Why Cloud?

Now that we have a firmer grip on what the cloud is, it is easier to see why it is such a big deal. The cloud is transforming and will continue to transform the computing industry. In the future we may talk about on-premise data centers in the same way as we do about mainframes today, as relics of the past that definitely work well enough, but are mostly used because moving away from them entirely was too difficult. Some applications will probably remain in the on-premise data center as they have on the mainframe, but no one will build new data centers of their own unless they have very particular reasons. The cloud will be the default place to go. In this chapter, we look at why the cloud is the future. We see how the market is already moving fast toward the cloud and also look at the main drivers of cloud computing.

The Cloud Is the Future

Even companies that have the scale to run a competitive on-premise data center like Netflix have moved to the cloud. Some like Spotify, who could have started in the cloud, initially thought an on-premise data center was necessary but gave up and moved to the cloud. For many this is an emotional subject. One should not forget that the majority of the IT workforce see their area of expertise tied not to the cloud but to on-premise technologies and running a data center, a point we will return to in a later chapter.

The same was the case for the drivers of horse-drawn carriages who were taken over by the automobile. Many arguments could and were made in favor of the horse against the car. The danger of the automobile was one of them, which incidentally is also one of the major arguments against the cloud that
we also return to in a later chapter. This was real enough. Early automobiles were dangerous but much of the danger was attributed to the fact that pedestrian behavior did not change with the advent of this new technology. The early days of the automobile saw the necessity for a shift in how to use the technology and how people adapted to it. It is a process and it is not finished yet.

It is, however, clear when we look at projections of cloud adoption that the cloud, like the automobile, is here to stay. A Gartner survey of IT leaders in 2019 revealed that the cloud comprised around 10% of total IT budgets. This means that the cloud is already mainstream but has a lot more opportunity to grow and this growth is fast. Gartner projects that the entire cloud market will grow to $355BN in 2022 from $197BN in 2018. The expected Compound Annual Growth Rate is 16% compared to 7% for traditional IT. The cloud is also in the top three areas where global CIOs will increase their spending.

If we look at the different segments in the cloud, the SaaS segment is the one with the fastest growth. It’s projected to almost double in the five-year period from 2018 to 2022, from $86BN to $151BN, according to Gartner. The second highest growth is projected in the IaaS segment, from $32BN in 2018 to $74BN in 2022. Although these numbers are from before the COVID-19 pandemic, it has only exacerbated the need for the cloud and the speed of adoption is expected to increase.

Drivers for Cloud Adoption

To better understand the reasons for this shift, let’s look at what drives adoption of the cloud. Because the cloud is a new type of technology it has to compete with existing alternatives. That alternative is to run an on-premise data center. That is similar to running your own generator compared to receiving your electricity from a utility company. There could be reasons for that for sure, but most organizations have little benefit from doing so. However, as we saw earlier, it is not as simple just to switch from your local data center to the cloud as it would be to switch from a local generator to a utility’s power. It is much messier and therefore the drivers have to be stronger. Different organizations have different priorities, but most will have one of the following as a key driver for adopting the cloud over on-premise solutions.

Economy

It is evident that economic concerns are primary drivers of cloud adoption. This goes both for the provider and the consumer side of the equation. Both stand to gain from increased adoption of the cloud.
CAPEX to OPEX

A frequently mentioned sentence is that the cloud helps users move from CAPEX to OPEX. This means that costs are moved from capital expenditures (CAPEX), that is initial up-front investments, to operational expenditures (OPEX), or ongoing costs. When you buy a lot of equipment in order to support future demand of computing resources, a few things happen.

First of all, it ties up capital that could have been spent on other investments in the organization. This impacts the cash flow of a company. This is one of the key metrics that companies are measured on by investors and shareholders. The better the cash flow, the better the valuation of the company. The reason for this is not trivial. If a company runs out of cash and cannot pay its bills, it could result in bankruptcy, no matter how many servers they bought and no matter how promising the software on it is.

