



TCSS 462/562: (SOFTWARE ENGINEERING FOR) CLOUD COMPUTING

Introduction to Cloud Computing - II

Wes J. Lloyd
School of Engineering and Technology
University of Washington - Tacoma



1

OBJECTIVES - 10/17

- **Questions from 10/12**
- Introduction to Cloud Computing II -From book #1 -
Chapter 3: Understanding Cloud Computing
Cloud Computing Concepts, Technology & Architecture
 - Benefits of cloud adoption
 - Risks of cloud adoption
- Background on AWS Lambda for the Term Project
- From Book #1:
Chapter 4: Cloud Computing Concepts and Models
- At the end: Open Discussion on the Term Project
 - Discussion
 - Team Planning

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.2
------------------	---	------

2

OFFICE HOURS – FALL 2023

- **Tuesdays:**
 - 2:30 to 3:30 pm - CP 229
- **Fridays**
 - 11:00 am to 12:00 pm – ONLINE via Zoom
- Or email for appointment

> Office Hours set based on Student Demographics survey feedback

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.3
------------------	---	------

3

ONLINE DAILY FEEDBACK SURVEY

- Daily Feedback Quiz in Canvas – Take After Each Class
- Extra Credit for completing

Announcements

Assignments

Discussions

Zoom

Grades

People

Pages

Files

Quizzes

Collaborations

UW Libraries

UW Resources

▼ Upcoming Assignments

- 📄 Class Activity 1 – Implicit vs. Explicit Parallelism
Available until Oct 11 at 11:59pm | Due Oct 7 at 7:50pm | -/10 pts
- 📄 Tutorial 1 - Linux
Available until Oct 19 at 11:59pm | Due Oct 15 at 11:59pm | -/20 pts

▼ Past Assignments

- 📄 **TCSS 562 - Online Daily Feedback Survey - 10/5**
Available until Dec 18 at 11:59pm | Due Oct 6 at 8:59pm | -/1 pts
- 📄 TCSS 562 - Online Daily Feedback Survey - 9/30
Available until Dec 18 at 11:59pm | Due Oct 4 at 8:59pm | -/1 pts

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.4
------------------	---	------

4

TCSS 562 - Online Daily Feedback Survey - 10/5
Started: Oct 7 at 1:13am
Quiz Instructions

Question 1 0.5 pts

On a scale of 1 to 10, please classify your perspective on material covered in today's class:

1 2 3 4 5 6 7 8 9 10
Mostly Review To Me Equal New and Review Mostly New to Me

Question 2 0.5 pts

Please rate the pace of today's class:

1 2 3 4 5 6 7 8 9 10
Slow Just Right Fast

October 17, 2023 TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma L6.5

5

MATERIAL / PACE

- Please classify your perspective on material covered in today's class (57 respondents):
 - 1-mostly review, 5-equal new/review, 10-mostly new
 - **Average - 6.29 (↓ - previous 6.53)**
- Please rate the pace of today's class:
 - 1-slow, 5-just right, 10-fast
 - **Average - 5.60 (↑ - previous 5.55)**
- **Response rates:**
 - TCSS 462: 37/44 - 84.1%
 - TCSS 562: 20/25 - 80.0%

October 17, 2023 TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma L6.6

6

FEEDBACK FROM 10/12

- *I understand SaaS and IaaS, but I'm not understanding what PaaS and what "platform" means in this context.*
- Platform-as-a-Service is a delivery model where the cloud provider provides a fully hosted platform for specific types of applications (often web apps)
- Often PaaS platforms offer some type of auto-scaling
- Examples:
 - Heroku provides web application hosting for a variety of languages
 - Heroku also provides databases for web applications
 - <https://devcenter.heroku.com/articles/deploying-java>
 - AWS Elastic Beanstalk provides web application hosting on AWS
 - <https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/Welcome.html>
 - https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/create_deploy_Java.html

October 17, 2023

TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L6.7

7

FEEDBACK - 2

- *What is the difference between autonomous cloud systems and self-organizing cloud systems?*
- **Autonomous (in cloud computing)** implies application heuristics, algorithms, and machine learning to automate application (or database) configuration, scaling, tuning, etc. An autonomous platform performs tasks without human intervention
 - Have you ever written code to automate something ?

October 17, 2023

TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L6.8

8

FEEDBACK - 3

- **A self-organizing (cloud) system** not only regulates or adapts its behavior (autonomous), but it creates its own organization with structure with function, Structure means components are arranged in a particular order with connections to integrate parts into a whole, and separations differentiate subsystems to avoid interference.
- Cloud computing systems are argued to be inherently self-organizing, but while they have some autonomy, they likely do not have or achieve the goal of reducing their own complexity
- A self-organizing system goes beyond automation of function to organizing structure as needed (on demand)
 - Structures could be collections of virtual servers and networks organized to satisfy user requests, host a robust application, etc.

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.9
------------------	---	------

9

AWS CLOUD CREDITS UPDATE

- **AWS CLOUD CREDITS ARE NOW AVAILABLE FOR TCSS 462/562**
- Credits provided on request with expiry of Sept 30, 2024
- Credit codes must be securely exchanged
- Request codes by sending an email with the subject “**AWS CREDIT REQUEST**” to wllloyd@uw.edu
- Codes can also be obtained in person (or zoom), in the class, during the breaks, after class, during office hours, by appt
 - All credit requests as of Oct 16 have been distributed
- To track credit code distribution, codes not shared via discord
- 51 students have completed AWS Cloud Credits Survey
 - 18 survey responses missing
- **NEXT:** instructor will work to create IAM user accounts
 - One IAM user request in queue

October 10, 2023	TCSS462/562: (Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L4.10
------------------	--	-------

10

TUTORIAL 0

- Getting Started with AWS
- http://faculty.washington.edu/wlloyd/courses/tcss562/tutorials/TCSS462_562_f2023_tutorial_0.pdf
- Create an AWS account
- Create account credentials for working with the CLI
- Install awsconfig package
- Setup awsconfig for working with the AWS CLI

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.11
------------------	---	-------

11

TUTORIAL 1

- **Introduction to Linux & the Command Line**
- https://faculty.washington.edu/wlloyd/courses/tcss562/tutorials/TCSS462_562_f2023_tutorial_1.pdf
- **Tutorial Sections:**
 1. The Command Line
 2. Basic Navigation
 3. More About Files
 4. Manual Pages
 5. File Manipulation
 6. VI – Text Editor
 7. Wildcards
 8. Permissions
 9. Filters
 10. Grep and regular expressions
 11. Piping and Redirection
 12. Process Management

