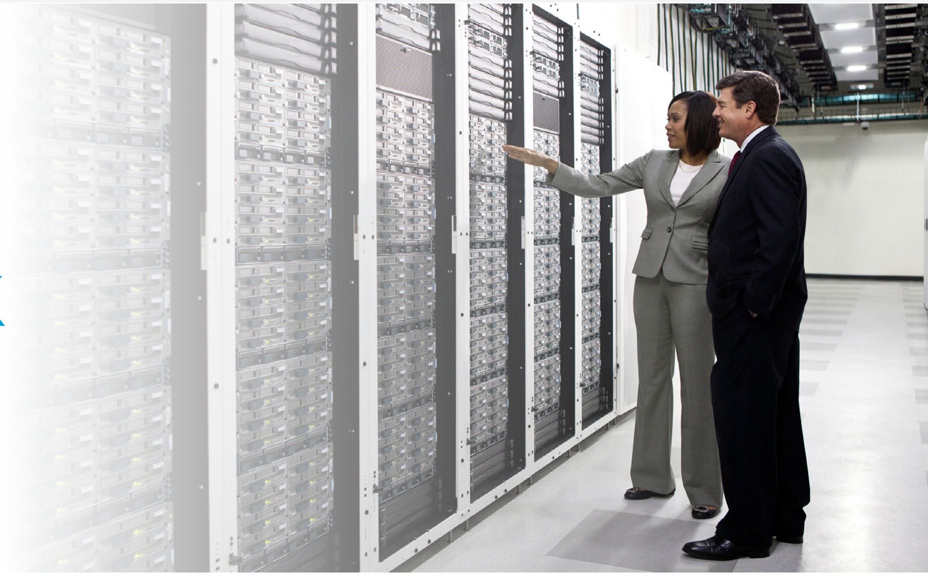
Cisco Global Cloud Index



Q What is the Cisco® Global Cloud Index (GCI), and how is it different from other Cisco IP traffic forecasts?

A Data center virtualization and cloud computing have become essential elements of business, education, government, and home communications and networking. Cisco conducts this study as an industry resource to provide IT professionals with new data to help them address increasingly complex data center operations and service delivery requirements. Understanding macro-level data center and cloud traffic trends can help organizations make strategic networking and business decisions. The Cisco GCI also measures and forecasts private network traffic, investigating trends within data centers, between data centers, and in content that ultimately travels from data centers to end users.

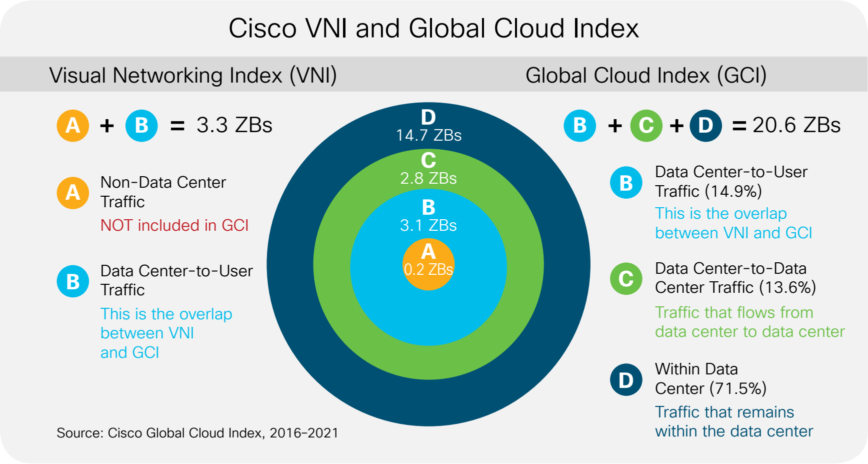
While the Cisco GCI and the other Cisco IP network traffic forecast (the Cisco Visual Networking Index™ [Cisco VNI™]) are distinct forecasts, there is some overlap (Figure 1). The Cisco VNI forecasts the amount of traffic crossing the Internet and IP WAN networks, while the Cisco GCI forecasts traffic within the data center, from data center to data center (east-west traffic), and from data center to user (north-south traffic). 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 is the overlap with the Cisco VNI), data center–to-data center traffic, and traffic within the data center. For more details about the Cisco VNI, see the Cisco Visual Networking Index: Forecast and Methodology, 2016–2021.

Q How does Cisco define a data center in the context of the GCI study?

