Scientific Computing/Engineers – 22844
COSC 594 – 004

Lecture 2:
Overview of High-Performance Computing

Jack Dongarra
Electrical Engineering and Computer Science Department
University of Tennessee

1

November 2011: The TOP10

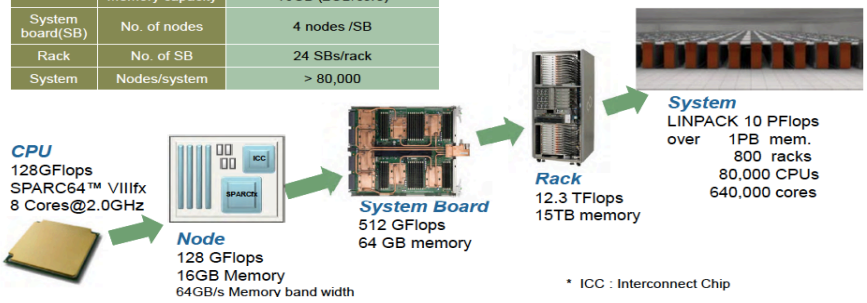
Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	MFlops /Watt
1	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx + custom	Japan	705,024	10.5	93	12.7	830
2	Nat. SuperComputer Center in Tianjin	Tianhe-1A, NUDT Intel + Nvidia GPU + custom	China	186,368	2.57	55	4.04	636
3	DOE / OS Oak Ridge Nat Lab	Jaguar, Cray AMD + custom	USA	224,162	1.76	75	7.0	251
4	Nat. Supercomputer Center in Shenzhen	Nebulea, Dawning Intel + Nvidia GPU + IB	China	120,640	1.27	43	2.58	493
5	GSIC Center, Tokyo Institute of Technology	Tusbame 2.0, HP Intel + Nvidia GPU + IB	Japan	73,278	1.19	52	1.40	865
6	DOE / NNSA LANL & SNL	Cielo, Cray AMD + custom	USA	142,272	1.11	81	3.98	279
7	NASA Ames Research Center/NAS	Pleiades SGI Altix ICE 8200EX/8400EX + IB	USA	111,104	1.09	83	4.10	265
8	DOE / OS Lawrence Berkeley Nat Lab	Hopper, Cray AMD + custom	USA	153,408	1.054	82	2.91	362
9	Commissariat a l'Energie Atomique (CEA)	Tera-10, Bull Intel + IB	France	138,368	1.050	84	4.59	229
10	DOE / NNSA Los Alamos Nat Lab	Roadrunner, IBM AMD + Cell GPU + IB	USA	122,400	1.04	76	2.35	446
500	IT Service	IBM Cluster, Intel + GigE	USA	7,236	.051	53		

Japanese K Computer

K computer Specifications

CPU (SPARC64 VIIIfx)	Cores/Node	8 cores (@2GHz)
	Performance	128GFlops
	Architecture	SPARC V9 + HPC extension
	Cache	L1(I/D) Cache : 32KB/32KB L2 Cache : 6MB
	Power	58W (typ. 30 C)
	Mem. bandwidth	64GB/s.
Node	Configuration	1 CPU / Node
	Memory capacity	16GB (2GB/core)
System board(SB)	No. of nodes	4 nodes /SB
Rack	No. of SB	24 SBs/rack
System	Nodes/system	> 80,000

Inter-connect	Topology	6D Mesh/Torus
	Performance	5GB/s. for each link
	No. of link	10 links/ node
	Additional feature	H/W barrier, reduction
	Architecture	Routing chip structure (no outside switch box)
Cooling	CPU, ICC*	Direct water cooling
	Other parts	Air cooling



New Linpack run with 705,024 cores at 10.51 Pflop/s (88,128 CPUs), 12.7 MW; 29.5 hours
Fujitsu to have a 100 Pflop/s system in 2014

3

China

- First Chinese Supercomputer to use a Chinese Processor
  - Sunway BlueLight MPP
  - ShenWei SW1600 processor, 16 core, 65 nm, fabbed in China
  - 125 Gflop/s peak
  - #14 with 139,364 cores, .796 Pflop/s & 1.07 Pflop/s Peak
  - Power Efficiency 741 Mflops/W

- Coming soon, Loongson (Godson) processor
  - 8-core, 65nm Loongson 3B processor runs at 1.05 GHz, with a peak performance of 128 Gflop/s

The New York Times

China Has Homemade Supercomputer Gain

By JOHN MARKOFF
Published: October 28, 2011

China has made its first supercomputer based on Chinese microprocessor chips, an advance that surprised high-performance computing specialists in the United States.

