

TCSS 562:
SOFTWARE ENGINEERING
FOR CLOUD COMPUTING

Cloud Computing –
How did we get here?

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

1

OBJECTIVES – 10/5

■ Questions from 9/30

■ Cloud Computing – How did we get here?
(Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition)

■ Data, thread-level, task-level parallelism & Parallel architectures

■ Class Activity 1 – Implicit vs Explicit Parallelism

■ SIMD architectures, vector processing, multimedia extensions

■ Graphics processing units

■ Speed-up, Amdahl's Law, Scaled Speedup

■ Properties of distributed systems

■ Modularity

■ Introduction to Cloud Computing – loosely based on book #1:
Cloud Computing Concepts, Technology & Architecture

October 5, 2021

TCSS562:Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.2

2

MATERIAL / PACE

- Please classify your perspective on material covered in today’s class (26 respondents):
 - 1-mostly review, 5-equal new/review, 10-mostly new
 - **Average – 6.15** (5.94, Fall 2021)
- Please rate the pace of today’s class:
 - 1-slow, 5-just right, 10-fast
 - **Average – 5.19** (5.5, Fall 2021)

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.3
-----------------	--	------

3

FEEDBACK FROM 9/30

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.4
-----------------	--	------

4

OBJECTIVES – 10/5

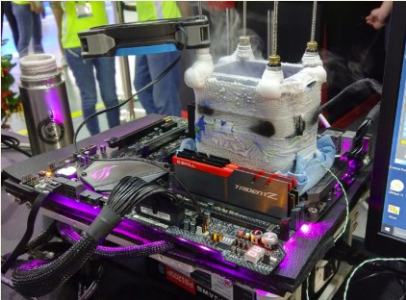
- Questions from 9/30
- Cloud Computing – How did we get here?
(Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition)
- Data, thread-level, task-level parallelism & Parallel architectures
- Class Activity 1 – Implicit vs Explicit Parallelism
- SIMD architectures, vector processing, multimedia extensions
- Graphics processing units
- Speed-up, Amdahl's Law, Scaled Speedup
- Properties of distributed systems
- Modularity
- Introduction to Cloud Computing – loosely based on book #1:
Cloud Computing Concepts, Technology & Architecture

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.5
-----------------	--	------

5

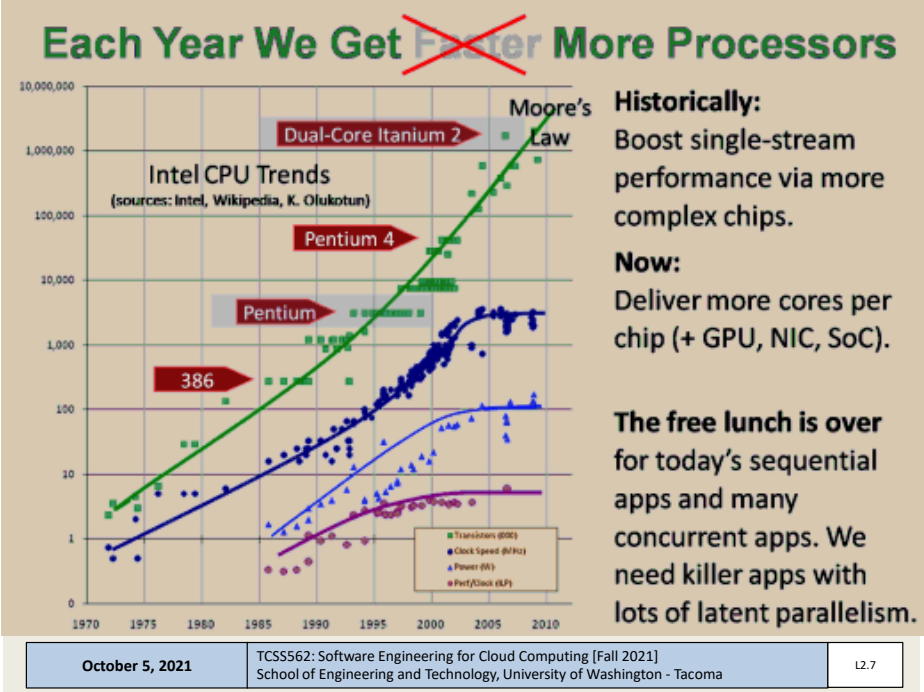
CLOUD COMPUTING:
HOW DID WE GET HERE?

- General interest in parallel computing
 - Moore's Law - # of transistors doubles every 18 months
 - Post 2004: heat dissipation challenges:
can no longer easily increase cloud speed
 - Overclocking to 7GHz takes more than just liquid nitrogen:
 - <https://tinyurl.com/y93s2yz2>
- Solutions:
 - Vary CPU clock speed
 - Add CPU cores
 - Multi-core technology



October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.6
-----------------	--	------

6



7

AMD'S 64-CORE 7NM CPUS

- Epyc Rome CPUs
- Announced August 2019
- EPYC 7H12 requires liquid cooling

AMD EPYC 7002 Processors (2P)						
	Cores Threads	Frequency (GHz)		L3*	TDP	Price
		Base	Max			
EPYC 7H12	64 / 128	2.60	3.30	256 MB	280 W	?
EPYC 7742	64 / 128	2.25	3.40	256 MB	225 W	\$6950
EPYC 7702	64 / 128	2.00	3.35	256 MB	200 W	\$6450
EPYC 7642	48 / 96	2.30	3.20	256 MB	225 W	\$4775
EPYC 7552	48 / 96	2.20	3.30	192 MB	200 W	\$4025

October 5, 2021 TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma L2.8

8

HYPER-THREADING

- Modern CPUs provide multiple instruction pipelines, supporting multiple execution threads, usually 2 to feed instructions to a single CPU core...
- Two hyper-threads are not equivalent to (2) CPU cores
- i7-4770 and i5-4760 same CPU, with and without HTT
- Example: → hyperthreads add +32.9%

4770 with HTT Vs. 4670 without HTT - 25% improvement w/ HTT

CPU Mark Relative to Top 10 Common CPUs
As of 7th of February 2014 - Higher results represent better performance

CPU	Score
Intel Core i7-4770 @ 3.40GHz	9,985
Intel Core i7-3770K @ 3.50GHz	9,542
Intel Core i7-3770 @ 3.40GHz	9,419
AMD FX-8350 Eight-Core	9,051
Intel Core i7-3820 @ 3.60GHz	9,015
Intel Core i7-2600K @ 3.40GHz	8,593
Intel Core i7-2600 @ 3.40GHz	8,316
AMD FX-8320 Eight-Core	8,121
Intel Core i5-4670 @ 3.40GHz	7,513

PassMark Software © 2008-2014

October 5, 2021

TCCS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.9

9

HYPER-THREADING - 2

- How do I use hyper-threading?
- Hyper-threading is automatic
- Modern CPUs expose each physical CPU core as two CPU cores
- `cat /proc/cpuinfo` command lists individual cores
- Operating system schedules processes & threads to run on a hyper-thread
- On CPUs with hyper-threading, each CPU core has two hyper-threads
- To the operating system they are seen as full-featured independent CPU cores