October 11, 2022	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L4.12
------------------	---	-------

12

TUTORIAL 2

- **Introduction to Bash Scripting**
- https://faculty.washington.edu/wlloyd/courses/tcss562/tutorials/TCSS462_562_f2023_tutorial_2.pdf
- Review tutorial sections:
- Create a BASH webservice client
 1. What is a BASH script?
 2. Variables
 3. Input
 4. Arithmetic
 5. If Statements
 6. Loops
 7. Functions
 8. User Interface
- Call service to obtain IP address & lat/long of computer
- Call weatherbit.io API to obtain weather forecast for lat/long

October 11, 2022

TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L4.13

13

TUTORIAL 3

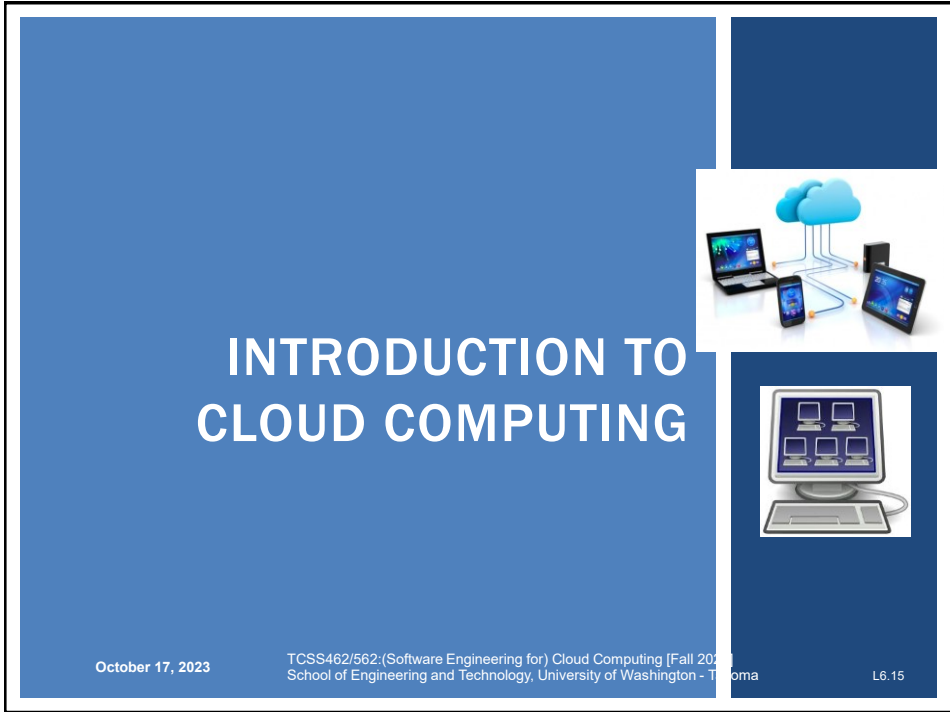
- **Best Practices for Working with Virtual Machines on Amazon EC2**
- http://faculty.washington.edu/wlloyd/courses/tcss562/tutorials/TCSS462_562_f2023_tutorial_3.pdf
- Creating a spot VM
- Creating an image from a running VM
- Persistent spot request
- Stopping (pausing) VMs
- EBS volume types
- Ephemeral disks (local disks)
- Mounting and formatting a disk
- Disk performance testing with Bonnie++
- Cost Saving Best Practices

October 17, 2023

TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L6.14

14

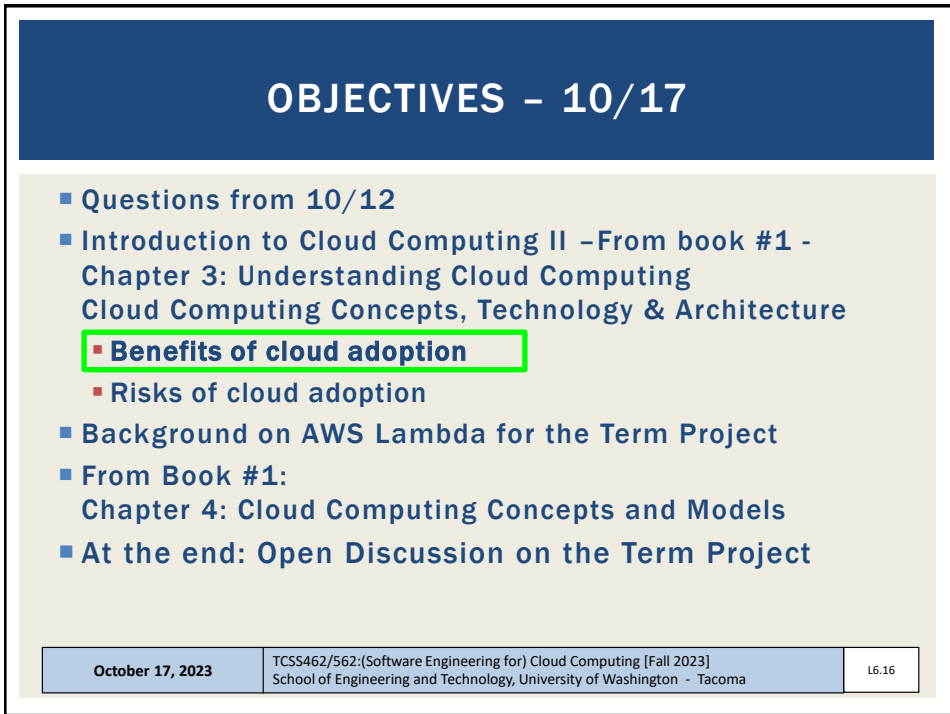


The slide features a large blue background with the title "INTRODUCTION TO CLOUD COMPUTING" in white, bold, uppercase letters. To the right, there are two smaller images: the top one shows a cloud icon connected to various devices (laptop, tablet, smartphone), and the bottom one shows a computer monitor displaying a network diagram. At the bottom, there is a footer with the date "October 17, 2023", the course information "TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma", and the slide number "L6.15".

INTRODUCTION TO CLOUD COMPUTING

October 17, 2023 TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma L6.15

15



The slide has a dark blue header with the title "OBJECTIVES - 10/17" in white, bold, uppercase letters. Below the header, there is a list of objectives in blue text. The objective "Benefits of cloud adoption" is highlighted with a green border. At the bottom, there is a footer with the date "October 17, 2023", the course information "TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma", and the slide number "L6.16".

