

Converged Networking for Next Generation Enterprise Data Centers

Casimer DeCusatis
Distinguished Engineer, IBM Corporation
Poughkeepsie, NY

Abstract

In this paper, we present the requirements for next generation enterprise data center networking and suggest possible evolutionary roadmaps to achieve these goals. In particular, the cloud computing environment and emerging standards for consolidation of storage and conventional local area network traffic known as Fibre Channel over Converged Enhanced Ethernet (FCoCEE) will be examined. The role of existing physical layer virtualization and convergence technologies, including optical wavelength division multiplexing (WDM), will be positioned in this context.

Introduction

Enterprise class data communication networks are characterized by their large size, potential for future scale-out, and resilient design. These networks can contain thousands of user connections, with a high degree of virtualization, and are designed for continuous availability, guaranteed data integrity, and error rates which are orders of magnitude lower than telecommunication networks ($10e-12$ or better, as opposed to $10e-9$ for voice traffic). These networks are also heterogeneous, multi-vendor environments, usually with specialized sub-networks devoted to communications, storage, and clustering applications. The widespread use of optical communication techniques and rapidly falling cost of bandwidth have combined to make it practical for multiple enterprise data centers to be interconnected over extended distances (100 km or more).

Within the next several years, constraints such as high power consumption and high operating or management costs will drive most of the world's Fortune 1000 companies to make significant changes in their data center architectures. Although many important changes are anticipated in servers, storage, and software, this paper will focus on the increasingly important role of the data center network. In particular, we consider the approach of converging multiple networks and protocols into a single, unified data center fabric, which is application fluent and extends across traditional organizational and geographic boundaries. First we describe the business forces motivating these new data center designs, and the cloud computing approach to delivery of information technology (IT) resources. Then we discuss the importance of a dynamic, virtualized network infrastructure in realizing the promise of cloud computing, and the recently proposed industry standards for convergence of storage area networks (SANs) and local area networks (LANs), as well as the means for transporting both emerging and legacy protocols across multiple data centers. Following a technical description of the new standards used in this approach, we recommend a strategy for adoption of this technology in various data center environments.

Cloud Computing for Next Generation Enterprise Data Centers

Modern internet-scale data centers take advantage of the low cost of computing power offered by rack-mounted and blade servers, and may contain upwards of 100,000 such servers in a single data center [1]. This translates into hundreds or thousands of network adapter ports and cables. The cost of networking equipment in this environment can approach 20 % or more of the total data center equipment budget. Worse yet is the high operating expense, which must be factored into total cost of ownership. This approach is straining the power and cooling limits of many data centers, some of which are unable to install new equipment due to these limitations rather than financial considerations. Faced with these constraints, over 70 % of the world's global 1000 enterprises will be building new data centers or significantly redesigning their existing data centers over the next 5 years [2]. As these redesigns proceed, enterprises will require their future data center infrastructures to better align IT with business objectives, to reduce capital expenses (capex) and operating expenses (opex) without sacrificing quality or performance, to reuse existing resources where possible, and to operate with lower energy consumption.

One possible approach for the next generation enterprise data center involves cloud computing, an emerging method of IT delivery in which applications, data, and resources are rapidly provisioned without requiring the end user to purchase or maintain IT resources (in the style of existing internet or telecommunication clouds). In a cloud computing data center, applications are provided as standardized offerings to end users over a dynamically reconfigurable network. This allows for a more flexible cost and pricing business model, and enables the enterprise to provision new resources much faster than current methods. There are several different cloud computing delivery models [3, 4]. So-called private clouds will exist within a corporate data center to address applications which are sensitive to security or performance issues, as well as for test and development environments. Public clouds will extend beyond the confines of a single enterprise, and share the underlying compute resources across multiple enterprises. Applications related to web hosting, software as a service, and others will benefit from this approach. Traditional enterprise IT delivery (including mainframe-based systems) will continue to exist for some time, especially in mission critical or highly regulated compliance environments. Mainframes and other types of servers will also be used in cloud computing, which holds the potential for disintermediation of the traditional IT delivery model.

