CS4823/CS6643 Parallel Programming

Wei Wang

Why Parallel Computing?

- Technology Requirements
  - Transistor scaling advances and challenges
- Application Requirements
  - Large data sets
  - Massive user requests
  - High-performance/scientific computing

Technology Requirements

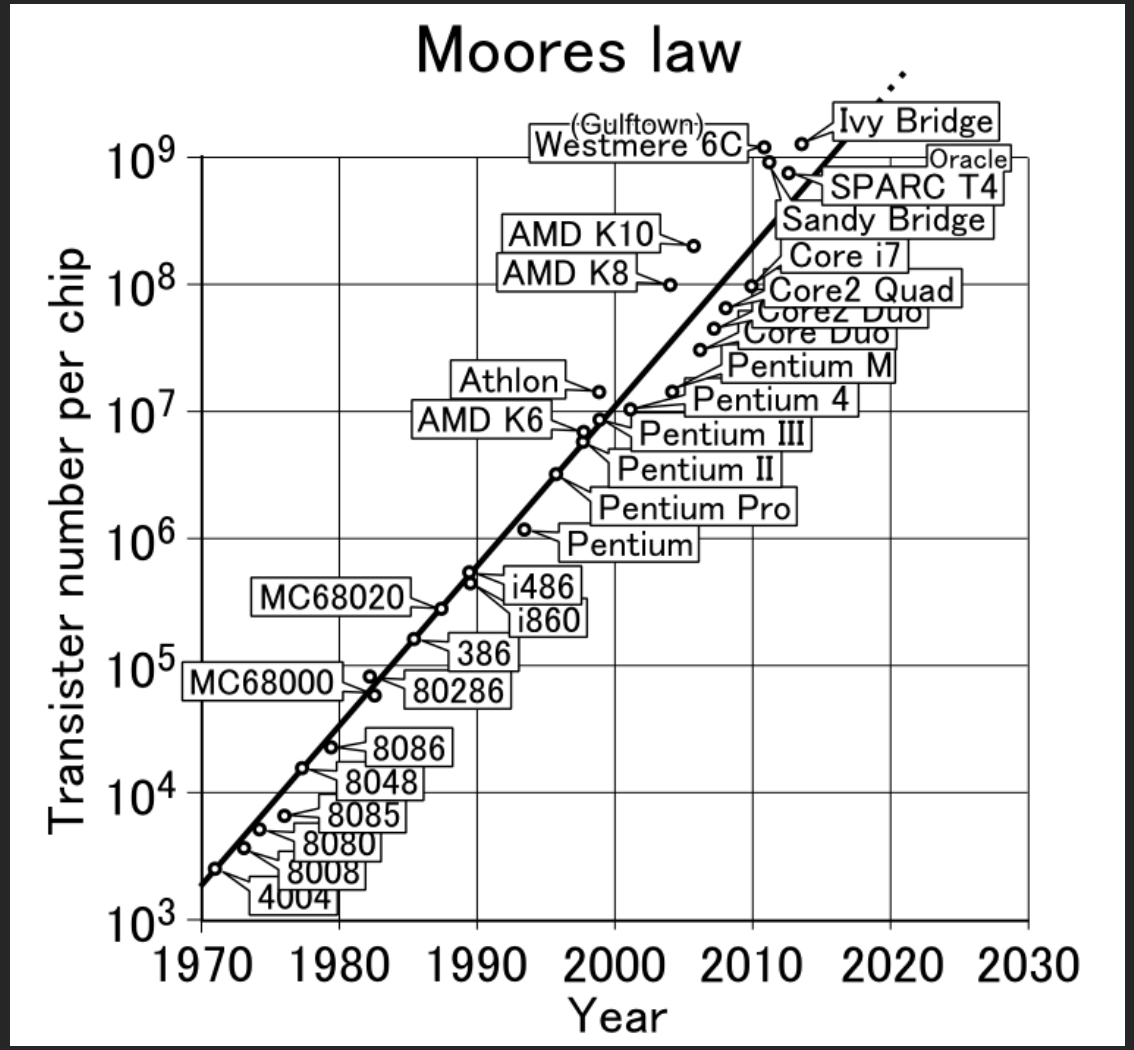
and The Rise of Multi-core Processors

Moore’s Law and Dennard Scaling

- The driver of modern computer industry can be explained by two observations on transistor scaling
  - Moore’s Law
  - Dennard Scaling

