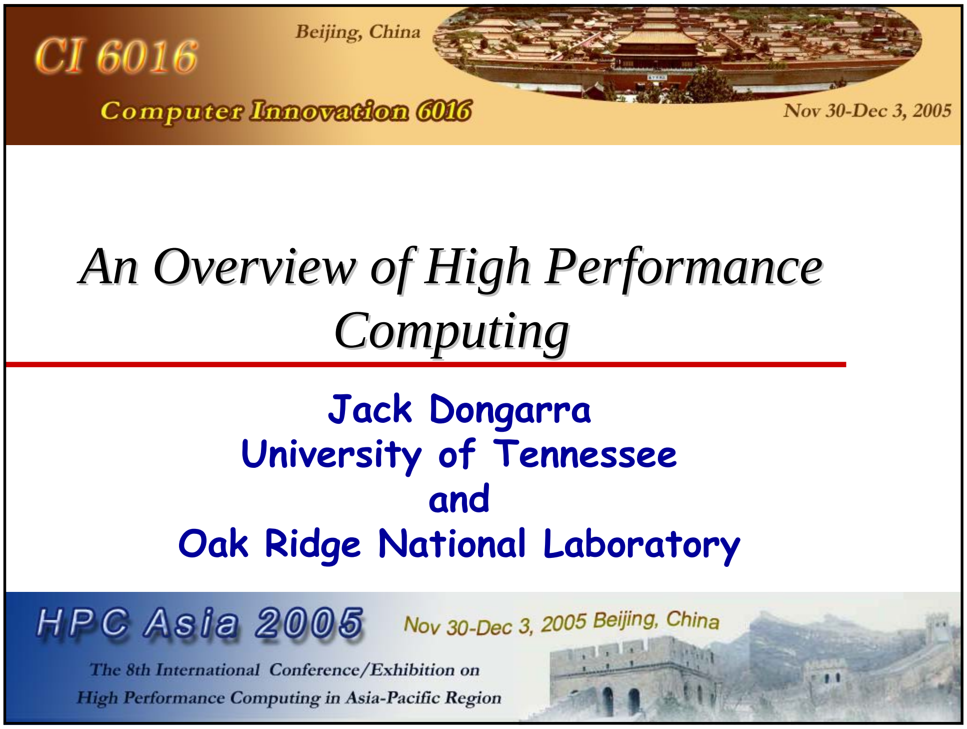


CI 6016

Beijing, China

Computer Innovation 6016

Nov 30-Dec 3, 2005

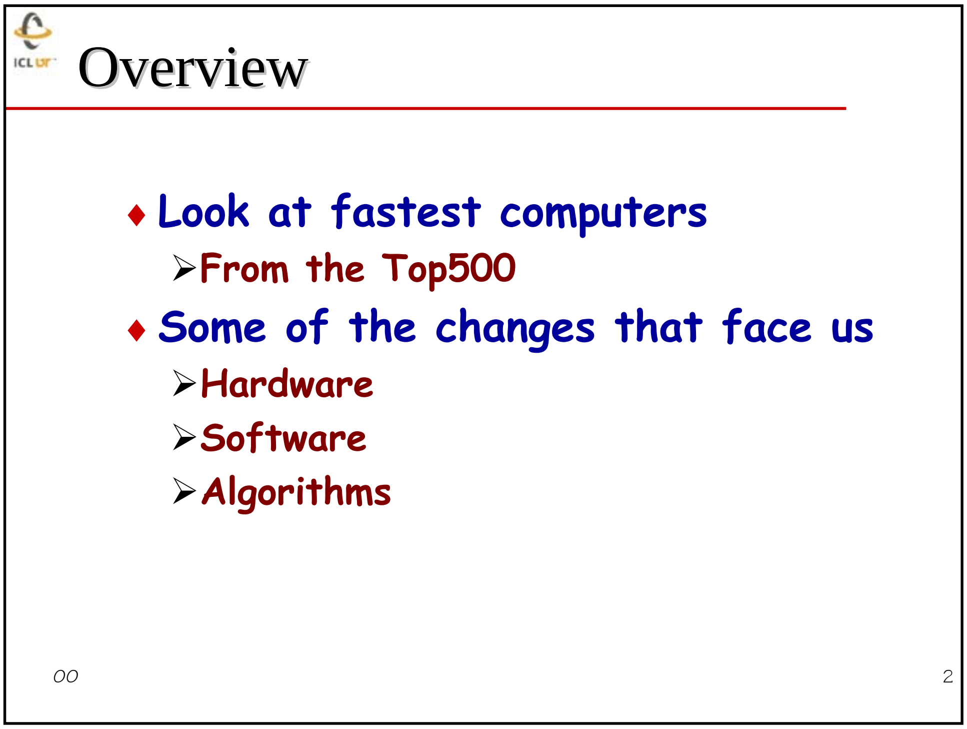
An Overview of High Performance Computing