Another point is that a server is what is called a depreciating asset. That means it is something that loses its value over time. We are familiar with this dynamic in cars. The reason that private persons are opting to lease cars is the same reason that companies are opting for the cloud rather than investing in the hardware: we have little use for assets that we know will lose their value with time. This is not the same for all asset types. Real estate, for example, is not the same. It might appreciate and if it depreciates it will rarely go as fast or as low as servers. A server bought five years ago is virtually worthless and there is no big market for vintage computers.

Cost is more predictable in the cloud. There is typically some sort of monthly fee, although that can vary with consumption for some types of cloud services. But compare this to the old world of IT driven by capital investment. Projects are delayed all the time and budgeting is a yearly exercise, so no one knows if the money will be there next year. People in the IT industry will know the year end frenzy to spend this year’s budget on something random that may or may not come in handy next year. Then come next year, the project may be cancelled altogether, and the investment proved to be wasted. This is a familiar dynamic.

Operational expenses are also easier to scale with changes in the market. If a company suddenly sees a downturn and loses revenue, a cloud model based on paying monthly fees based on consumption is easier to scale than a CAPEX model, where capital has been committed at the beginning of a multiyear period. In the cloud, the costs scale more easily up and down with the activity of the company.

A more subtle motivation for switching to an OPEX is that lower levels of capital investment will inevitably lead to lower levels of debt. And lower levels of debt in a company improve its credit rating, which means it is able to secure needed capital at a lower price. Consequently, investment in the cloud can be seen as part of a virtuous circle from an economic perspective.
Total Cost of Ownership

Although moving from CAPEX to OPEX is a win in itself for most private companies, it is not the same for all types of organizations. Governmental organizations work on a different set of premises than do private companies. The government is not concerned in the same way about cash flow or depreciating assets since they don’t work by the logic of the balance sheet in the same way. Governmental organizations often have the reverse desire to bring down the commitments of the future and to pay for everything up front and not worry about it anymore. It has been common practice to allocate big capital budgets rather than operational budgets. The reasons are several, but a common one is that no political administration wants to be seen as having high levels of operational spending, as this is viewed by the public as unnecessary overhead. Investments, on the other hand, are better sold as visionary and meaningful. Unfortunately, this doesn’t translate well to the cloud model of consumption.

Consequently, there has to be other economic reasons to invest in the cloud. One of them is total cost of ownership. Luckily it is indeed possible to bring down the operating expenses with the cloud, even compared to the traditional on-premise model. The reason is that, even though an IT asset, whether it is hardware or software, is paid for up front, buyers often also have to buy yearly support and an upgrade plan along with it. This is typically in the region of 10% to 20% of the initial purchase price. The price of many cloud services dwarf even that figure. I have been part of projects for public clients that saw reductions in operational costs by 80% to 90% without any capital expenditure—just by switching to the cloud.

Economies of Scale

The last economic driver is economies of scale and this applies to the supply side. The basic idea is that the cost of adding one unit of infrastructure decreases with the number of units already managed. This is a rule that also holds for other types of infrastructure like roads, telephone, gas, etc. The reasons are many, but to put it simply, let’s look at the example of a startup. If they buy 10 servers, they need one full-time employee to manage those servers. The first 10 servers come with a cost of one full-time IT specialist. This person is difficult to recruit because they are already scarce in the market and the company may have to hire a recruiter, which brings additional costs. If something weird happens to the servers, the company may have to hire consultants to fix it.
Now compare this to AWS. If they buy 100 extra servers, it still would not require them to hire an extra IT specialist. They would also likely get a substantial wholesale discount. They definitely don’t need a recruiter because people are lining up to work for them. If something weird happens to one of the servers, AWS likely already has an in-house specialist since the same weird thing happens much more frequently when you run thousands of servers and thus it makes sense to hire a specialist for this particular purpose.

Security

Although this is not the most common driver, it can be a powerful one. Many think of the cloud as insecure compared to on-premise, but recent advances in securing the cloud and changes in risk are changing that conception. If you have a data center that is completely cut off from the surrounding world you may feel like you have more control. But that is rarely if ever the case. Today’s world is by definition connected. Employees may access the company system from home and the company has an Internet site. This leaves the complete system vulnerable for the ever-emerging new threats from around the globe.

