CHAPTER 1

The Google Cloud Platform Difference

Cloud computing as a vision is just 54 years young in 2015 (much older than either of this book's authors!). In 1961, John McCarthy introduced the idea of "computation being delivered as a public utility." Over the next five decades, various technological innovations enabled today's cloud computing, including the following:

- In 1960s, J. C. R. Licklider developed ARPANET—the forerunner to the Internet and what is considered to be the biggest contributor to the history of cloud computing in this era.
- In 1971, Intel engineer Ray Tomlinson developed software that allowed users to send messages from one computer to another. This subsequently was recognized as the first e-mail.
- In 1976, Xerox's Robert Metcalfe introduced Ethernet, essentially standardizing the wired network interface in computers.
- In 1991, CERN released the World Wide Web for general (that is, noncommercial) use.
- In 1993, the Mosaic web browser allowed graphics to be shown on the Internet. In the same year, private companies were allowed to use the Internet for the first time.
- During the late 1990s and early 2000s (famously known as the dot-com era), the availability of multitenant architectures, widespread high-speed bandwidth, and global software interoperability standards created the right environment for cloud computing to finally take off.

The realization of a global high-speed network and a utilities-based business model are the two major driving principles behind cloud computing.

What Is Cloud Computing?

Cloud computing is about abstracting the computing infrastructure and other associated resources and offering them as service, usually on a pay-per-use basis, over the Internet. The service can be targeted for human consumption or consumption by other software systems. Users just need a web browser to access services; software systems can consume services using a web application programming interface (API). This abstraction is often realized through a technical process called *virtualization*.

WHAT IS VIRTUALIZATION?

Virtualization is a process through which a hardware resource (such as a server or network) is cloned as an in-memory resource and is used as the (virtual) foundation to support a software stack. Virtualization is not an entirely new concept; virtual memory, for example, is used extensively in modern operating system(s) for security, for process isolation, and to create an impression that more memory is available than is actually present. Virtualization also makes it easy to transfer a virtual resource to another system when the underlying hardware fails.

A good analogy to cloud computing is the electric grid that centralized the production, transmission, and distribution of electricity to consumers. Consumers simply plug in to the grid, consume power, and pay for what they use without worrying about the nitty-gritty details of how electricity is produced, transmitted, and distributed. (You may be interested to know that, before the electric grid was invented, each organization produced its own electricity. Obviously, this required a large capital expense and was affordable only for the elite and rich.)

Cloud technology standardizes and pools IT resources and automates many of the maintenance tasks done manually today. Cloud architectures facilitate elastic consumption, self-service, and pay-as-you-go pricing. *Cloud* in this context refers to cloud computing architecture, encompassing both public and private clouds. But the public cloud has its own distinct set of advantages, which are hard to replicate in a private setting. This chapter focuses on these from both technical and nontechnical perspectives.

Technical Benefits of Using a Public Cloud

Several key performance benefits may motivate you to migrate to the public cloud. This section covers a few of these benefits.

Uptime

Most public cloud providers have redundancy built in as part of their system design. This extends from foundational utilities like electricity, Internet, and air conditioning to hardware, software, and networking. As a result, providers typically can offer uptime of 99.9% or more. This translates to expected downtime of just 8.76 hours per year (~1/3 day). All businesses can benefit from such high uptime for their IT infrastructure.

As independent businesses, public cloud service providers are able to provide legally binding service-level agreements (SLAs) that state the guaranteed uptime for their infrastructure and the penalties when those guarantees are not met. Such SLAs are not typically available from internal IT departments. The following URLs are for the SLAs of some of the popular cloud platform products covered in this book. In general, once a product is out of beta and into general availability (GA), the corresponding SLA should be available at `https://cloud.google.com/<product>/sla`:

- `https://cloud.google.com/compute/sla`
- `https://cloud.google.com/appengine/sla`
- `https://cloud.google.com/sql/sla`
- `https://cloud.google.com/storage/sla`
- `https://cloud.google.com/datastore/sla`
- `https://cloud.google.com/bigquery/sla`

Resource Utilization

Many organizational applications' resource needs vary by time. (Here, *resource* is a generic term and may refer to CPU, RAM, disk traffic, or network traffic.) As an example, an employee-facing app may be used more during the day and require more resources; it uses fewer resources at night due to reduced demand. This time-of-day variability leads to low overall resource usage in a traditional data-center setup. When you use a public cloud infrastructure, more resources can be (instantly) deployed when required and released when not needed, leading to cost savings.

