

CSE 613: Parallel Programming

**Department of Computer Science
SUNY Stony Brook
Fall 2013**

*“We used to joke that
“parallel computing is the future, and always will be,”
but the pessimists have been proven wrong.”*

— Tony Hey

Course Information

- **Lecture Time:** TuTh 10:00 am - 11:20 am
- **Location:** Melville Library N4006, West Campus
- **Instructor:** Rezaul A. Chowdhury
- **Office Hours:** TuTh 12:30 pm - 2:00 pm, 1421 Computer Science
- **Email:** rezaul@cs.stonybrook.edu
- **TA:** Unlikely!
- **Class Webpage:**
<http://www.cs.sunysb.edu/~rezaul/CSE613-F13.html>

Prerequisites

- **Required:** Background in algorithms analysis
(e.g., CSE 373 or CSE 548)
- **Required:** Background in programming languages (C / C++)
- **Helpful but Not Required:** Background in computer architecture
- **Please Note:** This is not a course on
 - Programming languages
 - Computer architecture
- **Main Emphasis:** Parallel algorithms

Topics to be Covered

The following topics will be covered

- Analytical modeling of parallel programs
- Scheduling
- Programming using the message-passing paradigm and for shared address-space platforms
- Parallel algorithms for dense matrix operations, sorting, searching, graphs, computational geometry, and dynamic programming
- Concurrent data structures
- Transactional memory, etc.

Grading Policy

- Homeworks (three: lowest score 5%, others 10% each): 25%
- Exams (two: higher one 20%, lower one 10%): 30%
 - Midterm (in-class): Oct 15
 - Final (in-class): Dec 5
- Group project (one): 30%
 - Proposal: Sep 19
 - Progress report (in-class): Oct 22 - 24
 - Final report: Dec 6
- Scribe note (one lecture): 10%
- Class participation & attendance: 5%

Programming Environment

This course is supported by an educational grant from

- Extreme Science and Engineering Discovery Environment (XSEDE):

<https://www.xsede.org>

We have access to the following supercomputers

- **Stampede (Texas Advanced Comp. Center)**: 6,400 nodes;
16 cores (2 Intel Sandy Bridge) and 1/2 Intel Xeon Phi coprocessor(s) per node
- **Lonestar (Texas Advanced Comp. Center)**: 1,800+ nodes;
12 cores (2 Intel Westmere processors) per node
- **Trestles (San Diego Supercomputer Center)**: 300+ nodes;
32 cores (4 AMD Magny Cours processors) per node
- **Kraken (National Institute for Comp. Sciences)**: 9,000+ nodes;
12 cores (2 AMD Opteron Istanbul processors) per node
- **Keeneland KIDS (Georgia Tech)**: 120 nodes;
16 cores (2 Intel Sandy Bridge processors) and 3 NVIDIA Fermi GPUs per node

Programming Environment

World's Most Powerful Supercomputers in June, 2013

(www.top500.org)

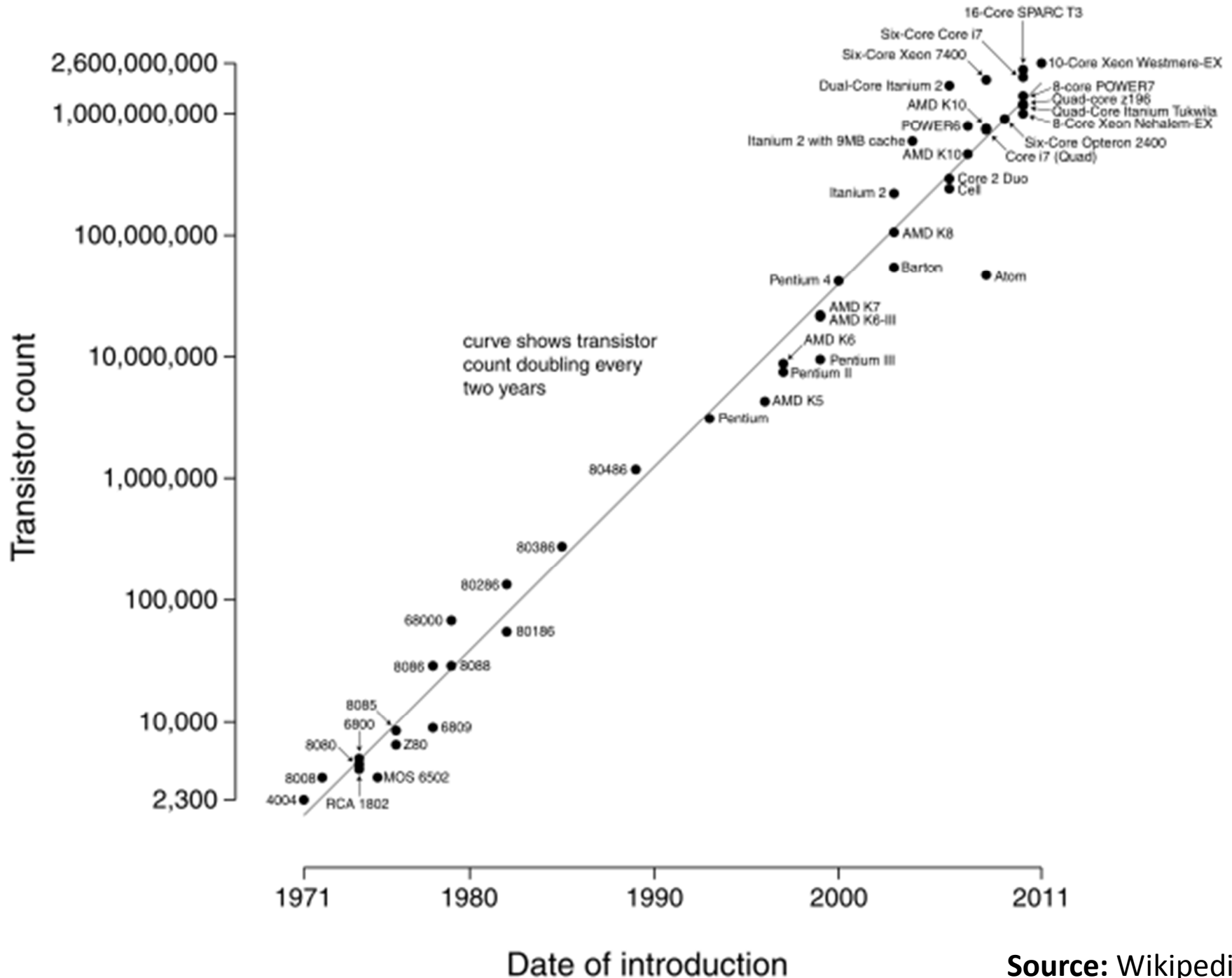
Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National University of Defense Technology China	Tianhe-2 (MilkyWay-2) - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3120000	33862.7	54902.4	17808
2	DOE/SC/Oak Ridge National Laboratory United States	Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560640	17590.0	27112.5	8209
3	DOE/NNSA/LLNL United States	Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1572864	17173.2	20132.7	7890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIx 2.0GHz, Tofu interconnect Fujitsu	705024	10510.0	11280.4	12660
5	DOE/SC/Argonne National Laboratory United States	Mira - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786432	8586.6	10066.3	3945
6	Texas Advanced Computing Center/Univ. of Texas United States	Stampede - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462462	5168.1	8520.1	4510
7	Forschungszentrum Juelich (FZJ) Germany	JUQUEEN - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	458752	5008.9	5872.0	2301
8	DOE/NNSA/LLNL United States	Vulcan - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393216	4293.3	5033.2	1972
9	Leibniz Rechenzentrum Germany	SuperMUC - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR IBM	147456	2897.0	3185.1	3423
10	National Supercomputing Center in Tianjin China	Tianhe-1A - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 NUDT	186368	2566.0	4701.0	4040

