BITS Pilani

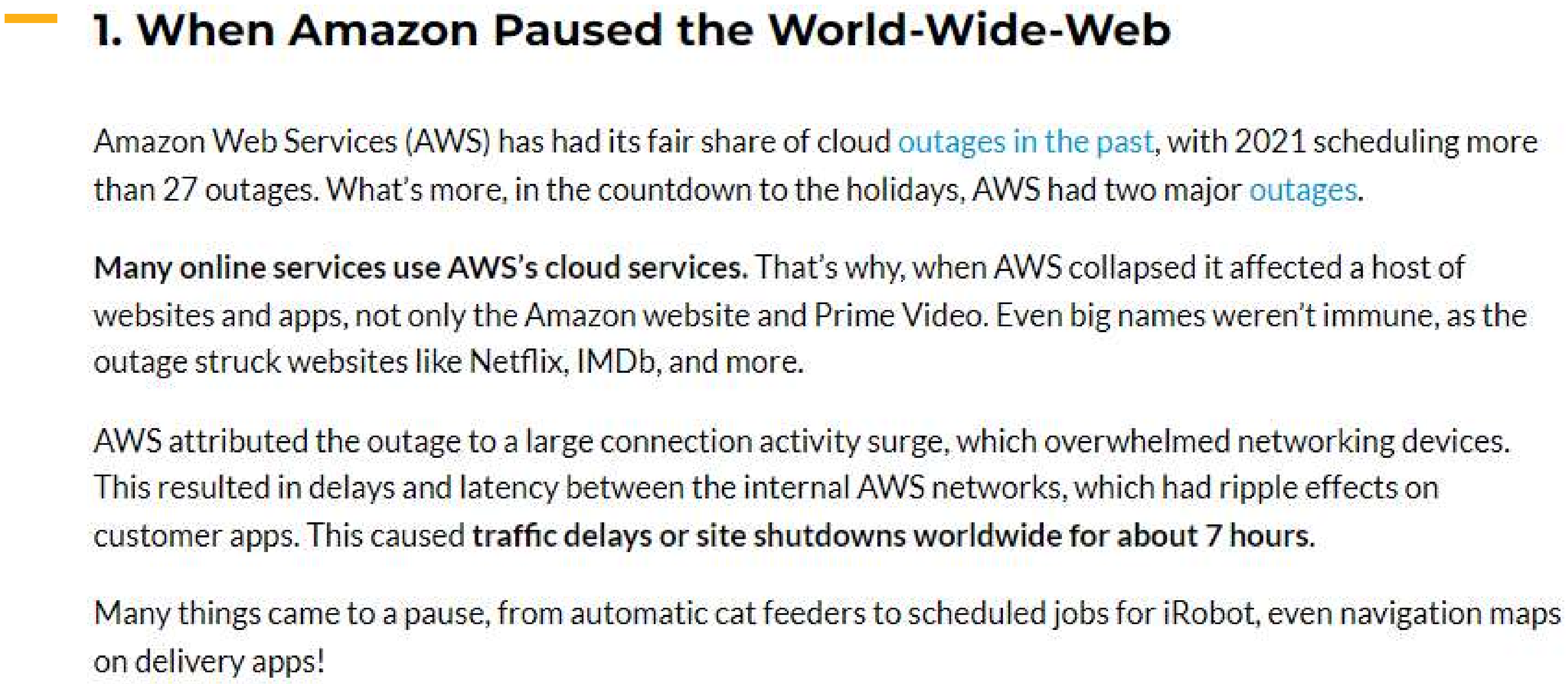
- High Availability
- Multi-tenancy
- Security

High Availability

Requirements in Cloud

- What is meant by High Availability?

- Cases impacting system availability
  - Service outage
  - Service degrade



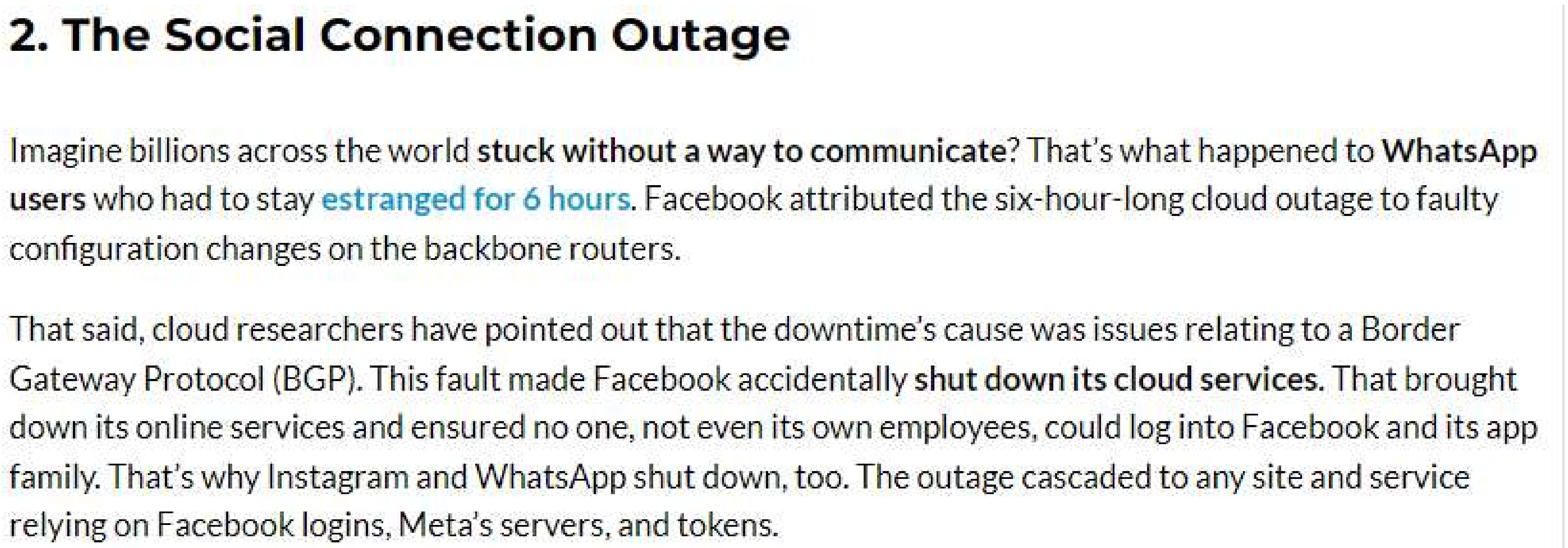
1. When Amazon Paused the World-Wide-Web

Amazon Web Services (AWS) has had its fair share of cloud outages in the past, with 2021 scheduling more than 27 outages. What's more, in the countdown to the holidays, AWS had two major outages.