October 5, 2021

TCCS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.10

10

CAT /PROC/CPUINFO

```
wlloyd@dlone:~/Dropbox/courses/tcss562$ cat /proc/cpuinfo | grep -C 20 ht
processor       : 0
vendor_id      : GenuineIntel
cpu family     : 6
cpu model      : 94
model name     : Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz
stepping       : 3
microcode      : 0xdc
cpu MHz        : 840.023
cache size     : 6144 KB
physical id    : 0
siblings       : 8
core id        : 0
cpu cores      : 4
apicid         : 0
initial apicid : 0
fpu            : yes
fpu_exception  : yes
cpuid level    : 22
wp             : yes
flags           : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts acpi mmx
fxsr sse sse2 ss ht tm pbe syscall nx pdpe1gb rdtscp lm constant_tsc art arch perfmon pebs bts rep_good nopl xt
opology nonstop_tsc aperfmperf pni pclmulqdq dtes64 monitor ds_cpl vmx est tm2 ssse3 sdbg fma cx16 xtpr pdcm pc
id sse4_1 sse4_2 x2apic movbe popcnt tsc_deadline_timer aes xsave avx f16c rdrand lahf_lm abm 3dnowprefetch epb
invtscid single_intel_pt ssbd ibrs ibpb stibp kaiser tpr_shadow vnmi flexpriority ept vpid fsgsbase tsc_adjust
bmi1 hle avx2 smep bmi2 erms invpcid rtm mpx rdseed adx snap clflushopt xsaveopt xsavec xgetbv1 dtherm ida arat
pln pts hwp hwp_notify hwp_act_window hwp_epp md_clear flush_lid
bugs           : cpu_meltdown spectre_v1 spectre_v2 spec_store_bypass l1tf mds swapgs taa itlb_multihit srbds
bogomips       : 5184.46
clflush size   : 64
cache_alignme  : 64
address sizes  : 39 bits physical, 48 bits virtual
power managem  :
```

If a CPU has hyper-threading enabled, the “ht” flag is listed

11

Hyper-Threading (HT) Technology

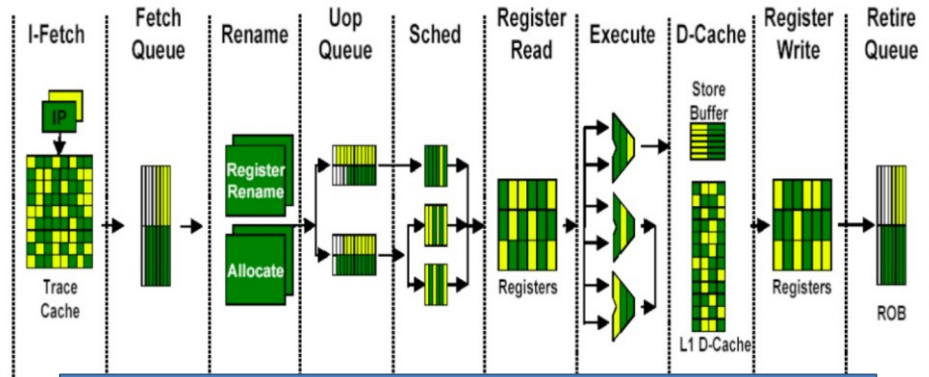
- Provides more satisfactory solution
- Single physical processor is shared as two logical processors
- Each logical processor has its own architecture state
- Single set of execution units are shared between logical processors
- N-logical PUs are supported
- Have the same gain % with only 5% die-size penalty.
- HT allows single processor to fetch and execute two separate code streams simultaneously.

Figure 2: Processors without Hyper-Threading Tech

Figure 3: Processors with Hyper-Threading Technology

12

Execution Pipeline



Each processor core consists of multiple stages

Hyper-threading is the idea to share the physical stages of a CPU core to execute two instructions at once

13

13

HYPER-THREADING - 3

■ When should we use hyper-threading, and when should not?

- For personal computing, hyper-threading helps improve system performance when many programs use only short bursts of CPU time
- Databases, HPC (science) applications, and others may benefit from disabling hyper-threading. Testing will help quantify performance.
- Disabling hyper-threading (HW setting), cuts the number of CPU cores available to operating system in half
 - Can be disabled in the System BIOS or UEFI (uniform extensible firmware interface) software
 - BIOS / UEFI is a small resident program that can be accessed by pressing a function-key when rebooting the computer
 - BIOS / UEFI is used to configure hardware options
 - Making changes requires rebooting the computer

October 5, 2021

TCCS562: Software Engineering for Cloud Computing [Fall 2021]
 School of Engineering and Technology, University of Washington - Tacoma

L10.14

14

CLOUD COMPUTING: HOW DID WE GET HERE? - 2

- To make computing faster, we must go “parallel”
- Difficult to expose parallelism in scientific applications
- Not every problem solution has a parallel algorithm
 - Chicken and egg problem...
- Many commercial efforts promoting pure parallel programming efforts have failed
- Enterprise computing world has been *skeptical* and less involved in parallel programming

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.15

15

CLOUD COMPUTING: HOW DID WE GET HERE? - 3

- **Cloud computing** provides access to “infinite” scalable compute infrastructure on demand
- Infrastructure availability is key to exploiting parallelism
- **Cloud applications**
 - Based on client-server paradigm
 - Thin clients leverage compute hosted on the cloud
 - Applications run many web service instances
 - Employ load balancing

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.16


16

CLOUD COMPUTING:
HOW DID WE GET HERE? - 4

- **Big Data** requires massive amounts of compute resources
- **MAP – REDUCE**
 - Single instruction, multiple data (SIMD)
 - Exploit data level parallelism
- **Bioinformatics example**

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.17
-----------------	--	-------

17

SMITH WATERMAN USE CASE

- Applies dynamic programming to find best local alignment of two protein sequences
 - Embarrassingly parallel, each task can run in isolation
 - Use case for GPU acceleration
- **AWS Lambda Serverless Computing Use Case:**
 - **Goal:** Pair-wise comparison of all unique human protein sequences (20,336)
 - Python client as scheduler
 - C Striped Smith-Waterman (SSW) execution engine

From: Zhao M, Lee WP, Garrison EP, Marth GT: SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. PLoS One 2013, 8:e82138

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.18
-----------------	--	-------

18

SMITH WATERMAN RUNTIME

- Laptop server and client (2-core, 4-HT): 8.7 hours
- AWS Lambda FaaS, laptop as client: 2.2 minutes
 - Partitions 20,336 sequences into 41 sets
 - Execution cost: ~ 82¢ (~237x speed-up)
- AWS Lambda server, EC2 instance as client: 1.28 minutes
 - Execution cost: ~ 87¢ (~408x speed-up)
- Hardware
 - Laptop client: Intel i5-7200U 2.5 GHz :4 HT, 2 CPU
 - Cloud client: EC2 Virtual Machine - m5.24xlarge: 96 vCPUs
 - Cloud server: Lambda ~1000 x Intel E5-2666v3 2.9GHz CPUs

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.19
-----------------	--	-------

19

CLOUD COMPUTING:
HOW DID WE GET HERE? - 5

- Compute clouds are large-scale distributed systems
 - Heterogeneous systems
 - Many services/platforms w/ diverse hw + capabilities
 - Homogeneous systems
 - Within a platform – illusion of identical hardware
 - Autonomous
 - Automatic management and maintenance- largely with little human intervention
 - Self organizing
 - User requested resources organize themselves to satisfy requests on-demand

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.20
-----------------	--	-------

20

CLOUD COMPUTING: HOW DID WE GET HERE? - 6

- Compute clouds are large-scale distributed systems
- Infrastructure-as-a-Service (IaaS) Cloud
 - Provide VMs on demand to users
 - *ec2instances.info* (AWS EC2)
- Clouds can consist of
 - Homogeneous hardware (servers, etc.)
 - Heterogeneous hardware (servers, etc.)
- Which is preferable?

September 25, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.21

21

HARDWARE HETEROGENEITY



- If providing IaaS, what are advantages/disadvantages of using homogeneous hardware?
 - Easier to provide same quality of service to end users
 - Less performance variance
 - Components with variable performance: CPUs, memory (speed differences), disks (SSDs, HDDs), network interfaces (caches?)
 - Homogeneous hardware (servers): components are interchangeable
 - As components fail, identical backups are immediately available
 - Example: blade servers
 - As clouds grow, why is HW homogeneity difficult to maintain?
- What are some advantages of using heterogeneous HW?

