Public cloud:
What, why, where, how, who?

Prepared by: Matthew Pettitt, Senior Security Consultant

Table of contents

1. Introduction

1.1 Are public cloud services safe, cost effective & reliable?

Whenever an outage on one of these cloud providers occurs, or a data breach of information held by them, the immediate press coverage starts asking whether they really are as secure and reliable as traditionally managed servers.

According to Gartner [2], “95 per cent of cloud security failures will be the customer’s fault” by 2022, and the cloud providers are all pushing to make it easier to apply security controls, in the hope that they can reduce the number of breaches that occur to data held on their systems.

Are the cloud providers really at fault for customer errors, or are they just a convenient scapegoat for companies trying to avoid bad publicity?

Are public cloud services safe, cost effective and reliable?

2. What are cloud services?

Part of the problem with defining whether cloud services are safe, cost effective and reliable comes from defining what counts as a cloud service. Does, for example, Office 365 count? It's hosted on servers run by someone else, with someone else dealing with technical maintenance and handling things such as server redundancy, but Microsoft class it under their "Business and Productivity" offerings rather than under their "Intelligent Cloud" offerings. Similarly, Gmail could be considered a traditional webmail provider, if used by individuals logging into their personal accounts, or a cloud service, when the same individuals log into a corporate account which is controlled by system administrators at work.

The US National Institute of Standards and Technology (NIST) defines cloud computing [3] as services which meet five characteristics: on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service. The UK National Cyber Security Centre (NCSC), meanwhile, defines [4] it as "A model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort, or service provider interaction."

Bearing these definitions in mind, cloud services are essentially a service which can be provisioned by clients themselves, accessed over a network, where the systems used to provide these services are shared between clients, with the potential to expand and reduce client capacity dynamically, and where these can be measured objectively.

The biggest players in the cloud services space are Amazon AWS, Microsoft Azure, Google Cloud Platform, IBM Cloud, Alibaba Cloud and Oracle Cloud. All of these providers offer a mix of generic services and specialist ones, but the specific markets in which they work vary somewhat.

AWS is by far the largest provider by virtually any metric: it alone holds about one third of the cloud market, operates worldwide, and provides a huge range of services. However, while having a large variety of services available allows for greater flexibility, this can also mean that running a service based on AWS can be more complex than running one on a different provider.

After that, it gets complicated. It is possible to argue that Azure and Google are either second or third in most metrics, although by including virtual private cloud deployments IBM could also claim to be in the top three. Additionally, in terms of specific business types or regions, Alibaba and Oracle both claim to lead in some areas.

There are also other cloud providers who either concentrate on specific regions, specific services, or whose services would previously not have been classed as cloud services. For example, Salesforce is best known as a software service provider, but also owns Heroku, a Platform-as-a-Service provider, which allow apps to be deployed to managed containers. While managed containers cannot always be used for the full range of applications which generic computing options allow, they can be a viable option for specific types of workload. They also introduce a slightly different set of risks to the Infrastructure-as-a-Service solutions, and have very different options for addressing those risks.

Even where the offerings seem to be almost equivalent on paper, there are some distinct differences in how they are actually used. For example, many companies already have some level of investment with Microsoft, so starting to use services like Office 365 can be a logical step. Many of the Azure services are also drop-in replacements for what have traditionally been on-premise systems, such as AD and Exchange. Microsoft have been pushing this, and as a result, many clients are using Azure as part of a hybrid cloud solution, where a mix of local and cloud systems are used.

By contrast, using AWS tends to require taking a more concrete step. It is often used initially for specific projects and then expanded if those projects go well. These tend to be self-contained solutions, meaning that there is more of a distinction between those services which are cloud-based and which remain on-premise.

This can mean that companies end up with multiple cloud providers almost by accident, especially if decisions to use specific technologies are made by teams that previously worked independently, such as productivity tools and web hosting. As a result, several companies have launched management platforms which can integrate with multiple cloud providers. One service worth highlighting, if only because of the provider, is Cloudyn [5]. This is a service provided by a Microsoft subsidiary, which is integrated into Azure but can also draw data from AWS and Google cloud systems. In some ways, this shows a difference in how the cloud providers are approaching competitors. Microsoft has been active in areas (such as database services) with other major players for a very long time, while both Amazon and Google are used to being dominant in their market areas. As a result, Microsoft has accepted that clients may want to use a range of providers, and provide some tools to facilitate this, while the other providers want clients to stick with one provider. This has led to Azure becoming a popular choice for managing cloud services, even where multiple providers are being used.

