Cloud Scale Networking Part I: Architecture Considerations

Abstract. We explore data center networks built for cloud practices. As external observers, we identify differences and commonalities across operators and derive solutions from semiconductor building block attributes. We cover emerging usage, topology, and link technologies.

April 2014
Introduction

Data center networks suitable for enterprise or cloud computing require a scalable infrastructure for traditional or distributed applications. Most large-scale data centers have adopted the “scale-out” network model [1]. This network model is analogous to the scale-out server model, for which the compute infrastructure is one-size-fits-all. Two-socket servers are used throughout instead of the n-tier server model, where fewer, more expensive multiprocessor configurations are used as database and application servers. We observe that distributed software is also a good fit for the network scale-out model.

In enterprise use cases, application instances execute on their respective physical (or virtual) servers and rely on the network to communicate with applications and network services on other systems. In public cloud use cases, applications tend to exceed (in scale) the capacity of any multiprocessor machine, making distributed applications the norm. These distributed applications are decomposed and deployed across multiple physical (or virtual) servers, introducing network demands for intra-application communications.

The cloud infrastructure further requires that multiple distributed applications coexist on the same server and network infrastructure.

Networks that do not satisfy distributed application workloads make suboptimal use of the data center infrastructure and may prevent the satisfaction of Service Level Agreements (SLAs). The goal is, therefore, to design a large-scale network that satisfies the workload requirements of cloud-distributed applications. This task is ideally accomplished by factoring in the knowledge cloud operators have about their own workloads.

Paradoxically, we find that the fear of emerging applications and the acquisition of new businesses that change workloads over a short period of time make cloud operators reluctant to rely on workload knowledge for network design. In cases of hosted applications (for example, “Infrastructure as a Service” (IaaS) models), the workload profile is beyond the operator's control. Operators prefer future-proof networks that are not very sensitive to the workload (conceptually, constant bisection bandwidth networks for arbitrary traffic patterns).

We examine how workload uncertainty affects main network design dimensions and what constitutes reasonable and optimal choices. The cost metric is approached as the cost per attached server, and we pursue a network that scales to larger node counts without impacting the cost per server. Workload uncertainty is translated into a network that favors symmetrical properties as opposed to locality assumptions. Specifically, the network throughput achievable between two servers in the infrastructure should not differ much based on the relative physical location of each node within the data center. This location independence is sometimes described as a “flat network”, even though we discuss other applications of the term “flat network” based on topology or address hierarchy.

Regarding topology, we present the application of large multistage fabrics to interconnect the physical server infrastructure and the role of various cable and transceiver technologies [2]. We also consider the capital acquisition costs and cable plant ramifications of the topology and link technology.

Another consideration is whether a topology allows incremental growth, with intermediate steps towards the upper limit that are reasonably optimal. We think this criterion is important given that part of the workload uncertainty is precisely demand growth uncertainty. We strive for solutions that are capital-efficient (in the sense that scale build-out is pushed out until absolutely necessary), not burdened by recabling disruptions or subject to high labor factors as the infrastructure grows.

The above observations apply generally across large cloud operators, as exemplified in [3]. In this paper, we drill down into the next level of detail and examine where the differences and commonalities across operators exist. We pursue a prescriptive analysis for building a network or designing the right semiconductor building blocks for future networks.
Network Design Dimensions

Initial network design considerations tend to focus on the choice of network topology. In this respect, there has been a universal transition to multilayer mesh networks with multipathing as the basis for total network capacity. These topologies are truly scale-out and, therefore, the cost per attached server is rather flat for any network size. But the cost is reasonable as long as the switch radix is sufficient for the average server attachment speed. These networks often take the form of a folded Clos network. In general, we describe them as leaf-spine configurations, where the size and attributes of a leaf and a spine switch do not have to be identical. For convenience, we assume that all server endpoints and external traffic (to/from backbone routers) are connected via leaf switch ports. Routing such topologies can be as simple as up-down routing, and the diameter of the network is the same for all traffic (up to the highest spine and, from there, down to a leaf switch port), assuming poor traffic locality.

We present Clos deployment tradeoffs as oversubscription (how much and where to oversubscribe), rate adaptation (speed ratio between leaf downlinks vs. leaf-to-spine links), and the relative sizes of leaf vs. spine switches.

Several alternatives to Clos topologies are worth covering. One alternative is toroidal: a type of direct-connect network, a Torus presumably saves cable costs by not requiring interior links. Other alternatives exist as multistage networks with a set of leaf-to-leaf links; these sacrifice Clos nonblocking properties to provide either a lower average diameter or a reduction in the number of required links.