Moore’s Law

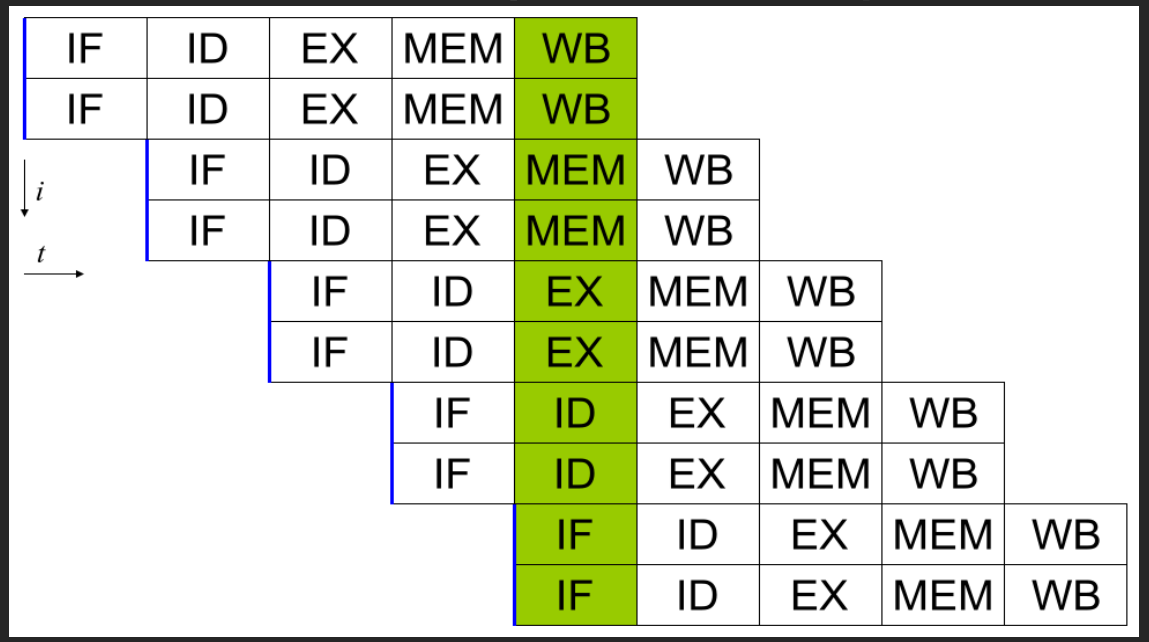
- Gordon Moore, Founder of Intel
- The number of transistors in a dense integrated circuit doubles approximately every two years



- By shigeru23 CC BY-SA 3.0 from Wikimedia Commons

The Benefits of Moore's Law

- More transistors ➔ more functional units (ALU, FPU etc.) ➔ can execute multiple codes simultaneously
- Implicit parallelization (a.k.a, instruction level parallelism)
  - Superscalar: fetch and execute multiple instructions simultaneously



- By Amit6 CC BY-SA 3.0

The Benefits of Moore's Law

- More transistors → more functional units (ALU, FPU etc.) → can execute multiple codes simultaneously
- Implicit parallelization (a.k.a., instruction level parallelism)
  - Superscalar: fetch and execute multiple instructions simultaneously
  - Out-of-order execution – additionally CPU components that maximize the chance of finding instructions that can be executed simultaneously
- Explicit parallelization (a.k.a., thread-level parallelism)
  - SMT (simultaneous multithreading), a.k.a. Hyper Threading for Intel
  - Multi-core and many-core

Dennard Scaling

- Robert Dennard, IBM fellow
- Dennard Scaling: As transistors get smaller their power density stays constant, so that the power use stays in proportion with area: both voltage and current scale (downward) with length

The Benefit of Dennard Scaling

- Power consumption of a CPU chip

$$p = c \times v^2 \times f$$

- *p* is the power consumption of a chip (in watt)
- *c* is the total number of transistors of a chip
- *v* is the power voltage (supply voltage) to run the chip
- *f* is the frequency of the CPU