4

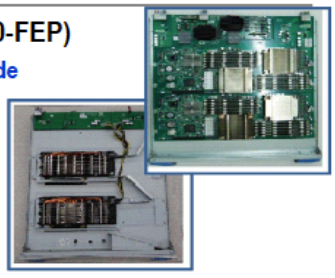
10+ Pflop/s Systems Planned in the States

- DOE Funded, Titan at ORNL, Based on Cray design w/AMD & Nvidia accelerators,
  - 20 Pflop/s, 2012
- DOE Funded, Sequoia at Lawrence Livermore Nat. Lab, Based on IBM's BG/Q,
  - 20 Pflop/s, 2012
- DOE Funded, BG/Q at Argonne National Lab, Based on IBM's BG/Q,
  - 10 Pflop/s, 2012
- NSF Funded, Blue Waters at University of Illinois UC, Based Cray XE6/XK6 hybrid,
  - 11.5 Pflop/s, 2012
- NSF Funded, U of Texas, Austin, Based on Dell/Intel MIC,
  - 10 Pflop/s, 2013

Tianhe-1A

Main configuration of TH-1A system

- 7,168 compute nodes (YH-X5670-FEP)
  - 2 six-core CPU and 1 GPU per node
  - CPU: Xeon X5670 (Westmere)
    - Processor speed - 2.93GHz
  - GPU: nVIDIA M2050
    - Connected with CPU by PCI-E
  - 32GB memory per node
  - 2U height



7168(nodes) × 2(CPU) ×2.93(GHz)× 6(Cores) ×4 =1.008PFlops

\+

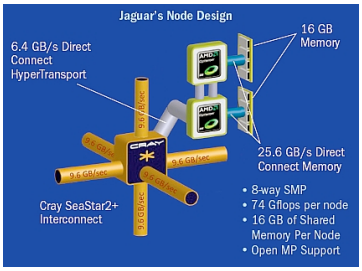
7168(nodes) × 1(GPU) ×1.15(GHz)*448(CUDA Cores) =3.692PFlops

Total: 4,701,061 GFlops

#3 ORNL's Jaguar System - Cray XT5

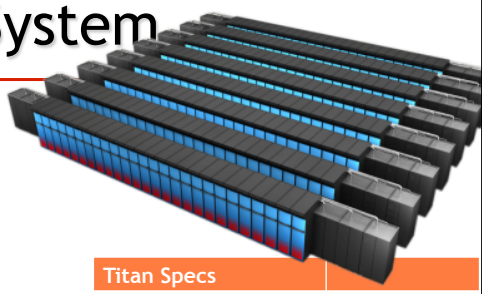
2.3 Pflop/s system with more than 224K processor cores using AMD's 6 Core chip.

Peak performance	**2.3 PF**
System memory	300 TB
Disk space	10 PB
Disk bandwidth	240+ GB/s
Interconnect bandwidth	374 TB/s



U.S. DEPARTMENT OF ENERGY | Office of Science

ORNL's "Titan" System



- Upgrade of existing Jaguar Cray XT5
- Cray Linux Environment operating system
- Gemini interconnect
  - 3-D Torus
  - Globally addressable memory
  - Advanced synchronization features
  - AMD Opteron 6200 processor (Interlagos)
- New accelerated node design using NVIDIA multi-core accelerators
  - 2011: 960 NVIDIA M2090 "Fermi" GPUs
  - 2012: 10-20 PF NVIDIA "Kepler" GPUs
- 10-20 PFlops peak performance
  - Performance based on available funds
- 600 TB DDR3 memory (2x that of Jaguar)

