# Facebook's Data Center Network Architecture

**Nathan Farrington and Alexey Andreyev**
*Facebook, Inc., 1601 Willow Road, Menlo Park, CA 94025, USA*
*Author e-mail address: farrington@fb.com*

**Abstract:** We review Facebook's current data center network architecture and explore some alternative architectures.
**OCIS codes:** (060.4250) Networks

## 1. Big data requires big networks

On a small website, the web server software and the database can reside and communicate on the same physical server, meaning that the only network communication is user traffic over the Internet. But Facebook's data sets are so large that it is impossible to deliver the Facebook experience using a single server. Instead, a front-end web server handles the original HTTP request, and then fetches data from a number of different cache, database, and backend servers in order to render the final page. All of this additional communication must traverse the internal data center network. In one measurement, a particular HTTP request required 88 cache lookups (648 KB), 35 database lookups (25.6 KB), and 392 backend remote procedure calls (257 KB), and took a total of 3.1 seconds for the page to completely load. If the user's original HTTP request was 1 KB, then this represents a 930x increase in internal data center traffic, only a subset of which will be returned to the user in the form of HTML and images.

In addition to the HTTP request amplification described above, Facebook has over 100 PB of data stored in its data warehouse [1]. This data has a number of uses, from building search indices [2] to capacity planning to optimizing product behavior. We use MapReduce (Hadoop and Hive) to execute queries over these large data sets. While the *map* phase of MapReduce runs locally on a given server's data, the *reduce* phase requires copying the results of the map phase to potentially thousands of other servers. If enough MapReduce jobs are run in parallel, the resulting network communication pattern can exhibit all-to-all characteristics.

## 2. Facebook's current "4-post" data center network architecture

Fig. 1 shows Facebook's current data center network. Each rack contains a rack switch (RSW) with up to forty-four 10G downlinks and four or eight 10G uplinks (typically 10:1 oversubscription), one to each cluster switch (CSW). A cluster is a group of four CSWs and the corresponding server racks and RSWs. Each CSW has four 40G uplinks (10G×4), one to each of four "FatCat" aggregation switches (typically 4:1 oversubscription). The four CSWs in each cluster are connected in an 80G protection ring (10G×8) and the four FC switches are connected in a 160G protection ring (10G×16). Intra-rack cables are SFP+ direct attach copper; otherwise MMF is used (10GBASE-SR).

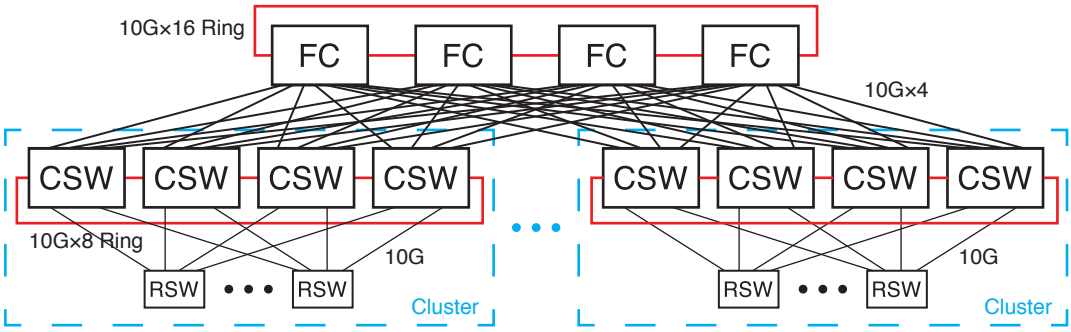


**Fig 1. Current "4-post" data center network architecture. RSW (rack switch), CSW (cluster switch), FC ("FatCat" aggregation switch)**

The current 4-post architecture solved a number of issues Facebook had encountered in the past. For example, network failures used to be one of the primary causes of service outages. The additional redundancy in 4-post has made such outages rare. In another example, traffic that needed to cross between clusters used to traverse expensive router links. The addition of the FC tier greatly reduced the traffic through such links.

The main disadvantages of 4-post are all a direct result of using very large, modular CSW and FC switches. First, a CSW failure reduces intra-cluster capacity to 75%; an FC failure reduces inter-cluster capacity to 75%. Second, the cluster size is dictated by the size of the CSW. This architecture results in a fewer number of very large clusters, making it more difficult to allocate resources among Facebook product groups. Third, large switches are produced in smaller volumes from fewer manufacturers and have higher per-bandwidth CAPEX and OPEX. Fourth, large switches are often oversubscribed internally, meaning that not all of the ports can be used simultaneously

without resulting in oversubscription. Fifth, large switches are often very proprietary. This can lead to months and years between bug fixes, make it impossible to deploy custom protocols, and severely complicate the task of managing, monitoring, and measuring a large data center network.