A Within the GCI forecast, there are several data center types defined by their size and function. From small server closets to large hyperscale deployments, data centers deliver IT services and provide storage, communications, and networking to consumers and their growing number of networked devices, as well as business users and processes. As part of our analysis about where and how data center servers are deployed globally, these three general storage and computing environments are considered:

- **Server room or closet:** An air-conditioned room devoted to the continuous operation of computer servers. These servers are generally managed internally and designed to support the computing and storage needs of a single company or organization.
- **Service provider data center:** Many traditional service providers offer data center environments and services that deliver processing, storage, networking, management, and distribution of data within enterprises and small-to-medium businesses. These data centers might also support some consumer services and storage.
- **Hyperscale data center:** Independent analyst firm Synergy (a syndicated research source for the Cisco GCI forecast) has identified 24 hyperscale operators. To be a hyperscale cloud operator, a company must meet one of the following Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), Software-as-a-Service (SaaS), or other cloud service revenue requirements:
  - >$1B in IaaS/PaaS (for example, Amazon/AWS, Rackspace, NTT, IBM)
  - >$2B in SaaS (for example, Salesforce, Google, Microsoft, Oracle)
  - >$4B in Internet/search/social networking (for example, Facebook, Apple, Tencent, Yahoo)
  - >$8B in e-commerce/payment processing (for example, Amazon, eBay, Alibaba)

Figure 1. Cisco VNI and Cisco GCI



Q What is big data?

A For the purposes of this study, big data is defined 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. However, data science techniques can be used on data of any size, and the quality of insights achieved with data science is not necessarily related to the size of the underlying data.

Q How is cloud defined in the Cisco GCI?

A The Cisco GCI aligns with the industry-standard cloud computing definition from the National Institute of Standards and Technology (NIST). The NIST definition lists five essential characteristics of cloud computing: on-demand self-service, broad network access, resource pooling, rapid elasticity or expansion, and measured service. Deployment models include private, public, and hybrid clouds (or a combination of these). These distinct forms of cloud computing support 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 in how we define 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 and the private network of an organization.

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 into this category. Also, multitenant consumer cloud services would be included in this category. If the cloud assets lie on the organization side of the demarcation line, then it 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.

Q Does Cisco GCI consider any cloud service models and how are they defined?

A The Cisco GCI forecasts cloud workload and compute instance splits across the three main cloud services models: SaaS, PaaS, and IaaS. They are defined in line with NIST's definitions:

- **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 does not manage or control 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 consumer-created or -acquired applications onto the cloud infrastructure. The applications are created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure, including network, servers, operating systems, or storage. But the consumer does have 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 can deploy and run diverse software. The software can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications and possibly limited control of selected networking components, such as host firewalls.

Cisco GCI categorizes a cloud application (IaaS, PaaS, or SaaS) based on how the service is ultimately used by the user, regardless of other cloud services types that might be involved in the final delivery of the service. For example, if an SaaS cloud service depends on some aspects of other cloud services, such as PaaS or IaaS, such a workload and compute instance is counted as SaaS only. If a PaaS workload and compute instance operates on top of IaaS, such a workload and compute instance is counted as PaaS only.

Q What is a workload and compute instance, and why is it important to understanding data center and cloud traffic?

A A server workload and compute instance is defined as a set of virtual or physical computer resources that is assigned to run a specific application or provide computing services for one or many users. A workload and compute instance 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 this study, if a server is not virtualized, then one workload and compute instance is equivalent to one physical server. When there is virtualization, one virtual machine or a container, used interchangeably, is counted as one workload and compute instance. The number of virtual machines per server will vary depending on various factors, which include the processing and storage requirements of a workload and compute instance as well as the type of hypervisor being deployed. In cloud environments, both nonvirtualized servers and virtualized servers, with many virtual machines on a single virtualized server, are deployed. The increasing migration of workloads and compute instances from end-user devices to remotely located servers and from premises-based networks to cloud networks creates new network requirements for operators of both traditional and cloud data center environments.

Q What is the difference between a virtual machine and a container?

A Each virtual machine runs a full copy of an operating system and a virtual copy of all the hardware that the operating system needs to run. In general, virtual machines are more resource intensive than containers. Containers require a minimal subset of an operating system and only the necessary supporting programs, libraries, and system resources to run a specific program. Both virtual machines and containers can play effective roles in data center virtualization, but the scope of the computing or processing that each does is essentially what differentiates these solutions. Containers are generally scoped to a specific application or operating system (streamlined). Virtual machines, in contrast, have a broader scope, with the ability to support multiple operating systems and applications. However, a complete set of resources is required to support these robust capabilities.

Q What is a hybrid cloud?

A 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. Cloudbursting is one example of a hybrid cloud. In this case, daily computing requirements are handled by a private cloud, but when there are sudden spurts of demand, the additional traffic demand (bursting) is handled by a public cloud. For the purposes of the Cisco GCI forecast, private cloud and public cloud are defined as distinct categories. Hybrid cloud is not currently broken out as a separate category, because it is simply a superset of the private and public clouds in varying degrees.

Q What is the methodology behind the Cisco GCI?

A The Cisco GCI incorporates a bottom-up and top-down approach to derive global and regional results. The methodology begins with the server shipments to different types of data centers (traditional, private cloud, and public cloud), calculating the installed base of workloads and compute instances. Then it applies the volume of bytes per workload and compute instance per month to obtain the traffic for current and future years within the forecast period. We have used a variety of data sources, such as Gartner, IDC, Synergy, Juniper Research, and Ookla, for GCI analysis. For traffic modeling and verification of data center traffic types and volumes, network data from various enterprise and Internet data centers was reviewed. A global cloud traffic forecast is provided, as well as six regional forecasts (Asia Pacific, Central and Eastern Europe, Western Europe, Middle East and Africa, North America, and Latin America). For specific details about our forecast methodology, see the Cisco Global Cloud Index: Forecast and Methodology, 2016–2021.

