CS546 Parallel and Distributing Processing

- Instructor: Professor Xian-He Sun
  - Email: sun@iit.edu, Phone: (312) 567-5260
  - Office hours: 4:30pm-5:30pm Tuesday, Thursday
    SB235C
- TA: Xiaoyang Lu
  - Email: xlu40@hawk.iit.edu
  - Office hour: SB003
- Blackboard:
  - http://blackboard.iit.edu
- Additional Web site
  - http://www.cs.iit.edu/~sun/cs546.html

What This Course Is About

- Parallelism
  - What is parallelism?
  - What can be parallelized?
  - Inhibitors and degradation of parallelism: dependences
- Different patterns of parallelism
  - Regular data parallelism
  - Irregular data parallelism
  - Task queue based parallelism
  - Pipeline parallelism
- Application and Algorithm

Question:

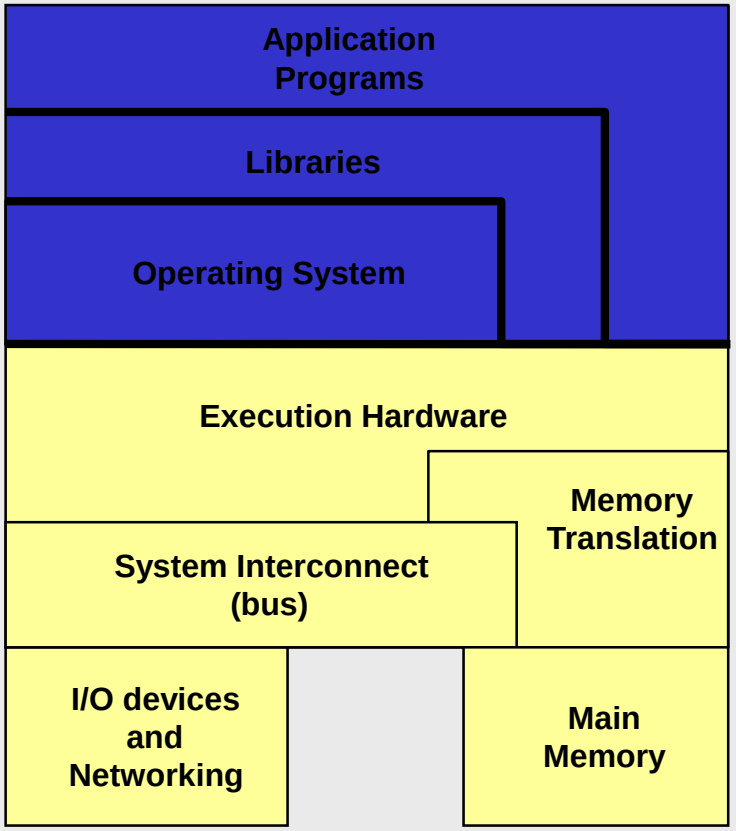
- What is the data-centric view?
- What is the machine?
- What is the machine layers?

The “Machine”

- Different perspectives on what the *Machine* is:
- OS developer

Instruction Set Architecture

- ISA
- Major division between hardware and software



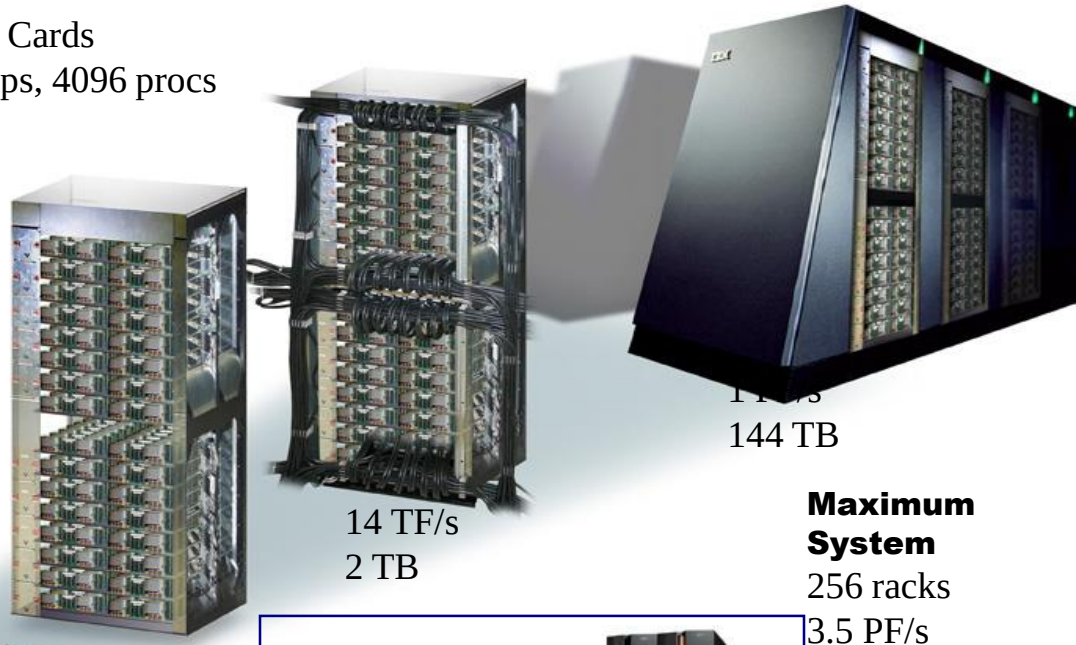
Evolution of Computing:
The biggest machine becomes even bigger

Petaflops System
72 Racks

Rack Cabled 8x8x16

IBM BG/P

32 Node Cards
1024 chips, 4096 procs



144 TB

Source: ANL ALCF

Node Board
(32 chips 4x4x2)
32 compute, 0-2 IO cards

14 TF/s
2 TB

Maximum System
256 racks
3.5 PF/s
512 TB

Compute Card
1 chip, 20 DRAMs

435 GF/s
64 GB

Chip
4 cores

13.6 GF/s
2.0 GB DDR
Supports 4-way SMP

850 MHz
8 MB EDRAM

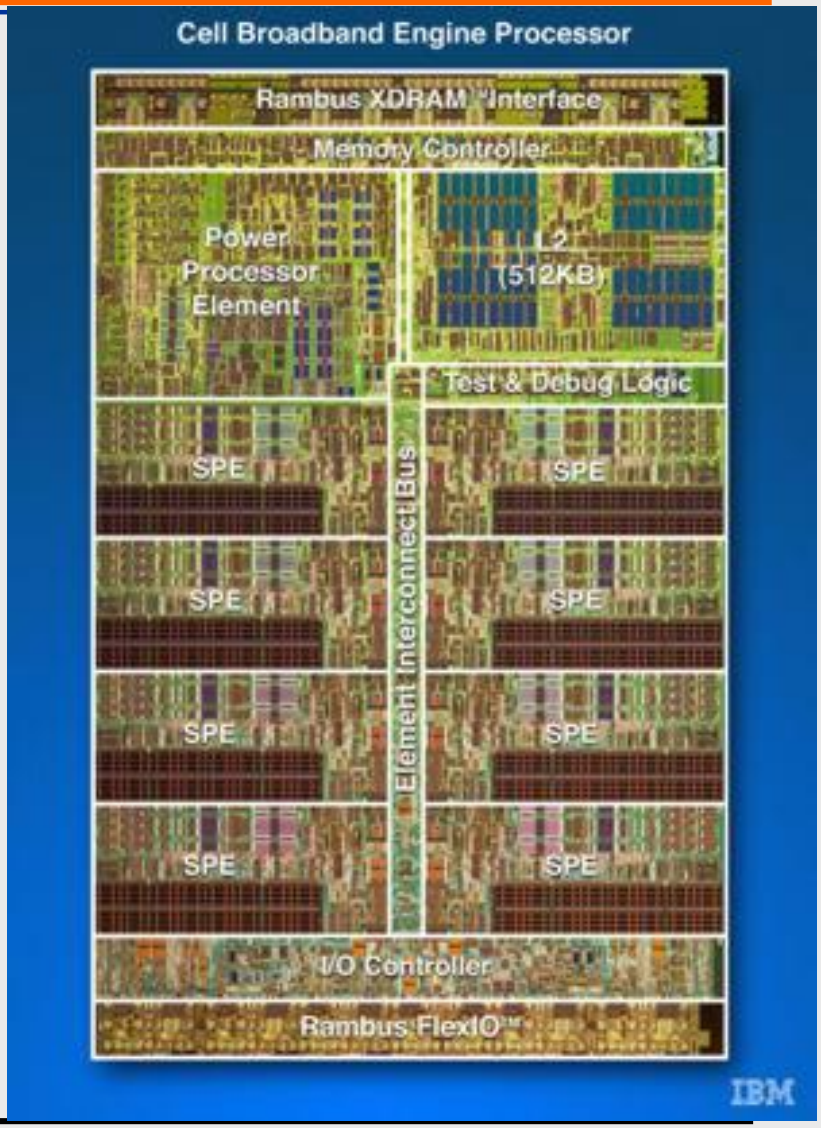
Front End Node / Service Node
System p Servers
Linux SLES10

HPC SW:
Compilers
GPFS
ESSL
Loadleveler

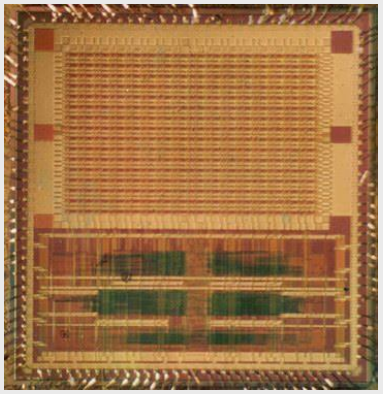
Multicore Add Another Dimension

IBM Multicore

- Cell
  - 1 PPE and 8 SPEs
  - Shared L2 cache
  - EIB
- Power6
  - Dual core
  - 5 GHz

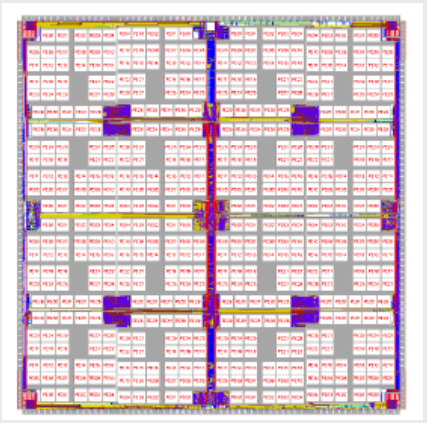


Many-Core Technology is Available

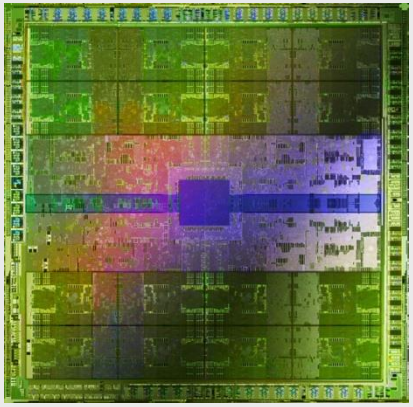


Kilocore: 256-core prototype
By Rapport Inc.