Many online services use AWS's cloud services. That's why, when AWS collapsed it affected a host of websites and apps, not only the Amazon website and Prime Video. Even big names weren't immune, as the outage struck websites like Netflix, IMDb, and more.

AWS attributed the outage to a large connection activity surge, which overwhelmed networking devices. This resulted in delays and latency between the internal AWS networks, which had ripple effects on customer apps. This caused **traffic delays or site shutdowns worldwide for about 7 hours.**

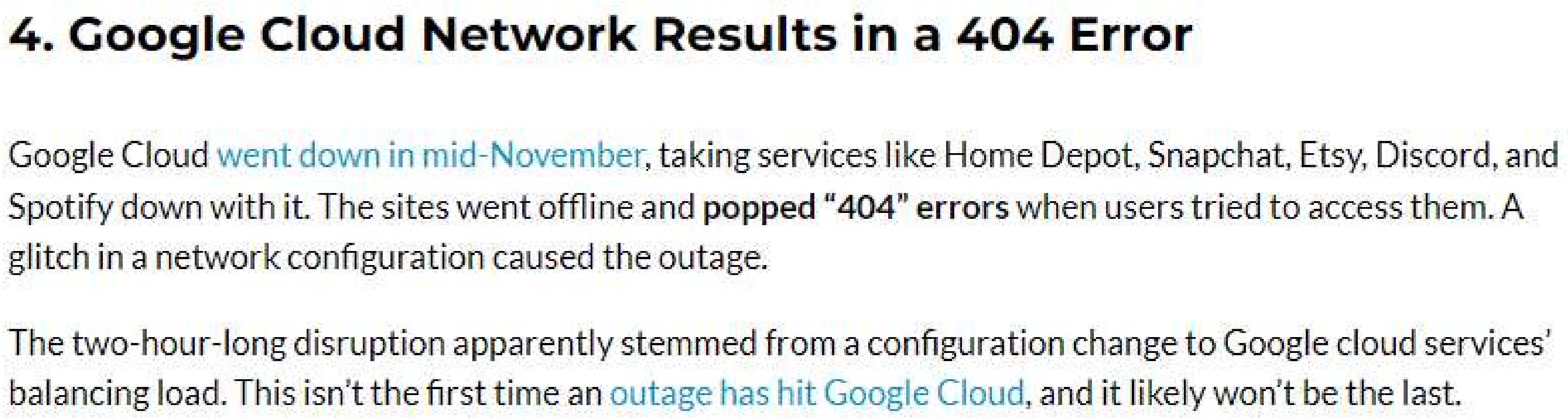
Many things came to a pause, from automatic cat feeders to scheduled jobs for iRobot, even navigation maps on delivery apps!



2. The Social Connection Outage

Imagine billions across the world **stuck without a way to communicate**? That's what happened to **WhatsApp users** who had to stay estranged for 6 hours. Facebook attributed the six-hour-long cloud outage to faulty configuration changes on the backbone routers.

That said, cloud researchers have pointed out that the downtime's cause was issues relating to a Border Gateway Protocol (BGP). This fault made Facebook accidentally **shut down its cloud services.** That brought down its online services and ensured no one, not even its own employees, could log into Facebook and its app family. That's why Instagram and WhatsApp shut down, too. The outage cascaded to any site and service relying on Facebook logins, Meta's servers, and tokens.



4. Google Cloud Network Results in a 404 Error

Google Cloud went down in mid-November, taking services like Home Depot, Snapchat, Etsy, Discord, and Spotify down with it. The sites went offline and **popped "404" errors** when users tried to access them. A glitch in a network configuration caused the outage.

The two-hour-long disruption apparently stemmed from a configuration change to Google cloud services' balancing load. This isn't the first time an outage has hit Google Cloud, and it likely won't be the last.

High Availability requirements in Cloud

High Availability SLA in Cloud Environment:

- Consolidation
- Sharing same infra
- Infra downtime
- Different SLA

High Availability requirements in Cloud

Architect a High Availability Cloud Infra

- How to design HA infra for Cloud?
  - Reduce unplanned outage
  - Configuration best practices
  - System migration

High Availability

Steps to achieve HA

- Build for Server failure
- Build for Zone failure
- Build for Cloud failure

- Automation

Key aspects of SLA

Web-application deploymen - **performance of the application**

Provisioning in those days involved

- Deciding hardware configuration
- Determining the number of physical machines
- Acquiring them upfront so that the overall business objectives could be achieved.

The web applications were **hosted** on

- Dedicated individual servers
- Or Within enterprises' own server rooms
-

Key aspects of SLA

- Service-level objectives

- Capacity planning

Key aspects of SLA

- **Complexity of managing the huge Data centres**
- **Legal agreement with the infrastructure service**
- **QoS parameters**

- Service-level agreement (SLA)

Key aspects of SLA

- Core time vs Non-Core time
- Issue response time
- Infrastructure SLAs
- **Application Service Providers (ASPs)**

Key aspects of SLA

- Problems associated with ASPs
  - Massive redundancies
  - Co-hosting applications

Key aspects of SLA

- Security guarantees
- Performance isolation

- Virtualization technologies

Key aspects of SLA

- Adoption of virtualization technologies
- **ASPs can allocate system resources more efficiently to these applications on-demand**
- Application SLAs
- **Managed Service Providers (MSP)**

