the Availability Digest

www.availabilitydigest.com

How Does Google Do It (part 2)

November 2012

Google has transformed our relationship with information. No longer do we go to the library to do our research or consult an encyclopedia. We type in a query for Google and instantly (often before we have even finished typing our query), Google gives us a long list of postings that can be found on the Web. True, this information has not been vetted; so we have to be careful about its accuracy and authenticity. But this has not slowed our adoption of the new, powerful information access technology that has rapidly evolved over the last decade.

Today, Google indexes twenty million web pages every day to support its searches. It handles three billion daily search queries. It provides free email to 425 million Gmail users. This level of activity is not supported by your everyday data center. It requires a massive network of close to a million computers spread all around the world.

How Google does this has been a closely guarded secret, and much of it still is. We reported on Google's technology for massive parallel computation in an earlier *Availability Digest* article entitled "How Does Google Do It?"[1]

Google has now opened its kimono a little bit more. Strongly committed to green practices, Google has made major inroads on energy savings in its data centers and feels that the extreme security regarding its progress in this area undercuts that commitment. So in 2009, Google opened its doors to provide an insight into its energy-conservation practices.[2]

Just recently, in October, 2012, Google provided a more detailed peek into its data centers and allowed the observations to be published.[3] We look at these reports in this article along with some other insights that have been published previously in the Availability Digest.

Google's Compute Capacity

Google is now one of the world's largest computer manufacturers. It builds all of its own servers and much of its networking equipment. No one seems to know exactly how many servers Google deploys in its worldwide data centers, but estimates indicate that there are several hundred thousand servers, perhaps approaching a million.

Google builds servers that are tailored for its use. They would not make good general-purpose computers. They are 2U (3.5") rack-mounted servers with no cases – they slide right into the racks. They

[1] How Does Google Do It?, *Availability Digest*; February 2008.
http://www.availabilitydigest.com/public_articles/0302/google.pdf
[2] Google uncloaks once-secret server, *CNet*; April 1, 2009.
[3] Steven Levy, Google Throws Open Doors to Its Top-Secret Data Center, *Wired*; October 17, 2009.

have no graphics boards (no monitors are driven by these servers) and no wireless capability. They are x86-based, but conjecture abounds that Google may someday make its own chips.

Central power supplies provide only twelve-volt power to the racks rather than the five-volt and twelve-volt feeds of commercial rack power supplies. The five-volt sources are generated on the mother board. This allows Google to run its power supplies efficiently at near peak power, and power can be distributed with smaller power buses because of the reduced amperage that they must carry due to the higher voltage. Both of these factors reduce the heat generated by a server.

Each server has its own twelve-volt battery to provide backup power in the event of a data center power outage. This is more efficient than a large, centralized UPS, which must have its own cooling system. In addition, reliability is increased because a battery failure will take down only one server rather than an entire data center.

Since 2005, Google's data centers have been built from standard shipping containers, or pods. Each pod is outfitted with 1,160 servers and provides its own cooling.

Google's system capability goes far beyond its data centers. Stung by some early failures of telecom operators, Google has bought up many abandoned fiber-optic networks for pennies on the dollar. In addition, it has started laying its own fiber. It has now built a mighty empire of glass circling the world in fiber.

Concerned about network capacity after acquiring YouTube, Google began to establish mini-data centers to store popular videos. The mini-data centers are often connected directly to ISPs like Comcast and AT&T. If you stream a video, you may not be receiving it from one of Google's giant data centers but rather from a mini-data center just a few miles away.

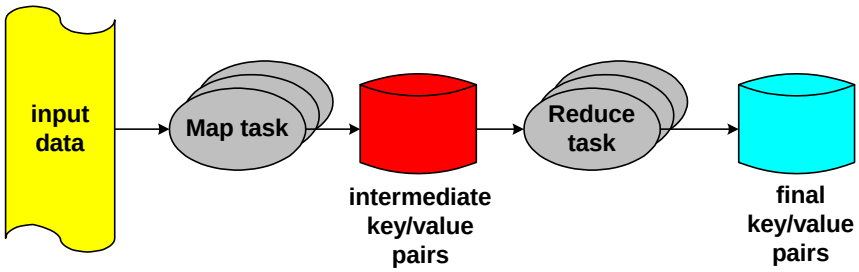
Managing All That Data

Search for "availability" on Google. How does Google give you 801,000,000 web references in 190 milliseconds? Through an extremely efficient, large-scale index of petabytes (that's millions of gigabytes) of web data.

A major problem Google faces is how to process the massive amount of worldwide data in a time that makes the indices useful. Google has solved this problem by building an equally massive parallel computing facility driven by its *MapReduce* application. MapReduce distributes applications crunching terabytes of data across hundreds or thousands of commodity PC-class machines to obtain results in seconds.[4] MapReduce takes care of partitioning the input data, scheduling the program's execution across the machines, handling balky and failed machines, balancing the load, and managing intermachine communications.

MapReduce comprises a set of Map tasks and a set of Reduce tasks. Many Map tasks are created, each working on a portion of the input data set. The Map tasks parse the input data and create *intermediate* key/value pairs. These are passed to the Reduce tasks, which merge and collate the results of the various Map tasks to generate the *final* key/value pairs.

[4] How Does Google Do It?, *Availability Digest*; February 2008.
http://www.availabilitydigest.com/public_articles/0302/google.pdf



The MapReduce Function

For instance, consider the search for the word "availability." The task is to determine the number of entities in the input data (the web-page indices) that contain the word "availability." The input data is partitioned across thousands of Map tasks. Each Map task generates a key/value pair for each word in its portion of the input data set. This key/value pair is simply the word (the key) followed by the number "1" (the value). Each key/value pair is written to an intermediate key/value store.