Jack Dongarra
University of Tennessee
and
Oak Ridge National Laboratory

HPC Asia 2005

Nov 30-Dec 3, 2005 Beijing, China

The 8th International Conference/Exhibition on
High Performance Computing in Asia-Pacific Region

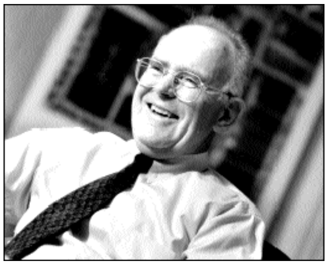


Overview

- ♦ Look at fastest computers
  - ➢From the Top500
- ♦ Some of the changes that face us
  - ➢Hardware
  - ➢Software
  - ➢Algorithms

00 2

Technology Trends: Microprocessor Chip Capacity



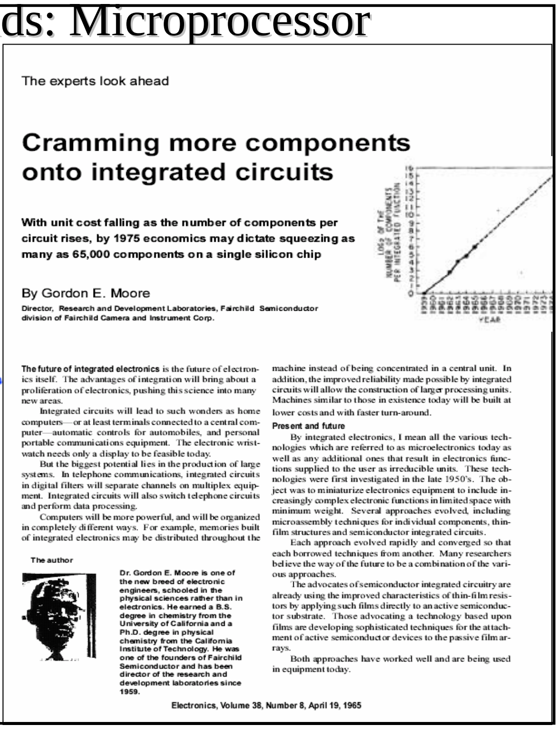
Gordon Moore
(co-founder of Intel)

Electronics Magazine, 1965

Number of devices/chip doubles every 18 months

2X transistors/Chip Every 1.5 years "Moore's Law"

OO



The experts look ahead

Cramming more components onto integrated circuits

With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip

By Gordon E. Moore

Director, Research and Development Laboratories, Fairchild Semiconductor division of Fairchild Camera and Instrument Corp.

4

The future of integrated electronics is the future of electronics itself. The advantages of integration will bring about a proliferation of electronics, pushing this science into many new areas.

Integrated circuits will lead to such wonders as home computers—or at least terminals connected to a central computer—automatic controls for automobiles, and personal portable communications equipment. The electronic wristwatch needs only a display to be feasible today.

But the biggest potential lies in the production of large systems. In telephone communications, integrated circuits in digital filters will separate channels on multiplex equipment. Integrated circuits will also switch telephone circuits and perform data processing.

Computers will be more powerful, and will be organized in completely different ways. For example, memories built of integrated electronics may be distributed throughout the machine instead of being concentrated in a central unit. In addition, the improved reliability made possible by integrated circuits will allow the construction of larger processing units. Machines similar to those in existence today will be built at lower costs and with faster turn-around.