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.22

22

OBJECTIVES – 10/5

- Questions from 9/30
- Cloud Computing – How did we get here?
(Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition)
- Data, thread-level, task-level parallelism & Parallel architectures
- Class Activity 1 – Implicit vs Explicit Parallelism
- SIMD architectures, vector processing, multimedia extensions
- Graphics processing units
- Speed-up, Amdahl's Law, Scaled Speedup
- Properties of distributed systems
- Modularity
- Introduction to Cloud Computing – loosely based on book #1:
Cloud Computing Concepts, Technology & Architecture

October 5, 2021	TCCS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.23
-----------------	--	-------

23

PARALLELISM

- Discovering parallelism and development of parallel algorithms requires considerable effort
- Example: numerical analysis problems, such as solving large systems of linear equations or solving systems of Partial Differential Equations (PDEs), require algorithms based on domain decomposition methods.
- How can problems be split into independent chunks?
- Fine-grained parallelism
 - Only small bits of code can run in parallel without coordination
 - Communication is required to synchronize state across nodes
- Coarse-grained parallelism
 - Large blocks of code can run without coordination

October 5, 2021	TCCS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.24
-----------------	--	-------

24

PARALLELISM - 2

- Coordination of nodes
- Requires message passing or shared memory
- Debugging parallel message passing code is easier than parallel shared memory code
- Message passing: all of the interactions are clear
 - Coordination via specific programming API (MPI)
- Shared memory: interactions can be implicit – *must read the code!!*
- Processing speed is orders of magnitude faster than communication speed (CPU > memory bus speed)
- Avoiding coordination achieves the best speed-up

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.25
-----------------	--	-------

25

TYPES OF PARALLELISM

- Parallelism:
 - Goal: Perform multiple operations at the same time to achieve a speed-up
- Types of parallelism:
- Thread-level parallelism (TLP)
 - Control flow architecture
- Data-level parallelism
 - Data flow architecture
- Bit-level parallelism
- Instruction-level parallelism (ILP)

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.26
-----------------	--	-------

26

THREAD LEVEL PARALLELISM (TLP)

- Number of threads an application runs at any one time
- Varies throughout program execution
- As a metric:
- **Minimum: 1 thread**
- Can measure average, maximum (peak)
- **QUESTION:** What are the consequences of average (TLP) for scheduling an application to run on a computer with a fixed number of CPU cores and hyperthreads?
- Let's say there are 4 cores, or 8 hyper-threads...
- **Key to avoiding waste of computing resources is knowing your application's TLP...**

October 5, 2021

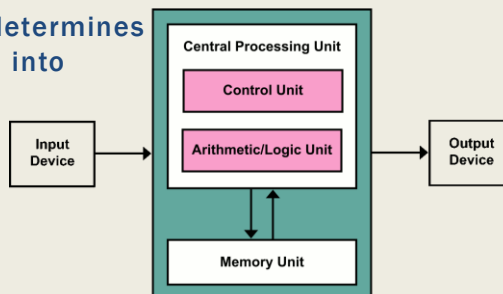
TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.27

27

CONTROL-FLOW ARCHITECTURE

- Typical architecture used today - w/ multiple threads
- By John von Neumann (1945)
- Also called the Von Neumann architecture
- Dominant computer system architecture
- Program counter (PC) determines next instruction to load into *instruction register*
- Program execution is sequential



October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.28

28

DATA-LEVEL PARALLELISM

- Partition data into big chunks, run separate copies of the program on them with little or no communication
- Problems are considered to be **embarrassingly parallel**
- Also perfectly parallel or pleasingly parallel...
- Little or no effort needed to separate problem into a number of parallel tasks
- MapReduce programming model is an example

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.29

29

DATA FLOW ARCHITECTURE

- **Alternate architecture** used by network routers, digital signal processors, special purpose systems
- Operations performed when input (data) becomes available
- Envisioned to provide much higher parallelism
- Multiple problems has prevented wide-scale adoption
 - Efficiently broadcasting data tokens in a massively parallel system
 - Efficiently dispatching instruction tokens in a massively parallel system
 - Building content addressable memory large enough to hold all of the dependencies of a real program

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.30

30

DATA FLOW ARCHITECTURE - 2

- Architecture not as popular as control-flow
- Modern CPUs emulate data flow architecture for dynamic instruction scheduling since the 1990s
 - Out-of-order execution – reduces CPU idle time by not blocking for instructions requiring data by defining execution windows
 - Execution windows: identify instructions that can be run by data dependency
 - Instructions are completed in data dependency order within execution window
 - Execution window size typically 32 to 200 instructions

Utility of data flow architectures has been much less than envisioned

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.31
-----------------	--	-------

31

BIT-LEVEL PARALLELISM

- Computations on large words (e.g. 64-bit integer) are performed as a single instruction
- Fewer instructions are required on 64-bit CPUs to process larger operands (A+B) providing dramatic performance improvements
- Processors have evolved: 4-bit, 8-bit, 16-bit, 32-bit, 64-bit

QUESTION: How many instructions are required to add two 64-bit numbers on a 16-bit CPU? (Intel 8088)

- 64-bit MAX int = 9,223,372,036,854,775,807 (signed)
- 16-bit MAX int = 32,767 (signed)
- Intel 8088 – limited to 16-bit registers

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.32
-----------------	--	-------

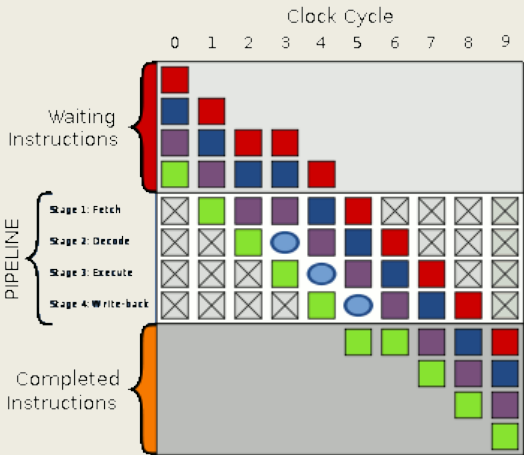
32

INSTRUCTION-LEVEL PARALLELISM (ILP)

- CPU pipelining architectures enable ILP
- CPUs have multi-stage processing pipelines
- Pipelining: split instructions into sequence of steps that can execute concurrently on different CPU circuitry
- Basic RISC CPU - Each instruction has 5 pipeline stages:
 - IF - *instruction fetch*
 - ID - *instruction decode*
 - EX - *instruction execution*
 - MEM - *memory access*
 - WB - *write back*

33

CPU PIPELINING



34

INSTRUCTION LEVEL PARALLELISM - 2

- RISC CPU:
 - After 5 clock cycles, all 5 stages of an instruction are loaded
 - Starting with 6th clock cycle, one full instruction completes each cycle
 - The CPU performs 5 tasks per clock cycle!
Fetch, decode, execute, memory read, memory write back
- Pentium 4 (CISC CPU) – processing pipeline w/ 35 stages!

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.35
-----------------	--	-------

35

OBJECTIVES – 10/5

- Questions from 9/30
- Cloud Computing – How did we get here?
(Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition)
- Data, thread-level, task-level parallelism & Parallel architectures
- Class Activity 1 – Implicit vs Explicit Parallelism
 - SIMD architectures, vector processing, multimedia extensions
 - Graphics processing units
 - Speed-up, Amdahl's Law, Scaled Speedup
 - Properties of distributed systems
 - Modularity
- Introduction to Cloud Computing – loosely based on book #1:
Cloud Computing Concepts, Technology & Architecture

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.36
-----------------	--	-------

36

ACTIVITY 1

- We will form groups of ~3 using Zoom breakout rooms
- Each group will complete a Google Doc worksheet
- Add names to Google Doc as they appear in Canvas
- The activity can be completed in class or after class
- The activity can also be completed individually
- When completed, one person should submit a PDF of the Google Doc to Canvas
- Instructor will score all group members based on the uploaded PDF file
- To get started:
 - Log into your UW Google Account (<https://drive.google.com>) using you UW NET ID
 - Follow the link:
<https://tinyurl.com/kp2jm9pj>

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.37
-----------------	--	-------

37

ACTIVITY 1

- Solutions to be discussed..