Recommended Textbooks

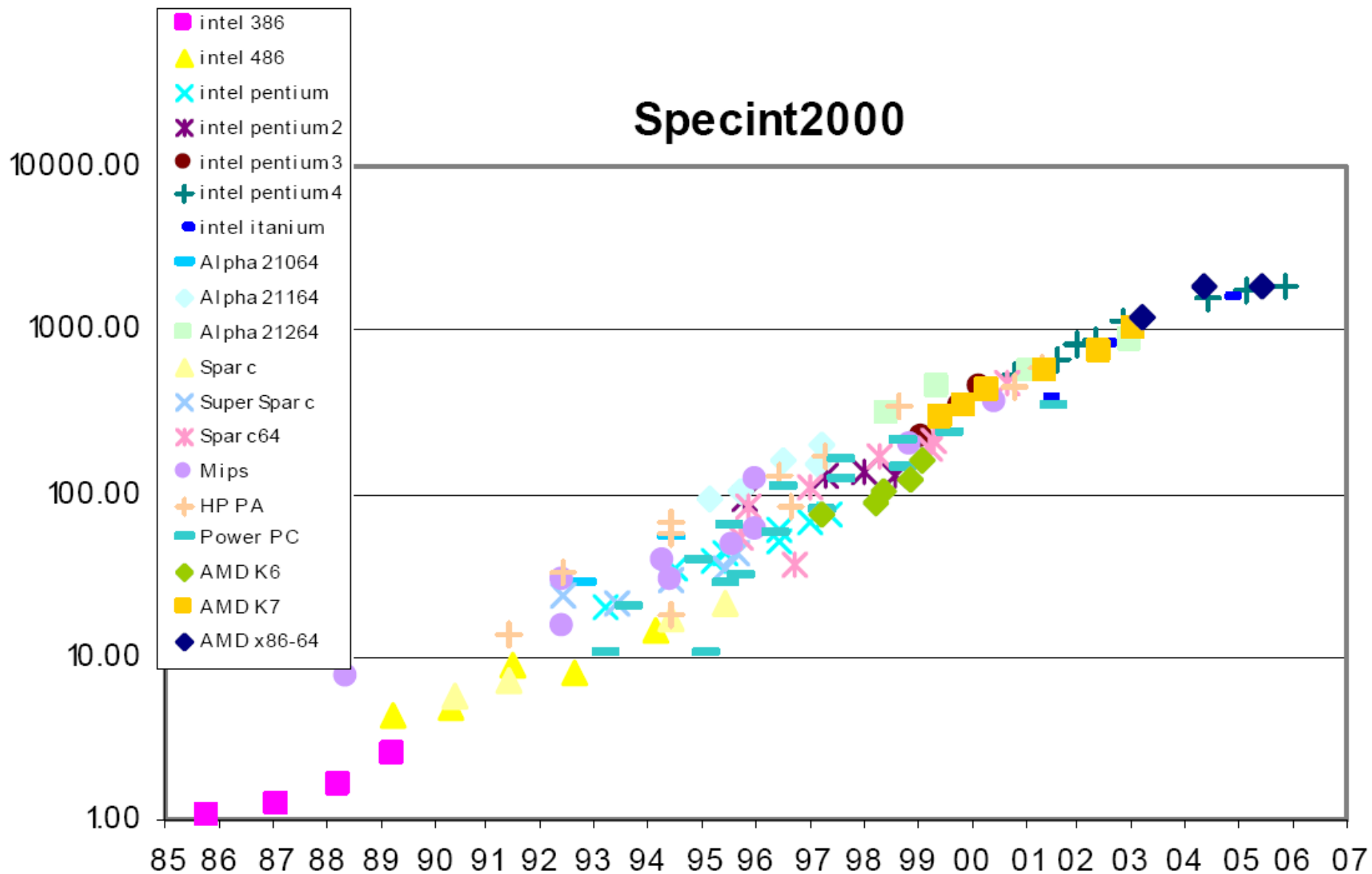
- A. Grama, G. Karypis, V. Kumar, and A. Gupta. ***Introduction to Parallel Computing*** (2nd Edition), Addison Wesley, 2003.
- J. JáJá. ***An Introduction to Parallel Algorithms*** (1st Edition), Addison Wesley, 1992.
- T. Cormen, C. Leiserson, R. Rivest, and C. Stein. ***Introduction to Algorithms*** (3rd Edition), MIT Press, 2009.
- M. Herlihy and N. Shavit. ***The Art of Multiprocessor Programming*** (1st Edition), Morgan Kaufmann, 2008.
- P. Pacheco. ***Parallel Programming with MPI*** (1st Edition), Morgan Kaufmann, 1996.

Why Parallelism?

Moore's Law



Unicore Performance



Source: Chung-Ta King, Department of Computer Science, National Tsing Hua University

Unicore Performance Has Hit a Wall!

Some Reasons

- Lack of additional ILP
(Instruction Level Hidden Parallelism)
- High power density
- Manufacturing issues
- Physical limits
- Memory speed

Unicore Performance: No Additional ILP

“Everything that can be invented has been invented.”

— Charles H. Duell

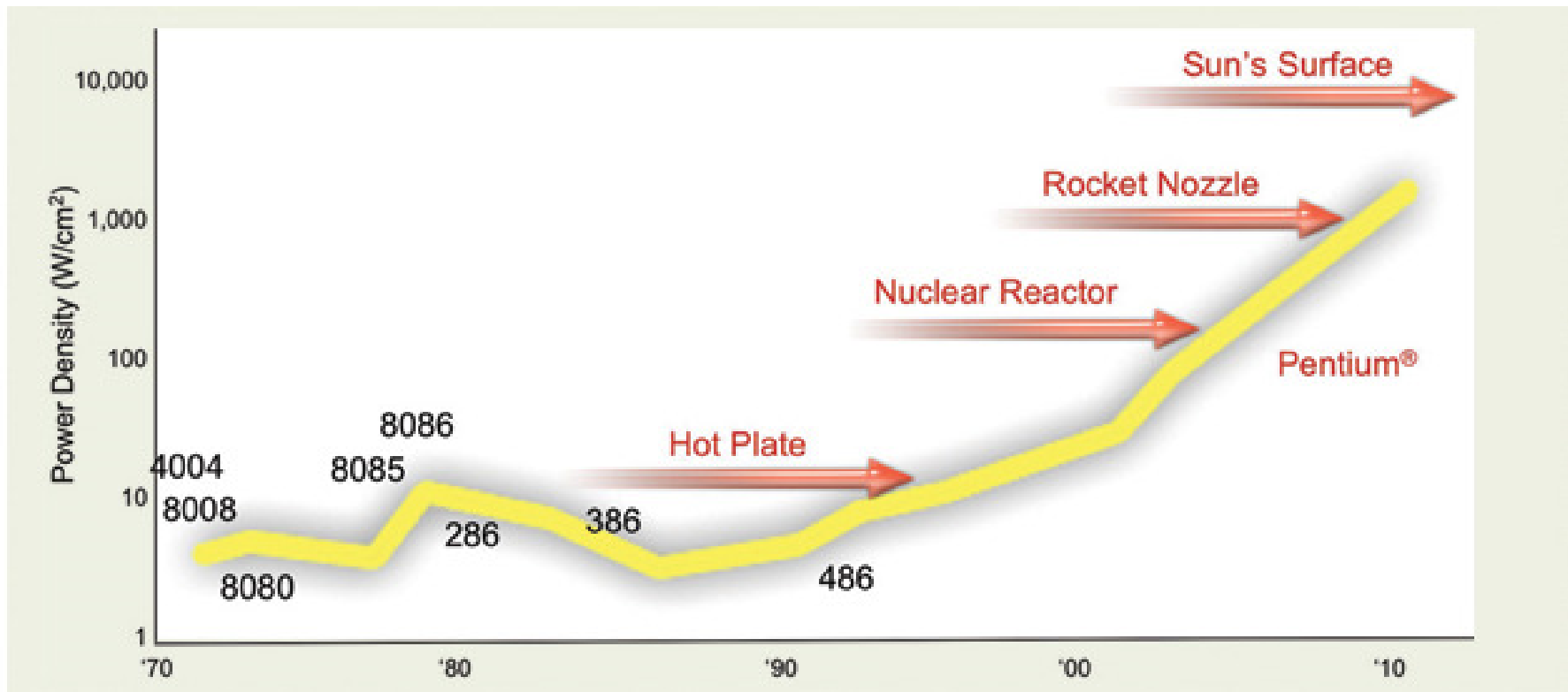
Commissioner, U.S. patent office, 1899

Exhausted all ideas to exploit hidden parallelism?

- Multiple simultaneous instructions
- Dynamic instruction scheduling
- Branch prediction
- Out-of-order instructions
- Speculative execution
- Pipelining
- Non-blocking caches, etc.

Unicore Performance: High Power Density

- Dynamic power, $P_d \propto V^2 f C$
 - $V = \text{supply voltage}$
 - $f = \text{clock frequency}$
 - $C = \text{capacitance}$
- But $V \propto f$
- Thus $P_d \propto f^3$



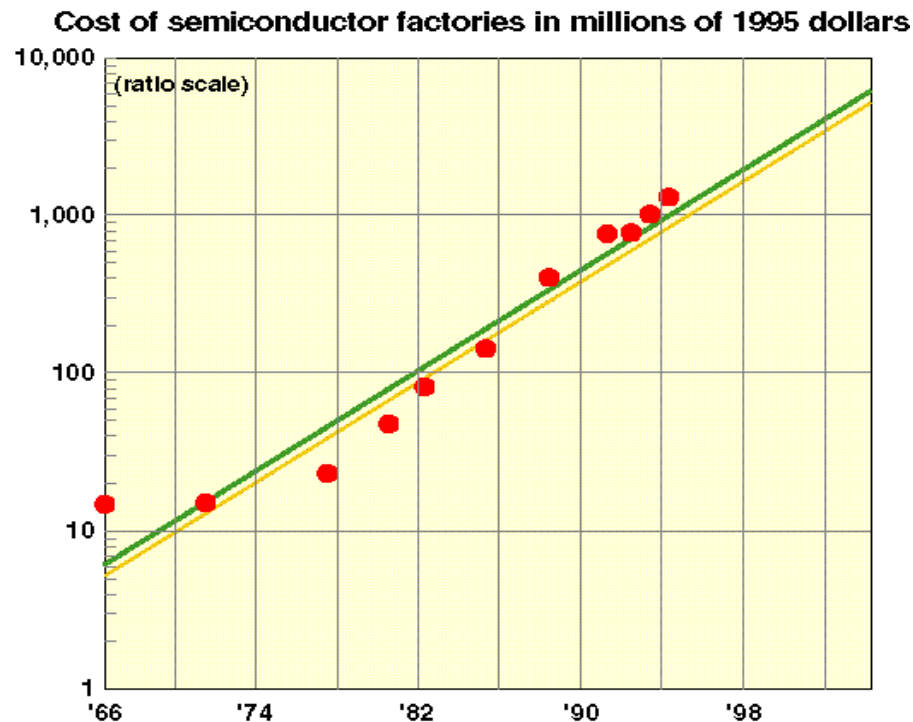
Source: Patrick Gelsinger, Intel Developer Forum, Spring 2004 (Simon Floyd)

Unicore Performance: Manufacturing Issues

- Frequency, $f \propto 1 / s$
 - $s = \text{feature size (transistor dimension)}$
- Transistors / unit area $\propto 1 / s^2$
- Typically, die size $\propto 1 / s$
- So, what happens if feature size goes down by a factor of x ?
 - Raw computing power goes up by a factor of x^4 !
 - Typically most programs run faster by a factor of x^3 without any change!