The Benefit of Dennard Scaling cont'd

The power consumption of an old generation of CPU

- $p_{old} = c_{old} \times v_{old}^2 \times f_{old}$

The power consumption of a new generation of CPU

- $p_{new} = c_{new} \times v_{new}^2 \times f_{new}$
- Based on Moore's law, $c_{new} = 2c_{old}$, i.e., transistor count doubles in the next generation
- Based on Dennard Scaling, $v_{new} = \frac{1}{2} v_{old}$, i.e., the supply voltage can be reduced by half
- If we double the frequency, i.e., $f_{new} = 2f_{old}$

The Benefit of Dennard Scaling cont'd

- The powers consumption of the new chip:
  - $p_{new} = c_{new} \times {v_{new}}^2 \times f_{new}$
    - Based on Moore's law, $c_{new} = 2c_{old}$, i.e., transistor count doubles in the next generation
    - Based on Dennard Scaling, $v_{new} = \frac{1}{2} v_{old}$, i.e., the supply voltage can be reduced by half
    - If we double the frequency, i.e., $f_{new} = 2f_{old}$
  - $p_{new} = c_{new} \times {v_{new}}^2 \times f_{new} = 2c_{old} \times (\frac{1}{2} v_{old})^2 \times 2f_{old}$
    $= c_{old} \times {v_{old}}^2 \times f_{old} = p_{old}$

The new CPU has the same power consumption as the old CPU with more transistors and doubled frequency

The Benefit of Dennard Scaling cont'd

- The benefit of Dennard Scaling:
  Every new generation of CPU doubles performance (i.e., doubles frequency) while retaining the same-level of power consumption
- Performance gain of every new generation of CPU is free – no extra cost of power and no extra cost of manufacture