GeForce RTX 2080 SUPER:
3072 CUDA cores, by NVDIA



GRAPE-DR chip:
512-core, By Japan

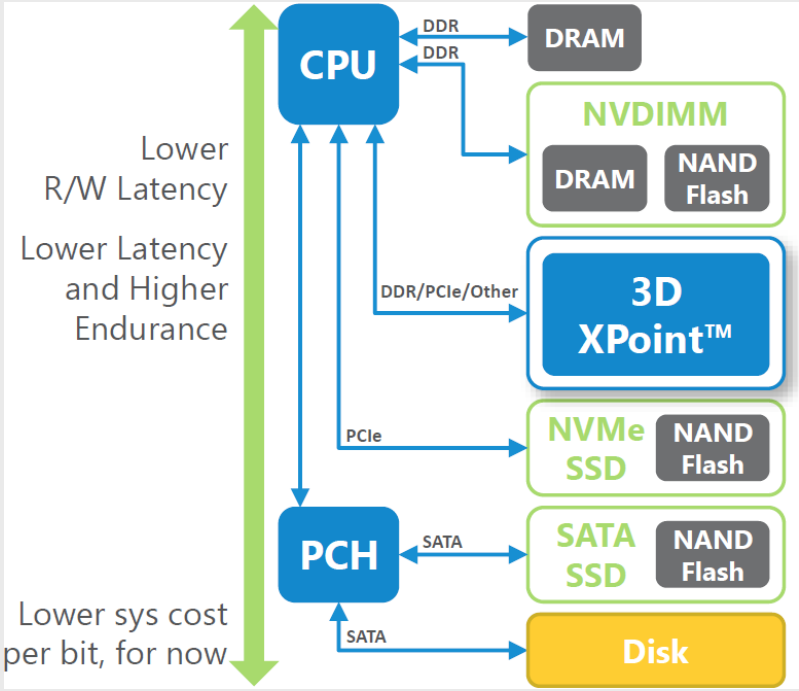


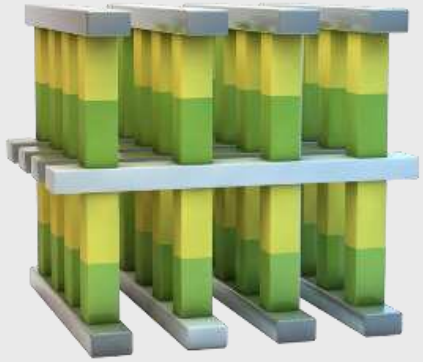
NVIDIA Fermi: 512 CUDA cores

Quadro M6000: 3,072 cores,
By NVDIA.

GRAPE-DR testboard

Nonvolatile Memories in Server Architectures





- 3D XPoint™ technology provides the benefit in the middle
- It is considerably faster than NAND Flash
- Performance can be realized on PCIe or DDR buses
- Lower cost per bit than DRAM while being considerably more dense

The View of Future Computing

Human-centered

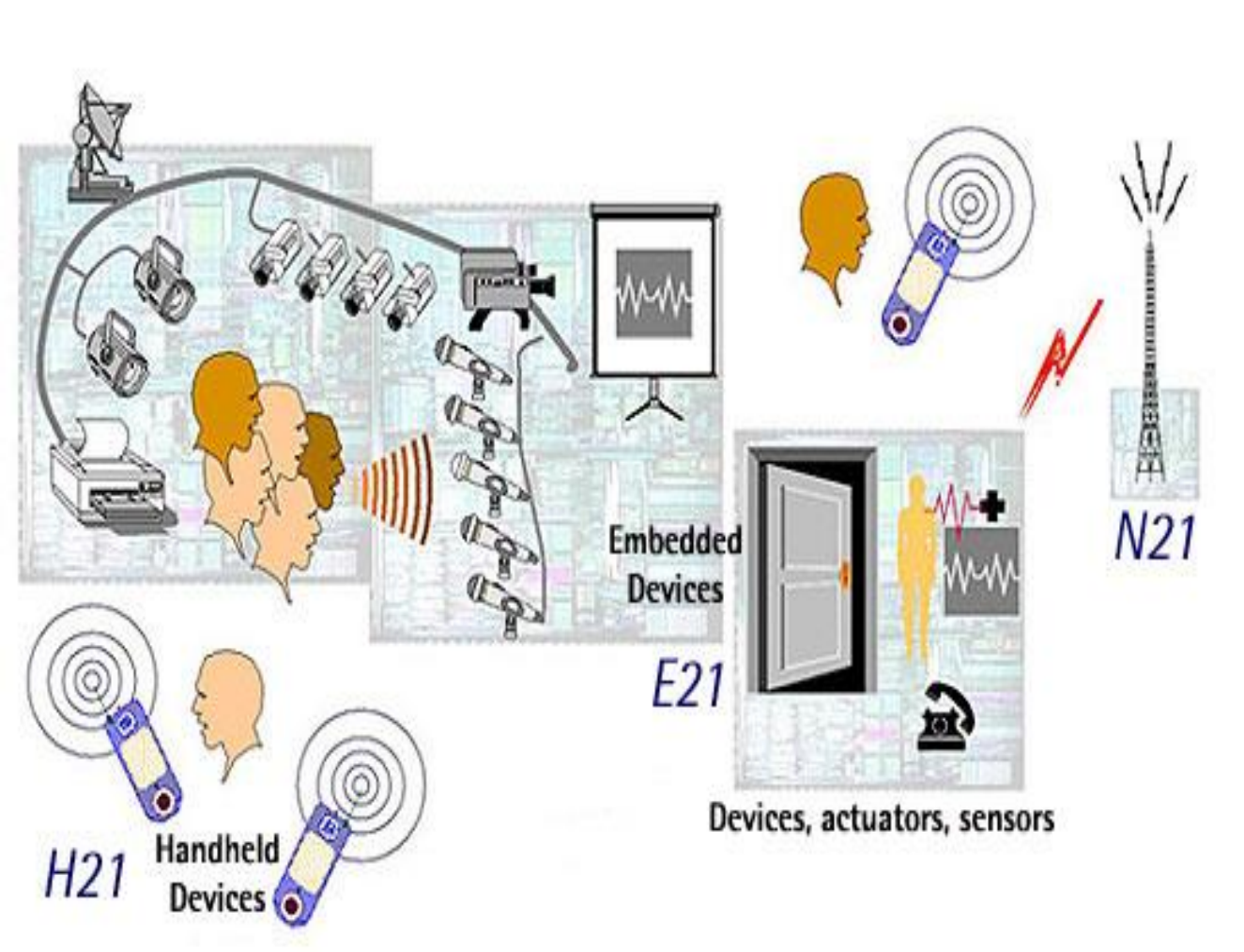
They are connected to form `smart space'

Devices become smaller and powerful

Cloud link `smart spaces' to support `global smartness'