Titan Specs	
Compute Nodes	18,688
Login & I/O Nodes	512
Memory per node	32 GB + 6 GB
NVIDIA "Fermi" (2011)	665 GFlops
# of Fermi chips	960
NVIDIA "Kepler" (2012)	>1 TFlops
Opteron	2.2 GHz
Opteron performance	141 GFlops
Total Opteron Flops	2.6 PFlops
Disk Bandwidth	~ 1 TB/s

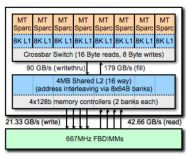
NSF Supercomputing Centers

University of Illinois - Blue Waters will be the powerhouse of the National Science Foundation's strategy to support supercomputers for scientists nationwide

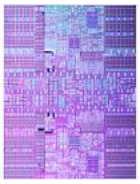
T1	Blue Waters	NCSA/Illinois	10 Pflop/s peak; 1 Pflop/s *sustained* per second in 2012
T2	Kraken	NICS/U of Tennessee	1 Pflop/s peak per second
T2	Ranger	TACC/U of Texas	504 Tflop/s peak per second
T3	Campuses across the U.S.	Several sites	50-100 Tflop/s peak per second

Today's Multicores

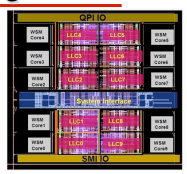
99% of Top500 Systems Are Based on Multicore



Sun Niagra2 (8 cores)

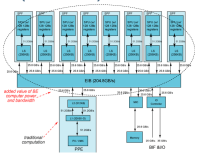


IBM Power 7 (8 cores)

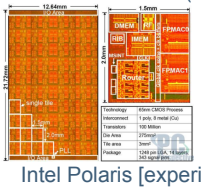


Intel Westmere (10 cores)

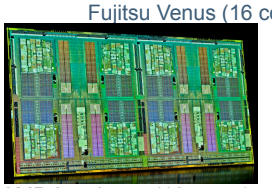
Fujitsu Venus (16 cores)



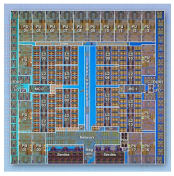
IBM Cell (9 cores)



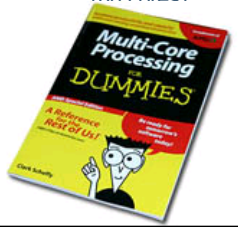
Intel Polaris [experimental]



AMD Interlagos (16 cores)

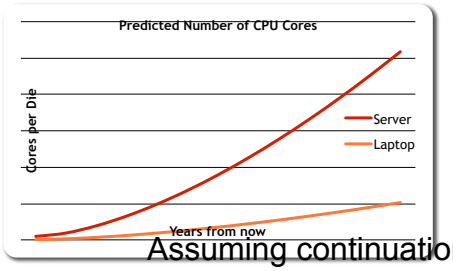


IBM BG/Q (18 cores)

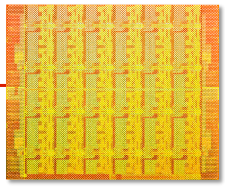


General Purpose CPU Concurrency Trends

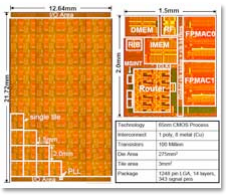
- Today
  - Typical server node chip ~ 8 cores
  - 1k node cluster ➔ 8,000 cores
  - Laptop ~ 2 cores (low power)
- By 2020
  - Typical server node chip ~400 cores
  - 1k node cluster ➔ 400,000 cores
  - Laptop ~ 100 cores (low power)



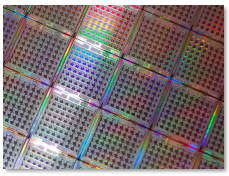
Assuming continuation of 58% historical density improvement per year



Intel SCC 48 cores



Intel 80 cores (teraflop



Tilera 100 GP cores

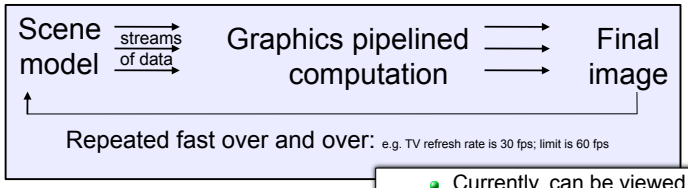
Future Computer Systems

- Most likely be a hybrid design
- Think standard multicore chips and accelerator (GPUs)
- Today accelerators are attached
- Next generation more integrated
- Intel's "Knights Corner" and "Knights Ferry" to come.
  - 48 x86 cores
- AMD's Fusion in 2011 - 2013
  - Multicore with embedded graphics ATI
- Nvidia's Project Denver plans to develop an integrated chip using ARM architecture in 2013.

