Design of Warehouse Scale Computers (WSC)

Akhilesh Kondra
Student #: 009380878
Computer Science Department
San Jose State University
San Jose, CA 95192
408-668-5975
kondra.akhilesh@gmail.com

ABSTRACT

Warehouse scale computers (WSC), by definition, are a large number of the hardware and software resources that work together to efficiently deliver good levels of Internet service performance. These Warehouse scale computers come together to form a Datacenter. WSC has the Common resource management flexibility, which is achieved by a holistic approach of design architects and deployment. The architectural organization of these systems has played a pivotal role in the last few years. Therefore, it is beneficial to understand the architecture at a high level as it sets the background for succeeding. This paper primarily focuses on the architecture of WSC, main factors that influence their designs, operation, and cost structure.

1. INTRODUCTION

There has been a staggering growth in the internet services such as Web-based email, search and social networks and the worldwide availability of high-speed connectivity ,which has forged a trend towards server-side or cloud computing. Computing and storage are moving from PC-like clients to large Internet services. The paradigm shift to server-side computing is driven primarily not only by the need for user experience improvements, such as ease of management like no configuration or backups needed and ubiquity of access, but also by the advantages it offers to vendors like Software as a service allows faster application development because it is simpler for software vendors to make changes and enhancements. Instead of updating many millions of client, vendors need only coordinate improvements and fixes inside their datacenters and can restrict their hardware deployment to a few well-tested configurations.

The trend toward for server-side computing and the burgeoning rise of Internet services, in terms of supply and consumerism, has created a new class of computing systems that we have named warehouse-scale computers, or WSCs. The name is meant to put forth the most distinguishing feature of these machines: the massive scale of their software infrastructure, data repositories, and hardware platform. This perspective is a departure from a view of the computing problem that implicitly assumes a model where one program runs in a single machine. In warehouse-scale computing, the program is an Internet service, which may consist of tens or more individual programs that interact to implement complex and myriad end-user services such as email, search, or maps. These programs might be implemented and maintained by different teams of engineers, perhaps distributed across organizational, geographic, and company boundaries.

2. ARCHITECTURAL OVERVIEW OF WSCs

No two hardware implementation of a WSCs' installations will be identical. There will certainly be variations so much so that even within a single organization, systems deployed in different years make use of different basic elements, reflecting the hardware improvements provided by the industry. On the other hand, the architectural organization of these systems has been relatively stable over the last few years. Therefore, it is imperative to describe this general architecture at a high level as it provides the necessary insights for subsequent discussions.

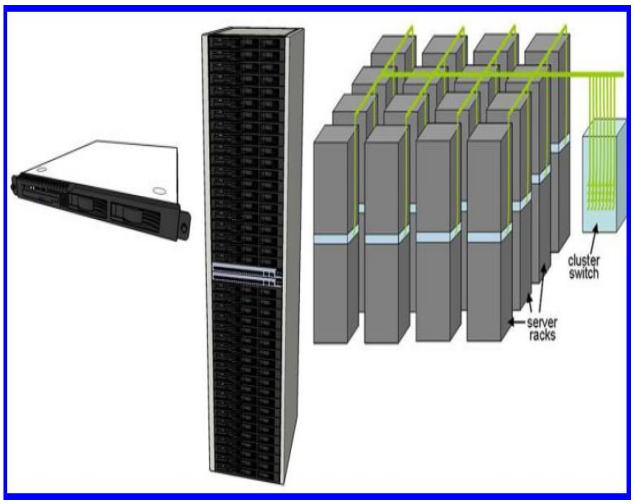

FIGURE 1.1: Architecture of WSC; Typical elements in warehouse-scale systems: 1U server (left), rack with Ethernet switch (middle), and diagram of a small cluster with a cluster-level Ethernet switch/router (right).