3. What can you do with cloud services?

There are cloud services for a huge range of purposes, from the obvious applications, such as hosting websites, to more specialist services, such as transcoding media files, or providing desktop environments which are not actually installed on local devices. Pretty much any task which can be performed by either a locally hosted server or a standard computer can be performed through one cloud service or another.

Some of the better-known services are:

- **Scalable virtual machines (VMs)**: Amazon Elastic Compute Cloud (EC2), Azure Virtual Machines and Google Compute Engine, among others, provide dynamically allocated virtual servers that can be used for general purpose computing. Typically, they can be treated similarly to any other Internet connected server, but can also be automatically spun up to cope with increased loads, configured for deployment in different regions to provide protection against network outages and deactivated without significant financial impact.
- **Storage**: Amazon Simple Storage Service (S3), Azure Storage and Google Cloud Storage offer a way to store and retrieve data. Rather than running servers, users can upload arbitrary data, which is then transparently distributed to multiple geographically separate physical devices to ensure that it can later be retrieved, even in the event of catastrophic systems failures. Users can also set access permissions, allowing other systems or users to access specific assets within the system.
- **Database hosting**: While it is possible to run database software on any VM, many providers offer specific managed database hosting as well. These services tend to handle the technical aspects of managing a database server while leaving the data management aspects to the client. For example, Amazon Relational Database Service (RDS) handles hardware configuration, software patching and backup functions, while allowing subscribers to run their choice of database management system (DBMS) from a selection of industry standard options. The Google equivalent offers similar, with either PostgreSQL or MySQL, while Microsoft offer their cloud-optimised Azure SQL Database, based on SQL Server compatibility as well. All these providers also offer specialised database options. Amazon have Aurora (a custom relational database), DynamoDB (a custom non-relational database), Neptune (a custom graph database) and ElastiCache (an in-memory data store). Meanwhile, Microsoft offer Cosmos (a custom multi-modal database), Table Storage (a semi-structured data store), and hosted Redis Cache (in-memory data store). Finally, Google have Cloud Bigtable (NoSQL column database), Cloud Spanner (custom relational database), Cloud Datastore (NoSQL document database), Cloud Memorystore (in-memory data store) and Firebase Realtime (realtime data sync and storage).
- **Networking**: Since all of the cloud providers have global network presence, they tend to offer ways to take advantage of this. These services include virtual networking, allowing connections between physical locations via the provider avoiding the need for dedicated connections between sites, content delivery services, aiming to speed up content delivery to end users, and DNS, ensuring that requests for domains are served quickly and give the optimal results based on the system behaviour.

These are by no means the only services available, even sticking with the big three providers. Some of the other products available include:

- **Serverless computing**: Instead of running a server all the time, serverless computing products allow customers to define processes which will run given a specific trigger, which can either be a direct call of some kind, or a condition which is reached by some other service. AWS Lambda, for example, allows for triggers to be called upon upload of files to an S3 bucket, then for processing of those files to take place, without requiring a client controlled server to monitor for uploads. Other providers similarly offer equivalent services linked to their storage, database and message queue functions.
- **Containerisation**: Running Docker or other containers without needing to manage the outer infrastructure that allows this. Services tend to handle upgrades of the containerisation software, isolation between containerised environments, and some form of automatic scaling.
- **Media processing**: Converting media files from one format to another, processing them prior to storage, or performing analysis against them. Providers offer managed services running on optimised hardware to allow fast processing of data as required. These services tie in closely with storage systems, so, for example, video content can be uploaded in one region, encoded for distribution, and be made available across multiple regions without further action by the customer.
- **Machine learning**: Services which allow for dynamically adjusting behaviour, based on training. Common services include natural language comprehension, aiming to automatically identify whether customer interactions with customer services, for example, are considered positively or negatively by the end users, voice recognition, which is often related to commercial offerings which the providers make (e.g. Alexa, Cortana and Google Now are all driven by the respective cloud machine learning engines), and image recognition.
- **Virtual desktop environments**: These allow for standard desktop applications to be run on remote servers and are commonly used for environments where end users can connect to controlled systems from virtually any device. This means that a higher level of control can be achieved over the work environment, without needing to dedicate a device for the purpose: users could connect to their work desktop from a personal laptop, yet not be able to transfer data to or from the work environment without further authorisation. They can also be used for cloud-based gaming, meaning that users who can't justify high end graphics hardware for gaming can rent an appropriate system and access it from a lower specification device. There are companies which specialise in both business use and hobby use of virtual desktop environments.

This wide range of services means that there are often multiple ways to achieve any given aim, and hence there may be widely varying costs, reliability impacts and security risks, even if the intended behaviour of systems is very similar.

