Distributed Systems
CS6421
Performance

Prof. Roozbeh Haghnazar

Slides Credit:

Prof. Tim Wood and Prof. Roozbeh Haghnazar

Final Project

Questions?

- Implementation phase!
  - Get coding!
  - Keep your code in GitHub
- Schedule meetings with us!
  - Especially if you realize you can't achieve what you originally planned
- Timeline
  - Milestone 0: Form a Team
  - Milestone 1: Select a Topic
  - Milestone 2: Literature Survey
  - Milestone 3: Design Document
  - **Milestone 4: Final Presentation**

Last Time…

- Replication and Consistency
  - Why replicate
  - What is consistency?
  - Consistency Models
  - Quorum Replication

- Exam
  - Avg: 90%

This Time…

- Performance in Dist. Systems
  - Introduction
  - Performance metrics
  - Models
  - Architectures

But first we need to finish a bit of consistency!

DistSys Challenges

- **Heterogeneity**
- Openness
- Security
- Failure Handling
- Concurrency
- **Quality of Service**
- **Scalability**
- Transparency

Performance Challenges

Problem

- Amazon: 100 ms extra latency costs 1% in sales , In 2020, Amazon Web Services (AWS) generated revenues of **45.37 billion U.S. dollars** with its cloud services.
- Bing: 2s slowdown would reduce the revenue by 4.3%
- Google: (August 16 2013) 5 minute down time cost 545k$
- Increasing 400ms web search latency caused a decrease of about 0.74% in Google's search frequency

What is Performance?

- Marriam-webster:
  - the execution of an action
  - the fulfillment of a claim, promise, or request
  - the manner of reacting to stimuli

- Wikipedia
  - In computing, computer performance is the amount of useful work accomplished by a computer system. Outside of specific contexts, computer performance is estimated in terms of accuracy, efficiency and speed of executing computer program instructions.

What is Performance?

- Performance considers:
  - Latency (transmission delay)
  - Bandwidth (maximal transmission capacity)
  - Throughput (average transmission rate)
  - Response time (time to see result of action)

WHAT IS THE FIRST STEP???

How can we choose the best service as a client?

How can we make sure that we have provide the best service as a provider?

How can we understand that what are we going to design as an engineer?

QoS & SLA

Service Level Agreements (SLA)

- Service Level Agreements are fundamental to an effective cloud utilization and especially business customers need them to ensure risks and service qualities are prevented respectively provided in the way they want.
- The confirmed SLAs serve as a basis for compliance and monitoring of the QoS.
- Due to the dynamic cloud character, the QoS attributes must be monitored and managed consistently

How can I describe and measure the QoS?

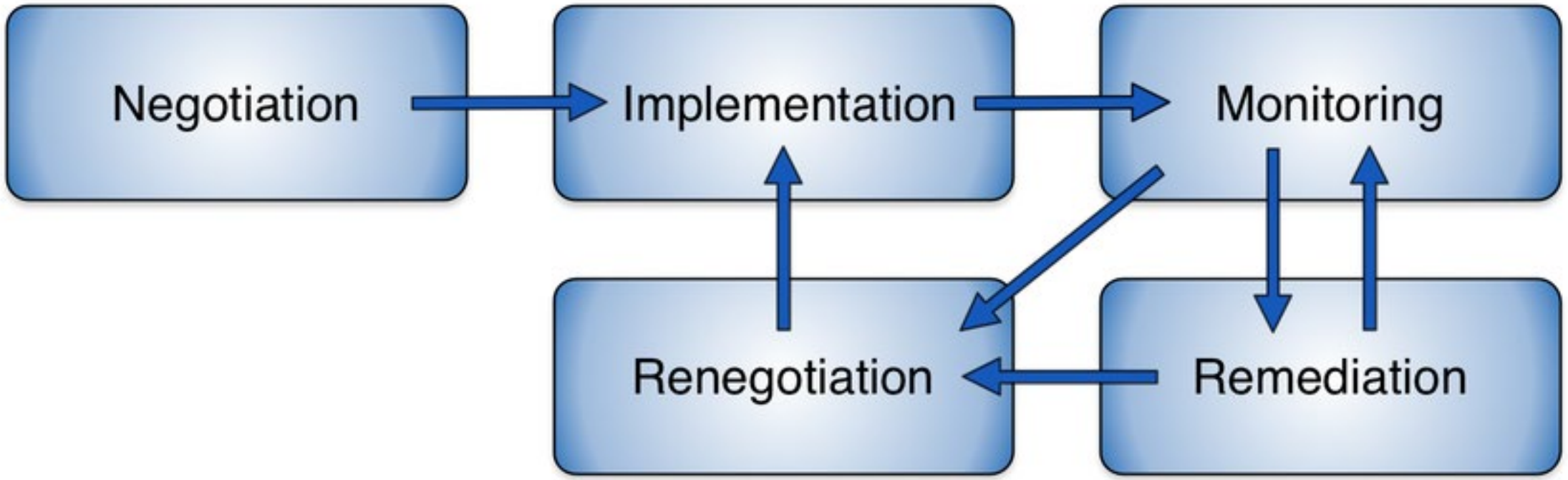
KPI

How can we define SLA

- NIST has pointed out the necessity of SLAs, SLA management, definition of contracts, orientation of monitoring on Service Level Objectives (SLOs) and how to enforce them. (https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication500-292.pdf)

- There are two major specification for describing SLAs
  - Web Service Level Agreement (WSLA) Language Specification, was developed by IBM with the focus on performance and availability metrics.
  - WS-Agreement (WS-A) was developed by the Open Grid Forum in 2007.
  - The newest update, which is based on the work of the European SLA@SOI project.