Keeping Up with Threats

A company relying on its own data center needs to keep on top of all developing threats and attacks, 24/7. This is costly and difficult. A local IT security department needs to be staffed with specialists. In the cloud, some threats are easier to mitigate. Let’s look at a couple of examples.

The classic distributed denial of service (DDoS) attack is where a number of machines controlled in a so-called bot network start sending multiple requests to a website. This can increase traffic 100 or 1000-fold, which very few websites are built to withstand. The effect is that it is taken offline and customers cannot place orders or get in contact with the company. In the cloud such attacks are easier to mitigate. There are systems that can detect infected machines and scale quite significantly to cope with the increased demand.

Keeping Information Safe

In terms of information security, recent legislation among others from the EU in the form of GDPR requires that companies are in much better control of personally identifiable information (PII). Cloud vendors now offer systems that will run reports and identify places where this information may be stored on internal resources and can also log who accessed the information when. There is also the possibility to automatically detect other vulnerabilities in the system configurations that would otherwise have taken an extensive audit to detect.
Encryption is now offered in the cloud by default, but is harder to set up on-premise. Encryption requires keys and they require adequate management for the system to be secure. Such systems are rarely in place on-premise but are easily available in the cloud and integrated into the services. Identity and Access Management (IAM), which makes sure that only authenticated and authorized users can access system resources and data, is built into the cloud by default, but has to be maintained in parallel on-premise.

As you can see, there are many reasons that security is a driver for the cloud. Vendors have a much better chance at keeping up with emerging threats and developing adequate ways of mitigating them.

Resilience

For most companies, business continuity is a key concern and because business operations are increasingly tied to IT, resilience becomes a key motivation. If an organization needs to make sure that its operations do not cease in the face of a natural or other disaster, they need to ensure the resilience of their IT operations.

The classical natural disasters like flooding, earthquake, hurricane, or volcanic eruption have the potential to completely destroy a data center. In order for this not to happen, a company has to build at least one backup site. This site needs to run exactly the same system setup as the primary one and there has to be automatic failover when the primary site fails. Not only is this expensive, it is also difficult to implement. To remain efficient, the company has to periodically conduct drills where the primary data center is shut down and the secondary center takes over. Many companies have such complicated setups that the drill in itself is far from a trivial risk.

In the cloud, this is in some cases virtually reduced to a few clicks that will give you even better geographical redundancy than was feasible for the company alone. Moreover, you can do this at the individual server level. If you want to make sure an earthquake in the U.S. West data center does not affect continuity, you can just choose to have it failover to a region in the U.S. East. No need to secure the connection and conduct drills. It is not even necessary to have the entire system running. It is possible to specify a template of the virtual machines that should spin up in the event of a server failure. This saves cost and energy for the company.

Another example of how the cloud can be used for resilience is for backups. It is not trivial to run a backup and recovery setup on-premise. Machines have to be added to the network and tested. In the cloud, storage for the backups is a fraction of on-premise costs and can be expanded with a few clicks. Consequently, some customers start with the cloud just for backups.
Scalability

Scalability is one of the most obvious benefits and one that is of particular importance for companies that grow quickly. We already learned that economically it required a lot of upfront payment to secure adequate system resources. But the cost is just one thing, the other is getting it fast enough and at the right locations to scale with demand. Especially in a global context, this can be a challenge.

Websites often see natural spikes in volume. Maybe the company has a sale, or it aired a successful commercial that generated a lot of interest. Maybe something went viral that prompted users to go to the website. Because website responsiveness is a key factor, it is necessary to be close to the customers geographically. For a small startup with global ambitions, it is not feasible to build up data center operations on all continents, but with content delivery networks this is again possible with just a few clicks. This is not just for companies. Cities may also want to make sure that a local disaster like a hurricane does not take down their website because residents want information. This can easily be scaled with the cloud.

Another example is how a traditional data center database administrator has to constantly monitor the storage available to a database and increase it or move it to a different server if it fills up. With cloud-based databases, that is no longer necessary. They will scale by themselves almost infinitely, depending on the particular service. It is also possible to scale up to hundreds of servers automatically if needed. This is why scalability is an important driver for many companies.