In general, we find that these non-Clos topologies are more sensitive to workload assumptions and, thus, go against the grain of future-proof topologies for evolving workloads.

A common method of multipathing in large data center networks is to use layer-3 (L3) (i.e., routers) everywhere, including the leaf switches. These boxes consume fewer layer-2 (L2) address entries (i.e., small subnets) at the expense of the larger consumption of subnet routes in their routing tables. In many cases, the addressing scheme is no longer flat, and overlays are utilized, particularly L2-over-L3 overlays (for example, VXLAN or NVGRE). In such cases, we get one level of indirection for actual endpoint addressing so that applications and tenants can be deployed anywhere (below any leaf switch). A side benefit is the reduced consumption of subnet route entries, since only the underlay subnets are stored in router tables.
Accurate cost models for large networks show that cables and transceivers represent a significant portion of the cost per endpoint attachment and that their incidence is increasing, as their costs do not necessarily improve in line with Moore's law. The total number of network links is derived once we select the number of endpoints, the topology, and the oversubscription ratio. The link media is an interesting design choice, where the leaf switch downlinks are short links that are hard to aggregate into bundles without constraining the server form factors. Traditionally, these short links have been dominated by the Direct Attached Copper (DAC) technology. The Top-of-Rack (ToR) switch is a leaf switch with a one-to-one association to a server rack. We conveniently view the ToR as an active patch panel of sorts, where the ToR not only starts the routed domain but also provides the opportunity to use different link technologies for downlinks and spine uplinks. Uplinks, which are generally longer-reach optical links, use pluggable transceivers suited for a particular choice of fiber in a data center.

The economics of downlinks and uplinks are positioned to evolve differently as the speed of the serial lane transitions from 10 Gbps to 25 Gbps, and as pluggable transceivers possibly are replaced by board-mounted optics and silicon photonics solutions. Downlinks limited in the 3-meter range present the opportunity for one or two more generations of copper-based solutions, where the dominant cost is in the cables and not as much in the endpoints. Uplinks, on the other hand, require longer distances, and the transition to optical is, therefore, imminent. Optical downlink solutions are also relevant, and their economics will presumably be driven by the cost of the copper variants, with a modest premium associated with the operational advantages of optical links in cable size and weight.

The Server Link

As we transition to next-generation server platforms, next-generation switches, and new network interfaces, we find that either a single 25 Gb link or 50 Gb link (two links of 25 Gb) are the “sweet spot” of server downlink attachment. The switch radices are sufficient to support flat networks with a 100 Gb uplink and 50 Gb downlinks, and the cable costs of 50 Gb transported over two lanes are less than those associated with 40 Gb over four 10 Gb lanes.

The ToR-to-server downlink is of particular interest beyond its impact on cost metrics. It is of interest in terms of network congestion and functional decomposition.

The topological symmetry of a Clos diagram can deceptively suggest that traffic and congestion are equally symmetrical and distributed. We point out two reasons that may not be so. The first reason is that the up and down directions are intrinsically different. Congestion on the path up to the spines can be “engineered away” either locally or globally by using alternate paths to the spine tier, as any spine can take traffic to any destination. Traffic down from a given spine follows a single possible path on fully developed Clos, and congestion at the ToR converges on a single output port for a given destination. It turns out that storage and Big Data patterns further exacerbate down congestion at the ToR, as the client requests produce self-synchronized types of responses from many servers to the same client. This pattern is generally described as “incast” traffic.

The above cloud use cases suggest that it is better to deploy oversubscribed ToR switches at faster downlinks than to follow the traditional enterprise practice of 10x rate adaptation between ToR uplinks and downlinks. A 10x rate adaptation aggravates ToR incast down congestion, whereas ToR oversubscription does not. Faster ToR downlinks helps move data towards the DRAM of the associated server even if the server processor cannot process data at the peak rate.
In terms of functional decomposition, we observe that significant innovation occurs on both sides of the server-to-ToR link. This observation follows the notion that rich functionality is best applied at the edge, with little impact to the slowly evolving spines. We have witnessed the emergence of L2-over-L3 tunnels, Edge Virtual Bridging, Service Chaining, complex multitenant policy, MPLS-style path controls, and congestion management—all of which are examples of edge functions. Most of these innovations are underpinned by a clear separation between the I/O and the network boundaries, and by the application of REST principles where functionality can be deployed in a modular fashion. The network interface card (NIC) and the virtual switch (Vswitch) represent the locus of innovation on the server side, while increasingly open switch environments represent a ToR switch innovation software foundation. The use of vendor proprietary I/O technologies for these links presents as many problems as the purported elimination of the ToR itself, along with the server downlink in some toroidal topologies.