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.38
-----------------	--	-------

38

IMPLICIT PARALLELISM

■ Applies to:

■ Advantages:

■ Disadvantages:

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.39

39

EXPLICIT PARALLELISM

■ Applies to:

■ Advantages:

■ Disadvantages:

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.40

40

PARALLELISM QUESTIONS

- 7. For bit-level parallelism, should a developer be concerned with the available number of virtual CPU processing cores when choosing a cloud-based virtual machine if wanting to obtain the best possible speed-up? (Yes / No)
- 8. For instruction-level parallelism, should a developer be concerned with the physical CPU's architecture used to host a cloud-based virtual machine if wanting to obtain the best possible speed-up? (Yes / No)

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.41
-----------------	--	-------

41

PARALLELISM QUESTIONS - 2

- 9. For thread level parallelism (TLP) where a programmer has spent considerable effort to parallelize their code and algorithms, what consequences result when this code is deployed on a virtual machine with too few virtual CPU processing cores?
- What happens when this code is deployed on a virtual machine with too many virtual CPU processing cores?

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.42
-----------------	--	-------

42

OBJECTIVES – 10/5

- Questions from 9/30
- Cloud Computing – How did we get here?
(Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition)
- Data, thread-level, task-level parallelism & Parallel architectures
- Class Activity 1 – Implicit vs Explicit Parallelism
- SIMD architectures, vector processing, multimedia extensions
- Graphics processing units
- Speed-up, Amdahl's Law, Scaled Speedup
- Properties of distributed systems
- Modularity
- Introduction to Cloud Computing – loosely based on book #1:
Cloud Computing Concepts, Technology & Architecture

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.43
-----------------	--	-------

43

MICHAEL FLYNN’S COMPUTER ARCHITECTURE TAXONOMY

- Michael Flynn’s proposed taxonomy of computer architectures based on concurrent instructions and number of data streams (1966)
- SISD (Single Instruction Single Data)
- SIMD (Single Instruction, Multiple Data)
- MIMD (Multiple Instructions, Multiple Data)
- LESS COMMON: MISD (Multiple Instructions, Single Data)
- Pipeline architectures: functional units perform different operations on the same data
- For fault tolerance, may want to execute same instructions redundantly to detect and mask errors – for task replication

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.44
-----------------	--	-------

44

FLYNN’S TAXONOMY

- **SISD (Single Instruction Single Data)**
Scalar architecture with one processor/core.
 - Individual cores of modern multicore processors are “SISD”
- **SIMD (Single Instruction, Multiple Data)**
Supports vector processing
 - When SIMD instructions are issued, operations on individual vector components are carried out concurrently
 - Two 64-element vectors can be added in parallel
 - Vector processing instructions added to modern CPUs
 - Example: Intel MMX (multimedia) instructions

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.45
-----------------	--	-------

45

(SIMD): VECTOR PROCESSING
ADVANTAGES

- Exploit data-parallelism: vector operations enable speedups
- Vectors architecture provide vector registers that can store entire matrices into a CPU register
- SIMD CPU extension (e.g. MMX) add support for vector operations on traditional CPUs
- Vector operations reduce total number of instructions for large vector operations
- Provides higher potential speedup vs. MIMD architecture
- Developers can think sequentially; not worry about parallelism

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.46
-----------------	--	-------

46

FLYNN’S TAXONOMY - 2

- **MIMD (Multiple Instructions, Multiple Data)** - system with several processors and/or cores that function asynchronously and independently
- At any time, different processors/cores may execute different instructions on different data
- Multi-core CPUs are MIMD
- Processors share memory via interconnection networks
 - Hypercube, 2D torus, 3D torus, omega network, other topologies
- MIMD systems have different methods of sharing memory
 - Uniform Memory Access (UMA)
 - Cache Only Memory Access (COMA)
 - Non-Uniform Memory Access (NUMA)

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.47
-----------------	--	-------

47

ARITHMETIC INTENSITY

- **Arithmetic intensity:** Ratio of work (W) to memory traffic r/w (Q)

$I = \frac{W}{Q}$

Example: # of floating point ops per byte of data read
- Characterizes application scalability with SIMD support
 - *SIMD can perform many fast matrix operations in parallel*
- **High arithmetic Intensity:**
Programs with dense matrix operations scale up nicely (many calcs vs memory RW, supports lots of parallelism)
- **Low arithmetic intensity:**
Programs with sparse matrix operations do not scale well with problem size (memory RW becomes bottleneck, not enough ops!)

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.48
-----------------	--	-------

48

ROOFLINE MODEL

- When program reaches a given arithmetic intensity performance of code running on CPU hits a “roof”
- CPU performance bottleneck changes from: memory bandwidth (left) → floating point performance (right)



Key take-aways:

When a program's has **low** Arithmetic Intensity, memory bandwidth limits performance..

With **high** Arithmetic intensity, the system has peak parallel performance...

→ *performance is limited by??*

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.49

49

OBJECTIVES – 10/5

- Questions from 9/30
- Cloud Computing – How did we get here?
(Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition)
- Data, thread-level, task-level parallelism & Parallel architectures
- Class Activity 1 – Implicit vs Explicit Parallelism
- SIMD architectures, vector processing, multimedia extensions
- Graphics processing units
- Speed-up, Amdahl's Law, Scaled Speedup
- Properties of distributed systems
- Modularity
- Introduction to Cloud Computing – loosely based on book #1:
Cloud Computing Concepts, Technology & Architecture

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.50

50

GRAPHICAL PROCESSING UNITS (GPUs)

- GPU provides multiple SIMD processors
- Typically 7 to 15 SIMD processors each
- 32,768 total registers, divided into 16 lanes (2048 registers each)
- GPU programming model:
single instruction, multiple thread
- Programmed using CUDA- C like programming language by NVIDIA for GPUs
- CUDA threads – single thread associated with each data element (e.g. vector or matrix)
- Thousands of threads run concurrently

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.51

51

OBJECTIVES – 10/5

- Questions from 9/30
- Cloud Computing – How did we get here?
(Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition)
- Data, thread-level, task-level parallelism & Parallel architectures
- Class Activity 1 – Implicit vs Explicit Parallelism
- SIMD architectures, vector processing, multimedia extensions
- Graphics processing units
- Speed-up, Amdahl's Law, Scaled Speedup
- Properties of distributed systems
- Modularity
- Introduction to Cloud Computing – loosely based on book #1:
Cloud Computing Concepts, Technology & Architecture

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.52

52

PARALLEL COMPUTING

- Parallel hardware and software systems allow:
 - Solve problems demanding resources not available on single system.
 - Reduce time required to obtain solution
- The *speed-up* (S) measures effectiveness of parallelization:

$$S(N) = T(1) / T(N)$$

$T(1)$ → execution time of total sequential computation

$T(N)$ → execution time for performing N parallel computations in parallel

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.53

53

SPEED-UP EXAMPLE

- Consider embarrassingly parallel image processing
- Eight images (multiple data)
- Apply image transformation (greyscale) in parallel
- 8-core CPU, 16 hyperthreads
- Sequential processing: perform transformations one at a time using a single program thread
 - 8 images, 3 seconds each: $T(1) = 24 \text{ seconds}$
- Parallel processing
 - 8 images, 3 seconds each: $T(N) = 3 \text{ seconds}$
- Speedup: $S(N) = 24 / 3 = 8x \text{ speedup}$
- Called “perfect scaling”
- Must consider data transfer and computation setup time

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.54

54

AMDAHL'S LAW

- Amdahl's law is used to estimate the speed-up of a job using parallel computing
 1. Divide job into two parts
 2. Part A that will still be sequential
 3. Part B that will be sped-up with parallel computing
- Portion of computation which cannot be parallelized will determine (i.e. limit) the overall speedup
- Amdahl's law assumes jobs are of a fixed size
- Also, Amdahl's assumes no overhead for distributing the work, and a perfectly even work distribution