A confluence of technology trends, including faster processors (multi-core designs with higher clock frequencies) and solid state memory, has contributed to dramatic cost reductions in processing power and storage. These, in turn, have helped make cloud computing models feasible. Consider, for example, the recently announced Intel Core i7 processor family, which promises a remarkable cost/performance ratio (around \$300 for a quad core, 2.66 GHz processor) [5]. Within the network, technology improvements combined with growing volumes have driven down network bandwidth costs to the point where communications technology usually reserved for telecommunication central offices has become practical to install in corporate data centers. As processing power, storage, and bandwidth become available anyone who needs them, businesses are able to more easily afford the computing resources they need to remain competitive. Moreover,

they no longer have to purchase, manage, and support these computing resources themselves; nearly free IT resources have made cloud computing possible by allowing providers (like IBM) to deploy massive amounts of computational power and make it available on demand to end users. While there are still technical issues to overcome, cloud computing has emerged as a viable technical option with demonstrated business value, including both reduced up-front costs and operational expenses, as well as improving agility in response to quickly shifting business conditions.

The first steps towards cloud deployment involve reducing infrastructure complexity and staffing requirements, thereby realizing reduced operational costs. This value proposition has led nearly one-third of companies currently using cloud computing to increase their investment in the technology, despite the weak global economy [6]. The next steps towards cloud computing transformation involve removing physical resource boundaries on key applications, sharing storage and computing resources, and implementing granular service metering and billing models. Not surprisingly, concerns with security and control of mission critical data are slowing more widespread adoption of cloud computing. In a recent survey, many clients cited their concerns with security as a major inhibitor to more rapid deployment of cloud computing models [6]. Recent highly publicized outages and business failures of some platform-as-a-service providers have only reinforced these concerns.

While most cloud computing discussions focus on processing power or software, the network architecture is playing an increasingly important role. Similar to grid computing models, some types of compute clouds are disaggregated across multiple data centers and may contain a mixture of mainframe, midrange, and other types of servers and storage. Just as optical fiber networks form the basis of telecommunication clouds, they will also be the underlying technology behind cloud computing. The cloud will be able to adapt optical routing, labeling, multiplexing, and switching techniques from related telecom practices. Applications such as virtual server migration are enabled by the joint management of network and server resources. Server partition mobility is facilitated by large layer 2 domains in the network, so high layer 2 port density becomes an important consideration for virtualized data centers. Conventional architectures statically map web services to Ethernet VLANs, where each VLAN is often constrained to a few hundred servers (due to overhead in the control plane). Spanning traffic across VLANs is a nontrivial problem, which is addressed in some cases by expensive, proprietary hardware and software. In this manner, data center resources are fragmented, and limitations on the network infrastructure (such as IP address hierarchies) make it difficult to dynamically reassign servers and applications. As noted previously, some approaches use a single, large layer 2 domain, but this approach doesn't scale well to large numbers of servers. Distributed computation is severely constrained by the low bisection bandwidth of the network, which can become a performance bottleneck over longer distances. Furthermore, conventional network designs do not scale well, since they concentrate traffic in a few large data center switches which must be disruptively upgraded to keep pace with traffic demand, and require configuring large amounts of stand-alone appliances dedicated to security, load balancing, and other functions. As workloads change frequently, intermittent congestion of the network occurs, which can reduce the aggregate computational power of the data center.

