State of the Art and Future Trends of Datacenter Networks

D

George Michelogiannakis
Lawrence Berkeley National Laboratory, USA

INTRODUCTION

Recent years have brought an unprecedented demand for storage and transport of data worldwide. This trend was caused by online services such as social networking and multimedia streaming, which have rapidly increased in popularity since their debut. For example, Facebook launched in 2004 and has grown to have 955 million active users in the year 2012, with 552 million of them being active daily by average (Associated Press). Moreover, YouTube provides free video streaming to an average of 490 million unique users per month, each of whom spends 5 hours and 50 minutes (Bullas, 2011). Except for bandwidth for accessing stored data, these services constantly need to store new data. In the case of YouTube, an average of 35 hours of video is uploaded every minute (Bullas, 2011). While multimedia and social networking services are prime examples, there are more applications such as Internet search engines like Google and services which provide computation power in remote servers, such as Amazon.

To meet the growing demands, companies built large datacenters in various parts of the world. Multiple datacenters have advantages compared to a single larger datacenter in terms of redundancy in case of failure, load balance, as well as locality since the datacenters are spread across the world. A single datacenter can be as large as 147,000 ft^2 (13,656 m^2) in the case of Facebook's datacenter in Prineville, USA, which is equivalent to 2.5 times a U.S. football field (Rogoway, 2011). While current datacenters contain approximately 50,000 processing cores, they are expected to increase to 100,000 cores in the near future (Davis, 2010). The distance between two processing cores can be from millimeters to hundreds of meters. Due to the vast amount of heat dissipated, the U.S. datacenter sector consumed $4.5 billion for cooling in the year 2006 (US Environmental protection agency, 2007).

DOI: 10.4018/978-1-4666-5888-2.ch181

Large physical distances combined with the high performance demands place very high pressure on datacenter networks. Computation demands keep increasing due to the increasing number of cores and also the increase of user load. In fact, it is projected that communication growth will be exponential, and that for every byte written or read to or from a disk, 10KB are transmitted over a datacenter network (Astfalk, 2009). Moreover, the estimated compound annual grown rate (CAGR) for the server count in a datacenter is 17%, while the storage growth CAGR is 52% (Astfalk, 2009).

At the same time, latency is also critical because high latencies negatively affect user experience in online services. The goal of increasing bandwidth and reducing latency conflicts with the goal of keeping the implementation cost reasonable. For these reasons, datacenter networks have been the focus of a large part of the research community.

This article presents an overview of the current state of the art in datacenter networks, and ends with a discussion of future trends and recommendations.

BACKGROUND

Physical Organization

Datacenters are organized in racks. Each rack typically contains 42 vertical 44.45 milimeter U slots (Nathan Farrington, 2009). Each U slot can currently hold 2- to 4-socket processor motherboards, or parts of a network switch (router) instead. With this configuration, racks are 0.6 meters wide, 1 meter deep and 2 meters high. Racks themselves are typically organized in rows. Each rack has a cold side where air enters for internal cooling and a hot side where hot air exits. The rack's cooling system must evacuate the heat generated by the processor sockets, DRAM and networking equipment. Racks are placed such as to form cold and hot isles, to

assist the critical problem of cooling in datacenters. Cooling systems must be designed to accommodate the worst-case power consumption at 100% utilization. Cold rows are approximately 1.22 meters and allow human access to blades but not the cables, whereas hot rows are 0.9 meters, contain cables, and are the key to the datacenter's heat extraction strategy. This configuration is illustrated in Figure 1.

Cooling is a primary consideration of the entire building the datacenter is housed in. Because cooling and other systems in the building consume power, power usage effectiveness (PUE) was devised as a metric of efficiency and equals the ratio of a datacenter's total power to the power actually used by the computing equipment. The average datacenter PUE is 2.0, and the most efficient 1.2 (Google Inc.).

Topology

The network topology is mapped on the array of rows of racks. Typically, each rack contains a top-of-rack (TOR) switch where each multi-core motherboard connects to using cables (channels). TOR switches connect to a larger end-of-row (EOR) switch serving the entire row of racks (Nathan Farrington, 2009). EOR switches then connect to core switches, which are the backbone of the network. This essentially translates to a fat tree network topology, illustrated in Figure 2 (Leiserson, 1985), where leaves are TOR switches, the next level above them are EOR switches, and the top level consists of one or multiple core switches. In a fat tree, the bandwidth between any two levels of the tree is held constant by using higher-bandwidth or multiple channels in the upper levels to compensate for the fewer input/output ports. The fat tree topology is widely used in today's datacenters because cables are easy to route within and across racks, whereas with more complicated topologies any TOR switch can connect with another rack's TOR switch.

Figure 1. Cold air enters racks from cold row for cooling and exits to the hot rows. The building's cooling system maintains airflow and supplies cold air.

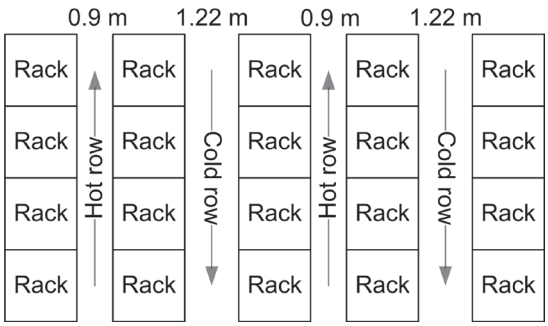
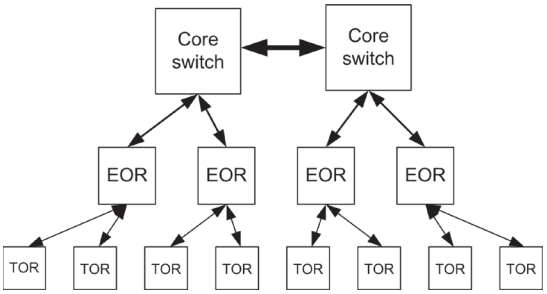


Figure 2. Switches connect to a pre-defined number of lower-level switches (in this example two) and one higher-level switch. The bandwidth between two levels is constant by increasing the channel bandwidth at higher levels.



Based on the baseline fat tree topology shown above, researchers have proposed topologies that provide more than one paths between a given source–destination pair. An example of such a topology is shown in Figure 3 (Mohammad Al-Fares, 2008). Path diversity can be critical in the presence of congestion in parts of the network (internal or at the edges), and allows packets from independent flows to proceed unaffected to their destinations. Furthermore, path diversity removes single points of failure and is necessary to provide system administrations the ability to power down nodes, switches, or remove channels without having to shut down part of the datacenter.

Topologies alternative to the fat tree have the disadvantage that their mapping to the 2D array of racks shown in Figure 1 is more complicated, but can reduce hop count using long (express) channels (Jung Ho Ahn, 2009). A popular example of a topology with express channels is the Dragonfly topology, which connects any source-destination pair with five hops with minimal routing, regardless of network size (John Kim, 2009).

Path Diversity

To make effective use of path diversity provided by the topology, various scheduling and flow control techniques have been proposed. Some techniques simply choose at random for each flow among all possible