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.55

55

AMDAHL'S LAW

$$S = \frac{1}{(1 - f) + \frac{f}{N}}$$

- S = theoretical speedup of the whole task
- f= fraction of work that is parallel (ex. 25% or 0.25)
- N= proposed speed up of the parallel part (ex. 5 times speedup)
- % improvement of task execution = $100 * (1 - (1 / S))$
- **Using Amdahl's law, what is the maximum possible speed-up?**

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.56

56

AMDAHL'S LAW EXAMPLE

- Program with two independent parts:

- Part A is 75% of the execution time
- Part B is 25% of the execution time

- Part B is made 5 times faster with parallel computing

- Estimate the percent improvement of task execution

- Original Part A is 3 seconds, Part B is 1 second

- $N=5$ (speedup of part B)

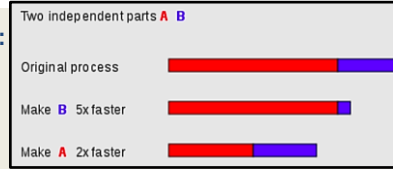
- $f=.25$ (only 25% of the whole job (A+B) will be sped-up)

- $S=1 / ((1-f) + f/S)$

- $S=1 / ((.75) + .25/5)$

- $S=1.25$

- % improvement = $100 * (1 - 1/1.25) = 20\%$



from Wikipedia

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
 School of Engineering and Technology, University of Washington - Tacoma

L2.57

57

GUSTAFSON'S LAW

- Calculates the scaled speed-up using “N” processors

$$S(N) = N + (1 - N) \alpha$$

N: Number of processors

α : fraction of program run time which can't be parallelized
 (e.g. must run sequentially)

- *Can be used to estimate runtime of parallel portion of program*

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
 School of Engineering and Technology, University of Washington - Tacoma

L2.58

58

GUSTAFSON'S LAW

- Calculates the scaled speed-up using “N” processors
$$S(N) = N + (1 - N) \alpha$$

N: Number of processors
 α : fraction of program run time which can't be parallelized (e.g. must run sequentially)
- Can be used to estimate runtime of parallel portion of program
- Where $\alpha = \sigma / (\pi + \sigma)$
- Where σ = sequential time, π =parallel time
- Our Amdahl's example: $\sigma= 3s$, $\pi =1s$, $\alpha =.75$

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.59
-----------------	--	-------

59

GUSTAFSON'S LAW

- Calculates the scaled speed-up using “N” processors
$$S(N) = N + (1 - N) \alpha$$

N: Number of processors
 α : fraction of program run time which can't be parallelized (e.g. must run sequentially)
- Example:
Consider a program that is embarrassingly parallel, but 75% cannot be parallelized. $\alpha=.75$
QUESTION: *If deploying the job on a 2-core CPU, what scaled speedup is possible assuming the use of two processes that run in parallel?*

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.60
-----------------	--	-------

60

GUSTAFSON'S EXAMPLE

■ **QUESTION:**
What is the maximum theoretical speed-up on a **2-core CPU** ?
 $S(N) = N + (1 - N) \alpha$
 $N=2, \alpha=.75$
 $S(N) = 2 + (1 - 2) .75$
 $S(N) = ?$

■ What is the maximum theoretical speed-up on a **16-core CPU**?
 $S(N) = N + (1 - N) \alpha$
 $N=16, \alpha=.75$
 $S(N) = 16 + (1 - 16) .75$
 $S(N) = ?$

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.61

61

GUSTAFSON'S EXAMPLE

■ **QUESTION:**
What is the maximum theoretical speed-up on a **2-core CPU** ?
 $S(N) = N + (1 - N) \alpha$
 $N=2, \alpha=$
 $S(N) = 2 + (1 - 2) .75$
 $S(N) = ?$

■ What is the maximum theoretical speed-up on a **16-core CPU**?
 $S(N) = N + (1 - N) \alpha$
 $N=16, \alpha=.75$
 $S(N) = 16 + (1 - 16) .75$
 $S(N) = ?$

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.62

62

MOORE’S LAW

- Transistors on a chip doubles approximately every 1.5 years
- CPUs now have billions of transistors
- Power dissipation issues at faster clock rates leads to heat removal challenges
 - Transition from: increasing clock rates → to adding CPU cores
- Symmetric core processor – multi-core CPU, all cores have the same computational resources and speed
- Asymmetric core processor – on a multi-core CPU, some cores have more resources and speed
- Dynamic core processor – processing resources and speed can be dynamically configured among cores
- Observation: asymmetric processors offer a higher speedup

October 5, 2021	TCCS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.63
-----------------	--	-------

63

OBJECTIVES – 10/5

- Questions from 9/30
- Cloud Computing – How did we get here?
(Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition)
- Data, thread-level, task-level parallelism & Parallel architectures
- Class Activity 1 – Implicit vs Explicit Parallelism
- SIMD architectures, vector processing, multimedia extensions
- Graphics processing units
- Speed-up, Amdahl's Law, Scaled Speedup
- Properties of distributed systems
- Modularity
- Introduction to Cloud Computing – loosely based on book #1:
Cloud Computing Concepts, Technology & Architecture

October 5, 2021	TCCS562:Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.64
-----------------	---	-------

64

DISTRIBUTED SYSTEMS

- Collection of autonomous computers, connected through a network with distribution software called “middleware” that enables coordination of activities and sharing of resources
- Key characteristics:
 - Users perceive system as a single, integrated computing facility.
 - Compute nodes are autonomous
 - Scheduling, resource management, and security implemented by every node
 - Multiple points of control and failure
 - Nodes may not be accessible at all times
 - System can be scaled by adding additional nodes
 - Availability at low levels of HW/software/network reliability

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.65
-----------------	--	-------

65

DISTRIBUTED SYSTEMS - 2

- Key non-functional attributes
 - Known as “ilities” in software engineering
- Availability – 24/7 access?
- Reliability - Fault tolerance
- Accessibility – reachable?
- Usability – user friendly
- Understandability – can under
- Scalability – responds to variable demand
- Extensibility – can be easily modified, extended
- Maintainability – can be easily fixed
- Consistency – data is replicated correctly in timely manner

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.66
-----------------	--	-------

66

TRANSPARENCY PROPERTIES OF
DISTRIBUTED SYSTEMS

- **Access transparency:** local and remote objects accessed using identical operations
- **Location transparency:** objects accessed w/o knowledge of their location.
- **Concurrency transparency:** several processes run concurrently using shared objects w/o interference among them
- **Replication transparency:** multiple instances of objects are used to increase reliability
 - *users are unaware if and how the system is replicated*
- **Failure transparency:** concealment of faults
- **Migration transparency:** objects are moved w/o affecting operations performed on them
- **Performance transparency:** system can be reconfigured based on load and quality of service requirements
- **Scaling transparency:** system and applications can scale w/o change in system structure and w/o affecting applications

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.67
-----------------	--	-------

67

OBJECTIVES – 10/5

- Questions from 9/30
- Cloud Computing – How did we get here?
(Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition)
- Data, thread-level, task-level parallelism & Parallel architectures
- Class Activity 1 – Implicit vs Explicit Parallelism
- SIMD architectures, vector processing, multimedia extensions
- Graphics processing units
- Speed-up, Amdahl's Law, Scaled Speedup
- Properties of distributed systems
- Modularity
- Introduction to Cloud Computing – loosely based on book #1:
Cloud Computing Concepts, Technology & Architecture

October 5, 2021	TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.68
-----------------	--	-------

68

TYPES OF MODULARITY

- **Soft modularity:** TRADITIONAL
 - Divide a program into modules (classes) that call each other and communicate with shared-memory
 - A procedure calling convention is used (or method invocation)
- **Enforced modularity:** CLOUD COMPUTING
 - Program is divided into modules that communicate only through message passing
 - The ubiquitous client-server paradigm
 - Clients and servers are independent decoupled modules
 - System is more robust if servers are stateless
 - May be scaled and deployed separately
 - May also FAIL separately!

