



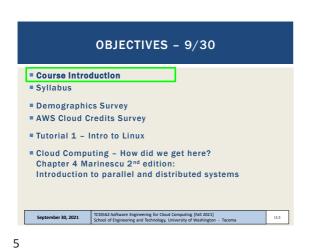
Where are you joining us from? (PUGET SOUND REGION)

INTRODUCTIONS: What is your name? nickname

W / alias? and list one or more areas of interest in

Computer Science:

3



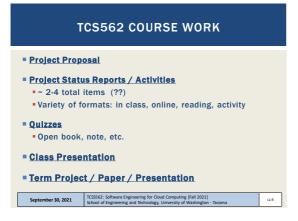
TCSS 562 - Fall 2021 Online is green... ■ 100% reduction of carbon footprint from transit ■ Saves commuting time TCSS 562 FALL 2021 Less fuel expenses ■ Easier to achieve perfect attendance -Lecture recording sessions are streamed LIVE for 24/7 availability • UW deletes content after ~90 days ■ 19 class meetings 2 Holidays: No Class on Nov 11 & Nov 25 ■ This course will not have exams ■ This course helps with preparation for TCSS 558 - Applied Distributed Computing

Slides by Wes J. Lloyd

6







CLASS PRESENTATION

■ Each student will make one presentation in a team of ~3

■ Technology sharing presentation

■ PPT Slides, demonstration

■ Provide technology overview of one cloud service offering

■ Present overview of features, performance, etc.

■ Cloud Research Paper Presentation

■ PPT slides, identify research contributions, strengths and weaknesses of paper, possible areas for future work

September 30, 2021 | TCSSSG2: Software Engineering for Cloud Computing [Fall 2021] | School of Engineering and Technology, University of Washington - Tacoma

9



TCSS562 TERM PROJECT - 2

Deliverables
Demo in class at end of quarter (TBD)
Project report paper (4-6 pgs IEEE format, template provided)
GitHub (project source)
How-To document (via GitHub markdown)

A standard project will be offered or propose your own
(Previous) Groups built an Extract-Transform-Load style serverless data processing pipeline combing AWS Lambda, S3, and Amazon Aurora Serverless DB

11 12

Slides by Wes J. Lloyd

Apache OpenWhisk, OpenFaaS, Fn, others

TCSS562: Software Engineering for School of Engineering and Technology

# CASE STUDY ALTERNATIVES Creative case studies are encouraged !!! Compare and contrast alternative application designs considering various cloud services, languages, platforms, etc. Examples: Application case study on cloud storage service trade-offs: Object/blob storage services Amazon S3, Google blobstore, Azure blobstore, vs. self-hosted App case study on cloud relational database service trade-offs: Amazon Relational Database Service (RDS), Aurora, Self-Hosted DB App case study on Platform-as-a-Service (PaaS) alternatives Amazon Elastic Beanstalk, Heroku, others App Case study on open source FaaS platforms

**CASE STUDY ALTERNATIVES - 2** 

- App case study on serverless storage alternatives
  - From AWS Lambda: Amazon EFS, S3, Containers, others
- App case study based on container platform hosting
- Amazon ECS/Fargate, AKS, Azure Kubernetes, Self-hosted Kubernetes cluster on cloud VMs
- App case study contrasting queueing service alternatives
  - Amazon SQS, Amazon MQ, Apache Kafka, RabbitMQ, Omq, others
- App case study on NoSQL database services comparison
  - DynamoDB, Google BigTable, MongoDB, Cassandra

September 30, 2021 TCSSS62: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma

13

September 30, 2021

### TERM PROJECT: KEY IDEA

- 1. BUILD A MULTI-FUNCTION SERVERLESS APPLICATION
  - Typically consisting of AWS Lambda Functions or Google Cloud Functions, etc. (e.g. FaaS platfrom)
- 2. CONTAST THE USE OF ALTERNATIVE CLOUD SERVICES TO INSTRUMENT SOME OR MULTIPLE ASPECTS OF THE APPLICATION
- CONDUCT A PERFORMANCE EVALUATION, REPORT ON YOUR FINDINGS IN A LIGHTNING TALK (5-minutes) AND TERM PAPER

September 30, 2021 TCSSS62: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma

14

### **KEY IDEA - 2**

- Application should involve multiple processing steps
- Implementation does not have to be FaaS
- Implementation involves use of external services (e.g. databases, object stores, queues)
- Case studies contrast alternate designs
- Which designs offer the fastest performance?
- Lowest cost?
- Best maintainability? In other words, have the least code? (Lines of Code metric)

■ Project cloud infrastructure support:

Standard AWS Account (RECOMMENDED)

Create standard AWS account with UW email

Instructor provides students with \$50 credit vouchers

Credits provided throughout Fall quarter (within reason)

September 30, 2021 TCSSS62: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma

15

16

### TERM PROJECT: RESEARCH

- Alternative I: conduct a cloud-related research project on any topic focused on specific research goals / questions
  - Can be used to help spur MS Capstone/Thesis work
  - If you're interested in this option, please talk with the instructor
  - First step is to identify 1 2 research questions
- Alternative II: conduct a gap-analysis literature survey of cloud computing research papers, produce a report which identifies open problems for future research in cloud computing that have tractable next steps
- Instructor will help guide projects throughout the quarter
- Project proposal approval based on team vision and preparedness for the project

September 30, 2021 TCSSS62: Software Engineering for Cloud Computing [Fall 2021]
September 30, 2021 September 30, 2021 TCSSS62: Software Engineering and Technology University of Michigantee Tocomo

AWS Educate

Includes up to \$100 in AWS credits via a restricted starter account

• When voucher is used up, request another voucher from instructor

**PROJECT SUPPORT** 

- Credits direct from Amazon, no instructor intervention necessary
- No Credit Card required

Credit card required

Spot instances not available (low-cost VMs)

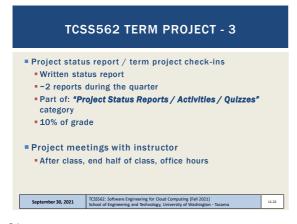
September 30, 2021 TCSSS62: Software Engineering for Cloud Computing [Fall 2021]

L1.18

17 18



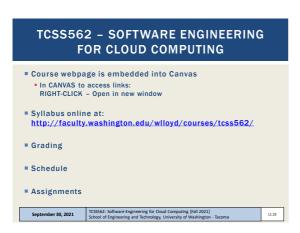
19 20



Course Introduction
 Syllabus
 Demographics Survey
 AWS Cloud Credits Survey
 Tutorial 1 - Intro to Linux
 Cloud Computing - How did we get here?
 Chapter 4 Marinescu 2<sup>nd</sup> edition:
 Introduction to parallel and distributed systems

| Toccompany | Tocc

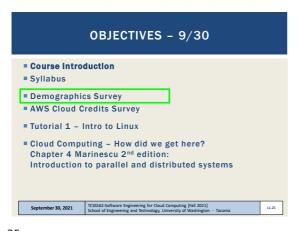
21

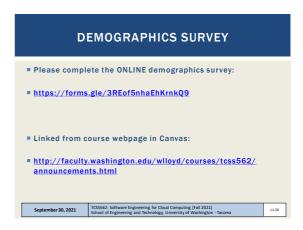


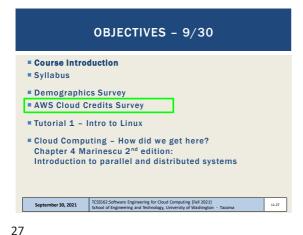
RECORDING BREAK
WILL RETURN AT 6:10PM

23 24

Slides by Wes J. Lloyd L1.4







AWS CLOUD CREDITS SURVEY

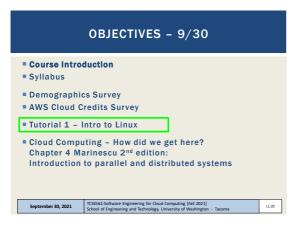
Please complete the ONLINE demographics survey:

https://forms.gle/uumXX9YGhQ34fm8x7

Linked from course webpage in Canvas:

http://faculty.washington.edu/wlloyd/courses/tcss562/announcements.html

...