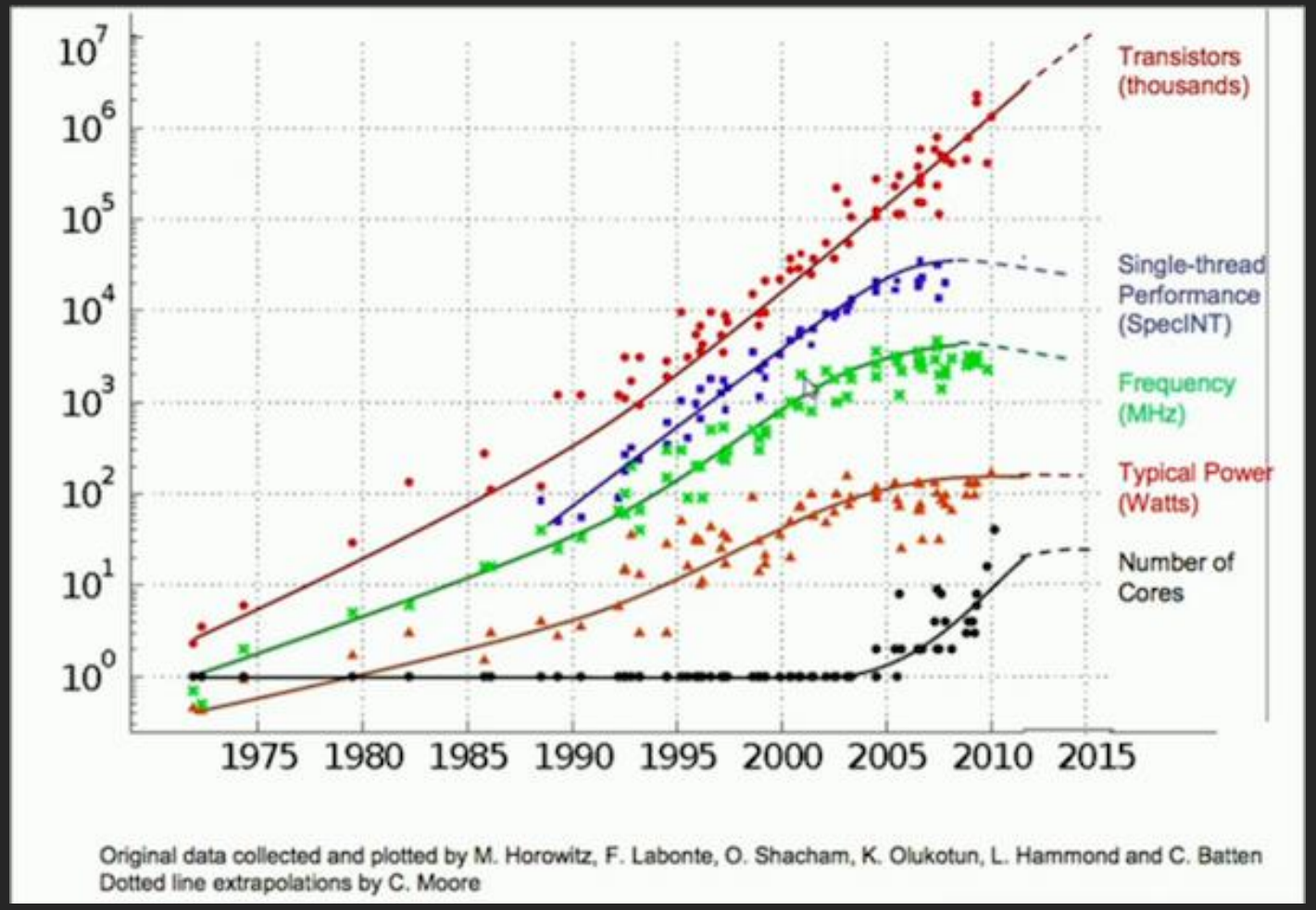
The Gradual Failing of Dennard Scaling

- Since 2005, Dennard Scaling started to fail
- No more half-reduction in the supply voltage
- If we double the performance, power consumption will be 4 times higher

$$p_{new} = c_{new} \times {v_{new}}^2 \times f_{new} = 2c_{old} \times {v_{old}}^2 \times 2f_{old}$$
$$= c_{old} \times {v_{old}}^2 \times f_{old} = 4p_{old}$$

- Don't forget more power == more heat, cooling is even more expensive
- Since 2005, instead of increasing frequency, the extra transistors are used to add more cores
  - Power consumption is still increasing, but at least not 4 times worse

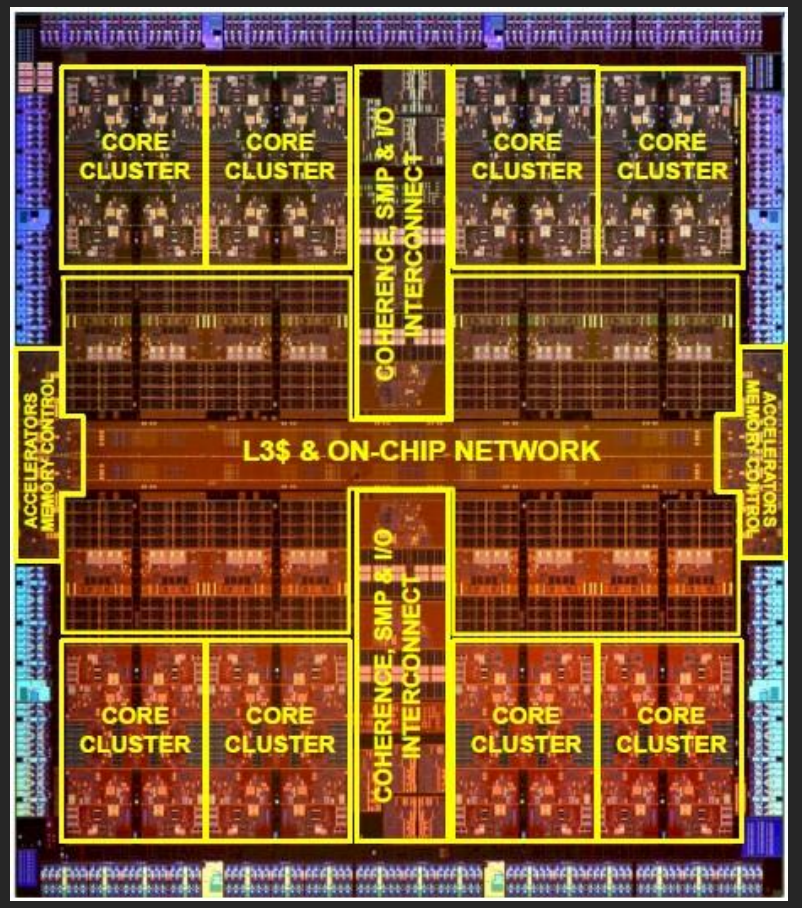
The Stall of CPU Frequency



-C. Moore: Data Processing in Exascale-Class Computer Systems, April 2011

Recent Processors

- 2022: AMD Threadripper — 64 cores
- 2018: AMD Ryzen2 — 32 cores
- 2015: Oracle SPARC M7 —32 cores
- 2014: Intel Haswell —18 cores
- 2013: SPARC M6 —12 cores
- 2012: AMD Trinity — 4 cores