## 3. Alternative data center network architectures

The 4-post architecture evolved over time to solve Facebook's specific networking challenges. It is worth considering alternative architectures used in industry or described in the literature.

The world's fastest supercomputer, the ORNL Titan [3], uses the Cray Gemini 3D Torus interconnect providing 11.96 Pb/s of capacity using 9,344 identical switches. A 3D Torus yields a number of benefits over 4-post. Being a direct network, there are no aggregation switches that can fail (only RSWs). The basic unit of resource allocation is a single rack. Only small, commodity switches (RSWs) are used, delivering lower per-bandwidth CAPEX and OPEX. And of the 56,064 links required, most are short neighbor links.

Unfortunately, the 3D Torus may not be ideally suited to Facebook's requirements. It is not possible to incrementally deploy a 3D Torus without taking some of the network offline. Sharing the network between different applications while achieving latency or throughput guarantees can be difficult. And the biggest challenge may be the need to create software that is network aware in order to properly route traffic and deal with bottlenecks and failures.

Another interesting topology is a 5-stage Folded Clos using a large number of small, commodity switches [4] (see Fig. 2). The FSW (fabric switches) and SSW (spine switches) are identical and differ only in their configuration and placement in the network. Like the 3D Torus, the use of small, commodity switches delivers lower per-bandwidth CAPEX and OPEX. Using $k$-port switches yields $\frac{k^3}{4}$ external-facing ports. For example, setting $k = 48$, 64, and 96, with a link rate of 40G yields 1.1 Pb/s, 2.6 Pb/s, and 8.8 Pb/s of aggregate capacity, respectively. A group of $\frac{k}{2}$ adjacent RSWs is called a pod. Consider the case of $k = 96$. An FSW failure reduces intra-pod capacity to 97.9% and an SSW failure reduces inter-pod capacity to 99.96%. The base unit of allocation, a pod, is only 48 racks. Pods can be deployed incrementally (i.e., scaled out) as the data center is built out, without requiring downtime or rewiring. But the biggest advantage may be that software need not be written to be network aware in order to get good performance.

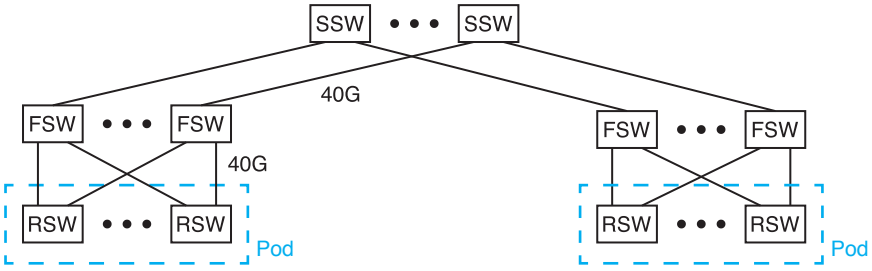


**Fig 2. Scale-out 40G data center network architecture. RSW (rack switch), FSW (fabric switch), SSW (spine switch)**

Unfortunately, the Folded Clos topology is well known for its daunting cabling complexity ( $\frac{k^3}{2} = 442{,}368$ internal links). Novel solutions for fiber management could be required to make the Folded Clos feasible. VCSELs now support rates of 25G, but switches are designed for 10G or 40G. Does the industry need a 25G switch? Another possibility is to move away from MMF towards SMF and take advantage of the additional channels available with WDM. Likewise, using more powerful codes to pack 2, 4, or more bits per symbol instead of today's 1 bit per symbol, would reduce the number of cables at the cost of increased power consumption. By carefully packaging the merchant silicon ASICs and building custom boards, it is possible to hide the cables as copper traces on backplanes [5]. Solving this cabling complexity challenge may ultimately prove to be an industry-wide problem.

## 4. References


[1] https://www.facebook.com/notes/facebook-engineering/under-the-hood-scheduling-mapreduce-jobs-more-efficiently-with-corona/10151142560538920

[2] https://www.facebook.com/notes/facebook-engineering/under-the-hood-building-out-the-infrastructure-for-graph-search/10151347573598920

[3] http://www.top500.org/featured/systems/titan-oak-ridge-national-laboratory/

[4] Al-Fares, M., Loukissas, A., and Vahdat, A. *A scalable, commodity data center network architecture.* Proceedings of ACM SIGCOMM, 2008.

[5] Farrington, N., Rubow, E., Vahdat, A. *Data Center Switch Architecture in the Age of Merchant Silicon*. Proceedings of IEEE Hot Interconnects, 2009.