OBJECTIVES - 10/17

- Questions from 10/12
- Introduction to Cloud Computing II -From book #1 - Chapter 3: Understanding Cloud Computing Cloud Computing Concepts, Technology & Architecture
 - **Benefits of cloud adoption**
 - Risks of cloud adoption
- Background on AWS Lambda for the Term Project
- From Book #1: Chapter 4: Cloud Computing Concepts and Models
- At the end: Open Discussion on the Term Project

October 17, 2023 TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma L6.16

16

GOALS AND BENEFITS

- **Cloud providers**
 - Leverage economies of scale through mass-acquisition and management of large-scale IT resources
 - Locate datacenters to optimize costs where electricity is low
- **Cloud consumers**
 - Key business/accounting difference:
 - **Cloud computing enables anticipated capital expenditures to be replaced with operational expenditures**
 - Operational expenditures always scale with the business
 - Eliminates need to invest in server infrastructure based on anticipated business needs
 - Businesses become more agile and lower their financial risks by eliminating large capital investments in physical infrastructure


October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.17
------------------	---	-------

17

CLOUD BENEFITS - 2

- On demand access to pay-as-you-go resources on a short-term basis (less commitment)
- Ability to acquire “unlimited” computing resources on demand when required for business needs
- Ability to add/remove IT resources at a fine-grained level
- Abstraction of server infrastructure so applications deployments are not dependent on specific locations, hardware, etc.
 - The cloud has made our software deployments more agile...

Before Cloud Computing?



October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.18
------------------	---	-------

18

CLOUD BENEFITS - 3

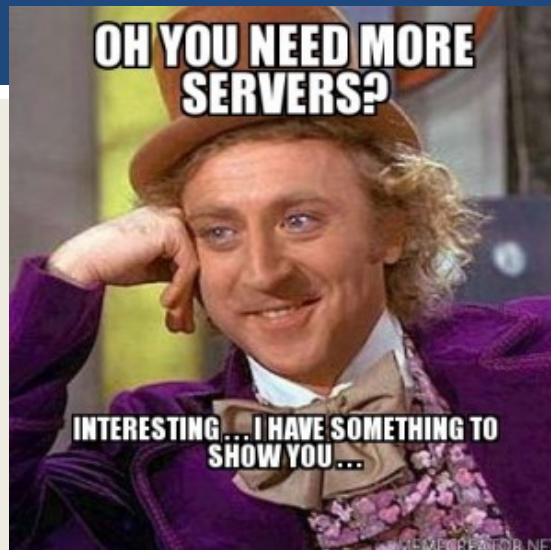
- Example: Using 100 servers for 1 hour costs the same as using 1 server for 100 hours
- Rosetta Protein Folding Use Case: Working with a UW-Tacoma graduate student, we deployed this science model across 5,900 compute cores on Amazon for 2-days...
- **What is the cost to purchase 5,900 compute cores?**
- Recent Dell Server purchase example:
20 cores on 2 servers for \$4,478...
- Using this ratio 5,900 cores costs \$1.3 million (purchase only)

October 17, 2023

TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L6.19

19

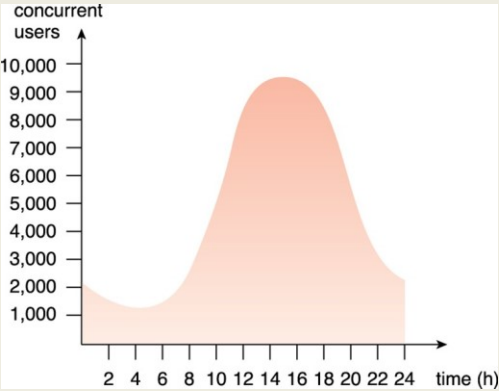


Gene Wilder, Charlie and the Chocolate Factory

20

CLOUD BENEFITS

- Increased scalability
 - Example demand over a 24-hour day →
- Increased availability
- Increased reliability



October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.21
------------------	---	-------

21

OBJECTIVES – 10/17

- Questions from 10/12
- Introduction to Cloud Computing II –From book #1 - Chapter 3: Understanding Cloud Computing
Cloud Computing Concepts, Technology & Architecture
 - Benefits of cloud adoption
 - Risks of cloud adoption
- Background on AWS Lambda for the Term Project
- From Book #1:
Chapter 4: Cloud Computing Concepts and Models
- At the end: Open Discussion on the Term Project

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.22
------------------	---	-------

22

CLOUD ADOPTION RISKS

- **Increased security vulnerabilities**
 - Expansion of trust boundaries now include the external cloud
 - Security responsibility shared with cloud provider
- **Reduced operational governance / control**
 - Users have less control of physical hardware
 - Cloud user does not directly control resources to ensure quality-of-service
 - Infrastructure management is abstracted
 - Quality and stability of resources can vary
 - Network latency costs and variability

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.23
------------------	---	-------

23

NETWORK LATENCY COSTS

The diagram shows two organizational boundaries. On the left, 'Organization A' is enclosed in a dashed box and contains a 'cloud service consumer' (represented by a blue rounded rectangle). Above it is a 'reliable network' (grey box). On the right, 'Cloud A' is enclosed in a dashed box and contains a 'cloud service' (represented by a yellow circle). Above it is another 'reliable network' (grey box). A horizontal line connects the 'cloud service consumer' to the 'cloud service'. A lightning bolt symbol is placed on this line, with a callout box labeled 'unreliable network connection' pointing to it. Below the dashed boxes are labels: 'organizational boundary of cloud consumer' and 'organizational boundary of cloud provider'.