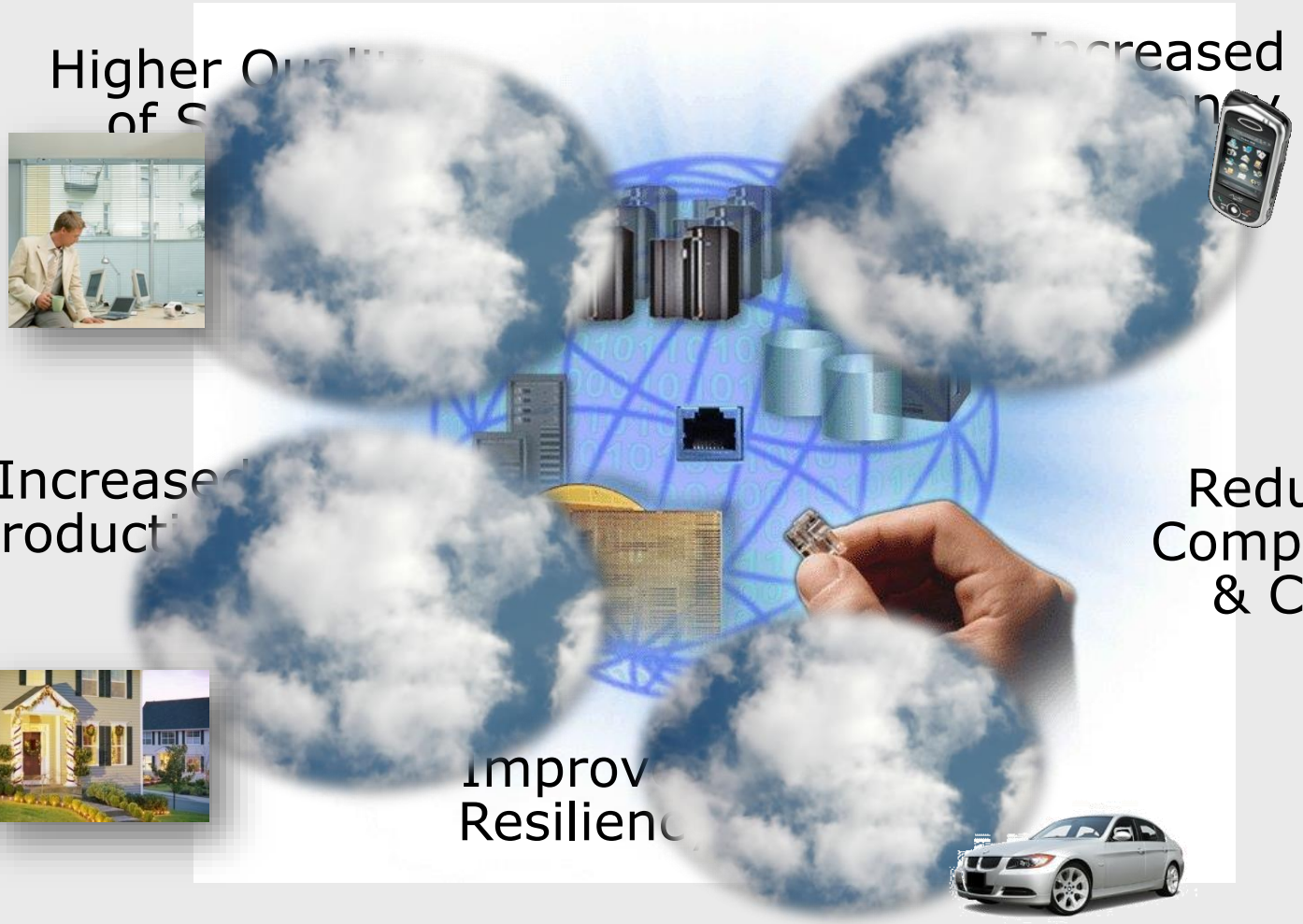
A device is an entry of the cyber world

Cyber Physical System – extended Smart Space

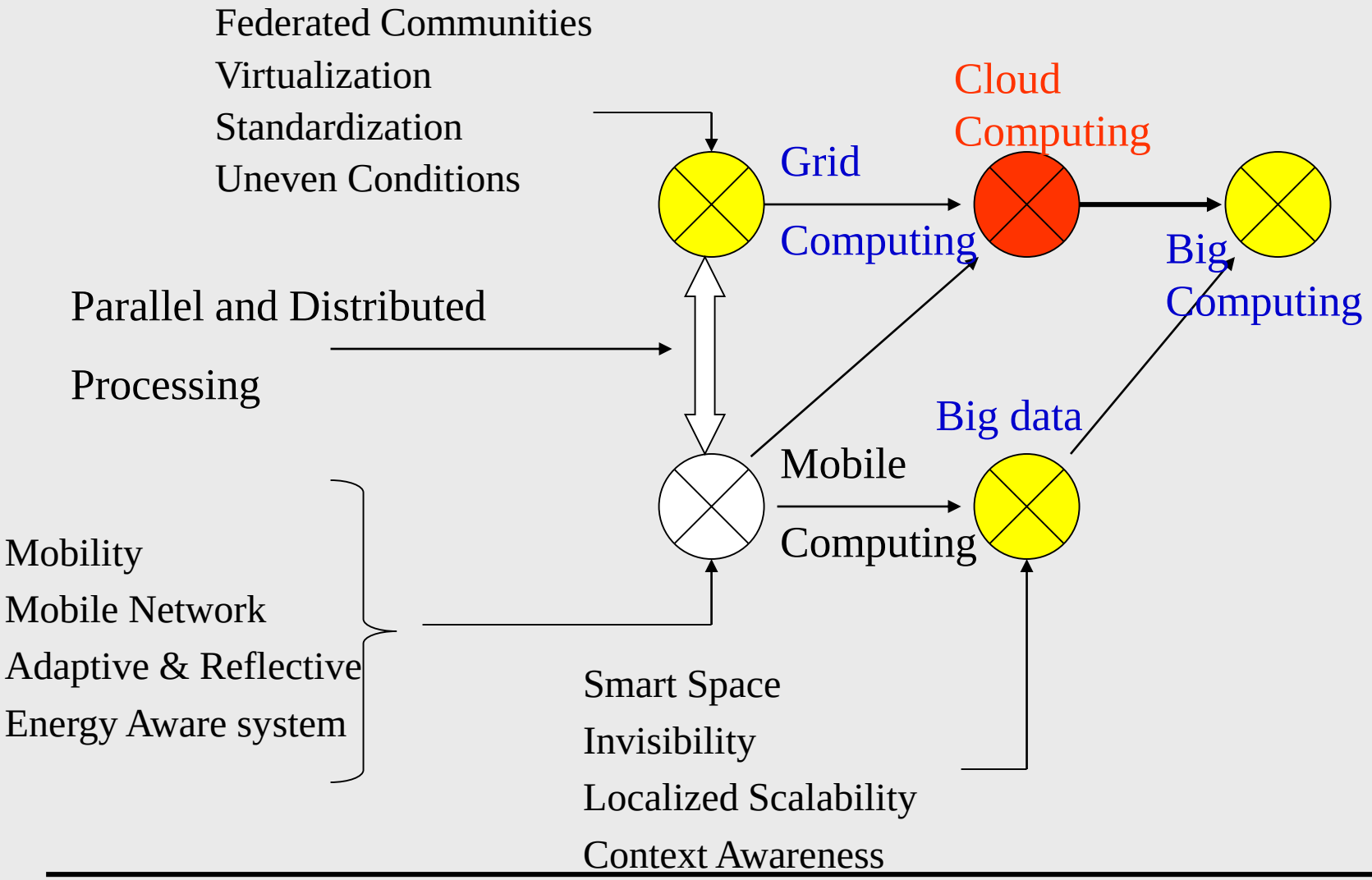


Edge Computing

Mimic the electrical power grid



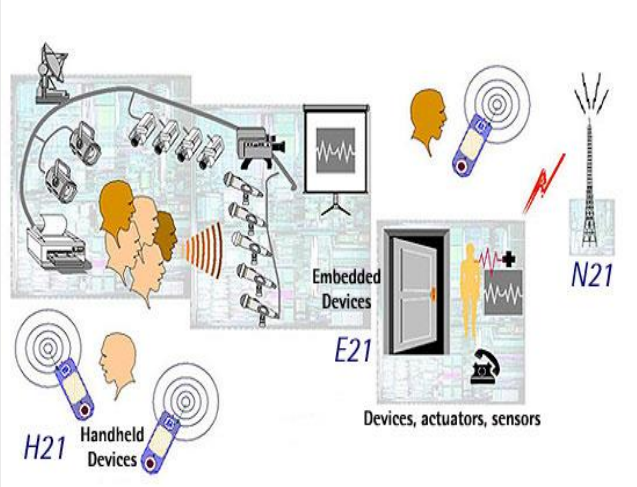
Evolution of Computing



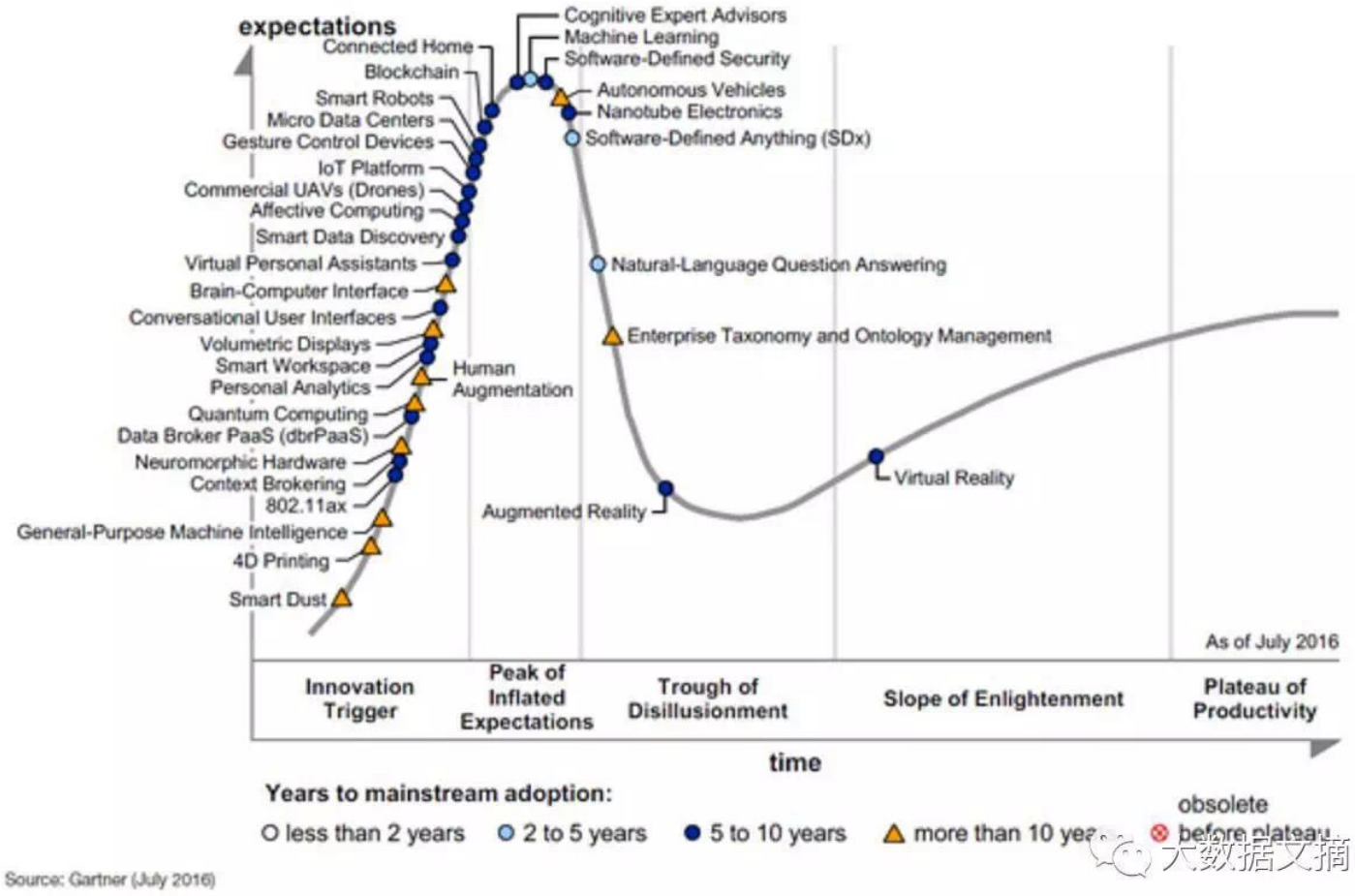
Pervasive Computing

- Big Computer becomes even bigger, Bigger computing power
- Small Computing becomes even smaller, Smart Space
- Smart Space, Sensor Network, bigdata
- Smart Space, Cyber Physic Systems
- Context Aware, Smartness (AI)
- AI is forward by big data and bigger computing power

  - Today (software/software connection): Cloud, CPS
  - Tomorrow (machine/human): pervasive



The Perspective of AI, or any others



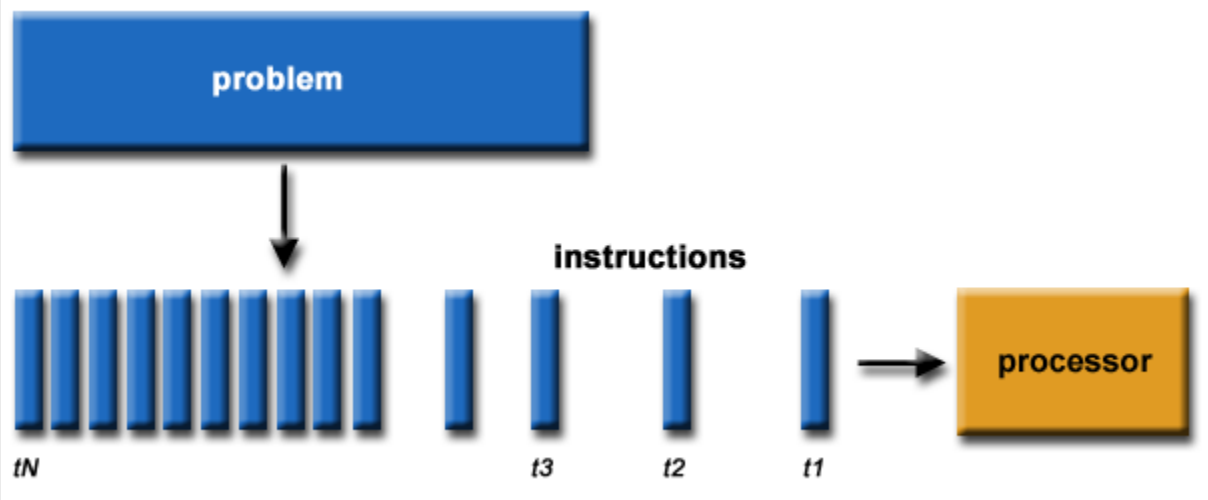
Any Questions?

Homework: Read through CS546 web site, especially the project part

Overview of Parallel Computing

Serial Computing

- Traditional, software has been written for serial computation
  - A problem is broken into a discrete series of inst.
  - Inst. are executed sequentially (one after another)
  - Executed on a single processor
  - Only one inst. executes at any moment in time