Dynamic Infrastructure using Converged Networks

The layers of a typical static data center network are illustrated in Figure 1. At the core and aggregation layer, as well as the access layer database tier, high capex and opex costs are driven by the need for two different fabrics (Ethernet and Fibre Channel) and the use of many lower bandwidth links (1-4 Gbit/s). At the access layer in both the web and application tiers, lightly utilized servers drive up capex and opex, while opex is further increased by manual management of virtualized servers and switches. Higher opex is driven in the service plane by the proliferation of security appliances and other single-purpose devices, and at the edge router and core layers by the large number of network, server, and storage managers. There is considerable market pressure to reduce and consolidate I/O adapters within many enterprise data centers, relieving these high expense points. The resulting next generation dynamic network infrastructure will be characterized by flattening of the hierarchy depicted in Figure 1, converged and highly virtualized networks, and dynamic re-provisioning of IT resources. Among other things, this approach will reduce capex through the use of converged fabrics with higher server and storage utilization, as well as lowering opex through integrated, automated management. Increased use of virtualization will eventually force the access layer to collapse into the server. This tier aggregation is driving the need for higher bandwidth connections from the server directly to the network aggregation or core layer in all application environments (SAN, LAN, and clustering fabrics). Tighter integration of virtual servers and networking technologies will also lead to integration of server and network management applications.

Emerging cloud computing environments are well positioned to take advantage of converged data center networks. The convergence of multiple protocols onto a common fabric is a long-standing goal of the data communication industry. Several previous attempts have been made to develop such a network, including Fibre Channel, Asynchronous Transfer Model (ATM), and InfiniBand. From these previous attempts, we have learned that in order for converged fabrics to succeed, they must address some basic criteria [7]. First, the converged fabric needs to fulfill all the requirements of the fabrics which it is replacing. Users will not accept a reduction in service or performance as a tradeoff for convergence; a converged fabric must meet or exceed all the technical requirements of any fabric it intends to displace. Second, a converged fabric must be based on open industry standards; previous efforts have suffered because of their reliance on proprietary features or attempts to lock users into a single vendor platform. Finally, since low cost is primarily driven by high volume in the networking market, a converged fabric must be designed around the highest volume existing fabric, and must insure that production volumes of converged traffic are sufficient to meet projected cost targets [8, 9]. Until recently, these conditions have all been successfully addressed only for physical layer interconnections between data centers, through applications such as optical wavelength division multiplexing (WDM), which we will describe later in more detail. The next step will involve convergence of higher level protocols within a single data center. The value proposition for these converged fabrics, at least in the initial deployment, results from the potential reduction of equipment (adapters, cables, and switches or switch ports). This brings an associated reduction in power consumption,

management complexity, and presumably improved reliability through simplification of the infrastructure.

The two highest volume protocols in the data center, which stand to benefit the most from convergence, are Ethernet [10] (most commonly used in local area networking and network attached storage applications) and Fibre Channel [11] (commonly used in enterprise-class storage area networks). Both of these protocols have been available for many years, and accommodate a range of different data rates as shown in Figure 2. There are several emerging industry standards which address convergence of these two protocols. Fiber Channel over Ethernet (FCoE) is an emerging set of industry standards which will allow dedicated Fibre Channel links to operate over Ethernet fabrics [12]. A separate effort known as Converged Enhanced Ethernet (CEE) is the result of an industry consortium called the CEE Authors group [13]. This group has proposed several enhancements which not only improve basic Ethernet transport but also make it more suitable as a convergence platform. In addition to being an open standard available today, CEE has been submitted to the IEEE for future consideration in the formal standards process. As an open industry standards approach, Fibre Channel over Converged Enhanced Ethernet (FCoCEE) represents the preferred path to fabric convergence. As shown in Figure 2, this convergence will initially occur at 10 Gbit/s data rates or higher, which will require optical links for distances exceeding about 10 meters.