Public cloud service providers have wide visibility on resource usage patterns across their customers and typically cluster them based on industry. Any application's resource usage may vary across individual system components; this is known as *multi-resource variability*. Resource usage patterns across industries are known as *industry-specific variability*.

Due to resource usage visibility, a public cloud service provider can reassign resources released by one customer to another customer, thereby keeping resource utilization high. If there is no demand for a particular resource, the provider may shut down the corresponding infrastructure to save operational costs. This way, the provider is able to handle applications whose resource needs are spiky in nature.

Expertise

Pubic cloud service providers have experienced system and network administrators along with 24×7 hardware maintenance personnel on site, owing to the tight SLAs they provide. By using a public cloud, companies can indirectly tap on this expert pool.

It would be challenging for a small or medium-size business to recruit, train, and maintain a top-notch team of domain experts, especially when deployment size is limited. Even larger companies are sometimes unable to match the deep expertise available at a public cloud service provider. For example, the well-known file-sharing company DropBox, which has millions of users, runs entirely on a public cloud.

Economic Benefits of Using a Public Cloud

In addition to the technical benefits of using a public cloud, there are several economic advantages to doing so. This section discusses the economic benefits of deploying on a public cloud, based on typical business yardsticks.

TCO

Total cost of ownership (TCO) refers to the total cost of acquiring, using, maintaining, and retiring a product. When you understand TCO, you will realize that many hidden costs usually are not accounted for. Specifically, TCO should include core costs such as the actual price of hardware/software and non-core costs such as time spent on pre-purchase research, operating costs including utilities, manpower, maintenance, and so on. Non-core costs typically are not included with traditional purchases and are bundled into administrative costs.

In the context of public cloud computing, TCO usually refers to software and/or hardware made available via lease. Interestingly, it avoids many non-core costs such as purchase-order processing, shipping, installation and so on.

Economies of Scale

Businesses (or customers) save more when they make a bulk purchase—the seller is willing to reduce its profit margin per unit for large sales. This is how big buyers, such as large companies, are able to get better deals compared to smaller companies in traditional transactions.

In the case of a public cloud, the buyer is the public cloud service provider such a Google Cloud Platform or Amazon Web Services. The larger the public cloud service provider, the more hardware it is likely to purchase from OEMs and the lower the price per unit. Public cloud service providers typically pass some of these savings to their customers (similar to a cooperative society model). This practice puts individual developers and companies of all sizes on the same level playing field, because they get the same low pricing for hardware/software.

CapEx and OpEx

Capital expenditures (CapEx) and *operational expenditures (OpEx)* are linked and refer to expenses incurred at different points in a product's consumption lifecycle. CapEx usually refers to large upfront expenses incurred before commencing use of a product, such as building a data center or acquiring hardware such as servers and racks and procuring Internet connectivity. OpEx refers to the associated operational expenses after a product is purchased and during its lifetime, such as manpower, utilities, and maintenance. The traditional wisdom is that high CapEx leads to low OpEx, whereas low CapEx leads to higher OpEx. Largely due to economies of scale, a public cloud service consumer enjoys low CapEx and low OpEx while transferring the large CapEx to the public cloud service provider, essentially creating a new economic model.

ROI and Profit Margins

Return on investment (ROI) and *profit margins* are strongly linked to one another and are key selling points for adopting a public cloud. ROI refers to the financial gain (or return) on an investment, and the profit margin is the ratio of income to revenue. By using a public cloud, an organization reduces its expenditures, and thus its ROI and profit margins are higher. Such higher returns are more visible in small and medium-sized businesses that have relatively high CapEx (because of low purchase quantities) when starting up.