For example, even a basic workload, such as hosting a website, can be achieved in multiple ways. Sticking with services available through AWS (although a similar range of options is available through other providers), some options could be:

- Host files on S3, expose through making the bucket publicly available.
- Run a traditional web server on an EC2 instance, as with a standard dedicated server.
- Run a web server in a Docker container, exposing only the required ports.
- Set up a serverless trigger in Lambda to output HTML when a page request is made.

- Use a pre-built CMS running on Lightsail, a simplified virtual server option.

In each case, the visible output to the end user would essentially be the same: an HTML page. However, both the costs for the single page load, and the overall costs for page loads over a month could vary massively. Similarly, some of these methods would be more robust than others, and the attack surface for each method would vary, both from the components mentioned above and from the security configuration of other steps in the process.

4. Security risks of cloud services

When considering the examples from the previous section, the impact of a security breach in each case could well be different too: the main risk could be to the site visitors, to the back end data, to other systems on the Internet, or even to the finances of the server user. It is unlikely that an attack against S3 or other block storage systems, for example, would result in an attacker being able to use the service to launch attacks against other sites. While the underlying servers would have this ability, the interfaces exposed through S3 deal with static data, which has to be requested by a remote service. There is no method to trigger a connection from an S3 bucket to a third party server. However, this would allow an attacker who could write to the bucket to use it as a storage location for malicious file content loaded by other compromised sites, which may have a financial impact on the legitimate subscriber and potentially cause problems for users of other unrelated sites.

By contrast, compromising an EC2 instance holding static files would allow additional attacks against the legitimate subscriber's AWS infrastructure and, depending on how this has been configured, potentially against non-AWS based infrastructure too. The compromised instance could also be used to launch Denial-of-service (DoS) attacks against third parties, or simply as a cryptocoin mining node, slowing down the legitimate activities it is being used for.

The responsibility for the security settings also varies. In most cases there are firewall levels limiting access to specific ports or restricting access to potentially dangerous services to known source IP addresses. However, with S3, Amazon only expose access controls for objects and buckets – users don't have to worry about the configuration of the underlying server. With EC2, if the user misconfigures a service running on an exposed port, firewalls won't help – the security is equivalent to a standalone server managed by the user.

The specific division of responsibility can vary with the type of service. Software-as-a-Service (SaaS) offerings tend to be mostly under the control of the provider, with the client only responsible for ensuring that data stored on the system is appropriately classified, and that access to accounts is appropriately limited. At the other end of the scale, on-premise cloud deployments can require the client to handle all aspects of securing the environment, from physical security to installing patches supplied by the vendor. Microsoft [5], Google [7] and Amazon [8] have documents detailing this division of responsibility, covering which patches are the responsibility of the provider and which the client, for example. It is important to understand both of these requirements, and to ensure that those responsibilities which belong to the client are actually performed. Many large companies struggle with applying security patches when released on traditionally hosted systems and while, in some cases, cloud providers take some of the burden, patches to deployed software on virtual machines generally remain the responsibility of the client.

This can also lead to issues which are difficult to diagnose, such as slowdowns that are due to specific patches being applied at the provider level, yet which affect client services. While a company could choose not to deploy patches with potential risks, it is unlikely that a cloud provider, with a much larger potential attack surface, can take the risk of not applying security fixes, even if there are performance impacts.

One of the key challenges unique to keeping cloud services secure is tracking precisely what services are being used. Since it is possible in many cases to launch additional services from code, or in response to pre-defined criteria being triggered, it is important to ensure that services started in this way are secured appropriately. Where security processes have been developed based on requirements for multiple stages of approval to deploy new systems, or on a limited number of systems being used, these can fall down when applied to dynamically changing infrastructures.

Another risk which is unique to cloud providers is due to the shared nature of the environment. If IP ranges relating to the cloud service are blocked from specific groups of users this can result in other, unrelated, services being blocked as well. A high profile instance of this was in April 2018, when Russia blocked a range of IP addresses used by the Telegram messaging service. These IP addresses corresponded to AWS and Google Cloud instances controlled by Telegram, but also resulted in other sites hosted on these platforms becoming inaccessible within Russia [9].

5. Instances of security breaches

There have been many breaches of cloud services and while the exact causes of the issues that allowed them to happen are not always publicised, it is often possible to make informed guesses about what happened. In other cases, the root problem can easily be identified, although the specific reason for that problem can remain unclear.