Unicore Performance: Manufacturing Issues

- Manufacturing cost goes up as feature size decreases
 - Cost of a semiconductor fabrication plant doubles every 4 years (Rock's Law)
- CMOS feature size is limited to 5 nm (at least 10 atoms)



Source: Kathy Yelick and Jim Demmel, UC Berkeley

Unicore Performance: Physical Limits

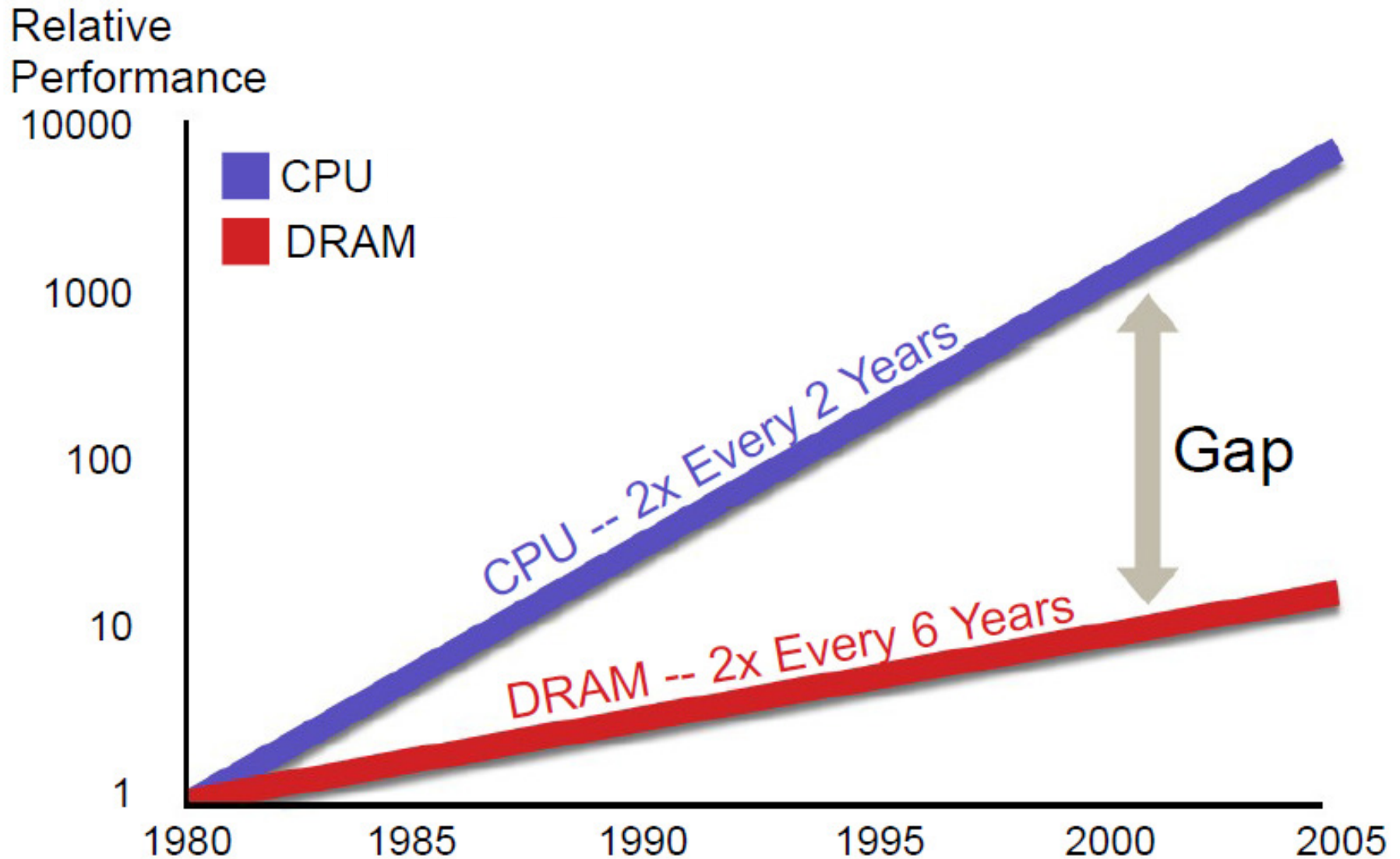
Execute the following loop on a serial machine in 1 second:

for (i = 0; i < 10¹²; ++i)

z[i] = x[i] + y[i];

- We will have to access 3×10^{12} data items in one second
- Speed of light is, $c \approx 3 \times 10^8$ m/s
- So each data item must be within $c / 3 \times 10^{12} \approx 0.1$ mm from the CPU on the average
- All data must be put inside a 0.2 mm \times 0.2 mm square
- Each data item (≥ 8 bytes) can occupy only 1 Å² space!
(size of a small atom!)

Unicore Performance: Memory Wall



Source: Sun World Wide Analyst Conference Feb. 25, 2003

Source: Rick Hetherington, Chief Technology Officer, Microelectronics, Sun Microsystems

Unicore Performance Has Hit a Wall!

Some Reasons

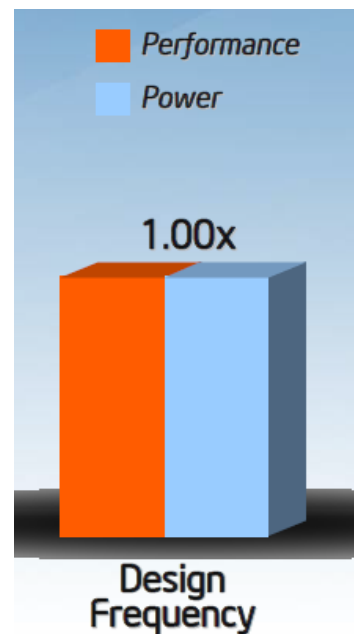
- Lack of additional ILP
(Instruction Level Hidden Parallelism)
- High power density
- Manufacturing issues
- Physical limits
- Memory speed

“Oh Sinnerman, where you gonna run to?”

— Sinnerman (recorded by Nina Simone)

Where You Gonna Run To?

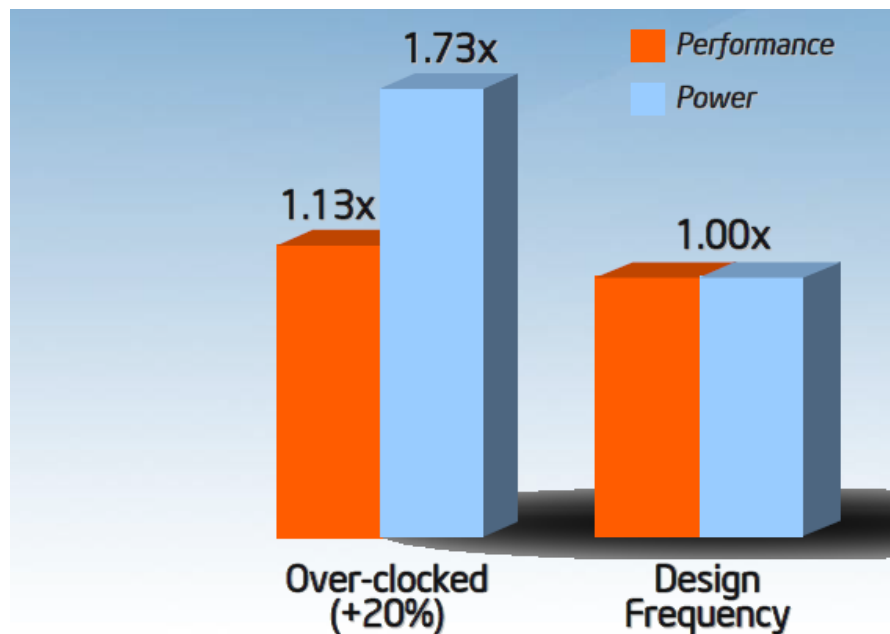
- Changing f by 20% changes performance by 13%
- So what happens if we overclock by 20%?
- And underclock by 20%?



Source: Andrew A. Chien, Vice President of Research, Intel Corporation

Where You Gonna Run To?

