BITS Pilani
Pilani Campus

Introduction

K Hari Babu
Department of Computer Science & Information Systems

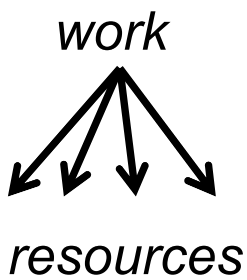
What is a parallel computer?

BITS Pilani

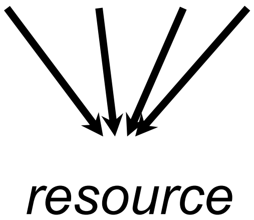
- A parallel computer is a type of computer that uses multiple processing elements, such as processors or cores, to perform computations simultaneously. These processing elements work together to solve a problem or perform a task faster than a single processing element would be able to on its own.

Parallelism vs. Concurrency

- Parallelism:
  - Use extra resources to solve a problem faster



- Concurrency:
  - Correctly and efficiently manage access to shared resources



Parallel Computing vs Distributed Computing

- Both try solving a problem with multiple processing elements.
- Main differences:
  - The level of communication and coordination between the processing elements or nodes.
    - In parallel computing, the processing elements are tightly coupled and share a common memory space. They communicate with each other frequently and quickly, and are often part of the same computer or device.
    - In distributed computing, on the other hand, the nodes are typically loosely coupled and have their own memory and processing power. They communicate over a network, and the communication is generally slower and less frequent.
  - Parallel computing is mainly used to increase the performance of a single application by dividing the work among multiple processors, while distributed computing is used to build large-scale systems that can provide services and share resources over a network.

Parallel Computing vs Distributed Computing

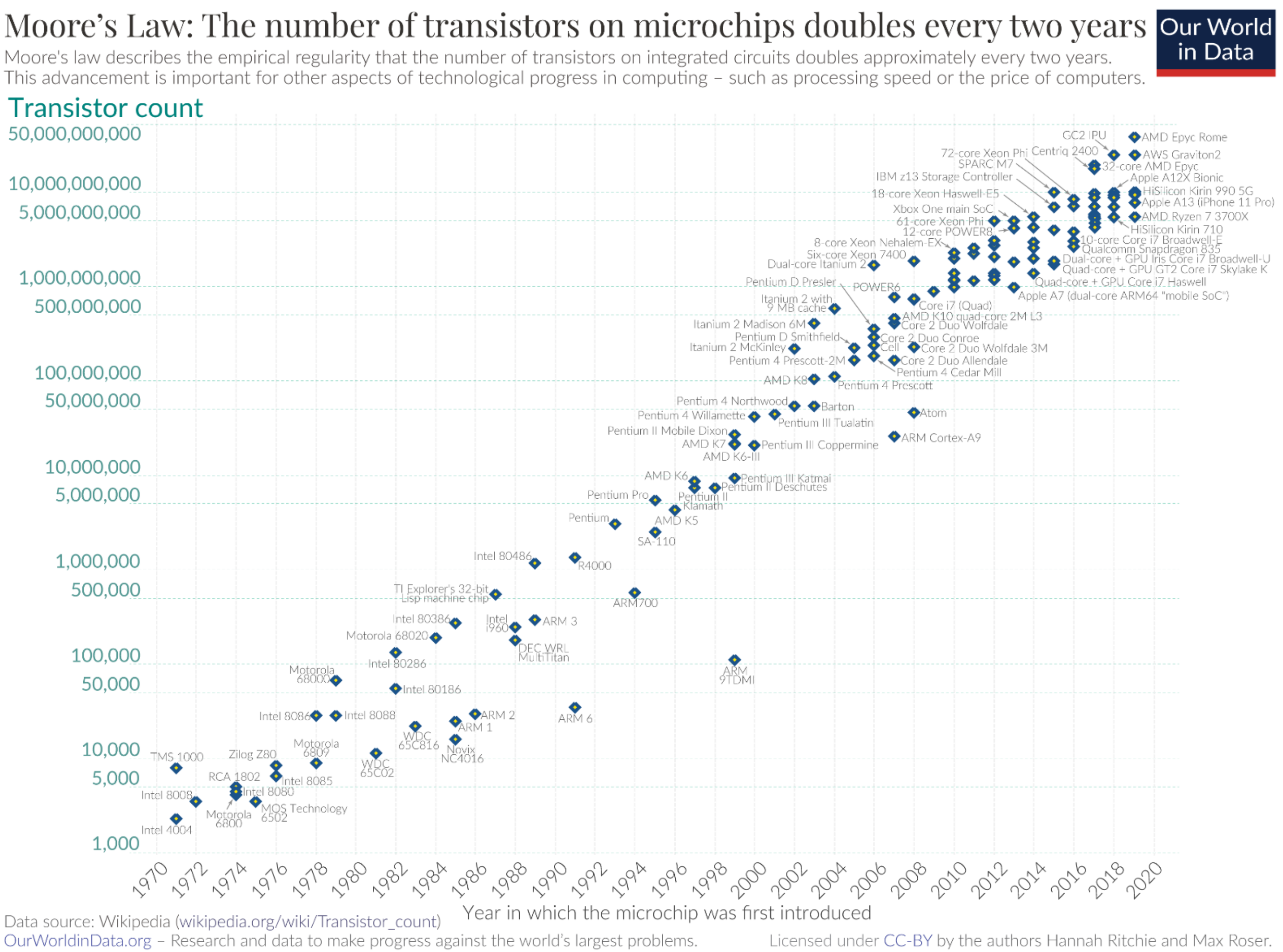
- Lines are blurring, but broadly:
  - Tightly coupled vs. Loosely coupled
  - Shared memory vs. Message passing
  - Bus/Switch vs. Network Interconnect
  - Monolithic vs. Composite
- Clusters form the dividing line
  - Treated as Parallel or Distributed depending on context.