There are a number of technical enhancements which have been introduced to enable converged fabrics [14]. For example, the current IEEE 802.3x Ethernet standard defines a mechanism to pause all traffic from a data source to prevent the onset of network congestion and the nucleation of congestion trees which can propagate across the entire network. This results in data packets being dropped by the network during periods of congestion. Normally this is managed by a higher level protocol such as TCP in a pure Ethernet environment. Fibre Channel protocols, as well as some types of server-to-server communications, have a strict requirement to deliver all packets in order and cannot tolerate dropped packets; this is incompatible with the standard Ethernet approach. To prevent dropped packets, CEE introduces a separate flow control mechanism for different classes of traffic. This is known as priority-based flow control (PBFC), and makes it possible to pause low priority traffic, such as Ethernet, while allowing high priority traffic, such as Fibre Channel, to remain unaffected. This avoids network congestion without sacrificing higher priority traffic.

Another CEE feature is Enhanced Transmission Selection (ETS). Standard IEEE 802.1p protocols do not allow network bandwidth to be allocated to different traffic priorities. To avoid this limitation, ETS maps the IEEE 802.1p priority levels into priority groups, each of which can contain a different class of traffic (LAN or SAN) which is defined as lossless or lossy, and permits different amounts of link bandwidth to be specified for each priority group. In this manner, storage traffic can be guaranteed a share of available network bandwidth.

Finally, CEE defines data center bridging (DCB), which uses link layer discovery protocol (LLDP) defined in IEEE 802.1ab to advertise connectivity and management information, so that CEE compliant devices can be discovered, initialized, and managed.

Converged Fabrics for Enterprise Applications

When planning for the introduction of converged fabrics into the enterprise environment, there are several important business and technical considerations. These include the value proposition which converged fabrics bring to the enterprise customer as well as the technical readiness of the enterprise for converged fabrics. For some types of data centers, converged fabrics offer a clear business advantage. These include cases where an enterprise needs to move a lot of raw data over a network for storage backup; greenfield opportunities where FCoCEE's performance, reliability and latency levels would be acceptable; or large, Internet-scale data centers interested in reducing servers, adapters and cables. In addition, environments where low-cost connectivity is a higher priority than reliability—social-networking sites and search engines, for instance—also might make a good fit for cloud computing with converged fabrics. These sites operate on a low-cost commodity model, and FCoCEE could provide a fairly reliable transport without requiring any initial investment in a separate Fibre Channel SAN. Initial deployments will likely be within a single rack or blade server. One deployment model would involve using relatively small (1-rack-unit high chassis) top-of-rack (TOR) switches to aggregate traffic from within a single server rack, later connecting with a larger core switch to encompass multiple racks. Most converged fabrics require new host bus adapters (HBAs) in the server, while preserving the legacy HBAs on storage; this facilitates migration from legacy SANs, since clients do not have to initially upgrade all their storage. In the short term, this usage model can deliver significant capital and operational cost savings. In the long term, as FCoCEE standards, hardware, software and management evolve, they may eventually enable full data center fabric convergence, with associated economies of scale and resource savings. In the meantime, 8 Gbit/s Fibre Channel is an option on some of the latest switches which provides customers the flexibility to migrate to converged fabrics at their own pace, following the roadmap illustrated in figure 2.

Within an enterprise data center based on mainframes and high-end storage control units, the situation is more complex. These environments may also contain large amounts of legacy protocols (ESCON links), lower data rate Fibre Channel (1-4 Gbit/s) and high availability disaster recovery architectures which require other communication protocols (Geographically Dispersed Parallel Sysplex or GDPS) [8, 15]. Further, they can employ the FICON protocol as a subset of their SAN; this is essentially similar to Fibre Channel protocol, but contains enhancements at layer 4 and above to improve data integrity [8, 11]. Because so many different types of communication links are used, protocol independent WDM is used to interconnect multiple data centers over extended distance (10-100 km or more). As shown in figure 3, WDM technology allows a single optical fiber channel to transport 64 or more independent data channels, by multiplexing each data channel on a different optical wavelength. There are different types of WDM, depending on the channel count and wavelength spacing [8]. This transport has become an integral part of many enterprise networks, and will continue to accommodate converged fabrics as required in the future. Since converged HBA will appear first on rack-mounted or blade servers, which will also have requirements for extended distance interconnections, the

traditional role of WDM will expand from mainframe-based solutions to encompass many different types of servers and storage.

