Realtime publishers

The Definitive Guide to Cloud Acceleration

Dan Sullivan

sponsored by

CDNetworks

Copyright Statement

Chapter 4: Multiple Data Centers and Content Delivery

Deploying applications to a global user base involves several technical as well as many potential legal issues. When you support a user base on multiple continents, you have all the requirements of a more localized user base—reliable access to the application—as well additional considerations, such as concerns about latency and the impact of network performance on application response time. Although content moves easily to any point on the Internet, local regulations and laws introduce a layer of complexity to the deployment of applications and services internationally. Content that is readily available in Western Europe or North America may be restricted in China, for example. As you consider how to optimize your application performance and content delivery for a global user base, keep in mind legal requirements as well. Ideally, you will be able to deploy an application and network solution that addresses both technical and legal requirements within an integrated service.

This chapter examines several topics related to data centers and content delivery:

- Appeal of deploying multiple data centers
- Challenges to maintaining multiple data centers
- Data centers combined with content delivery networks
- Country-specific issues to multiple data centers and content delivery, particularly in China

Replicating data across multiple data centers is part of a solution for cloud application acceleration, but more than simply the addition of data centers is required to meet the needs of a global user base.

Appeal of Deploying Multiple Data Centers

There are many benefits to deploying multiple data centers to support enterprise applications, particularly redundancy and failover advantages as well as the potential for reduced latency. Resiliency is a crucial feature of enterprise applications. Customers, employees, contractors, and business collaborators expect to have access to the applications and data they need. At the same time, applications are subject to a wide range of potential disruptive events:

- Application maintenance
- Data loss
- Network disruption
- Disruption in environmental controls in data centers

Hosting your data and application in multiple data centers can help mitigate the impact of each of these events (see Figure 4.1).

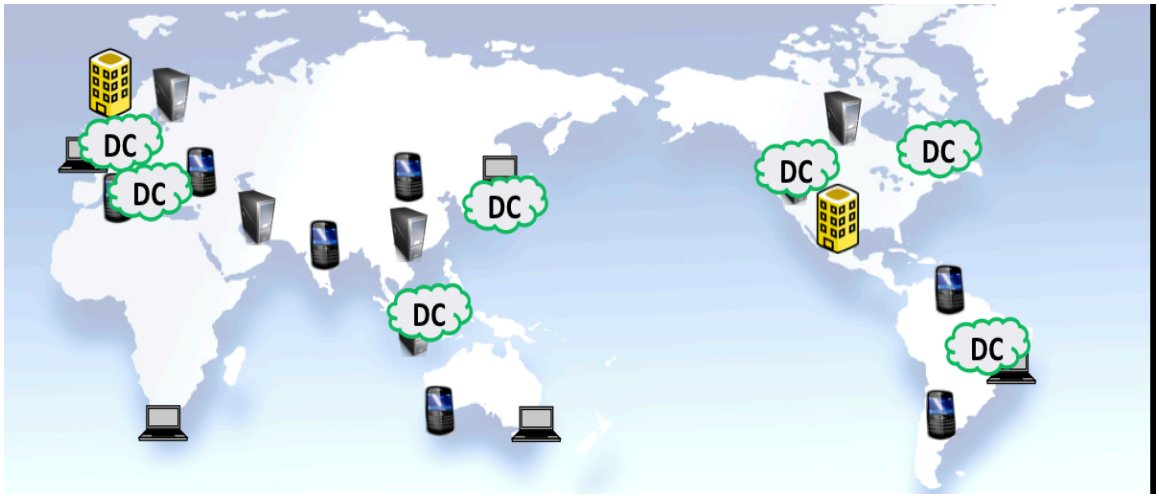


Figure 4.1: The implementation of multiple data centers offers a number of advantages, including resiliency to several potential problems (Image source: CDNetworks).

Application Maintenance

Applications have to be shut down for maintenance, sometimes due to needed upgrades or patches to an operating system (OS) or application code. In other cases, equipment in the data center needs to be replaced or reconfigured, making them unavailable to support your enterprise application. If the application and data are replicated to another data center, user traffic can be routed to the alternative data center, allowing users to continue working with the system.

This type of replication can also be done locally, within a data center. Failover servers and redundant storage arrays can allow for resiliency within the application. The most obvious risk with this approach, however, is that a data center–wide disruptive event would render the failover servers and storage inaccessible.

Data Loss

Data loss can occur as a result of hardware failures, software errors, and user mistakes. Depending on the type of failure, having data and applications replicated in multiple data centers can aid in recovery.

Data Loss Due to Hardware Failure

Consider the case of hardware failing within one data center and corrupting a set of data. The data might have been altered or may be unrecoverable because of a hardware failure. In such cases, it is highly unlikely that the hardware in other data centers experienced the exact same type of failure and lost the same data.

> **Note**
>
> This scenario ignores the highly improbable but theoretically possible case of a design flaw in the hardware that fails in exactly the same way, at the same time, and with the same data.

If the data had been replicated in at least one other data center, the data could be recovered and restored in the data center that is experiencing a failure. This scenario would be comparable to restoring data from a backup. There would, of course, be a time delay between the time of the failure and the time that data is restored. In cases where the time between failure and restoration must be as short as possible, application designers can replicate both data and applications between data centers. In this circumstance, in the event of a hardware failure, users would be routed to an application running in the alternative data center. Users might experience degraded performance due to an increased number of users on the application servers; they might also face longer network latency if the alternative data center is significantly farther away then the location of the failed hardware data center.

Data Loss Due to Software Failure and Human Error

Data loss due to software failure and human error can be more difficult to address than data loss due to hardware failure. The additional challenge stems from the fact that similar, if not identical, software runs in multiple data centers when applications are replicated. The primary reason to run multiple instances of applications in multiple data centers is to have identical, redundant systems that can take over the workload of the other in the event of a failure. This setup requires identical or near identical configurations, which creates the root of a potential problem: a significant software error in every instance of an application.