SLA Life cycle



Service Level Objectives

- Service Level Objectives (SLOs) are a central element of every service level agreements (SLA), which include:
  - negotiated service qualities (service level)
  - corresponding Key Performance Indicators.
- SLOs contain the specific and measurable properties of the service, such as ***availability***, ***throughput or response time*** and often consist of ***combined or composed attributes***

SLO Characteristics

- SLOs should thereby have the following characteristics:
  - Repeatable
  - Measurable
  - Understandable
  - Significant
  - Controllable
  - Affordable
  - Mutually acceptable
  - Influential

We need KPI

SLO Characteristics

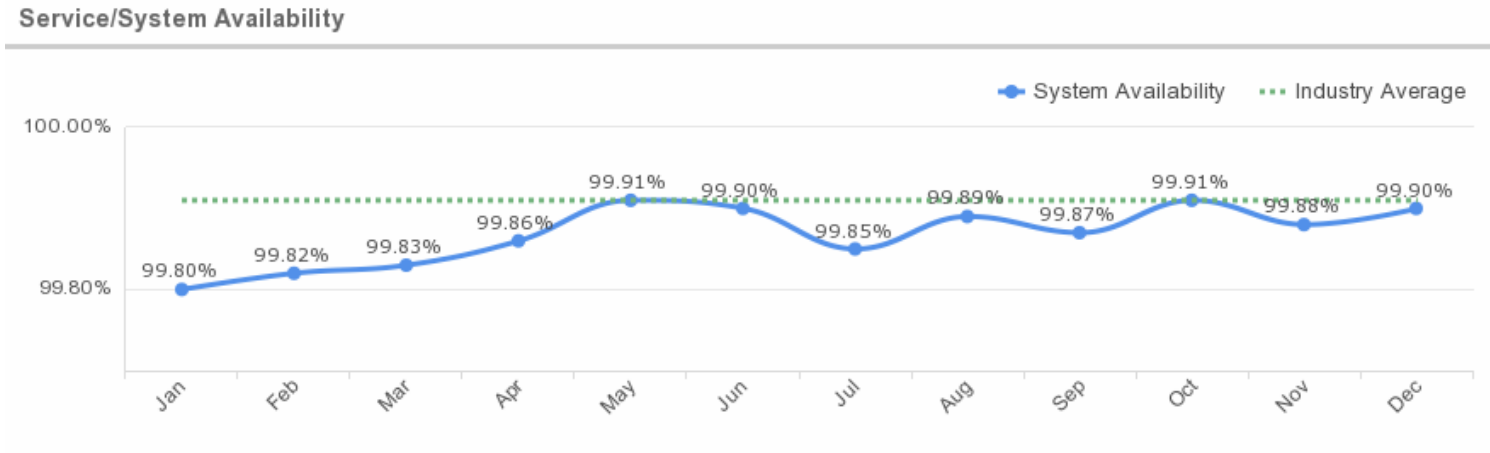
- A valid SLO specification might, for instance, look like this:
  - *The IT system should achieve an availability of 98% over the measurement period of one month. The availability represents thereby the ratio of the time in which the service works with a response time of less than 100ms plus the planned downtime to the total service time, measured at the server itself.*

KEY PERFORMANCE METRICS

- General Service KPIs
- Network Service KPIs
- Cloud Storage KPIs
- Backup and Restore KPIs
- Infrastructure as a Service KPIs