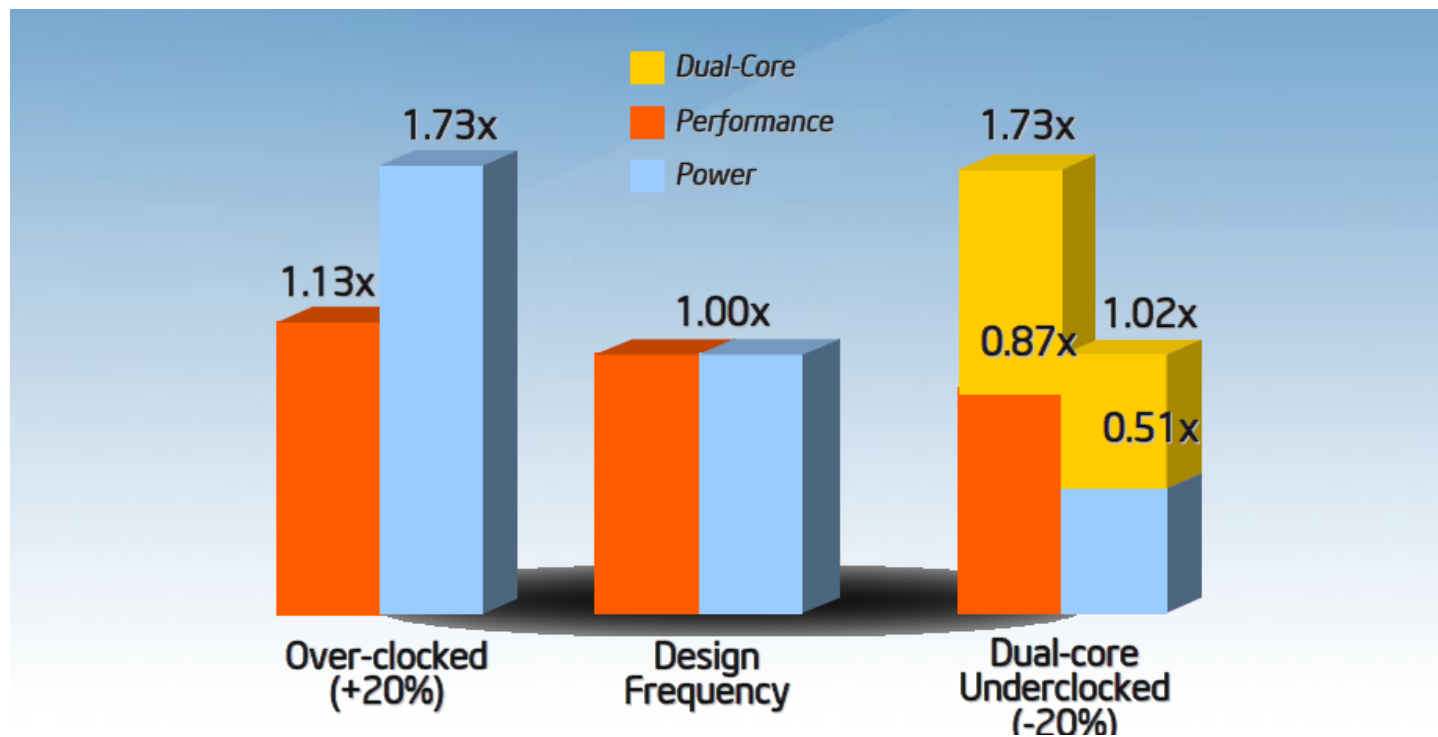
- Changing f by 20% changes performance by 13%
- So what happens if we overclock by 20%?
- And underclock by 20%?



Source: Andrew A. Chien, Vice President of Research, Intel Corporation

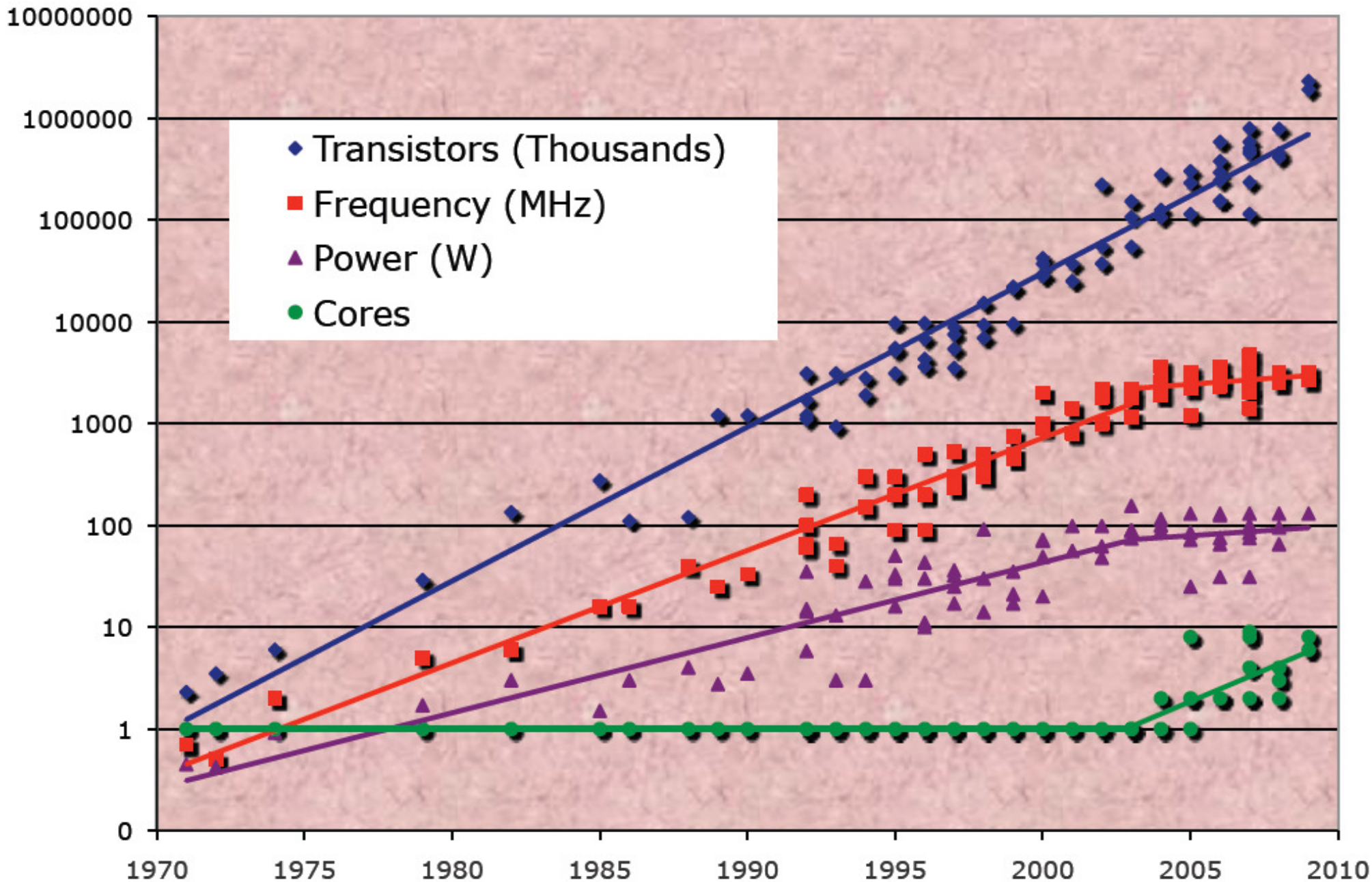
Where You Gonna Run To?

- Changing f by 20% changes performance by 13%
- So what happens if we overclock by 20%?
- And underclock by 20%?