Present and future

By integrated electronics, I mean all the various technologies which are referred to as microelectronics today as well as any additional ones that result in electronics functions supplied to the user as irreducible units. These technologies were first investigated in the late 1950's. The object was to miniaturize electronics equipment to include increasingly complex electronic functions in limited space with minimum weight. Several approaches evolved, including microassembly techniques for individual components, thin-film structures and semiconductor integrated circuits.

Each approach evolved rapidly and converged so that each borrowed techniques from another. Many researchers believe the way of the future to be a combination of the various approaches.

The advocates of semiconductor integrated circuitry are already using the improved characteristics of thin-film resistors by applying such films directly to an active semiconductor substrate. Those advocating a technology based upon films are developing sophisticated techniques for the attachment of active semiconductor devices to the passive film arrays.

Both approaches have worked well and are being used in equipment today.

The author

Dr. Gordon E. Moore is one of the new breed of electronic engineers, schooled in the physical sciences rather than in electronics. He earned a B.S. degree in chemistry from the University of California and a Ph.D. degree in physical chemistry from the California Institute of Technology. He was one of the founders of Fairchild Semiconductor and has been director of the research and development laboratories since 1959.

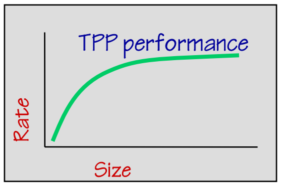
Electronics, Volume 38, Number 8, April 19, 1965