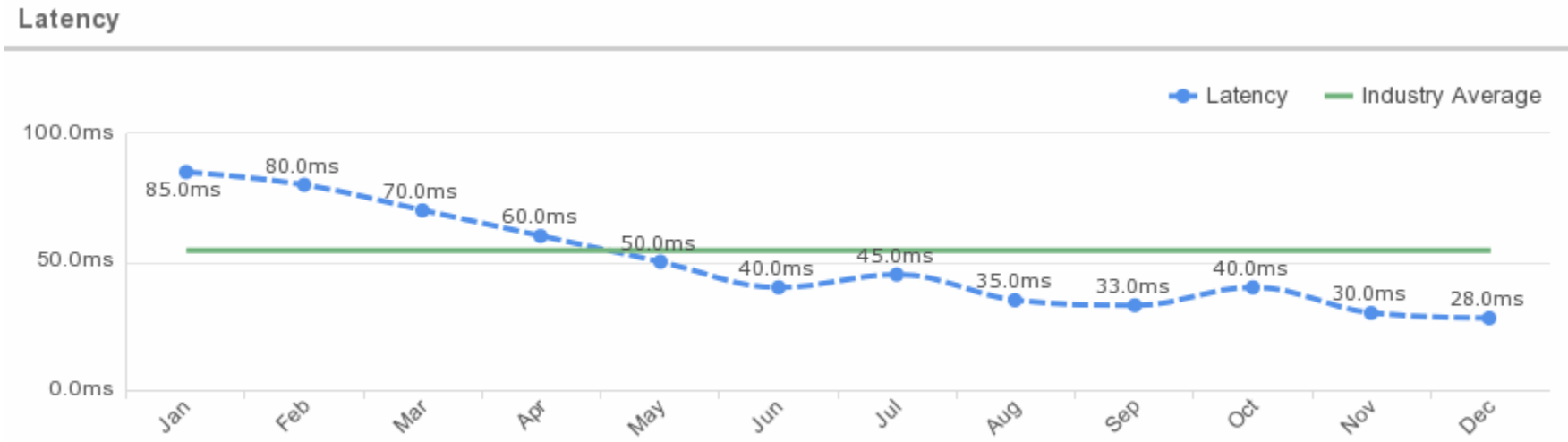
General Service KPIs

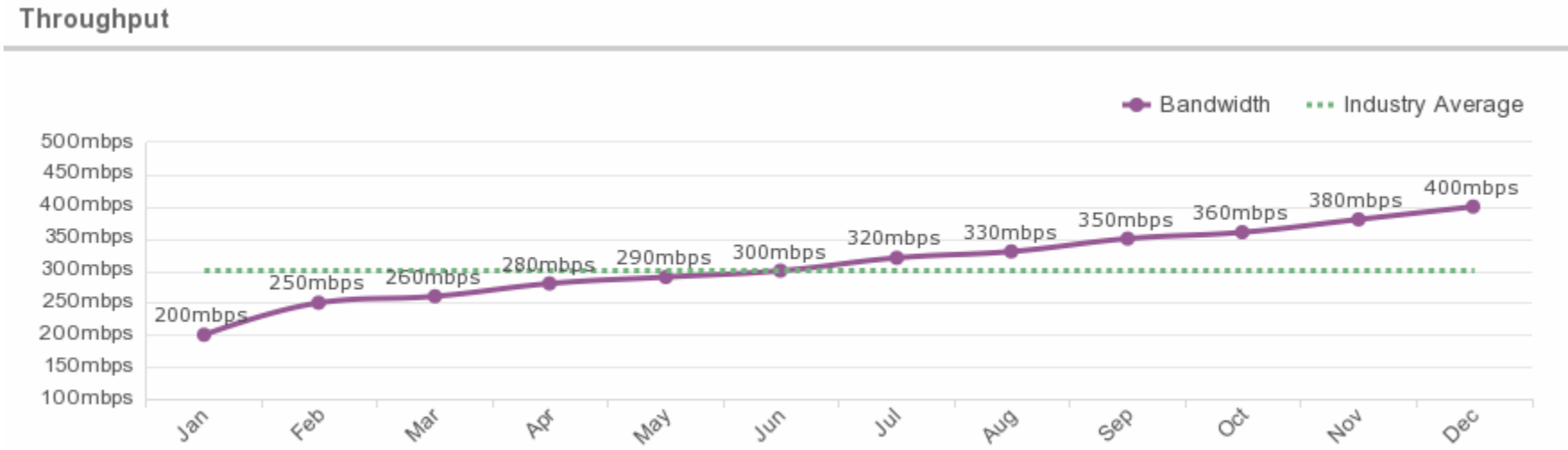
- Basic Services
- Security
- Service and Helpdesk
- Monitoring
- Etc.



Network Service KPIs

- Round Trip Time
- Response Time
- Packet Loss
- Bandwidth
- Throughput
- Network Utilization
- Latency
- Etc.





Cloud Storage KPIs

- Response Time
- Throughput
- Average Read Speed
- Average Write Speed
- Random Input / Outputs per second (IOPS)
- Sequential Input / Outputs per second (IOPS)
- Free Disk Space
- Provisioning Type
- Average Provisioning Time