- **Cloud Platform**

Key aspects of SLA

Key Components of a Service-Level Agreement

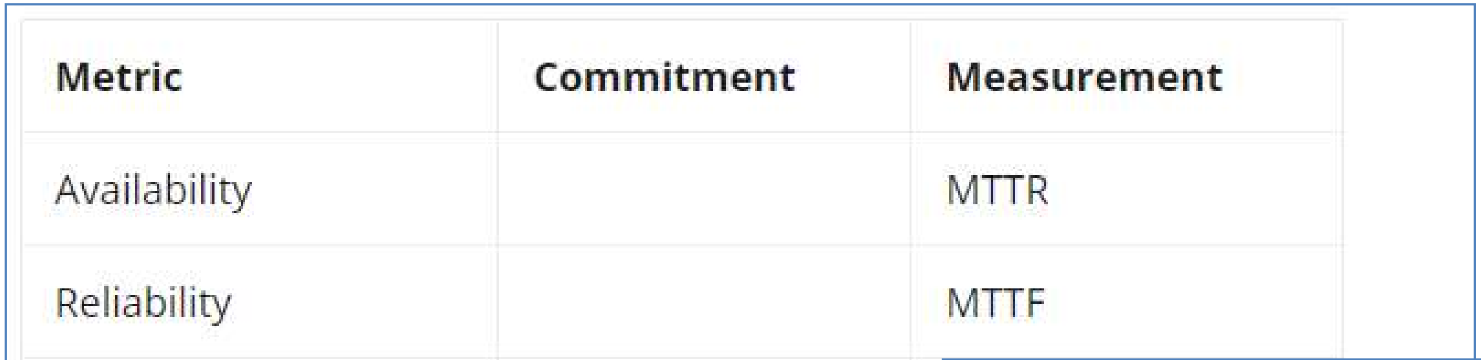
- **Service Level Parameter**
- **Metrics**
- **Function**
- **Measurement directives**

Key aspects of SLA

What Are SLA Metrics?

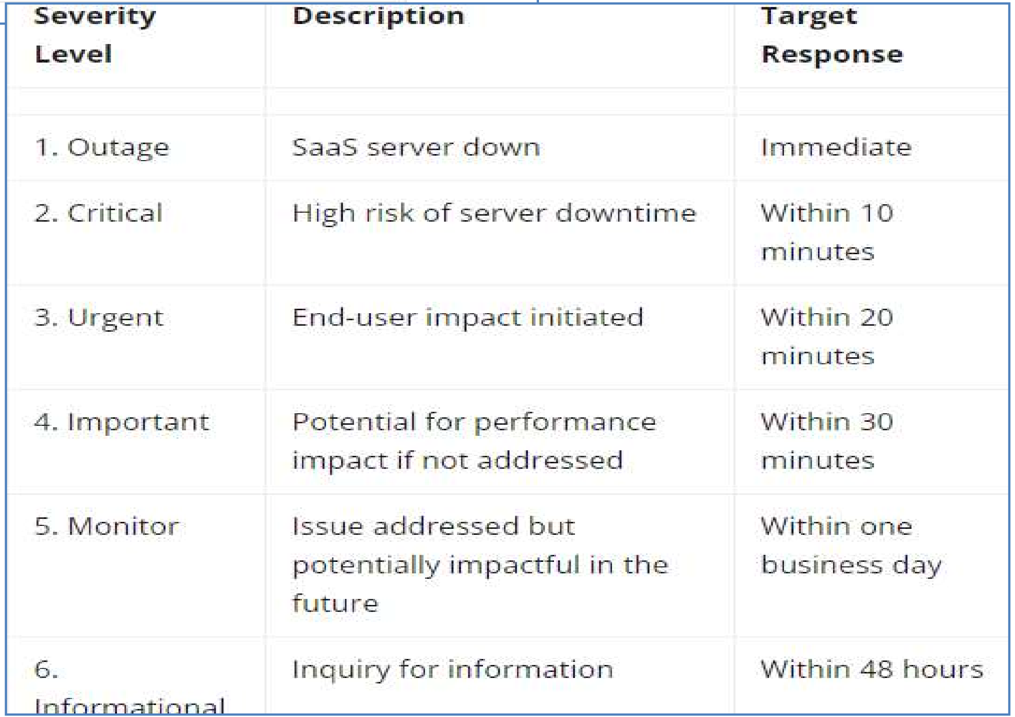
- Availability
  - Uptime
  - Service Availability
- Response Time
  - MTTR
  - Transaction Response Time
- Throughput
  - Disk Write Bytes
  - Link Throughput
- Errors
  - HTTP Errors
  - Disk Read Errors
- Utilization
  - Disk Utilization
  - Memory Utilization

KPIs and Metrics



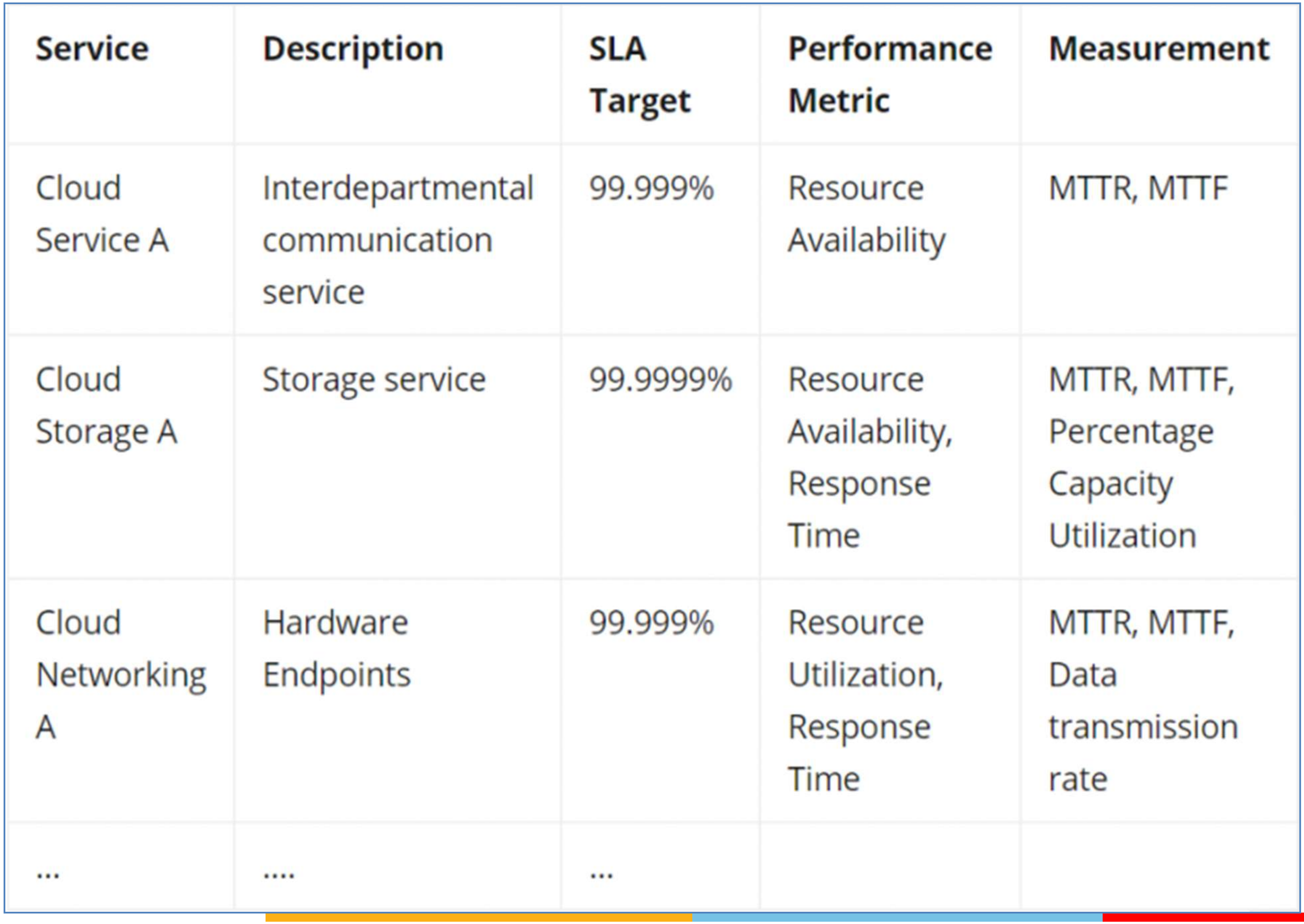
Metric	Commitment	Measurement
Availability		MTTR
Reliability		MTTF