October 5, 2021	TCCS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.69
-----------------	--	-------

69

CLOUD COMPUTING – HOW DID WE GET HERE?
SUMMARY OF KEY POINTS

- Multi-core CPU technology and hyper-threading
- What is a
 - Heterogeneous system?
 - Homogeneous system?
 - Autonomous or self-organizing system?
- **Fine grained vs. coarse grained parallelism**
- Parallel message passing code is easier to debug than shared memory (e.g. p-threads)
- Know your application’s max/avg **Thread Level Parallelism (TLP)**
- **Data-level parallelism:** Map-Reduce, (SIMD) Single Instruction Multiple Data, Vector processing & GPUs

October 5, 2021	TCCS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma	L2.70
-----------------	--	-------

70

CLOUD COMPUTING – HOW DID WE GET HERE?
SUMMARY OF KEY POINTS - 2

- **Bit-level parallelism**
- **Instruction-level parallelism** (CPU pipelining)
- **Flynn’s taxonomy:** computer system architecture classification
 - **SISD** – Single Instruction, Single Data (modern core of a CPU)
 - **SIMD** – Single Instruction, Multiple Data (Data parallelism)
 - **MIMD** – Multiple Instruction, Multiple Data
 - MISD is RARE; application for fault tolerance...
- **Arithmetic Intensity:** ratio of calculations vs memory RW
- **Roofline model:**
Memory bottleneck with low arithmetic intensity
- **GPUs:** ideal for programs with high arithmetic intensity
 - SIMD and Vector processing supported by many large registers

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.71

71

CLOUD COMPUTING – HOW DID WE GET HERE?
SUMMARY OF KEY POINTS - 3

- **Speed-up (S)**
 $S(N) = T(1) / T(N)$
- **Amdahl’s law:**
 $S = 1 / \alpha$
 α = percent of program that must be sequential
- **Scaled speedup with N processes:**
 $S(N) = N - \alpha(N-1)$
- Moore’s Law
- Symmetric core, Asymmetric core, Dynamic core CPU
- Distributed Systems Non-function quality attributes
- Distributed Systems – Types of Transparency
- Types of modularity- Soft, Enforced

October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma



L2.72

72

INTRODUCTION TO CLOUD COMPUTING

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma



L2.73

73

OBJECTIVES – 10/5

- Questions from 9/30
- Cloud Computing – How did we get here?
(Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition)
- Data, thread-level, task-level parallelism & Parallel architectures
- Class Activity 1 – Implicit vs Explicit Parallelism
- SIMD architectures, vector processing, multimedia extensions
- Graphics processing units
- Speed-up, Amdahl's Law, Scaled Speedup
- Properties of distributed systems
- Modularity
- Introduction to Cloud Computing – loosely based on book #1:
Cloud Computing Concepts, Technology & Architecture

October 5, 2021

TCSS562:Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.74

74

OBJECTIVES – 10/5

- Introduction to Cloud Computing
 - Why study cloud computing?
 - History of cloud computing
 - Business drivers
 - Cloud enabling technologies
 - Terminology
 - Benefits of cloud adoption
 - Risks of cloud adoption

September 30, 2019	TCSS562: Software Engineering for Cloud Computing [Fall 2019] School of Engineering and Technology, University of Washington - Tacoma	L2.75
--------------------	--	-------

75

WHY STUDY CLOUD COMPUTING?

- LINKEDIN - TOP IT Skills from job app data
 - #1 Cloud and Distributed Computing
 - <https://learning.linkedin.com/week-of-learning/top-skills>
 - #2 Statistical Analysis and Data Mining
- FORBES Survey – 6 Tech Skills That’ll Help You Earn More
 - #1 Data Science
 - #2 Cloud and Distributed Computing
 - <http://www.forbes.com/sites/laurencebradford/2016/12/19/6-tech-skills-thatll-help-you-earn-more-in-2017/>

September 30, 2019	TCSS562: Software Engineering for Cloud Computing [Fall 2019] School of Engineering and Technology, University of Washington - Tacoma	L2.76
--------------------	--	-------

76

WHY STUDY CLOUD COMPUTING? - 2

■ Computerworld Magazine

TECH FORECAST 2017 SPECIAL REPORT

Hot Skills

Top 10 skills respondents plan to hire for in the next 12 months:

Source: Computerworld's Forecast 2017 survey of 196 IT managers, directors and executives.

Base: 57 respondents who expect to increase IT head count in the next 12 months.

Programming/application development35%

Help desk/tech support35%

Security/compliance/governance26%

Cloud/SaaS26%

Business intelligence/analytics26%

Web development26%

Database administration25%

Project management25%

Big data25%

Mobile applications and device management21%

© COMPUTERWORLD

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.77

77

OBJECTIVES – 10/5

■ Introduction to Cloud Computing

■ Why study cloud computing?

■ History of cloud computing

■ Business drivers

■ Cloud enabling technologies

■ Terminology

■ Benefits of cloud adoption

■ Risks of cloud adoption

September 30, 2019


TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.78

78

A BRIEF HISTORY OF CLOUD COMPUTING

- John McCarthy, 1961
 - Turing award winner for contributions to AI
- “If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility... The computer utility could become the basis of a new and important industry...”



September 30, 2019	TCSS562: Software Engineering for Cloud Computing [Fall 2019] School of Engineering and Technology, University of Washington - Tacoma	L2.79
--------------------	--	-------

79

CLOUD HISTORY - 2

- Internet based computer utilities
 - Since the mid-1990s
 - Search engines: Yahoo!, Google, Bing
 - Email: Hotmail, Gmail
 - 2000s
 - Social networking platforms: MySpace, Facebook, LinkedIn
 - Social media: Twitter, YouTube
 - Popularized core concepts
 - Formed basis of cloud computing

September 30, 2019	TCSS562: Software Engineering for Cloud Computing [Fall 2019] School of Engineering and Technology, University of Washington - Tacoma	L2.80
--------------------	--	-------

80

Cloud History: Services - 1

- Late 1990s – Early Software-as-a-Service (SaaS)
 - Salesforce: Remotely provisioned services for the enterprise
- 2002 -
 - Amazon Web Services (AWS) platform: Enterprise oriented services for remotely provisioned storage, computing resources, and business functionality
- 2006 – Infrastructure-as-a-Service (IaaS)
 - Amazon launches Elastic Compute Cloud (EC2) service
 - Organization can “lease” computing capacity and processing power to host enterprise applications
 - Infrastructure

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.81

81

Cloud History: Services - 2

- 2006 – Software-as-a-Service (SaaS)
 - Google: Offers Google DOCS, “MS Office” like fully-web based application for online documentation creation and collaboration
- 2009 – Platform-as-a-Service (PaaS)
 - Google: Offers Google App Engine, publicly hosted platform for hosting scalable web applications on google-hosted datacenters

September 30, 2019


TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.82

82

CLOUD COMPUTING
NIST GENERAL DEFINITION

“Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (networks, servers, storage, applications and services) that can be rapidly provisioned and reused with minimal management effort or service provider interaction”...



September 30, 2019	TCSS562: Software Engineering for Cloud Computing [Fall 2019] School of Engineering and Technology, University of Washington - Tacoma	L2.83
--------------------	--	-------

83

MORE CONCISE DEFINITION

“Cloud computing is a specialized form of distributed computing that introduces utilization models for remotely provisioning scalable and measured resources.”

