 percent of the public cloud server installed base in 2020 and 86 percent of public cloud workloads.

While only seven of these 24 companies are headquartered outside of the United States, their data center footprint is much more geographically diverse.

Figure 2.   Data Center Growth: Regional View

Note: Percentages within parentheses refer to relative share for 2015 and 2020.

Source: Cisco Global Cloud Index, 2015–2020; Synergy Research.

At the end of 2016, these 24 hyperscale operators will in aggregate have 297 data centers, with North America having the largest share, at 51 percent, followed by Asia Pacific, with 29 percent, Western Europe, with 17 percent, and Latin America, with 3 percent.

Asia Pacific has been the fastest growing region in terms of hyperscale data center location and will continue to grow more rapidly over the next five years, although North America will still account for 43 percent of hyperscale data centers by the end of 2020.

As with servers, hyperscale data centers represent a large portion of overall data, traffic, and processing power in data centers. Traffic within hyperscale data centers will quintuple by 2020. Hyperscale data centers already account for 34 percent of total traffic within all data centers and will account for 53 percent by 2020. Hyperscale data centers will also represent 57 percent of all data stored in data centers and 68 percent of total data center processing power.

Figure 3.   The Scale of Hyperscale

By 2020, Hyperscale Data Centers Will House:		Today:
47%	of all data center **servers**	21%
68%	of all data center **processing power**	39%
57%	of all **data** stored in data centers	49%
53%	of all data center **traffic**	34%

Source: Cisco Global Cloud Index, 2015–2020.

## Global Data Center IP Traffic: Three-Fold Increase by 2020

Most Internet traffic has originated or terminated in a data center since 2008, when peer-to-peer traffic

(which does not originate from a data center but instead is transmitted directly from device to device) ceased to dominate the Internet application mix. Data center traffic will continue to dominate Internet traffic for the foreseeable future, but the nature of data center traffic is undergoing a fundamental transformation brought about by cloud applications, services, and infrastructure. The importance and relevance of the global cloud evolution are highlighted by one of the top-line projections from this updated forecast: by 2020 more than 90 percent of data center traffic will be cloud traffic.

The following sections summarize not only the volume and growth of traffic entering and exiting the data center, but also the traffic carried between different functional units within the data center, cloud versus traditional data center segments, and business versus consumer cloud segments.

Figure 4 summarizes the forecast for data center IP traffic growth from 2015 to 2020.

Figure 4.   Global Data Center IP Traffic Growth

Source: Cisco Global Cloud Index, 2015–2020.
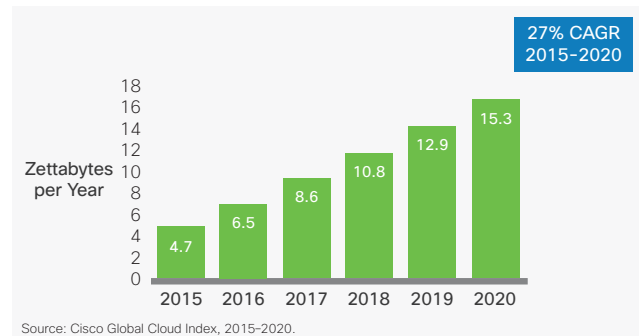
Although the amount of global traffic crossing the Internet and IP WAN networks is projected to reach 2.3 ZB per year by 2020[1], the amount of annual global data center traffic in 2015 is already estimated to be 4.7 ZB and by 2020 will triple to reach 15.3 ZB per year. This increase represents a 27 percent CAGR. The higher volume of data center traffic is due to the inclusion of traffic inside the data center (typically, definitions of Internet and WAN traffic stop at the boundary of the data center).

The global data center traffic forecast, a major component of the Cisco GCI report, covers network data centers worldwide operated by service providers as well as enterprises. Refer to Appendix A for more details about the methodology of the data center and cloud traffic forecasts and Appendix B for the positioning of the GCI Forecast relative to the Cisco VNI Global IP Traffic Forecast.

[1] Refer to Cisco Visual Networking Index: Forecast and Methodology, 2015–2020.

Table 1 provides details for global data center traffic growth rates.

Table 1.   Global Data Center Traffic, 2015–2020

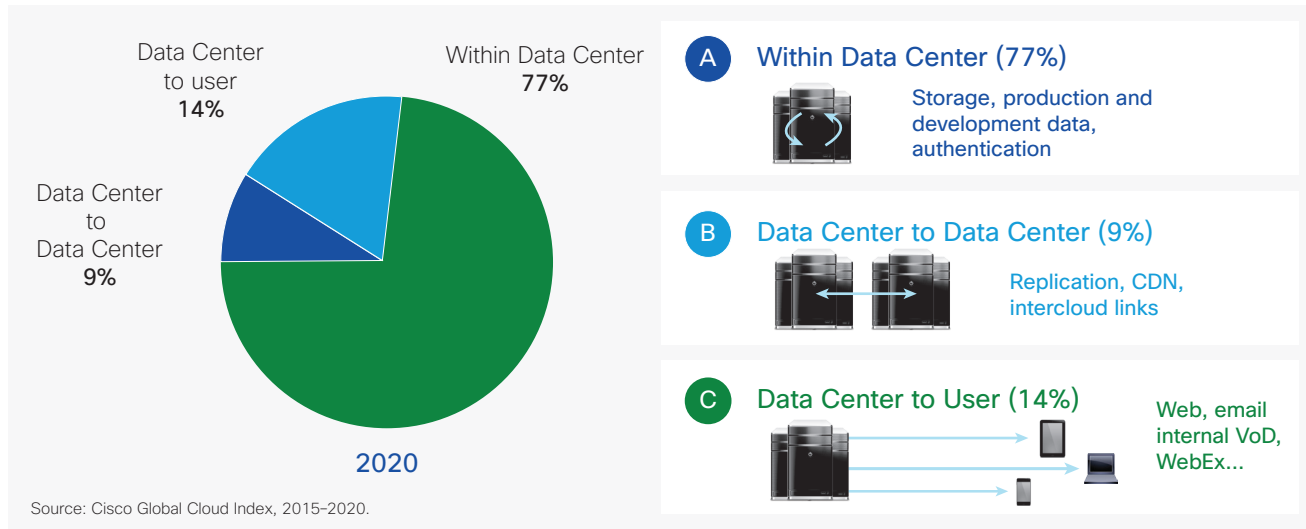| Data Center IP Traffic, 2015–2020							
	2015	2016	2017	2018	2019	2020	CAGR 2015–2020
By Type (EB per Year)							
Data center to user	744	933	1,164	1,438	1,772	2,183	24.0%
Data center to data center	346	515	713	924	1,141	1,381	31.9%
Within data center	3,587	5,074	6,728	8,391	10,016	11,770	26.8%
By Segment (EB per Year)							
Consumer	2,997	4,304	5,836	7,435	9,075	10,906	29.5%
Business	1,681	2,218	2,768	3,318	3,853	4,429	21.4%
By Type (EB per Year)							
Cloud data center	3,851	5,636	7,712	9,802	11,850	14,076	29.6%
Traditional data center	827	885	892	951	1,078	1,259	8.8%
Total (EB per Year)							
Total data center traffic	4,678	6,522	8,604	10,753	12,928	15,335	26.8%

Definitions:

- **Data center to user:** Traffic that flows from the data center to end users through the Internet or IP WAN
- **Data center to data center:** Traffic that flows from data center to data center
- **Within data center:** Traffic that remains within the data center, excludes traffic within the rack
- **Consumer:** Traffic originating with or destined for consumer end users
- **Business:** Traffic originating with or destined for business end users
- **Cloud data center:** Traffic associated with cloud data centers
- **Traditional data center:** Traffic associated with noncloud data centers

Data Center Traffic Destinations: Most Traffic Remains Within the Data Center

Consumer and business traffic flowing through data centers can be broadly categorized into three main areas (Figure 5):

- **Traffic that remains within the data center:** For example, moving data from a development environment to a production environment within a data center, or writing data to a storage array
- **Traffic that flows from data center to data center:** For example, moving data between clouds, or copying content to multiple data centers as part of a content distribution network
- **Traffic that flows from the data center to end users through the Internet or IP WAN:** For example, streaming video to a mobile device or PC

Figure 5.　Global Data Center Traffic by Destination in 2020

Data Center to user 14%

Within Data Center 77%

Data Center to Data Center 9%

2020

A Within Data Center (77%)
Storage, production and development data, authentication

B Data Center to Data Center (9%)
Replication, CDN, intercloud links

C Data Center to User (14%)
Web, email internal VoD, WebEx...

Source: Cisco Global Cloud Index, 2015–2020.

The portion of traffic residing within the data center will remain at the same level over the forecast period, accounting for 77 percent of data center traffic in 2015 and about 77 percent by 2020. The totals for within the data center do not include rack-local traffic (traffic that remains within a given server rack). Rack-local traffic is approximately twice the size of the "within data center" volumes shown in the forecast. The inclusion of rack-local traffic would change our traffic distribution to show more than 90 percent of traffic remaining local to the data center.

Big data is a significant driver of traffic within the data center. While much of big data traffic is rack-local, enough exits the rack that big data will be responsible for 17 percent of all traffic within the data center by 2020, up from 10 percent in 2015. Video does not drive a large volume of traffic within the data center, since minimal processing is done on the video relative to the large size of the video stream.

Traffic between data centers is growing faster than either traffic to end users or traffic within the data center, and by 2020, traffic between data centers will account for almost 9 percent of total data center traffic, up from 7 percent at the end of 2015. The high growth of this segment is due to the increasing prevalence of content distribution networks, the proliferation of cloud services and the need to shuttle data between clouds, and the growing volume of data that needs to be replicated across data centers.

Overall, east-west traffic (traffic within the data center and traffic between data centers) represents 86 percent of total data center by 2020,

and north-south traffic (traffic exiting the data center to the Internet or WAN) is only 14 percent of traffic associated with data centers.

Global Data Center and Cloud IP Traffic Growth

Data center traffic on a global scale will grow at a 27 percent CAGR (Figure 4), but cloud data center traffic will grow at a faster rate (30 percent CAGR) or 3.7-fold growth from 2015 to 2020 (Figure 6).

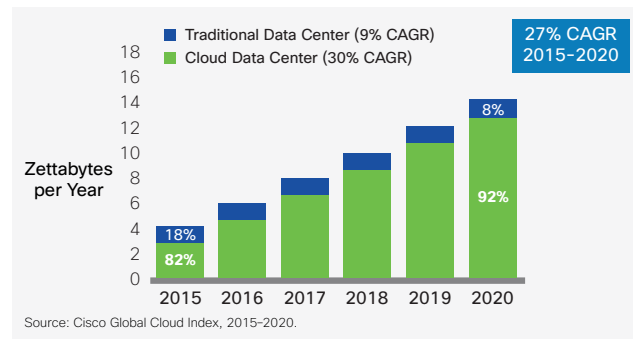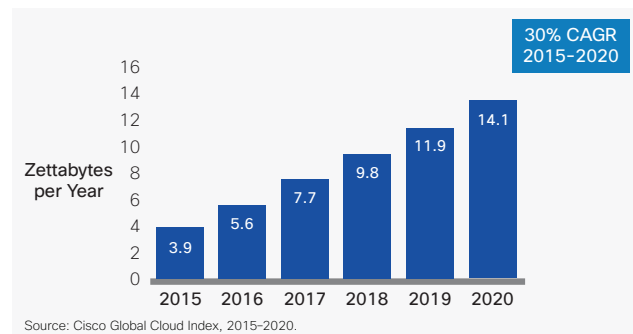Figure 6.　Total Data Center Traffic Growth

- Traditional Data Center (9% CAGR)
- Cloud Data Center (30% CAGR)

27% CAGR 2015–2020

Zettabytes per Year

2015: 18% / 82%
2020: 8% / 92%

Source: Cisco Global Cloud Index, 2015–2020.

