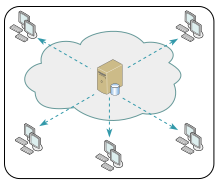
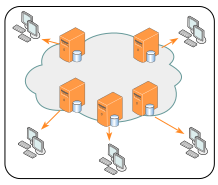
Content delivery network

A **content delivery network**, or **content distribution network** (**CDN**), is a geographically distributed network of proxy servers and their data centers. The goal is to provide high availability and performance by distributing the service spatially relative to end users. CDNs came into existence in the late 1990s as a means for alleviating the performance bottlenecks of the Internet[1][2] as the Internet was starting to become a mission-critical medium for people and enterprises. Since then, CDNs have grown to serve a large portion of the Internet content today, including web objects (text, graphics and scripts), downloadable objects (media files, software, documents), applications (e-commerce, portals), live streaming media, on-demand streaming media, and social media sites.[3]





(Left) Single server distribution
(Right) CDN scheme of distribution

CDNs are a layer in the internet ecosystem. Content owners such as media companies and e-commerce vendors pay CDN operators to deliver their content to their end users. In turn, a CDN pays Internet service providers (ISPs), carriers, and network operators for hosting its servers in their data centers.

CDN is an umbrella term spanning different types of content delivery services: video streaming, software downloads, web and mobile content acceleration, licensed/managed CDN, transparent caching, and services to measure CDN performance, load balancing, Multi CDN switching and analytics and cloud intelligence. CDN vendors may cross over into other industries like security, DDoS protection and web application firewalls (WAF), and WAN optimization.

Technology

CDN nodes are usually deployed in multiple locations, often over multiple Internet backbones. Benefits include reducing bandwidth costs, improving page load times, and increasing the global availability of content. The number of nodes and servers making up a CDN varies, depending on the architecture, some reaching thousands of nodes with tens of thousands of servers on many remote points of presence (PoPs). Others build a global network and have a small number of geographical PoPs.[4]

Requests for content are typically algorithmically directed to nodes that are optimal in some way. When optimizing for performance, locations that are best for serving content to the user may be chosen. This may be measured by choosing locations that are the fewest hops, the lowest number of network seconds away from the requesting client, or the highest availability in terms of server performance (both current and historical), to optimize delivery across local networks. When optimizing for cost, locations that are least expensive may be chosen instead. In an optimal scenario, these two goals tend to align, as **edge servers** that are close to the end user at the edge of the network may have an advantage in performance or cost.

Most CDN providers will provide their services over a varying, defined, set of PoPs, depending on the coverage desired, such as United States, International or Global, Asia-Pacific, etc. These sets of PoPs can be called "edges", "edge nodes", "edge servers", or "edge networks" as they would be the closest edge of CDN assets to the end user.[5]

Security and privacy

CDN providers profit either from direct fees paid by content providers using their network, or profit from the user analytics and tracking data collected as their scripts are being loaded onto customers' websites inside their browser origin. As such these services are being pointed out as potential privacy intrusions for the purpose of behavioral targeting[6] and solutions are being created to restore single-origin serving and caching of resources.[7]

CDNs serving JavaScript have also been targeted as a way to inject malicious content into pages using them. Subresource Integrity mechanism was created in response to ensure that the page loads a script whose content is known and constrained to a hash referenced by the website author.[8]

Content networking techniques

The Internet was designed according to the end-to-end principle.[9] This principle keeps the core network relatively simple and moves the intelligence as much as possible to the network end-points: the hosts and clients. As a result, the core network is specialized, simplified, and optimized to only forward data packets.

Content Delivery Networks augment the end-to-end transport network by distributing on it a variety of intelligent applications employing techniques designed to optimize content delivery. The resulting tightly integrated overlay uses web caching, server-load balancing, request routing, and content services.[10]

Web caches store popular content on servers that have the greatest demand for the content requested. These shared network appliances reduce bandwidth requirements, reduce server load, and improve the client response times for content stored in the cache. Web caches are populated based on requests from users (pull caching) or based on preloaded content disseminated from content servers (push caching).[11]

Server-load balancing uses one or more techniques including service-based (global load balancing) or hardware-based (i.e. layer 4–7 switches, also known as a web switch, content switch, or multilayer switch) to share traffic among a number of servers or web caches. Here the switch is assigned a single virtual IP address. Traffic arriving at the switch is then directed to one of the real web servers attached to the switch. This has the advantage of balancing load,

increasing total capacity, improving scalability, and providing increased reliability by redistributing the load of a failed web server and providing server health checks.

A content cluster or service node can be formed using a layer 4–7 switch to balance load across a number of servers or a number of web caches within the network.