Business Benefits of Using a Public Cloud

In addition to the technical and economic benefits, there are several business-process advantages to using a public cloud. This section describes a few of them.

Time to Market

Responsiveness is crucial in today's business environment. Business opportunities often arrive unannounced and are short-lived. Winners and losers are often determined by who is able to move faster and grab opportunities. Such opportunities typically require new/additional IT resources, such as computational power or bandwidth. A cloud service provider can provide these almost instantaneously. Hence, by using a public cloud, any business can reduce the time it takes to bring a product to market. In comparison, using the traditional route of building/acquiring infrastructure first, introducing a new product would require days if not weeks of onsite deployment.

Using a public cloud leads to reduced opportunity costs, increases agility, and makes it easy to respond to new opportunities and threats. The same quick response times also apply to shedding unneeded capacity. In summary, public cloud computing enables just-in-time procurement and usage for just as long as needed.

Self-Service

One of the hallmarks of the public cloud is the easy-to-use, remotely accessible interface based on modern web standards. All large public cloud service providers offer at least three interfaces: a web-based, graphical, point-and-click dashboard; a console-based command-line tool; and APIs. These enable customers to deploy and terminate IT resources anytime. These facilities make it easy for customers to perform self-service and further reduce time to market. In a traditional setting, even if IT deployment is outsourced to a third party, there is usually a lot of paperwork to be done, such as a request for quotes, purchase orders, and invoice processing.

Pay per Use

One of the promises of a public cloud is no lock-in through contracts. *No lock-in* means no upfront fees, no contractual time period, no early termination penalty, and no disconnection fees. Customers can move to another public cloud provider or simply take things onsite.

Google Cloud Platform adopts this definition and charges no upfront fees, has no contractual time period, and certainly charges no termination/disconnection fees. But Amazon Web Services offers a contract-like reservation plan that requires an initial payment to reserve resources and have lower usage costs during the reservation period. The downside of this reservation plan is that the promised savings are realized only if the same resource type is used nonstop the entire time.

The pay-per-use business model of a public cloud allows a user to pay the same for 1 machine running for 1,000 hours as they would for 1,000 machines running for 1 hour. Today, a user would likely wait 1,000 hours or abandon the project. In a public cloud, there is virtually no additional cost to choosing 1,000 machines and accelerating the user's processes.

WHAT IS SCALABILITY?

Scalability is a process through which an existing resource can be expanded on an on-demand basis either vertically or horizontally. An example of vertical scalability would be to upgrade a server's RAM from 2GB to 4GB, whereas horizontal scalability would add a second server with 2GB RAM. Scalability can be automatic or manual, but the end user should be able to update resources on an on-demand basis using either a web-based dashboard or an API.

Uncertain Growth Patterns

All organizations wish for exponential growth, but they can't commit sufficient IT infrastructure because they are not certain about the future. In a traditional setup, such scenarios result in unused capacity when growth is less than predicted or result in unhappy customers when the installed capacity is not able to handle additional load. Arbitrary loads are best handled by using public cloud deployments.

Why Google Cloud Platform?

Google Cloud Platform is built on the same world-class infrastructure that Google designed, assembled, and uses for corporate products like Google search, which delivers billions of search results in milliseconds. Google has also one of the largest, most geographically widespread, most advanced computer networks in the world. Google's backbone network comprises thousands of miles of fiber-optic cable, uses advanced software-defined networking, and is coupled with edge-caching services to deliver fast, consistent, scalable performance. Google is also one of the few companies to own a private fiber-optic cable under the Pacific Ocean.

Google Cloud Platform empowers software application developers to build, test, deploy, and monitor applications using Google's highly scalable and reliable infrastructure. In addition, it enables system administrators to focus on the software stack while allowing them to outsource the challenging work of hardware assembly, maintenance, and technology refreshes to experts at Google.

