CSCI 5409 Adv. Topics in Cloud Computing – Fall, 2023
Week 12 – Lecture 2 (Nov 24, 2023)

Business Considerations in Cloud Computing (2)

Dr. Lu Yang
Faculty of Computer Science
Dalhousie University
luyang@dal.ca

Housekeeping and Feedback

- Start recording
- Starting working on the term project. Ask Purvesh and Rahul questions.
- SLEQ

Objectives

- Understand service-level agreements and the service quality metrics used to audit cloud computing service performance

Contents

Service Quality Metrics & SLAs

Overview

- **SLAs** issued by cloud providers are "human-readable documents that describe quality-of-service features, guarantees, and limitations of one or more cloud-based IT resources."[1]

  https://aws.amazon.com/legal/service-level-agreements/

- SLAs use **service quality metrics** to express measurable quality-of-service characteristics[1]:
  - Availability
  - Reliability
  - Performance
  - Scalability
  - Resiliency
- SLA example:
  https://www.bmc.com/blogs/sla-template-examples/

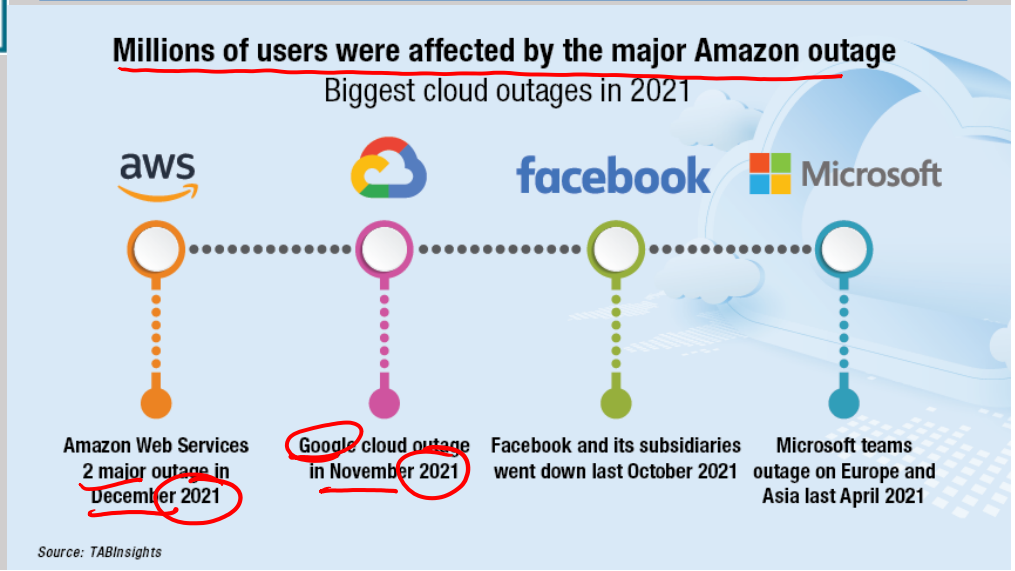
- Service quality metrics are defined by these characteristics[1]:
  - **Quantifiable** – There is an appropriate unit of measure
  - **Repeatable** – Metrics yield identical results when repeated under identical conditions
  - **Comparable** – Standardized to allow comparison with competitors
  - **Easily Obtainable** – Easily retrieved and understood by the cloud consumer

[1] Cloud Computing (T. Erl, Z. Mahmoud, R. Puttini, 2013) pg. 404 - 405

Availability

Availability (%)	Downtime/Week (Seconds)	Downtime/Month (Seconds)	Downtime/Year (Seconds)
99.5	3024	216	158112
99.8	1210	5174	63072
99.9	606	2592	31536
99.95	302	1294	15768
99.99	60.6	259.2	3154
99.999	6.05	25.9	316.6
99.9999	0.605	2.59	31.5

https://www.theasianbanker.com/updates-and-articles/cloud-services-outage-effects-on-the-global-cloud-market