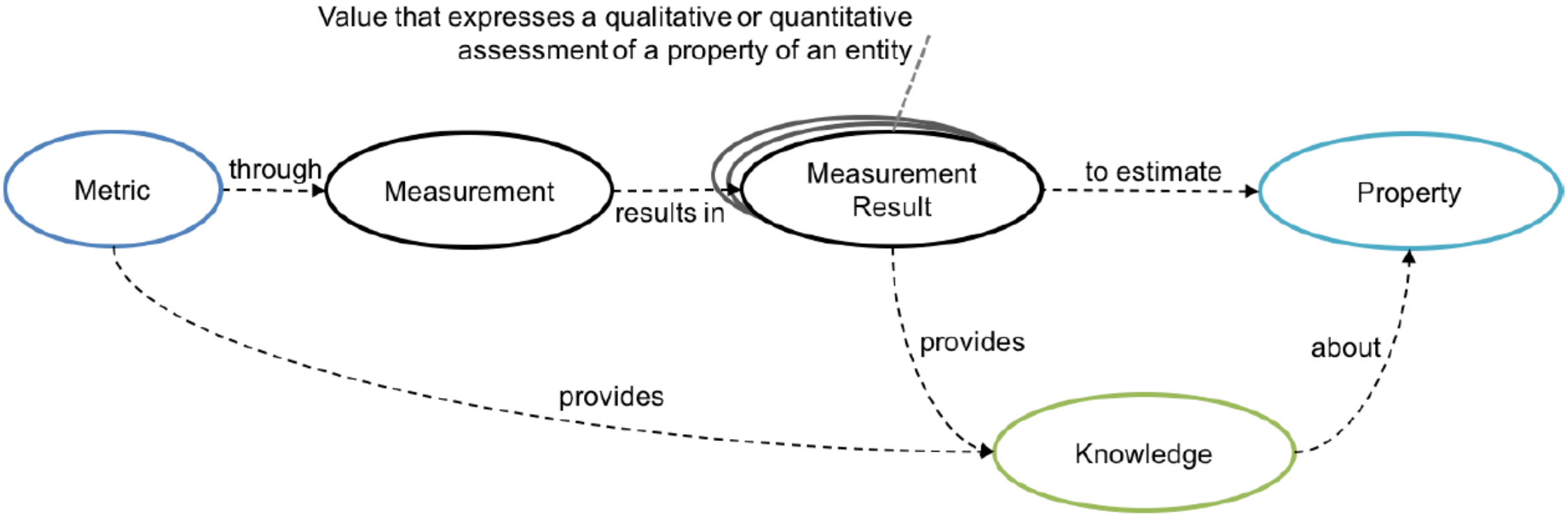
Backup and Restore KPIs

- Backup Interval
- Backup Type
- Time To Recovery
- Backup Media
- Backup Archive

Infrastructure as a Service KPIs

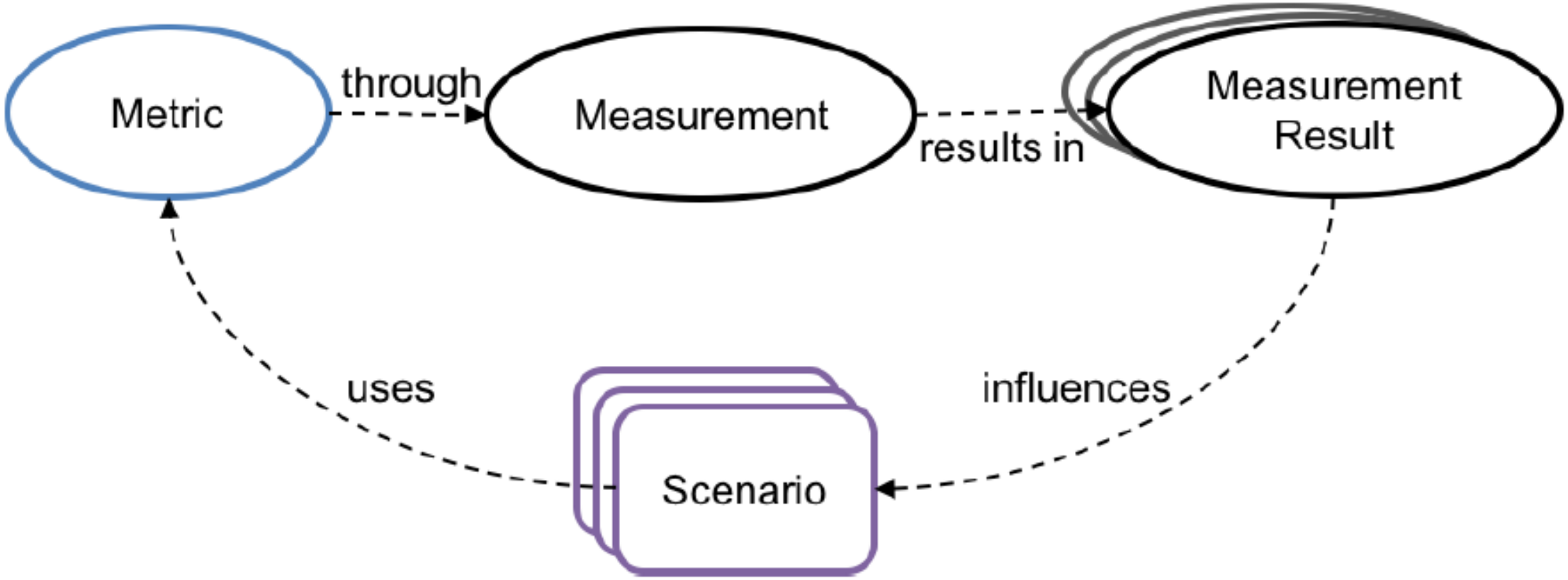
- VM CPUs
- CPU Utilization
- VM Memory
- Memory Utilization
- Minimum Number of VMs
- Migration Time
- Migration Interruption Time
- Logging