However, the value proposition for converged fabrics is somewhat different for mainframe servers, which already employ sophisticated virtualization technology (indeed, the virtual machine or VM partition has been a staple of IBM's System 360, 390, and Z series for many years). In addition, many of these servers already have relatively high levels of link utilization compared with other environments. In fact, it is considered best practice to limit link utilization to 50% or less for performance and reliability reasons. Thus, there are fewer opportunities to achieve bandwidth efficiencies by merging several lower utilization links onto a single 10 Gbit/s converged fabric link; convergence on these servers may need to wait for future 40 Gbit/s links to be introduced (see figure 2). Further, many enterprise servers already support 10 Gbit/s Ethernet and 8 Gbit/s Fibre Channel or FICON links, so there is little incremental improvement achieved by a 10 Gbit/s converged fabric link. Many analysts agree that factors other than raw data rate are limiting current enterprise server architectures (for example, performance can be limited by start rate, or the number of new jobs that can be initiated on the I/O subsystem in a given time). Virtualized servers increase the use of shared resources, such as internal buses, slots, and host bus adapters. It will be necessary to alleviate possible contention for resources in highly virtualized environments, using solutions which go beyond higher data rates. New approaches have been suggested to address these performance issues, including more extensive use of PCIe adapters and switches as a way to extend a single physical port across multiple virtual ports, enabling connection to virtual servers [16].

We also note that many mainframe data centers historically prefer to wait until new technology has been proven and tested before using it in production data centers (this is one reason why WDM qualification programs have become popular with these clients). It must be proven that converged fabrics can deliver best of breed performance compared with dedicated, single-purpose Ethernet or Fibre Channel fabrics. There are also technical concerns to be addressed, such as switch cascading and interoperability of different vendor products, or state change propagation across converged fabrics. Note that some converged fabric products offer significant scale-out capabilities, through the use of stacked switches or virtual chassis backplanes which make several physical switch elements appear as a single entity to the management system. This has the potential to scale converged fabrics to very high port counts in the future, well beyond what large SAN environments have currently deployed. Stress testing of 100-1000 port or larger fabrics will be an important differentiator for this market; new test methodologies may be required, and the full impact of deploying such large scale switches into the data center architecture remains to be seen.

Mainframe customers are more accustomed to chassis or rack mounted cabinet-based switch form factors in the SAN environment. Migration to a TOR switch form factor is less appealing when the mainframe server itself is cabinet-based and requires the TOR to be mounted in an adjacent rack. The mainframe environment is also reluctant to compromise performance, which is one reason why FICON continues to be provisioned as a private sub-networks within the larger Fibre Channel SAN, rather than being converged onto a single switch fabric (despite the fact that switches have existed for

many years which support both protocols). The more complex step of converging SAN and LAN traffic must be considered in this context. Given these market realities, converged fabric adoption will appear first in those environments which present the greatest short-term value proposition, and later in traditional mainframe data centers once reliability, performance, and other features have been demonstrated.

The FCoCEE body of standards is expected to complete public review and be formally adopted sometime in 2009. Some companies will likely introduce products to market before the standard is finalized or proprietary versions of converged fabrics with trademarked names. Pre-standard solutions should obviously be required to bring their offerings into compliance with the final version of the standards when available. Proprietary or non-standard solutions may not interoperate with standards-based solutions or may require additional upgrades in the future. Early adopters of such products should be careful to insure that their infrastructure switches support the CEE Authors Group open standards and the T11 FC-BB-5 industry standard for Fibre Channel converged fabrics. The fabric should also support intermediary switches (creation of a CEE cloud). At the very least, noncompliant switches should offer upgrades at no cost to future de-facto standards and clouds, as a way to protect investments in the network infrastructure.