Why Parallel or Distributed Computing?

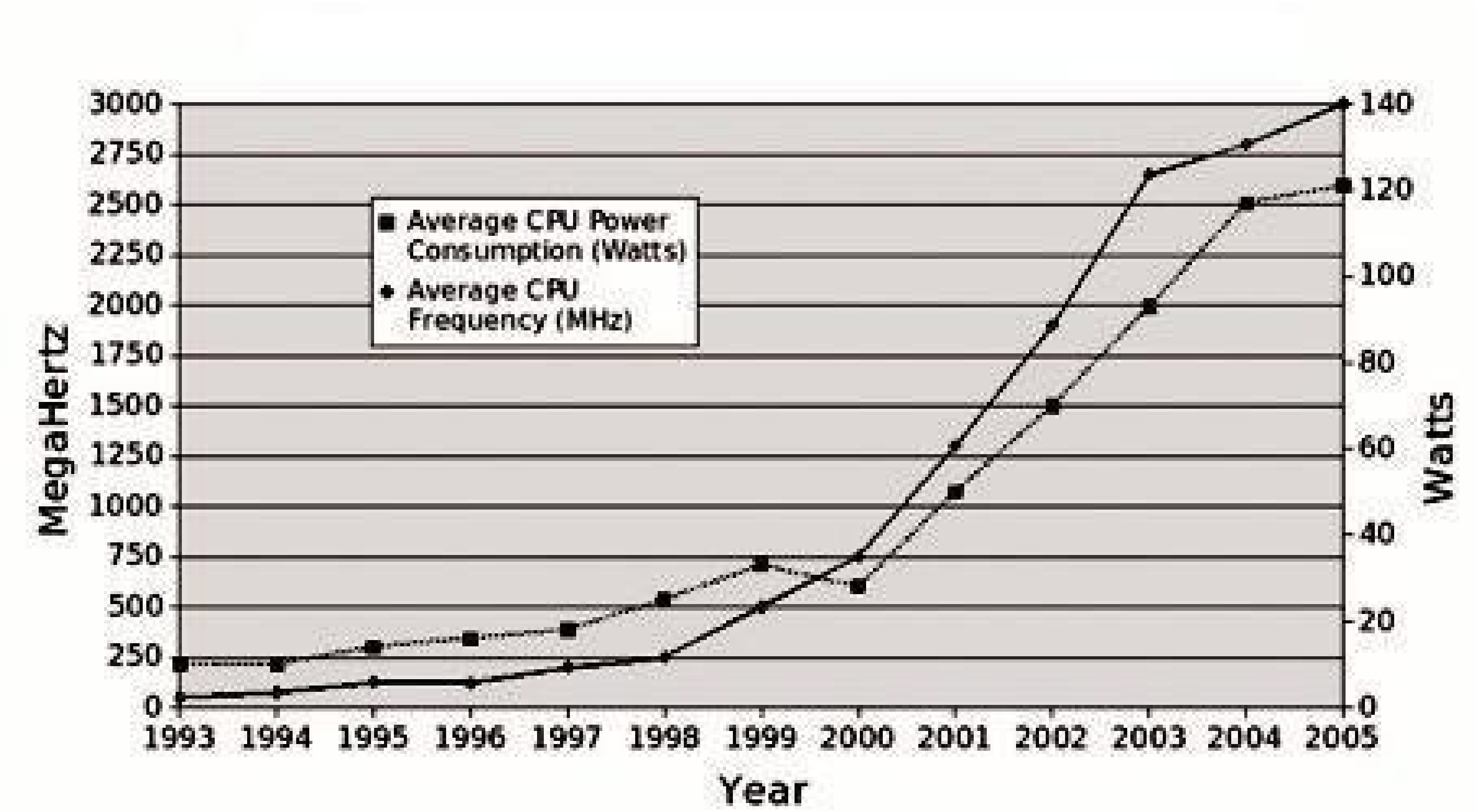
BITS Pilani

- Limitations of Sequential computer:
  - By using multiple processing elements or nodes, a computation can be divided into smaller parts that can be executed simultaneously, which can lead to a significant reduction in the time it takes to complete the computation.
- Handling large-scale data:
  - As data sets continue to grow in size, parallel and distributed computing provide a way to process, analyze, and store large amounts of data in a reasonable amount of time. With parallel and distributed computing, data can be divided into smaller chunks and processed by multiple processors or nodes, which can help to speed up the analysis process.
- Cost-effectiveness:
  - Parallel and distributed computing can also be cost-effective, as it allows organizations to use less powerful, less expensive processors or nodes to perform a computation, rather than relying on a single, expensive supercomputer.