From Cloud Computing Concepts, Technology, and Architecture
Z. Mahmood, R. Puttini, Prentice Hall, 5th printing, 2015

September 30, 2019	TCSS562: Software Engineering for Cloud Computing [Fall 2019] School of Engineering and Technology, University of Washington - Tacoma	L2.84
--------------------	--	-------

84

OBJECTIVES – 10/5

- Introduction to Cloud Computing
 - Why study cloud computing?
 - History of cloud computing
 - Business drivers
 - Cloud enabling technologies
 - Terminology
 - Benefits of cloud adoption
 - Risks of cloud adoption

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.85

85

BUSINESS DRIVERS
FOR CLOUD COMPUTING

- Capacity planning
- Cost reduction
- Operational overhead
- Organizational agility

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.86

86

BUSINESS DRIVERS FOR CLOUD COMPUTING

- Capacity planning
 - Process of determining and fulfilling future demand for IT resources
 - Capacity vs. demand
 - Discrepancy between capacity of IT resources and actual demand
 - Over-provisioning: resource capacity exceeds demand
 - Under-provisioning: demand exceeds resource capacity
 - Capacity planning aims to minimize the discrepancy of available resources vs. demand

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.87

87



Dwight, The Office TV sitcom

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.88

88

BUSINESS DRIVERS FOR CLOUD - 2

- Capacity planning
 - Over-provisioning: is costly due to too much infrastructure
 - Under-provisioning: is costly due to potential for business loss from poor quality of service
- Capacity planning strategies
 - Lead strategy: add capacity in anticipation of demand (pre-provisioning)
 - Lag strategy: add capacity when capacity is fully leveraged
 - Match strategy: add capacity in small increments as demand increases
- Load prediction
 - Capacity planning helps anticipate demand fluctuations

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.89

89

CAPACITY PLANNING

Capacity vs. Usage
(Traditional Data Center)

Compute Power

Time

Planned Capacity

Actual Usage

Waste

Customer Dissatisfaction

amazon
aws services

September 30, 2019

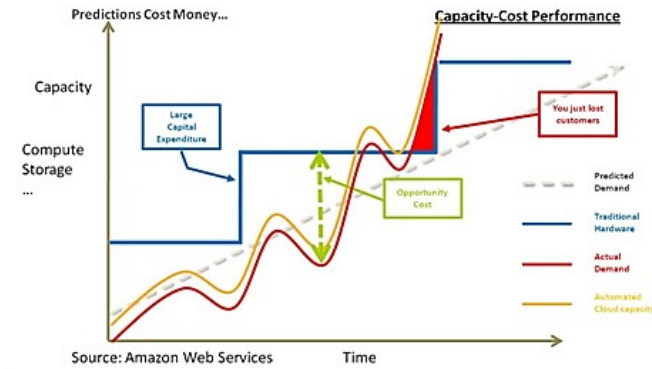
TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.90

90

CAPACITY PLANNING - 2

■ Ca



September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
 School of Engineering and Technology, University of Washington - Tacoma

L2.91

91

BUSINESS DRIVERS FOR CLOUD - 3

■ Cost reduction

- IT Infrastructure acquisition
- IT Infrastructure maintenance

■ Operational overhead

- Technical personnel to maintain physical IT infrastructure
- System upgrades, patches that add testing to deployment cycles
- Utility bills, capital investments for power and cooling
- Security and access control measures for server rooms
- Admin and accounting staff to track licenses, support agreements, purchases

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
 School of Engineering and Technology, University of Washington - Tacoma

L2.92

92

BUSINESS DRIVERS FOR CLOUD - 4

- Organizational agility
 - Ability to adapt and evolve infrastructure to face change from internal and external business factors
 - Funding constraints can lead to insufficient on premise IT
 - Cloud computing enables IT resources to scale with a lower financial commitment

September 30, 2019	TCSS562: Software Engineering for Cloud Computing [Fall 2019] School of Engineering and Technology, University of Washington - Tacoma	L2.93
--------------------	--	-------

93

OBJECTIVES – 10/5

- Introduction to Cloud Computing
 - Why study cloud computing?
 - History of cloud computing
 - Business drivers
 - Cloud enabling technologies
 - Terminology
 - Benefits of cloud adoption
 - Risks of cloud adoption

September 30, 2019	TCSS562: Software Engineering for Cloud Computing [Fall 2019] School of Engineering and Technology, University of Washington - Tacoma	L2.94
--------------------	--	-------

94

TECHNOLOGY INNOVATIONS
LEADING TO CLOUD

- Cluster computing
- Grid computing
- Virtualization
- Others

September 30, 2019


TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.95

95

CLUSTER COMPUTING

- Cluster computing (clustering)
 - Cluster is a group of independent IT resources interconnected as a single system
 - Servers configured with homogeneous hardware and software
 - Identical or similar RAM, CPU, HDDs
 - Design emphasizes redundancy as server components are easily interchanged to keep overall system running
 - Example: if a RAID card fails on a key server, the card can be swapped from another redundant server
 - Enables warm replica servers
 - Duplication of key infrastructure servers to provide HW failover to ensure high availability (HA)




September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.96

96

GRID COMPUTING



- On going research area since early 1990s
- Distributed heterogeneous computing resources organized into logical pools of loosely coupled resources
- For example: heterogeneous servers connected by the internet
- Resources are heterogeneous and geographically dispersed
- Grids use middleware software layer to support workload distribution and coordination functions
- Aspects: load balancing, failover control, autonomic configuration management
- Grids have influenced clouds contributing common features: networked access to machines, resource pooling, scalability, and resiliency

September 30, 2019

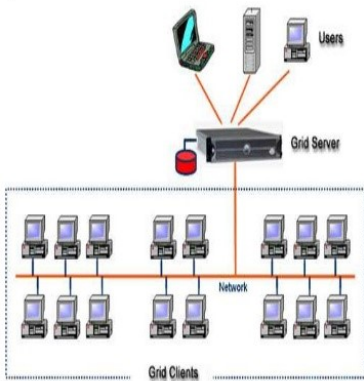
TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.97

97

GRID COMPUTING - 2

How Grid computing works ?



In general, a grid computing system requires:

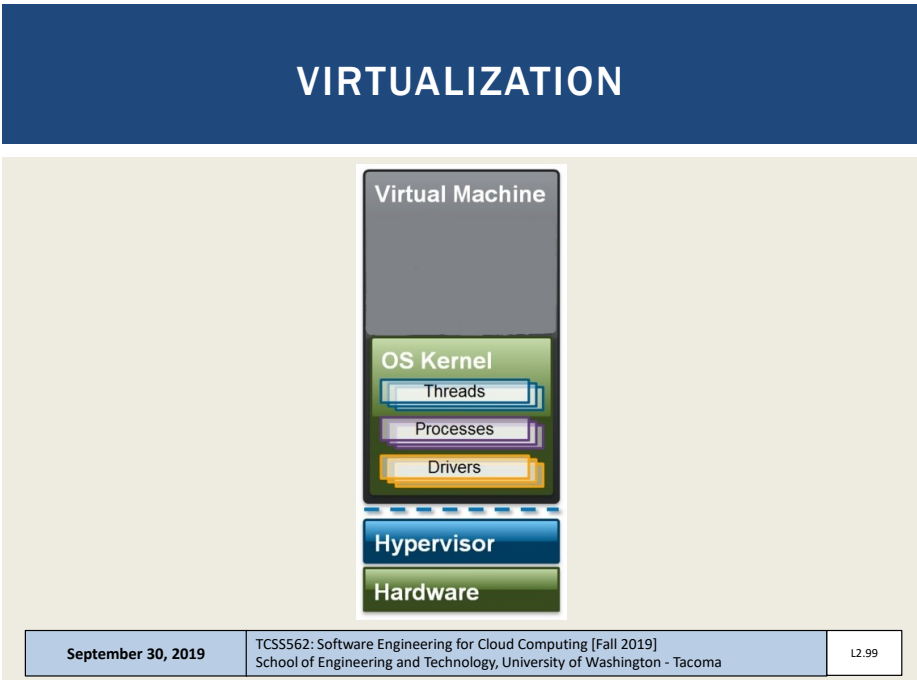
- At least one computer, usually a server, which handles all the administrative duties for the System
- A network of computers running special grid computing network software.
- A collection of computer software called middleware