Finally, there are practical issues of management tools, behavior, and organization structure which affect the convergence conversation. The management tools need to be vetted in real-world high availability environments, and preferably require learning a single set of tools which will scale with the enterprise network. Today, the SAN and LAN are managed by two different organizations, often with different legacy tools, operating budgets, and requirements. Fabric convergence implies a merger between these two organizations, as well. The not-inconsequential business and political ramifications of such a transformation across the enterprise must be managed carefully, gracefully and gradually. Managers of large data centers should work with an IT company that has the service skills necessary to undergo this organizational transformation in careful steps.

When planning for the eventual use of converged fabrics, enterprise data centers should leverage their existing server investments where possible. Deployment of virtual disk and server management is therefore important groundwork for future data center initiatives. As always, the challenge faced by data center managers will involve matching technologies with applications to most affordably satisfy technical requirements and business objectives. Data center managers must select their converged fabric switch provider with care. For at least some switches, convergence of the SAN and LAN will require an upgrade of the existing Ethernet infrastructure with new low-latency Ethernet switches. This could amount to a major, disruptive and costly overhaul of the enterprise network core. Data center managers will need to carefully re-examine their infrastructure partners with migration issues in mind, to obtain the best value with minimal disruption. Companies would be well advised to partner with a data center server provider who understands when FCoCEE is appropriate, rather than those who tout this technology as one size fits all. Finally, it's important to consider the end-to-end capabilities of the data center infrastructure. While it may be tempting to impose geographic limitations on the

boundaries of the data center, there are mission-critical applications such as business continuity which require the interconnection of multiple, geographically dispersed locations. Protocol-agnostic channel extension through WDM will continue to serve as the convergence cloud for an array of uniquely valuable protocols linking together multiple enterprise-class data centers, as well as transporting converged fabrics where required.

Conclusions

Cloud computing models are emerging which hold the potential for more rapid, low cost deployment of IT resources. This, in turn, is driving significant changes in the data center network and compute model. The traditional data center compute model, especially in the case of rack-mounted x86-based servers, has consisted of lightly utilized servers running a minimal operating system or a hypervisor with a small number of virtual machines. In this environment, lower bandwidth links and a fairly static infrastructure were sufficient for attachment to server resources. The industry is moving towards a higher bandwidth, dynamic infrastructure model that involves highly utilized servers running many virtual machines per server, using higher bandwidth links to communicate with virtual storage. This compute model offers several tactical business advantages. These include lower capital expenses through higher utilization of midrange servers, storage, and converged network fabrics, as well as lower operational expenses which are achieved through automated and integrated end-to-end management. The preferred convergence fabric will be based on the FCoCEE open industry standard, which achieves lossless packet transport, guaranteed bandwidth, and end-to-end infrastructure management. By contrast, enterprise servers which already offer high link utilization and make extensive use of virtual machine partitions will realize more long-term benefits. For the foreseeable future, mainframe data centers will continue to manage a variety of protocols when it comes to interconnecting over extended distance. For these environments, WDM effectively becomes the convergence foundation for FCoCEE and other protocols.

References

- 1) A. Greenberg, P. Lahiri, D. Maltz, P. Patel, S. Sengupta, "Towards a next generation data center architecture: scalability and commoditization", Proceedings ACM PRESTO '08, Seattle, WA (Aug. 22, 2008)
- 2) Gartner Group report, "Next Generation Enterprise Data Centers", September 2007
- 3) Forrester report, "There are two types of compute clouds", available from www.forrester.com (November 2008)
- 4) Gartner Group report, "You can't do cloud computing without the right cloud (network)", available from www.gartner.com (June 2008)