Hardware Innovations

Whereas a typical cloud service provider's strategy is wholesale-to-retail using standard hardware and software components, Google's approach has been to innovate at every level: hardware, networking, utilities, and software. This is evident from the multitude and variety of innovations that Google has introduced over the years. Needless to say, Google Cloud Platform benefits from all these innovations and thus differentiates itself from the competition:

- *Highly efficient servers:* In 2001, Google designed energy-efficient servers using two broad approaches: it removed unnecessary components like video cards, peripheral connections, and casing; and it used energy-efficient power supplies (that do AC-to-DC conversion) and power regulators (DC-to-DC conversion) and backup batteries on server racks.
- *Energy-efficient data centers:* In 2003, Google designed portable data centers using shipping containers that held both servers and cooling equipment. This modular approach produced better energy efficiency compared to traditional data centers at the time. Since 2006, Google has achieved the same efficiency using alternate construction methods.
- *Carbon neutrality:* In 2007, Google became a carbon-neutral Internet company, and it remains so today. Its data centers typically use 50% less energy compared to traditional data centers.
- *Industry-leading efficiency:* The cost of electricity is rapidly increasing and has become the largest element of TCO (currently 15%–20%). Power usage effectiveness (PUE) tends to be significantly lower in large facilities than in smaller ones. Google's data centers have very low PUE: it was 1.23 (23% overhead) in Q3 2008 and came down to 1.12 (12% overhead) in Q4 2014. This is significantly lower than the industry average of 1.7 (70% overhead).

All of these hardware innovations result in lower operational costs for Google, and the difference is passed to Google Cloud Platform users. This means customers save on costs.

Software Innovations

Infrastructure innovation is not just about hardware. Google has also led the industry with innovations in software infrastructure:

- *Google File System:* In 2002, Google created the Google File System (GFS), a proprietary distributed file system designed to provide efficient, reliable access to data using a large cluster of commodity hardware.
- *MapReduce:* In 2004, Google shared the MapReduce programming model that simplifies data processing on large clusters. The Apache Hadoop project is an open source implementation of the MapReduce algorithm that was subsequently created by the community.

- *BigTable:* In 2006, Google introduced the BigTable distributed storage system for structured data. BigTable scales across thousands of commodity servers and is used by several Google applications.
- *Dremel:* In 2008, Google shared the details of a system called Dremel that has been in production since 2006. Dremel is a scalable, interactive, ad hoc query system for analyzing read-only nested data that is petabytes in size. Dremel combines multilevel execution trees, uses a columnar data layout, and is capable of running aggregation queries over trillion-row tables in seconds. Dremel is the backend of Google BigQuery.
- *Pregel:* In 2009, Google created a system for large-scale graph processing. The principles of the system are useful for processing large-scale graphs on a cluster of commodity hardware. Examples include web graphs, among other things.
- *FlumeJava:* In 2010, Google introduced FlumeJava. FlumeJava is a pure Java library that provides a few simple abstractions for programming data-parallel computations. These abstractions are higher-level than those provided by MapReduce and provide better support for pipelines. FlumeJava makes it easy to develop, test, and run efficient data-parallel pipelines of MapReduce computations.
- *Colossus:* In 2010, Google created the successor to GFS. Details about Colossus are slim, except that it provides a significant performance improvement over GFS. New products like Spanner use Colossus.
- *Megastore:* In 2011, Google shared the details of Megastore, a storage system developed to meet the requirements of today's interactive online services. Megastore blends the scalability of a NoSQL datastore with the convenience of a traditional RDBMS in a novel way, and provides both strong consistency guarantees and high availability. Megastore provides fully serializable ACID semantics within fine-grained data partitions. This partitioning allows Megastore to synchronously replicate each write across a wide area network with reasonable latency and support seamless failover between datacenters.
- *Spanner:* In 2012, Google announced this distributed database technology. Spanner is designed to seamlessly operate across hundreds of datacenters, millions of machines, and trillions of rows of information.
- *Omega:* In 2013, Google introduced Omega—a flexible, scalable scheduler for large-scale compute clusters. Google wanted to move away from current schedulers, which are monolithic by design and limit new features. Omega increases efficiency and utilization of Google's compute clusters.
- *Millwheel:* In 2014, Google introduced Millwheel, a framework for fault-tolerant stream processing at Internet scale. Millwheel is used as a platform to build low-latency data-processing applications within Google.