The complexity of the flaws can vary massively too: some of the biggest breaches resulted from trivial flaws. For instance, details of nearly 200 million American voters collated by the Republican National Committee (RNC) were stored on an S3 bucket which had no access restrictions in place [10]. Where a bucket is set to allow public access, visiting the URL <bucketname>.<region>.amazonaws.com will return an XML file containing a list of the bucket contents. <bucketname> is a user selected string, while the <region> is one of a known list of AWS regions which support S3 services. In this case, the bucket was named “dra-dw”, and the file names of the contents suggested that this stood for “Deep Root Analytics Data Warehouse” – pointing to the company and the purpose of the store. This one S3 bucket had about 25 terabytes (TB) of data held in it, with about 1.1 TB of this fully accessible to anyone who found the bucket name.

Configuring an S3 bucket to full public access on creation is down to choosing a non-default value in a drop down: switching the “Manage public permissions” setting from “Do not grant public read access to this bucket (Recommended)” to “Grant public read access to this bucket”. It is hard to see any potential confusion from the behaviour of this setting – and the setting for individual objects is similarly labelled. Even when making an object publicly accessible at a later stage, using the S3 management console, any attempt will result in a prominent warning message, explaining what public users will be able to do. While some of these warnings may have been added as a result of well publicised issues, there are still plenty of buckets being created which are publicly accessible – using a tool [11] which passively monitors for bucket creation over a five minute period resulted in three publicly accessible buckets being identified, and that was only looking for fully accessible buckets created during that period. In other words, any buckets which had been previously created and then made public, or which were configured to allow access to specific files were not detected.

However, it’s entirely possible that the exposed RNC data was made public not by someone choosing options in the AWS console, but through programmatic errors. Most cloud services are fully scriptable, meaning that it is possible to set permissions and even start new services by sending messages from code running elsewhere, assuming you have the appropriate permissions. Given this, a minor typo in a script could result in files being made public, rather than permissions being given to another AWS account, for instance.

The RNC are not the only group to have suffered from this kind of misconfiguration [12]: phone provider Verizon [13] exposed both customer data (through a third party) and proprietary technology information (directly), the WWE [14] wrestling entertainment company exposed customer data and defence contractor Booz Allen Hamilton [15] exposed remote login keys and credentials for military systems, as well as imagery from battlefields. All of these were from 2017 and were fairly widely reported due to involving well known or potentially sensitive companies. However, the same issues are still turning up in 2018 – data relating to a proposed infrastructure on AWS for use by GoDaddy was exposed in August [16], for one example, and “terabytes of data” gathered by a company marketing phone spyware [17], including selfies and text messages, for another.

Although many of the highest profile breaches have been using this service, misconfiguration of cloud services is not limited to S3. Google’s Firebase service for example is a popular choice for developing mobile applications and includes a database system. Researchers scanning mobile apps found about 2,250 misconfigured database systems, allowing access to anyone who found the URL [18]. These databases contained a variety of data, ranging from plaintext passwords to health related discussions with medical services.

Another common flaw is a failure to limit access to administrative access services: Secure Shell (SSH) and Remote Desktop Protocol (RDP) in particular. Using Shodan, a database of online devices showing which ports are open on them, over 100,000 systems within the Azure IP address ranges were found to have port 3389, used for RDP connections, open. While some of these may have legitimate uses, it is important to note that Shodan is unlikely to be a legitimate source of administrative actions for all these systems. A similar search on port 22, used for SSH, on AWS IP ranges revealed nearly 1.5 million systems exposing it, and nearly 700,000 systems on Google’s cloud ranges. While it is possible that these systems only allow access using key based authentication, which would decrease the risk of compromise, it is possible that there are flaws in SSH server implementations exposed. For example, CVE-2016-6615 can cause denial of service to OpenSSH versions below 7.3 by forcible attempting login with a very long password – that would apply to over 100,000 of the Google-based systems exposing port 22.

This kind of issue has been suggested as contributing to data breaches such as that suffered by Deloitte [19], [20] in November 2016, where data stored on Azure servers was stolen. According to reports, this was possible since administrative accounts were accessible from anywhere and did not require multi-factor authentication.

One flaw which is clearly out of the hands of cloud providers, although which they have generally been responsive to, is the use of API keys which have been published through services like GitHub. API keys can be used in place of a username/password combination to perform actions within cloud management services – they’re usually intended to allow custom code to access other services, without needing a password to be included in the code. In most cases, it is possible to limit precisely what each API key allows access to, but where keys with permissions to trigger the launch of services are found, they can be used to spin up instances on cloud providers, which can then be used by the attackers. One common use for such servers is for mining cryptocoins. This takes a lot of computing resources, but doesn’t tend to require a lot of ongoing access, meaning that if a server owner spots the problem and terminates the instance, only a limited amount of effort is lost. This type of mistake is common: Tesla [21], OlinData [22], and DXC Technologies [23] are among companies who got high bills from AWS due to keys exposed in published code repositories or in unsecured consoles for other services.