12

Evolution of GPUs

GPUs: excelling in graphics rendering



- Currently, can be viewed as **multithreaded multicore vector units**

This type of computation:

- Requires **enormous computational power**
- Allows for **high parallelism**
- Needs **high bandwidth *vs* low latency**

(as low latencies can be compensated with deep graphics pipeline)

Obviously, this pattern of computation is common with many other applications

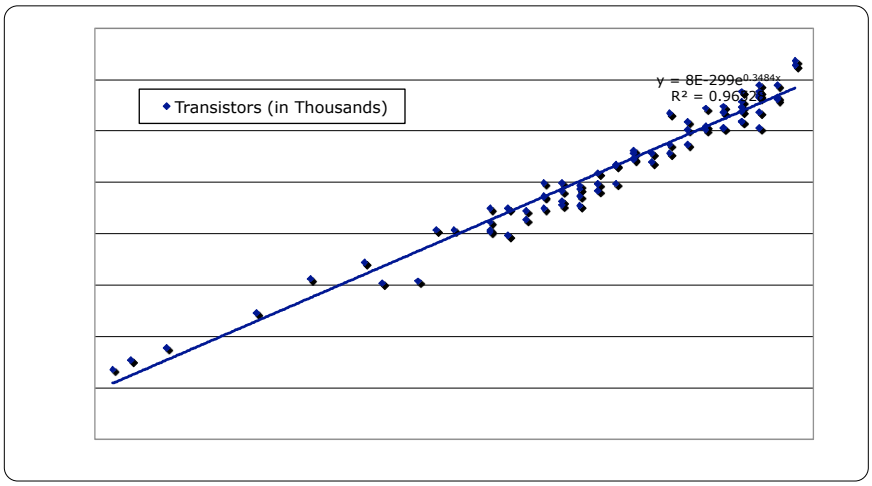
Challenges of using GPUs

- **High levels of parallelism**
  Many GPU cores, serial kernel execution
  [e.g. 240 in the Nvidia Tesla; up to 512 in *Fermi* - to have concurrent kernel execution]

- **Hybrid/heterogeneous architectures**
  Match algorithmic requirements to architectural strengths
  [e.g. small, non-parallelizable tasks to run on CPU, large and parallelizable on GPU]

- **Compute *vs* communication gap**
  Exponentially growing gap; persistent challenge
  [Processor speed improves 59%, memory bandwidth 23%, latency 5.5%]
  [on all levels, e.g. a GPU Tesla C1070 (4 x C1060) has compute power o O(1,000)
  Gflop/s but GPUs communicate through the CPU using O(1) GB/s connection]