The Reduce function reads the intermediate key/value store and merges by key all of the intermediate key/value pairs that have been generated by the Map tasks. In this simple case, the Reduce function will simply add up the number of key/value pairs for each "availability" key and will store the final count in the final key/value store. The final key/value will be "availability/801,000,000."

Of course, the actual search is a little more complex than this. As part of the search, Google evaluates the relevance of the web page to the search criteria and organizes the list of references in priority order. This is what the user sees.

A typical configuration might have 200,000 Map tasks and 5,000 Reduce tasks spread among 2,000 worker machines. If a computer fails during the search task, its Map and Reduce tasks are simply assigned to other computers and repeated.

MapReduce has been so successful that an open-source version, Hadoop, has become an industry standard.

Energy Efficiency

Google is focused on energy efficiency of its data centers. Data centers consume 1.5% of all the energy in the world. Google's focus is evident in several areas in which it has reduced power and costs.

One area is the servers themselves. We have noted earlier that Google goes so far as to distribute only twelve volts to its servers, generating the five-volt sources on the motherboard, so that smaller power buses generating less heat can be used.

The local batteries associated with each server preclude the need for a UPS system. A UPS system has its own power requirements during normal operations and generates its own share of heat that must be removed from the data center.

Cooling the data center is perhaps the second biggest consumer of energy next to the computing equipment itself. Google has done much to reduce the energy needed for cooling, which we discuss in the next section.

Power Usage Effectiveness (PUE) is a measure of how efficiently a data center uses power. It is the ratio of the total power used by the data center, including cooling, lighting, and other overhead, as compared to the amount of power used by just the computing equipment. For a long time, a PUE of 2.0 was considered reasonable for a data center (that is, the power used by the computing equipment was half of

the total power consumed by the data center). Google has blown through that number. It typically achieves a PUE of 1.2 for its data centers. Its best showing is a PUE of 1.12 – the total power overhead is only 12% of that used by the computing equipment.

Cooling a Massive Data Center

In its early data centers, Google had a radical insight into data-center cooling. It found that a data center did not have to be cooled to arctic temperatures by giant air conditioners such that system operators needed to wear sweaters. Rather, it made extensive experiments and found that system reliability was quite sufficient if data centers were maintained at about 80° F (T-shirts and shorts weather).[5] This realization allowed Google to dramatically reduce the amount of energy needed for cooling – the primary factor in the PUE.

The cooling facilities for the computing equipment are built into the pods. The servers are organized into aisles. The front aisles, or cold aisles, are the access aisles for computer operators. The servers are removed and inserted via the front aisles. All of the cables and plugs are accessible to the front aisles. The front aisles are kept at a balmy temperature of 81° F.

The rear aisles are the hot aisles. The hot aisles are a tightly enclosed space – access is blocked with metal baffles at either end - and can reach temperatures of 120° F. (If maintenance personnel must enter a hot aisle, the servers are first shut down; and the aisle is allowed to cool.) The heat generated by the computer equipment is absorbed by water-filled coils and is pumped outside the building to cool.

This is where another advance has been made by Google. Rather than cooling water with energy-consuming chillers, it uses cooling towers and allows the water to drizzle down. Some water evaporates and cools the remaining water. The water is replaced from a local source such as a river or canal.

Thus, Google's data centers do not use large, energy-hungry computer room air conditioners. They basically just have to pump some water. Google has taken this a step further in some of its new data centers. It is locating data centers in cooler climes and is using the outside cold air to cool these data centers to 80° F.[6] It only has to turn on its air conditioners for the few days in the year when the outside air temperature rises to the point that this *air economizer* technique won't work.

Reliability

Google is very concerned about reliability. From a server viewpoint, its applications are all distributed across hundreds or thousands of servers. If one server fails, another one automatically picks up the load. The failed server is pulled from the rack, and a new one is inserted. The new server is then automatically added back into the server pool. This is different from today's virtual environments – there is no migration of a virtual machine from one physical server to another. Recovery time is unnoticeable as the other hundreds or thousands of servers involved in the distributed processing continue on with their tasks.

Security of data is a major concern. Not only is data backed up at several levels, but failed disk drives are physically destroyed to prevent the compromise of data stored on them.

Google deploys a Site Reliability Engineering team. Comprising normal Google engineers who spend most of their time writing production code, the SREs are like a geek SEAL team. Every year, the appointed SREs get leather jackets with military-style insignia. Their job is to wage a simulated (and sometimes real) war against Google to try to take down portions of it and to gauge the response of Google's incident managers. The war is called DIRT – Disaster-Recovery Testing.

[5] The Brain of the Beast: Google Reveals The Computers Behind The Cloud, *NPR*; October 17, 2012.
[6] Data Center Cooling Nature's Way, *Availability Digest*; May 2010.
http://www.availabilitydigest.com/public_articles/0505/cooling.pdf

Prior attacks have included causing leaks in water pipes, staging protests outside the gates of a Google facility in an attempt to distract attention from intruders trying to steal disks from servers, and cutting most of Google's fiber connections to Asia. If the incident response team cannot figure out the fixes, the attack is aborted so as not to inconvenience real users.

Summary

Google manages its massive data-processing requirements by building large data centers that behave as a single computer. Applications are distributed across the entire server floor. Its data centers are so large that Google provides its maintenance staff with personal transportation devices – foot scooters, which don't increase the PUE.

Google has a particular interest in energy efficiency. It achieves industry-leading PUEs with particular attention to cooling, server design, battery backup, and other initiatives such as foot scooters to minimize the energy overhead required to operate a major data center.