Figure 1.1 depicts some of the critical building blocks for WSCs. A set of low-end servers, typically in a 1U or blade enclosure format, are mounted within a rack and interconnected using a local Ethernet switch. These rack-level switches, which can use 1- or 10-Gbps links, have a number of uplink connections to one or more cluster-level (or datacenter-level) Ethernet switches. Servers are in 1U format [1]. These are mounted in the rack. And they are interconnected. 1U provides the height of the server that is place in a rack of 19inch or 23inch.

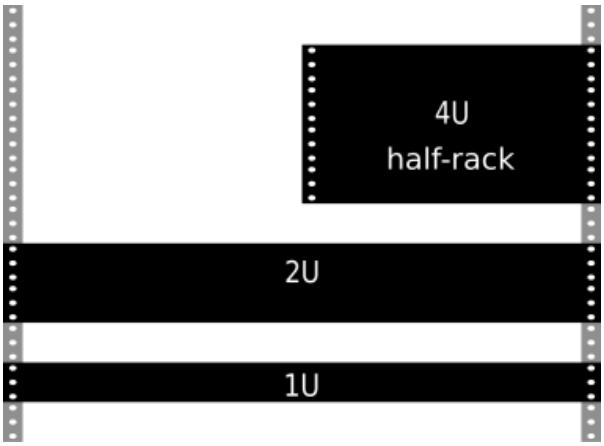

FIGURE 1.2: Different height of servers placed in a rack

2.1 Storage

FIGURE 1.3: Disk drives connected to servers directly

FIGURE 1.4: Network Attached Storage (NAS) devices that are directly connected to the cluster-level switching fabric

A important aspect of the architecture is ,that the Disk drives, shown in Fig. 1.3, can be connected directly to each individual server and be managed by a global distributed file system [2] or rather they can be part of Network Attached Storage (NAS) devices that in-turn are directly connected to the cluster-level switching fabric. There are significant differences in an NAS and collections of disks, they are: A NAS tends to be a simpler solution to deploy initially because it leverages the responsibility for data management and integrity to a NAS appliance vendor. Whereas, using the collection of disks directly attached to server nodes requires a fault-tolerant file system at the cluster level. This would be cumbersome to implement but whittle down the hardware costs and networking fabric utilization [5]. The replication model for these two approaches is also fundamentally different. A NAS provides extra reliability through replication or error correction capabilities within each appliance, on the contrary, systems with global distributed file system implement replication across different machines and consequently will use more networking bandwidth to complete write operations. Added to that, global distributed file systems keep the data available even after the loss of an entire server enclosure or rack and may allow higher aggregate read bandwidth because the same data can be sourced from multiple replicas. An additional advantage of having disks collocated with compute servers is that it enables distributed system software to exploit data locality.

2.2 Networking Fabric

Opting for a networking fabric for WSCs comes with a trade-off between speed, scale and cost. 1-Gbps Ethernet switches with up to 48 ports are more often a commodity component, costing less than $30/Gbps per server to connect a single rack [7]. As a result, bandwidth within a rack of servers tends to have a monolithic profile. However, network switches with high port counts, which are needed to interlace WSC clusters, have an altogether different price structure and are ten or more times expensive (per 1-Gbps port) than commodity switches. In other words, a switch that has 10 times the bi-section bandwidth costs about 100 times as much. As a result of this cost disconnectedness, the networking fabric of WSCs is often organized as the two-level hierarchy depicted in Figure 1.1. Commodity switches in each rack provide a fraction of their bi-section bandwidth for inter-rack communication through a handful of uplinks to the more costly cluster-level switches.