Focus

Because the cloud provider takes over much of the management of the IT infrastructure and because managing an IT infrastructure is not the business that most companies find themselves in, it can allow a company to focus its energy on its core competences. Today similarly, most companies don’t want to build their own buildings, roads, and power stations, either because these are rarely aspects that provide a competitive advantage.

In business strategy, much attention is given to identifying and cultivating a company’s competitive advantage. This is something that this company can and wants to do better than their competitors. For Apple, for example, this is the design of user-friendly products, not manufacturing the products. This is why Apple designs their products at its headquarters in California and outsources the production to another company, Foxconn in China, whose competitive advantage is not design, but manufacturing electronics products.
Chapter 2 | Why Cloud?

In much the same way that Apple doesn’t have a competitive advantage in manufacturing their products, most companies do not have any competitive advantage in managing their own IT infrastructure. Using the cloud allows companies to focus on what they do best.

Agility

The importance of the ability to react fast to changing market conditions came into full view during the COVID-19 pandemic, where suddenly cloud solutions came even more into focus. With everyone in quarantine, the easiest way to work was in the cloud due to its universal access.

Elasticity means that consumption not only expands when necessary but also that it contracts when the need is gone again. Otherwise it would be necessary to keep paying for system resources that were no longer needed.

This dynamic is difficult to implement in an on-premise data center since you can’t return the servers or the software licenses once you are done with them. Companies that see great spikes or uneven usage have a particular need for elasticity, since that allows them not to pay continuously for the peak load.

Another important part of agility is that system resources are available on demand when needed. If a particular database is needed, it shouldn’t be necessary first to contact sales at the software company, get a quote, purchase it, and then install it. In the cloud all of this is reduced to searching for it on the cloud provider’s website and then clicking it. Licenses are included in the metered usage of the service.

With this kind of usage, it is also easier to support innovation. If this was just an experiment, a proof of concept, it can be provisioned and used. Once it is no longer needed, it can be deleted and no further payments have to be made. This makes innovation and experimentation much cheaper.

Sustainability

A rarer but emerging driver is that the cloud is more sustainable when it comes to energy consumption. There are different reasons for that. First of all, a cloud provider may ensure a much better utilization of resources.

When you turn on a computer, its CPU runs no matter if it is used or not. Mostly CPUs idle and don’t do anything. At a company where people work from 9 to 5, the computing resources will be used only in the time frame and typically not fully utilized at any point. That means that a traditional server running 24/7 is using three times as much energy as is needed. In the cloud, with elasticity and pooled resources, it is possible to secure a much better utilization since multiple customers can use the same machines. When utilization drops, machines will turn off and come back up again with increased use.
Another reason that the cloud is more sustainable is that it is easier to build efficient energy consumption into larger data centers. Energy for cooling can be reduced by up to 40%. Excess heat can be captured and turned into hot water. Because data centers also take up a considerable area of land, it is possible for them to set up solar cells to harvest energy. For example, Microsoft has vowed to be carbon negative by 2030, which means that they will produce more clean energy than they consume.

**Summary**

As we can see, investments in the cloud are already substantial and they are growing more than twice as fast as investment in other types of information technology. The market is converging toward the cloud as the default model. The reasons for this differ across different types of organizations.

Economically, the ability to shift from up-front commitment of capital to ongoing costs is attractive for private companies. The total cost of ownership will appeal to everyone, but is particularly valuable to public customers seeking to lower their costs, while economies of scale ensures that cloud providers can continuously offer cloud services cheaper as the volume expands.

The cloud also allows for increased control of security and mitigation of risks. Due to the setup of modern cloud platforms, it is possible to increase the resilience of the IT infrastructure of any organization compared to an on-premise setup.

Scaling system resources is attractive and, along with the agility of cloud consumption, it makes it attractive for companies that have uneven utilization patterns or need for technological innovation.

The fact that a lot of capital is freed up and IT infrastructure work becomes obsolete allows companies to focus on their competitive advantage. Most companies do not have a need to run a data center and can therefore focus on their core competences by using the cloud.

A final emerging driver is sustainability, since professional cloud providers are much more likely to manage a data center that efficiently uses energy and may even supply it themselves with wind or solar power.

There are many drivers and different ones will appeal to different organizations, but even though the cloud is already mainstream, it still seems poised for further growth.