Service Levels, Rankings, and Priority



Severity Level	Description	Target Response
1. Outage	SaaS server down	Immediate
2. Critical	High risk of server downtime	Within 10 minutes
3. Urgent	End-user impact initiated	Within 20 minutes
4. Important	Potential for performance impact if not addressed	Within 30 minutes
5. Monitor	Issue addressed but potentially impactful in the future	Within one business day
6. Informational	Inquiry for information	Within 48 hours

Service Response



Service	Description	SLA Target	Performance Metric	Measurement
Cloud Service A	Interdepartmental communication service	99.999%	Resource Availability	MTTR, MTTF
Cloud Storage A	Storage service	99.9999%	Resource Availability, Response Time	MTTR, MTTF, Percentage Capacity Utilization
Cloud Networking A	Hardware Endpoints	99.999%	Resource Utilization, Response Time	MTTR, MTTF, Data transmission rate
...	....	...		

Key aspects of SLA

Key contractual components of an application SLA:

Service level parameter metric

- Web site response time (e.g., max of 3.5 sec per user request)
- Latency of web server (WS) (e.g., max of 0.2 sec per request)
- Latency of DB (e.g., max of 0.5 sec per query)

Function:

- Average latency of WS (latency of web server 1+latency of web server 2) /2
- Web site response time: Average latency of web server+ latency of database

Measurement directive

- DB latency available via http://mgmtserver/em/latency
- WS latency available via http://mgmtserver/ws/instanceno/latency

Service level objective - Service assurance

- web site latency , 1 sec when concurrent connection , 1000 ms single
- Penalty 1000 USD for every minute while the SLO was breached

Key aspects of SLA

Key Contractual Elements of an Infrastructural SLA

- **Hardware availability**: 99% uptime in a calendar month
- **Power availability**: 99.99% of the time in a calendar month
- **Data center network availability**: 99.99% of the time in a calendar month
- **Backbone network availability**: 99.999% of the time in a calendar month
- **Service credit for unavailability**: Refund of service credit prorated on downtime period
- **Outage notification guarantee**: Notification of customer within 1 hr of complete downtime
- **Internet latency guarantee**: When latency is measured at 5 min intervals to an upstream provider, the average doesn't exceed 60 msec

Packet loss guarantee: Shall not exceed 1% in a calendar month

Key aspects of SLA

- Each SLA goes through a sequence of steps:
  - Identification of terms and conditions
  - Activation
  - Monitoring of the stated terms and conditions
  - Termination of contract once the hosting relationship ceases to exist.
- **SLA life cycle** and consists of the following five phases:

1. Contract definition
2. Publishing and discovery
3. Negotiation
4. Operationalization
5. De-commissioning

Key aspects of SLA

Some of the parameters

- SLA class (Platinum, Gold, Silver, etc
- Penalty associated with SLA breach
- Threshold of breaching OR already breached the SLA
- The number of applications belonging to the same customer that has breached SLA.
- The number of applications belonging to the same customer about to breach SLA
- The type of action to be performed to rectify the situation.

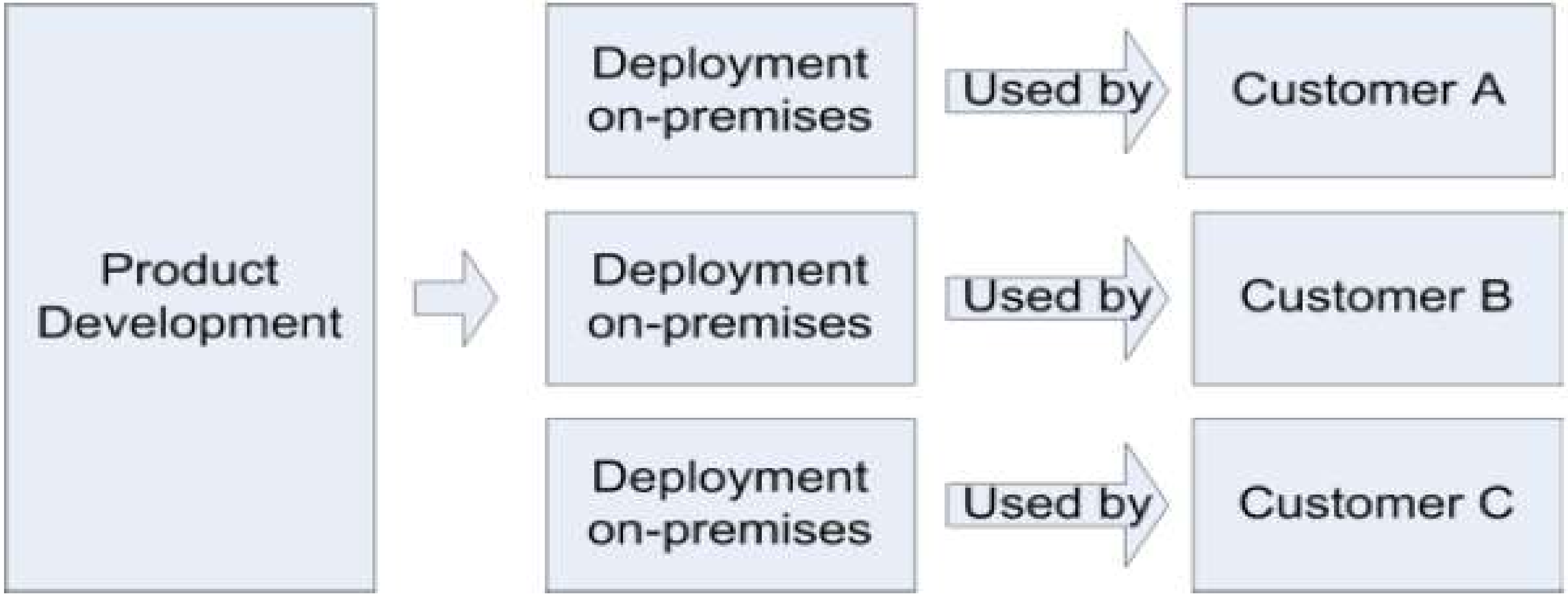
Multitenancy – What is it?