2.3 Storage Hierarchy

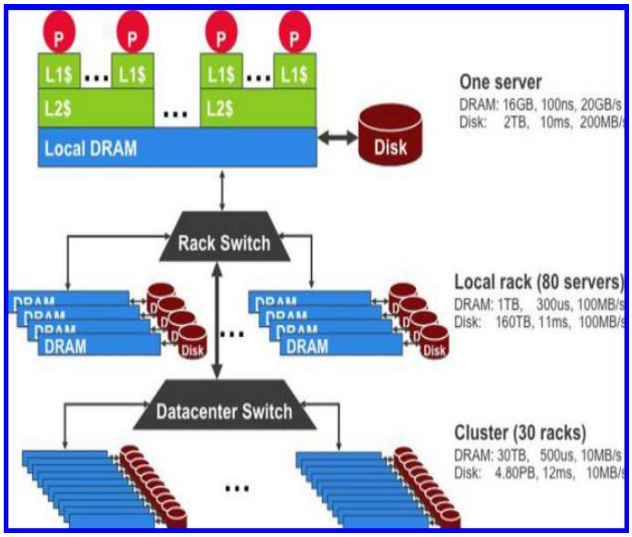

FIGURE 1.5: Storage hierarchy of a WSC.

Figure 1.5 portrays a programmer's perspective of storage hierarchy of a typical WSC. A server, in its composition, has 'n' number of processor sockets, each with a multi core CPU and its internal cache hierarchy, local shared and coherent DRAM, and a number of directly attached disk drives [9]. The DRAM and disk resources within the rack are accessible through the first-level rack switches (assuming some sort of remote procedure call API to them), and all resources in all racks are accessible via the cluster-level switch.

2.4 Quantifying Latency, Bandwidth and Capacity

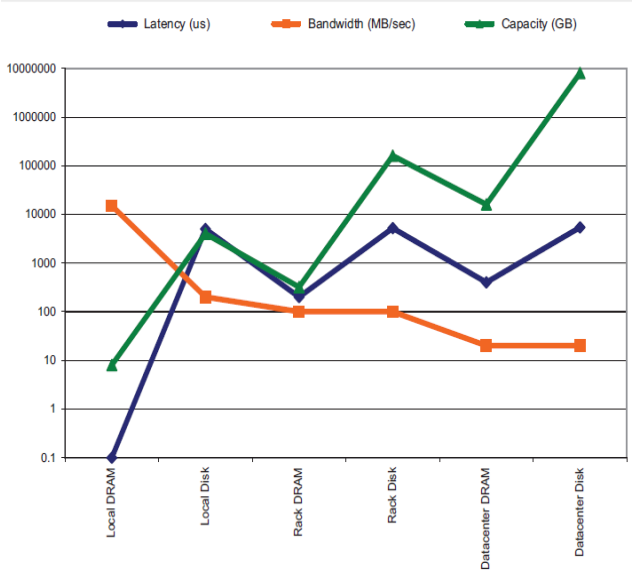

FIGURE 1.6: Latency, bandwidth, and capacity of a WSC.

Figure 1.6 makes an attempt to quantify the typical attributes of a WSC: latency, bandwidth, and capacity. For illustration we suppose a system with 2,000 servers, each with 8 GB of DRAM and four 1-TB disk drives. Each group of 40 servers is connected through a 1-Gbps link to a rack-level switch that has an additional eight 1-Gbps ports used for connecting the rack to the cluster-level switch (an oversubscription factor of 5).

Assumptions: 1) For Network latency numbers-A socket-based TCP-IP transport. 2) For networking bandwidth values-Each server behind an oversubscribed set of uplinks is using its fair share of the available cluster-level bandwidth. 3) The rack- and cluster-level switches themselves are not internally oversubscribed. 4) For disks, we show typical commodity disk drive (SATA) latencies and transfer rates.

The above graph depicts the relative latency, bandwidth, and capacity of each resource pool. For instance, the bandwidth available from local disks is 200 MB/s, while the bandwidth from off-rack disks is just 25 MB/s via the shared rack uplinks. On the flip side, total disk storage in the cluster is almost ten million times larger than local DRAM [7].

A humongous application that makes avail of many more servers than can fit on a single rack must deal effectively with these large discrepancies in latency, bandwidth, and capacity [6]. These discrepancies are much larger than those seen on a single machine, making it more intricate to program a WSC.

For the architects of WSCs the underlying challenge here is to level these discrepancies in a cost-cutting manner. Conversely, a crucial task for software architects is to bring about abstraction be developing cluster infrastructure and services that hides most of this complexity from application developers.