All of these innovations are used to make Google Cloud Platform products, just as they are used to build Google's internal products. By using Google Cloud Platform, customers get faster access to Google innovations, thereby distinguishing the effectiveness of applications hosted on Google Cloud Platform.

Figure 1-1 shows a few important innovations from the above list to help visualize the continuous innovations by Google.

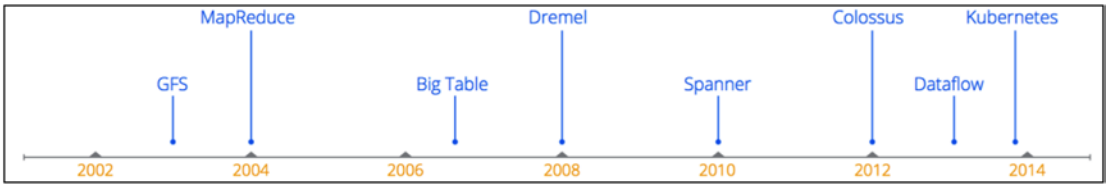

Figure 1-1. *Google's software innovations that are actively used in Google Cloud Platform*

Economic Innovations

In addition to making technical and infrastructure innovations, Google has also taken a fresh look at how to charge for cloud computing resources. Let's consider the economic innovations that Google has introduced in Google Cloud Platform, many of which benefit cloud platform users.

Typical public cloud providers, in particular Amazon Web Services, provide two types of pricing options for products: *on-demand* and *reserved* pricing. The guiding principle behind these two types of pricing options is to secure longer-term commitments from users. In the on-demand pricing model, the customer is free to use the resource for as long as needed and is free to leave anytime. There is no time contract or penalty for termination; this is typical of cloud hosting. In the reserved price model, the customer is required to pay a nonrefundable upfront fee and select the type of resource. As a result, the customer enjoys lower hosting charges for the specified time period.

There are several shortcomings in the reserved pricing model. First, because lower pricing is tied to the resource type, if the customer decides to switch resource types (say, due to different traffic patterns than expected), they are thrown back to the higher pricing model. Second, the upfront fees are time bound and not based on the number of hours of usage. Third, the upfront fees are not refundable if the customer decides to terminate early. In essence, the onus of choosing the right resource type and time duration is with the customer; there is no reconciliation if the actual workload is different from the expected workload.

Google's approach is that customers should be want to host on Google Cloud Platform due to its meritocracy and technical superiority. They should be able to leave anytime and not be tied through contract-like approaches. They should also be able to switch resource types anytime, as their needs change. Finally, while customers are hosting on Google Cloud Platform, they should enjoy the best pricing, on par with the industry.

To realize these objectives, Google has created a new type of pricing model called a *sustained-use discount*. Under this model, Google Cloud Platform automatically applies discounts to resources that run for a significant time. The discount is based on the cumulative amount of time a resource of a particular type is up rather than being tied to a single instance. This means two instances of equivalent specs running simultaneously or concurrently are given the same discount as long as the cumulative hosting period is above a threshold. Sustained-use discounts combined with per-minute billing ensure that customers get the best deal. The following list shows the sustained-use discounts as of this writing (March 2015):

- *0%–25%:* 100% of base rate
- *25%–50%:* 80% of base rate
- *50%–75%:* 60% of base rate
- *75%–100%:* 40% of base rate

Google has ventured to decipher the sometimes-complex world of cloud pricing by explaining how to calculate the cost of a cloud deployment. See the following post in the official Google cloud platform blog for details: `http://googlecloudplatform.blogspot.sg/2015/01/understanding-cloud-pricing.html`.