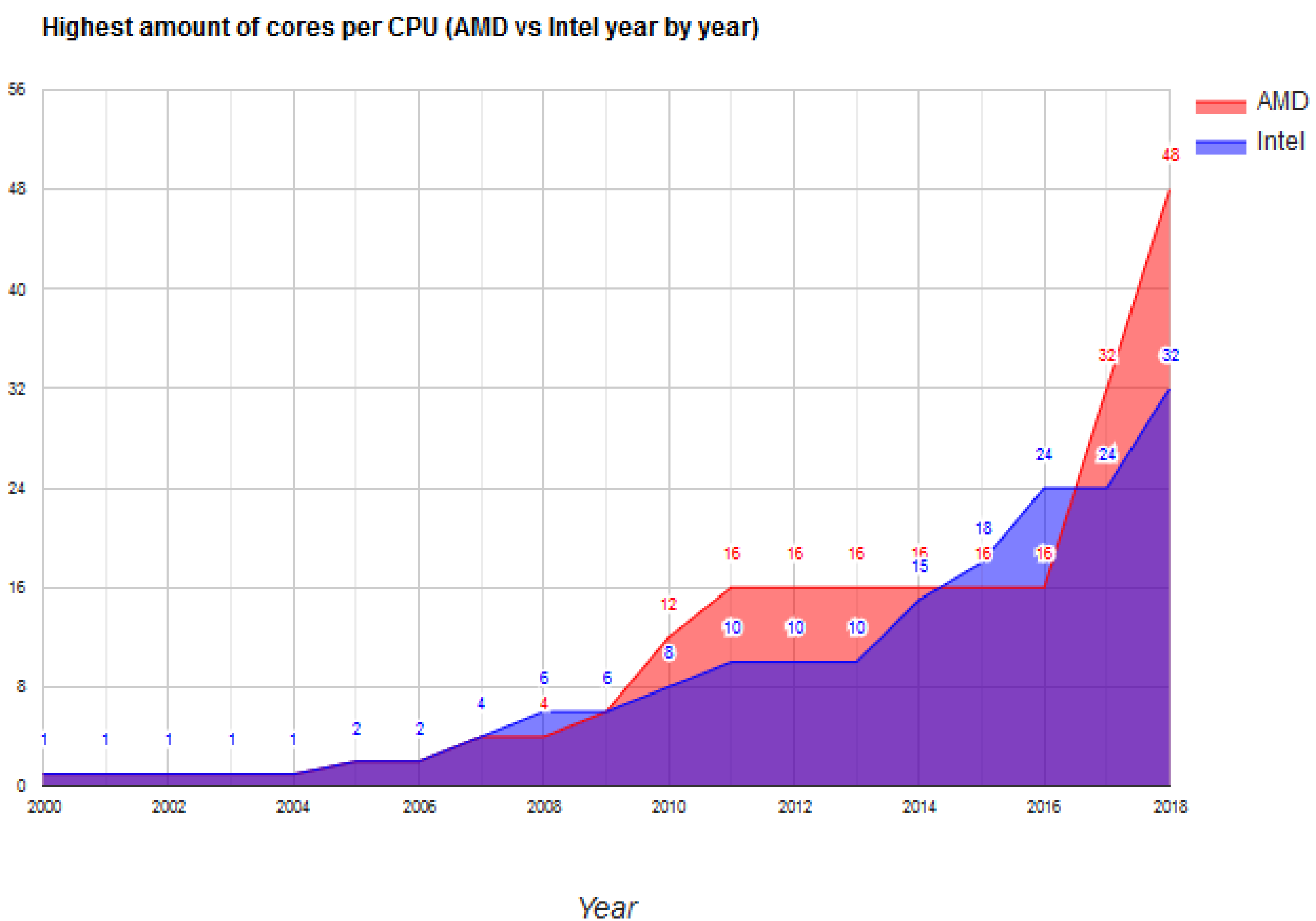
Moore's Law


Moore's Law: The number of transistors on microchips doubles every two years
Our World in Data
Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years.
This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.
Transistor count
50,000,000,000
10,000,000,000
5,000,000,000
1,000,000,000
500,000,000
100,000,000
50,000,000
10,000,000
5,000,000
1,000,000
500,000
100,000
50,000
10,000
5,000
1,000
1970
1972
1974
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Year in which the microchip was first introduced
Intel 4004
Intel 8008
TMS 1000
RCA 1802
Intel 8080
Motorola 6800
MOS Technology 6502
Zilog Z80
Intel 8085
Motorola 6809
Intel 8086
Intel 8088
Motorola 68000
Intel 80186
Intel 80286
WDC 65C02
WDC 65C816
Motorola 68020
Intel 80386
TI Explorer's 32-bit Lisp machine chip
ARM 1
ARM 2
Novix NC4016
Intel i960
DEC WRL MultiTitan
ARM 3
Intel 80486
R4000
ARM 6
ARM700
Pentium
SA-110
AMD K5
Pentium Pro
AMD K6
Pentium II Klamath
Pentium II Deschutes
Pentium III Katmai
ARM 9TDMI
AMD K6-III
AMD K7
Pentium II Mobile Dixon
Pentium III Coppermine
Pentium 4 Willamette