6. Precautions

Most of the precautions for keeping cloud services secure are simply extensions of steps that should be taken to keep self-managed services secure, although, as with self-managed services, not all users of cloud services take these steps.

One key step in keeping cloud deployments secure is protecting the root or super administrator account from compromise. This account has the permissions to assign roles to other accounts, make changes to billing limits, and to adjust security permissions for all services available through the provider. As a result, it is recommended to avoid using this account for purposes that can be delegated to lower privileged accounts once the cloud infrastructure has been set up. Examples of tasks that can't be delegated are closing an account with a provider, changing passwords for administrator accounts without the existing password, requesting authorisation for penetration testing of the cloud deployment and making changes to technical support plans [24].

For AWS, documentation relating to the root user regularly states things such as "We strongly recommend that you do not use the root user for your everyday tasks, even the administrative ones." Azure and Office 365 suggest users to "Create dedicated global administrator accounts with very strong passwords and use them only when necessary." In most cases, a Global Administrator created in either Azure or Office 365 will have the same permissions in the other service, if the organisation uses both. Google takes a slightly different approach, in that a standard Google account (such as those created when signing up for Gmail or through a corporate GSuite subscription) is used for access, but is initially given only the privileges required to assign permissions to perform other activities.

In all cases, the providers recommend the use of multi-factor authentication for all administrator level accounts, ideally using some form of virtual token generator running on a mobile device, and it is generally possible to enforce this, as well to as set up alerts if these settings are modified.

It is also possible to configure role-based access control (RBAC) for non-administrative functions in many cases. Using this allows for test engineers to access only test environments, which may make use of multiple individual products from the provider, while preventing access to production environments, for example. Since the RBAC system is tightly integrated with deployment tools this can extend to allowing some roles to create additional instances, without needing additional technical approval, with the knowledge that all instances created by test engineers will only be accessible from within the business, avoiding the worry of test systems being exposed to the wider Internet. Obviously, there would be a cost implication to allowing arbitrary instance creation, but it is often possible to limit the amount which can be spent by members of a given role too.

All cloud providers offer tools which are designed to help with dynamic deployment of systems: security groups, so new systems can be assigned to groups with specific permissions, IAM configuration, so access rules can be automatically applied to new systems, and budget alerts, so users are warned when the cost of service they are using is likely to exceed a pre-defined limit. However, it is important to ensure that these are configured correctly and that access to these tools is restricted to an appropriate set of users – there is no point in allowing all users to make changes to access control settings, for example.

Patch management is important in keeping software running on virtual machines secure, and both traditional methods and cloud-specific methods can be used to help with this. For Linux servers on any of the cloud providers, tools like apt and yum can still run on a scheduled basis, pulling updated packages from the vendor and automatically applying security fixes, although in most cases, there are local mirrors of the package repositories available on the cloud provider network. Alternatively, there are various patch management systems offered by the cloud providers, which allow clients to specify a minimum patch level for systems and then attempt to keep all configured systems up to that level, alerting the client if there are problems, such as incompatibilities with other installed software. Similarly, Windows servers can still use tools like Windows Server Update Services (WSUS) and Windows Update, and are often integrated with patch management tools too.

Beyond the cloud-specific points, most advice about securing self-managed systems still applies. Making use of firewalls to restrict access to systems and ports is still important, although in many cases the rules should be applied to groups of instances, rather than individual instances, and creation rules configured to place new instances in a highly restricted group if not otherwise specified.

System logs are still important and most services will automatically log actions that are controlled by the provider to a centralised log. It is sensible to link any system-level logs to centralised logging, since this can help with detecting issues in instances deployed through automatic scaling processes. This is mostly a concern with virtual server products, where the actual software is managed by the client, rather than with managed services, which will tend to default to centralised logging. For example, a MySQL database running on a Compute Engine instance will log locally, unless configured specifically, while one running on Cloud SQL will log centrally. Since the client can run any software on a virtual server, the specifics of configuring logging to an alternative location will vary.

It may also be important to consider the access of the cloud provider to the data. While the providers have physical access to the systems, and hence should be considered as technically able to access the content, most providers will ensure that all access to customer systems is logged in a customer accessible manner. In some businesses, such as healthcare, it may be important to monitor all access from provider staff.

7. Reliability

The big three cloud providers are generally very reliable when each is taken as a single unit. It is very rare that any of them are completely inaccessible, meaning that for large customers who have applications distributed across multiple zones and regions they can offer uptime statistics of well over 99 per cent. In the 30 days prior to writing, only Azure had a failure which impacted a whole region, lasting 14 minutes, giving a 99.97 per cent uptime across the period [25].