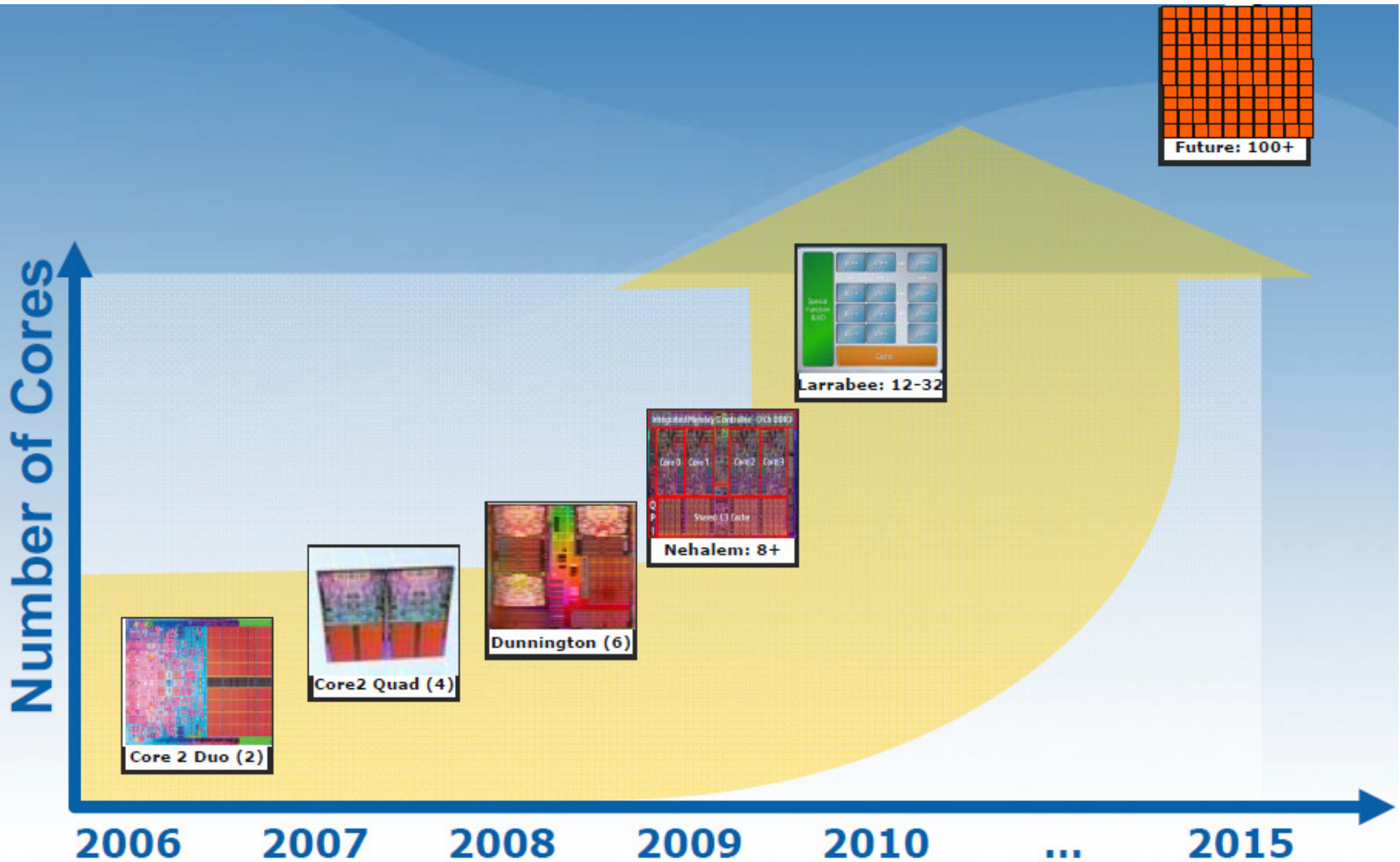
Source: Andrew A. Chien, Vice President of Research, Intel Corporation

Moore's Law Reinterpreted



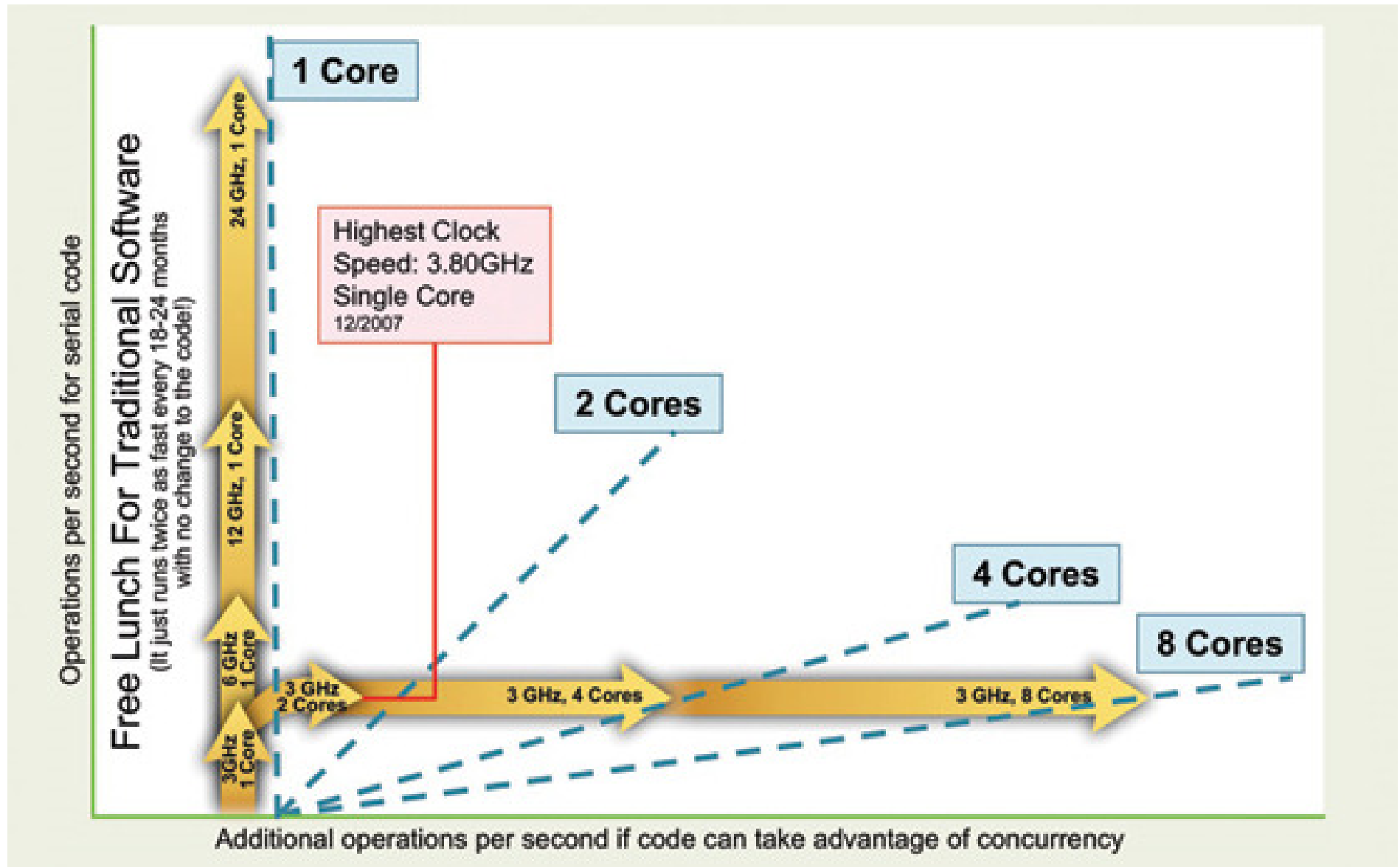
Source: Report of the 2011 Workshop on Exascale Programming Challenges

Cores / Processor (General Purpose)



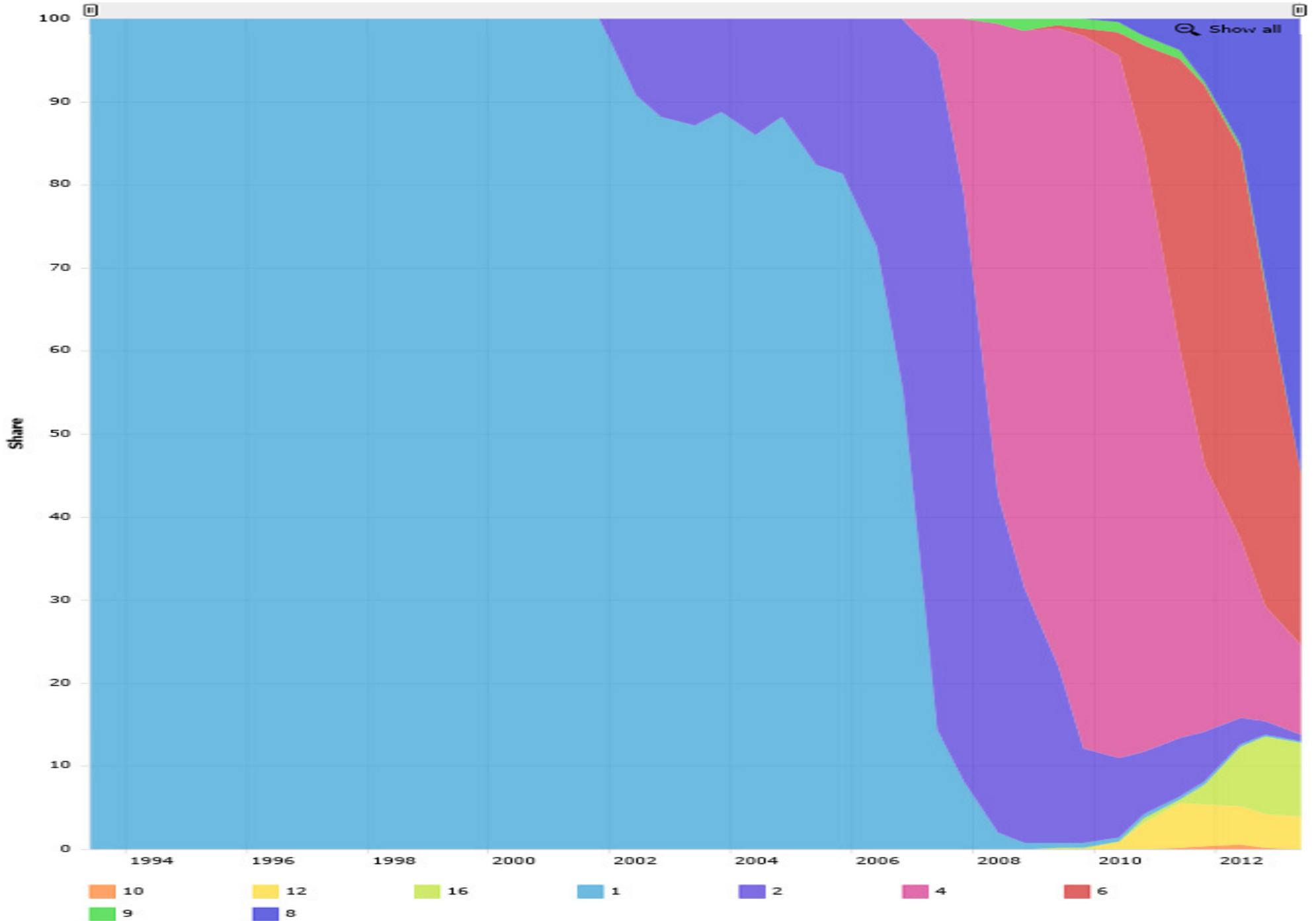
Source: Andrew A. Chien, Vice President of Research, Intel Corporation

