Architecture of Warehouse-Scale Computers

- Networks are the connective tissue that binds 50,000 servers together. WSCs use a hierarchy of networks. Ideally, the combined network would provide nearly the performance of a custom high-end switch for 50,000 servers at nearly the cost per port of a commodity switch designed for 50 servers (fig below)

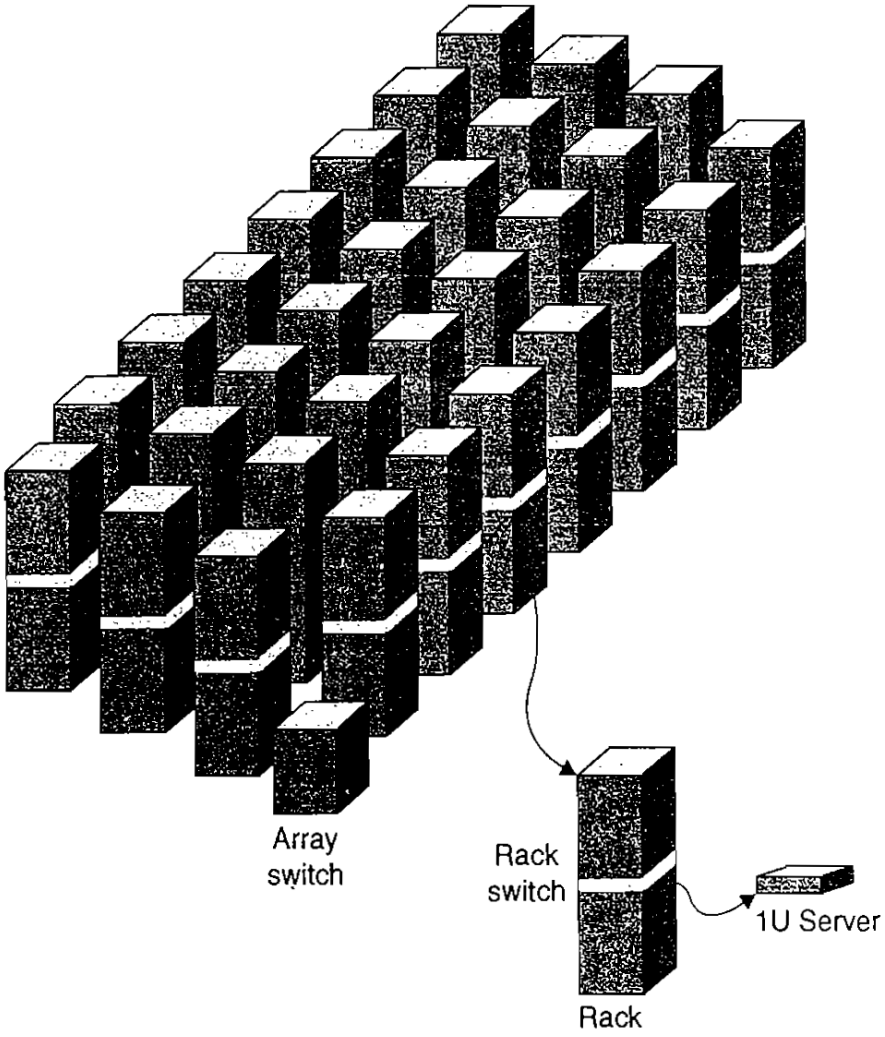

Fig: Hierarchy of switches in a WSC

- The 19-inch (48.26-cm) rack is still the standard framework to hold servers, despite this standard going back to railroad hardware from the 1930s. Servers are measured in the number of rack units (U) that they

occupy in a rack. One U Is 1.75inches(4.45cm) high, and that is the minimum space a server can occupy.

- A 7-foot (213.36-cm) rack offers 48 U, so it's not a coincidence that the most popular switch for a rack is a 48-port Ethernet switch. This product has become a commodity that costs as little as $30 per port for a 1 Gbit/sec Ethernet link in 2011 [Barroso and Holzle 2009]. Note that the bandwidth within the rack is the same for each server, so it does not matter where the software places the sender and the receiver as long as they are within the same rack.