Availability

- **July AWS Region Outage**
  - On July 28, 2022, Seattle-based AWS experienced a power loss in a single availability zone of the U.S. East 2 region—located in Ohio—that lasted about 20 minutes but knocked out third-party services for up to three hours, according to a report from ThousandEyes.
  - The loss of power started at 9:57 a.m. Pacific and was restored at 10:19 a.m. Pacific, according to the report. And customers with multiple availability zone redundancy likely failed over to a working zone.
  - The outage affected Amazon's Elastic Compute Cloud (EC2), Webex and Okta, among other services and ISVs, according to the report.
  - In its own report on the incident, Metrist said the outage affected AWS' CloudFront, CloudWatch, Amazon Elastic Kubernetes Service (EKS) and Lambda services, among others.
  - The report also credited the outage with affecting service from ISVs including Zoom and New Relic, according to Metrist.

https://www.crn.com/news/cloud/the-15-biggest-cloud-outages-of-2022/12

Reliability

- **Reliability** is "the probability that an IT resource can perform its intended function under pre-defined conditions without experiencing failure."[1]
- Usually tracked by **mean-time between failures**
- May also be tracked by service outcomes:
  - Example:
    - 5 / 5 performed as expected, 100% reliability
    - 4/5 performed as expected, 80% reliability

[1] Cloud Computing (T. Erl, Z. Mahmoud, R. Puttini, 2013) pg. 407

Performance

- Service **performance** "refers to the ability of an IT resource to carry out its functions within expected parameters."[1]
- Measured differently for each resource:
  - Network capacity: bandwidth
  - Storage device capacity: size in GB
  - Server capacity: number of CPUs, amount of RAM and storage
  - Web app capacity: number of requests/minute
  - Instance starting time
  - Response time of an asynchronous task
  - Completion time of an asynchronous task

[1] Cloud Computing (T. Erl, Z. Mahmoud, R. Puttini, 2013) pg. 407 - 408

Scalability

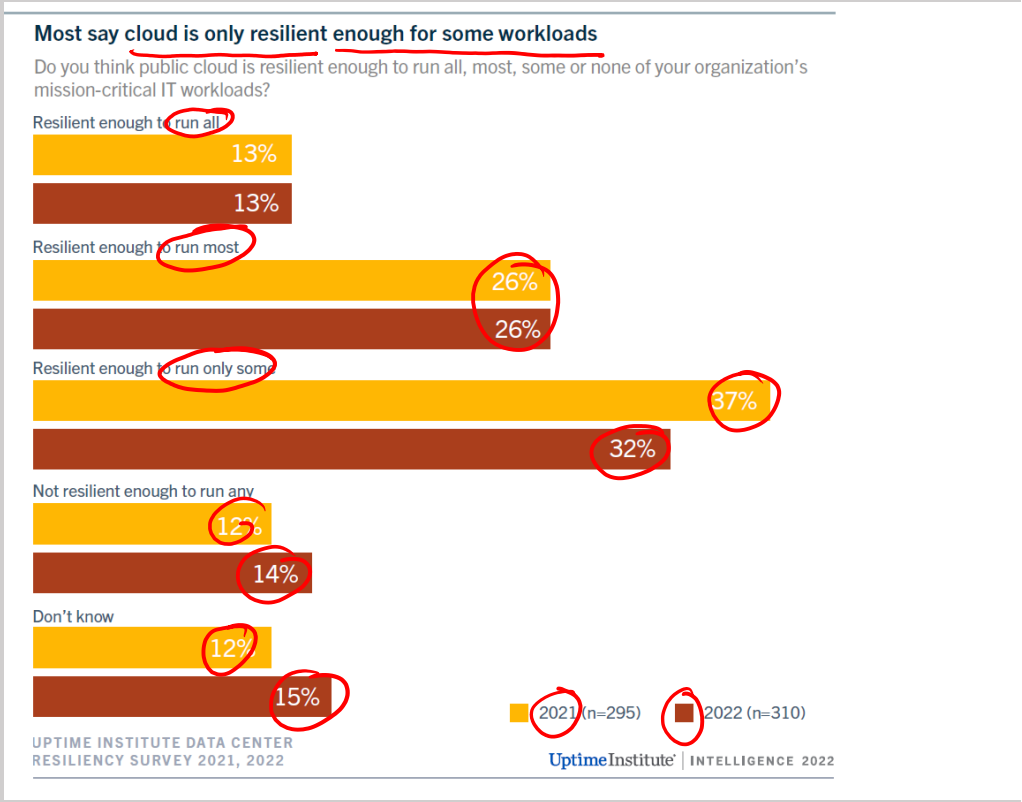
- Service **scalability** metrics "are related to IT resource elasticity capacity, which is related to the maximum capacity that an IT resource can achieve, as well as measurements of its ability to adapt to workload fluctuations."[1]
- Metrics:
  - **Storage scalability (Horizontal)**: Maximum potential storage capacity
  - **Server scalability (Horizontal)**: Minimum and maximum virtual servers
  - **Server scalability (Vertical)**: Maximum number of CPUs, amount of RAM