Oracle SPARC M7

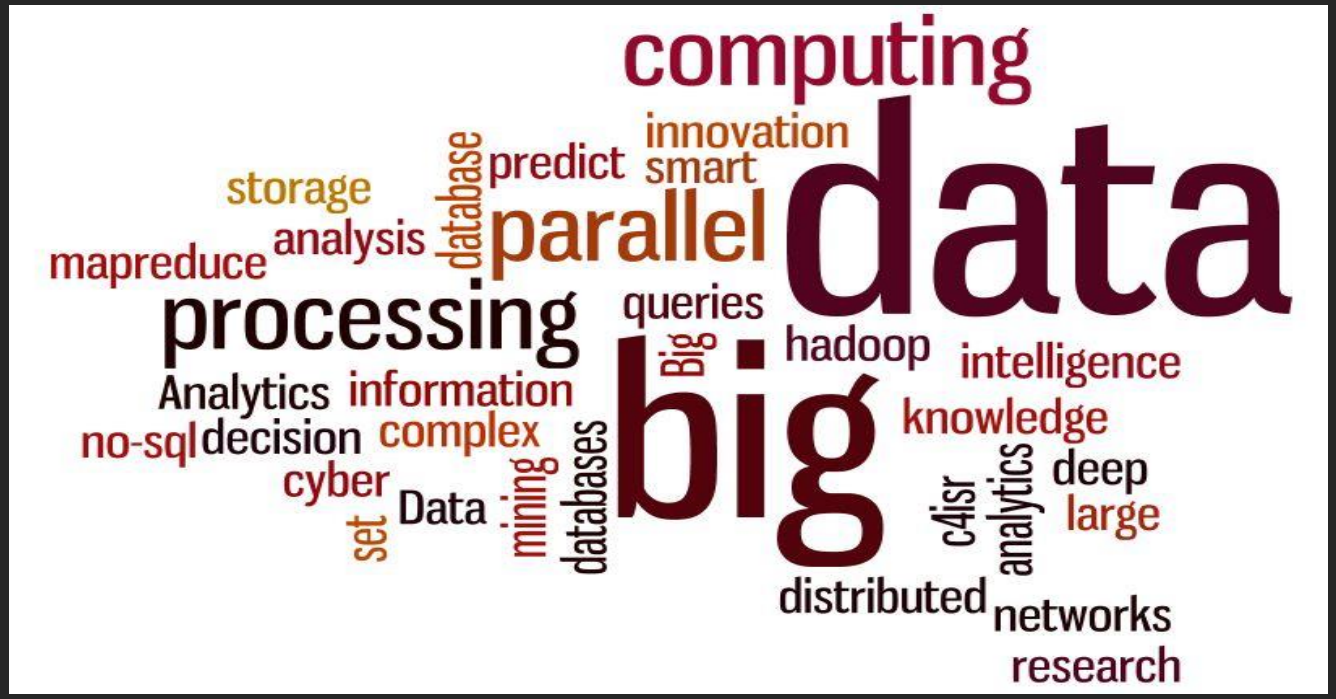
Technology Requirement: Moral of the Story

- Because of the failing of Dennard Scaling, parallelization becomes the primary means to keep improving performance
- Most programs we write today will be executed in a parallel style, so it is important to understand the parallel computation

Application Requirements

Application Requirements: Large Data Sets

- Data Analytics
  - Customer Relationship
  - Social media analysis
- Machine Learning
  - Fraud detection
  - Medical diagnostics
  - Large language/image models
- Data usually too large for the memory and disk of a single machine
- Analysis algorithms are usually heavy and complex