However, this hides issues which affect a subset of services or users. For example, Google Compute Engine has had nine outages in the year to date [26], with a total of nearly 40 hours of disruption; although individual clients are unlikely to have been affected by all of these instances, which each affected a specific subset of systems. Similarly, users of AWS S3 had issues due to power loss incidents [27], but only in a single region – nevertheless, this caused disruption to a wide range of services which relied on this.

As a result, it is very difficult to compare the potential impact on a non-cloud infrastructure with a cloud one. For a non-cloud based deployment, the client is responsible for things like ensuring failover works properly, ensuring that there are no single points of failure, and, where running physical infrastructure, even things like ensuring that there are multiple independent power supplies to each server. With a cloud service, many of these are the responsibility of the provider. They let clients make choices such as running systems in multiple availability zones, but the technical details of what happens when one goes down are abstracted away.

Similarly, the clients do not need to worry about power issues, cabling problems, or physical intrusions, but these are still concerns that could affect the services. If the cloud provider makes a mistake, the client service is still affected. These mistakes do happen: Microsoft had an outage in September 2017 due to an accident involving fire suppressant systems [28], while Google had an outage in June 2018 which resulted from new VM instances being assigned duplicate IP addresses [29] by a Google controlled system, just a couple of weeks after AWS had a combination of power issues and networking problems affecting two zones at the same time [30].

In each case, there was little that clients whose systems were affected could do: if they had previously configured failover systems in other regions, these would take over, but for any business which had relied on a single region they just had to wait for the services to be restored.

These issues are still possible with self-managed servers hosted in non-cloud provider data centres, and do happen, but the number of systems affected tends to be much smaller. This comes both from the scale of the involved systems, and in differences in how the data centres are managed. A traditional server provider can easily identify exactly what clients would be affected by any given server having problems, while a cloud provider has a much harder job, since the customers affected by a given server could vary from minute to minute.

It is possible for a non-cloud service to have similar levels of uptime as a cloud provider, but the amount of effort required to maintain this means it tends to only be possible for very large companies who are able to commit large amounts of time and money to the problem. With the scale of cloud providers, the sheer number of people able to work on finding improved solutions to potential problems spread across the massive numbers of clients means that the cost is distributed across a much larger group, with the benefits feeding back to all clients.

8. Costs

Any discussion of specific costs relating to cloud services is likely to become outdated within days of writing – all of the big three providers regularly adjust prices as hardware changes and specific service offerings change. In general, prices for cloud services tend to either drop or remain stable, with very few services increasing in cost once they have become chargeable. It is typical for newly introduced services to be free or very low cost for an introductory period, with the caveat that there may be issues relating to these – in some cases the providers recommend against using these services for production workloads, while in others they suggest ensuring a fall-back position is maintained, using better established options.

In many cases it is also possible to obtain lower prices by making a commitment to use a specific level of resource for a longer period, usually between one and three years. This involves paying either an upfront commitment fee or a monthly subscription, which usually applies whether or not services are used within the applicable period, then paying a lower hourly rate for services which are used. For stable workloads, these agreements can save large amounts over paying for on-demand instances, but are less useful where demand is less predictable, and providers vary in flexibility should requirements change.

At the other extreme, it is also possible to obtain lower prices by being flexible as to when workloads run. By using processing capacity which is not currently required by standard priority instances, with the trade-off that the cloud provider can interrupt the job with a minimal level of notice, it is possible to get anything up to 90 per cent reductions in costs for instances. This is ideal for non-time critical processing – things like 3D rendering, where tasks can be queued and processed at low cost periods, on the basis that they will probably be completed by the time they are required, and if they do become time critical, it would be possible to move the remaining tasks to higher cost on-demand instances.

When calculating costs for cloud services, it is important to take all chargeable elements into account. For example, while inbound data traffic tends to be free, outbound traffic is often chargeable, with different rates applying depending on where the data is going. Transfers between availability zones are usually charged at a different rate to transfers going to the wider Internet, and even data transferred between clients on the same cloud provider may be treated as external for the purposes of billing. Similarly, storage costs tend be ongoing, rather than a one off purchase as with a traditional server, where, once a hard drive is installed, it can be used with no additional costs until full.

It is also important to ensure that the appropriate services are being used for the workloads. A relatively simple example of this is in running a database management system (DBMS): there can be a significant price difference between a managed database (DB) service, such as Azure Database for MySQL or Amazon RDS, and running the same DBMS package on a generic VM, and the difference can be in favour of either option depending on the specific requirements. As with reliability, however, the amount of effort required from the client varies between the options too – managing database backups is included in the managed DB packages, but is down to the client in a self-managed VM install.