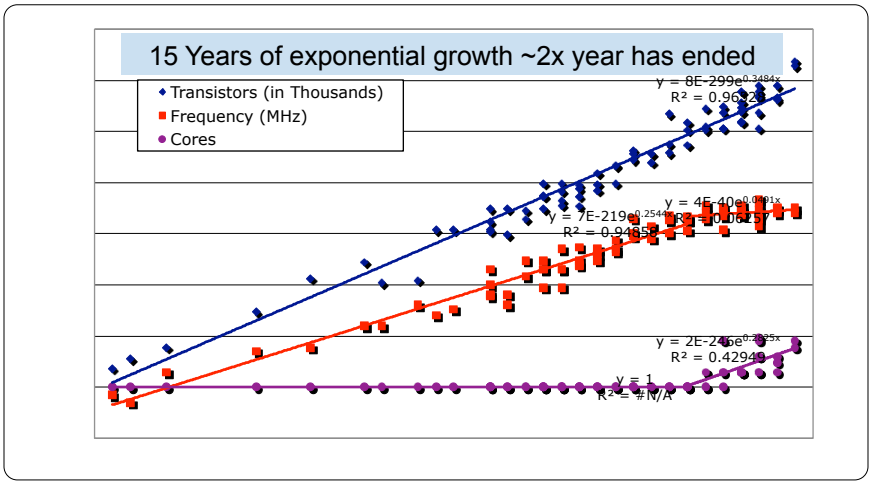
Moore's Law is Alive and Well



Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç

Slide from Kathy Yelick

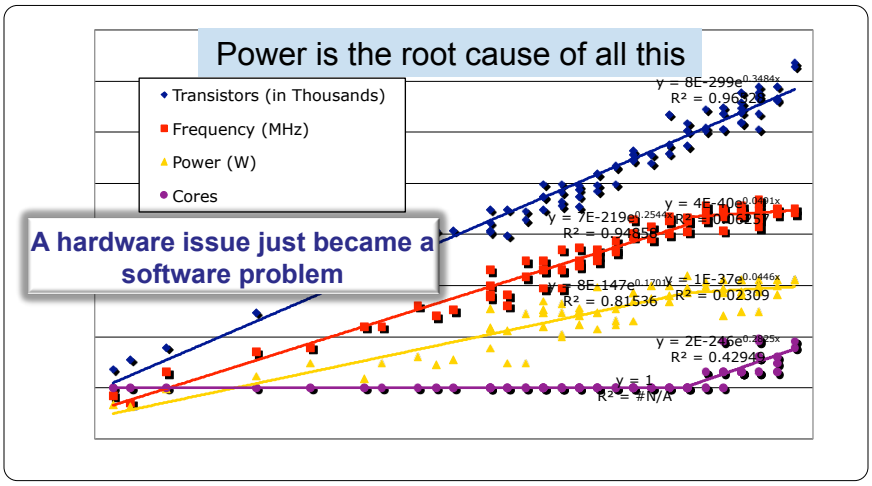
But Clock Frequency Scaling Replaced by Scaling Cores / Chip



Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç

Slide from Kathy Yelick

Performance Has Also Slowed, Along with Power



Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç

Slide from Kathy Yelick

Power Cost of Frequency

- Power ∝ Voltage^2 x Frequency (V^{2}F)
- Frequency ∝ Voltage
- Power ∝ Frequency^3

	Cores	V	Freq	Perf	Power	PE (Bops/watt)
Superscalar	1	1	1	1	1	1
"New" Superscalar	1X	1.5X	1.5X	1.5X	3.3X	0.45X

8

Power Cost of Frequency

- Power ∝ Voltage^2 x Frequency (V^{2}F)
- Frequency ∝ Voltage
- Power ∝ Frequency^3

	Cores	V	Freq	Perf	Power	PE (Bops/watt)
Superscalar	1	1	1	1	1	1
"New" Superscalar	1X	1.5X	1.5X	1.5X	3.3X	0.45X
Multicore	2X	0.75X	0.75X	1.5X	0.8X	1.88X

(Bigger # is better)

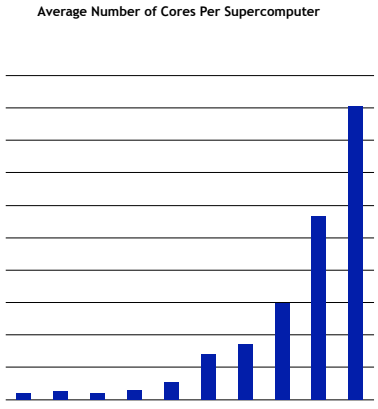
50% more performance with 20% less power

Preferable to use multiple slower devices, than one superfast device

19

Moore's Law Reinterpreted

- Number of cores per chip doubles every 2 year, while clock speed decreases (not increases).
  - Need to deal with systems with millions of concurrent threads
    - Future generation will have billions of threads!
  - Need to be able to easily replace inter-chip parallelism with intro-chip parallelism
- Number of threads of execution doubles every 2 year