Pros and Cons

	House	Apartment
Effective use of land	-	+
Privacy	+	-
Infrastructure sharing	-	+
Maintenance cost sharing	-	+
Freedom	+	-

House: Privacy and freedom
Apartment: Cost efficiency

Traditional Deployment Model



Multitenancy – Introduction

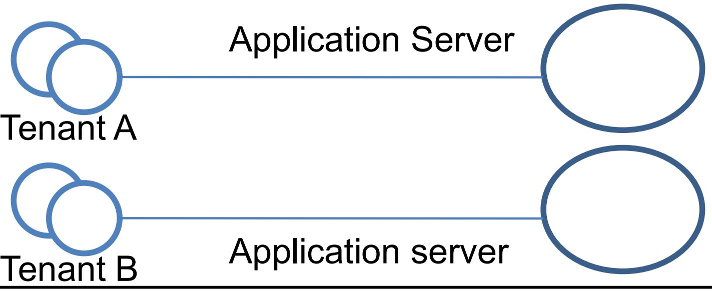
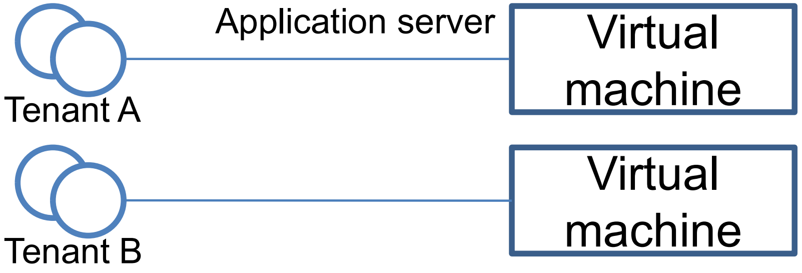
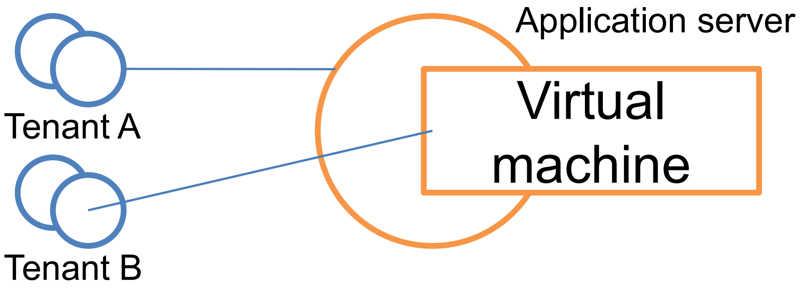
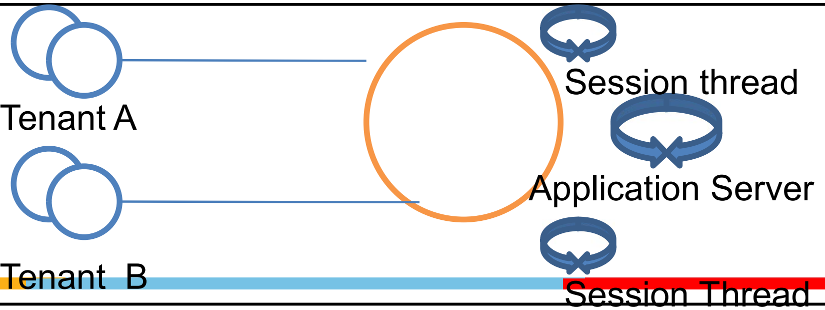
- Multi-tenancy
  - Tenant
  - Customize some parts of the application
  - Cannot customize the application's code
- A software-as-a-service (SaaS) app
- Multi-tenancy is an architectural pattern
- Multi – User vs Multitenancy

Multitenancy – key aspects

- Share the same hardware resources
- Configure the application

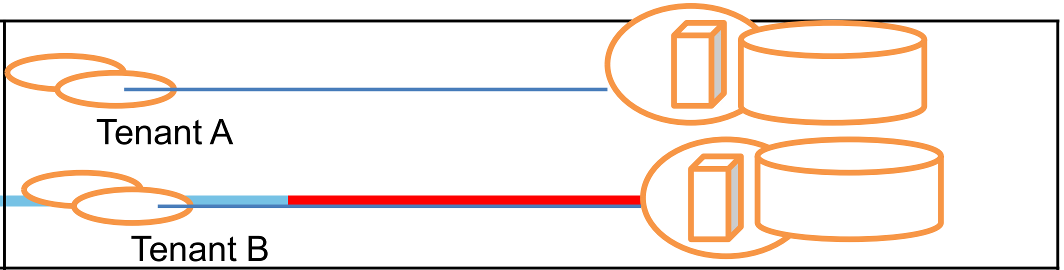
- These definition focus on what we believe to be the key aspects of multi tenancy:
  i. The ability of the application to share hardware resources.
  ii. The offering of a high degree of configurability of the software.
  iii. The architectural approach in which the tenants make use of a single application and database instance.