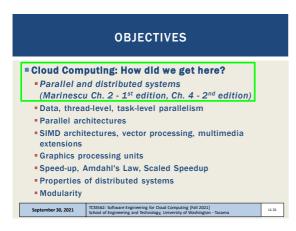


Course Introduction
Syllabus
Demographics Survey
AWS Cloud Credits Survey
Tutorial 1 - Intro to Linux
Cloud Computing - How did we get here?
Chapter 4 Marinescu 2<sup>nd</sup> edition:
Introduction to parallel and distributed systems

TCSS62-Software Engineering for Cloud Computing [Fall 2021]
Software Engineering for Cloud Computing [Fall 2021]
Software Engineering and Inchnology, University of Washington - Tecoma

29 30

Slides by Wes J. Lloyd L1.5



**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

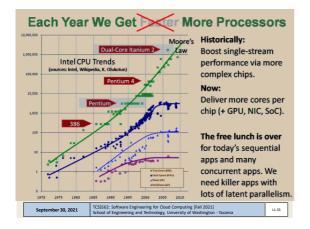
September 30, 2021

32

34

31

33



AMD'S 64-CORE 7NM CPUS ■ Fnvc Rome CPUs Announced August 2019 ■ EPYC 7H12 requires liquid cooling AMD EPYC 7002 Pro 64 / 128 3.35 EPYC 7642 48 / 96 3.20 256 MB \$4775 2.30 225 W EPYC 7552 48 / 96 2.20 3.30 192 MB \$4025 September 30, 2021

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 4770 with HTT Vs. 4670 without HTT - 25% improvement w/ HTT are not equivalent to (2) CPU cores CPU Mark Relative to Top 10 Common CPUs ■ i7-4770 and i5-4760 same CPU, with and without HTT ■ Example: → hyperthreads add +32.9% September 30, 2021

**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 September 30, 2021

36 35

### **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

37

- Based on client-server paradigm
- Thin clients leverage compute hosted on the cloud
- Applications run many web service instances
- Employ load balancing

September 30, 2021 TCSSS62: Software Engineering for Clo School of Engineering and Technology

## **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

38

September 30, 2021

## **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

September 30, 2021

39

### **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 Intel E5-2666v3 2.9GHz CPUs

40

September 30, 2021

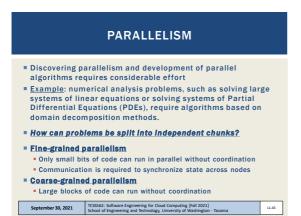
### **CLOUD COMPUTING: HOW DID WE GET HERE? - 3**

- Compute clouds are large-scale distributed systems
  - Heterogeneous systems
  - Homogeneous systems
  - Autonomous
  - Self organizing

September 30, 2021

41 42

### **OBJECTIVES** Cloud Computing: How did we get here? Parallel and distributed systems (Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition) Data, thread-level, task-level parallelism Parallel architectures SIMD architectures, vector processing, multimedia extensions Graphics processing units Speed-up, Amdahl's Law, Scaled Speedup Properties of distributed systems Modularity September 30, 2021



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

  september 30, 2021 | CSSSS62: Software Engineering for Cloud Computing [Fall 2021]

44

43

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

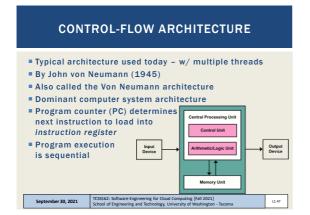
THREAD LEVEL PARALLELISM (TLP)

- Number of threads an application runs at any one time
- Varies throughout program execution
- As a metric:
- <u>Minimum</u>: 1 thread
- Can measure <u>average</u>, <u>maximum (peak)</u>
- QUESTION: What are the consequences of <u>average</u> (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...

September 30, 2021 TCSSS62: Software Engineering for Cloud Computing [Fall 2021]
School of Engineering and Technology University of Washington - Taronna

46

45



Partition data into big chunks, run separate copies of the program on them with little or no communication

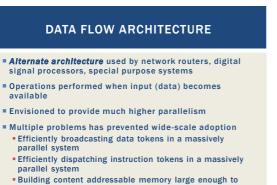
Problems are considered to be embarrassingly parallel

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

47 48



**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 size typically 32 to 200 instructions

Utility of data flow architectures has been much less than envisioned

September 30, 2021

49

September 30, 2021

### **BIT-LEVEL PARALLELISM**

Computations on large words (e.g. 64-bit integer) are performed as a single instruction

hold all of the dependencies of a real program

- 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

TCSS562: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Ta September 30, 2021

50

### 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

September 30, 2021

52

### 51

September 30, 2021

53

## **CPU PIPELINING**

RISC CPU:

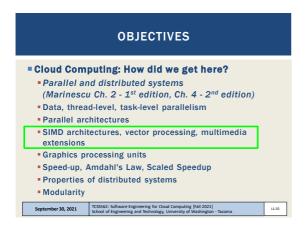
After 5 clock cycles, all 5 stages of an instruction are loaded

**INSTRUCTION LEVEL PARALLELISM - 2** 

- Starting with 6<sup>th</sup> 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!