Software errors can corrupt data in many ways. A database operation designed to update a particular set of records might unintentionally update more than the target data set. An error in writing data to disk can cause a pointer to a data structure to be lost, leading to unrecoverable data. A miscalculation could write the wrong data to a database. In each of these cases, a data loss occurs and, unlike hardware-related data loss, the same type of error is likely to occur in replicated instances as well.

Application designers can plan for many types of potential human errors, from invalid data entry to changes that violate business rules. It is difficult to distinguish an intentional change from an unintentional change when data validation rules, business rules, or other criteria for assessing users' actions are not violated. Changes made by users that do not violate application rules are considered valid; by default, they will be accepted and eventually replicated to other instances of the data stores.

There are advantages to having multiple data centers to mitigate the risk of data loss. However, it is important to understand the limitations of this strategy's ability to recover from different causes of data loss.

Network Disruption

Network disruption at a data center level can adversely impact a large number of users. The importance of reliable access to the Internet for a data center cannot be overstated. Data centers typically contract with multiple Internet providers for access to the Internet. If one of the providers experiences a network disruption, traffic can be routed over the other providers' connections. The assumption, of course, is that the redundant services of multiple providers will not fail at the same time.

This assumption is reasonable for the most part except when you consider major disruptions due to natural disasters. Severe storms that disrupt power for days or earthquakes that damage cables can leave entire data centers disconnected from the Internet for extended periods.

Avoiding areas prone to natural disasters can be a challenge. As Figure 4.2 shows, areas of high risk for seismic activity exist in the United States on the West coast, in the Midwest, and in small areas of the Southeast. The Midwest and Gulf Coast are, in general, at low risk of seismic activity but are prone to tornadoes and hurricanes, respectively. For this reason, using multiple data centers located in areas with different risk profiles is a reasonable approach to mitigating the risk of data center–level network disruptions.

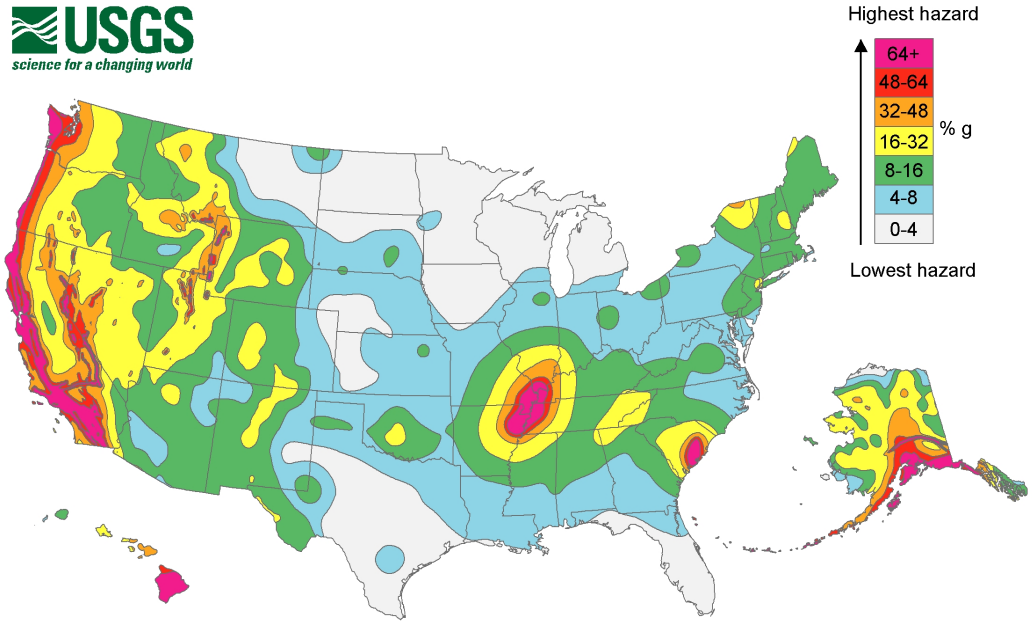


Figure 4.2: Seismic hazard map of the United States indicates locations of highest risk on the West Coast and parts of the Midwest and the Southeast (Source: Earthquake.usgs.gov).

Disruption of Environmental Controls in Data Center

There is an old proverb dating back at least to the 14th century that starts "For want of a nail, the shoe was lost. For want of a shoe, the horse was lost…" and ends with a line about the loss of a kingdom. The gist of the proverb is that small events can have large consequences. This same kind of potential problem exists in data centers with regards to environmental controls.