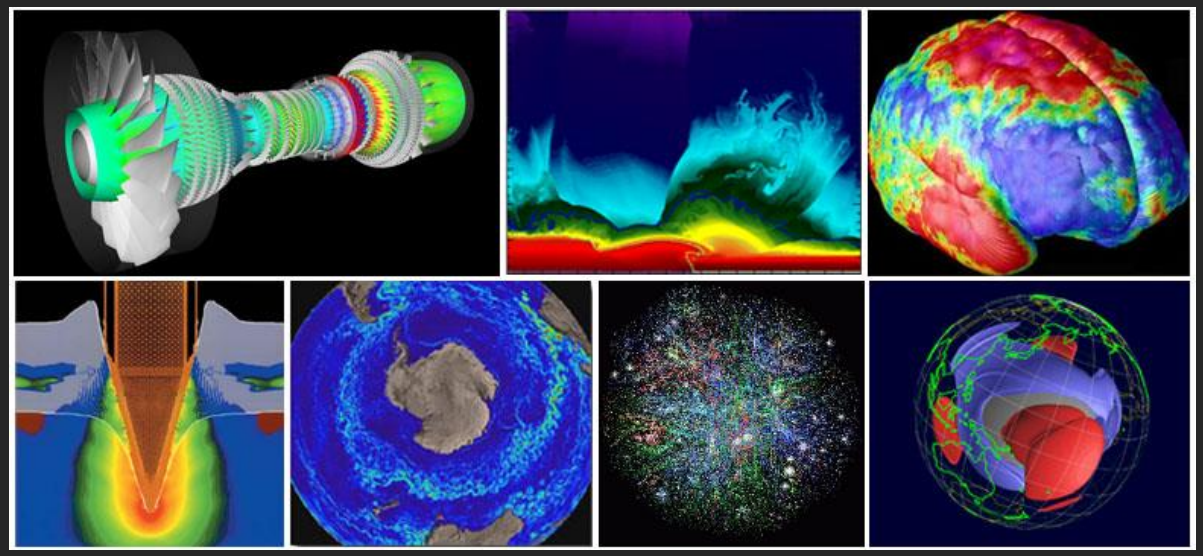


Application Requirements: Massive User Requests

- 7,000+ Tweets / sec
- 50,000+ Google Searches /sec
- 100,000+ Youtube Videos /sec
- 398 items /sec sold on Amazon on Prime day 2015
- To satisfy the basic performance requirements, data centers with tens of thousands of servers are employed
- A big challenge to coordinate the computations of thousands of servers

Application Requirements: High-Performance/Scientific Computing

- Analyzing DNA sequences
- Simulate earth quake, ocean circulation
- Computational fluid dynamics and airplane design
- Climate prediction
- Advanced material design
- Complex computations that only can be finished with massive numbers of computers running in parallel