Figure 7.　Cloud Data Center Traffic Growth

30% CAGR 2015–2020

Zettabytes per Year

| 2015	2016	2017	2018	2019	2020
3.9	5.6	7.7	9.8	11.9	14.1

Source: Cisco Global Cloud Index, 2015–2020.

Cloud will represent more than 90 percent of all data center traffic will be based in the cloud. (For regional cloud traffic trends, refer to Appendix C) Significant promoters of cloud traffic growth include the rapid adoption of and migration to cloud architectures and the ability of cloud data centers to handle significantly higher traffic loads. Cloud data centers support increased virtualization, standardization, and automation. These factors lead to better performance as well as higher capacity and throughput.

The Evolution of Data Center Architecture: SDN/NFV

Three technology trends are transforming the data center: leaf-spine architectures (which flatten the tiered architecture of the data center), software-defined networks (SDNs, which separate the control and forwarding of data center traffic), and network function virtualization (NFV, which virtualizes a variety of network elements).

Most major hyperscale data centers already employ flat architectures and software-defined network and storage management, and adoption of SDN/NFV within large-scale enterprise data centers has been rapid. As a portion of traffic within the data center, SDN/NFV is already transporting 23 percent, growing to 44 percent by 2020 (Figure 8).

Figure 8. SDN/NFV Traffic Within the Data Center
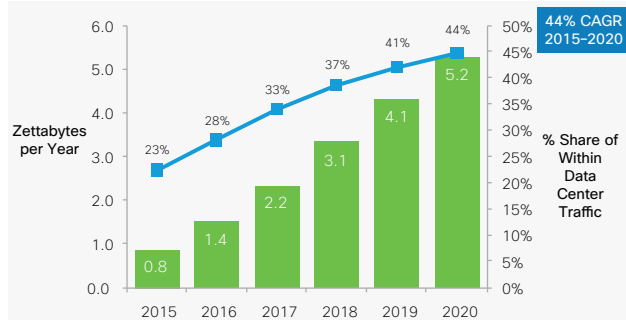
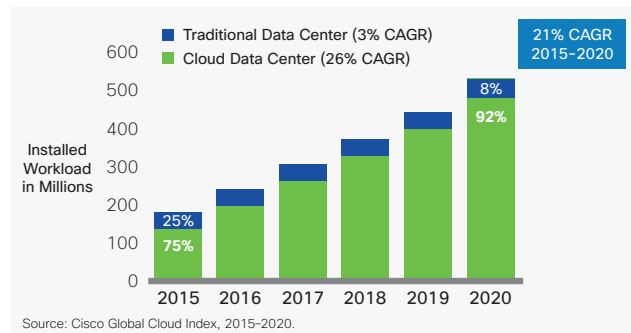Source: Cisco Global Cloud Index, 2015–2020.

SDN and NFV, along with flat architectures, might streamline traffic flows with the data center such that traffic is routed more efficiently in the future than it is routed today. In theory, SDN allows for traffic handling policies to follow virtual machines and containers, so that those elements can be moved within a data center in order to minimize traffic in response to bandwidth bottlenecks. However, there are also ways in which SDN/NFV can lead to an increase in both data center traffic and in general Internet traffic:

- **Big data:** Traffic engineering enabled by SDN/NFV supports "elephant"[2] data flows without compromising "mouse"[3] data flows, making it safe to transport large amounts of data to and from big data clusters.

- **Video bitrates:** SDN will allow video bitrates to increase, because SDN can seek out highest bandwidth available even midstream, instead of lowering the bitrate according the available bandwidth for the duration of the video, as is done today.

- **Cloud gaming:** SDN can decrease latency within the data center, decreasing delay experiences by end-users in cloud gaming applications, which might help increase cloud gaming adoption by both content providers and end users.

Trend 2: Continued Global Data Center Virtualization

A server workload is defined as a virtual or physical set of computer resources, including storage, that are assigned to run a specific application or provide computing services for one to many users. A workload is a general measurement used to describe many different applications, from a small lightweight SaaS application to a large computational private cloud database application. For the purposes of quantification, we consider each workload being equal to a virtual machine or a container. In fact, containers are one of the factors enabling a steady increase in the number of workloads per server deployed. The Cisco Global Cloud Index forecasts the continued transition of workloads from traditional data centers to cloud data centers. By 2020, 92 percent of all workloads will be processed in cloud data centers (Figure 9). For regional distributions of workloads, refer to Appendix D.

Figure 9. Workload Distribution: 2015–2020

Source: Cisco Global Cloud Index, 2015–2020.

[2] Elephant flows have varying definitions in the industry, but refer to flows of traffic that carry a disproportionate amount of traffic in terms of bytes, usually greater than 1% of total traffic in a time period.

[3] Mouse flows generate average-or-below traffic, but might have strict requirements in terms of latency.

Cloud workloads are expected to more than triple (grow 3.2-fold) from 2015 to 2020, whereas traditional data center workloads are expected to see a global decline, at a negative 3 percent CAGR from 2015 to 2020. Traditionally, one server carried one workload. However, with increasing server computing capacity and virtualization, multiple workloads per physical server are common in cloud architectures. Cloud economics, including server cost, resiliency, scalability, and product lifespan, along with enhancements in cloud security, are promoting migration of workloads across servers, both inside the data center and across data centers (even data centers in different geographic areas). Often an end-user application can be supported by several workloads distributed across servers. Table 2 provides details about the shift of workloads from traditional data centers to cloud data centers.

Table 2.   Workload Shift from Traditional Data Centers to Cloud Data Centers

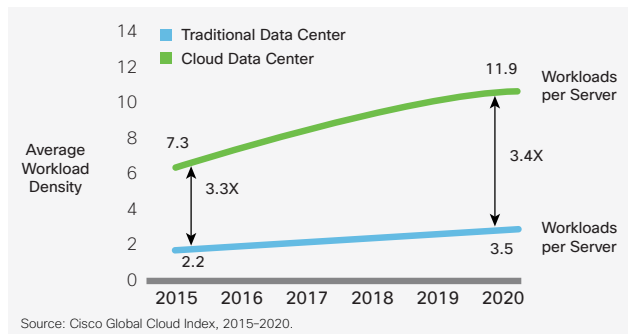| Global Data Center Workloads in Millions							
	2015	2016	2017	2018	2019	2020	CAGR 2015–2020
Traditional data center workloads	44.9	45.1	44.2	43.4	41.5	38.8	–3%
Cloud data center workloads	136.0	189.8	255.4	322.0	383.3	440.0	26%
Total data center workloads	180.9	234.9	299.7	365.4	424.8	478.8	21%
Cloud workloads as a percentage of total data center workloads	75%	81%	85%	88%	90%	92%	–
Traditional workloads as a percentage of total data center workloads	25%	19%	15%	12%	10%	8%	–

One of the main factors prompting the migration of workloads from traditional data centers to cloud data centers is the greater degree of virtualization (Figure 10) in the cloud space, which allows dynamic deployment of workloads in the cloud to meet the dynamic demand of cloud services. This greater degree of virtualization in cloud data centers can be expressed as workload density. Workload density measures average number of workloads per physical server. The workload density for cloud servers will grow from 7.3 in 2015 to 11.9 by 2020. In comparison, the workload density in traditional data center servers will grow from 2.2 in 2015 to 3.5 by 2020.

Figure 10. Increasing Cloud Virtualization

Source: Cisco Global Cloud Index, 2015–2020.

Public vs. Private Cloud[4]

We look into the growth of public cloud vs. private cloud through workload analysis. Public cloud, as indicated by the workloads growth, is growing faster than the private cloud. As the business sensitivity to costs associated with dedicated IT resources grows along with demand for agility, we can see a greater adoption of public cloud by the businesses, especially with strengthening of public cloud security. Although many mission-critical workloads might continue to be retained in the traditional data centers or private cloud, the public cloud adoption is increasing along with the gain in trust in public cloud. Some enterprises might adopt a hybrid approach to cloud. In a hybrid cloud environment, some of the cloud computing resources are managed in-house by an enterprise and some are provided by an external provider. Cloud bursting is an example of hybrid cloud where daily computing requirements are handled by a private cloud, but for sudden spurts of demand the additional traffic demand (bursting) is handled by a public cloud.

While the overall cloud workloads are growing at a CAGR of 26 percent from 2015 to 2020 (Figure 11), the public cloud workloads are going to grow at 35 percent CAGR from 2015 to 2020, and
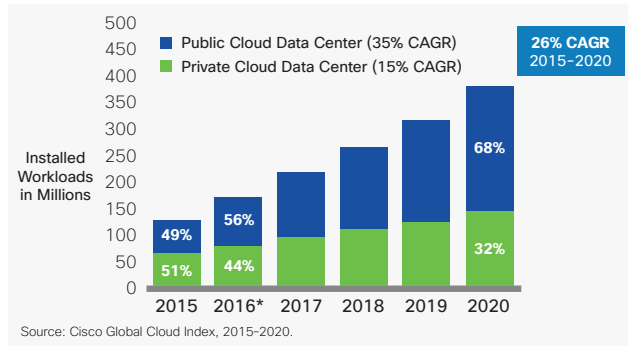
[4] For definition of public and private cloud, refer to Appendix E.

private cloud workloads will grow at a slower pace of 15 percent CAGR from 2015 to 2020. By 2016 there will be more workloads (56 percent) in the public cloud as compared to private cloud (44 percent).

Figure 11. Public vs. Private Cloud Growth

Source: Cisco Global Cloud Index, 2015–2020.

This growth of workloads in the public cloud space is also reflected in the growth of virtualization, as shown in Figure 12. The workload density in public cloud data centers will overtake that in private cloud data centers by 2016.

Figure 12. Public Cloud Virtualization Gaining Momentum

Source: Cisco Global Cloud Index, 2015–2020.
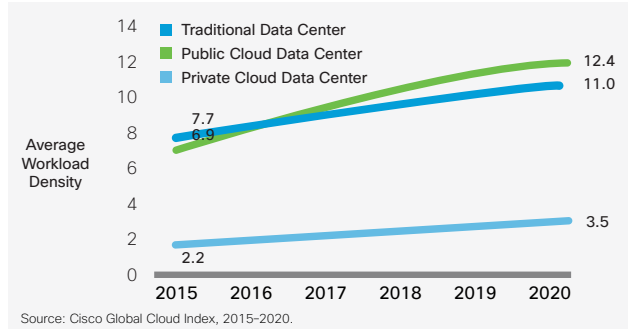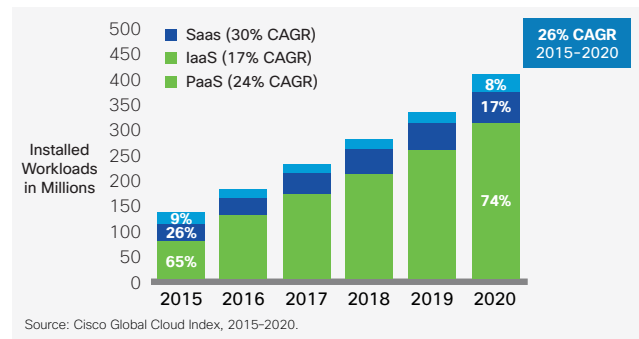
Trend 3: Cloud Service Trends

This section reviews the growth of the three different cloud service categories: IaaS, PaaS, and SaaS[5]. Although numerous other service categories have emerged over time, they can be aligned within the IaaS, PaaS, and SaaS categorization. For example, business process as a service (BPaaS) is considered part of SaaS. For simplicity we can think of these three service models as different layers of cloud with infrastructure at the bottom, the platform next, and finally software at the top.