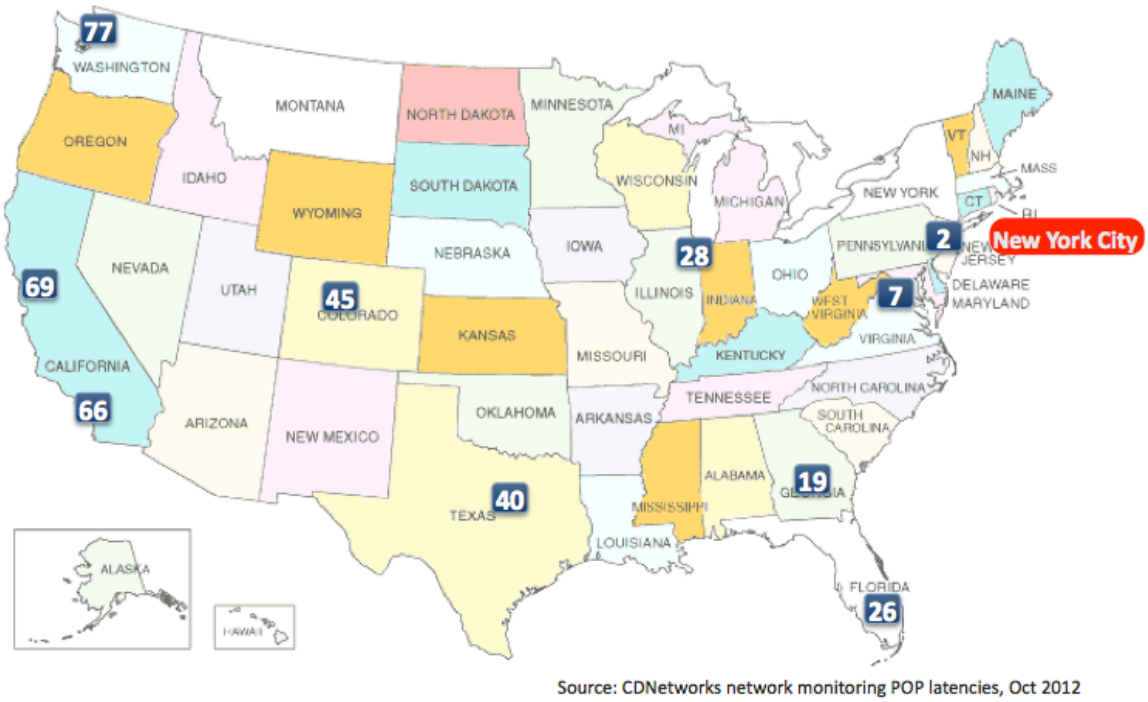
Managing environmental controls, especially cooling, is an important part of data center operation management. When you consider risks to data center operations, it is easy to focus on the major components, such as server hardware and network connections. However, it is important to not forget the less visible but still important elements such as environmental controls—the proverbial nails that hold together data centers. Failures in environmental controls can lead to a partial shutdown of hardware until environmental controls systems are functioning again.

Redundant data centers can help mitigate several risks of service disruption, ranging from application maintenance and software failure to natural disasters that destroy data centers and loss of environment controls that diminish operating capacity. The advantages of multiple data centers are not limited to minimizing the impact of disruptions.

Reduced Latency

Latency, or the round-trip time from a server to a client device and back to the server, can vary significantly for different users. As Figure 4.3 shows, a customer in the Pacific Northwest accessing an application hosted in New York will experience latencies almost three times that of a customer in Florida.

One way to address these wide variations in latency is to deploy applications to multiple data centers and architect applications to serve customers from the closest data center.



Source: CDNetworks network monitoring POP latencies, Oct 2012

Figure 4.3: Latencies within a country can be substantially different for customers in different locations when those customers are served from a single data center (Image source: CDNetworks).

Clearly there are advantages to having multiple data centers host your applications. Multiple data centers provide for redundancy and improve the resiliency of applications. Hardware failures, network disruptions, hardware-based data loss, and risks of natural disasters can all be addressed to some degree with multiple data centers. It would seem obvious that we should all deploy large-scale, mission-critical applications to multiple data centers, but there are drawbacks. It is often said that there are no free lunches in economics. Similarly, there are no free solutions to IT.

Challenges to Maintaining Multiple Data Centers

Some of the challenges to maintaining data centers are well known. Others, especially with respect to addressing content delivery challenges, are not as obvious. In the interest of a broad assessment of the challenges facing businesses and organizations considering multiple data center deployments, the following list highlights factors to consider:

- Costs of data centers/rising costs of colocation
- Need for specialized expertise
- Software errors
- Synchronization issues
- Unaddressed content delivery challenges

Outlining each of these challenges will reveal how deploying multiple data centers might not be sufficient to meet global content delivery needs.

Costs of Data Centers/Rising Costs of Colocation

The costs of data centers can be divided into the initial costs to construct and equip and the ongoing costs to operate. Construction costs will vary by location for any type of building construction, but data centers have specialized requirements. For example, compared with typical buildings, buildings that house data centers require higher levels of security, substantial cooling systems, redundant power supply lines, and depending on the area, remediation for regional natural disaster risks. For example, data centers built in areas with substantial risk of seismic activity may require earthquake resistant structures.

Figure 4.4: Data centers, such as this one housed in Oregon in the US, are major investments to both build and operate (Source: By Tom Raftery (Flickr) [CC-BY-SA-2.0 (http://creativecommons.org/licenses/by-sa/2.0), via Wikimedia Commons).

Operating costs in a typical data center are dominated by the costs of servers and power. Networking equipment, power distribution systems, cooling systems, and other infrastructure are also major cost categories.

These costs have to be weighed against the benefits of deploying multiple data centers, which are, as previously described, are substantial. There are alternative ways to achieve some of the same benefits, especially with regards to content delivery, without incurring the substantial step-wise costs of adding data centers.

Need for Specialized Expertise

A fully staffed data center requires an unusual combination of skills. The physical plant aspects of a data center require individuals knowledgeable in heating ventilation and cooling (HVAC), power systems, and physical security. Network professionals working in a data center need the skills to configure complex networks, analyze performance data, and tune configurations. A data center might support the use of multiple hypervisors and host operating systems (OSs), which requires knowledgeable systems administrators.

IT managers will be in demand as well. Teams of professionals may be needed to diagnose and resolve problems in a data center. For example, power fluctuations can adversely affect cooling systems, which in turn can cause problems for servers that first come to the notice of systems administrators. Coordinating ad hoc teams to address one-time problems is just part of a manger's responsibility. Ongoing operations require services such as a Help desk as well as the ability to communicate with business process owners who might have limited knowledge of data center operations.