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.24
------------------	---	-------

24

CLOUD RISKS - 2

- **Performance monitoring of cloud applications**
 - Cloud metrics (AWS cloudwatch) support monitoring cloud infrastructure (network load, CPU utilization, I/O)
 - Performance of cloud applications depends on the health of aggregated cloud resources working together
 - User must monitor this aggregate performance
- **Limited portability among clouds**
 - Early cloud systems have significant “vendor” lock-in
 - Common APIs and deployment models are slow to evolve
 - Operating system containers help make applications more portable, but containers still must be deployed
- **Geographical issues**
 - Abstraction of cloud location leads to legal challenges with respect to laws for data privacy and storage

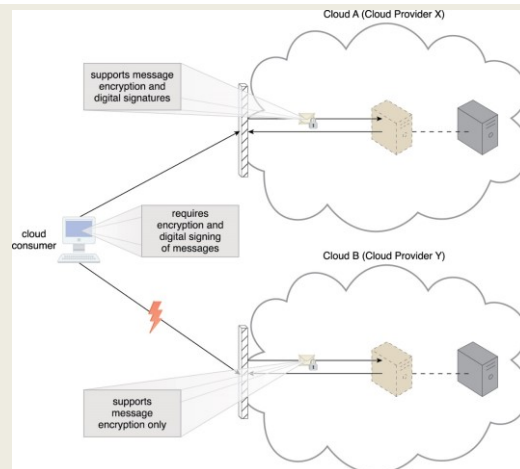
October 17, 2023

TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L6.25

25

CLOUD: VENDOR LOCK-IN



October 17, 2023

TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L6.26

26



OBJECTIVES – 10/17

- Questions from 10/12
- Introduction to Cloud Computing II –From book #1 -
Chapter 3: Understanding Cloud Computing
Cloud Computing Concepts, Technology & Architecture
 - Benefits of cloud adoption
 - Risks of cloud adoption
- **Background on AWS Lambda for the Term Project**
- From Book #1:
Chapter 4: Cloud Computing Concepts and Models
- At the end: Open Discussion on the Term Project

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.27
------------------	---	-------

27

TCSS 462/562 TERM PROJECT



October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.28
------------------	---	-------

28

SERVERLESS – KEY CONCEPTS

- **Function-as-a-Service (FaaS) platform**
 - A platform where developers deploy “functions” written in common languages (e.g. Java, Python, Go, Node.js) that run as microservices
 - AWS Lambda is a FaaS platform
 - We will discuss platform limitations
- **Function Instances**
 - This is an instantiation of a running function
 - A function instance is created when a client (user) calls the serverless function
 - Each concurrent (parallel) call to AWS Lambda to the same function will create a unique function instance to handle the request
 - The default maximum number of concurrently running function instances in your account is 10.
 - The default was originally 1,000 when the platform was introduced, and was dropped to 100, then 50, and is now just 10 in response to the growing popularity of AWS Lambda (they are running out of servers??)
 - You will want to request an increase in your AWS account's default concurrency. A minimum of 100 is recommended

October 17, 2023

TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L6.29

29

AWS LAMBDA

- Lambda functions can be invoked by creating an HTTP REST endpoint that responds to HTTP POST requests
- A json object is provided as a request object to the function
- In the function code, the request object can be accessed to interpret how the user parameterized the function call
- The function generates a JSON response object
- AWS Lambda is introduced in detail in Tutorial 4

October 17, 2023

TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L6.30

30

TYPES OF FUNCTION CALLS: SYNCHRONOUS

- **Serverless Computing:**
 - AWS Lambda supports synchronous and asynchronous function calls
 - Clients typically orchestrate synchronous calls and pipelines
 - Asynchronous calls are often made via events
- **Synchronous web service:**
 - Client calls service
 - Client blocks (freezes) and waits for server to complete call
 - Connection is maintained in the "OPEN" state
 - Problematic if service runtime is long!
 - Connections are notoriously dropped
 - System timeouts reached
 - Client can't do anything while waiting unless using threads

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.31
------------------	---	-------

31

TYPES OF FUNCTION CALLS: ASYNCHRONOUS

- **Asynchronous web service**
 - Client calls service
 - Server responds to client with OK message
 - Client closes connection
 - Server performs the work associated with the service
 - Server posts service result in an external data store
 - AWS: S3, SQS (queueing service), SNS (notification service)

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.32
------------------	---	-------

32

AWS LAMBDA PLATFORM LIMITATIONS

- Maximum 10 GB memory per function instance
- Maximum 15-minutes execution per function instance
- 500 MB of /tmp disk space for local I/O (default)
- Up to 10 GB /tmp ephemeral storage (for additional charge)
 - <https://aws.amazon.com/blogs/aws/aws-lambda-now-supports-up-to-10-gb-ephemeral-storage/>
- Access up to 6 vCPUs depending on memory reservation size

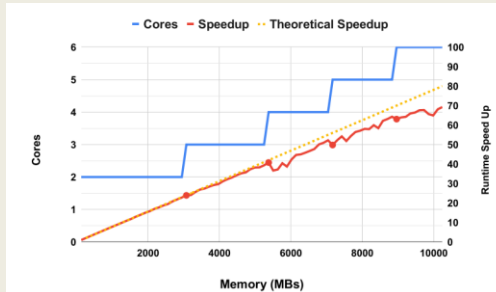


Figure 1: AWS Lambda Performance Speedup for Sysbench Prime Number Generation vs. Function Memory

October 17, 2023

TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L6.33

33

AWS LAMBDA PLATFORM LIMITATIONS - 2

- 10 concurrent function executions inside account (default)
- Function payload: 6MB (synchronous), 256KB (asynchronous)
- Deployment package: 50MB (compressed), 250MB (unzipped)
- Container image size: 10 GB
- Processes/threads: 1024
- File descriptors: 1024
- Function instances run Amazon Linux 2
 - Pending upgrade to Amazon Linux 2023 ?
- See: <https://docs.aws.amazon.com/lambda/latest/dg/gettingstarted-limits.html>


October 17, 2023

TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L6.34

34

CPUSTEAL



- *CpuSteal*: Metric that measures when a CPU core is ready to execute but the physical CPU core is busy and unavailable
- Symptom of over provisioning physical servers in the cloud
- Factors which cause *CpuSteal*: (x86 hyperthreading)
 1. Physical CPU is shared by too many busy VMs
 2. Hypervisor kernel is using the CPU
 - On AWS Lambda this would be the Firecracker MicroVM which is derived from the KVM hypervisor
 3. VM's CPU time share <100% for 1 or more cores, and 100% is needed for a CPU intensive workload.
- Man procs - press "/" - type "proc/stat"
 - CpuSteal is the 8th column returned
 - Metric can be read using SAAF in tutorial #4

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.35
------------------	---	-------