Multi-tenants Deployment Modes for Application Server

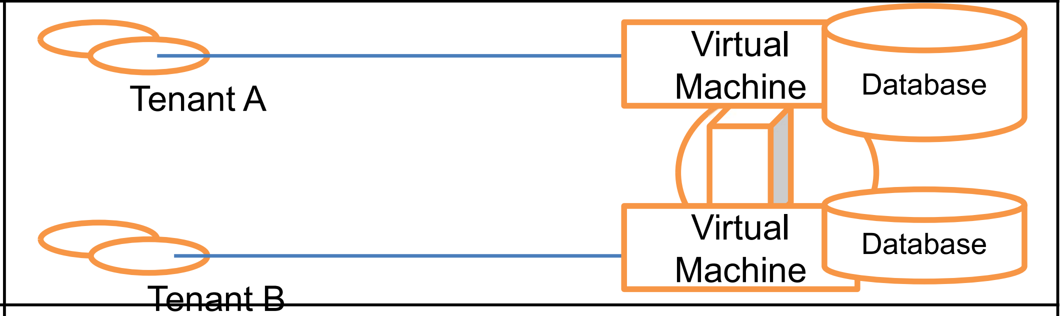
Fully isolated Application server	 Application Server Tenant A Tenant B Application server
Virtualized Application Server	 Application server Virtual machine Tenant A Virtual machine Tenant B
Shared Virtual Server	 Application server Tenant A Virtual machine Tenant B
Shared Application Server	 Tenant A Session thread Application Server Tenant B Session Thread