METRIC AND PROPERTY



https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.500-307.pdf

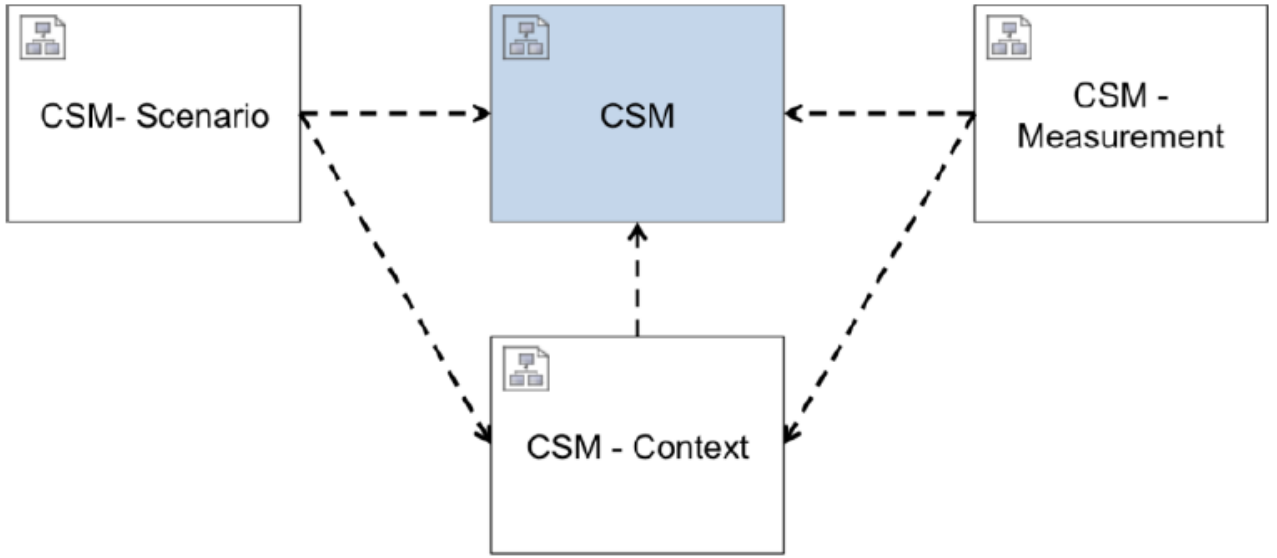
Scenario and metric



https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.500-307.pdf

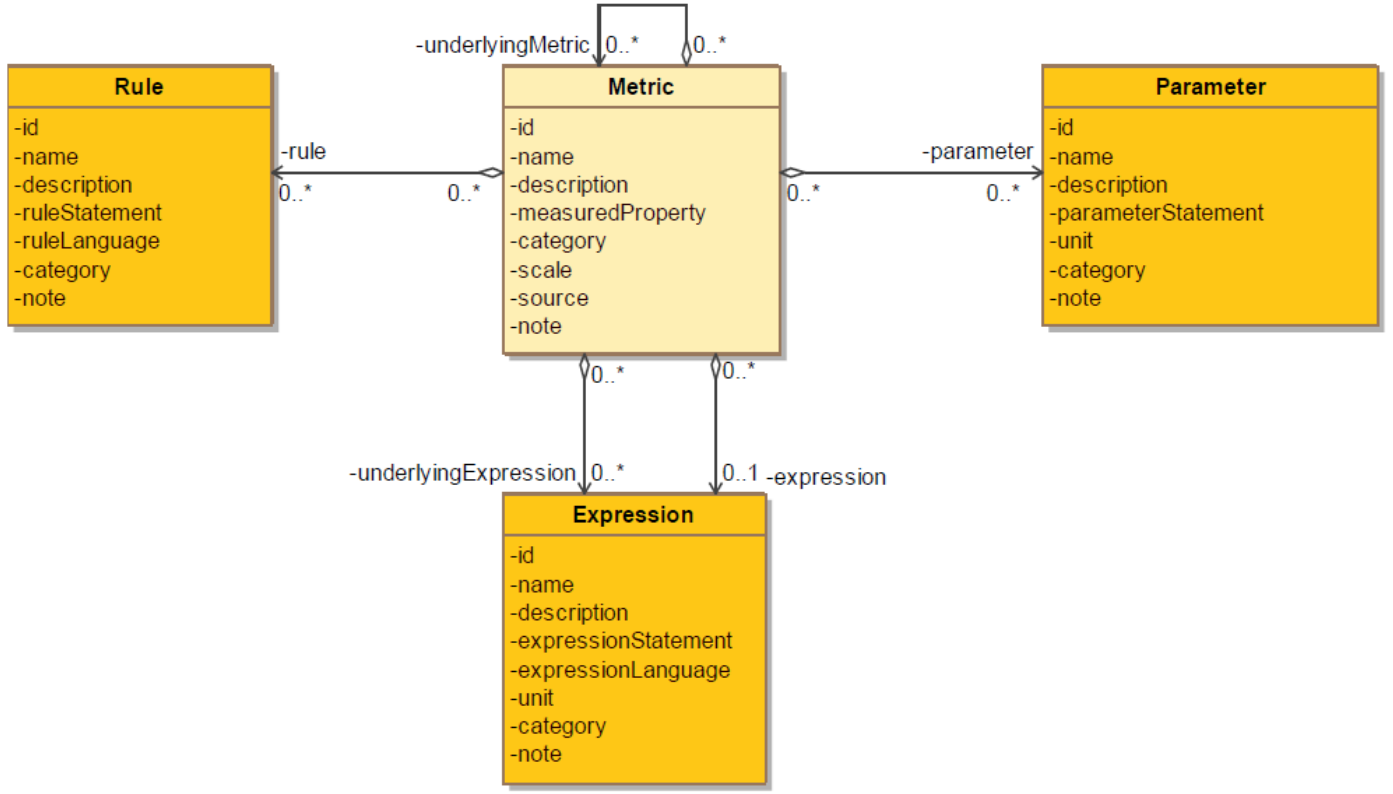
Cloud Service Metric Ecosystem Model

- CSM: The description and definition of a standard of measurement (e.g. metric for customer response time)
- CSM Context: The context related to using the CSM in a specific scenario. (e.g. objectives and applicability conditions of the customer response time metric)
- CSM Measurement: The use of the CSM to make measurements (e.g. the measurement of response time property based on the customer response time metric)
- CSM Scenario: The use of the CSM in a scenario (e.g. the selection and use of the customer response time metric in an SLA)



https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.500-307.pdf

Cloud Service Metric (CSM)



https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.500-307.pdf