35

EFFECTS OF SCALING FUNCTION MEMORY ON CPU TIME SHARE

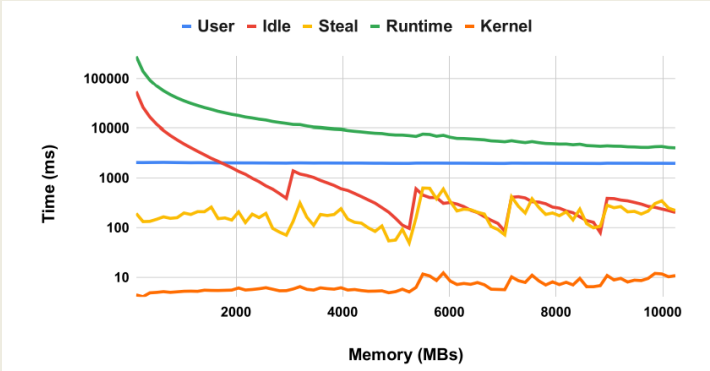


Figure 2: Linux CPU Utilization (log scale) vs. Function Memory for Sysbench Prime Number Generation

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.36
------------------	---	-------

36

EFFECTS OF SCALING FUNCTION MEMORY ON CPU TIME SHARE

Key observations:

- Runtime decreases as vCPUs and CPU time share increase
- CPU user time remains constant for the prime number generation task – work doesn't change
- CPU idle time gradually decreases as memory and vCPUs increase (the idle time is becoming active time)
- When the 4th vCPU is added, cpuSteal tracks closely with cpuidle time (hyperthreading effect)
- There is more cpu Kernel time after the 4th vCPU is added

October 17, 2023 TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma L6.37

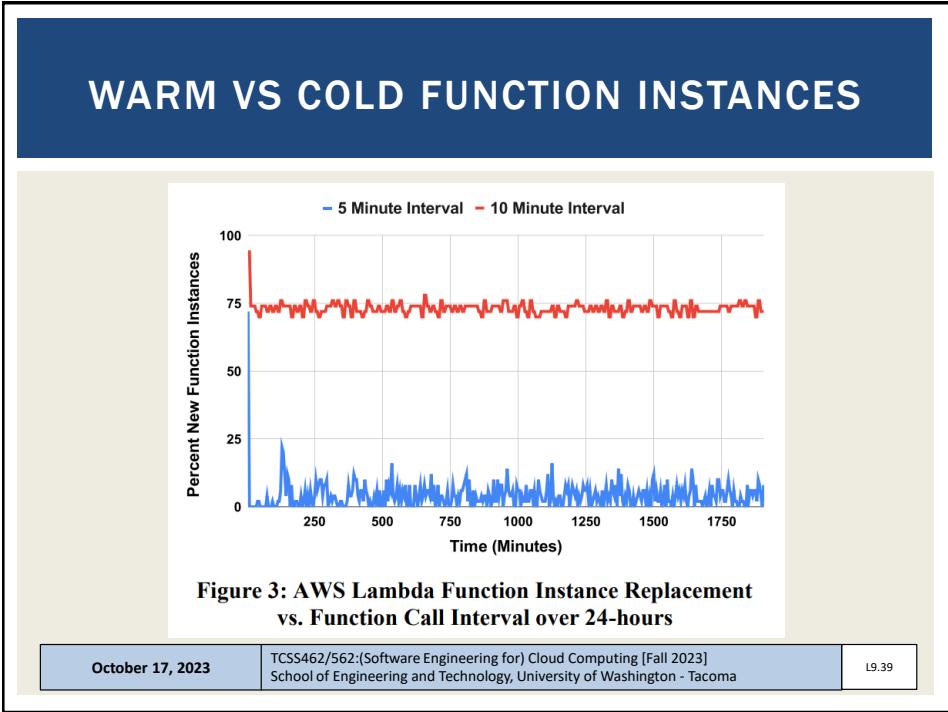
37

FUNCTION INSTANCE LIFE CYCLES

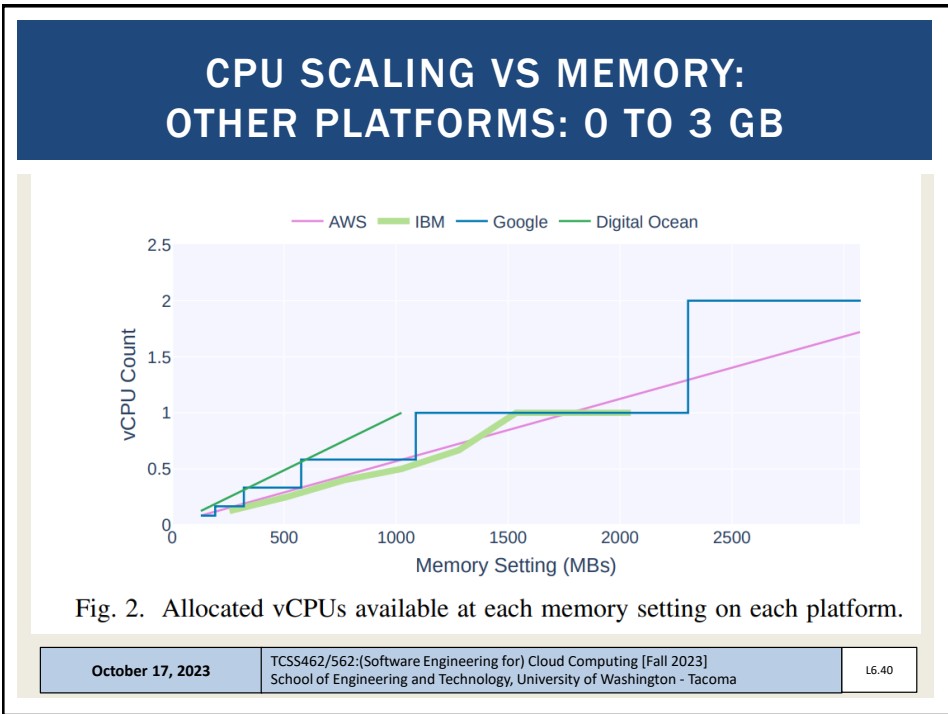
- **Function states:**
- **COLD:** brand new function instance just initialized to run the request (more overhead)
 - Platform cold (first time ever run)
 - Host cold (function assets cached locally on servers)
- **WARM:** existing function instance that is reused
- All function instances persist for ~5 minutes before they begin to be “garbage collected” by the platform
 - 100% garbage collection may take up to ~30-40 minutes
- AWS Lambda appears to “recycle” infrastructure faster than other FaaS platforms
 - Presumably because of need, because the platform is busy