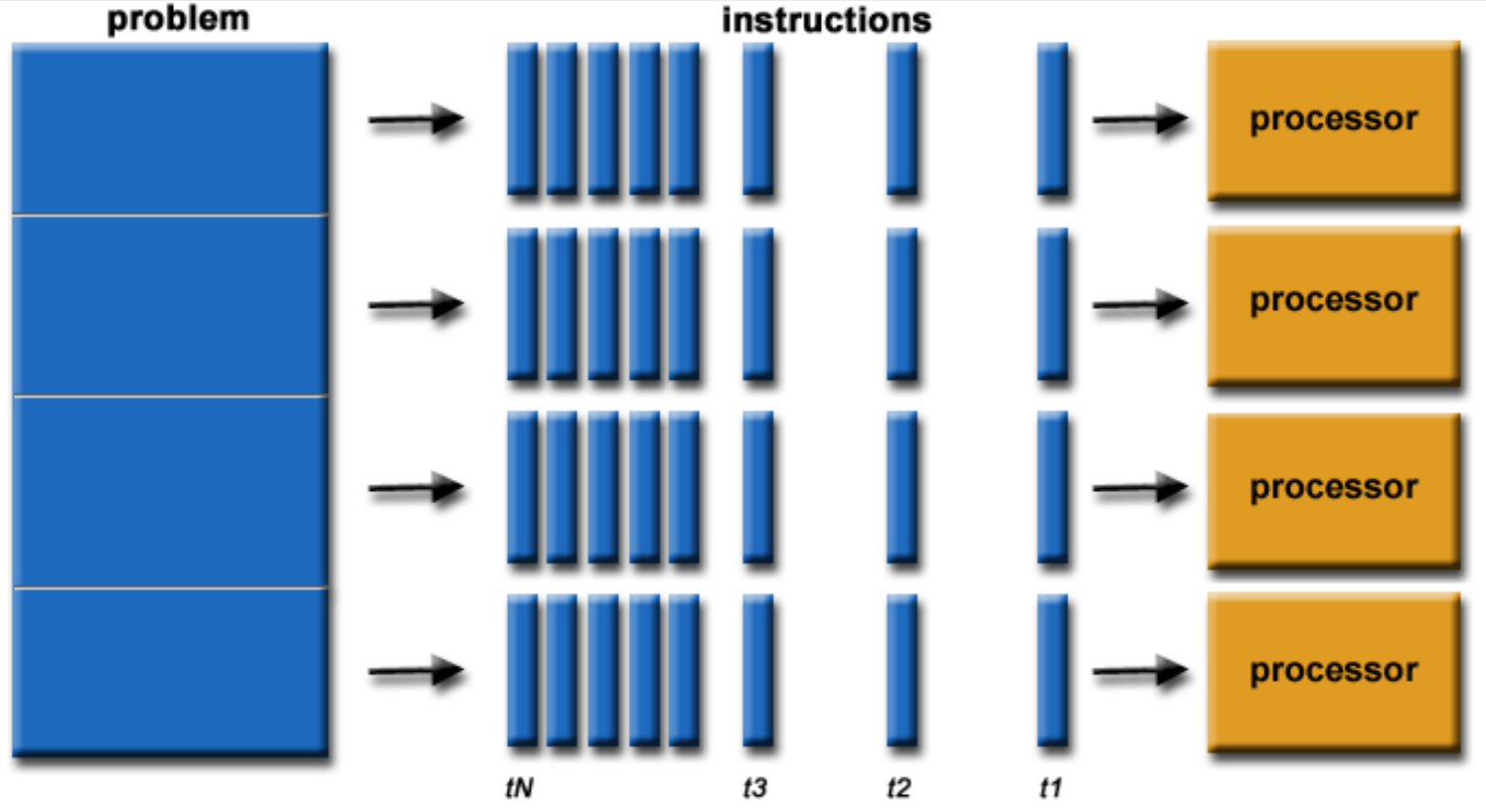


What is Parallel Processing

- Parallel Processing
  - Several working entities work together toward a common goal
- Parallel Processing
  - A kind of information processing that emphasizes the concurrent manipulation of data elements belonging to one or more processes solving a single problem
- Parallel Computer
  - A computer designed for parallel processing

- Supercomputer (high performance computer, high end computer, advanced computer)
  - A general-purpose computer capable of solving individual problems at extremely high computation speed
- New Terms
  - Cloud computing, Date intensive computing
  - Big computing (supercomputing and super data processing)

What is Parallel Computing

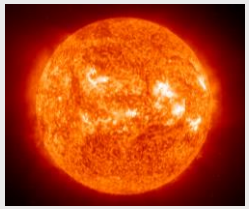


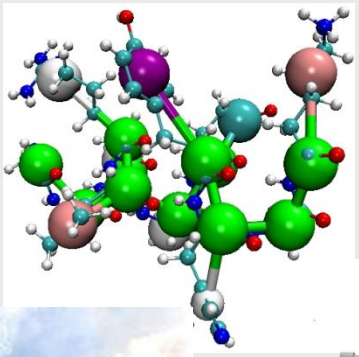
The Need for Parallel Processing

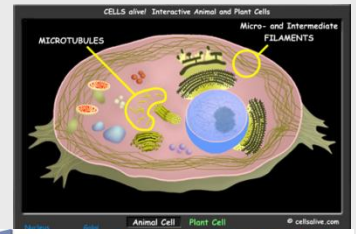
- The Fastest Computer Is Never Fast Enough
- Applications
  - Numerical computation: weather modeling, fluid flows, simulation, life science & medical applications
  - Real time multifaceted problems: speech recognition, image processing, computer vision
  - Large memory and I/O intensive problems: database & transaction systems, data mining, Web search
  - Graphics and design systems: CAD, virtual world
  - AI: knowledge based systems, inferencing engines
  - New applications, e.g. life science, remote surgery

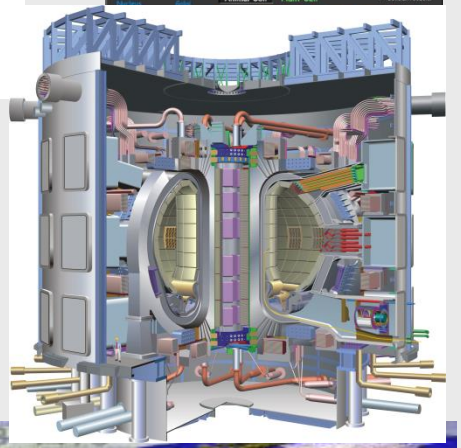
Application of Supercomputers

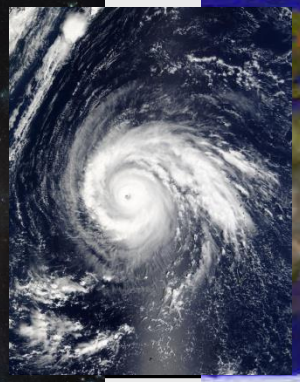
- Physical Sciences
- Technology
- Biology and Medical Science
- Energy
- Meteorology and Climate
- Materials and Nanotechnology
- National Security

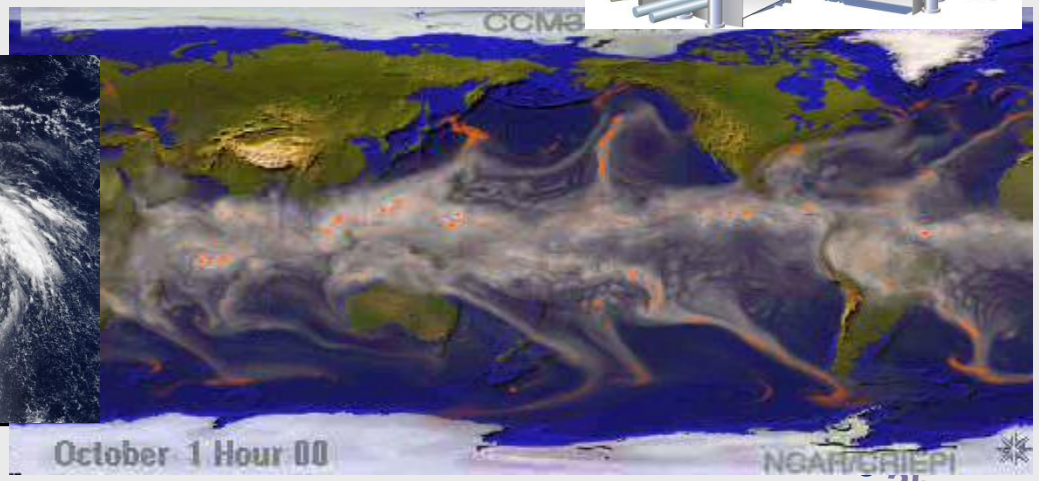








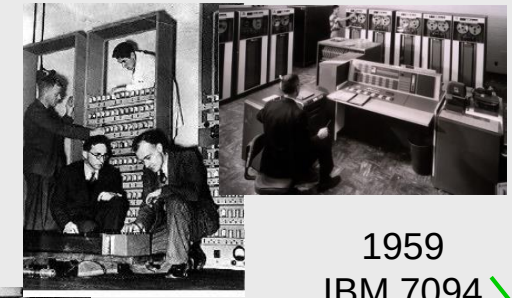


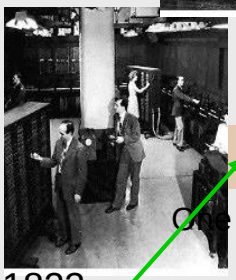


Parallel Computers

- ~25 years ago
  - $1*10^6$ Floating Point Ops/sec (Mflop/s)
    - Scalar based
- ~15 years ago
  - $1*10^9$ Floating Point Ops/sec (Gflop/s)
    - Vector & shared memory computing, bandwidth aware
    - Block partitioned, latency tolerant
- ~5 years ago
  - $1*10^{12}$ Floating Point Ops/sec (Tflop/s)
  - Highly parallel, distributed processing, message passing, network based
  - Data decomposition, communication/computation
- Today
  - $1*10^{15}$ Floating Point Ops/sec (Pflop/s) (in fact, 100 Pflop/s)
  - Many more levels MH, combination/grids &HPC
  - More adaptive, LT and BW aware, fault tolerant, extended precision, …

Evolution of HPC





1959
IBM 7094

1976
Cray 1

1991
Intel Delta

1996
T3E

2003
Cray X1

1949
Edsac

1 10^3 10^6 10^9 10^{12} 10^{15}

One OPS KiloOPS MegaOPS GigaOPS TeraOPS PetaOPS

1823
Babbage Difference Engine

1943
Harvard Mark 1

1951
Univac 1

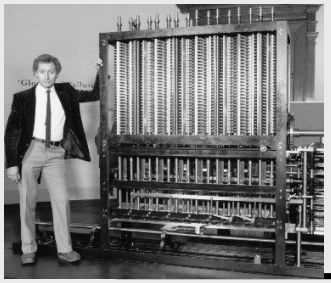
1964
CDC 6600

1982
Cray XMP

1988
Cray YMP

1997
ASCI Red

2001
Earth Simulator



TOP 10 Machines (6/2019)

Rank	Site	Computer	cores	TFlop/s	Country
1	Oak Ridge National Laboratory	Summit - IBM Power System	2414592	148600	United States
2	Lawrence Livermore National Laboratory	Sierra - IBM Power System	1572480	94640	United States
3	National Supercomputing Center in Wuxi	Sunway TaihuLight - Sunway MPP	10649600	93014.6	China
4	National Super Computer Center in Guangzhou	Tianhe-2A - TH-IVB-FEP Cluster	4981760	61444.5	China
5	Texas Advanced Computing Center/Univ. of Texas	Frontera - Dell C6420,	448448	23516.4	United States
6	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC50	387872	21230	Switzerland
7	Los Alamos National Lab	Trinity - Cray XC40	979072	20158.7	United States
8	National Institute of Advanced Industrial Science and Technology (AIST)	AI Bridging Cloud Infrastructure (ABCI) Fujitsu	391680	19880	Japan
9	Leibniz Rechenzentrum	SuperMUC-NG - ThinkSystem	305856	19476.6	Germany
10	DOE/NNSA/LLNL	Lassen - IBM Power System	288288	18200	United States

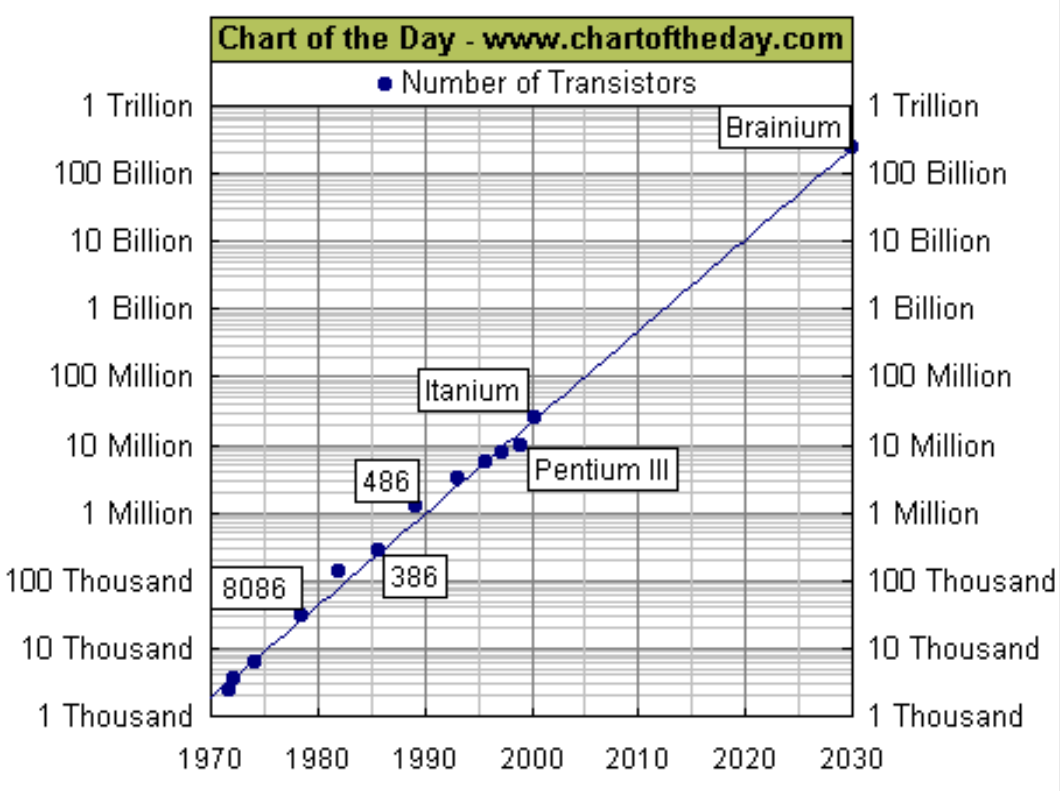
http://www.top500.org/list/2019/06/

Where Does Performance Come From?