2.5 Power Usage

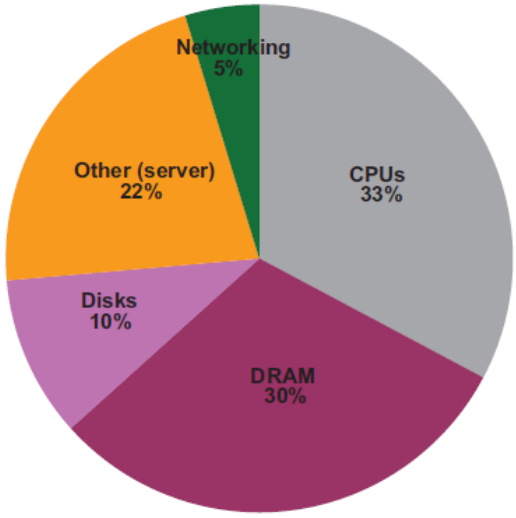

FIGURE 1.7: Approximate distribution of peak power usage by hardware subsystem in one of Google's datacenters (circa 2007).

There are two important facets that need to be considered in the design of WSCs: Energy and Power Usage, this because energy-related costs have become an indispensable component of the total cost of ownership of this class of systems. Figure 1.7 throws light on how energy is used in modern IT equipment by breaking down the peak power usage of one generation of WSCs deployed at Google [3] categorized by main component group.

However, this breakdown is subject to significant changes depending on how systems' configuration for a given workload domain, the graph indicates that CPUs can no longer be the sole focus of energy efficiency improvements because no one subsystem dominates the overall energy usage profile [8].

2.6 Handling Failures

The magnitude and sheer scale of WSCs at large, mandates that Internet services should possess software adequate tolerance for relatively high component fault rates. Disk drives, as a matter-of-fact, can exhibit annualized failure rates higher than 4% [10]. Different deployments have reported between 1.2 and 16 average server-level restarts per year. Given such immense component

failure rates, an application running across thousands of machines may need to be agile in their response to failure conditions by reacting on an hourly basis.

3. DATACENTERS

Datacenters essentially are buildings wherein multiple servers and communication gear are collocated because of they have a similar thread of environmental requirements and physical security needs, and also for the ease of maintenance. Warehouse scale computers can simply be referred to as datacenters, where scale is the only distinguishing feature.

Typically, Traditional datacenters are home to a large number of relatively small- or medium-sized applications, each simultaneously running on a dedicated hardware infrastructure that is de-coupled and protected from other systems in the same facility. Those datacenters host hardware and software for multiple organizational units or even different companies, which often have precious little in common in terms of hardware, software, or maintenance infrastructure, and tend not to communicate with each other at all.

A key characteristic of Datacenter economics is that is allows many application services to run at a low cost per user. For example, servers may be shared among thousands of active users, resulting in better utilization. In the similar lines, the computation itself may become cheaper in a shared service (e.g., an email attachment received by multiple users can be stored once rather than many times). All in all, servers and storage in a datacenter can be easier to manage than the desktop or laptop equivalent because they are under control of a single, knowledgeable entity.

3.1 How WSCs differ from a Datacenter

There are seemingly different aspects between WSCs and traditional datacenters: WSCs belong to a single organization, use a relatively monochrome hardware and system software platform, and share a common systems management layer. Also WSCs run a smaller number of very large applications (or Internet services), and the common resource management infrastructure allows the convenience of deployment flexibility. The requirements of homogeneity, single-organization control, and enhanced focus on cost efficiency motivate designers to take new approaches in constructing and operating these systems.

4. CHALLENGES OF WSCs

4.1 Cost Efficiency

There is a cost-overhead in building and operation of a large computing platform, and the quality of a service may be directly proportional to aggregate processing and storage capacity available, further leading to sky-rocketing costs and thus shifting our towards monetary savings measures . Let's consider information retrieval systems such as Web search, the growth of computing needs is driven by three main factors.