October 17, 2023 TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma L6.38

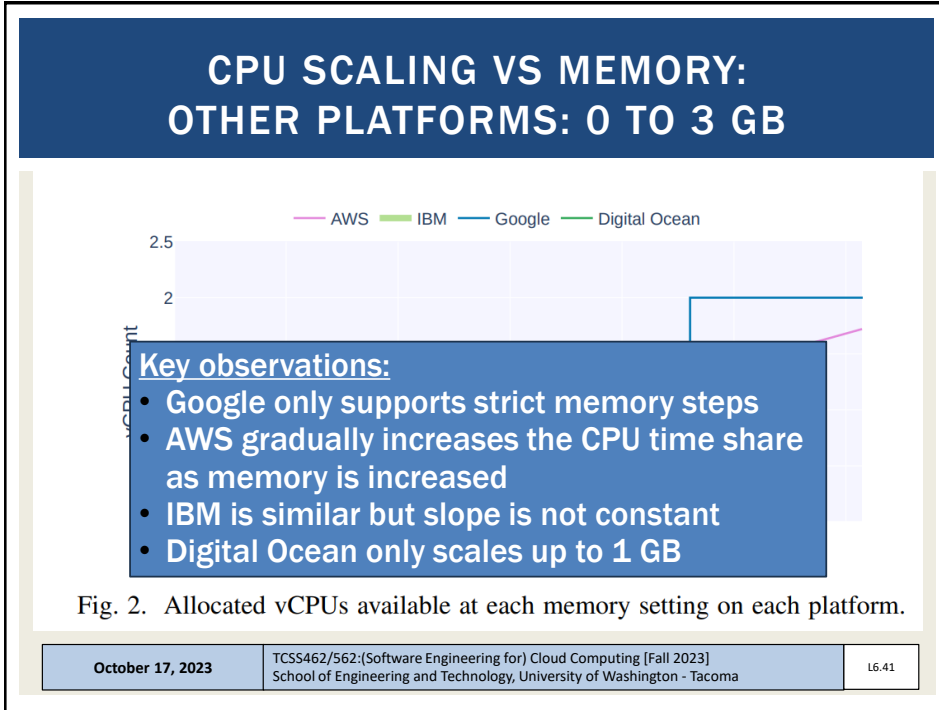
38



39



40



41

ELASTIC FILE SYSTEM (AWS EFS)

- Traditionally AWS Lambda functions have been limited to 500MB of storage space
- Recently the Elastic File System (EFS) has been extended to support AWS Lambda
- The Elastic File System supports the creation of a shared volume like a shared disk (or folder)
 - EFS is similar to NFS (network file share)
 - Multiple AWS Lambda functions and/or EC2 VMs can mount and share the same EFS volume
 - Provides a shared R/W disk
 - Breaks the 500MB capacity barrier on AWS Lambda
- **Downside: EFS is expensive: ~30 \$/GB/month**
- **Project: EFS performance & scalability evaluation on Lambda**

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.42
------------------	---	-------

42

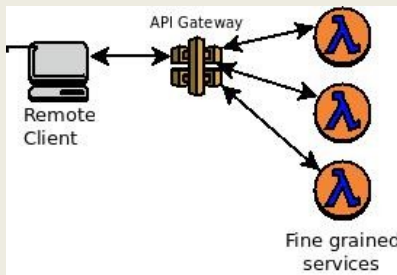
SERVERLESS FILE STORAGE COMPARISON PROJECT

- **Elastic File System (EFS):**
 Performance, Cost, and Scalability Evaluation in the context of AWS Lambda / Serverless Computing
 - EFS provides a file system that can be shared with multiple Lambda function instances in parallel
- **Using a common use case, compare performance and cost of extended storage options on AWS Lambda:**
 - Docker container support (up to 10 GB) – read only
 - Ephemeral /tmp (up to 10 GB) – read/write
 - EFS (unlimited, but costly) – read/write
 - image integration with AWS Lambda – performance & scalability

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.43
------------------	---	-------

43

SERVICE COMPOSITION



Remote Client API Gateway Fine grained services

A	B	C	3 services Full Service Isolation
A	B	C	2 services
A	B	C	2 services
A B C			1 service Full Service Aggregation

Other possible compositions: group by library, functional cohesion, etc.

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.44
------------------	---	-------

44

SWITCH-BOARD ARCHITECTURE

- Single deployment package with consolidated codebase (Java: one JAR file)**
- Entry method contains "switchboard" logic**
 Case statement that route calls to proper service
- Routing is based on data payload**
 Check if specific parameters exist, route call accordingly
- Goal: reduce # of COLD starts to improve performance**

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.45
------------------	---	-------

45

APPLICATION FLOW CONTROL - 3

Client flow control

(a) Microservices

Microservice as controller

(c) Microservices

AWS Step Function

(b) Microservices

Asynchronous

(d)

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.46
------------------	---	-------

46

PROGRAMMING LANGUAGE COMPARISON

- FaaS platforms support hosting code in multiple languages
- AWS Lambda- common: Java, Node.js, Python
 - Plus others: Go, PowerShell, C#, and Ruby
- Also Runtime API (“BASH”) which allows deployment of binary executables from any programming language
- August 2020 – Our group’s paper:
 - <https://tinyurl.com/y46eq6np>
- If wanting to perform a language study either:
 - Implement in C#, Ruby, or multiple versions of Java, Node.js, Python
 - OR implement different app than TLQ (ETL) data processing pipeline

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.47
------------------	---	-------

47

FAAS PLATFORMS

- Many commercial and open source FaaS platforms exist
- TCSS562 projects can choose to compare performance and cost implications of alternate platforms.

- Supported by SAAF:
 - AWS Lambda
 - Google Cloud Functions
 - Azure Functions
 - IBM Cloud Functions
 - Apache OpenWhisk (*open source, deploy your own FaaS*)
 - Open FaaS (*open source, deploy your own FaaS*)

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.48
------------------	---	-------

48

DATA PROVISIONING

- Consider performance and cost implications of the data-tier design for the serverless application
- Use different tools as the relational datastore to support service #2 (LOAD) and service #3 (EXTRACT)
- **SQL / Relational:**
 - Amazon Aurora (serverless cloud DB), Amazon RDS (cloud DB), DB on a VM (MySQL), DB inside Lambda function (SQLite, Derby)
- **NO SQL / Key/Value Store:**
 - Dynamo DB, MongoDB, S3

October 17, 2023

TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L6.49

49

PERFORMANCE VARIABILITY

- Cloud platforms exhibit performance variability which varies over time
- Goal of this case study is to measure performance variability (i.e. extent) for AWS Lambda services by hour, day, week to look for common patterns
- Can also examine performance variability by availability zone and region
 - Do some regions provide more stable performance?
 - Can services be switched to different regions during different times to leverage better performance?
- Remember that performance = cost
- If we make it faster, we make it cheaper...