- Device Technology
  - Logic switching speed and device density
  - Memory capacity and access time
  - Communications bandwidth and latency
- Computer Architecture
  - Instruction issue rate
    - Execution pipelining
    - Reservation stations
    - Branch prediction
    - Cache management
  - Parallelism
    - Parallelism – number of operations per cycle per processor
      - Instruction level parallelism (ILP)
      - Vector processing
    - Parallelism – number of processors per node
    - Parallelism – number of nodes in a system

Gordon Moore's Law

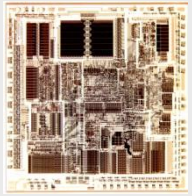
"the number of transistors that can be fabricated on a single integrated circuit at a reasonable cost doubles every year…"



- How?
  - Material techniques such as extreme ultraviolet lithography (<100 nm)
- Impact:
  - Increase in manufacturing yield, more dies per wafer
  - Smaller transistors consume less power => Higher speed for same power per unit area
  - More complex devices can be created in same die area
- Corollary
  - Processor speed doubles at same rate

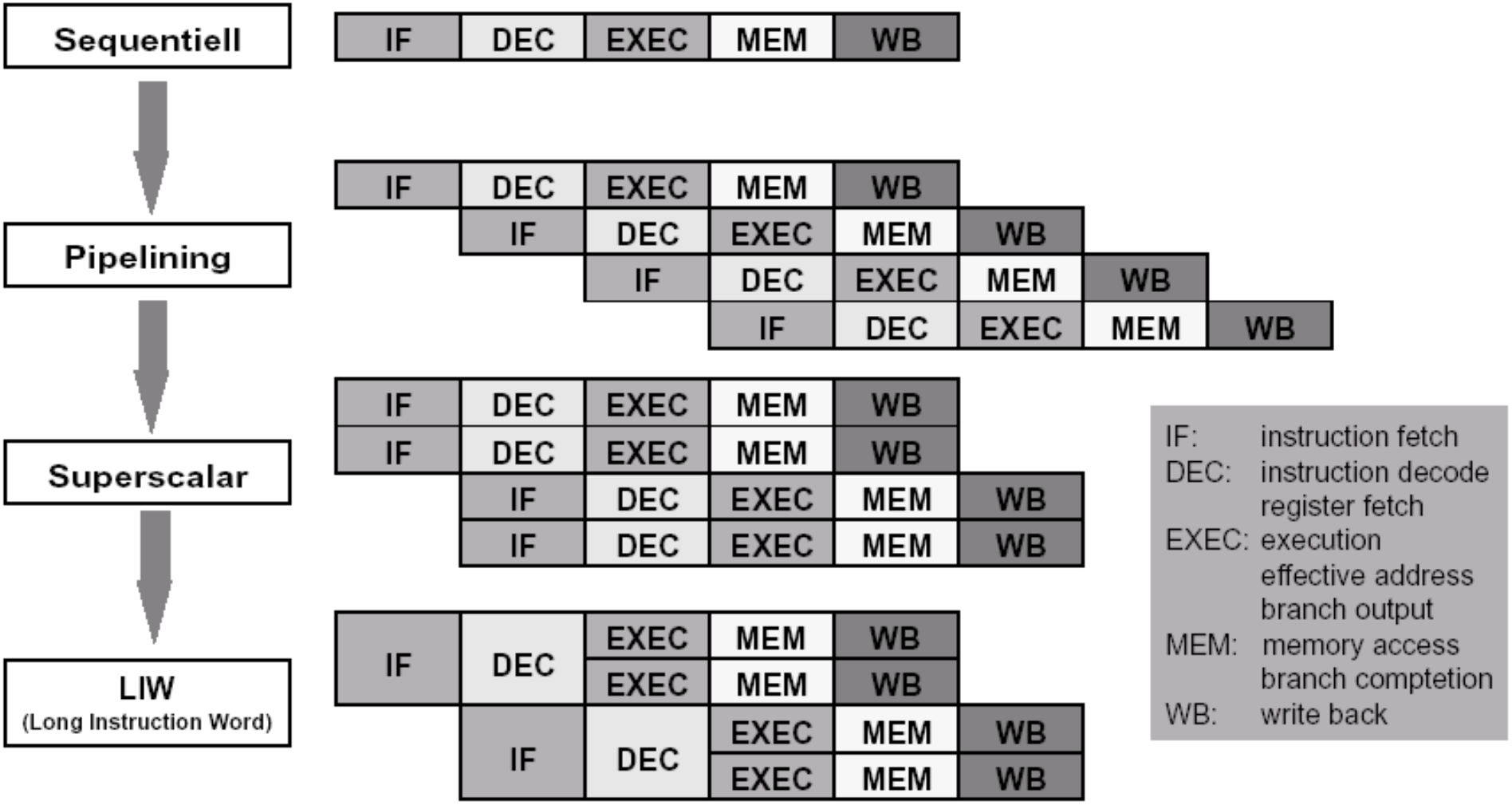
CPUs: Archaic (Nostalgic) v. Modern (Newfangled)

- 1982 Intel 80286
- 12.5 MHz
- 2 MIPS (peak)
- Latency 320 ns
- 134,000 xtors, 47 mm^2
- 16-bit data bus, 68 pins
- Microcode interpreter, separate FPU chip
- (no caches)

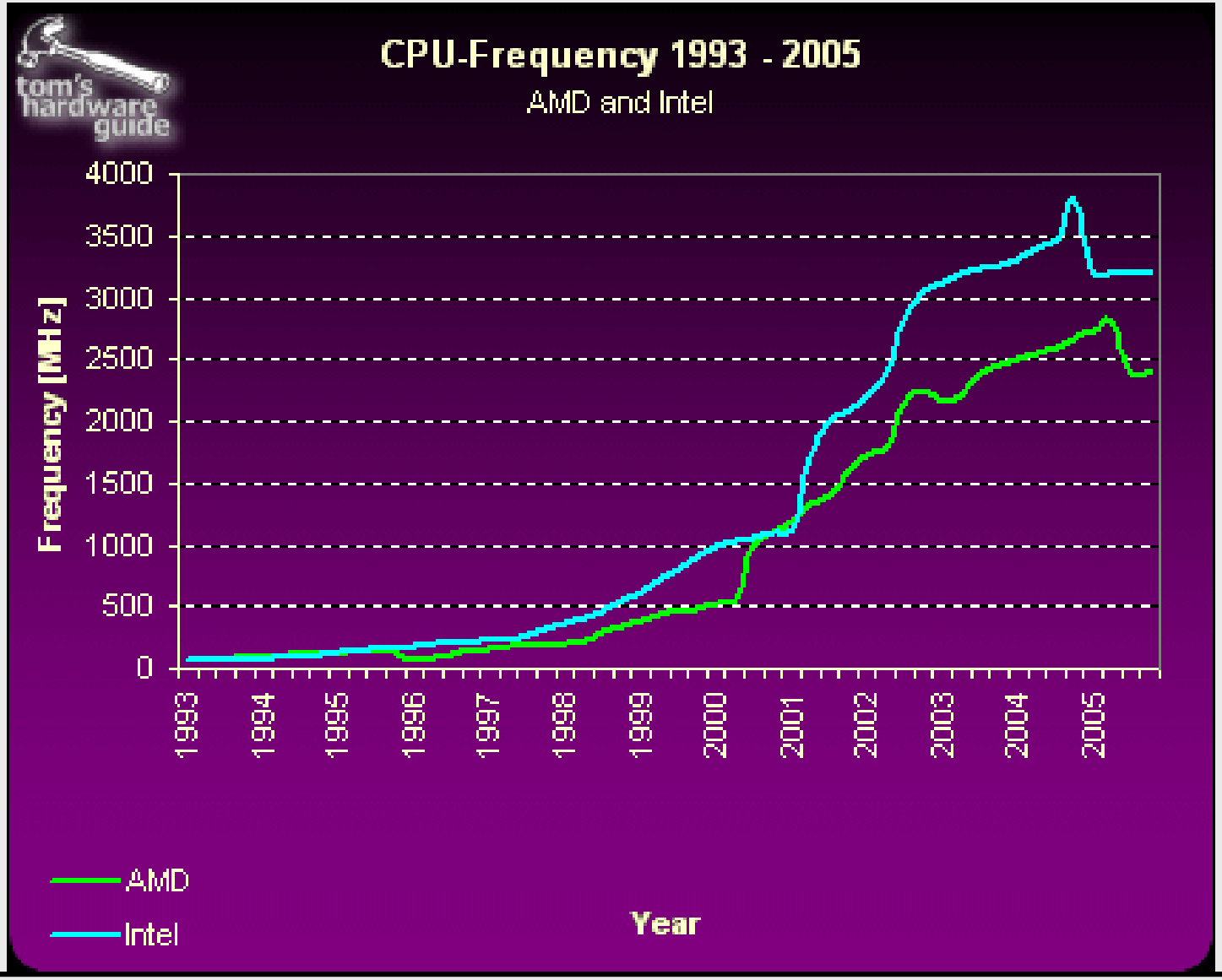


- 2001 Intel Pentium 4
- 1500 MHz **(120X)**
- 4500 MIPS (peak) **(2250X)**
- Latency 15 ns **(20X)**
- 42,000,000 xtors, 217 mm^2
- 64-bit data bus, 423 pins
- 3-way superscalar, Dynamic translate to RISC, Superpipelined (22 stage), Out-of-Order execution
- On-chip 8KB Data caches, 96KB Instr. Trace cache, 256KB L2 cache