Summary

- SLA: The agreement you make with the clients and users
- SLO: The objectives your team must hit to meet that agreement
- KPI: Key performance indicators

REFERENCES:

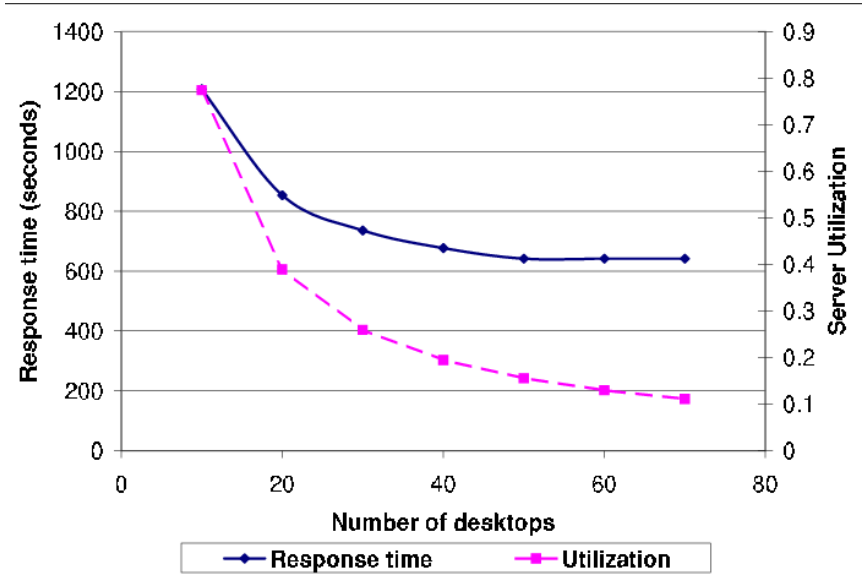
- http://www.thinkmind.org/articles/emerging_2013_3_30_40082.pdf
- https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.500-307.pdf
- https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication500-292.pdf
- https://www.cocop-spire.eu/content/key-performance-indicator-kpi-and-impact-evaluation-distributed-production-systems-%E2%80%93-importance-feedback
- https://www.atlassian.com/incident-management/kpis
- https://www.atlassian.com/incident-management/kpis/sla-vs-slo-vs-sli
- https://cloud.google.com/blog/products/gcp/sre-fundamentals-slis-slas-and-slos
- https://cloud.google.com/blog/products/gcp/availability-part-deux-CRE-life-lessons

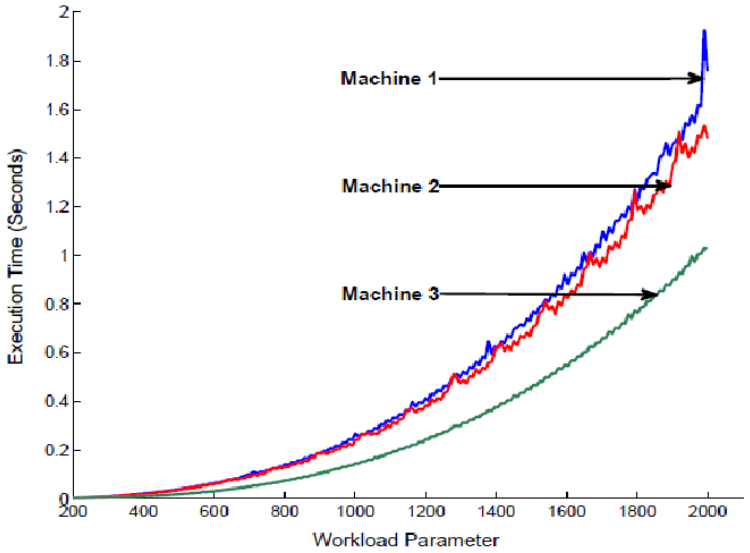
Prof. Tim Wood & Prof. Roozbeh Haghnazar