Request routing directs client requests to the content source best able to serve the request. This may involve directing a client request to the service node that is closest to the client, or to the one with the most capacity. A variety of algorithms are used to route the request. These include Global Server Load Balancing, DNS-based request routing, Dynamic metafile generation, HTML rewriting,[12] and anycasting.[13] Proximity—choosing the closest service node—is estimated using a variety of techniques including reactive probing, proactive probing, and connection monitoring.[10]

CDNs use a variety of methods of content delivery including, but not limited to, manual asset copying, active web caches, and global hardware load balancers.

Content service protocols

Several protocol suites are designed to provide access to a wide variety of content services distributed throughout a content network. The Internet Content Adaptation Protocol (ICAP) was developed in the late 1990s[14][15] to provide an open standard for connecting application servers. A more recently defined and robust solution is provided by the Open Pluggable Edge Services (OPES) protocol.[16] This architecture defines OPES service applications that can reside on the OPES processor itself or be executed remotely on a Callout Server. Edge Side Includes or ESI is a small markup language for edge-level dynamic web content assembly. It is fairly common for websites to have generated content. It could be because of changing content like catalogs or forums, or because of the personalization. This creates a problem for caching systems. To overcome this problem, a group of companies created ESI.

Peer-to-peer CDNs

In *peer-to-peer (P2P)* content-delivery networks, clients provide resources as well as use them. This means that, unlike client–server systems, the content-centric networks can actually perform better as more users begin to access the content (especially with protocols such as Bittorrent that require users to share). This property is one of the major advantages of using P2P

networks because it makes the setup and running costs very small for the original content distributor.[17][18]

Private CDNs

If content owners are not satisfied with the options or costs of a commercial CDN service, they can create their own CDN. This is called a private CDN. A private CDN consists of PoPs (points of presence) that are only serving content for their owner. These PoPs can be caching servers,[19] reverse proxies or application delivery controllers.[20] It can be as simple as two caching servers,[19] or large enough to serve petabytes of content.[21]

Large content distribution networks may even build and set up their own private network to distribute copies of content across cache locations.[22][23] Such private networks are usually used in conjunction with public networks as a backup option in case the capacity of the private network is not enough or there is a failure which leads to capacity reduction. Since the same content has to be distributed across many locations, a variety of multicasting techniques may be used to reduce bandwidth consumption. Over private networks, it has also been proposed to select multicast trees according to network load conditions to more efficiently utilize available network capacity.[24][25]

CDN trends

Notable content delivery service providers

See also

References

Further reading

- Buyya, R.; Pathan, M.; Vakali, A. (2008). *Content Delivery Networks* (https://web.archive.org/web/20170927070358/http://www.gridbus.org/cdn/book/) . Springer. doi:10.1007/978-3-540-77887-5_1 (https://doi.org/10.1007%2F978-3-540-77887-5_1) . ISBN 9783540778868. Archived from the original (http://www.gridbus.org/cdn/book/) on 2017-09-27. Retrieved 2008-07-07.

- Hau, T.; Burghardt, D.; Brenner, W. (2011). "Multihoming, Content Delivery Networks, and the Market for Internet Connectivity" (https://zenodo.org/record/895901) . *Telecommunications Policy*. **35** (6): 532–542. doi:10.1016/j.telpol.2011.04.002 (https://doi.org/10.1016%2Fj.telpol.2011.04.002) .
- Majumdar, S.; Kulkarni, D.; Ravishankar, C. (2007). "Addressing Click Fraud in Content Delivery Systems" (http://www.cs.ucr.edu/~ravi/Papers/NWConf/clickfraud.pdf) (PDF). *Infocom*. IEEE. doi:10.1109/INFCOM.2007.36 (https://doi.org/10.1109%2FINFCOM.2007.36) .
- Nygren., E.; Sitaraman R. K.; Sun, J. (2010). "The Akamai Network: A Platform for High-Performance Internet Applications" (http://www.akamai.com/dl/technical_publications/network_overview_osr.pdf) (PDF). *ACM SIGOPS Operating Systems Review*. **44** (3): 2–19. doi:10.1145/1842733.1842736 (https://doi.org/10.1145%2F1842733.1842736) . S2CID 207181702 (https://api.semanticscholar.org/CorpusID:207181702) . Retrieved November 19, 2012.
- Vakali, A.; Pallis, G. (2003). "Content Delivery Networks: Status and Trends". *IEEE Internet Computing*. **7** (6): 68–74. doi:10.1109/MIC.2003.1250586 (https://doi.org/10.1109%2FMIC.2003.1250586) . S2CID 2861167 (https://api.semanticscholar.org/CorpusID:2861167) .

Last edited 6 days ago by Werg57