Even when businesses have the resources to absorb the costs of data center construction, maintenance, and management, the deployment of multiple data centers does not always address global reach network issues.

Software Errors

Complex systems, such as data centers and cloud infrastructures, are subject to failure. Even when systems are designed for resiliency, they can experience problems due to software errors or unanticipated cascading effects that propagate through a complex system.

Consider some of the major cloud service outages in the past several years as examples of what can go wrong:

- In December 2012, Amazon Web Services experienced an outage in its US eastern data center due to a problem with the Elastic Load Balancing service.
- A software bug forced Google to resort to backups to restore mail for 150,000 users whose data had been lost in one of a series of outages between 2008 and 2009.
- In February 2013, Microsoft Azure storage service was unavailable due to problems with expired digital certificates

Amazon, Google, and Microsoft are all major cloud providers with sufficient resources to deploy and manage multiple data centers. Yet even these major providers experience substantial disruptions due to software errors.

Synchronization Issues

When data has to be available in multiple data centers, you have to contend with synchronization issues, including both technical and cost considerations. Synchronizing large volumes of data can consume substantial amounts of bandwidth. Like the data sent back and forth to end users of applications, data sent to other data centers is subject to network congestion, long latencies, and lost packets. Sub-optimal network conditions can lead to extended synchronization times.

In a worst-case scenario, delays in synchronizing data can adversely affect application performance. For example, a server with stale data could report inaccurate information to a user, while another user issuing the same query but receiving data from a server in a different data center might get the correct, up-to-date information.

Unaddressed Content Delivery Challenges

In addition to the challenges of deploying multiple data centers, businesses still have to contend with the fact that not all their content delivery needs are met by this setup. In spite of deploying the best technology deployed in state-of-art data centers that are managed by highly skilled professional, these businesses still have to address issues like non-synchronized content, lost packets, and inefficient TCP/IP traffic.

Some content requires a single source. For example, airline tickets must have a single source record so that a seat on a flight is not sold more than once. In these types of applications, multiple data centers can only be used if specialized techniques, such as two-phase commits, are employed.

Lost packets are often a product of congested networks. Congestion can be episodic—for example, during periods of peak demand for bandwidth—or it might be chronic—for example, due to peering agreements that lead to less than optimal amounts of traffic across networks. Deploying applications to multiple data centers can help avoid or minimize the number of hops a packet must travel, but when data packets travel over congested networks, the packets are at risk of being lost.

Perhaps one of the most significant unaddressed challenges of deploying multiple data centers is the fact that traffic might still be using non-optimized TCP configurations.

These unaddressed needs and disadvantages to multiple data centers are not presented as deterrents from employing multiple data centers. Instead, the point is made to use a more balanced approach that combines the benefits of multiple data centers with the benefits of a content delivery network that maximizes the pros of the two while limiting the costs and disadvantages of each.

Combining Data Centers, Content Delivery Network, and Application Acceleration

Constructing and running multiple data centers addresses some of the significant challenges businesses face, especially the need for redundancy. Data centers can help improve latency in some cases, but the overall benefit may not be sufficient for that alone to justify the capital and operational costs of data centers. Instead of relying on a single method to address multiple problems (for example, latency and redundancy), an optimal solution is based on combining a small number of data centers with the use of content delivery networks and application acceleration techniques.

Multiple data centers provide redundancy but at a substantial cost and increased complexity. Nonetheless, having at least two data centers is difficult to avoid if you are looking to maintain application availability in the event of catastrophic failure at a single data center. One can reasonably ask whether one backup data center is enough. There could conceivably be catastrophic failures at two data centers or a catastrophic failure at one and a less significant but still performance-degrading event at the other data center. The right solution depends on your requirements and tolerance for risk.

If your risk profile allows for two data centers rather than more, you can reduce the overall capital and operational expenses for data centers. Two data centers employing a geo-load-balancing method can share the application traffic between the two data centers. This setup will likely reduce the latency for some users in close proximity to the data center but in general, data centers do not solve the latency problem. As this scenario considers only two data centers, the number of users benefiting from this arrangement is limited.