No Free Lunch for Traditional Software



Source: Simon Floyd, Workstation Performance: Tomorrow's Possibilities (Viewpoint Column)

Top 500 Supercomputing Sites (Cores / Socket)

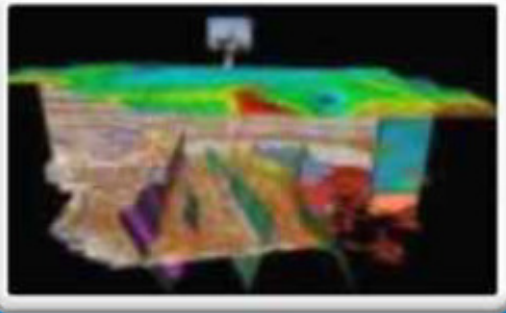


Source: www.top500.org

Insatiable Demand for Performance



Weather Prediction



Oil Exploration



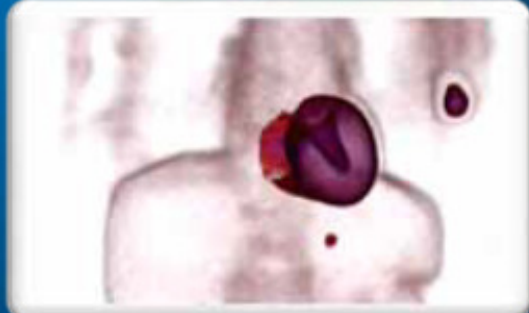
Design Simulation



Genomics Research



Financial Analysis



Medical Imaging

Numerical Weather Prediction

Problem: (*temperature, pressure, ..., humidity, wind velocity*)
← *f(longitude, latitude, height, time)*

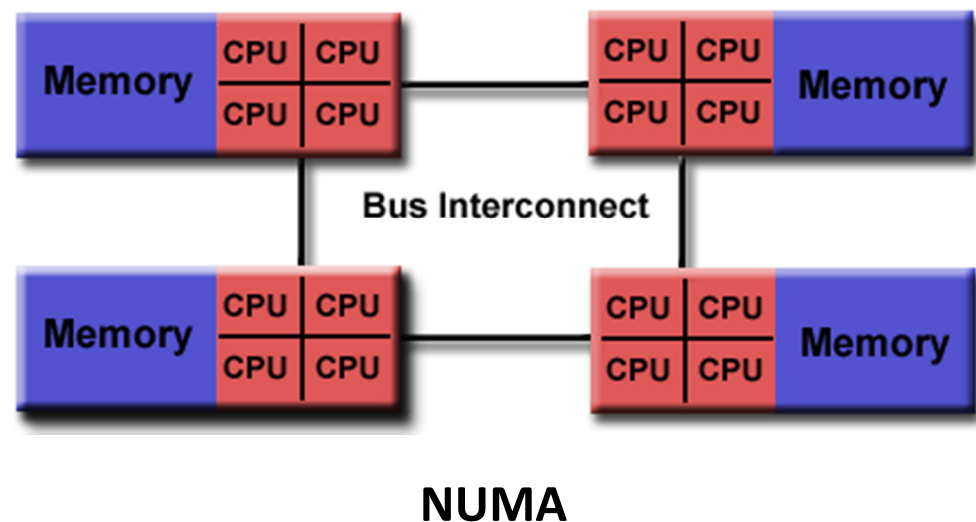
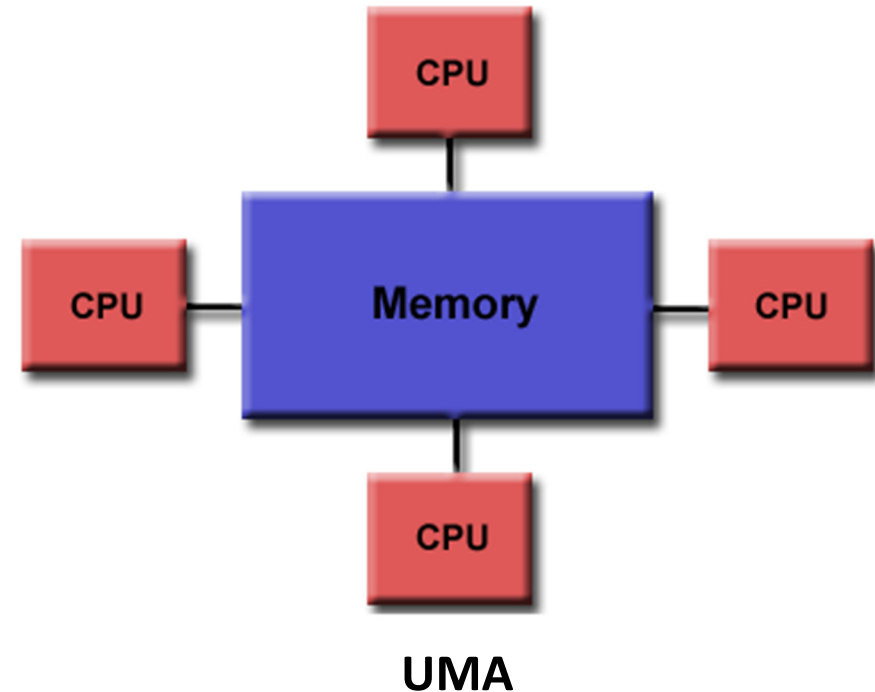
Approach (very coarse resolution):

- Consider only modeling fluid flow in the atmosphere
- Divide the entire global atmosphere into cubic cells of size 1 mile × 1 mile × 1 mile each to a height of 10 miles
≈ 2×10^9 cells
- Simulate 7 days in 1 minute intervals
≈ 10^4 time-steps to simulate
- 200 floating point operations (flop) / cell / time-step
≈ 4×10^{15} floating point operations in total
- To predict in 1 hour ≈ 1 Tflop/s (Tera flop / sec)

Some Useful Classifications of Parallel Computers

Parallel Computer Memory Architecture (Shared Memory)

- All processors access all memory as global address space
- Changes in memory by one processor are visible to all others
- Two types:
 - Uniform Memory Access (UMA)
 - Non-Uniform Memory Access (NUMA)



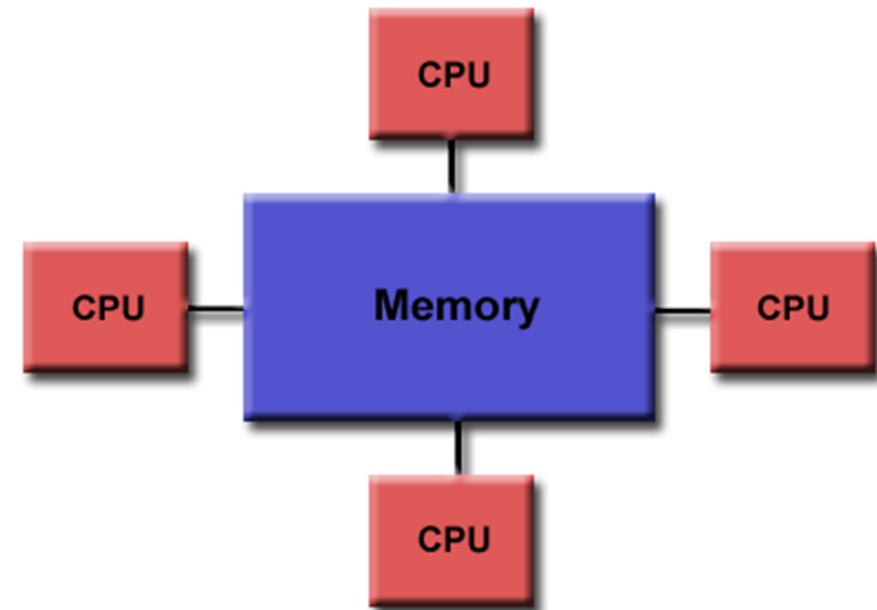
Parallel Computer Memory Architecture (Shared Memory)