Parallelism within a Processor

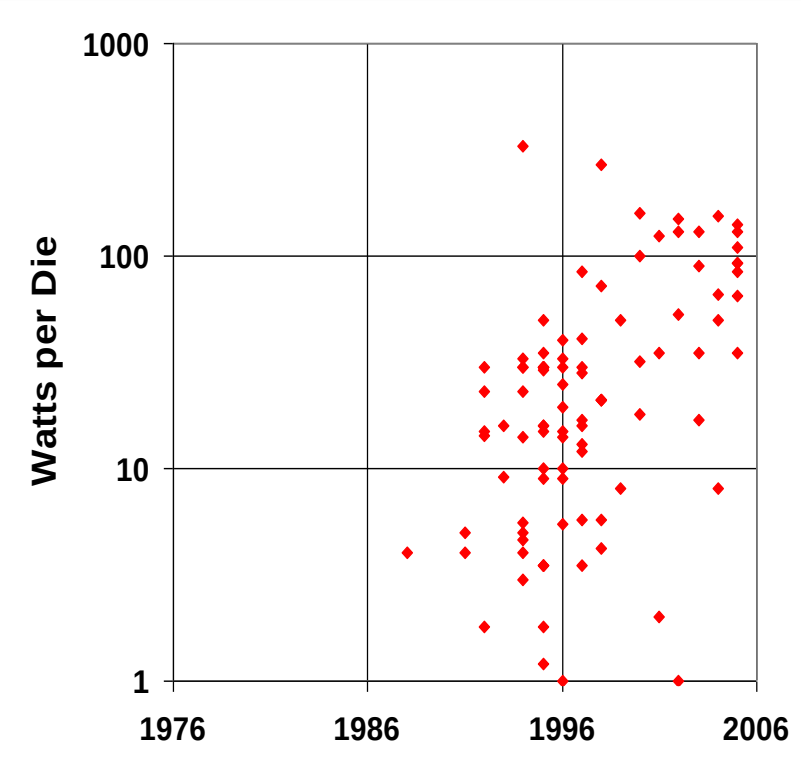


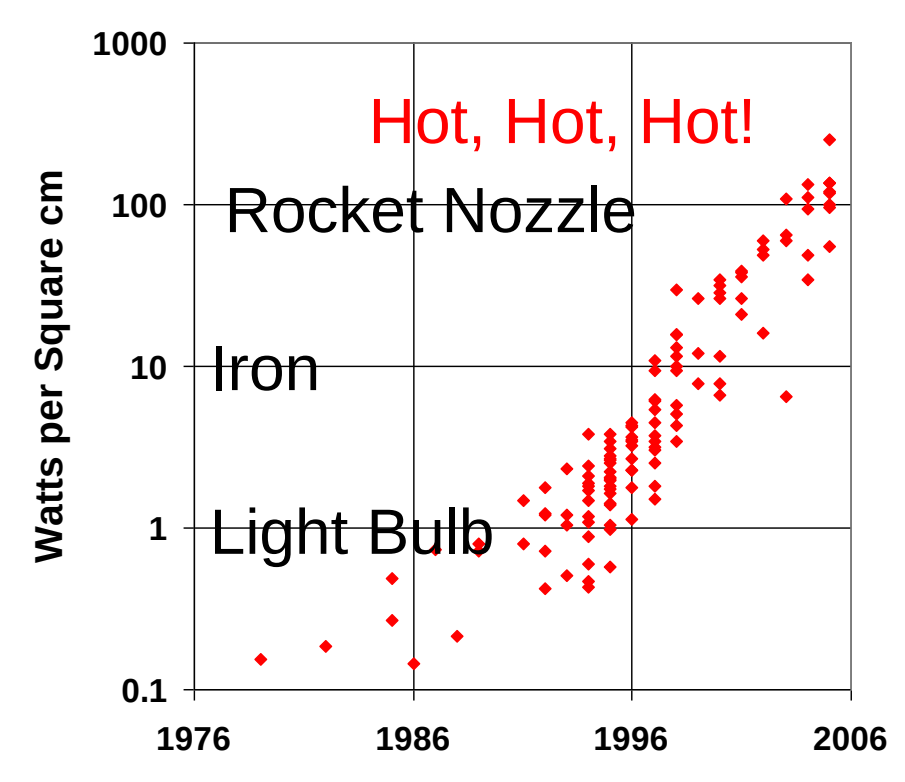
Clock Rate


CPU-Frequency 1993 - 2005
AMD and Intel
tom's hardware guide
Frequency [MHz]
4000
3500
3000
2500
2000
1500
1000
500
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
AMD
Intel
Year

Why the Clock Flattening?

POWER

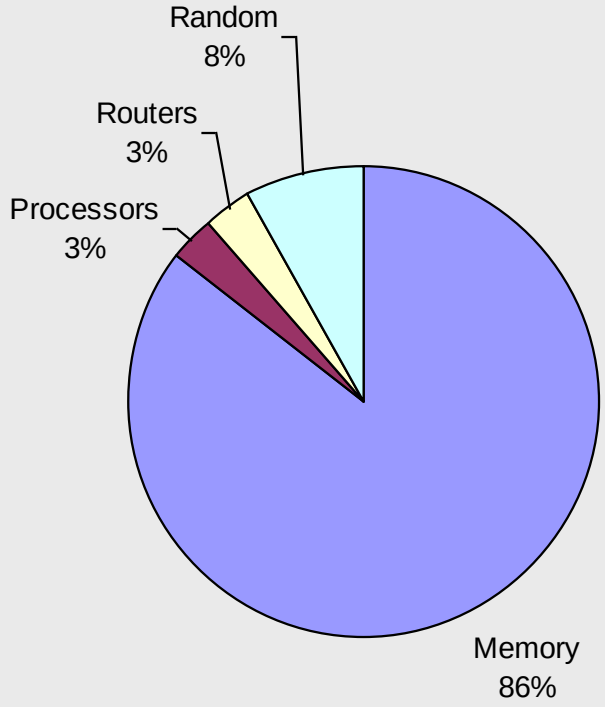




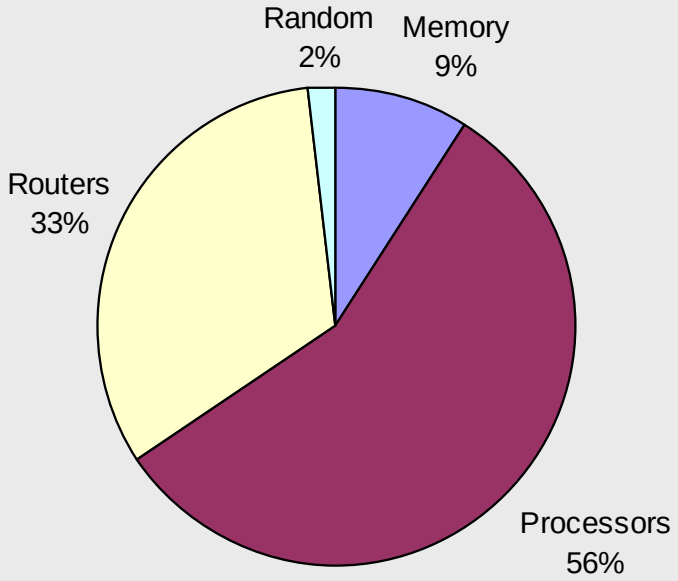
Courtesy of Peter Kogge, UND

What Are We Doing with the Total System Silicon?

Silicon Area Distribution



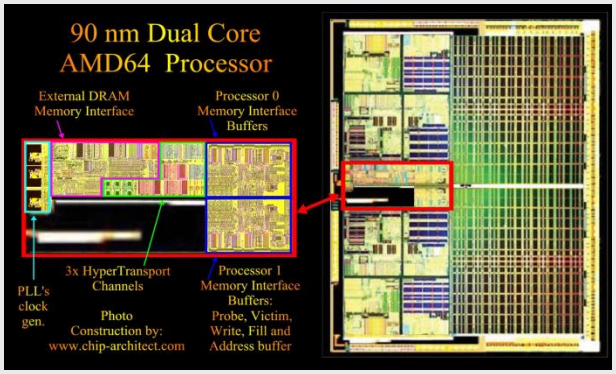
Power Distribution

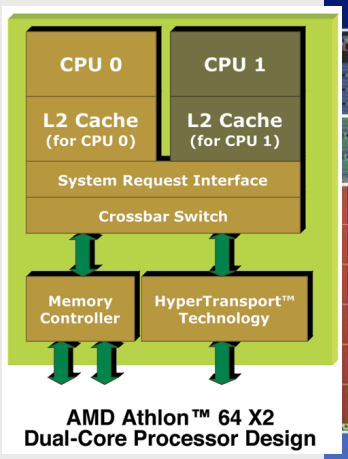


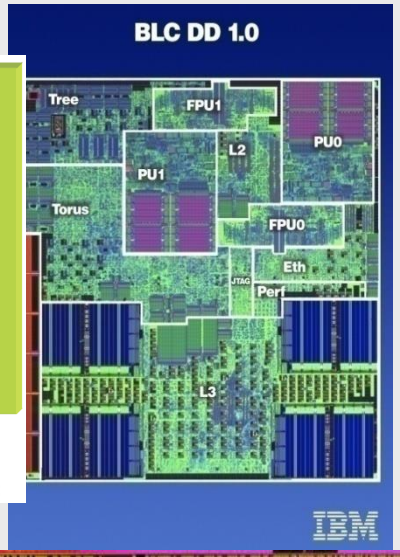
Courtesy of Peter Kogge, UND

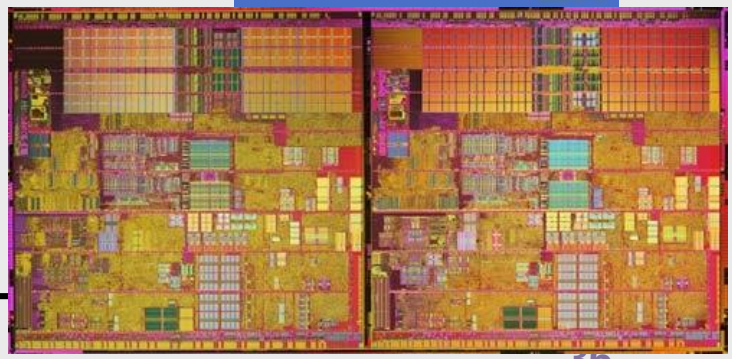
Multi-Core

- Motivation for Multi-Core
  - Exploits improved feature-size and density
  - Increases functional units per chip (spatial efficiency)
  - Limits energy consumption per operation
  - Constrains growth in processor complexity
- Challenges resulting from multi-core
  - Aggravates memory wall
    - Memory bandwidth
      - Way to get data out of memory banks
      - Way to get data into multi-core processor array
    - Memory latency
    - Fragments L3 cache
  - Relies on effective exploitation of multiple-thread parallelism
    - Need for parallel computing model and parallel programming model
  - Pins become strangle point
    - Rate of pin growth projected to slow and flatten
    - Rate of bandwidth per pin (pair) projected to grow slowly
  - Requires mechanisms for efficient inter-processor coordination
    - Synchronization
    - Mutual exclusion
    - Context switching