54

September 30, 2021



## 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

September 30, 2021 TCSSS62: Software Engineering for Cloud Computing [Fall School of Engineering and Technology, University of Wasl

56

55

57

### **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

September 30, 2021 TCSSS62: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma

## (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

September 30, 2021 TCSSS62: Software Engineering for Cloud Computing [Fall 2021] School of Engineering and Technology, University of Washington - Tacoma

**ARITHMETIC INTENSITY** 

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

Programs with dense matrix operations scale up nicely

(many calcs vs memory RW, supports lots of parallelism)

Programs with sparse matrix operations do not scale well

Ratio of work (W) to memory traffic r/w (Q)

58

 MIMD (Multiple Instructions, Multiple Data) - system with several processors and/or cores that function asynchronously and independently

FLYNN'S TAXONOMY - 2

- 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)

September 30, 2021 TCSSS62: Software Engineering for Cloud Computing [Fall 2021]

Subset of Continencian and Technologies, University of Mechinetee Towns

(memory RW becomes bottleneck, not enough ops!)

September 30, 2021 TCSSS62: Software Engineering for Cloud Computing [fall 2021]
School of Engineering and Technology, University of Washington - Tacoma

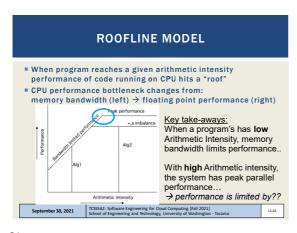
Arithmetic intensity:

High arithmetic Intensity:

Low arithmetic intensity:

with problem size

59 60



Cloud Computing: How did we get here?

Parallel and distributed systems
(Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition)

Data, thread-level, task-level parallelism

Parallel architectures

SIMD architectures, vector processing, multimedia extensions

Graphics processing units

Speed-up, Amdahl's Law, Scaled Speedup

Properties of distributed systems

Modularity

| TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | September 30, 2021 | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | September 30, 2021 | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2021] | TGSSG2: Software Engineering for Cloud Computing [Fall 2

61 62

# 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

Cloud Computing: How did we get here?

Parallel and distributed systems
(Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition)

Data, thread-level, task-level parallelism

Parallel architectures

SIMD architectures, vector processing, multimedia extensions

Graphics processing units

Speed-up, Amdahl's Law, Scaled Speedup

Properties of distributed systems

Modularity

Tossics: Software Engineering for Cloud Computing [fall 2021] should of Engineering and Technology, University of Washington - Taccoms

63

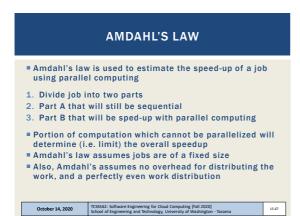
# PARALLEL COMPUTING ■ Parallel hardware and software systems allow: ■ Solve problems demanding resources not available on single system. ■ Reduce time required to obtain solution ■ The speed-up (S) measures effectiveness of parallelization: S(N) = T(1) / T(N) T(1) → execution time of total sequential computation T(N) → execution time for performing N parallel computations in parallel September 30, 2021 | TSSSG2: Software Engineering for Cloud Computing [Fall 2021] | School of Engineering and Technology, University of Washington - Tacoma | 11.55

SPEED-UP EXAMPLE

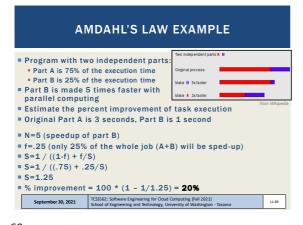
Consider embarrassingly parallel image processing
Eight images (multiple data)
Apply image transformation (greyscale) in parallel
S-core CPU, 16 hyper threads
Sequential processing: perform transformations one at a time using a single program thread
Simages, 3 seconds each: T(1) = 24 seconds
Parallel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

65 66

Slides by Wes J. Lloyd L1.11



67



GUSTAFSON'S LAW

Calculates the scaled speed-up using "N" processors S(N) = N + (1 - N) α

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

Cotober 14, 2020

TCSSSG2: Software Engineering for Cloud Computing [Fall 2020] School of Engineering and Technology, University of Washington - Taccoma