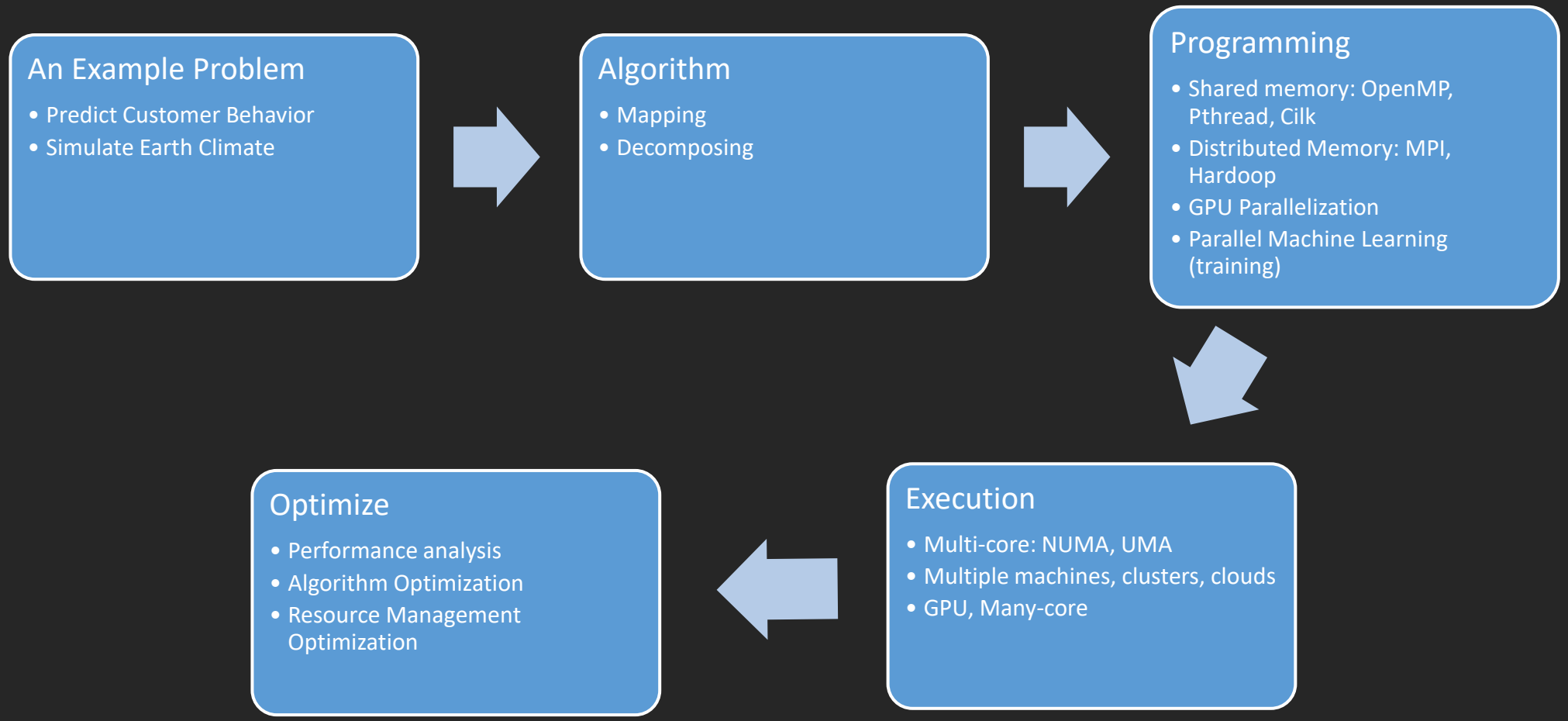


Application Requirements VS Technology Challenges

- While single core performance growth has stalled, the sizes of our problems and data sets are keep growing.
- To handles these larger problems, we are forced to rely on parallel computation.
- As problem size keeps growth, our parallel systems are getting more complex. Therefore, not only do we need to learn parallel programming, we also need to learn managing and optimizing parallel applications.
- Traditional parallel hardware and programming models may not be enough for future problems

Goal of This Class

The Essential Parts of Parallel Computing



What I Hope You Learn In the End

- Understand parallel computer architectures.
- Learn the Operating System and Middle-ware support for parallel computation.
- Understand parallel algorithm designs.
- Learn to program parallel applications with various software libraries on different parallel architectures.
- Understand parallel performance theory and learn to analyze and optimize parallel applications.

Class Information

Class Information

- Prerequisites: CS 3343 and CS 3423; be able to write C/C++ programs
- CS4823
  - One in-class midterm exam (20%) – openbook.
- CS4823/6643
  - One in-class final programming exam (25%)
- Programming Assignments (CS4823 50%, CS6643 70%)
  - 5-7 programming assignments
  - CS6643 has two more required assignments than CS4823.
- Class participation, in class quizzes, (5%)
- Special and extra credit events (3%)
- The percentages may be adjusted depending on the number of assignments and quizzes

Class Information cont'd

- Textbook: Introduction to Parallel Computing (2nd Edition), by Ananth Grama, George Karypis, Vipin Kumar, Anshul Gupta
  - Optional
- Late homework is docked 10%
  - If it is more than one week late, the assignment will not be accepted
- Office Hours
  - NPB 3.342
  - Online/In-person: Monday 1:00pm to 2:00pm; Tuesday 2pm to 4pm
  - and by appointment
- For e-mail contact, always include “CS4823” or “CS6643” in the subject line.

Class Information cont'd

- Class site:
  - I post my slides, supplement materials, class schedules and syllabus on the class site
  - Canvas
    - Assignments
- Read emails!
- Grader: Yuntong Zhang, yuntong.zhang@utsa.edu

Student Participation

- Attendance (I know it is hard)
  - Missing many lectures will result in an F grade
  - If you really having trouble traveling to SPI, please let me know.
- In-class discussion
- Ask questions
- Let me know what do you think about this course
- Any feedback is welcome