Multi-tenants Deployment Modes in Data Centers

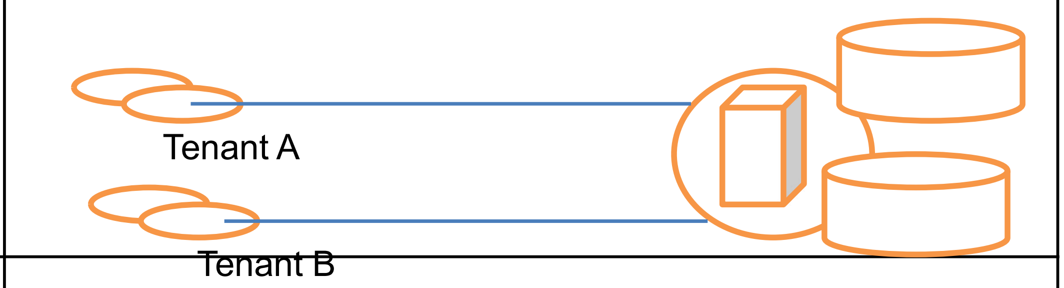
Fully isolated data center



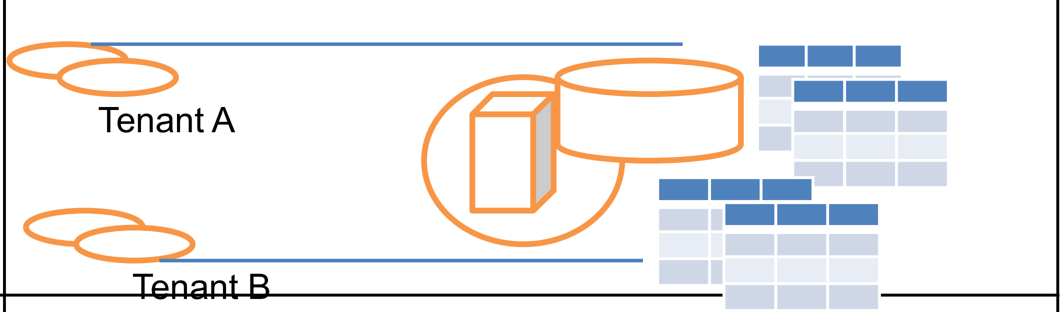
Virtualized servers



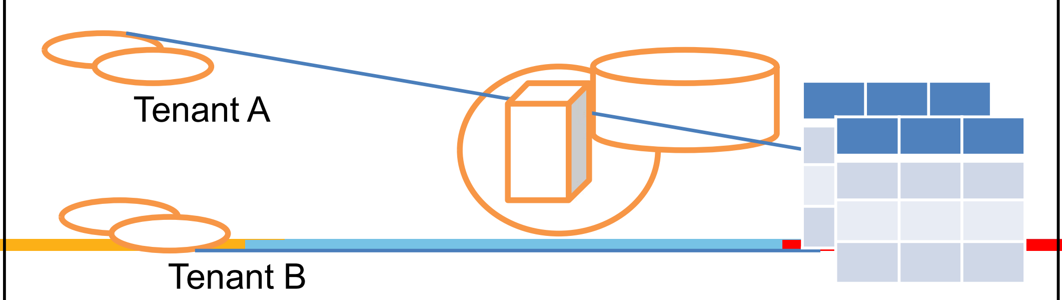
Shared Server



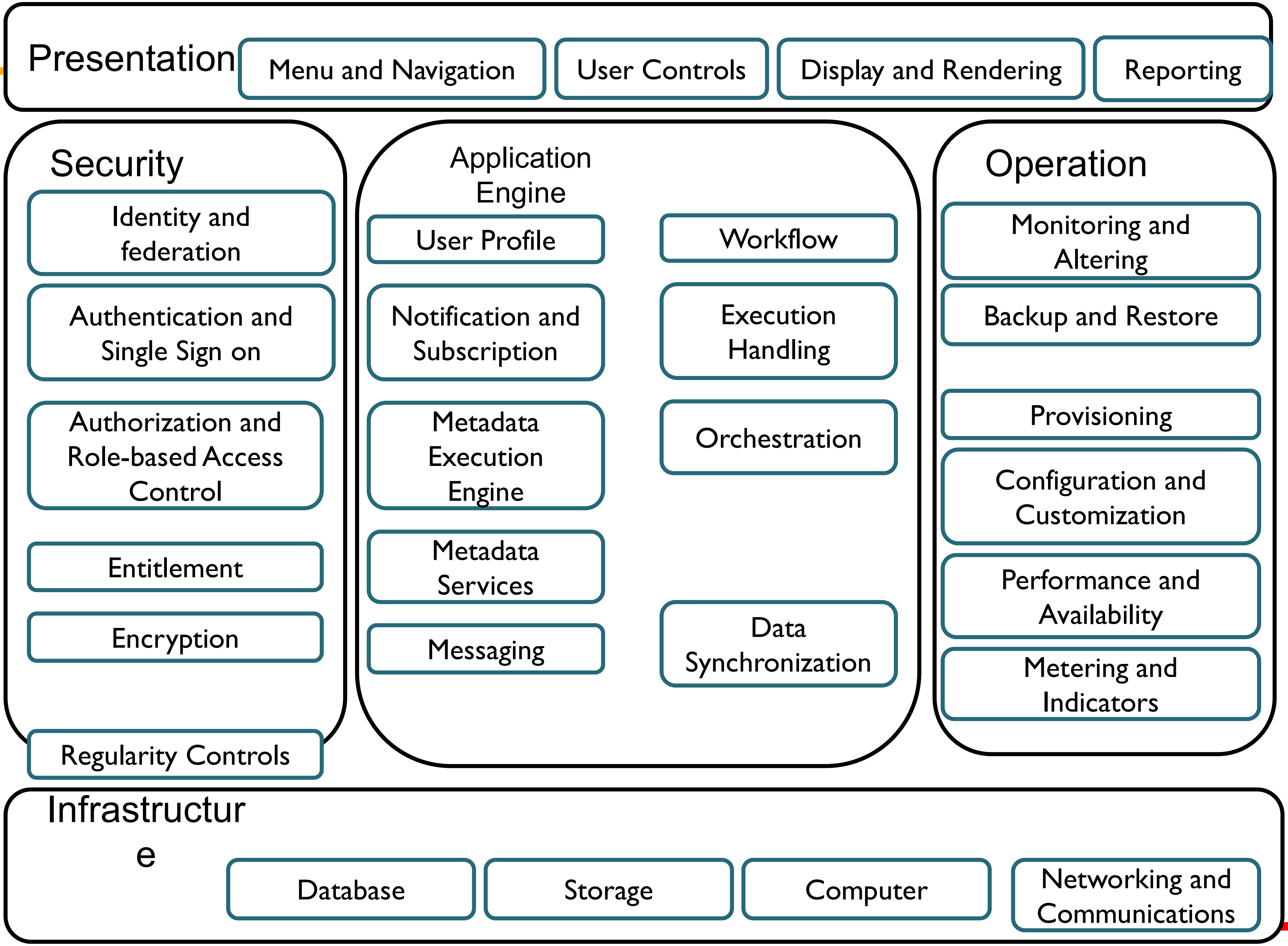
Shared Database



Shared Schema



Conceptual framework of Software as a Service

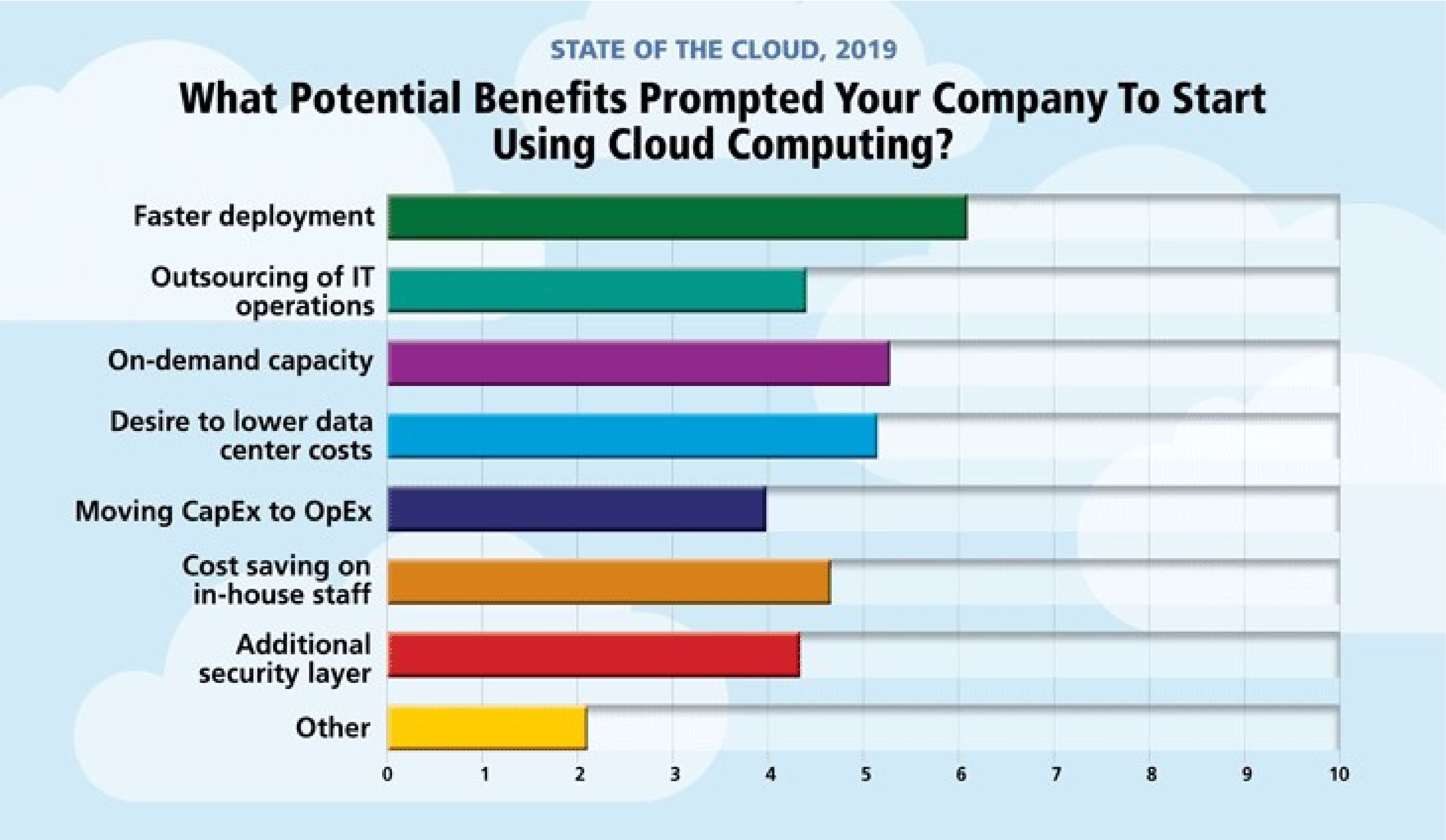


Introduction to cloud security

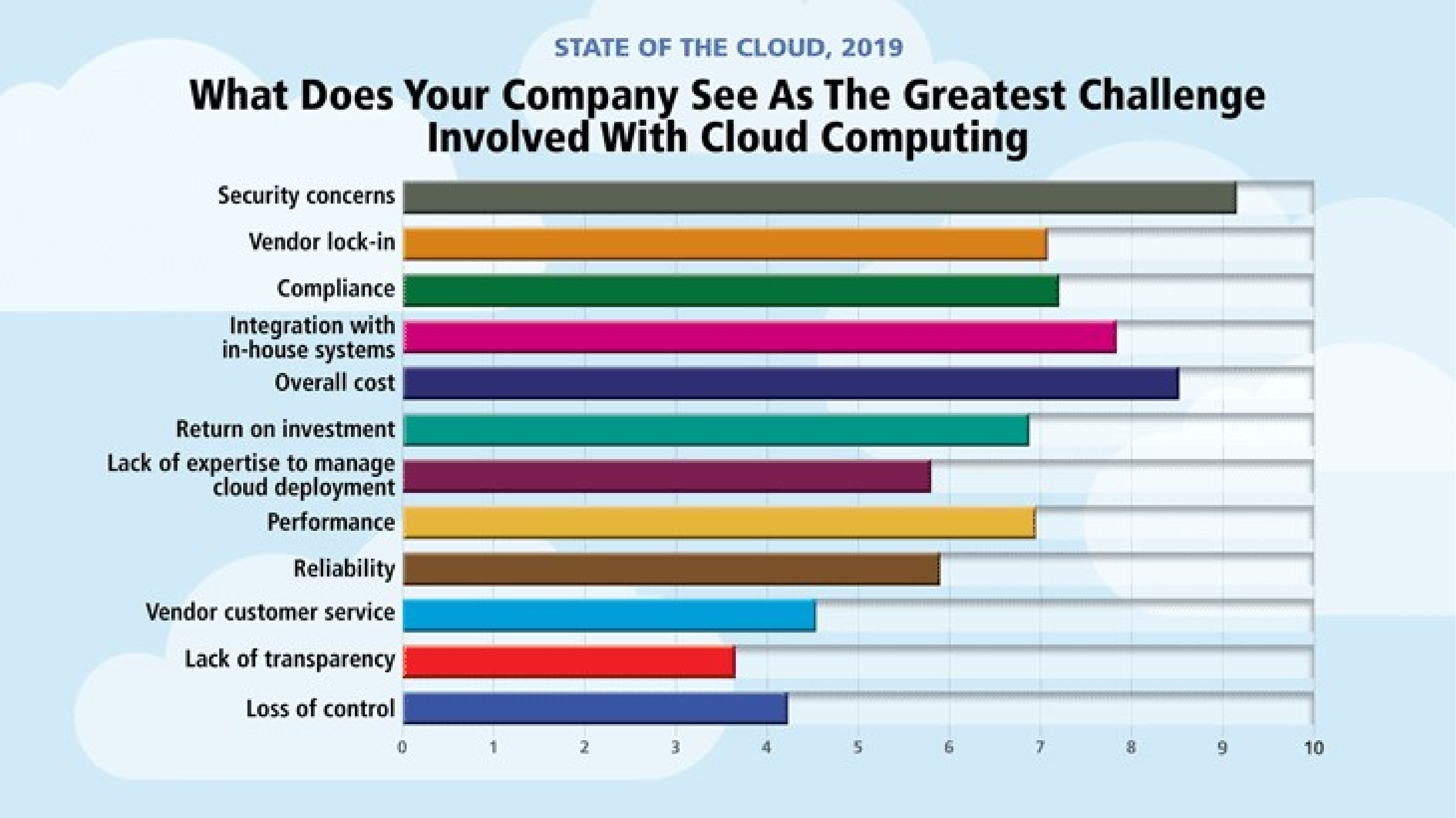
If cloud computing is so great, why isn't everyone doing it?

- The cloud - big black box
- Clouds are still subject to traditional data confidentiality, integrity, availability, and privacy issues

Benefits for company to start using cloud



Companies are still afraid to use clouds



Cloud Security Issues

- Most security problems stem from:
  - Loss of Control
    - Take back control
      - Data and apps may still need to be on the cloud
      - But can they be managed in some way by the consumer?
  - Lack of trust
    - Increase trust (mechanisms)
      - Technology
      - Policy, regulation
      - Contracts (incentives)
  - Multi-tenancy
    - Private cloud
      - Takes away the reasons to use a cloud in the first place
    - VPC: its still not a separate system
    - Strong separation
- These problems exist mainly in 3rd party management models
  - Self-managed clouds still have security issues, but not related to above

Loss of Control in the Cloud

Consumer's loss of control

- Data, applications, resources are located with provider
- User identity management is handled by the cloud
- User access control rules, security policies and enforcement are managed by the cloud provider
- Consumer relies on provider to ensure
  - Data security and privacy
  - Resource availability
  - Monitoring and repairing of services/resources

Multi-tenancy Issues in the Cloud

- Conflict between tenants' opposing goals
  - Tenants share a pool of resources and have opposing goals
- How does multi-tenancy deal with conflict of interest?
  - Can tenants get along together and 'play nicely' ?
  - If they can't, can we isolate them?
- How to provide separation between tenants?
- Cloud Computing brings new threats
  - Multiple independent users share the same physical infrastructure