September 30, 2019

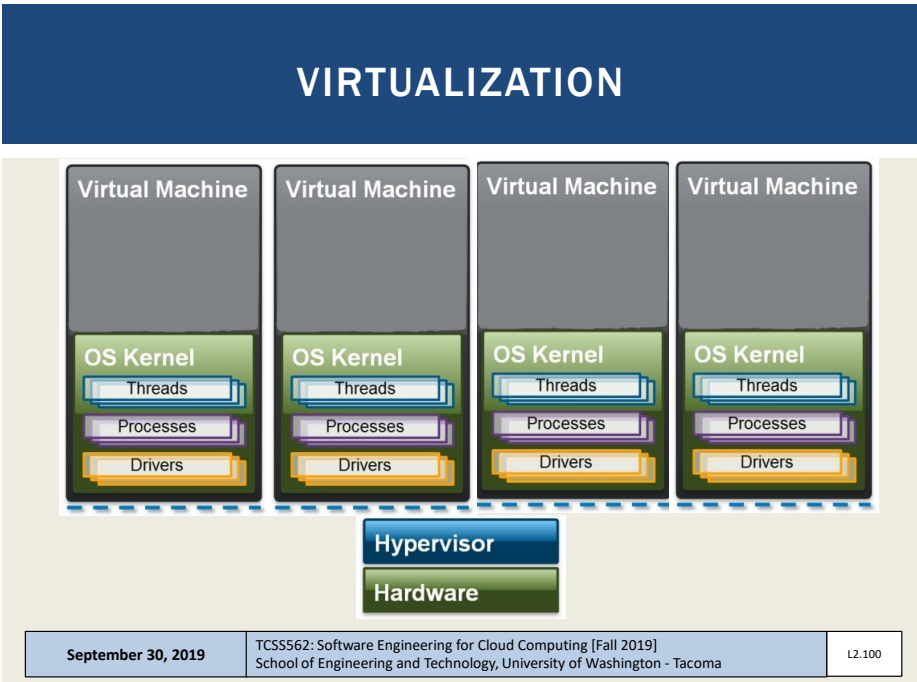
TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.98

98



99



100

VIRTUALIZATION

- Simulate physical hardware resources via software
 - The virtual machine (virtual computer)
 - Virtual local area network (VLAN)
 - Virtual hard disk
 - Virtual network attached storage array (NAS)
- Early incarnations featured significant performance, reliability, and scalability challenges
- CPU and other HW enhancements have minimized performance GAPS

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.101

101

OBJECTIVES – 10/5

- Introduction to Cloud Computing
 - Why study cloud computing?
 - History of cloud computing
 - Business drivers
 - Cloud enabling technologies
 - Terminology
 - Benefits of cloud adoption
 - Risks of cloud adoption

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.102

102

KEY TERMINOLOGY

- **On-Premise Infrastructure**
 - Local server infrastructure not configured as a cloud
- **Cloud Provider**
 - Corporation or private organization responsible for maintaining cloud
- **Cloud Consumer**
 - User of cloud services
- **Scaling**
 - **Vertical scaling**
 - Scale up: increase resources of a single virtual server
 - Scale down: decrease resources of a single virtual server
 - **Horizontal scaling**
 - Scale out: increase number of virtual servers
 - Scale in: decrease number of virtual servers

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.103

103

VERTICAL SCALING

- Reconfigure virtual machine to have different resources:
 - CPU cores
 - RAM
 - HDD/SDD capacity
- May require VM migration if physical host machine resources are exceeded

The diagram shows a vertical arrow labeled "vertical scaling". To the left of the arrow, there are two server icons. The bottom icon is labeled "A" and is associated with "2 CPUs". The top icon is labeled "B" and is associated with "4 CPUs". An upward-pointing arrow connects the two server icons, indicating the transition from 2 CPUs to 4 CPUs.

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.104

104

HORIZONTAL SCALING

- Increase (scale-out) or decrease (scale-in) number of virtual servers based on demand

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.105

105

HORIZONTAL VS VERTICAL SCALING

Horizontal Scaling	Vertical Scaling
Less expensive using commodity HW	Requires expensive high capacity servers

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.106

106

HORIZONTAL VS VERTICAL SCALING

Horizontal Scaling	Vertical Scaling
Less expensive using commodity HW	Requires expensive high capacity servers
IT resources instantly available	IT resources typically instantly available

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.107

107

HORIZONTAL VS VERTICAL SCALING

Horizontal Scaling	Vertical Scaling
Less expensive using commodity HW	Requires expensive high capacity servers
IT resources instantly available	IT resources typically instantly available
Resource replication and automated scaling	Additional setup is normally needed

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.108

108

HORIZONTAL VS VERTICAL SCALING

Horizontal Scaling	Vertical Scaling
Less expensive using commodity HW	Requires expensive high capacity servers
IT resources instantly available	IT resources typically instantly available
Resource replication and automated scaling	Additional setup is normally needed
Additional servers required	No additional servers required

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.109

109

HORIZONTAL VS VERTICAL SCALING

Horizontal Scaling	Vertical Scaling
Less expensive using commodity HW	Requires expensive high capacity servers
IT resources instantly available	IT resources typically instantly available
Resource replication and automated scaling	Additional setup is normally needed
Additional servers required	No additional servers required
Not limited by individual server capacity	Limited by individual server capacity

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.110

110

KEY TERMINOLOGY - 2

- Cloud services
 - Broad array of resources accessible “as-a-service”
 - Categorized as Infrastructure (IaaS), Platform (PaaS), Software (SaaS)
- Service-level-agreements (SLAs):
 - Establish expectations for: uptime, security, availability, reliability, and performance

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.111

111

OBJECTIVES – 10/5

- Introduction to Cloud Computing
 - Why study cloud computing?
 - History of cloud computing
 - Business drivers
 - Cloud enabling technologies
 - Terminology
 - Benefits of cloud adoption
 - Risks of cloud adoption

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.112

112

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

September 30, 2019


TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.113

113

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...



September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.114

114

CLoud BENEFITS - 3

- Example: Using 100 servers for 1 hour costs the same as using 1 server for 100 hours
- Rosetta Protein Folding: Working with a UW-Tacoma graduate student, we recently 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)


September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.115

115

OH YOU NEED MORE SERVERS?
INTERESTING... I HAVE SOMETHING TO SHOW YOU...



Gene Wilder, Charlie and the Chocolate Factory

116

CLOUD BENEFITS

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

Time (h)	Concurrent Users
2	2,000
4	1,500
6	1,800
8	2,500
10	4,000
12	7,000
14	9,000
16	9,500
18	8,000
20	4,000
22	2,800
24	2,500

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.117

117

OBJECTIVES – 10/5

- Introduction to Cloud Computing
 - Why study cloud computing?
 - History of cloud computing
 - Business drivers
 - Cloud enabling technologies
 - Terminology
 - Benefits of cloud adoption
 - Risks of cloud adoption

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.118

118

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

September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.119

119

NETWoRK LATENCY CoSTS

The diagram illustrates the network latency costs between Organization A and Cloud A. Organization A is represented by a dashed box containing a blue cube labeled 'cloud service consumer'. Cloud A is represented by a dashed box containing a yellow cloud labeled 'cloud service'. A horizontal line connects the 'cloud service consumer' to the 'cloud service'. Above Organization A is a grey box labeled 'reliable network'. Above Cloud A is a grey box labeled 'reliable network'. Between the two dashed boxes is a grey box labeled 'unreliable network connection' with a red lightning bolt symbol on the line connecting the two boxes. Below Organization A is the text 'organizational boundary of cloud consumer'. Below Cloud A is the text 'organizational boundary of cloud provider'.


September 30, 2019

TCSS562: Software Engineering for Cloud Computing [Fall 2019]
School of Engineering and Technology, University of Washington - Tacoma

L2.120

120

QUESTIONS



October 5, 2021

TCSS562: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

L2.123