GCI categorizes a cloud workload as IaaS, PaaS, or SaaS based upon how the user ultimately uses the service, regardless of other cloud services types that might be involved in the final delivery of the service. As an example, if a cloud service is a SaaS type but it also depends on some aspects of other cloud services such as PaaS or IaaS, such a workload is counted as SaaS only. As another example, if a PaaS workload operates on top of an IaaS service, such a workload is counted as PaaS only.
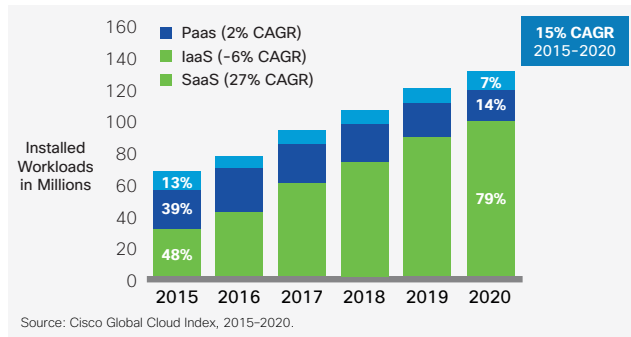
At the aggregate cloud level, we find that SaaS workloads maintain majority share throughout the forecast years, and by 2020 will have 74 percent share of all cloud workloads, growing at 30 percent CAGR from 2015 to 2020 (Figure 13). PaaS will have the second-fastest growth, although it will lose the share of total cloud workloads from 9 percent in 2015 to 8 percent by 2020.

Figure 13. SaaS Most Highly Deployed Global Cloud Service from 2015 to 2020

Source: Cisco Global Cloud Index, 2015–2020.

In order to understand the reasons behind this trend, we have to analyze the public and private cloud segments a bit more deeply. In the private cloud, initial deployments were predominantly IaaS. Test and development types of cloud services were the first to be used in the enterprise; cloud was a radical change in deploying IT services, and this use was a safe and practical initial use of private cloud for enterprises. It was limited, and it did not pose a risk of disrupting the workings of IT resources in the enterprise. As trust in adoption of SaaS or mission-critical applications builds over time with technology enablement in processing power, storage advancements, memory advancements, and networking advancements, we foresee the adoption of SaaS type applications to accelerate over the forecast period (Figure 14), while shares of IaaS and PaaS workloads decline.

---

[5] For definition of IaaS, PaaS, and SaaS, refer to Appendix E

Figure 14. SaaS Gains Momentum in Private Cloud

Source: Cisco Global Cloud Index, 2015–2020.

In the public cloud segment the first cloud services to be deployed were SaaS. SaaS services offered in the public cloud were often a low-risk and easy-to-adopt proposition, with some clear financial and flexibility benefits to users. The first users of SaaS were the consumer segment, followed by some small and medium-sized businesses. As public SaaS solutions become more sophisticated and robust, larger enterprises are adopting these services as well, beginning with less-critical services. Enterprises, especially large ones, will be carefully weighing the benefits (scalability, consistency, cost optimization, and so on) of adopting public cloud services against the risks (security, data integrity, business continuity, and so on) of adopting such services.

As shown in Figure 15, IaaS and PaaS have gone beyond the initial stages of deployment in the public cloud. Spend on public IaaS and PaaS is still small compared with spend on enterprise data center equipment, data center facilities, and associated operating expenses. These cloud services will gain momentum over the forecast period as more competitive offers come to the market and continue to build enterprise trust for outsourcing these more technical and fundamental services.

Figure 15. IaaS and PaaS Gain Public Cloud Share of Workloads

Source: Cisco Global Cloud Index, 2015–2020.
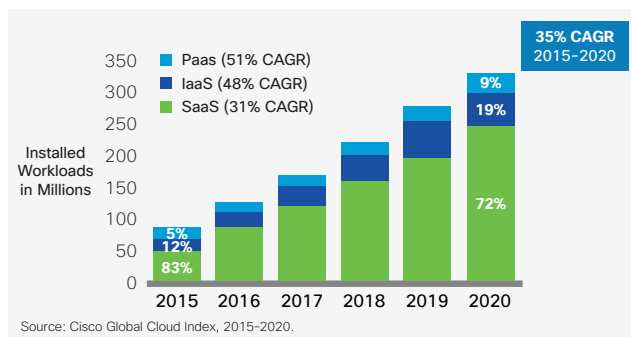
## Trend 4: Workloads by Application

This is the first year we are looking at workloads split by applications. We estimate that in 2015 enterprises (including SMB, government, and public sector) accounted for 79 percent of workloads and consumers 21 percent. Consumer share of the total will grow to 28 percent by 2020, while enterprise sector share will decline to 72 percent (Figure 16).

Figure 16. Global Data Center Workloads: Consumer vs. Enterprise Applications

Source: Cisco Global Cloud Index, 2015–2020.
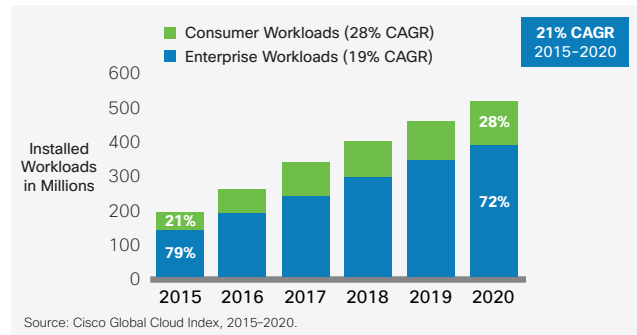
Within enterprise, compute/IaaS and collaboration are the two main contributors to workload totals, while on the consumer side social networking and video/media streaming are the biggest contributors. While the percentage mix will change, those will remain the biggest contributors to workload totals over the next five years (Figure 17). For definitions of the applications, see Appendix F.

Figure 17. Global Data Center Workloads by Applications


Source: Cisco Global Cloud Index, 2015–2020.
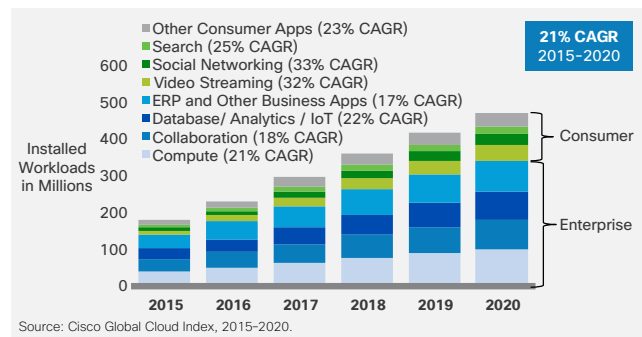
If we look at the application split of workloads across traditional, public, and private cloud data centers, then we find that public cloud data centers have the largest share of consumer application workloads, while traditional and private cloud data centers have a larger share in the business/enterprise segment (Figure 18).

Figure 18. Global Data Center Workloads by Applications: Traditional vs. Cloud (2015)

Source: Cisco Global Cloud Index, 2015–2020.
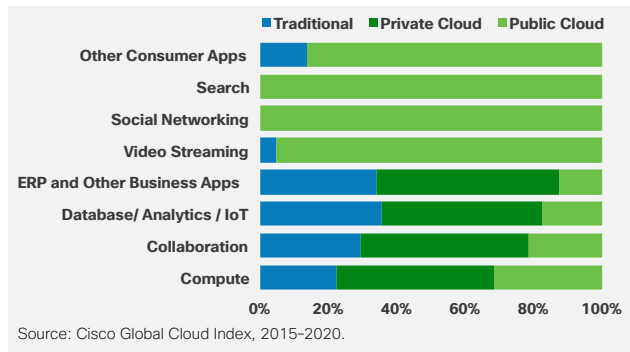
However, by 2020, traditional and private cloud data centers will lose share to public clouds across all applications except for database/analytics/IoE and other business application categories. In the latter categories, private clouds will mostly maintain their share (Figure 19).

Figure 19. Global Data Center Workloads by Applications: Traditional vs. Cloud (2020)

Source: Cisco Global Cloud Index, 2015–2020.
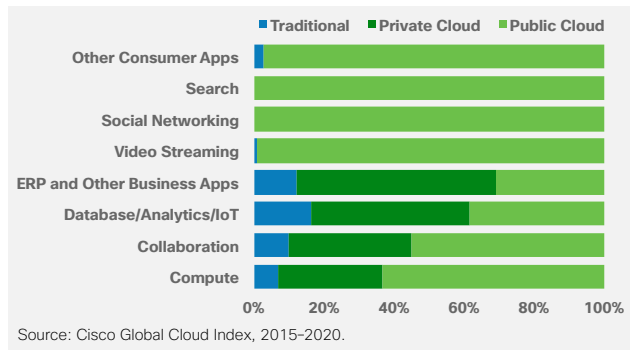
## Trend 5: Data Center and Cloud Storage[6]: Capacity and Utilization

Global Data Center Storage Installed Capacity

This year, for the first time, we have looked at the installed storage capacity in global data centers. We estimate that total data center storage capacity will grow nearly 5-fold from 2015 to 2020, growing from 382 EB in 2015 to 1.8 ZB by 2020. Cloud will account for 88 percent of the total storage capacity (Figure 20).

Figure 20. Global Data Center Storage Capacity: Traditional vs. Cloud

Source: Cisco Global Cloud Index, 2015–2020.
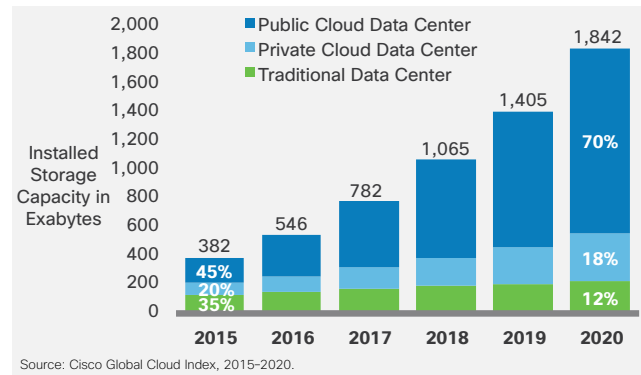
### Global Data Center Storage Utilization

Also for the first time we have estimated the total amount of actual data stored within data centers. Globally, the data stored in data centers will quintuple by 2020 to reach 915 EB by 2020, up 5.3-fold (a CAGR of 40%) from 171 EB in 2015 (Figure 21).

Figure 21. Actual Data Stored in Data Centers
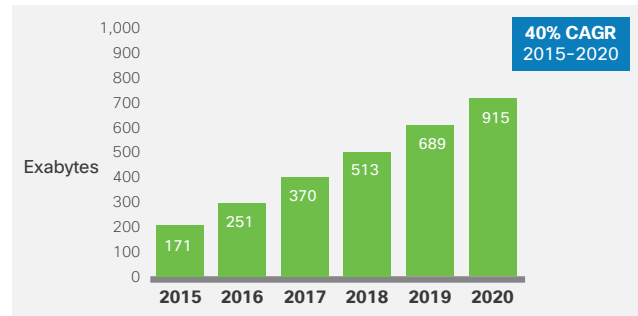
Big data is a key driver of overall growth in stored data. Big data will reach 247 EB by 2020, up almost 10-fold from 25 EB in 2015. Big data alone will represent 27 percent of data stored in data centers by 2020, up from 15 percent in 2015 (Figure 22).
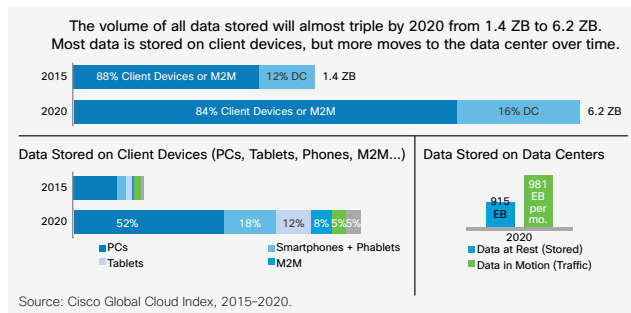
Figure 22. Big Data Volumes

---
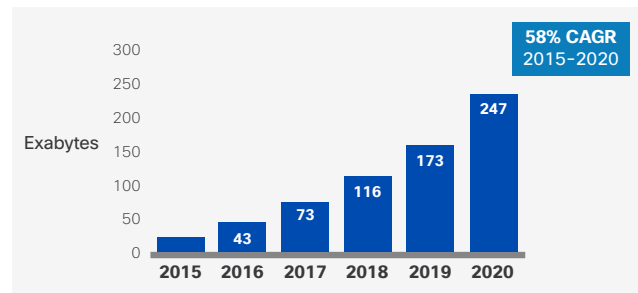
[6] Storage does not include archival media such as tape.