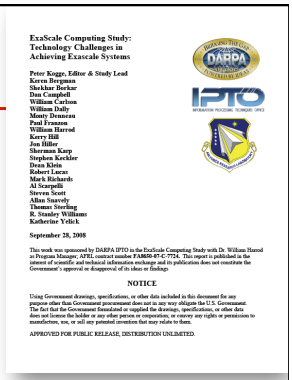


Major Changes to Software

- **Must rethink the design of our software**
  - **Another disruptive technology**
    - Similar to what happened with cluster computing and message passing
  - **Rethink and rewrite the applications, algorithms, and software**

21

Exascale Computing


ExaScale Computing Study:
Technology Challenges in
Achieving Exascale Systems

Peter Kogge, Editor & Study Lead
Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snavely
Thomas Sterling
R. Stanley Williams
Katherine Yelick

September 28, 2008

This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod as Program Manager; AFRL contract number FA8650-07-C-7724. This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

NOTICE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government furnished or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED.

- Exascale systems are likely feasible by 2017±2
- 10-100 Million processing elements (cores or mini-cores) with chips perhaps as dense as 1,000 cores per socket, clock rates will grow more slowly
- 3D packaging likely
- Large-scale optics based interconnects
- 10-100 PB of aggregate memory
- Hardware and software based fault management
- Heterogeneous cores
- Performance per watt — stretch goal 100 GF/watt of sustained performance $\Longrightarrow$ >> 10 - 100 MW Exascale system
- Power, area and capital costs will be significantly higher than for today's fastest systems

22

Google: exascale computing study

Factors that Necessitate Redesign of Our Software

- Steepness of the ascent from terascale to petascale to exascale
- Extreme parallelism and hybrid design
  - Preparing for million/billion way parallelism
- Tightening memory/bandwidth bottleneck
  - Limits on power/clock speed implication on multicore
  - Reducing communication will become much more intense
  - Memory per core changes, byte-to-flop ratio will change
- Necessary Fault Tolerance
  - MTTF will drop
  - Checkpoint/restart has limitations

Software infrastructure does not exist today

www.exascale.org

Conclusions

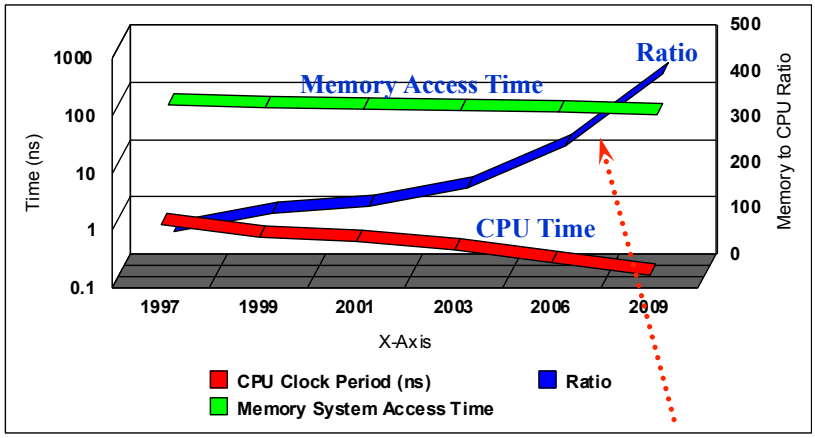
- For the last decade or more, the research investment strategy has been overwhelmingly biased in favor of hardware.
- This strategy needs to be rebalanced - barriers to progress are increasingly on the software side.
- Moreover, the return on investment is more favorable to software.
  - Hardware has a half-life measured in years, while software has a half-life measured in decades.
- High Performance Ecosystem out of balance
  - Hardware, OS, Compilers, Software, Algorithms, Applications
    - No Moore's Law for software, algorithms and applications

Why Fast Machines Run Slow

- Latency
  - Waiting for access to memory or other parts of the system
- Overhead
  - Extra work that has to be done to manage program concurrency and parallel resources the real work you want to perform
- Starvation
  - Not enough work to do due to insufficient parallelism or poor load balancing among distributed resources
- Contention
  - Delays due to fighting over what task gets to use a shared resource next. Network bandwidth is a major constraint.