Storage (Vertical)
database

[1] Cloud Computing (T. Erl, Z. Mahmoud, R. Puttini, 2013) pg. 409 - 410

Resilience

- "The ability of an IT resource to recover from operational disturbances"[1]
- Cloud provider guarantees are achieved through redundancy and disaster recovery systems
- Two metrics for measurement[1]:
  - **Mean-Time to Switchover**: The time expected to complete switchover from a severe failure to a replicated instance in a different geographical area
  - **Mean-Time System Recovery**: The time expected for a resilient system to perform a completed recovery from a severe failure



https://journal.uptimeinstitute.com/is-concern-over-cloud-and-third-party-outages-increasing/

[1] Cloud Computing (T. Erl, Z. Mahmoud, R. Puttini, 2013) pg. 411 - 413

Guideline

- Best practices for working with SLAs[1]:
  - **Map Business Cases to SLAs**
    - Identify the necessary quality-of-service requirements for each project, then link them to concrete guarantees in the SLA for responsible IT resources
  - **Document Guarantees at Appropriate Granularity**
    - If your organization has specific requirements, the corresponding level of detail should be used to describe the guarantee.
  - **Define Penalties for Non-Compliance**
  - **Disclosing Cross-Cloud Dependencies**
    - If cloud providers are leasing IT resources from other cloud providers, this results in a loss of control over their own guarantees. This is a risk you need to be aware of.

[1] Cloud Computing (T. Erl, Z. Mahmoud, R. Puttini, 2013) pg. 413

Risk – Contract Complexity

- "Another aspect to contractual obligations is where the lines are drawn between cloud consumer and cloud provider"[1]
- Anything you build in the cloud becomes technology architected from artifacts owned by the cloud consumer (you) and the cloud provider
- If something happens how is blame determined?
- Where are disputes arbitrated?

[1] Cloud Computing (T. Erl, Z. Mahmoud, R. Puttini, 2013) pg. 132

Risk – Legacy Spaghetti

- "Moving an enterprise's legacy IT into the cloud is difficult because it forces tough decisions about consolidation and standardization. Most organizations that have been around awhile have a hodgepodge of hardware, operating systems, and applications often described as **'legacy spaghetti'**, few are willing to give up their portion of it just so their company can move to the cloud."[1]
- When we build large, complicated systems that have long lifetimes you will find them held together by duct tape, patches, propped up by old programming books
  - It's very difficult to take these things apart, and then reassemble them with modern tools, languages and cloud services
  - It is also very difficult to estimate the time, and therefore the **cost** of this process

[1] https://hbr.org/2011/11/what-every-ceo-needs-to-know-about-the-cloud

2

Cost Optimization Best Practices

Cost Optimization Best Practices[1]

- Train your organization to understand the new model, long approval, procurement and deployment cycles are a thing of the past, organizations must "speed up". It is the "whole team" approach including business people, not just the technology team
- Conversely, the technology team must strive to understand business objectives and priorities, all technology must be directly connected to achieving objectives with the **business's** priorities in mind, not the developers.
- Costs must be categorized and attributed to individual departments or product owners to know which products are truly profitable and make more informed budget allocation decisions.
- By employing a checks-and-balances approach, you can innovate without overspending.
- Implement a decommissioning policy through change control and resource management from project inception to end-of-life. This makes sure you terminate unused resources to reduce waste.
- Use the appropriate instances and resources for your workload. Do the math! Is horizontal scaling better than vertical? Run the numbers! Virtualization and on-demand usage makes this simple.
- Decide whether **managed services** are more cost effective than administrative and operational overhead
- Evaluate pricing models (reserved vs. on-demand)
- Architect to minimize data transfer

[1] https://docs.aws.amazon.com/wellarchitected/latest/cost-optimization-pillar/


Kahoot!