Advantages

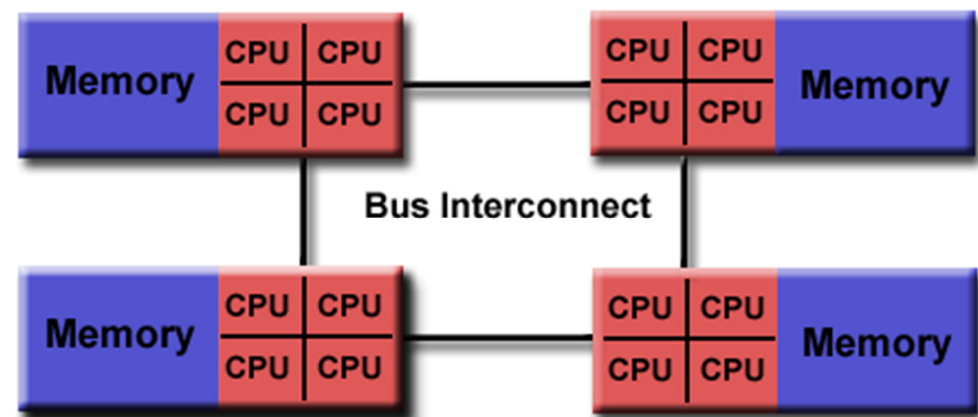
- User-friendly programming perspective to memory
- Fast data sharing

Disadvantages

- Difficult and expensive to scale
- Correct data access is user responsibility



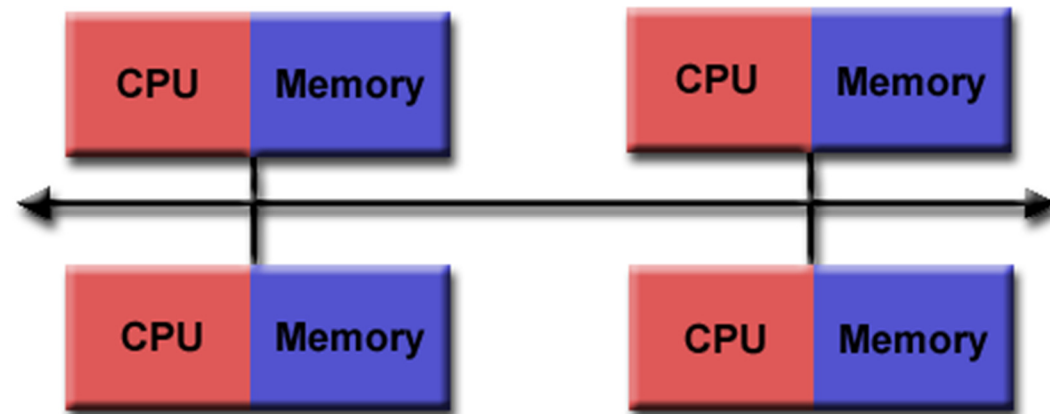
UMA



NUMA

Parallel Computer Memory Architecture (Distributed Memory)

- Each processor has its own local memory — no global address space
- Changes in local memory by one processor have no effect on memory of other processors
- Communication network to connect inter-processor memory

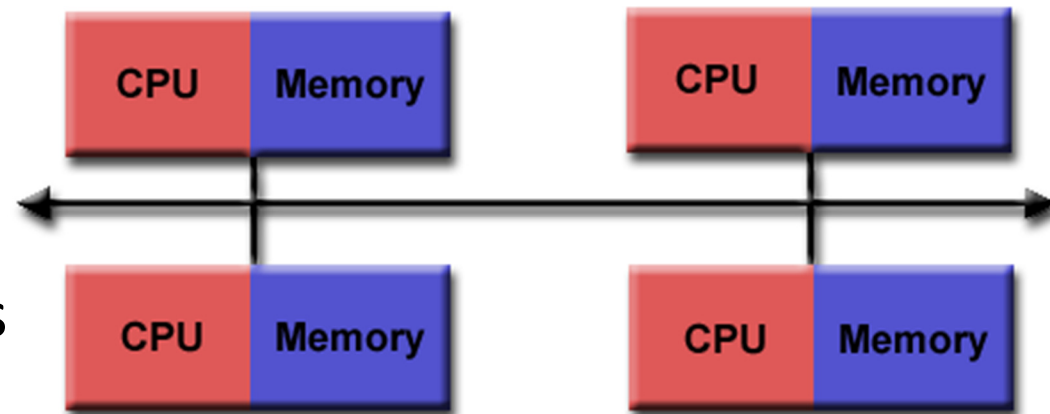


Source: Blaise Barney, LLNL

Parallel Computer Memory Architecture (Distributed Memory)

Advantages

- Easily scalable
- No cache-coherency needed among processors
- Cost-effective



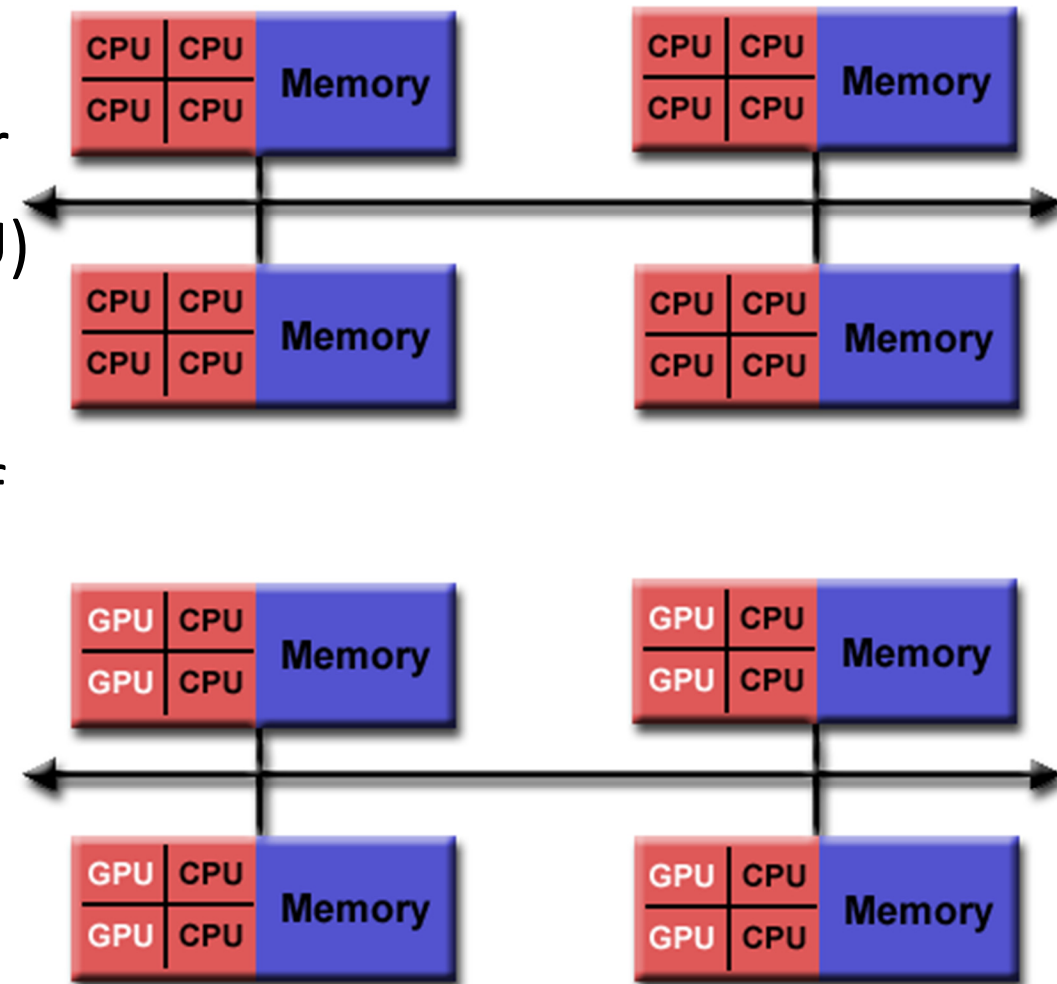
Source: Blaise Barney, LLNL

Disadvantages

- Communication is user responsibility
- Non-uniform memory access
- May be difficult to map shared-memory data structures to this type of memory organization

Parallel Computer Memory Architecture (Hybrid Distributed-Shared Memory)

- The share-memory component can be a cache-coherent SMP or a Graphics Processing Unit (GPU)
- The distributed-memory component is the networking of multiple SMP/GPU machines
- Most common architecture for the largest and fastest computers in the world today



Flynn's Taxonomy of Parallel Computers

Flynn's classical taxonomy (1966):

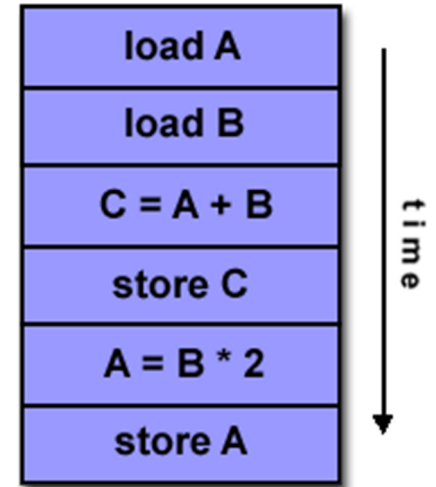
Classification of multi-processor computer architectures along two independent dimensions of *instruction* and *data*.