- Surge in service popularity that amounts to higher request loads.
- The size of the problem keeps snowballing as the Web is multiplying by millions of pages per day. As a repercussion, this increases the cost of building and serving a Web index.
- Despite achieving stability in throughput and data repository, the ever-changing competitive nature of this market continuously fuels innovations to improve the quality of results retrieved and the frequency with which the index is updated. Although some quality improvements can be made by smarter algorithms alone, most substantial improvements exhort additional computing resources for every request. Assume a search system that also purveys synonyms of the search terms in a query. In here, retrieving results is substantially more expensive. Either the search needs to retrieve documents that match a more complex query that includes the synonyms or the synonyms of a term need to be replicated in the index data structure for each term.

The unremitting need for more computing capabilities places cost efficiency a higher ground thus making it a primary metric of interest in the design of WSCs. A huge computing platform may be expensive, and the quality entirely depends on the aggregate processing and storage capacity available, further creating in spurt in costs and demanding an unavoidable focus on cost efficiency.

4.2. Not just a collection of servers

Datacenters ruling the roost in many of today's successful Internet services are not merely a collection of machines co-existing in a facility and interweaved with wires. It is this large cluster of servers that needs to be considered as a single component or a computing unit. WSCs have an additional layer of complexity beyond systems consisting of individual servers or small groups of server. WSCs introduce a significant new gambit to programmer productivity. This additional complexity arises invariably from the larger scale of the application domain and manifests itself as a deeper and less homogeneous storage hierarchy, higher fault rates and possibly higher performance vacillation.

5. SUMMARY/CONCLUSION

Computation is treading a path into the cloud & thereby into WSCs. Both software and hardware architects must be fully aware of the end-to-end systems to devise good solutions. We cannot do by developing individual single-server applications, and we can no longer neglect the physical and economic mechanisms that loom large in a warehouse full of computers. At one level, WSCs are simply put, a few thousand cheap servers connected via a LAN. In real-time scenario, building a cost-efficient large-scale computing platform that has the required reliability, security and programmability requirements for the next generation of cloud-computing workloads is an unqualified challenge that calls in question the required skillset. We hope researches on this type of machine make computer scientists in helping them understand this relatively new area, and we believe that in the years to come, their persistent efforts will cater to the variety yet fascinating problems arising from warehouse-scale systems.

6. REFERENCES

[1] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity datacenter network architecture," in Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication, Seattle, WA, August 17–22, 2008.

[2] Google Inc., "Efficient computing—step 2: efficient datacenters". Available at http://www .google.com/corporate/green/datacenters/step2.html.

[3] Google Inc., "Efficient Data Center Summit, April 2009". Available at http://www.google .com/corporate/green/datacenters/summit.html.

[4] Google Inc., "Efficient Data Center, Part 1". Available at http://www.youtube.com/watch?v=Ho1GEyftpmQ, starting at 0:5930.

[5] J. Elerath and S. Shah, "Server class disk drives: how reliable are they?", IEEE Reliability and Maintainability, 2004 Annual Symposium—RAMS, January 2004. doi:10.1109/ RAMS.2004.1285439

[6] E. V. Carrera, E. Pinheiro, and R. Bianchini, "Conserving disk energy in network servers," in Proceedings of the 17th Annual International Conference on Supercomputing, San Francisco, CA, June 23–26, 2003. ICS '03. doi:10.1145/782814.782829

[7] An Introduction to the Design of Warehouse-Scale Machineshttp://ieeexplore.ieee.org/ebooks/6813234/6813235.pdf ?bkn=6813234&pdfType=book

[8] Computer Architecture Techniques for Power-Efficiency Stefanos Kaxiras and Margaret Martonosi 2008

[9] B. Schroeder, E. Pinheiro, and W.-D. Weber, "DRAM errors in the wild: a large-scale field study," to appear in the Proceedings of SIGMETRICS, 2009.

[10] E. Pinheiro, W.-D. Weber, and L. A. Barroso, "Failure trends in a large disk drive population," in Proceedings of 5th USENIX Conference on File and Storage Technologies (FAST 2007), San Jose, CA, February 2007.