Big data is defined here as data deployed in a distributed processing and storage environment (such as Hadoop). Generally speaking, distributed processing is chosen as a data architecture when the data is big in volume (more than 100 terabytes), velocity (coming in or going out at more than 10 gigabytes per second), or variety (combining data from a dozen or more sources). Big data is sometimes used interchangeably with data analytics or data science, but data science techniques can be used on data of any size, and the quality of insights achieved with data science is not related to the size of the underlying data.

As large as the data stored in data centers will be (nearly 1 zettabyte by 2020), the amount of data stored on devices will be 5 times higher: 5.3 ZB by 2020. Out of the combined 6.2 ZB of stored data in the world, most stored data will continue to reside in client devices, as it does today. Today, only 12 percent of total stored data is stored in the data center, but more data will move to the data center over time (Figure 23). In addition to larger volumes of stored data, the stored data will be coming from a wider range of devices by 2020. Currently, 61 percent of data stored on client devices resides on PCs. By 2020, stored data on PCs will reduce to 52 percent, with a greater portion of data on smartphones, tablets, and machine-to-machine (M2M) modules. Stored data associated with M2M grows at a faster rate than any other device category.

Figure 23. Data Center Storage Analysis

The volume of all data stored will almost triple by 2020 from 1.4 ZB to 6.2 ZB. Most data is stored on client devices, but more moves to the data center over time.

Source: Cisco Global Cloud Index, 2015–2020.

Over time, cloud-based services will enable consumers and businesses alike to move more of their stored data to a central repository that can provide ubiquitous access to content and applications through any device at any location. The following section covers the growth in consumer cloud storage.
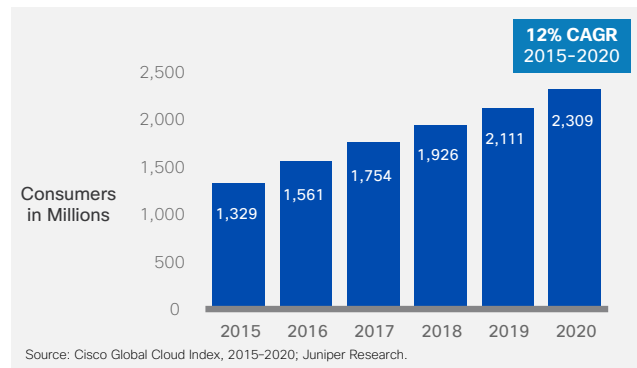
Consumer Cloud Storage Growth

Along with the growth in consumer Internet population and multidevice ownership devices, we are seeing a significant growth in the use of cloud services such as consumer cloud storage, also called personal content lockers. In personal content lockers, users can store and share music, photos, and videos through an easy-to-use interface at relatively low or no cost. Furthermore, the proliferation of tablets, smartphones, and other mobile devices allows access to personal content lockers in a manner convenient to the user.
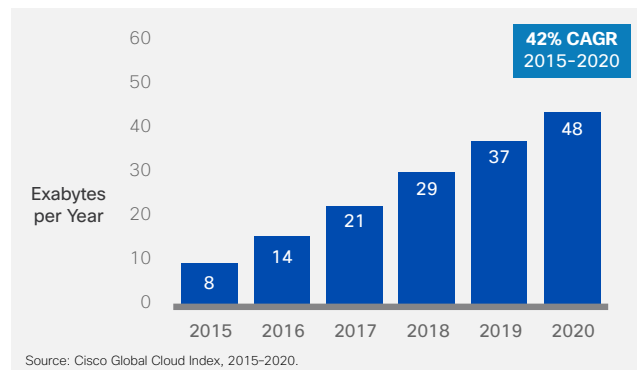
Cisco GCI estimates that by 2020, 59 percent (2.3 billion) of the consumer Internet population will use personal cloud storage, up from 47 percent (1.3 billion users) in 2015 (Figure 24).

Figure 24. Personal Cloud Storage: Growth in Users

Source: Cisco Global Cloud Index, 2015–2020; Juniper Research.

Cisco GCI forecasts that global consumer cloud storage traffic will grow from 8 EB annually in 2015 to 48 EB by 2020 at a CAGR of 42 percent (Figure 25). This growth translates to per-user traffic of 1.7 gigabytes (GB) per month by 2020, compared to 513 MB per month in 2015.

Figure 25. Consumer Cloud Storage Traffic[7] Growth

Source: Cisco Global Cloud Index, 2015–2020.

---

[7] Consumer cloud storage traffic includes personal content lockers, cloud backup, and so on, and does not include cloud DVR.

Trend 6: Global Digitization: Impact of IoE

Potential Effects of IoE on Global Data Centers

Cloud services are accelerated in part by the unprecedented amounts of data being generated by not only people but also machines and things. Cisco GCI estimates that 600 ZB will be generated by all people, machines, and things by 2020, up from 145 ZB generated in 2015.

Figure 26 shows examples of the amounts of data that will be generated by planes, automobiles, and buildings, among other things and systems.

Figure 26. Smart City: Multiple Applications Create Big Data

What Makes a Smart City?
Multiple Applications Create Big Data

Connected Plane	**Intelligent Building**
40 TB per day (0.1% transmtted)	275 GB per day (1% transmitted)
Connected Factory	**Smart Hospital**
1 PB per day (0.2% transmitted)	5 TB per day (0.1% transmitted)
Public Safety	**Smart Car**
50 PB per day (<0.1% transmitted)	70 GB per day (0.1% transmitted)
Weather Sensors	**Smart Grid**
10 MB per day (5% transmitted)	5 GB per day (1% transmitted)

A city of one million will generate 200 million gigabytes of data per day by 2020

Source: Cisco Global Cloud Index, 2015–2020

Most of the more than 600 ZB that will be generated by 2020 will be ephemeral in nature and will be neither saved nor stored. Much of this ephemeral data is not useful to save, but we estimate that approximately 10 percent is useful, which means that there will be 10 times more useful data being created (60 ZB, 10 percent of the 600 total) than will be used (6 ZB) in 2020. Edge or fog computing might help bridge this gap.
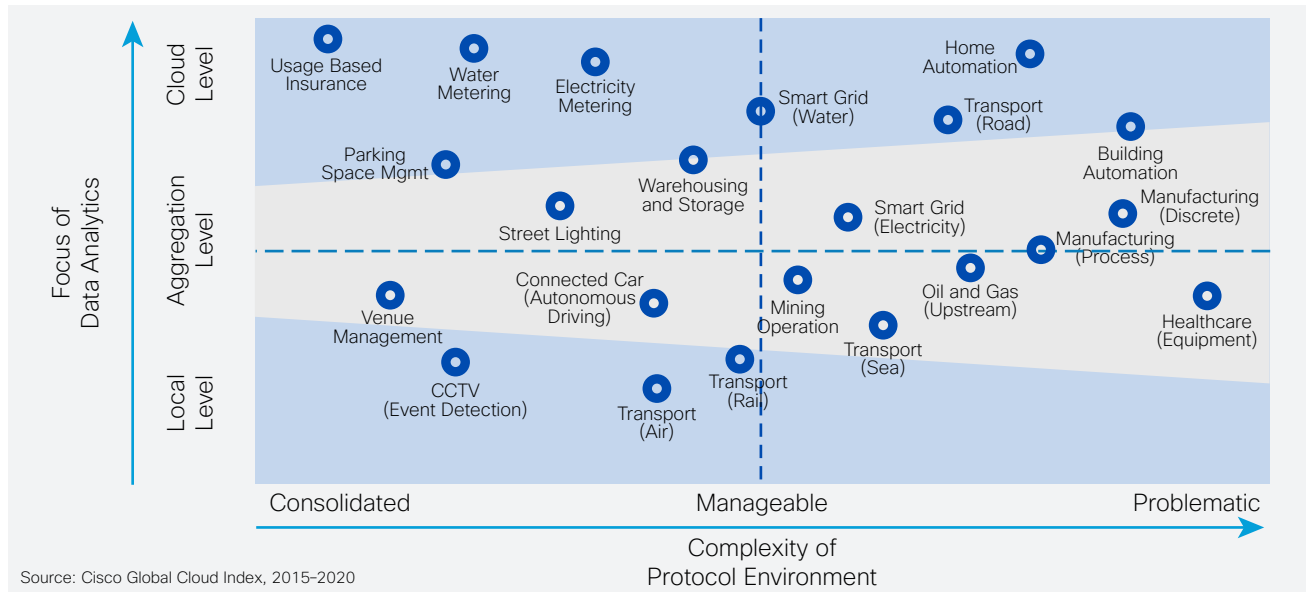
M2M Data Analytic Requirements Drive Fog/Cloud Computing

Growth in machine-to-machine connections and applications is also driving new data analytics needs. Although not all M2M applications promote a lot of traffic, the sheer number of these connections is capable of delivering intelligent, actionable information if the data from them can be analyzed. Figure 27 maps several of such M2M applications for their complexity of protocol environment and data analytic needs. IoT applications have very different characteristics. In some cases application analytics and management can occur at the edge device level whereas for others it is more appropriately handled centrally, typically hosted in the cloud.

The opportunity for fog propositions (distributed intelligence) and the associated intelligent gateways is the strongest in markets that meet two conditions. First, their focus of data analytics is on the so-called aggregation level, and, second, they are subject to a problematic degree of protocol complexity. Especially the first aspect is set to be decisive in determining the demand for intelligent gateways as a product segment. In general, such markets reside in the Industrial IoT domain – involving verticals such as manufacturing, extractive industries, and healthcare. Applications such as smart metering can benefit from real-time analytics of aggregated data that can optimize the usage of resources such as electricity, gas, and water. Local level analytics is suited for those applications that require the data to be stored and analyzed locally due to either regulatory reasons or because the cost of transportation of the data upstream and associated wait-time for analysis is prohibitive.

A key issue for IoT in the coming years is subsidiarity, i.e. performing the data analysis at the appropriate level. In most cases, there will be a blend of approaches and the functionality to manage local as well as central application management will be increasingly critical.

According to the National Institute of Technology (NIST), cloud computing can be divided into three main service types (refer to the section "Trend 3: Cloud Service Trends"): Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), and each affects data control and governance a little differently. With IaaS, the customer might have full control of the actual server configuration, granting them more risk management control over the environment and data. In PaaS, the provider manages the hardware and underlying operating system, limiting enterprise risk management capabilities on those components. With SaaS, both the platform and the infrastructure are fully managed by the cloud provider, meaning if the underlying operating system or service isn't configured appropriately the data in the higher layer application might be at risk.

Cybercrime damages will cost the world $6 trillion annually by 2021, up from $3 trillion in 2015. The cybercrime costs prediction includes damage and destruction of data, stolen money, lost productivity, theft of intellectual property, theft of personal and financial data, embezzlement, fraud, postattack disruption to the normal course of business, forensic investigation, restoration and deletion of hacked data and systems, and reputational harm.[9] Cyberthreats have evolved from targeting and harming computers, networks, and smartphones to people, cars, railways, planes, power grids, and anything with a heartbeat or an electronic pulse, all powered by the cloud. The last several years have undoubtedly been the most eventful period from a cloud security threat perspective, with various instances of massive breaches and escalating distributed denial-of-service (DDoS) amplification attacks. Many network security vendors and other networking device manufacturers raced to patch their appliances against Heartbleed, a serious vulnerability in the popular OpenSSL cryptographic software library, and Shellshock, an open-source vulnerability, which set off a series of patch releases, among many others. The response showcased the effectiveness of security vendors to provide support and assistance to customers in need. It also shed light on the broad adoption and very quickly evolving landscape of cloud security and management. Across all documents uploaded to file sharing services, the most common type of sensitive content is confidential company data (for example, financial records, business plans, source code, trading algorithms, and so on). A total of 7.6 percent of documents in file sharing services contain confidential data. That's followed by personally identifiable information (for example, Social Security numbers, tax ID numbers, phone numbers, addresses, and so on) at 4.3 percent of all documents. Next, 2.3 percent of documents contain payment data (for example, credit card numbers, debit card numbers, bank account numbers, and so on). Finally, 1.6 percent of documents contain protected health information (for example, patient diagnoses, medical treatments, medical record IDs, and so on).[10]

Users expect their online experiences to be always available and always secure—and their personal and business assets to be safe. As more data, business processes, and services move to the cloud, organizations are challenged to protect websites and infrastructure without sacrificing performance for security.