Q What is the difference between a traditional data center and a cloud data center?

A The main differences are in levels of virtualization, standardization, automation, and security. Cloud data centers offer increased performance, higher capacity, and greater ease of management compared with traditional data centers. Virtualization serves as a catalyst for hardware and software consolidation, greater automation, and an integrated security approach.

Q How does the Cisco Global Cloud Index differentiate between cloud traffic and noncloud traffic?

A Cloud traffic can be identified as the traffic generated from cloud servers and workloads and compute instances. Cloud traffic is generated as a result of cloud services: easily deployed services that are accessible through the Internet, have elastic and scalable provisioning and usage-based pricing, and can be delivered on demand. Cloud traffic is measured and then subtracted from total data center traffic to obtain noncloud traffic estimates.

Q What is meant by cloud readiness, and what characteristics are used to assess regions' ability to support cloud services?

A The cloud readiness segment of the Cisco GCI study offers regional and some country-level views of the fundamental performance factors required for broadband and mobile networks to deliver next-generation cloud services. Increased adoption of business-grade and consumer-grade cloud computing often depends on these enhancements and the reliability of these performance factors. For instance, it is important for consumers to be able to download music and videos on the road, as well as for business users to have continuous access to videoconferencing and mission-critical Customer Relationship Management (CRM) and Enterprise Resource Planning (ERP) systems. Download speeds, upload speeds, and latencies are vital metrics when assessing the network capabilities of cloud readiness.

More than 300 million records from Ookla, along with inputs from Ovum/Informa, Synergy Research, PointTopic, United Nations (UN), World Bank, NetCraft, International Telecommunication Union (ITU), Internal Labor Organization (ILO), and other sources, from more than 200 countries were analyzed to understand cloud readiness. Regional cloud readiness values (calculated as an average of country-level values within a particular region) are included in the Cisco Global Cloud Index: Forecast and Methodology, 2016–2021. Individual countries might have slightly or significantly higher or lower averages compared to the regional averages for download speed, upload speed, and network latency. For country-level data, refer to the Cisco GCI Supplement: Cloud Readiness Regional Details. The major cloud readiness characteristics and performance factors included in this study are as follows:

- **Internet ubiquity:** This indicator measures fixed and mobile Internet penetration while considering population demographics to assess the pervasiveness and expected connectivity in various regions.
- **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 and containers, CRM and ERP cloud platforms for businesses, and video download and content retrieval cloud services for consumers.
- **Upload speed:** With the increased adoption of virtual machines, tablets, and videoconferencing 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.
- **Network latency:** Delays experienced with Voice over IP (VoIP), viewing and uploading videos, online banking on mobile broadband, or viewing hospital records in a healthcare setting are caused by high latencies (usually reported in milliseconds). Reducing delay in delivering packets to and from the cloud is crucial to delivering today's advanced services.

The study has traditionally focused on average or mean download, upload, and latency characteristics. However, to better understand the distribution of speeds within a country, the median download speed, median upload speed, and median latency are included in the study. In most countries, median speeds are lower than mean or average speeds. This is caused by the higher occurrence of lower speeds in the lower 50th percentile, compared to the longer tail of distribution of the higher speeds.

For any set of numbers, the median is the midpoint, where half the numbers are lower and half the numbers are higher. The average of a set of numbers is the total of all the numbers in the set, divided by the number of items in that set. For further details, see the sample speed distribution curve in Cisco GCI Supplement: Cloud Readiness Regional Details.

Q Are there any other factors, besides those already listed, that might affect the end-user cloud experience?

A Yes, besides the upload and download speeds and latency of the Internet Service Provider (ISP) network, the location of the content providers' servers or Content Distribution Network (CDN) and their distance from the ISP network are factors. It is estimated that latency increases by 1 millisecond for every additional 150 miles travelled by a video stream.

Q May I or my organization or company use or publish Cisco GCI forecast data?

A Yes. Cisco welcomes and encourages press, analysts, service providers, and other interested industry parties (business, regulatory, or academic) to use or publish the data. We do require that proper Cisco attribution be given for any and all Cisco GCI data that is published or shared in private or public, print and electronic forms (for example, source: Cisco Global Cloud Index [or Cisco GCI], 2016–2021). No further signatures or consents are required to reference our publicly available white papers and reports. We are always interested in the context in which our data is used and would appreciate it if parties that use our content would share copies of their completed work containing Cisco GCI insertions. You can find answers to frequently asked questions or connect with analysts through the VNI/GCI community page.

Q Can you share the data center and cloud data you used to construct the Cisco GCI forecast projections?

A We are not able to share the specific source data that serves as a primary input to our forecast methodology.

Q Where can I direct questions about the Cisco GCI forecast?

A Cisco GCI forecast questions may be submitted through the VNI/GCI community page.

C67-738065-02 01/18