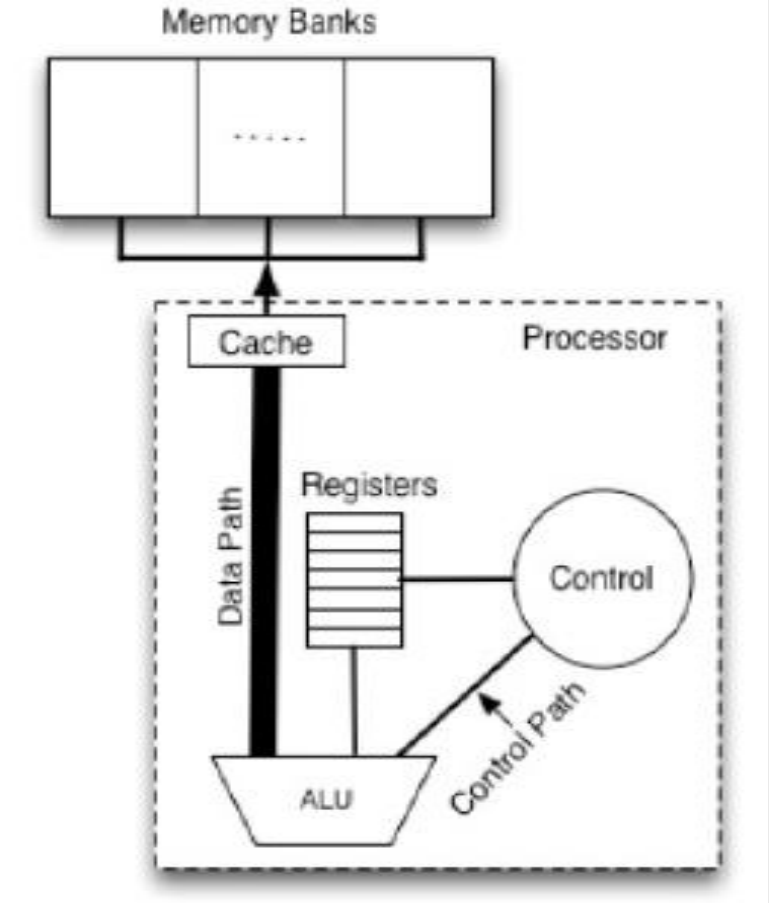






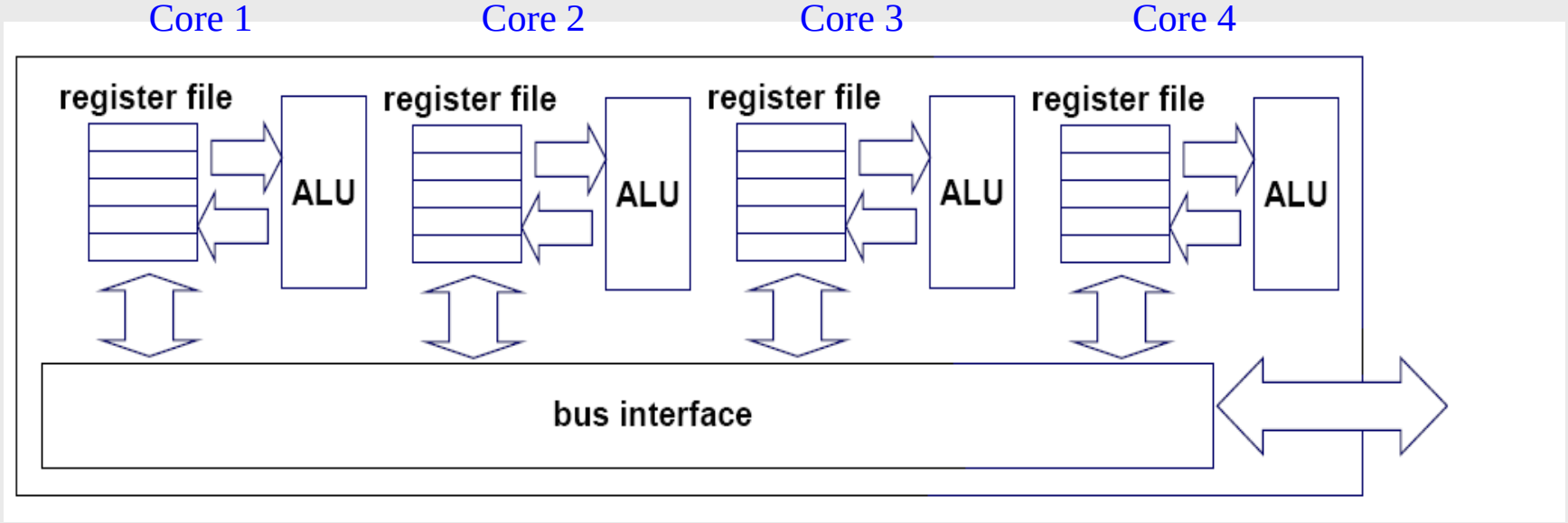
Basic Uni-Processor Architecture elements

- Cache hierarchy
- Arithmetic Logic Units
- Register Sets
- Control
- Memory Interface
- I/O Interface
- Execution pipeline