Pentium III Tualatin
Pentium 4 Northwood
Barton
AMD K8
Pentium 4 Prescott
Pentium 4 Prescott-2M
Itanium 2 McKinley
Pentium D Smithfield
Itanium 2 Madison 6M
Itanium 2 with 9 MB cache
Pentium D Presler
Dual-core Itanium 2
POWER6
Core 2 Duo Conroe
Cell
Core 2 Duo Wolfdale 3M
Core 2 Duo Allendale
Pentium 4 Cedar Mill
Core 2 Duo Wolfdale
AMD K10 quad-core 2M L3
Core i7 (Quad)
Atom
ARM Cortex-A9
Six-core Xeon 7400
8-core Xeon Nehalem-EX
12-core POWER8
61-core Xeon Phi
Xbox One main SoC
18-core Xeon Haswell-E5
IBM z13 Storage Controller
SPARC M7
72-core Xeon Phi
Centriq 2400
GC2 IPU
AMD Epyc Rome
AWS Graviton2
32-core AMD Epyc
Apple A12X Bionic
HiSilicon Kirin 990 5G
Apple A13 (iPhone 11 Pro)
AMD Ryzen 7 3700X
HiSilicon Kirin 710
10-core Core i7 Broadwell-E
Qualcomm Snapdragon 835
Dual-core + GPU Iris Core i7 Broadwell-U
Quad-core + GPU GT2 Core i7 Skylake K
Quad-core + GPU Core i7 Haswell
Apple A7 (dual-core ARM64 "mobile SoC")
Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)
OurWorldinData.org – Research and data to make progress against the world's largest problems.
Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.