	Single Data (SD)	Multiple Data (MD)
Single Instruction (SI)	SISD	SIMD
Multiple Instruction (MI)	MISD	MIMD

Flynn's Taxonomy of Parallel Computers

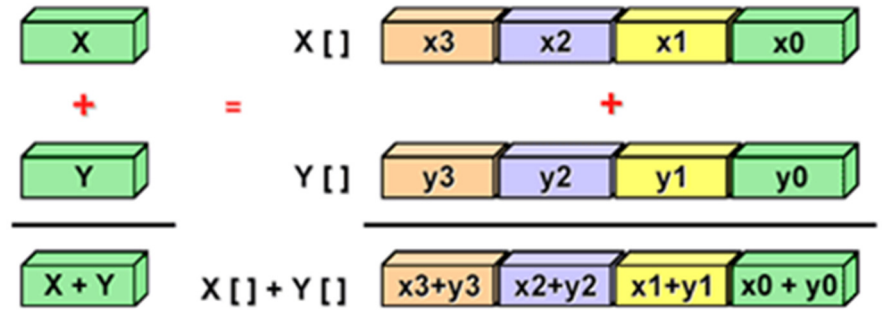
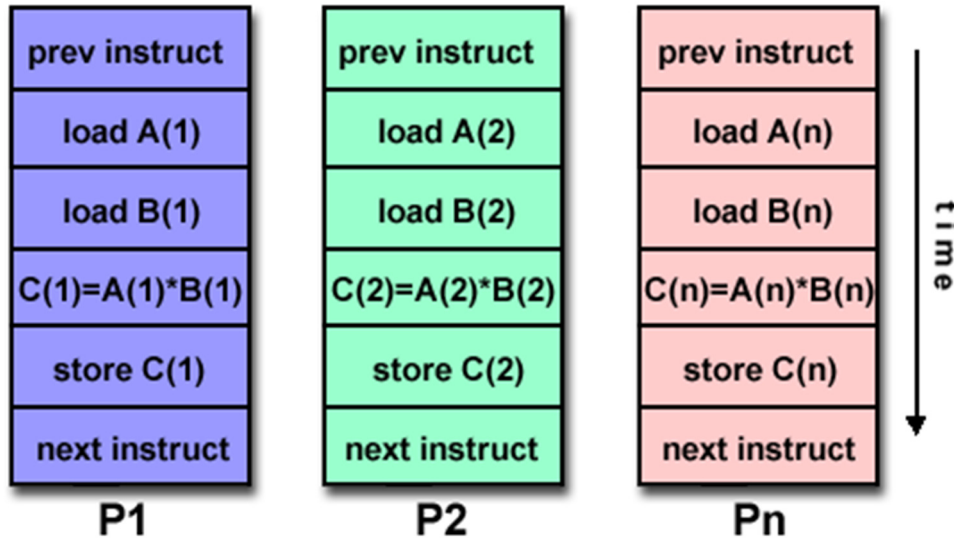
SISD

- A serial (non-parallel) computer
- The oldest and the most common type of computers
- Example: Uniprocessor unicore machines



Source: Blaise Barney, LLNL

Flynn's Taxonomy of Parallel Computers



Source: Blaise Barney, LLNL

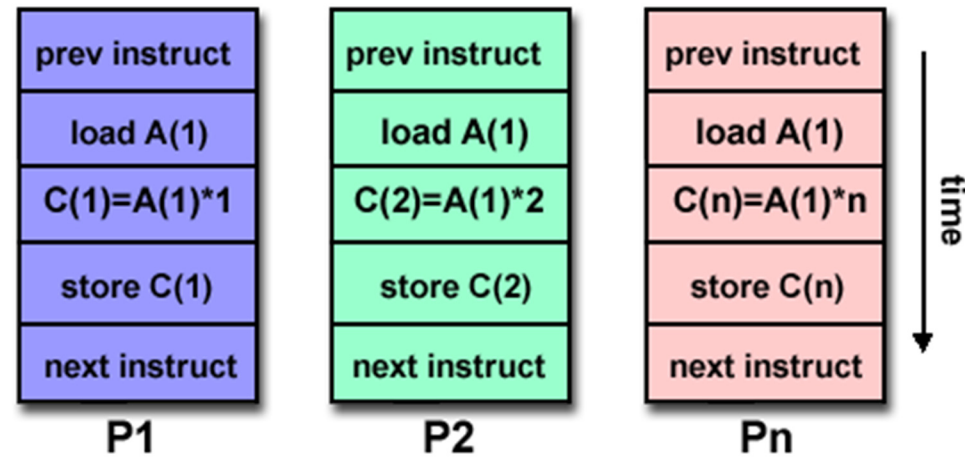
SIMD

- A type of parallel computer
- All PU's run the same instruction at any given clock cycle
- Each PU can act on a different data item
- Synchronous (lockstep) execution
- Two types: processor arrays and vector pipelines
- Example: GPUs (Graphics Processing Units)

Flynn's Taxonomy of Parallel Computers

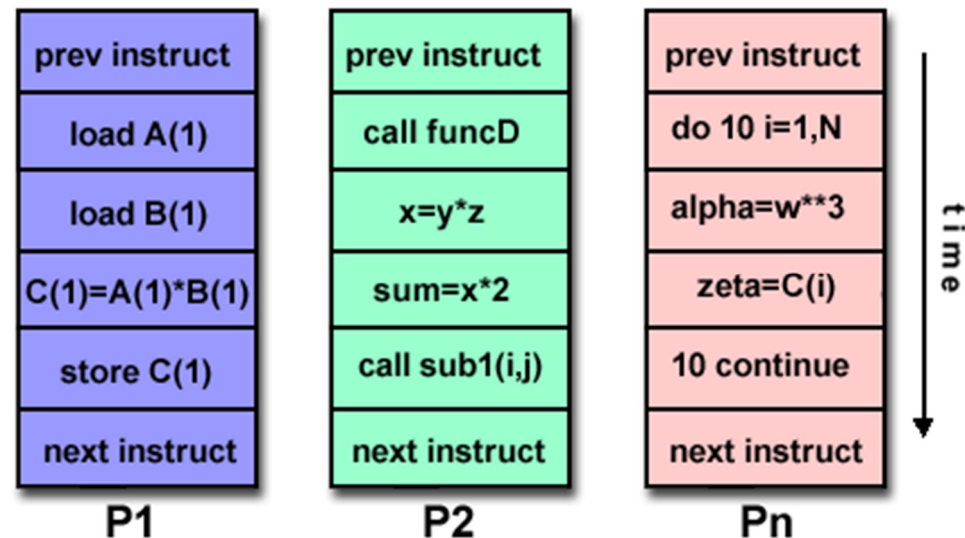
MISD

- A type of parallel computer
- Very few ever existed



MIMD

- A type of parallel computer
- Synchronous /asynchronous execution
- Examples: most modern supercomputers, parallel computing clusters, multicore PCs



Parallel Algorithms

Warm-up

“The way the processor industry is going, is to add more and more cores, but nobody knows how to program those things. I mean, two, yeah; four, not really; eight, forget it.”

— Steve Jobs, NY Times interview, June 10 2008

Parallel Algorithms Warm-up (1)

Consider the following loop:

for i = 1 to n do

$C[i] \leftarrow A[i] \times B[i]$

- Suppose you have an infinite number of processors/cores
- Ignore all overheads due to scheduling, memory accesses, communication, etc.
- Suppose each operation takes a constant amount of time
- How long will this loop take to complete execution?

Parallel Algorithms Warm-up (1)

Consider the following loop:

for i = 1 to n do

C[i] ← A[i] × B[i]

- Suppose you have an infinite number of processors/cores
- Ignore all overheads due to scheduling, memory accesses, communication, etc.
- Suppose each operation takes a constant amount of time
- How long will this loop take to complete execution?
 - *O(1) time*

Parallel Algorithms Warm-up (2)

Now consider the following loop:

$c \leftarrow 0$

for $i = 1$ *to* n *do*

$c \leftarrow c + A[i] \times B[i]$

- How long will this loop take to complete execution?

Parallel Algorithms Warm-up (2)

Now consider the following loop:

$c \leftarrow 0$

for $i = 1$ *to* n *do*

$c \leftarrow c + A[i] \times B[i]$

- How long will this loop take to complete execution?
 - $O(\log n)$ time

Parallel Algorithms Warm-up (3)

Now consider quicksort:

QSort(A)

if $|A| \leq 1$ return A

else $p \leftarrow A[\text{rand}(|A|)]$

return QSort($\{ x \in A: x < p \})$

{ p }

QSort($\{ x \in A: x > p \})$

- Assuming that A is split in the middle everytime, and the two recursive calls can be made in parallel, how long will this algorithm take?

Parallel Algorithms Warm-up (3)

Now consider quicksort:

QSort(A)

if $|A| \leq 1$ return A

else $p \leftarrow A[\text{rand}(|A|)]$

return QSort($\{ x \in A: x < p \}$)

{ p }

QSort($\{ x \in A: x > p \}$)

- Assuming that A is split in the middle everytime, and the two recursive calls can be made in parallel, how long will this algorithm take?
 - $O(\log^2 n)$ (if partitioning takes logarithmic time)
 - $O(\log n)$ (but can be partitioned in constant time)