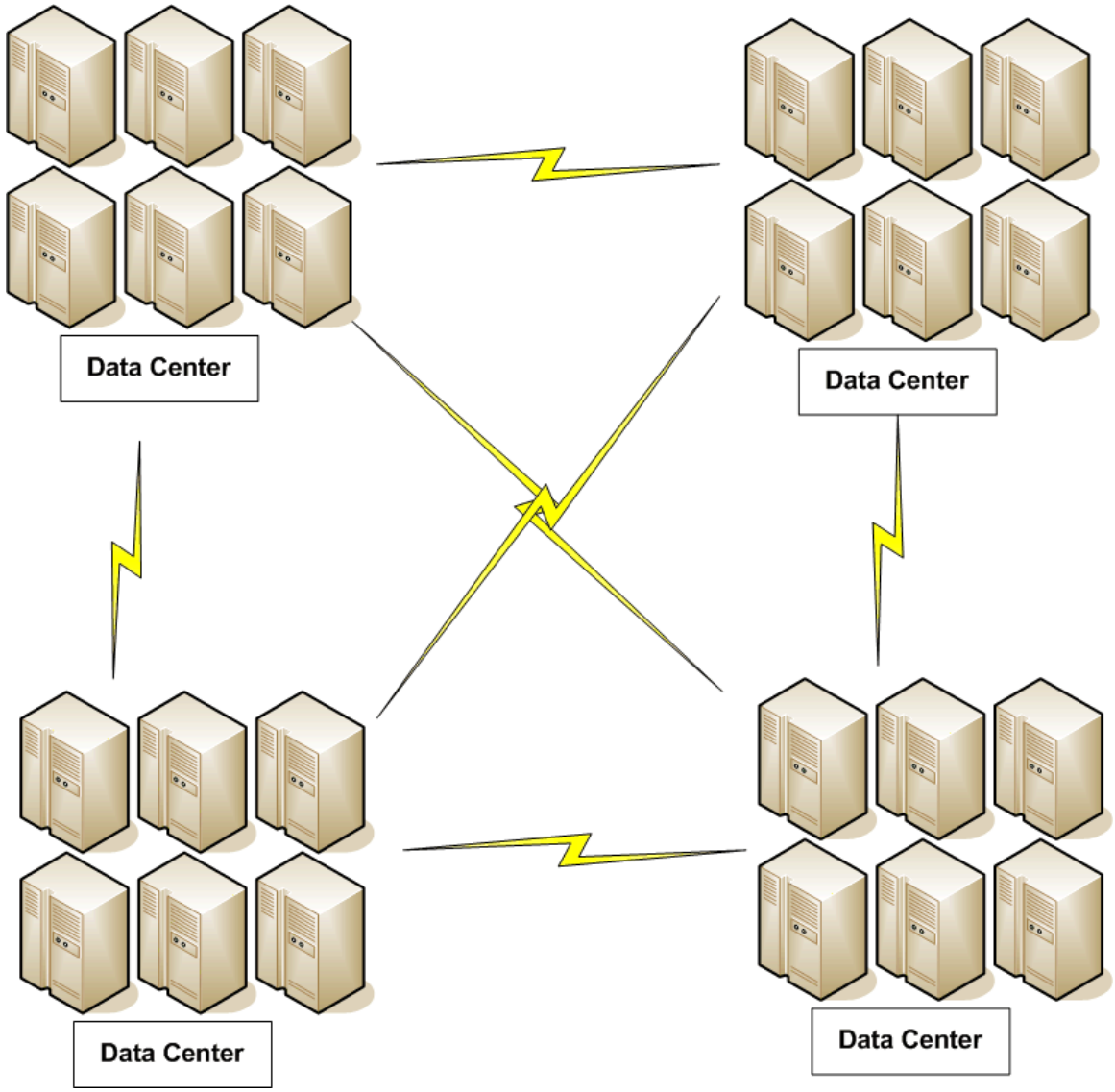


Figure 4.5: Multiple data centers increases redundancy but at additional cost and management complexity.

If you combine two data centers with caching of static content along with dynamic application traffic optimization techniques focused on TCP and HTTP, you can improve end user experience by enabling lower latency along with improved availability.

Optimizing Network Traffic in the Middle Mile

Network traffic moves from servers in the data center and across the Internet until it reaches the target device. The segment of the network from the Internet Service Provider's (ISP's) facility to the end user's device is known as the "last mile." The speed of the network in the last mile is dependent on the type of medium in place locally to transmit signals. The segment of the network between the source data center and an the edge of the network prior to the last mile is known as the "middle mile."

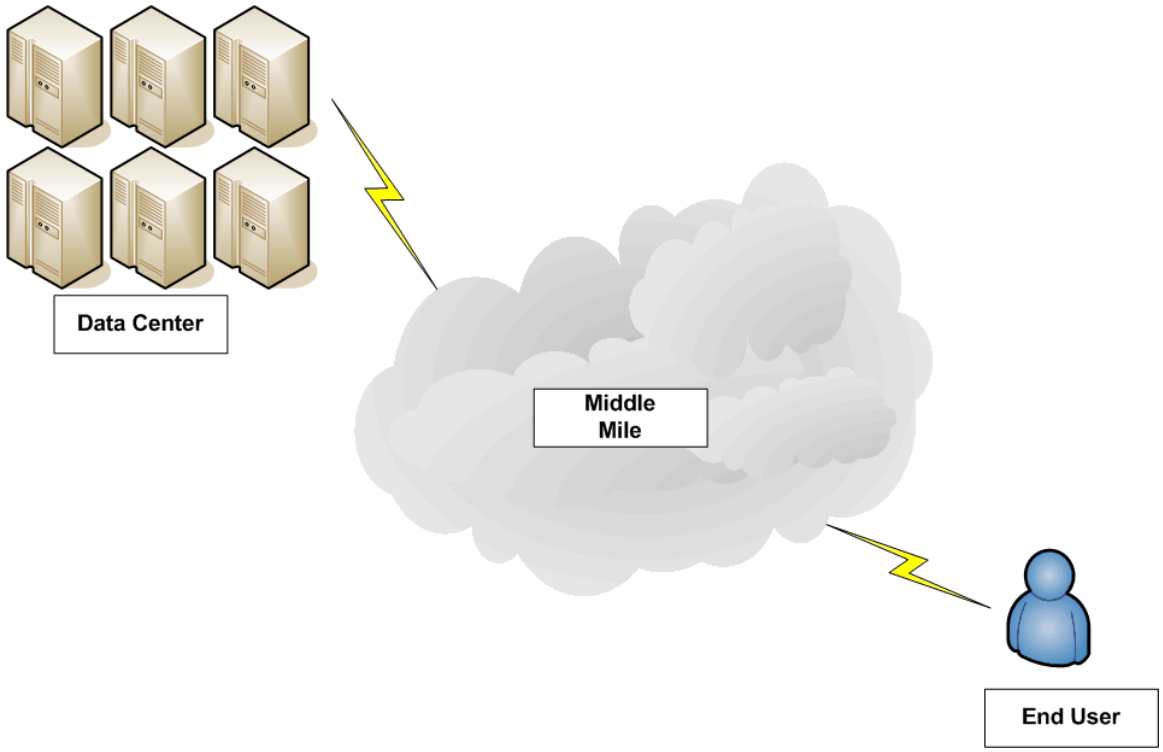


Figure 4.6: The middle mile between data centers and the edge of the last mile can be optimized to improve TCP performance. The first mile is the segment originating at the data center while the last mile is the segment ending at the end user.