25

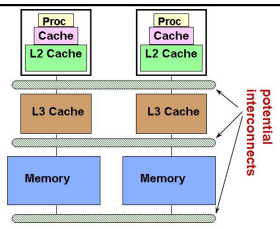
Latency in a Single System



THE *WALL*

26

Memory hierarchy



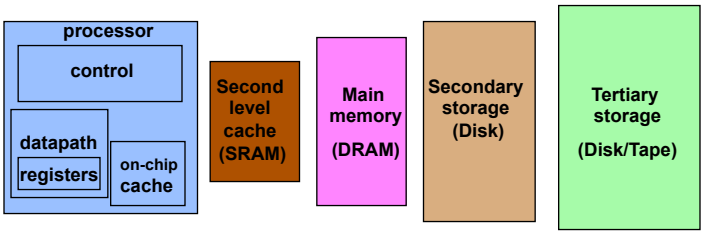
- Typical latencies for today's technology

Hierarchy	Processor clocks
Register	1
L1 cache	2-3
L2 cache	6-12
L3 cache	14-40
Near memory	100-300
Far memory	300-900
Remote memory	$O(10^3)$
Message-passing	$O(10^3)$-$O(10^4)$

27

Memory Hierarchy

- Most programs have a high degree of locality in their accesses
  - spatial locality: accessing things nearby previous accesses
  - temporal locality: reusing an item that was previously accessed
- Memory hierarchy tries to exploit locality



Speed	1ns	10ns	100ns	10ms	10sec
Size	B	KB	MB	GB	TB

Percentage of peak

- A rule of thumb that often applies
  - A contemporary RISC processor, for a spectrum of applications, delivers (i.e., sustains) 10% of peak performance
- There are exceptions to this rule, in both directions
- Why such low efficiency?
- There are two primary reasons behind the disappointing percentage of peak
  - IPC (in)efficiency
  - Memory (in)efficiency

29

Different Architectures

- Parallel computing: single systems with many processors working on same problem
- Distributed computing: many systems loosely coupled by a scheduler to work on related problems
- Grid Computing: many systems tightly coupled by software, perhaps geographically distributed, to work together on single problems or on related problems

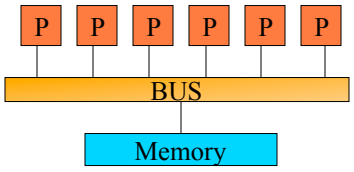
30

Types of Parallel Computers

- **The simplest and most useful way to classify modern parallel computers is by their memory model:**
  - **shared memory**
  - **distributed memory**

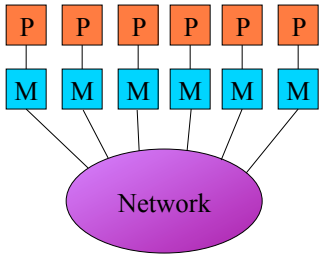
31

Shared vs. Distributed Memory



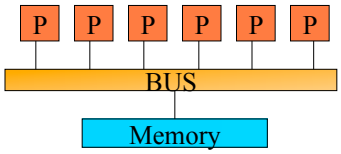
Shared memory - single address space. All processors have access to a pool of shared memory. (Ex: SGI Origin, Sun E10000)

Distributed memory - each processor has it's own local memory. Must do message passing to exchange data between processors. (Ex: CRAY T3E, IBM SP, clusters)



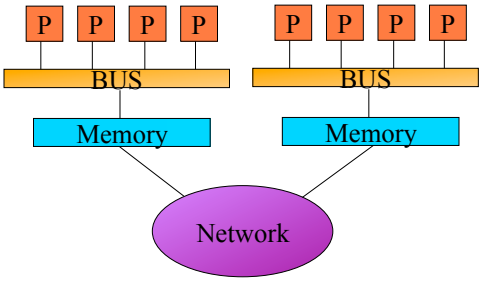
32

Shared Memory: UMA vs. NUMA



Uniform memory access (UMA): Each processor has uniform access to memory. Also known as symmetric multiprocessors (Sun E10000)

Non-uniform memory access (NUMA): Time for memory access depends on location of data. Local access is faster than non-local access. Easier to scale than SMPs (SGI Origin)



33