October 17, 2023

TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L6.50

50

CPU STEAL CASE STUDY

- On AWS Lambda (or other FaaS platforms), when we run functions, how much CpuSteal do we observe?
- How does CpuSteal vary for different workloads? (e.g. functions that have different resource requirements)
- How does CpuSteal vary over time hour, day, week, location?
- How does CpuSteal relate to function performance?

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.51
------------------	---	-------

51

CLOUD COMPUTING: CONCEPTS AND MODELS



October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.52
------------------	---	-------

52

OBJECTIVES - 10/17

- From: Cloud Computing Concepts, Technology & Architecture: Chapter 4: Cloud Computing Concepts and Models:
 - **Roles and boundaries**
 - Cloud characteristics
- **At the end: Open Discussion on the Term Project**
 - Discussion
 - Team Planning

October 17, 2023	TCSS562:Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L6.53
------------------	---	-------

53

ROLES

- **Cloud provider**
 - Organization that provides cloud-based resources
 - Responsible for fulfilling SLAs for cloud services
 - Some cloud providers “resell” IT resources from other cloud providers
 - Example: Heroku sells PaaS services running atop of Amazon EC2
- **Cloud consumers**
 - Cloud users that consume cloud services
- **Cloud service owner**
 - Both cloud providers and cloud consumers can own cloud services
 - A cloud service owner may use a cloud provider to provide a cloud service (e.g. Heroku)

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.54
------------------	---	-------

54

ROLES - 2

- **Cloud resource administrator**
 - Administrators provide and maintain cloud services
 - Both cloud providers and cloud consumers have administrators
- **Cloud auditor**
 - Third-party which conducts independent assessments of cloud environments to ensure security, privacy, and performance.
 - Provides unbiased assessments
- **Cloud brokers**
 - An intermediary between cloud consumers and cloud providers
 - Provides service aggregation
- **Cloud carriers**
 - Network and telecommunication providers which provide network connectivity between cloud consumers and providers

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.55
------------------	---	-------

55

ORGANIZATION BOUNDARY

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.56
------------------	---	-------

56

TRUST BOUNDARY

trust boundary

Organization A Cloud A

cloud service consumer cloud service

organizational boundary organizational boundary

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.57
------------------	---	-------

57

OBJECTIVES – 10/17

- From: Cloud Computing Concepts, Technology & Architecture: Chapter 4: Cloud Computing Concepts and Models:
 - Roles and boundaries
 - **Cloud characteristics**
- **At the end: Open Discussion on the Term Project**
 - Discussion
 - Team Planning

October 17, 2023	TCSS562:Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L6.58
------------------	---	-------

58

CLOUD CHARACTERISTICS

- On-demand usage
- Ubiquitous access
- Multitenancy (resource pooling)
- Elasticity
- Measured usage
- Resiliency

■ Assessing these features helps measure the value offered by a given cloud service or platform

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.59
------------------	---	-------

59

ON-DEMAND USAGE

- The freedom to self-provision IT resources
- Generally, with automated support
- Automated support requires no human involvement
- Automation through software services interface

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.60
------------------	---	-------

60

UBIQUITOUS ACCESS

- Cloud services are widely accessible
- Public cloud: internet accessible
- Private cloud: throughout segments of a company's intranet
- 24/7 availability

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.61
------------------	---	-------

61

MULTITENANCY

- Cloud providers pool resources together to share them with many users
- Serve multiple cloud service consumers
- IT resources can be dynamically assigned, reassigned based on demand
- Multitenancy can lead to performance variation

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.62
------------------	---	-------

62

SINGLE TENANT MODEL

> Isolation <

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.63
------------------	---	-------

63

MULTITENANT MODEL

- Resource is “multiplexed” and share amongst multiple users
- Goal is to increase utilization
- Often server resources are underutilized
- There are many “sunk costs” whether usage is 0% or 100%
- Cloud computing tries to maximize “sunk cost” investments through **multi-tenancy**


shared cloud storage device

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.64
------------------	---	-------

64

MULTITENANT DATABASE

Isolated

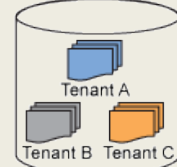


Tenant A
Tenant B Tenant C

Separate database

E1

Semi-shared

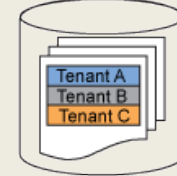


Tenant A
Tenant B Tenant C

**Shared database
Separate schema**

E2

Shared



Tenant A
Tenant B
Tenant C

**Shared database
Shared schema**

E3

- Many users on a single database instance
- What issues may occur when sharing a single database instance?**

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.65
------------------	---	-------

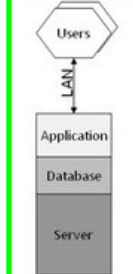
65

MULTITENANCY OF RESOURCES

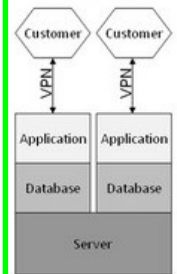
■ Where is the multitenancy?

■ >> What is shared? What is isolated?

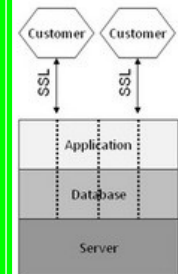
Traditional On Premise



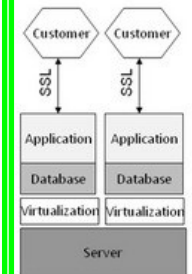
Single Tenant (Hosted)



Multi-Tenant



Virtual Appliance



October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.66
------------------	---	-------

66

RESOURCE CONTENTION FROM MUTLI-TENANCY

- Despite best efforts at isolation, co-resident VMs on a single cloud server running identical benchmarks simultaneously do not perform equally.

From Han, X., Schooley, R., Mackenzie, D., David, O., Lloyd, W., Characterizing Public Cloud Resource Contention to Support Virtual Machine Co-residency Prediction, 2020 8th IEEE International Conference on Cloud Engineering (IC2E 2020), Apr 21-24, 2020.

VM Tenants	sysbench (CPU)	y-cruncher (CPU)	pgbench (CPU + I/O)	iperf (network I/O)
0	100%	100%	100%	100%
10	95%	90%	92%	40%
20	90%	85%	88%	25%
30	85%	80%	82%	18%
40	80%	75%	78%	15%
48	75%	70%	72%	12%

Up to 48 VMs sharing same server !!