A slightly more subtle case of this occurs in long term storage solutions, for things like archiving records which must be kept for regulatory purposes. While it is possible to use standard storage options from each provider, they also offer long term storage systems, where in return for additional latency in being able to access the data, ongoing storage costs can be massively reduced. The data transfer costs between the different types of storage can also vary, however, so it can be important to carefully define what the access requirements for each type of data are before choosing a solution.

Another cost which can be ignored in some discussions of cloud services is staffing. While cloud services can help reduce the need for some types of engineer to be available full time, they do not remove the need for technical staff to be part of client businesses. For one example, Database Administrators (DBAs) tend to mix management of the underlying DBMS and management of the structure of the data held within it. In managed database services, the underlying DBMS tends to be the responsibility of the provider, but the structure of data remains the responsibility of the client, and this can be key to obtaining expected performance from systems, as well as keeping costs down. By making intelligent choices about how to structure data, fewer database servers can be used to provide the same performance as more servers which have poorly structured data on. However, the appropriate structure will depend heavily on the specific workflow that the system requires – patterns of reads and writes, and the type of data being stored – hence needs to be handled by a DBA familiar with the requirements.

When there is a requirement to have data held in specific regions, this can also affect costs. For instance, running systems in Australia tends to be substantially more expensive than running them in the USA, with systems in South America costing even more. It is sometimes possible to mitigate this to some extent through the use of content delivery network (CDN) services, or by running computationally heavy tasks in other regions, but where this is not an option, the headline costs for cloud deployments may bear little resemblance to actual costs.

There can also be costs which are accidentally incurred, either through maintenance mistakes, programmer error, or malicious activity against the account. Services which are started but then not removed, or data which is no longer required, but has been left in storage can both result in unexpected charges. As mentioned in the instances of security breaches section, attackers have been known to launch large numbers of high power instances and then use these for cryptocoin mining, or as jump off points for further attacks against other targets. By timing attacks carefully, attackers can aim to get the maximum run time from the hijacked account before the legitimate owner notices: for example, small businesses may not have 24 hour monitoring of systems, so launching a number of instances on a Friday evening could give until Monday morning before anyone notices there is a problem. This is less of a problem for larger businesses, as long as cloud services are integrated into monitoring systems, and the staff working at those times have the appropriate access and authority to be able to take action.

9. Conclusion

One of the key points about cloud security is that it is not about handing over security responsibilities to a cloud provider. Even if a company is only using SaaS offerings they will still have responsibility for the data entered into the system and to ensure that access groups and user controls are applied correctly. Additionally, users of Infrastructure-as-a-Service (IaaS) offerings will retain responsibility for almost all the same controls as would be required for a traditionally hosted offering. These requirements mean that skilled and trusted staff who understand both the business and the cloud provider offerings can be key to having a successful and secure cloud deployment.

It is also important that staff are able to keep up to date with changes in the available offerings. For one historical example, prior to October 2009 AWS clients could run MySQL on EC2 instances, where they were responsible for security patches, while the release of Amazon RDS meant that it was possible to move to a managed database service, where AWS took on the responsibility for applying patches. More recently, Google released Cloud Armor [31], a service protecting against DoS attacks, with the ability to implement custom rules to protect services. While this works in a default state for some systems, ensuring that appropriate protection has been applied to custom services is important, and could easily be missed if staff do not keep up with service release news.

All of the major players offer some form of certification showing competency in using and administering their services, usually with a security specialism available. As a vendor specific service, these can include details of specific tools available from the providers, which can be valuable if companies are using a single provider, but may be of less use where multiple providers are being used for redundancy or lock-in avoidance reasons. There are also non-vendor specific certifications, such as Certified Cloud Security Professional (CCSP) from (ISC)², or Cloud+ from CompTIA. These tend to focus more on the principles of keeping systems secure, which can then be used on any provider, but won’t always show the most appropriate method of achieving a given result on all providers, if only because the services themselves change regularly.

The requirements for knowledgeable staff whose knowledge is kept up to date can easily be forgotten in the drive to move systems to cloud providers, especially with all providers promising quick and easy methods to start using their platforms. While it is possible to deploy an application in ten minutes, this hides the complexity of interactions between systems on the same provider infrastructure, such as between managed database services and services using these.

These requirements have a cost implication, which is not always included in billing calculators offered by cloud providers, and while it is likely that a smaller team of staff may be required, these staff may require more training than those supporting a less fluid infrastructure. Additionally, they may require higher levels of authority to be able to escalate problems which are the responsibility of the provider to appropriate levels, with a commensurate impact on expected pay.