- These switches typically offer two to eight uplinks, which leave the rack to go to the next higher switch in the network hierarchy. Thus, the bandwidth leaving the rack is 6 to 24 times smaller-48/8 to 48/2-than the bandwidth within the rack. This ratio is called ***oversubscription.***

Storage :

- A natural design is to fill a rack with servers, minus whatever space you need for the commodity Ethernet rack switch. This design leaves open the question of where the storage is placed. From a hardware construction perspective, the simplest solution would be to include disks inside the server, and rely on Ethernet connectivity for access to information on the disks of remote servers. The alternative would be to use network attached storage (NAS), perhaps over a storage network like Infiniband. The NAS solution is generally more expensive per terabyte of storage, but it provides many features, including RAID techniques to improve dependability of the storage.
- WSCs generally rely on local disks and provide storage software. that handles connectivity and dependability. For example, GFS uses local disks and maintains at least three replicas to overcome dependability problems. This redundancy covers not just local disk failures, but also power failures to racks and to whole clusters.

Array Switch :

- The switch that connects an array of racks is considerably more expensive than the 48-port commodity Ethernet switch. This cost is due in part because of the higher connectivity and in part because the bandwidth through the switch must be much higher to reduce the oversubscription problem. Barroso and Holzle [2009] reported that a switch that has 10 times the bisection bandwidth-basically, the worst-case internal bandwidth-of a rack switch costs about 100 times as much. One reason is that the cost of switch bandwidth for n ports can grow as n^2.

- Network switches are major users of content-addressable memory chips and of field-programmable gate arrays (FPGAs), which help provide these features, but the chips themselves are expensive.

WSC Memory Hierarchy :

	Local	Rack	Array
DRAM latency (microseconds)	0.1	100	300
Disk latency (microseconds)	10,000	11,000	12,000
DRAM bandwidth (MB/sec)	20,000	100	10
Disk bandwidth (MB/sec)	200	100	10
DRAM capacity (GB)	16	1040	31,200
Disk capacity (GB)	2000	160,000	4,800,000

Fig: Latency, bandwidth, and capacity of the memory hierarchy of a WSC

- Each server contains 16 GBytes of memory with a 100-nanosecond access time and transfers at 20 GBytes/sec and 2 terabytes of disk that offers a 10-millisecond access time and transfers at 200 MBytes/sec.

There are two sockets per board, and they share one I Gbit/sec Ethernet port.

- Every pair of racks includes one rack switch and holds 80 2U servers. Networking software plus switch overhead increases the latency to DRAM to 100 microseconds and the disk access latency to 11 milliseconds. Thus, the total storage capacity of a rack is roughly 1 terabyte of DRAM and 160 terabytes of disk storage. The 1 Gbit/sec Ethernet limits the remote bandwidth to DRAM or disk within the rack to 100 MBytes/sec.

- The array switch can handle 30 racks, so storage capacity of an array goes up by a factor of 30: 30 terabytes of DRAM and 4.8 petabytes of disk. The array switch hardware and software increases latency to DRAM within an array to 500 microseconds and disk latency to 12 milliseconds. The bandwidth of the array switch limits the remote bandwidth to either array DRAM or array disk to 10 MBytes/sec

Example What is the average memory latency assuming that 90% of accesses are local to the server, 9% are outside the server but within the rack, and 1% are outside the rack but within the array?

Answer The average memory access time is

$$(90\% \times 0.1) + (9\% \times 100) + (1\% \times 300) = 0.09 + 9 + 3 = 12.09 \text{ microseconds}$$

or a factor of more than 120 slowdown versus 100% local accesses. Clearly, locality of access within a server is vital for WSC performance.

Example How long does it take to transfer 1000 MB between disks within the server, between servers in the rack, and between servers in different racks in the array? How much faster is it to transfer 1000 MB between DRAM in the three cases?

Answer A 1000 MB transfer between disks takes:

Within server = 1000/200 = 5 seconds

Within rack = 1000/100 = 10 seconds

Within array = 1000/10 = 100 seconds

A memory-to-memory block transfer takes

Within server = 1000/20000 = 0.05 seconds

Within rack = 1000/100 = 10 seconds

Within array = 1000/10 = 100 seconds