A Quick Comparison to AWS

This section highlights a few select features of Google Cloud Platform and how they compare with the incumbent public cloud provider, Amazon Web Services:

- Google Compute Engine, the Internet-as-a-service (IaaS) product from Google Cloud Platform, adopts a per-minute charging model except for the initial minimum 10-minute tier. On the other hand, AWS charges on an hourly-basis.

  Let's consider two example use cases. First, if you use an instance for 11 minutes, you pay for 11 minutes in Google Cloud Platform, but you pay for 60 minutes with Amazon Web Services. Second, if you use an instance for 1 minute, you pay for 10 minutes in Google Cloud Platform or 60 minutes in Amazon Web Services. In either case, you can see that Google Cloud Platform is cheaper than Amazon Web Services.

- Google Compute Engine is better suited to handle traffic spikes. This is because the Compute Engine load balancers don't require pre-warming, unlike AWS load balancers. In addition, pre-warming AWS load balancer requires customers to subscribe to AWS support. Compute Engine load balancers are able to scale instantly when they notice a sudden traffic spike.

  In 2013, Google demonstrated that its load balancers could serve 1 million requests per second on a sustained basis and within 5 seconds after setup. You are advised to read the full article at http://googlecloudplatform.blogspot.in/2013/11/compute-engine-load-balancing-hits-1-million-requests-per-second.html.

- Compute Engine's persistent disks (PDs) support a larger disk size (currently 10TB) compared with AWS. In addition, Google includes the I/O costs in the cost of the PD, thereby giving customers predictable costing. In the case of AWS, the cost of I/O is separate from the cost of the raw disk space. Moreover, other nice features include the ability to mount a PD to multiple VMs as read-only or a single VM in read-write mode.

- Compute instances are hosted as virtual machines in IaaS. Periodically, the IaaS service provider needs to do maintenance (host OS or hardware) on the platform. The hardware may also fail occasionally. In such cases, it is desirable to have the VM automatically migrate to another physical host. Compute Engine can do live migration.

- Google App Engine, the platform-as-a-service (PaaS) product from Google Cloud Platform, is in our view a pure PaaS product when compared with Beanstalk from Amazon Web Services. This is because Beanstalk is a management layer built on top of AWS EC2. The implication of this design choice is that Beanstalk needs to have at least one EC2 instance up all the time, which adds to hosting costs. App Engine, on the other hand, charges only when there is traffic and includes a monthly free tier.

- BigQuery, the big-data analytics product from Google Cloud Platform, is an integrated and fully hosted platform that scales to thousands of nodes and charges only for space and computation time. In comparison, the AWS equivalent (Red Shift) requires users to configure the system and also charges by the hour rather than based on usage.

- Google data centers (that host Google Cloud Platform's regions and zones) are spread globally and interconnected by Google's private fiber network. This means network traffic between two cloud platform zones passes through a private network and not over the public Internet. As of today, AWS does not have such a private network.

Overall, Google's approach with Google Cloud Platform is not to achieve feature parity with Amazon Web Services but to build products that are by far the best in the industry and in the process fill in the gaps in the AWS portfolio. Hence, the question to ask is whether your needs are being met by what Google Cloud Platform has today, rather than talking about what Google Cloud Platform doesn't have.

When talking about the strengths of Google Cloud Platform, it is important to acknowledge that Amazon Web Services currently has a broader portfolio of products and services than Google Cloud Platform. This is primarily due to the fact that AWS started much earlier, while Google was busy building the fundamentals right, as shown in the list of major software innovations earlier in this chapter.

Summary

We started this chapter by defining the concept of cloud computing. Following this, we leaped into public clouds, which we cover in this book. We shared with you the advantages of a public cloud from several perspectives: technical, economic, and business. Following this, we highlighted several Google research publications that are used to build the strong foundation of Google Cloud Platform. We concluded this chapter by listing the strengths of Google Cloud Platform when compared with Amazon Web Services.

The promise of the public cloud is not just cheaper computing infrastructure, but also faster, easier, more flexible, and ultimately more effective IT.