To help meet user expectations, more secure Internet servers are being deployed worldwide. This situation creates a growing infrastructure footprint that provides more stringent authorization and authentication processes and better serves end users with secure transactions and communication. The percentage of secure Internet servers that conduct encrypted transactions over the Internet using a Secure Sockets Layer (SSL) to the total number of web facing servers is shown in Figure 29. In the past year, North America and Western Europe led with the percentage of secure Internet servers compared to web-facing Internet servers.

Figure 29. Percentage of Secure Internet Servers to Total Web-Facing Internet Servers by Region and Increase from End of Year 2014 to 2015

Source: Cisco GCI 2016, UN, NetCraft, Synergy Research.
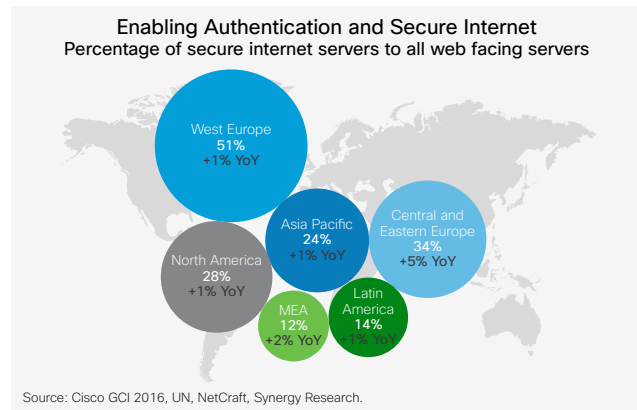
Although end-user security concerns exist, the time of amateur hackers is long over, and hacking is now an organized crime or state-sponsored event. DDoS attacks against customers remain a major operational threat to service providers. Attacks against infrastructure continue to grow in prominence. Phishing and malware threats occur on a daily basis.

[9] http://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016

[10] https://www.skyhighnetworks.com/cloud-report/

According to the Cisco 2016 Annual Security report, the government has emerged as the number one high risk industry for web malware encounters in 2015. To track high-risk verticals for web malware encounters, the relative volumes of attack traffic (block rates) and normal or expected traffic was examined. From January to March 2015, government was the vertical with the highest block rate activity. From March to May, it was electronics. In midsummer, professional services saw the most blocks. And in the fall of 2015, healthcare was leading all verticals in the number of block rates, per Figure 30.

Figure 30. Vertical Block Rates by Month, November 2014–September 2015
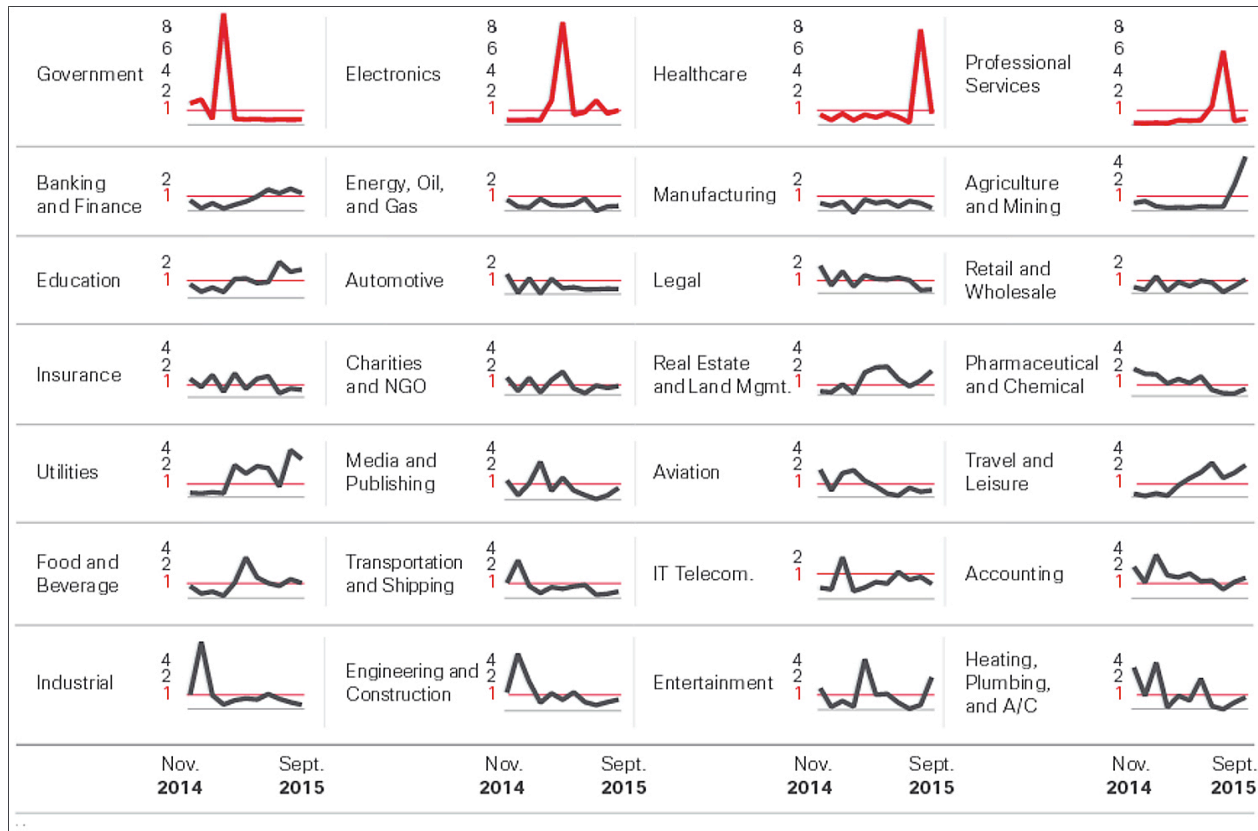
Source: Cisco Security Research.

Encrypted traffic, particularly HTTPS, has reached a tipping point. While not yet the majority of transactions, it will soon become the dominant form of traffic on the Internet. It consistently represents over 50 percent of bytes transferred (Figure 31) due to the HTTPS overhead and larger content that is sent via HTTPS, such as transfers to file storage sites.

Figure 31. SSL/HTTPS Percentages
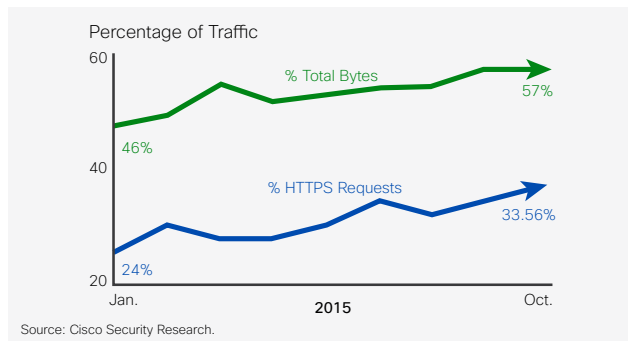
Source: Cisco Security Research.

IoE and big data requirements are starting a new wave of security discussions and technology convergence. As enterprises and service providers move to public and private clouds and modernize data centers with SDN or consume IT as a Service (ITaaS), security becomes an even more complex concern. Besides hardware appliances, virtual machines, and server software, innovative services that use SDN and NFV will help to improve the data integrity and security of cloud infrastructures.

Network Speeds and Latency Analysis
The cloud-readiness study offers a regional view of the requirements for broadband and mobile networks to deliver next-generation cloud services. The enhancements and reliability of these networks will support the increased adoption of cloud computing solutions that deliver basic as well as advanced application services. For example,

consumers expect to be able to communicate with friends as well as stream music and videos at any time, any place. Business users require reliable access to business communications along with mobile solutions for video conferencing and mission-critical customer and operational management systems.

The study also explores the ability of each global region (Asia Pacific, Central and Eastern Europe, Latin America, Middle East and Africa, North America, and Western Europe) to support a sample set of basic, intermediate, and advanced business and consumer cloud applications. Each region's cloud readiness is assessed with relation to the sample services based on download and upload fixed and mobile network speeds as well as associated network latencies (segmented by business and consumer connections). Download and upload speeds as well as latencies are essential measures to assess network capabilities for cloud readiness. Figure 32 provides the business and consumer cloud service categories and the corresponding network requirements used for this study. Tables 3 through 5 describe the requirements and define a sample set of applications from each of the readiness categories. Note that the concurrent use of applications can further influence the user experience and cloud accessibility.

Figure 32. Sample Business and Consumer Cloud Service Categories

Basic Cloud Apps	Intermediate Cloud Apps	Advanced Cloud Apps
Network Requirements:	Network Requirements:	Network Requirements:
Download Speed: Up to 750 kbps	Download Speed: 751–2,500 kbps	Download Speed: Higher than 2,500 kbps
Upload Speed: Up to 250 kbps	Upload Speed: 251–1,000 kbps	Upload Speed: Higher than 1,000 kbps
Latency: Above 160 ms	Latency: 159–100 ms	Latency: Less than 100 ms
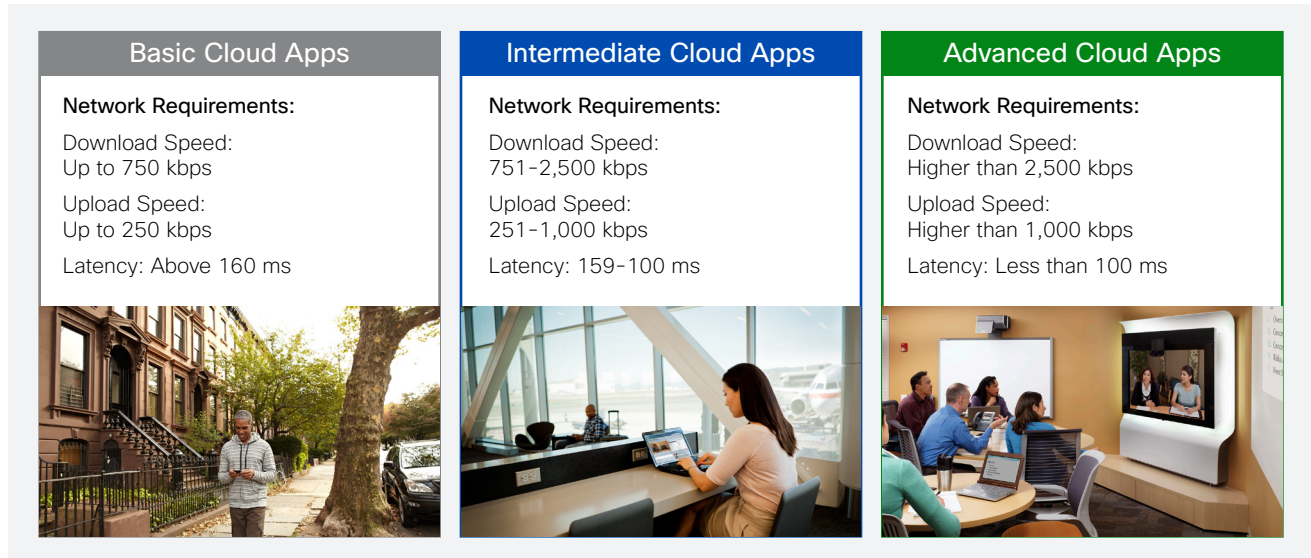
Table 3. Sample Basic Applications