The distance that packets must travel in the middle mile can be significant. Anything that can reduce the number of packets sent can help improve performance. These techniques include:

- Data compression to reduce the amount of payload data
- TCP optimizations to reduce the number of round trips made
- HTTP optimizations to reduce connection management overhead

Content delivery networks and application delivery network services can use these and other techniques to optimize traffic between data centers.

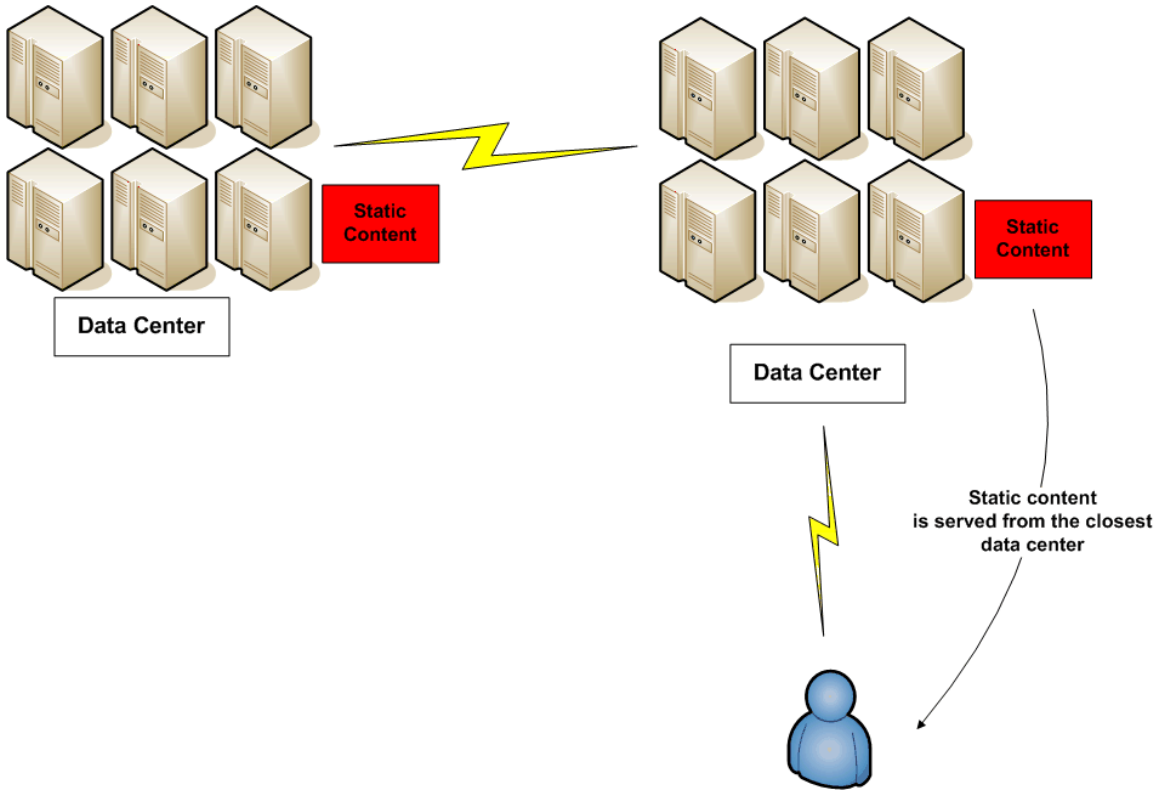


Figure 4.7: Static content is cached at data centers. When a user requests content, it is served from the closest CDN data center. If the content is not currently in the cache, it is requested from the origin data center.

Data compression can help reduce the number of packets that must be sent between the data center and client devices by reducing the size of the payload. TCP breaks a stream of data into individual units of length that depend on the medium; for example, Ethernet packets can be as long as 1500 bytes in length, including header information. The media defines the maximum transmission unit on a network, so a valuable technique is to compress the payload data before it is transmitted.

TCP has a number of configuration parameters that allow for tuning. Characteristics such as buffer size and the settings controlling the retransmission of lost packets can be adjusted to maximize throughput. For example, one tuning technique employs a more efficient way of detecting a lost packet and reduces the number of packets that might be retransmitted unnecessarily.

These types of techniques are especially important when optimizing non-static, application traffic. Static pages can be cached by content delivery networks and served to users within their region. Application-generated content, such as responses to database queries, search results, or reporting and analysis tool output, will vary from one user to another.

Caching

Acceleration techniques reduce latency and packet loss between data centers, but there are additional benefits of caching.