### The Boxes

After considering the network design dimensions, the types of boxes, and their gradual evolution over time, we focus on the semiconductor devices in the boxes that ultimately represent the network building blocks.

Although the packet processing aspects of routers are defined by standards, their queuing architectures differ from vendor to vendor and across product lines. Therefore, queuing, scheduling, and buffering attributes depend on the ASICs and box designs used in a specific network.

Using ASICs with embedded buffers provides a higher radix building block than using those that require external memory interfaces. Historically, embedded buffer devices have been used in ToR boxes, and either shallow embedded buffers or deep external buffers have been used in spines. These choices remain a matter of preference, and sometimes have to do with the technical attributes of the spines in terms of size, scalability, buffering, and virtual output queuing capabilities. We view a spine as two stages of switching where, in some cases, the stages are traditional packet switching and, in others, they are cell switching, and we strive to find each its optimal role within a cloud data center network.

Cloud-specific alternatives have emerged for packaging ASICs into systems, including multi-ASIC systems that behave as one large switch, multi-ASIC systems that behave as discrete individual routers, and multi-ASIC systems that deliberately segregate their ports into disjoint networks. These packaging considerations are closely related to the topology decisions, as opposed to traditional enterprise principles that define each switch/router box as a versatile entity whose usage in the network is not known a priori.
The Racks

Since the early days of mega data centers, we have observed that one of the most effective weapons against large-scale design complexity is homogeneity. Reducing the number of choices reduces variability and simplifies qualification matrices. We do not suggest that a data center is homogeneous. We assume a homogeneous pattern for a given deployment in time, leading to the data center having islands of homogeneity that are deployed for different purposes or at different points in time.

This notion of homogeneity is being extended in the industry to the compute infrastructure (including local storage and leaf switch functions) by the creation of rack-level designs, where the mechanical and functional unit of deployment is the entire rack as described in [4]. The concept is not new, as even larger shipping container-based deployment and homogeneity have already been proposed and practiced [5]. The rack-level effort is interesting, as it exposes the boundaries and fault lines between traditional networking and server I/O. We do believe that intra-rack interconnects are best architected with pure network vs. server I/O approaches, such as those based on PCI Express (PCIe) specifications [6]. We designate our preference “homogeneity with choice” because one can create a homogeneous design while retaining the freedom to use best-of-breed components for each element in the design, for example, the server network interfaces, the switches, and the rack storage targets.

Ethernet-based network interfaces also make a much more efficient use of cable bandwidth—an expensive resource—because Ethernet is a streaming interface that “pushes” data, as opposed to PCIe being a load/store/memory interface that must pull data from a root complex. These I/O bus inefficiencies, along with the need to consume bandwidth for memory descriptors, are very notorious for short packet transactions and, generally, result in more lanes being used on the PCIe side of a device than its network side, often by a factor of four.

In light of the changes in the traditional storage hierarchy introduced by nonvolatile flash memory, along with new practices for distributed applications and compute building blocks such as very dense microserver systems, the role and location of storage are being reexamined. At the risk of sounding tautological, we conclude that the best technology for transporting networking traffic is a network link, and we set out to explore how to extend the benefits of a network link to various forms of storage needed in cloud scale: rack-level arrays, storage clusters, and even object storage.

References

WHITE PAPER: Cloud Scale Networking Part I: Architecture Considerations

© 2014 by Broadcom Corporation. All Rights Reserved. Broadcom®, the pulse logo, Connecting everything®, and the Connecting everything logo are among the trademarks of Broadcom Corporation and/or its affiliates in the United States, certain other countries and/or the EU. Any other trademarks or trade names mentioned are the property of their respective owners.

Broadcom Corporation reserves the right to make changes without further notice to any products or data herein to improve reliability, function, or design.

Information furnished by Broadcom Corporation is believed to be accurate and reliable. However, Broadcom Corporation does not assume any liability arising out of the application or use of this information, nor the application or use of any product or circuit described herein, neither does it convey any license under its patent rights nor the rights of others.

Broadcom Corporation
5300 California Avenue
Irvine, CA 92617
Phone: 949-926-5000
Fax: 949-926-5203

info@broadcom.com  broadcom.com

CSN-ETP100-R • April 2014