Apps	Definitions	Download	Upload	Latency
Stream basic video and music	Deliver sound and video without the need to download files of different audio or video formats using computer servers connected to the Internet to access information.	High	Low	Medium
Text communications (email and instant messaging)	A cross-platform messaging application that allows the exchange of messages without having to pay for Short Message Service (SMS), using an Internet data plan.	Low	Low	Medium
Voice over IP (VoIP) (Internet telephony)	A broad range of services transmitting voice over the Internet.	Low	Low	Medium
Web browsing	Accelerate web experiences and searching through cloud computing using technology to shift the workload to the cloud servers.	Low	Low	Medium

Apps	Definitions	Download	Upload	Latency
Web conferencing	A cloud application used to interact with other participants and have that live and in-person feeling for attendees; it offers services such as collaborative web browsing and application sharing.	Medium	Medium	Medium
Cloud-based learning management system	This app provides the user with the flexibility of being able to access and collaborate with others in a centralized environment. With information housed in a virtual storage environment, it allows work to be completed outside the boundaries of the formal learning or work institutions.	High	Medium	Medium

Table 4.   Sample Intermediate Applications

Apps	Definitions	Download	Upload	Latency
Enterprise resource planning (ERP) and customer relationship management (CRM)	ERP and CRM systems allow businesses to manage their business and business relationships and the data and information associated with them.	Medium	Low	Medium
High-definition (HD) video streaming	Deliver HD video without the need to download files of HD video formats using computer servers connected to the Internet to access information.	High	Low	Low
Augmented reality (AR) gaming applications	Augmented reality (AR) games involve a live direct or indirect view of a physical, real-world environment whose elements are augmented (or supplemented) by computer-generated sensory input such as sound, video, graphics, or GPS data.	High	Medium	Low
Web electronic health records (EHRs)	EHRs are designed to contain and share information from all providers involved in a patient's care in a structured format allowing patient information to be easily retrieved and transferred in ways that can aid patient care.	Medium	High	Low
Voice over LTE (VoLTE)	This standardized system allows for transferring traffic for VoLTE.	Low	Low	Low
Personal content locker	Asynchronous storage enables applications that use compound files to efficiently render their content when accessed by means of existing Internet protocols, with a single request to a server triggering the download of nested objects contained within a webpage, eliminating the need to separately request each object.	High	High	Low

Table 5.   Sample Advanced Applications

Apps	Definitions	Download	Upload	Latency
Telemedicine	Telemedicine is the use of medical information exchanged from one site to another through electronic communications to improve a patient's clinical health status and includes using two-way video, email, smartphones, wireless tools, and other forms of telecommunications technology.	Medium	Medium	Low
HD video conferencing	Two-way interactive HD video communication is delivered using Internet technologies that allow people at different locations to come together for a meeting.	High	High	Low
Ultra HD video streaming	This app delivers Ultra HD video without the need to download files of different video formats using computer servers connected to the Internet to access information.	High	High	Low
Virtual reality (VR) streaming	Streaming of realistic and immersive simulation of a three-dimensional environment, created using interactive software and hardware, and experienced or controlled by movement of the body or as an immersive, interactive experience generated by a computer.	High	High	Low
High-frequency stock trading	These apps support the rapid turnover of positions through the use of sophisticated trading algorithms, which process hundreds of trades in fractions of a second on the basis of changing market conditions.	Low	Low	Low
Connected vehicles safety applications	These apps involve the development and deployment of a fully connected transportation system that makes the most of multimodal, transformational applications requiring a combination of well-defined technologies, interfaces, and processes that, combined, help ensure safe, stable, interoperable, reliable system operations that minimize risk and maximize opportunities.	Low	Low	Low

Regional network performance statistics were ranked by their ability to support these three cloud service categories. More than 300 million records from Ookla's Speedtest[11] along with data from Ovum-Informa, Synergy Research, Point Topic, United Nations (UN), World Bank, NetCraft, International Telecommunication Union (ITU), International Labor Organization (ILO), and other sources were analyzed from more than 200 countries to understand cloud readiness. The regional averages of these measures are included as follows and in Appendix G.

[11] Measured by Speedtest.net, small binary files are downloaded and uploaded between the web server and the client to estimate the connection speed in kilobits per second (kbps).

The cloud readiness characteristics follow.

Network Access

- **Internet ubiquity:** This indicator measures fixed and mobile Internet penetration while considering population demographics to understand the pervasiveness and expected connectivity in various regions.

Network Performance

- **Download speed:** With increased adoption of mobile and fixed bandwidth-intensive applications, end-user download speed is an important characteristic. This indicator will continue to be critical for the quality of service delivered to virtual machines, CRM, and ERP cloud platforms for businesses, video download, and content-retrieval cloud services for consumers.

- **Upload speed:** With the increased adoption of virtual machines, tablets, and video conferencing in enterprises as well as by consumers on both fixed and mobile networks, upload speeds are especially critical for delivery of content to the cloud. The importance of upload speeds will continue to increase over time, promoted by the dominance of cloud computing and data center virtualization, the need to transmit many millions of software updates and patches, the distribution of large files in virtual file systems, and the demand for consumer cloud game services and backup storage.

- **Network latency:** Delays experienced with VoIP, viewing and uploading videos, online banking on mobile broadband, or viewing hospital records in a healthcare setting are due to high latencies (usually reported in milliseconds). Reducing delay in delivering packets to and from the cloud is crucial to delivering today's advanced services (and making sure of a high-quality end-user experience).

Global Average Download and Upload Speed Overview (2016)

Download and upload speeds as well as latencies are important measures to assess network capabilities for cloud readiness. The Cisco GCI Supplement provides additional country-level details for download speeds, upload speeds, and latencies. To support cloud services and applications, the quality of the broadband connection is critical. Although theoretical speeds offered by fixed and mobile operators can seem adequate, many extraneous factors are involved in the actual network measurements. Speeds and latencies vary within each country and region, based on urban and rural deployment of fixed and mobile broadband technology, proximity to traditional and cloud data centers, and the quality of customer premises equipment (CPE).

Lesser variability in download speeds, upload speeds, and latency will allow consumers to access advanced cloud applications consistently throughout the country. To measure this variability, we have also included the median download speeds and median upload speeds, along with the update to the mean or average download speeds and upload speeds, all measured and typically represented in kilobits per second (kbps) or megabits per second (Mbps).

Key Results

- The global average fixed download speed is 29.5 Mbps, and the global median fixed download speed is 19.1 Mbps.

- The global average fixed upload speed is 14.3 Mbps, and the global median upload speed is 6.8 Mbps.

- The global average mobile download speed is 15.1 Mbps, and the global median mobile download speed is 8.3 Mbps.

- The global average mobile upload speed is 7.6 Mbps, and the global median mobile upload speed is 3.5 Mbps.

Regional Fixed Download and Upload Speeds

- **Average fixed download speeds:** Asia Pacific leads with 33.9 Mbps, and North America follows with 32.9 Mbps.

- **Average fixed upload speeds:** Central and Eastern Europe leads with 19.3 Mbps, and Asia Pacific follows with 18.9 Mbps (Figure 33). For further details, refer to Appendix G and the Cisco GCI Supplement.

- **Median fixed download and upload speeds:** Median speeds are lower than the average/mean speeds, as shown in Figure 33, because of a higher distribution of speeds in the region that are lower than the mean. Besides the required network characteristics for advanced cloud application, for an optimal end-user experience in larger user bases with cloud services, the majority of speeds must also be closer to the mean. This factor is a critical factor. To understand speed distribution patterns in detail for a select list of countries, refer to the Cisco GCI Supplement.

Figure 33. Regional Average Fixed Speeds, 2016

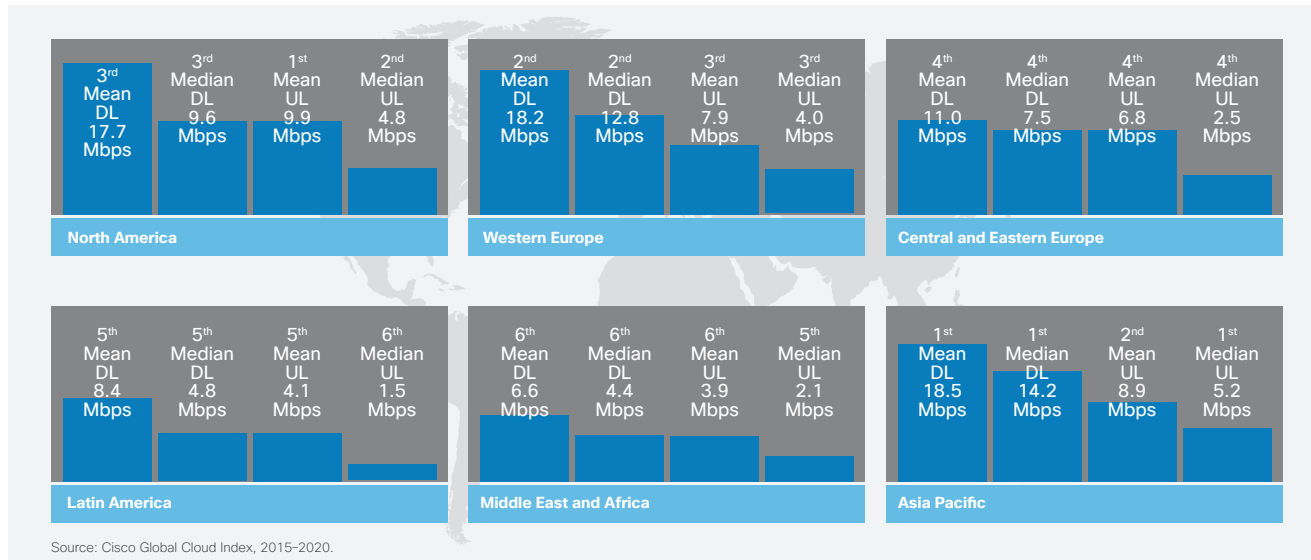|			
2nd Mean DL 32.9 Mbps	1st Median DL 23.4 Mbps	3rd Mean UL 11.6 Mbps	3rd Median UL 5.9 Mbps
North America			
3rd Mean DL 30.2 Mbps	3rd Median DL 18.0 Mbps	4th Mean UL 11.0 Mbps	4th Mean UL 4.8 Mbps
Western Europe			
4th Mean DL 29.1 Mbps	4th Median DL 16.2 Mbps	1st Mean UL 19.3 Mbps	1st Median UL 10.7 Mbps
Central and Eastern Europe			
5th Mean DL 9.3 Mbps	5th Median DL 6.2 Mbps	6th Mean UL 3.3 Mbps	6th Median UL 1.0 Mbps
Latin America			
6th Mean DL 7.8 Mbps	6th Median DL 5.7 Mbps	5th Mean UL 3.9 Mbps	5th Median UL 1.4 Mbps
Middle East and Africa			
1st Mean DL 33.9 Mbps	2nd Median DL 22.4 Mbps	2nd Mean UL 19.0 Mbps	2nd Median UL 8.8 Mbps
Asia Pacific			

Source: Cisco Global Cloud Index, 2015–2020.

Regional Average Mobile Download and Upload Speeds

- **Average mobile download speeds:** Asia Pacific leads with 18.5 Mbps, and Western Europe follows with 18.2 Mbps.

- **Average mobile upload speeds:** North America leads with 9.9 Mbps, and Asia Pacific follows with 8.9 Mbps (Figure 34). For further details, refer to Appendix G and the Cisco GCI Supplement.

- **Median mobile download and upload speeds**: Median speeds are lower than mean mobile speeds within all regions, with the distribution of speeds in the regional population tending to be lower than the average.