Where businesses have already moved some systems to cloud providers, ensuring that the appropriate tools are being used can help ensure that security is maximised, that costs are minimised, and, as with all security, this should be considered an ongoing activity. Where businesses are wanting to move systems to cloud services, care should be taken to scope solutions in detail, including how testing and development activities will be performed. For new systems being built from scratch to use cloud services, or for new businesses, the potential to scale out easily makes cloud services very attractive, but care should be taken that security options are properly used, to minimise the risks of unexpected bills or massive reputational damage in the event of data being obtained by hackers.

Finally, there are some cases where cloud services are not appropriate. Systems which have to run without internet connections, or with limited connectivity, tend to be a poor fit. This can apply to both cases where the specific application must run without connections elsewhere, or where support systems must be available: while it is obvious that a cloud x-ray machine would be of limited use, there may be sensible arguments for keeping servers allowing storage of x-ray images locally too. Where there are systems which must be kept locally as well as a desire to move others to the cloud, ensuring that support for the local systems is available may involve providing additional training to existing technical staff, enabling them to support both these systems and any new cloud deployments.

About NCC Group

NCC Group is a global expert in cyber security and risk mitigation, working with businesses to protect their brand, value and reputation against the ever-evolving threat landscape.

With our knowledge, experience and global footprint, we are best placed to help businesses identify, assess, mitigate & respond to the risks they face.

We are passionate about making the Internet safer and revolutionising the way in which organisations think about cyber security.

Headquartered in Manchester, UK, with over 35 offices across the world, NCC Group employs more than 2,000 people and is a trusted advisor to 15,000 clients worldwide.

References

[1] https://www.forbes.com/sites/johnkoetsier/2018/04/30/cloud-revenue-2020-amazons-aws-44b-microsoft-azures-19b-google-cloud-platform-17b/#4b0cfc617ee5

[2] https://www.gartner.com/smarterwithgartner/is-the-cloud-secure/

[3] https://csrc.nist.gov/publications/detail/sp/800-145/final

[4] https://www.ncsc.gov.uk/guidance/cloud-security-standards-and-definitions

[5] https://docs.microsoft.com/en-us/azure/cost-management/

[6] https://gallery.technet.microsoft.com/Shared-Responsibilities-81d0ff91

[7] https://cloud.google.com/security/overview/

[8] https://aws.amazon.com/compliance/shared-responsibility-model/

[9] https://www.bbc.co.uk/news/technology-43797176

[10] https://www.upguard.com/breaches/the-rnc-files

[11] https://github.com/eth0izzle/bucket-stream

[12] https://businessinsights.bitdefender.com/worst-amazon-breaches

[13] https://www.infosecurity-magazine.com/news/verizon-hit-by-another-amazon-s3/

[14] https://www.forbes.com/sites/thomasbrewster/2017/07/06/massive-wwe-leak-exposes-3-million-wrestling-fans-addresses-ethnicities-and-more/#5f4e9e9b75dd

[15] https://www.cyberscoop.com/booz-allen-hamilion-amazon-s3-chris-vickery/

[16] https://www.engadget.com/2018/08/09/amazon-aws-error-exposes-31-000-godaddy-servers/

[17] https://motherboard.vice.com/en_us/article/9kmj4v/spyware-company-spyfone-terabytes-data-exposed-online-leak

[18] https://www.bleepingcomputer.com/news/security/thousands-of-apps-leak-sensitive-data-via-misconfigured-firebase-backends/

[19] https://www.theregister.co.uk/2017/09/25/deloitte_email_breach/

[20] https://www.theguardian.com/business/2017/sep/25/deloitte-hit-by-cyber-attack-revealing-clients-secret-emails

[21] http://fortune.com/2018/02/20/tesla-hack-amazon-cloud-cryptocurrency-mining/

[22] https://www.olindata.com/en/blog/2017/04/spending-100k-usd-45-days-amazon-web-services

[23] https://www.theregister.co.uk/2017/11/14/dxc_github_aws_keys_leaked/

[24] https://docs.aws.amazon.com/general/latest/gr/aws_tasks-that-require-root.html

[25] https://cloudharmony.com/status-of-compute-and-storage-group-provider

[26] https://status.cloud.google.com/summary

[27] https://www.computerweekly.com/news/252436193/AWS-outage-Datacentre-power-cut-knocks-out-hundreds-of-internet-services

[28]
https://www.theregister.co.uk/2017/10/03/faulty_fire_systems_take_down_azure_across_northern_europe/

[29] https://status.cloud.google.com/incident/compute/18005#18005007

[30] http://www.datacenterdynamics.com/content-tracks/security-risk/power-event-at-aws-data-center-disrupts-us-east-1/100213.fullarticle

[31] https://cloud.google.com/armor/