TOP500 superCOMPUTER

H. Meuer, H. Simon, E. Strohmaier, & JD

- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

Ax=b, dense problem



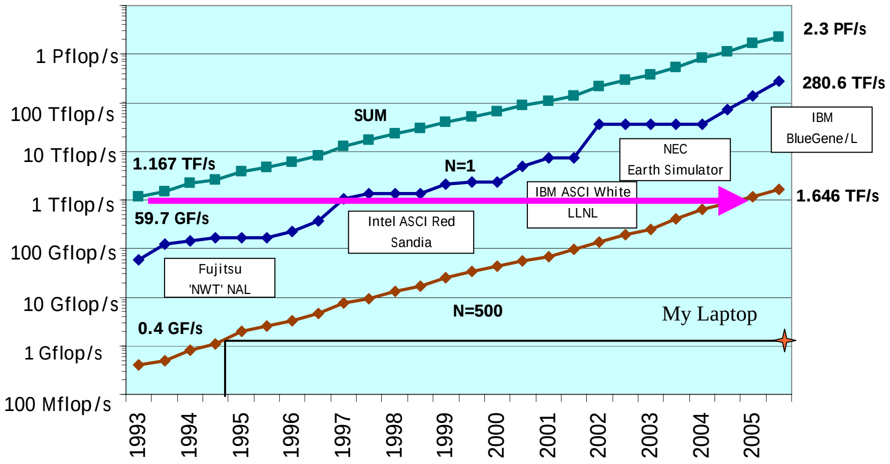
- Updated twice a year

  SC'xy in the States in November

  Meeting in Germany in June

OO- All data available from **www.top500.org**

Performance Development

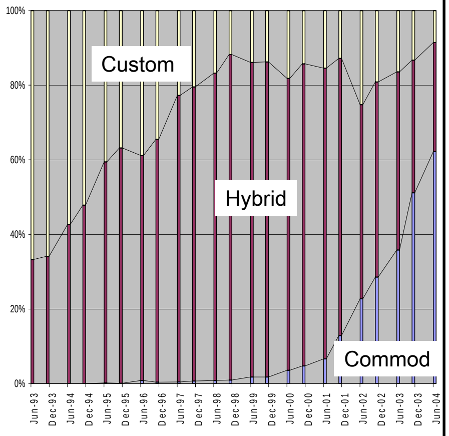


Architecture/Systems Continuum

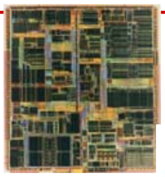
Tightly Coupled

Loosely Coupled

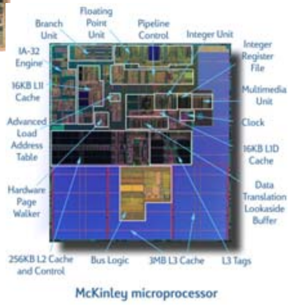
- ♦ Custom processor with custom interconnect
  - ➢ Cray X1
  - ➢ NEC SX-8
  - ➢ IBM Regatta
  - ➢ IBM Blue Gene/L
- ♦ Commodity processor with custom interconnect
  - ➢ SGI Altix
    - ➢ Intel Itanium 2
  - ➢ Cray XT3, XD1
    - ➢ AMD Opteron
- ♦ Commodity processor with commodity interconnect
  - ➢ Clusters
    - ➢ Pentium, Itanium, Opteron, Alpha
    - ➢ GigE, Infiniband, Myrinet, Quadrics
  - ➢ NEC TX7
  - ➢ IBM eServer
  - ➢ Dawning



Commodity Processors

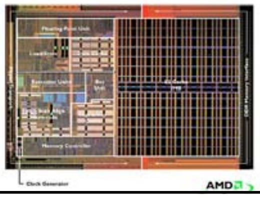


- ♦ **Intel Pentium Nocona**
  - ➢ 3.6 GHz, peak = 7.2 Gflop/s
  - ➢ Linpack 100 = 1.8 Gflop/s
  - ➢ Linpack 1000 = 4.2 Gflop/s

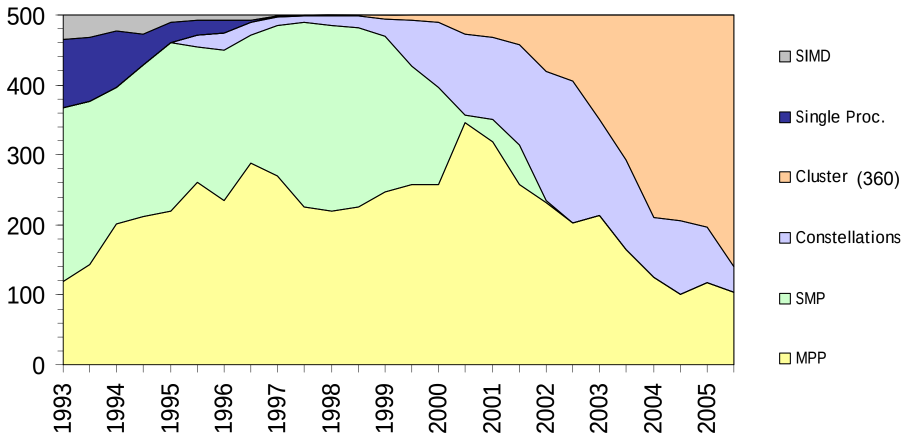


- ♦ **Intel Itanium 2**
  - ➢ 1.6 GHz, peak = 6.4 Gflop/s
  - ➢ Linpack 100 = 1.7 Gflop/s
  - ➢ Linpack 1000 = 5.7 Gflop/s

- ♦ **AMD Opteron**
  - ➢ 2.6 GHz, peak = 5.2 Gflop/s
  - ➢ Linpack 100 = 1.6 Gflop/s
  - ➢ Linpack 1000 = 3.9 Gflop/s