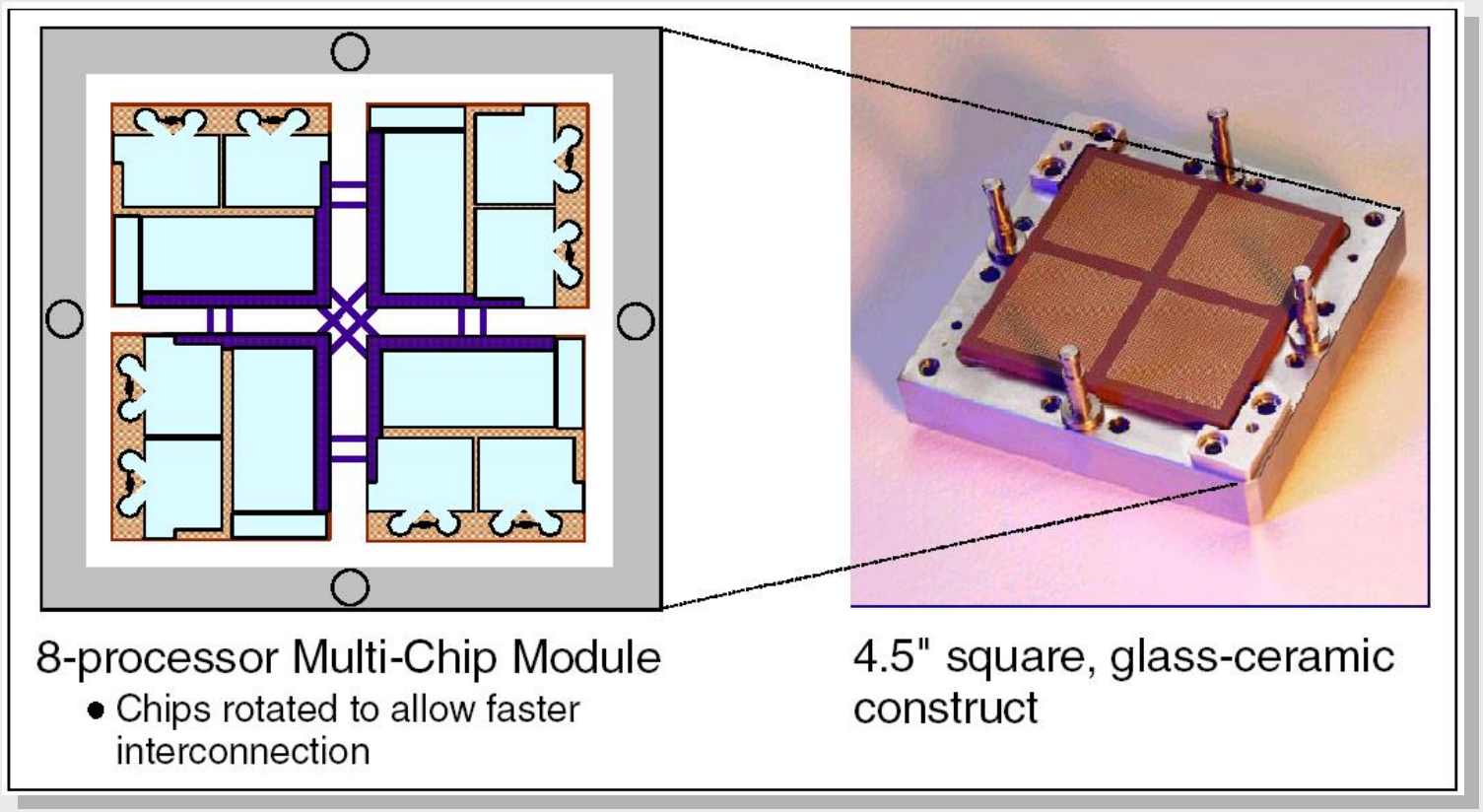


Multicore CPUs

- Most of current Intel and AMD multicore CPUs just put multiple cores



8 Processor Module

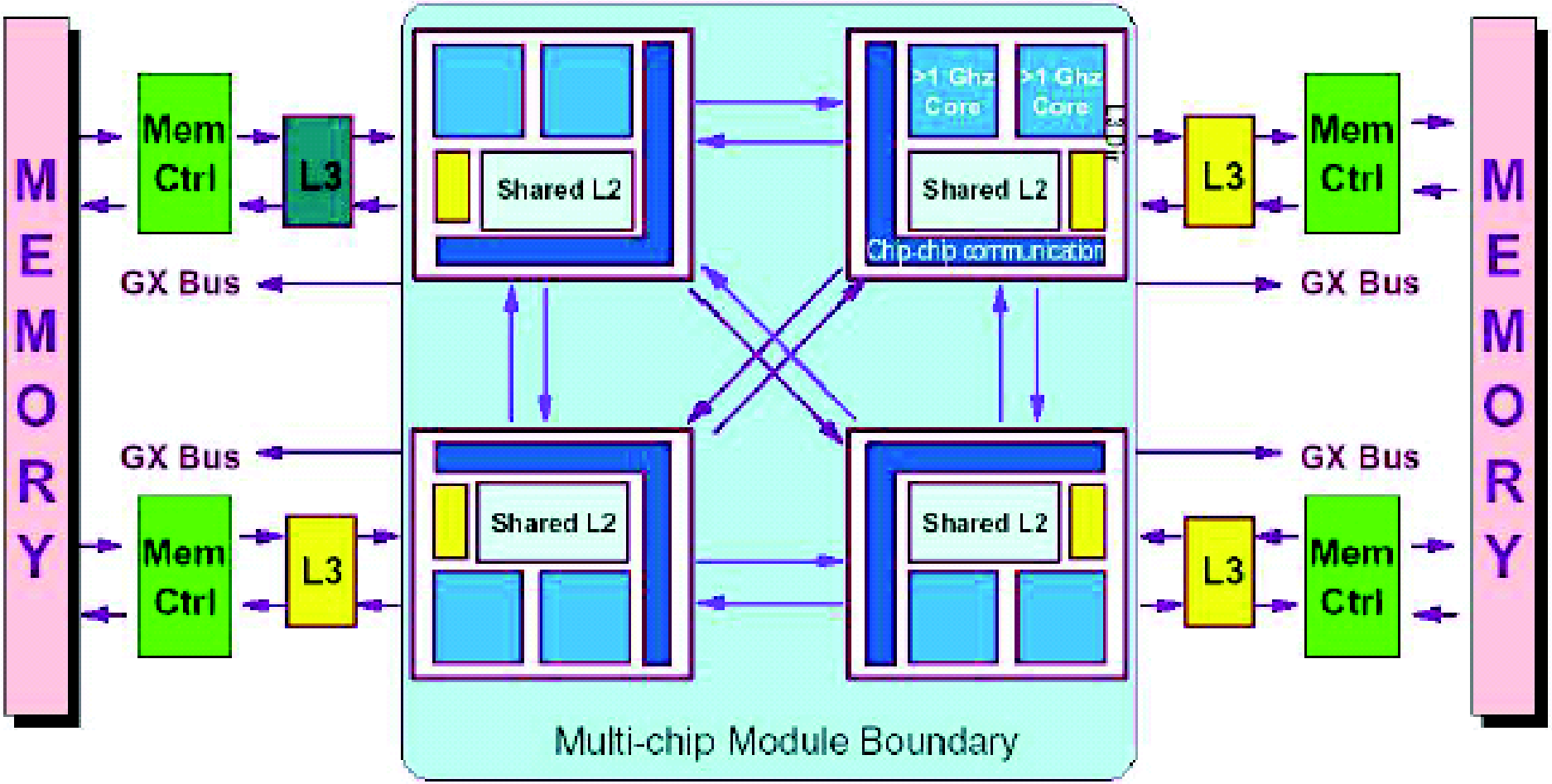


8-processor Multi-Chip Module

- Chips rotated to allow faster interconnection

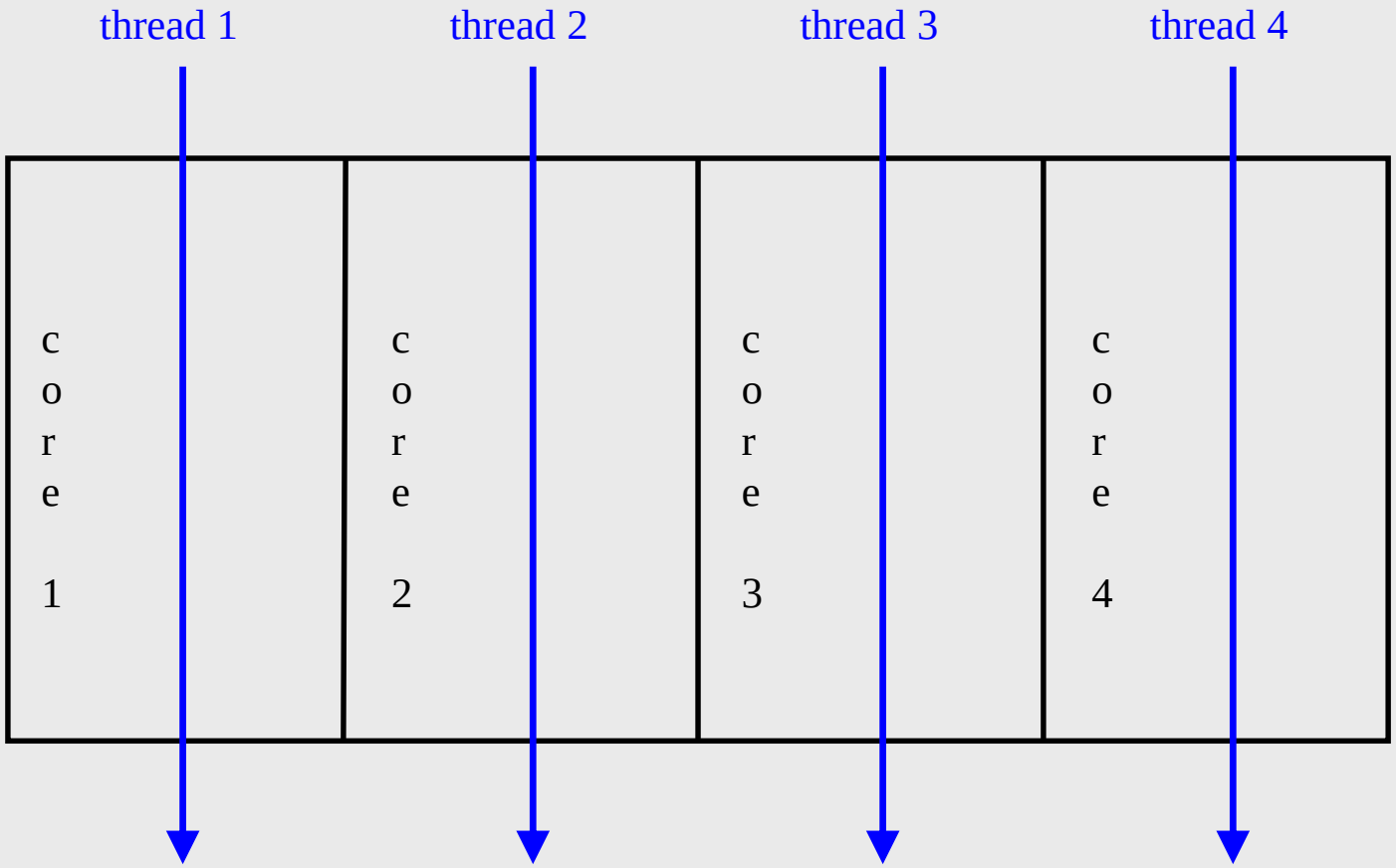
4.5" square, glass-ceramic construct

Multi-Chip Module

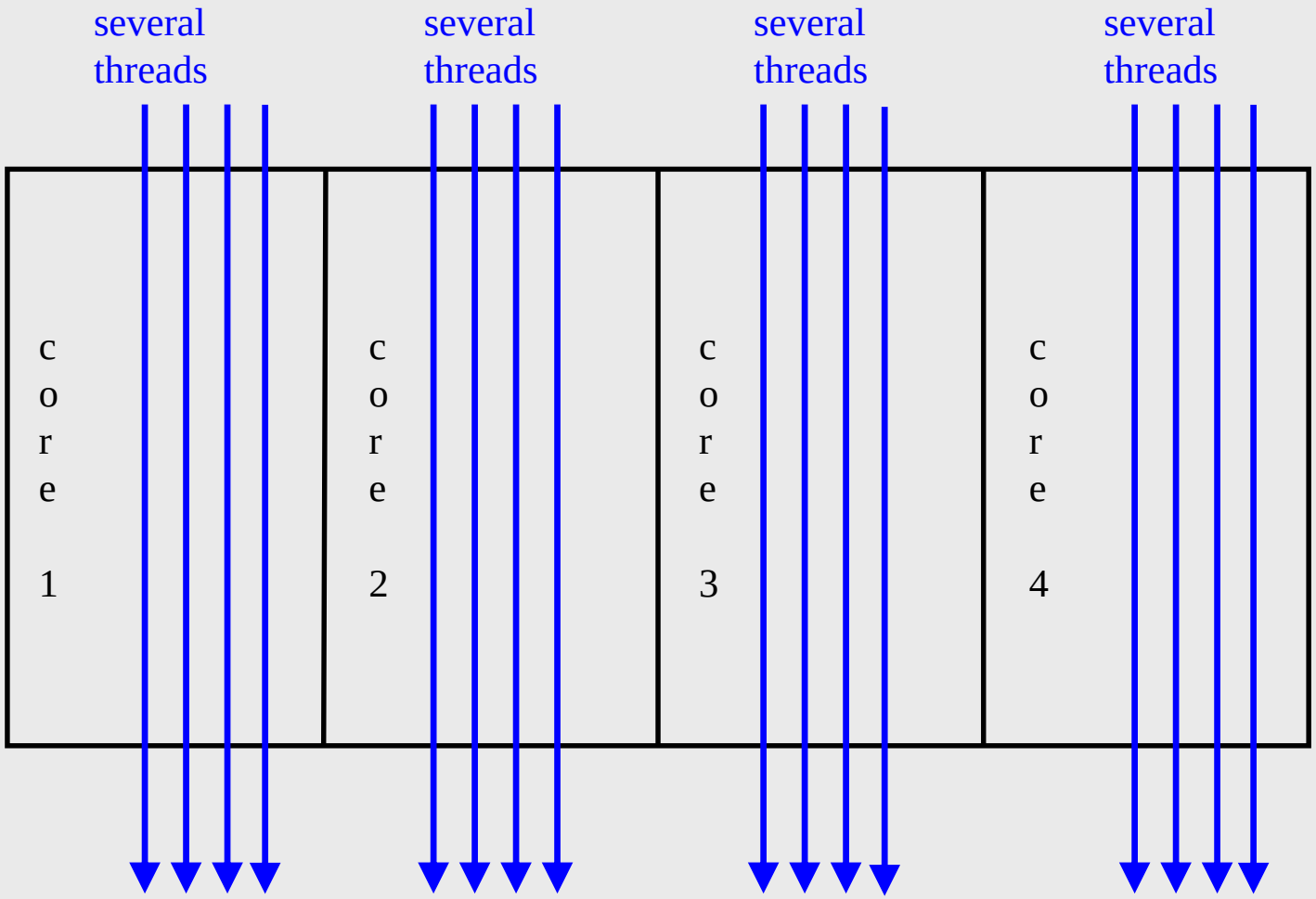


- L3 cache shared across all processors

Single-threaded Multicore CPUs

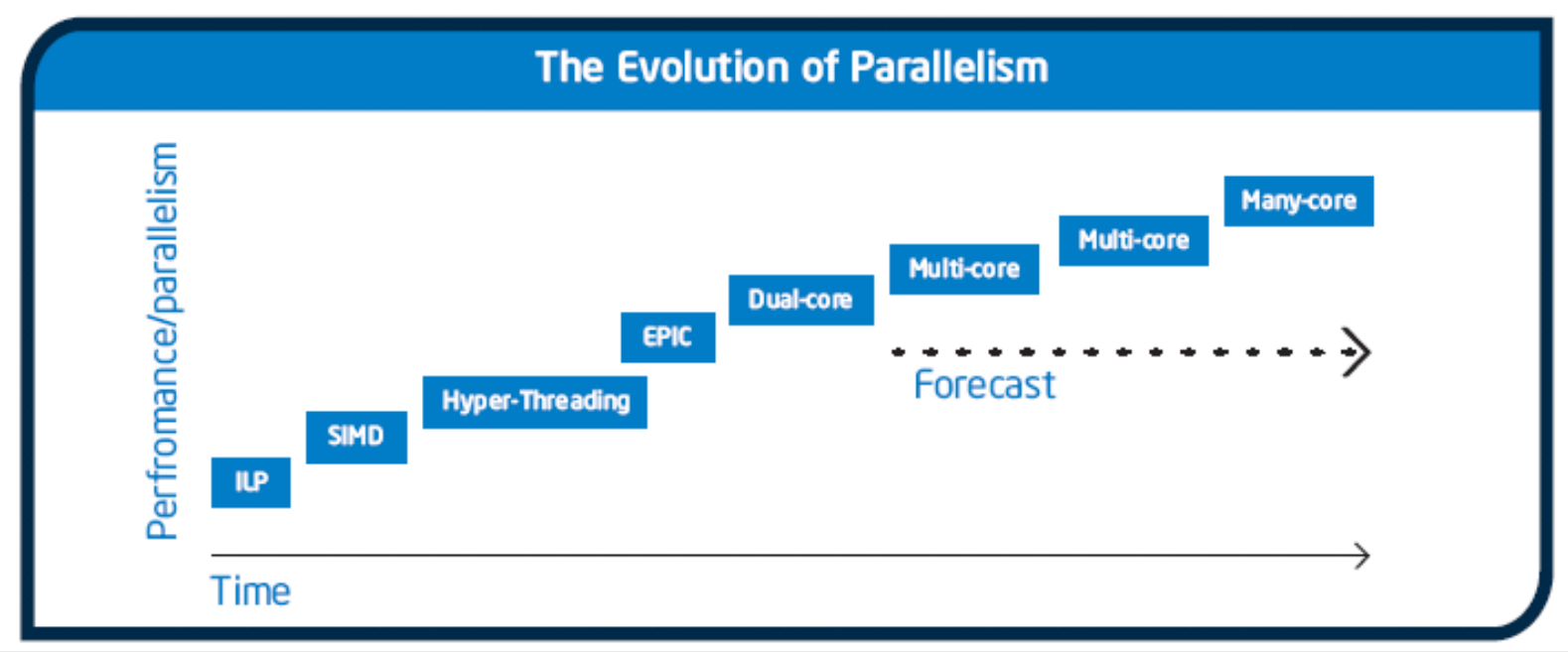


Multi-threaded Multicore CPUs



Evolution of Parallelism

- According to Intel



Source: Intel Website

Challenges of Parallelism

- "Automatic" parallelism in modern machines
  - Bit level parallelism – within floating point operations
  - Instruction level parallelism (ILP) – multiple instructions per clock cycle
  - Memory system parallelism – overlap of memory accesses with computation, **concurrent data access**
  - OS parallelism – multiple jobs run in parallel on SMPs

- There are limitations to all of these!
- To achieve high performance, the programmer needs to identify, schedule and coordinate parallel tasks and data!

Challenges of Parallelism

- Applications are often very sophisticated
  - E.g., adaptive algorithms may require dynamic load balancing
- Algorithm development is harder
  - Complexity of specifying and coordinating concurrent activities
  - Algorithmic scalability losses
  - Serialization and load imbalance
  - Computing and **data access**
- Software development is much harder
  - Lack of development tools and programming models
  - Subtle program errors: race conditions
  - Multilevel parallelism is difficult to manage
  - Communication and/or IO bottlenecks
- Rapid pace of change in computer system architecture
  - A parallel algorithm for one machine may not be a good match for another
  - Homogeneous multicore processors vs GPGPU
- Extreme scale exacerbates inefficiencies

Summary

- What is parallel computing
- Why parallel computing
- Different levels of parallelism
- Challenge and opportunities of parallelism

- Reading:
  - Kumar – ch 1; ch 2