Figure 34. Regional Average Mobile Speeds, 2016

3rd Mean DL 17.7 Mbps	3rd Median DL 9.6 Mbps	1st Mean UL 9.9 Mbps	2nd Median UL 4.8 Mbps
North America			
2nd Mean DL 18.2 Mbps	2nd Median DL 12.8 Mbps	3rd Mean UL 7.9 Mbps	3rd Median UL 4.0 Mbps
Western Europe			
4th Mean DL 11.0 Mbps	4th Median DL 7.5 Mbps	4th Mean UL 6.8 Mbps	4th Median UL 2.5 Mbps
Central and Eastern Europe			
5th Mean DL 8.4 Mbps	5th Median DL 4.8 Mbps	5th Mean UL 4.1 Mbps	6th Median UL 1.5 Mbps
Latin America			
6th Mean DL 6.6 Mbps	6th Median DL 4.4 Mbps	6th Mean UL 3.9 Mbps	5th Median UL 2.1 Mbps
Middle East and Africa			
1st Mean DL 18.5 Mbps	1st Median DL 14.2 Mbps	2nd Mean UL 8.9 Mbps	1st Median UL 5.2 Mbps
Asia Pacific			

Source: Cisco Global Cloud Index, 2015–2020.

Network Latency

- Global average fixed latency is 36 ms.

- Asia Pacific leads in average fixed latency with 26 ms, followed by Central and Eastern Europe with 30 ms.

- Global average mobile latency is 81 ms.

- Western Europe leads in average mobile latency with 57 ms, followed by Asia Pacific with 73 ms.

As cloud data centers are built and distributed around the world and content is closer to the user, the length of the time it takes for a small packet of data to be sent and received will get lesser. The other factor that can also affect latency is congestion, which leads to a lower throughput. Latency has significantly improved in both fixed and mobile networks. Figures 35 and 36 show the latency improvements from 2014 through 2016 in average fixed latency in ms as well as average mobile latency in ms by region.

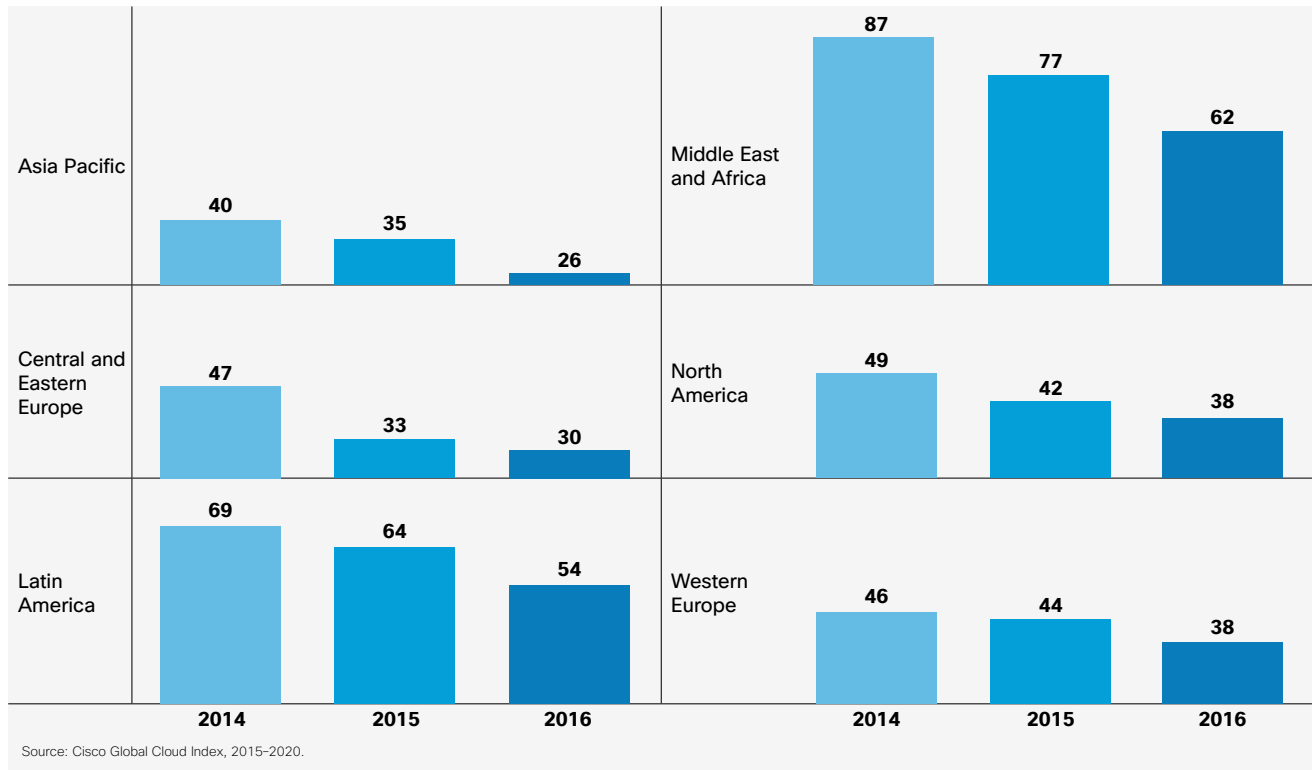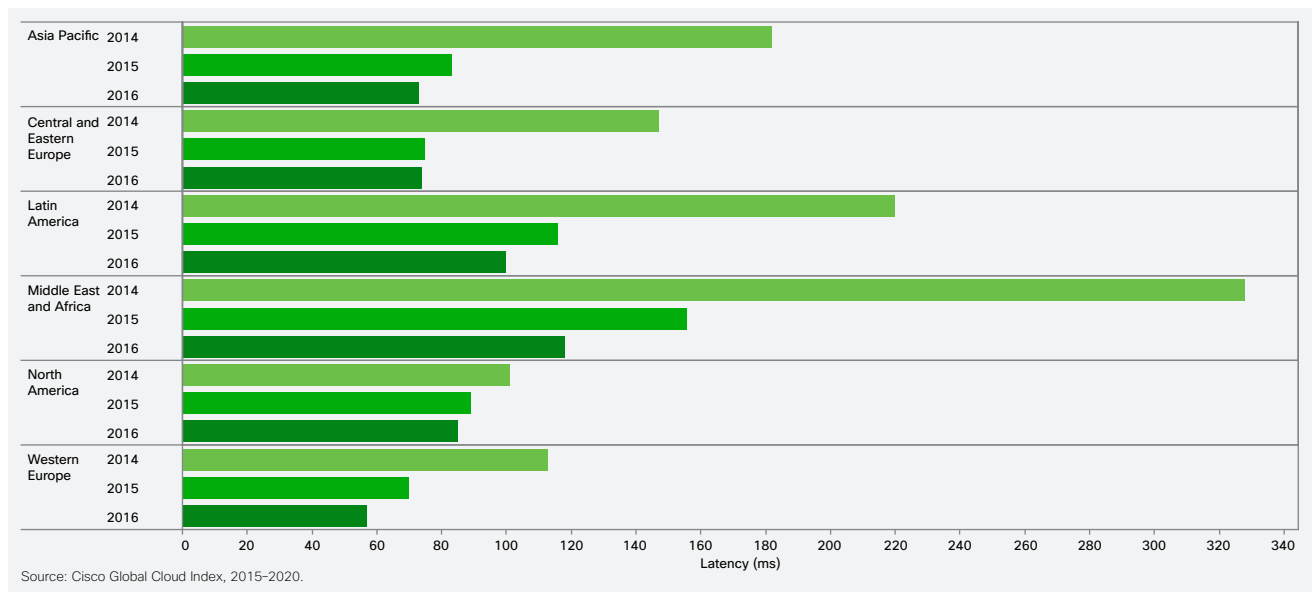**Figure 35.** Improvements in Average Fixed Latency in ms, 2014–2016

Source: Cisco Global Cloud Index, 2015–2020.

Figure 36. Improvements in Average Mobile Latency in ms, 2014–2016

Source: Cisco Global Cloud Index, 2015–2020.

For further details, refer to Appendix G and the Cisco GCI Supplement.

Conclusion

In summary, we can draw several main conclusions from the updated Cisco GCI 2015–2020 report.

Global data center traffic is firmly in the zettabyte era and will more than triple from 2015 to reach 15.3 ZB annually by 2020. Not only is the data center traffic growing, but it is also getting streamlined with architectural innovations such as SDN and NFV, which offer new levels of optimization for data centers. A rapidly growing segment of data center traffic is cloud traffic, which will nearly quadruple over the forecast period and represent 92 percent of all data center traffic by 2020. An important traffic enabler in the rapid expansion of cloud computing is increasing data center virtualization, which provides services that are flexible, fast-to-deploy, and efficient. By 2020, 92 percent of all workloads will be processed in the cloud.

Within the cloud segment private cloud will have significantly more workloads than the public cloud initially, but public cloud will grow faster than the private cloud over the forecast period, and by 2016 the majority share of workloads will transition to public cloud. This will also be depicted in the degree of virtualization: the workload density in public cloud will outpace that in private cloud by 2016 as well. As the business sensitivity to costs associated with dedicated IT resources grows along with demand for agility, we can see a greater adoption of public cloud by the businesses, especially with strengthening of public cloud security. Many enterprises will adopt a hybrid approach to cloud as they transition some workloads from internally managed private clouds to externally managed public clouds. All three types of cloud service delivery models—IaaS, PaaS, and SaaS—will continue to grow as more and more businesses realize the benefits of moving to a cloud environment.

Additional trends influencing the growth of data center and cloud computing include increasing digitization, the widespread adoption of multiple devices and connections or the IoE phenomenon, and the growth of mobility. An extraordinary amount of data is being generated by IoE applications—to the tune of 600 ZB by 2020. However, only a relatively very small portion of that content (about 6.2 ZB), will be stored. Over time, more and more of the data resident on client devices will move to the data center. Total data center storage capacity will grow nearly 5-fold from 2015 to 2020, growing from 352 EB in 2015 to 1.8 ZB by 2020. Cloud will accounts for nearly 90 percent of the installed storage capacity by 2020. To address rising user demands, cloud-based services such as consumer cloud storage are gaining momentum. By 2020, nearly 60 percent (59%) of the consumer Internet population will be using personal cloud storage. Growth in M2M connections and applications is also driving new data analytics needs.

This study also considers the importance of cloud readiness. Based on the analysis of regional average download and upload speeds and latencies for business and consumer mobile and fixed networks, all regions have made significant strides to reach a capable level of supporting basic and intermediate cloud services. The focus now turns to continuing to improve network capabilities to support the advanced cloud applications that organizations and end users expect and rely upon.

For More Information

For more information, visit www.cisco.com/go/cloudindex.

Appendix A: Data Center Traffic Forecast Methodology

Figure 37 outlines the methodology used to forecast data center and cloud traffic. The methodology begins with the installed base of workloads categorized by workload type and implementation and then applies the volume of bytes per workload per month to obtain the traffic for current and future years.

Figure 37. Data Center Traffic Forecast Methodology
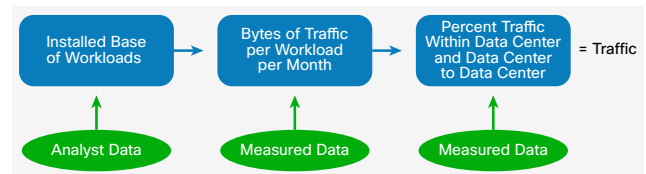
Analyst Data

Data from several analyst firms and international agencies (including Gartner, IDC, Juniper Research, Ovum, Synergy, ITU, and the United Nations) was used as inputs to the Global Cloud Index analysis. For example, analyst data was considered to calculate an installed base of workloads by workload type and implementation (cloud or noncloud). The analyst input consisted of server shipments with specified workload types and implementations. Cisco then estimated the installed base of servers and the number of workloads per server to obtain an installed base of workloads.

Measured Data

Network data was collected from a variety of enterprise and Internet centers. The architectures of the data centers analyzed vary, with some having a three-tiered and others a two-tiered architecture. For three-tiered data centers, data was collected from four points: the link from the access routers to the aggregation routers, the link from the aggregation switches or routers to the site or regional backbone router, the WAN gateway, and the Internet gateway. For two-tiered data centers, data was collected from three points: the link from the access routers to the aggregation routers, the WAN gateway, and the Internet gateway.