- 5) J. Eijmberts, "Understanding Intel Itanium architecture values", proc. HP Dutchworld 2008 / Interex User Group Conference (November 20, 2008)
- 6) C. Zaragoza, "Global study: cloud computing provides real business benefits...", Kelton Research study, available from www.avanade.com (February 2009)
- 7) C. DeCusatis, "The new enterprise data center: global networking with zero downtime", Proc. Enterprise Technology Summit, New York, NY (May 15, 2008)
- 8) C. DeCusatis, *Handbook of Fiber Optic Data Communication*, 3rd edition, Academic Press, NY (2008)
- 9) C. DeCusatis and T. Bundy, "What role will FCoE play in data center convergence ?", Virtualization Journal, to be published, April 2009 (see <http://www.asiaing.com/magazine/13.html>)
- 10) IEEE standards 802.3x, Ethernet full duplex and flow control standard (1997)
- 11) INCITS (international committee for information technology standards) Fibre Channel standard, information available from <http://www.T11.org>
- 12) Carlson et.al, "call for investigation of a Fibre Channel over CEE (FCoCEE) standard", <http://www.t11.org/ftp/t11/pub/fc/bb-4/07-181v1.pdf>
- 13) CEE Authors group, <http://tech.groups.yahoo.com/group/ceeauthors>
- 14) Priority based flow control ver 0 is available from <http://www.ieee802.org/1/files/public/docs2008/bb-pelissier-pfc-proposal-0508.pdf> ; see also Data Center Bridging Exchange Protocol ver 0, available from <http://www.ieee802.org/1/files/public/docs2008/az-wadekar-dcbx-capability-exchange-discovery-protocol-0608-v1.0.pdf> ; see also Enhanced Transmission Selection ver 0, available from <http://www.ieee802.org/1/files/public/docs2008/az-wadekar-ets-proposal-rev-1.0.pdf>
- 15) E. Chencinski, C. DeCusatis, T. Gregg, K. Pandey, L. Mohr, T. Schlipf, and M. Helms, "IBM System Z I/O subsystem overview", IBM Journal of Research and Development, to be published (April 2009)
- 16) Gartner Group report "Fibre Channel over Ethernet", available from www.gartner.com (October 2008)

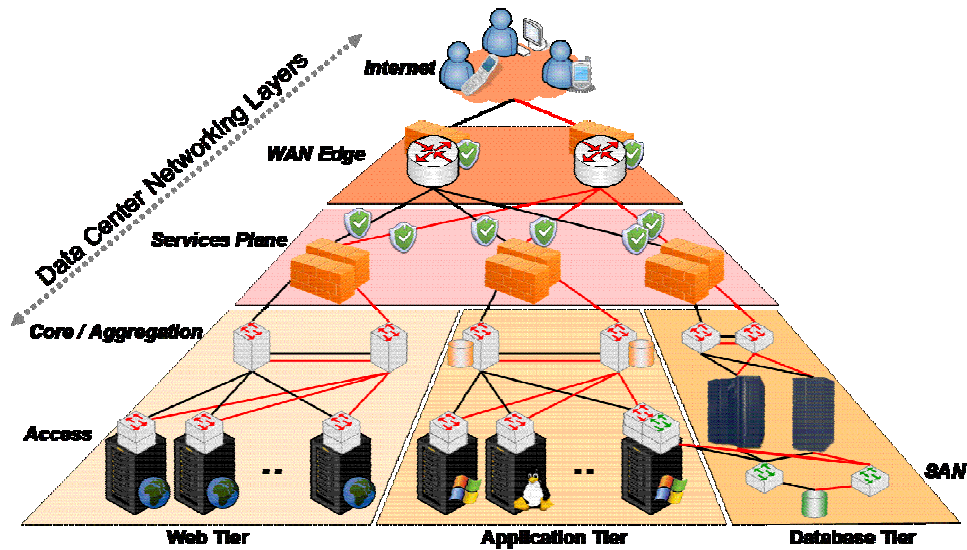


Figure 1 - Network model for conventional enterprise data centers. Among other benefits, converged fabrics promise to flatten the core, aggregation, and access layers into a single stage, and to collapse the web, application, and data base tiers into a common dynamic network infrastructure.