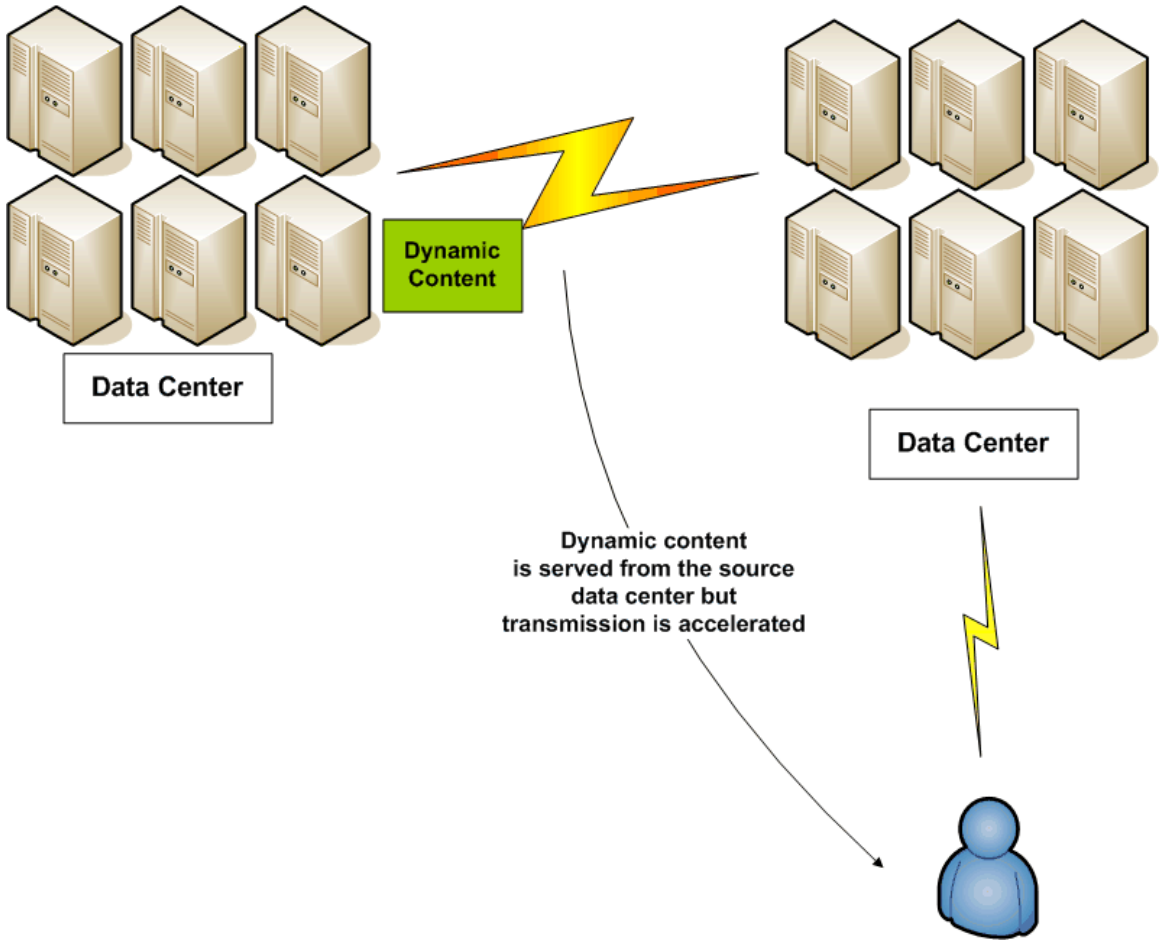


Figure 4.8: Dynamic content cannot be cached and must be sourced from the data center hosting the application generating the content. Acceleration improves throughput and allows for a more responsive application experience. Here, the user has accelerated access to dynamic content from a distant data center and fast access to static content from a closer data center.

Load Balancing

By using a content delivery network, the workload for serving content is distributed to multiple points of presence around the globe. Users in different parts of the world viewing the same Web page will see the same content even though that content is served from different locations. This load balancing is known as geo-load balancing and is one type of load balancing supported by content delivery network providers.

In addition to geo-load balancing, content delivery network providers can load balance within a point of presence. Clusters of servers can be deployed to serve static content so that a single Web server does not become a bottleneck. It would be unfortunate if after globally distributing your content and deploying TCP optimizations and other acceleration techniques, a single Web server slows the overall throughput of your application.

Monitoring Server Status

Content delivery network providers can monitor server status and load and adjust the server resources dedicated to your application as needed. For example, during periods of peak demand, additional servers can be added to a load-balanced cluster and when demand subsides, those extra servers can be returned to a resource pool of cloud resources.

Fault Tolerant Clusters of Servers

Clusters of servers used for content delivery can be configured as a fault tolerant cluster. In the event of a failure in one of the servers, the other servers will automatically process content requests that under normal operating conditions would be processed by the failed server.

Virtual IP Address and Network Failover

Servers are not the only potential point of failure. Network failures that leave some servers inaccessible could disrupt content distribution. Using virtual IP addresses allows content delivery network providers to create a more fault tolerant network and route traffic to accessible servers.

Benefits of Multiple Data Centers, Content Delivery Networks, and Application Delivery Acceleration

The combination of multiple data centers, content delivery networks, and application delivery acceleration addresses significant challenges to delivering static and dynamic content in a reliable, low latency way, including:

- Redundant data centers provide for reliable access to content in the event of a failure at another data center
- Acceleration techniques reduce latency and packet loss resulting in more responsive applications from the end users' perspective
- Acceleration techniques reduce the time required to keep static content synchronized across data centers
- Content delivery networks maintain copies of static content in multiple data centers, reducing the distance between end users and content
- Content delivery providers can offer load balanced and fault tolerant clusters of servers
- Content delivery providers can monitor server status and the load on systems and adjust resources as needed to maintain acceptable performance levels

These benefits apply to most content delivery use cases but they do not capture all the challenges an organization faces when distributing content globally. China, in particular, presents additional requirements that are worth considering as you evaluate content delivery and application delivery acceleration providers.

China: Country-Specific Issues in Content Distribution

It is difficult to describe China without using superlatives. China has the world's largest population and occupies the second largest landmass of any country. It has the largest number of Internet users in Asia. Since 2000, its gross domestic product has grown between 8% and 11.4% per year (Data source: http://www.chinability.com/GDP.htm). The business opportunities presented by China is well documented and is a key driver motivating businesses to provide their goods and services to Chinese markets. An important part of accessing markets and delivering goods and services is providing access to information and applications.

Figure 4.9: China presents many business opportunities but regulation and cultural differences need to be considered (Source: By Cacahuate, amendments by Peter Fitzgerald and ClausHansen (Own work based on the map of China by PhiLiP) [CC-BY-SA-3.0-2.5-2.0-1.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons).