For enterprise data centers, any traffic measured northbound of the aggregation also carries non–data center traffic to and from the local business campus. For this reason, to obtain ratios of the volume of traffic carried at each tier, it was necessary to measure the traffic by conversations between hosts rather than traffic between interfaces, so that the non–data center conversations could be eliminated. The hosts at either end of the conversation were identified and categorized by location and type. To be considered data center traffic, at least one of the conversation pairs had to be identified as appearing in the link between the data center aggregation switch or router and the access switch or router.

In addition, as noted in this paper, the methodology for the estimation of cloud data center traffic has changed since the last release of the Cisco Global Cloud Index. The previous methodology included all storage traffic in the noncloud traffic category. The updated methodology includes storage traffic associated with cloud workloads in the cloud traffic category. For example, storage traffic associated with cloud application development would be counted as cloud traffic in the updated methodology, but would have been excluded in the previous methodology.

Appendix B: Global Cloud Index and Visual Networking Index

The Cisco Global Cloud Index (GCI) and Cisco Visual Networking Index (VNI) are distinct forecasts that have an area of overlap. The Cisco VNI forecasts the amount of traffic crossing the Internet and IP WAN networks, whereas the Cisco GCI forecasts traffic within the data center, from data center to data center, and from data center to user. The Cisco VNI forecast consists of data center-to-user traffic, along with non–data center traffic not included in the Cisco GCI (various types of peer-to-peer traffic).

The Cisco GCI includes data-center–to-user traffic (this area is the overlap with the Cisco VNI) data-center–to–data center traffic, and traffic within the data center. The Cisco VNI forecasts the amount of traffic crossing the Internet and IP WAN networks (Figure 38).

Figure 38. Cisco VNI and Global Cloud Index

Source: Cisco Global Cloud Index, 2015–2020.
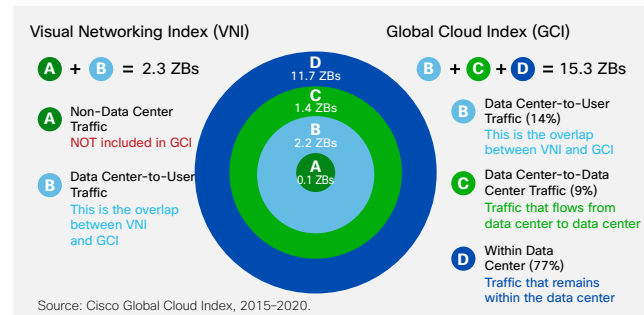
Appendix C: Regional Cloud Traffic Trends

The Cisco Global Cloud Index includes regional forecast data for cloud traffic growth (Table 6).

- In 2015, North America generated the most cloud traffic (1,891 EB annually), followed by Asia Pacific (908 EB annually) and Western Europe (718 EB annually).

- By 2020, North America will generate the most cloud traffic (3.6 ZB annually), closely followed by Asia Pacific (2.3 ZB annually) and Western Europe (1.5 ZB annually).

- From 2015 to 2020, the Middle East and Africa is expected to have the highest cloud traffic growth rate (41 percent CAGR), followed by Central and Eastern Europe (38 percent CAGR) and North America (33 percent CAGR).

Table 6. Cloud Traffic Growth by Region, in EB

Region	2015	2016	2017	2018	2019	2020	CAGR 2015–20
Asia Pacific	908	1,367	1,871	2,387	2,923	3,469	31%
Central and Eastern Europe	124	168	221	291	379	485	31%
Latin America	140	190	242	304	371	448	26%
Middle East and Africa	69	105	145	191	242	304	34%
North America	1,891	2,771	3,838	4,860	5,779	6,844	29%
Western Europe	718	1,035	1,395	1,768	2,157	2,528	29%

Source: Cisco Global Cloud Index, 2015–2020.

Appendix D: Workload Distribution by Region

Tables 7, 8, and 9 summarize data center workloads by type and region.

Table 7. Regional Distribution of Total Data Center Workloads, in Millions

Total Data Center Workloads in Millions							
	2015	2016	2017	2018	2019	2020	CAGR 2015–2020
Asia Pacific	48.6	68.1	91.3	115.7	39.7	163.4	27.4%
Central and Eastern Europe	5.2	6.6	8.3	10.0	11.6	12.9	20.1%
Latin America	6.5	8.4	10.6	13.0	15.1	17.1	21.3%
Middle East and Africa	4.7	6.1	7.8	9.5	11.0	12.4	21.6%
North America	78.5	99.1	124.1	149.3	170.8	189.1	19.2%
Western Europe	37.4	46.6	57.5	67.9	76.6	83.9	17.5%

Source: Cisco Global Cloud Index, 2015–2020.

Table 8. Regional Distribution of Cloud Workloads, in Millions

Cloud Data Center Workloads in Million							
	2015	2016	2017	2018	2019	2020	CAGR 2015–2020
Asia Pacific	38.3	57.0	79.6	103.6	127.6	151.5	31.6%
Central and Eastern Europe	4.0	5.5	7.3	9.1	10.7	12.2	25.1%
Latin America	5.2	7.2	9.6	12.0	14.1	16.2	25.5%
Middle East and Africa	3.6	5.1	6.9	8.7	10.3	11.8	26.6%
North America	57.2	77.9	103.6	129.5	152.3	172.1	24.6%
Western Europe	27.7	37.1	48.4	59.1	68.4	76.3	22.5%

Source: Cisco Global Cloud Index, 2015–2020.

Table 9. Regional Distribution of Traditional Data Center Workloads, in Millions

| Traditional Data Center Workloads in Millions							
	2015	2016	2017	2018	2019	2020	CAGR 2015–2020
Asia Pacific	10.3	11.1	11.7	12.1	12.1	11.9	2.9%
Central and Eastern Europe	1.2	1.1	1.0	0.9	0.9	0.8	-8.5%
Latin America	1.3	1.2	1.1	1.0	1.0	0.9	-7.4%
Middle East and Africa	1.0	1.0	0.9	0.8	0.7	0.6	-10.3%
North America	21.3	21.2	20.5	19.8	18.6	17.0	-4.4%
Western Europe	9.7	9.5	9.1	8.7	8.3	7.6	-4.8%

Source: Cisco Global Cloud Index, 2015–2020.

Appendix E: Cloud Definitions

Cloud Definition

The Cisco GCI aligns with the industry-standard cloud computing definition from the National Institute of Technology (NIST). The NIST definition lists five essential characteristics of cloud computing (Figure 39):

- On-demand self-service
- Broad network access
- Resource pooling
- Rapid elasticity or expansion
- Measured service

**Figure 39.** Essential Characteristics of Cloud
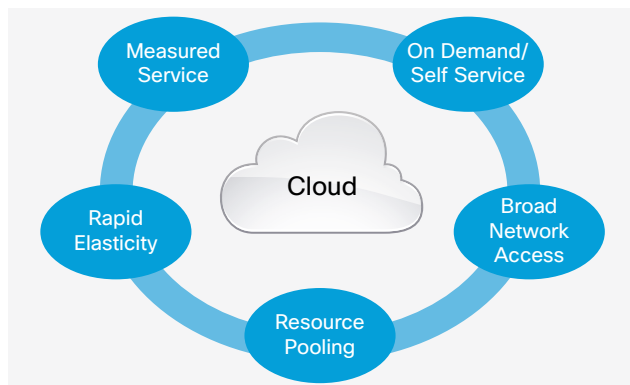
Cloud Deployment Models

Cloud deployment models include private, public, and hybrid clouds (or a combination of them). These distinct forms of cloud computing enable a variety of software, platform, and infrastructure services. Cloud data centers can be operated by service providers as well as private enterprises.

However, there is a slight variation from the NIST definition on how we classify private and public clouds. A cloud service could be public or private, depending on the demarcation line—the physical or virtual demarcation—between the public telecommunications network (WAN) and the private network of an organization (LAN) (Figure 40).

Figure 40. Cloud Deployment Models
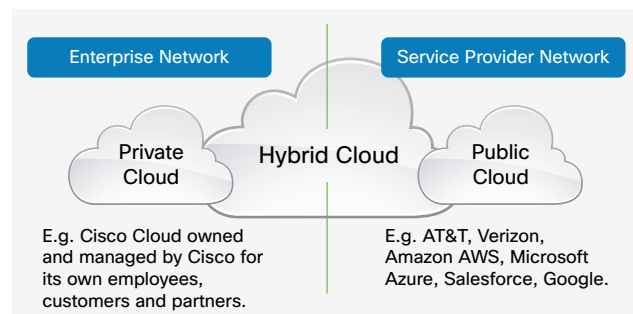
If the cloud assets lie on the service provider side of the demarcation line, then it would be considered a public cloud service. Virtual private cloud (VPC) falls in this category. Also the multitenant consumer cloud services would fall in this category.

If the cloud assets lie on the organization side of the demarcation line, then the service would be considered a private cloud service. In general, a dedicated cloud owned and managed by an organization's IT would be considered a private cloud.

Hybrid cloud, as the name suggests, is a combination of public and private clouds. In a hybrid cloud environment, some of the cloud computing resources are managed in-house by an enterprise and some are managed by an external provider. We define private and public as distinct categories; we do not separately break out the hybrid cloud because it is simply a superset of the private and public clouds in varying degrees.

Cloud Service Models (IaaS, PaaS, and SaaS)

The Cisco GCI forecast for cloud workload splits across the three main cloud services models: SaaS, PaaS, and IaaS (Figure 41). They are defined in line with NIST's definitions.

Figure 41. Cloud Service Models: IaaS, PaaS, and SaaS
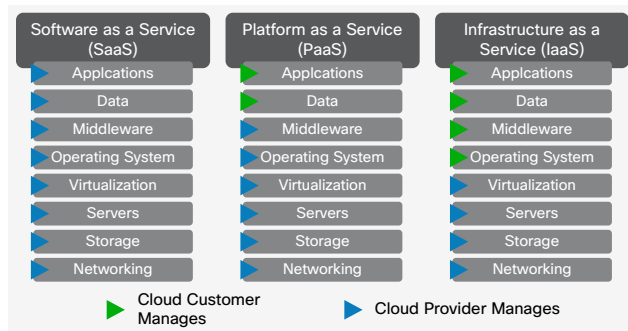
SaaS

The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin-client interface, such as a web browser (for example, web-based email) or a program interface. The consumer neither manages nor controls the underlying cloud infrastructure, including networks, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

PaaS

The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or -acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer neither manages nor controls the underlying cloud infrastructure, including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

IaaS

The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer neither manages nor controls the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (for example, host firewalls).

Appendix F: Workload Application Definitions

Following is the list of application definitions used for segmenting workloads in trend 4:

Enterprise Workload Applications

- Compute: mainly covers cloud IaaS
- Collaboration: email, conferencing, enterprise social networking, file sharing, content management
- Database/analytics/IoE: on-premises database apps, big data apps, business intelligence, data-oriented PaaS services
- Other ERP and enterprise apps: CRM, HCM, finance apps, storage, service/systems/operations management, and so on

Consumer Workload Applications

- Search: search
- Social Networking: Facebook, Twitter, Tencent, LinkedIn, Google+, Snapchat, and so on
- Video/streaming media: video sharing and video streaming services
- Other consumer apps: email, messaging, storage, file sharing, music services, e-commerce, news, and so on

Appendix G: Regional Cloud Readiness Summary

Table 10 summarizes cloud readiness for businesses and consumers by region, considering download and upload speeds, and latency. For more details, refer to the Cisco GCI Supplement.

Table 10. Regional Cloud Readiness

Network	Region	Average Download Speeds (Mbps)	Average Upload Speeds (Mbps)	Average Latency (ms)
Fixed	Asia Pacific	28.1	15.9	35
	Central and Eastern Europe	28.3	20.9	33
	Latin America	7.6	2.4	64
	Middle East and Africa	7.0	2.2	77
	North America	25.4	8.8	42
	Western Europe	22.8	7.4	44
Mobile	Asia Pacific	12.1	6.1	83
	Central and Eastern Europe	10.9	7.7	75
	Latin America	5.7	2.0	116
	Middle East and Africa	4.5	2.3	156
	North America	16.3	6.5	63
	Western Europe	13.7	4.8	70

Source: Cisco Global Cloud Index, 2015–2020.