Heat problem



Number of cores per CPU



Nowadays, all computers are essentially parallel

1. Parallelism is present deep in the processor microarchitecture
   - Pipelining
2. Any commercial computer, tablet, and smartphone contain a processor with multiple cores
3. Many servers contain several multi-core processors
4. Even consumer-level computers contain graphic processors capable of running hundreds or even thousands of threads in parallel

Reasons for making modern computers parallel

- First, it is not possible to increase processor and memory frequencies indefinitely
  - to increase computational power of computers, new architectural and organizational concepts are needed
- Second, power consumption rises with processor frequency while the energy efficiency decreases.
  - if the computation is performed in parallel at lower processor speed, this can be avoided
- Finally, parallelism has become a part of any computer and this is likely to remain unchanged

Approaches to parallel programming

- Three different approaches to parallel programming exist:
  - Threads model for shared memory systems
  - Message passing model for distributed systems
  - Stream-based model for GPUs

Why is Parallel Computing Hard?

- Communication and coordination
- Portability – knowledge of underlying architecture often required

Why Study Parallel computing?

BITS Pilani

- It's a critical skill for high-performance computing: With the increasing demand for computing power in many fields, such as scientific simulation, data analytics, artificial intelligence and machine learning, the ability to write parallel programs is becoming an essential skill for computer scientists and engineers.

Why Study Parallel computing?

- It's needed for large scale data processing: The amount of data being generated and collected is growing rapidly, and the ability to process this data in parallel is critical for many applications such as big data analytics and machine learning
- It's needed to solve large and complex problems: Parallel computing allows for solving large and complex problems that would be otherwise intractable with a single processor.
- It's needed to take advantage of new technologies: Advances in hardware technologies, such as GPUs and many-core processors, as well as new parallel programming models and languages, such as OpenMP, MPI, CUDA, OpenCL, and others, are making parallel computing more accessible and easier to use.

Applications

BITS Pilani

- Parallel computing has a wide range of applications, including:
  - Scientific simulations: can be used to simulate complex physical systems, such as weather patterns, fluid dynamics, and chemical reactions.
  - Data analysis: can be used to process and analyze large sets of data, such as in data mining and machine learning.
  - Computer graphics: can be used to render complex 3D images and animations in real-time.
  - Drug discovery: can be used to simulate the interactions between potential drug compounds and biological systems, helping researchers to identify new drugs more quickly.
  - Climate modeling: can be used to simulate and predict future climate patterns and to understand the causes and effects of climate change.

Applications

BITS Pilani

- Parallel computing has a wide range of applications, including:
  - Computer Vision: can be used to process images and videos in real-time, using techniques such as object detection, image recognition, and facial recognition.
  - Artificial Intelligence: can be used to train and deploy large AI models, such as deep learning neural networks, and to speed up the inference process.
  - Financial modeling: can be used to perform complex financial simulations, such as risk analysis and portfolio optimization.

Q&A

Thank You