Technical Challenges to Delivering Content in China

Content delivery considerations are only a part of a strategy for doing business in China, but they should not be dismissed. China is geographically large, which can mean long latencies between distant cities.

As with Internet service in any part of the globe, peering relationships can substantially affect latency and packet loss. ISPs in China may have peering agreements with other ISPs that lead to less than optimal routing of network traffic. This can lead to longer latencies and increased packet loss due to congestion. Due to issues with peering, businesses may experience less than 100% availability of network devices.

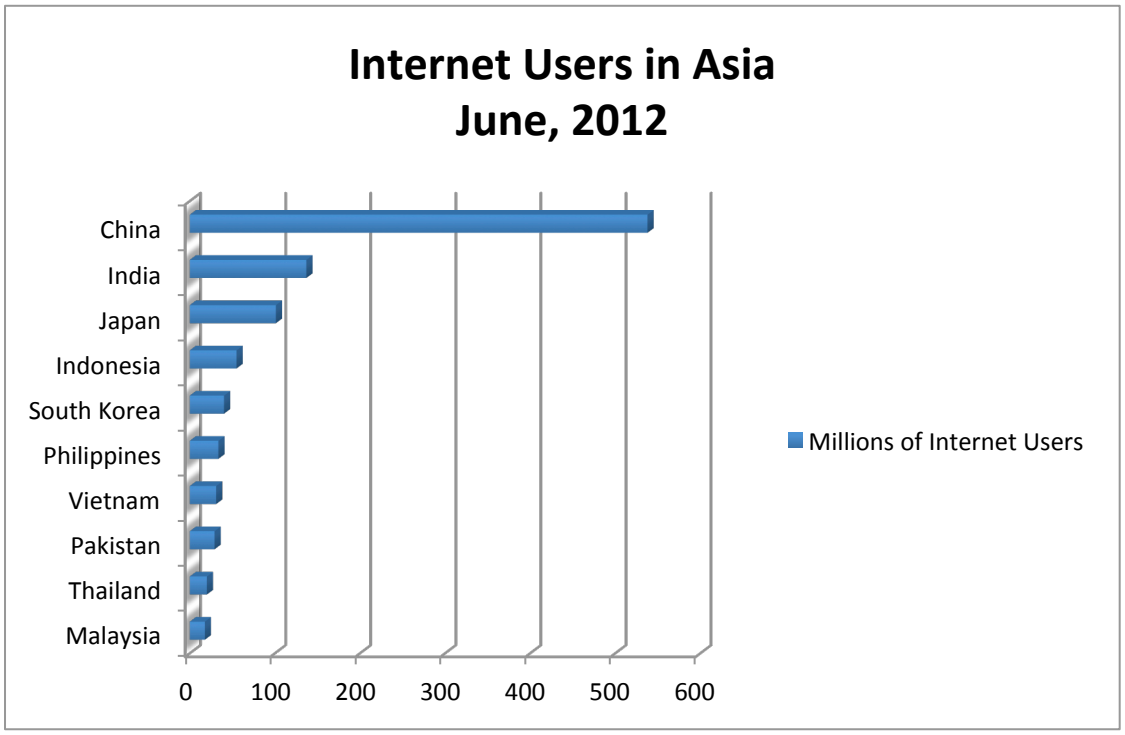


Figure 4.10: The number of Internet users in China far exceeds those in any other country in Asia (Data source: Internet World Stats, http://www.internetworldstats.com/stats3.htm).

One way to address these technical issues is to have multiple points of presence in China. This setup allows for static content caching to more sites and therefore closer to more users. Having multiple points of presence can help reduce the number of networks that must be traversed to deliver content, and thus avoid some of the negative consequences of poor peering. In addition to these technical challenges, there are regulatory and legal issues that must be considered when delivering content in China.

The Great Firewall of China

China regulates and restricts Internet content using a set of controls commonly known as the Great Firewall of China. The GreatFirewallofChina.org estimates that as many as 30,000 civil servants review Internet content and block material deemed undesirable. In some cases, entire sites are blocked if the material is categorized as undesirable.

Figure 4.11: Web content and sites are regulated in China and sites readily available to users outside of China are inaccessible within that country (Source: GreatFirewallofChina.org).

The Great Firewall of China has two implications for delivering content in China: technical and legal. Downloading content within the Great Firewall of China can be significantly slower than doing so outside the firewall. In addition to long distances and poor peering, businesses need to consider the impact of censoring technologies on network performance.

From a legal perspective, businesses might find themselves self-censoring content delivered in China. This activity might lead some to maintain two sets of content: one for general use and one for use in China. Also, businesses will need to have procedures in place to respond to take-down orders from the government should some of the business' content be considered undesirable. Sites or content about politics and gambling are considered high risk but even news and user-generated content may be subject to scrutiny. Chinese regulations require that content providers comply with Chinese law, including not posting prohibited content, having all necessary licenses, and monitoring the site to ensure banned content is not posted.

Content Delivery Network Considerations in China

Given the technical and legal issues with delivering content in China, it is clear that specialized services are needed to comply with local regulations. Content delivery network providers may be in a position to assist their customers with compliance if the providers have local staff that are familiar with regulations and understand procedures for responding to take-down notices and other government orders.

Summary

Multiple data centers, content delivery networks, and application delivery networks with network acceleration can improve application performance and responsiveness from an end user's perspective. Data centers provide essential redundancy needed for reliable access to applications and content. They are, however, costly and complex to operate. By combining a small number of data centers with application and network acceleration technologies, businesses can realize improved application performance without the cost of additional data centers.

Delivering content to global markets requires attention to a variety of national regulations and cultural expectations. China presents opportunities for business but requires compliance with laws governing the types of content that should be available to Internet users within China.