  - Thus an attacker can legitimately be in the same physical machine as the target

Taxonomy of Fear

- Confidentiality
  - Fear of loss of control over data
    - Will the sensitive data stored on a cloud remain confidential?
    - Will cloud compromises leak confidential client data
  - Will the cloud provider itself be honest and won't peek into the data?
- Integrity
  - How do I know that the cloud provider is doing the computations correctly?
  - How do I ensure that the cloud provider really stored my data without tampering with it?

Taxonomy of Fear

Availability

- Will critical systems go down at the client, if the provider is attacked in a Denial of Service attack?
- What happens if cloud provider goes out of business?
- Would cloud scale well-enough?
- Often-voiced concern
  - Although cloud providers argue their downtime compares well with cloud user's own data centers

Taxonomy of Fear

- Privacy issues raised via massive data mining
  - Cloud now stores data from a lot of clients, and can run data mining algorithms to get large amounts of information on clients
- Increased attack surface
  - Entity outside the organization now stores and computes data, and so
  - Attackers can now target the communication link between cloud provider and client
  - Cloud provider employees can be phished

Taxonomy of Fear

- Audit-ability and forensics (out of control of data)
  - Difficult to audit data held outside organisation in a cloud
  - Forensics also made difficult since now clients don't maintain data locally
- Legal issues
  - Who is responsible for complying with regulations?
  - e.g., SOX, HIPAA, GLBA ?
  - If cloud provider subcontracts to third party clouds, will the data still be secure?

Threat Model

- A threat model helps in analysing a security problem, design mitigation strategies, and evaluate solutions
- Steps:
  - Identify attackers, assets, threats and other components
  - Rank the threats
  - Choose mitigation strategies
  - Build solutions based on the strategies

Threat Model

- Basic components
  - Attacker modelling
    - Choose what attacker to consider
      - insider vs. outsider?
      - single vs. collaborator?
    - Attacker motivation and capabilities
  - Attacker goals
  - Vulnerabilities / threats

Thank you