Performance Modeling

Performance Models

- Measuring performance is not always enough
- We want to Predict these metrics in advance
  - How will response time change if my workload doubles
  - How much memory CPU do I need to get a target throughput

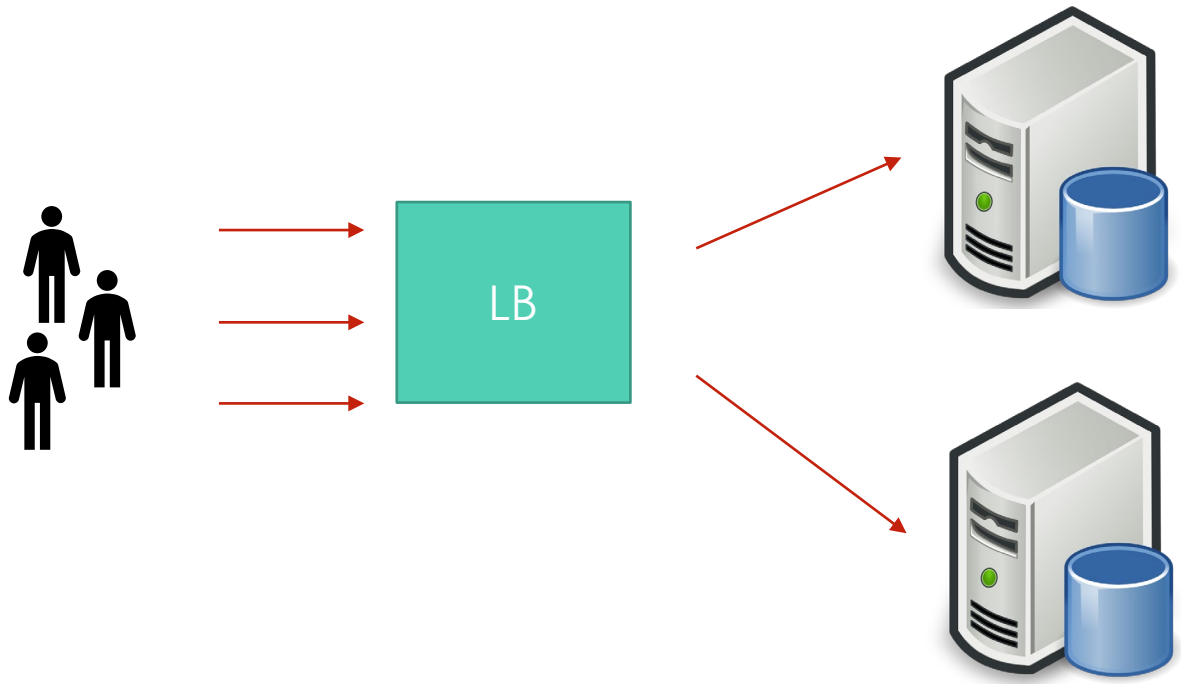




Model Scenario

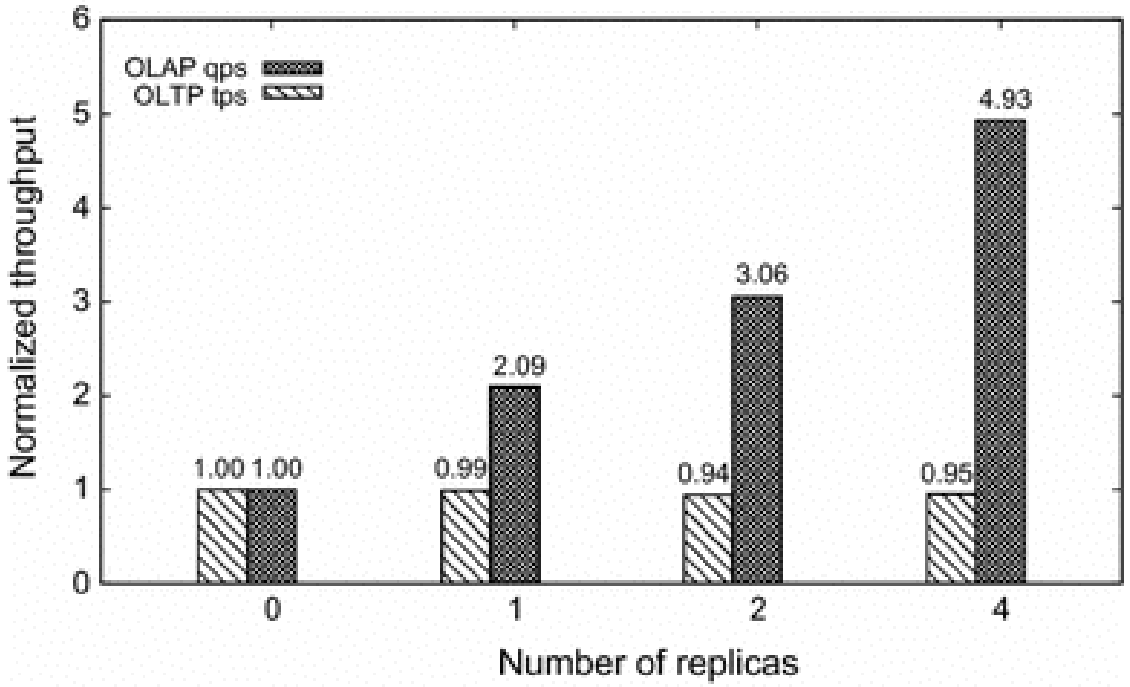
- Consider this web application

- What will affect its performance
  - Clients + Workload
  - Network + Latency
  - Application details
  - Server details
  - # Servers
  - LB Algorithms



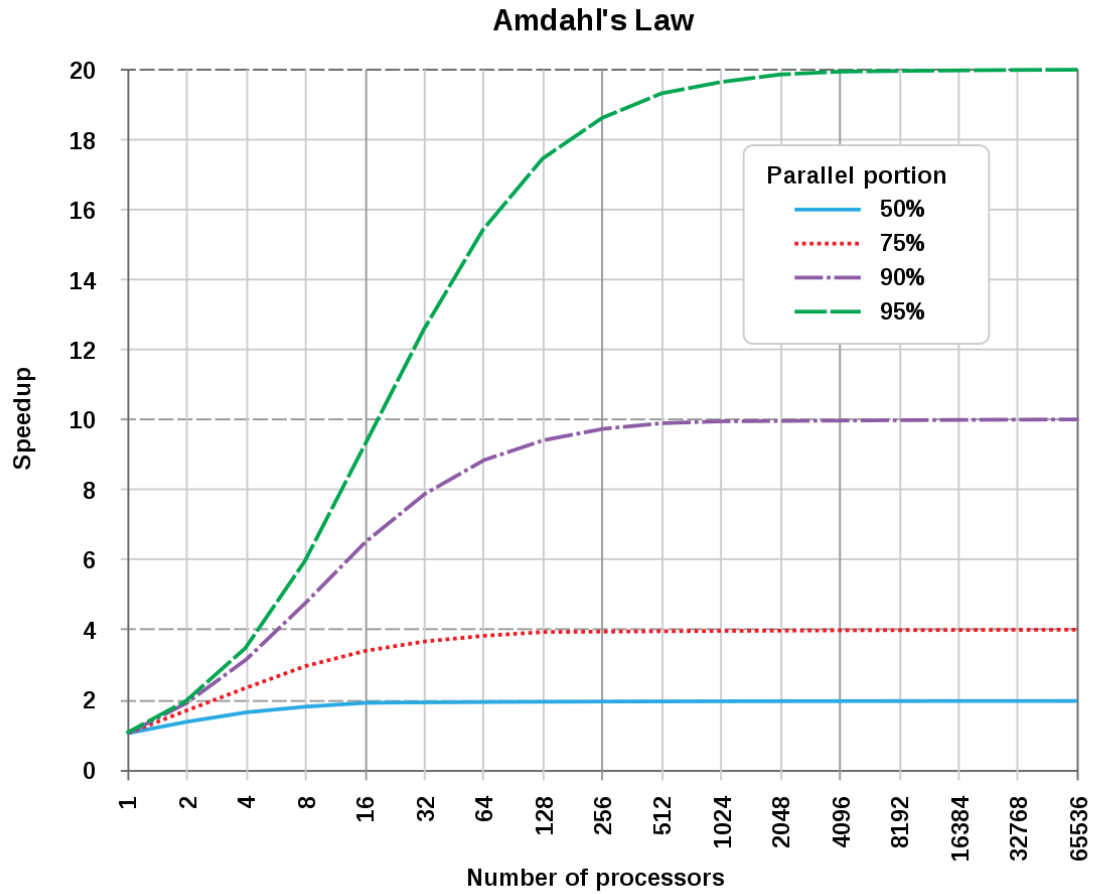
THROUGHPUT MODELING

- How can we predict the maximum throughput **(Req/Sec)** of the system?
- Service Time □ 500 ms on webserver
- Throughput per replica ≈ $\frac{\#\ \mathrm{Th}reads}{Service\ time}$



AMDAHLS LAWN

- How will our maximum throughput change if we add a database
- In general terms, **Amdahl's Law** states that in parallelization, if **P** is the proportion of a system or program that can be made parallel, and **1-P** is the proportion that remains serial, then the maximum speedup **S(N)** that can be achieved using N processors is:
  - **S(N)=1/((1-P)+(P/N))**
- **If 95% of processing is in the web server, then max speedup is 1/1-0.95 = 20X**



Other Models

- Response Time
- Queuing Delay
- Queuing theory
- Performance Models
- ML Models

Prof. Tim Wood & Prof. Roozbeh Haghnazar

Architectures for Performance

Case Study: Microservices