October 17, 2023
TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma
L6.67

67

RESOURCE CONTENTION FROM MUTLI-TENANCY - 2

- Performance variation from multi-tenancy is increasing as cloud servers add more CPU cores
- Running many idle operating system instances can impose significant overhead for some workloads

Maximum potential resource contention (i.e. worst-case scenario) →

From Han, X., Schooley, R., Mackenzie, D., David, O., Lloyd, W., Characterizing Public Cloud Resource Contention to Support Virtual Machine Co-residency Prediction, 2020 8th IEEE International Conference on Cloud Engineering (IC2E 2020), Apr 21-24, 2020.

EC2 Instance Family	y-cruncher (CPU)	sysbench (CPU)	pgbench (CPU + I/O)	iperf (network)	Total Variance (%)
c3	19.2%	0.3%	84.6%	19.2%	19.2%
c4	0.2%	0.6%	38.1%	42.1%	42.1%
z1d	0.2%	0.6%	8.2%	84.6%	84.6%
m5d (t)	48.0%	20.8%	33.0%	94.6%	94.6%

October 17, 2023
TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma
L6.68

68

ELASTICITY

- Automated ability of cloud to transparently scale resources
- Scaling based on runtime conditions or pre-determined by cloud consumer or cloud provider
- Threshold based scaling
 - CPU-utilization > threshold_A, Response_time > 100ms
 - Application agnostic vs. application specific thresholds
 - Why might an application agnostic threshold be non-ideal?
- Load prediction
 - Historical models
 - Real-time trends

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.69
------------------	---	-------

69

PREDICTABLE DEMAND

- AWS EC2 Scaling Example:

Auto-Scaling Example: Netflix

From: Kejarjwal, A., 2013, March. Techniques for optimizing cloud footprint. In 2013 IEEE Int. Conf. on Cloud Engineering (IC2E), pp. 258-268.

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.70
------------------	---	-------

70

MEASURED USAGE

- Cloud platform tracks usage of IT resources
- For billing purposes
- Enables charging only for IT resources actually used
- Can be time-based (millisec, second, minute, hour, day)
 - Granularity is increasing...
- Can be throughput-based (data transfer: MB/sec, GB/sec)
- Can be resource/reservation based (vCPU/hr, GB/hr)

- Not all measurements are for billing
- Some measurements can support auto-scaling
- For example CPU utilization

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.71
------------------	---	-------

71

EC2 CLOUDWATCH METRICS

EC2 Instance: i-1267037f

Description Monitoring Tags

Graphs are for 1 instance that has monitoring enabled. Times are displayed in UTC. Time Range: Last Hour Refresh

Avg CPU Utilization (Percent)

Avg Disk Reads (Bytes)

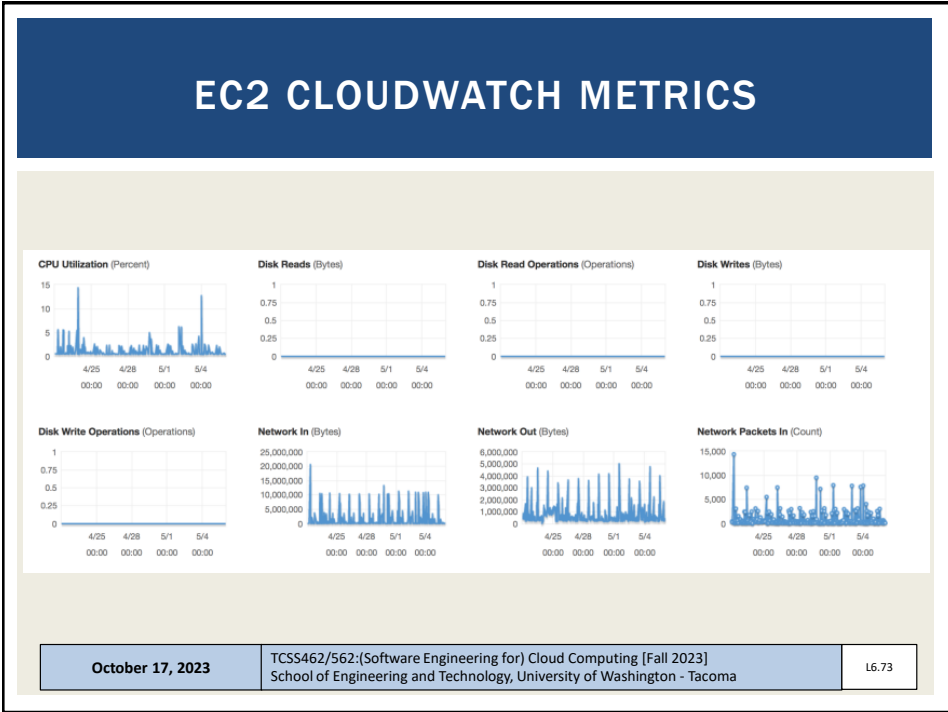
Avg Disk Writes (Bytes)

Max Network In (Bytes)

Max Network Out (Bytes)

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.72
------------------	---	-------

72



73

RESILIENCY

- Distributed redundancy across physical locations (regions on AWS)
- Used to improve reliability and availability of cloud-hosted applications
- Very much an engineering problem
- No “resiliency-as-a-service” for user deployed apps
- Unique characteristics of user applications make a one-size fits all service solution challenging

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.74
------------------	---	-------

74


OBJECTIVES – 10/17

- Questions from 10/12
- Introduction to Cloud Computing II –From book #1 -
Chapter 3: Understanding Cloud Computing
Cloud Computing Concepts, Technology & Architecture
 - Benefits of cloud adoption
 - Risks of cloud adoption
- Background on AWS Lambda for the Term Project
- From Book #1:
Chapter 4: Cloud Computing Concepts and Models
- **At the end: Open Discussion on the Term Project**
 - Discussion
 - Team Planning

October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.75
------------------	---	-------

75

TCSS 462/562 TERM PROJECT



October 17, 2023	TCSS462/562:(Software Engineering for) Cloud Computing [Fall 2023] School of Engineering and Technology, University of Washington - Tacoma	L6.76
------------------	---	-------

76

QUESTIONS

October 17, 2023

TCSS462/562: (Software Engineering for) Cloud Computing [Fall 2023]
School of Engineering and Technology, University of Washington - Tacoma

L6.77

77