69

GUSTAFSON'S LAW	
■ Calculates th	te <b>scaled speed-up</b> using "N" processors $S(N) = N + (1 - N) \alpha$
	processors program run time which can't be parallelized n sequentially)
■ Can be used a program ■ Where α = σ	to estimate runtime of parallel portion of $(\pi + \sigma)$
	quential time, $\pi$ =parallel time
Our Amdahl's	s example: $\sigma$ = 3s, $\pi$ =1s, $\alpha$ =.75
October 14, 2020	TCSSS62: Software Engineering for Cloud Computing [Fall 2020] School of Engineering and Technology, University of Washington - Tacoma

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. α=.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?

September 30, 2021

TGSSGS: Software Engineering for Cloud Computing [Fall 2021]
September 30, 2021

TGSSGS: Software Engineering for Cloud Computing [Fall 2021]
TGSSGS Software Engineering for Cloud Computing [Fall 2021]

71 72

Slides by Wes J. Lloyd L1.12

68

# 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) = ?\*\*September 30, 2021 | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: Software Engineering for Cloud Computing [Fall 2021] | TCSSG2: So

GUSTAFSON'S EXAMPLE

# QUESTION:
What is the maximum theoretical speed-up on a 2-core CPU?  $S(N) = N + (1 - N) \alpha$   $N = 2, \alpha = S(N) = 2$ For 2 CPUs, speed up is 1.25x S(N) = ?For 16 CPUs, speed up is 4.75x

# What is the maximum theoretical speed-up on a 2-core CPU?  $S(N) = N + (1 - N) \alpha$   $N = 16, \alpha = .75$  S(N) = 16 + (1 - 16) .75 S(N) = ?# September 30, 2021

\*\*TOCOMORDING\*\* [Tail 2021]

\*\*TOCOMORDING\*\* [Tail 2021]

\*\*TOCOMORDING\*\* [Tail 2021]

\*\*TOCOMORDING\*\* [Tail 2021]

73

# 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 | September 30, 2021 | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021] | INCSSE2: Software Engineering for Cloud Computing [Fall 2021]

Cloud Computing: How did we get here?

Parallel and distributed systems
(Marinescu Ch. 2 - 1st edition, Ch. 4 - 2nd edition)

Data, thread-level, task-level parallelism

Parallel architectures

SIMD architectures, vector processing, multimedia extensions

Graphics processing units

Speed-up, Amdahl's Law, Scaled Speedup

Properties of distributed systems

Modularity

| TCSSS2: Software Engineering for Cloud Computing [Fall 2021] | School of Engineering and Technology, University of Washington - Taccoma | 11.76

75

# 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 September 30, 2021 TSSSG2: Software Engineering for load computing [Fal 2021] September 30, 2021 LTTS LTT

77 78

Slides by Wes J. Lloyd L1.13

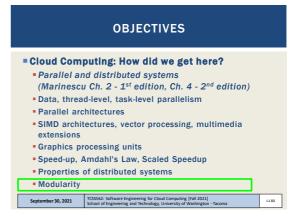
74

## TRANSPARENCY PROPERTIES OF **DISTRIBUTED SYSTEMS**

- Access transparency: local and remote objects accessed using identical operations
- Location transparency: objects accessed w/o knowledge of
- 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
- Fallure 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

are Engineering for Cloud Co eering and Technology, Univ September 30, 2021

79



### TYPES OF MODULARITY

- Divide a program into modules (classes) that call each other
- 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!

September 30, 2021

82

L1.81

80

■ What is a

Heterogeneous system?

Homogeneous system?

speed-up (S)
S(N) = T(1) / T(N)

## **Soft modularity:** TRADITIONAL

- and communicate with shared-memory

81

### **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

September 30, 2021

83 84

## • 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 September 30, 2021

**CLOUD COMPUTING - HOW DID WE GET HERE?** 

**SUMMARY OF KEY POINTS - 3** 

**CLOUD COMPUTING - HOW DID WE GET HERE?** 

SUMMARY OF KEY POINTS

■ Multi-core CPU technology and hyper-threading

■ Amdahl's law: S=1 / ((1-f) + f/N),s=latency, f=parallel fraction, N=speed-up 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 September 30, 2021