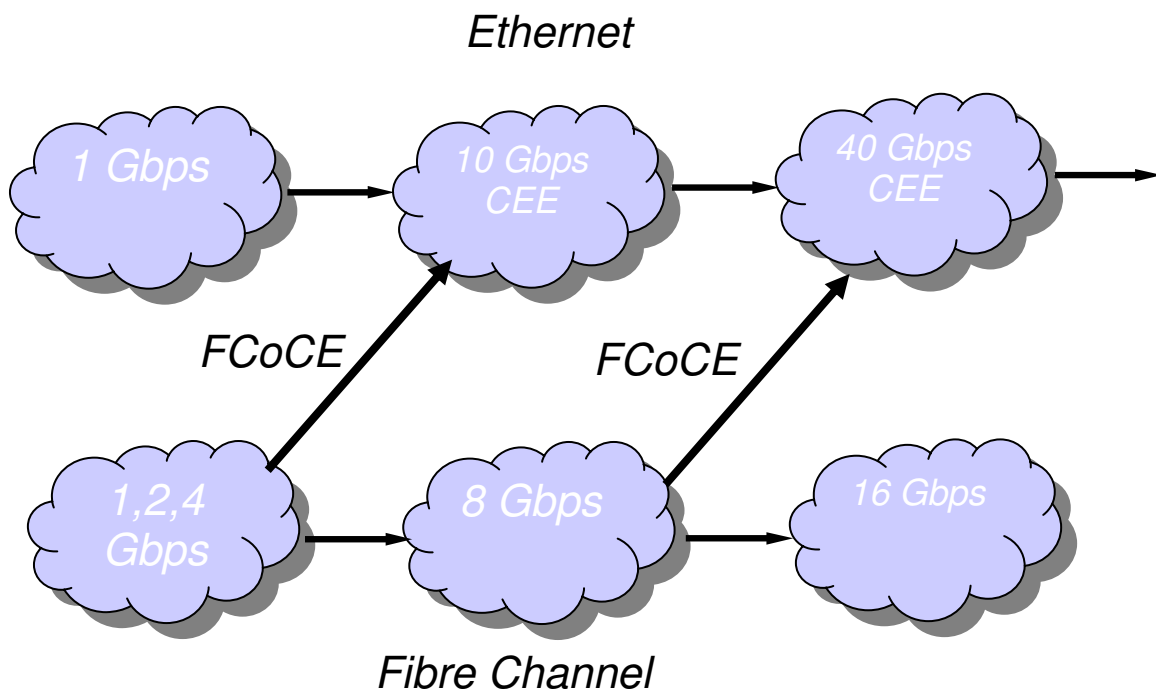


Figure 2 – Evolution of Ethernet, Fibre Channel, and converged networking protocols

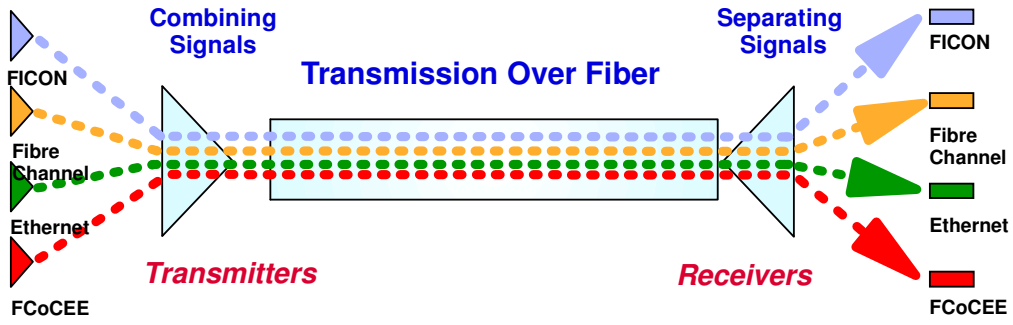


Figure 3 – Schematic of wavelength division multiplexing (WDM) technology