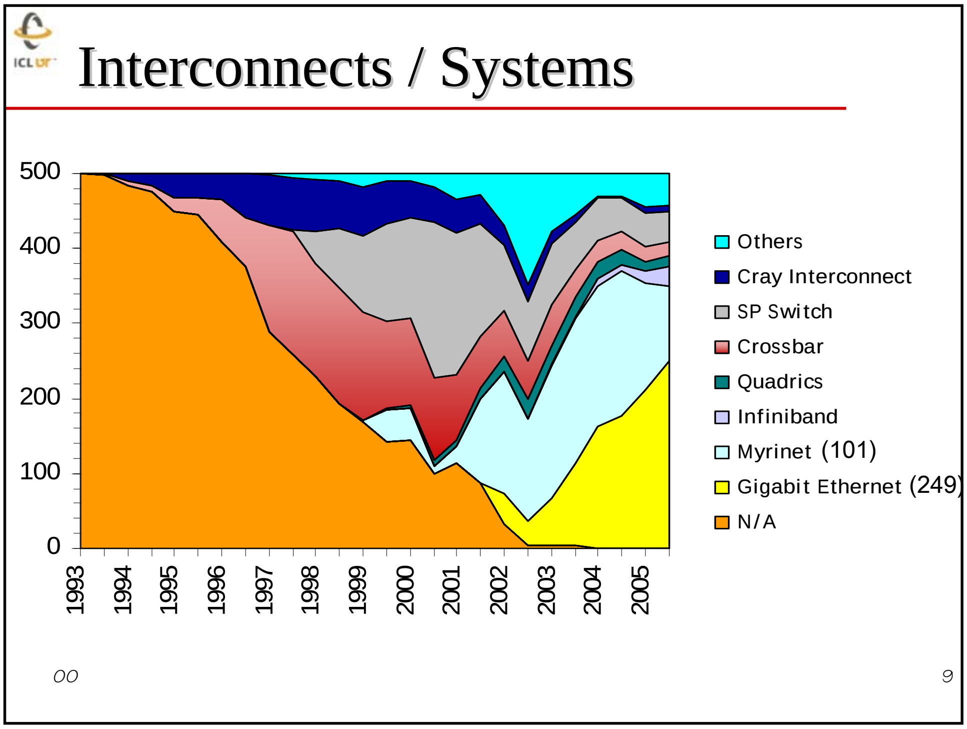


Architectures / Systems

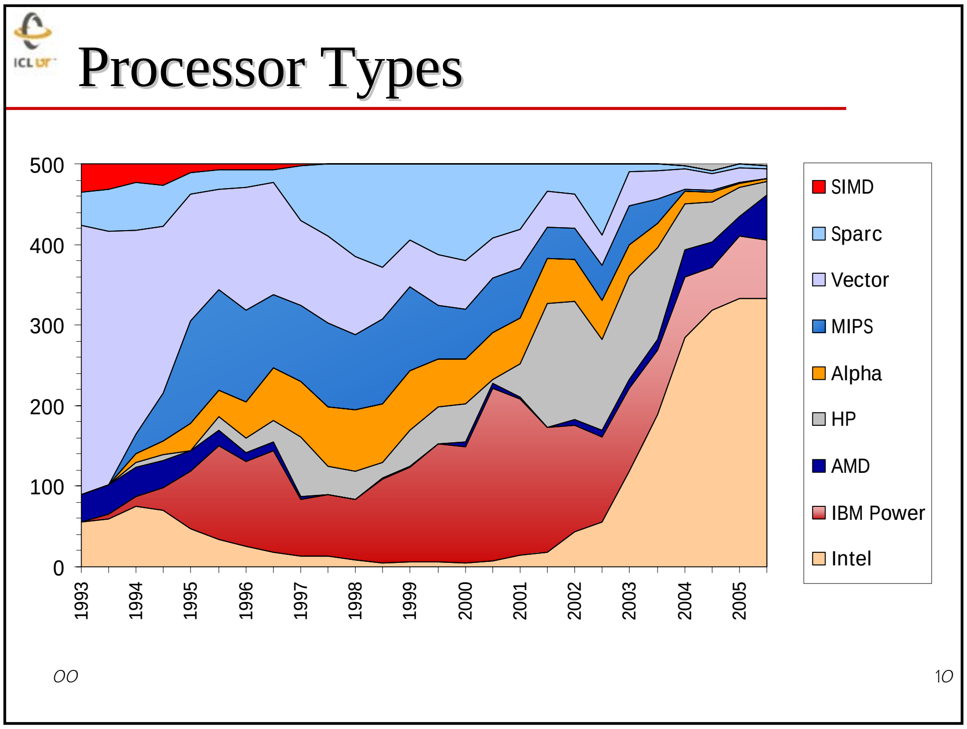


Cluster: Commodity processors & Commodity interconnect

Constellation: # of procs/node ≥ nodes in the system



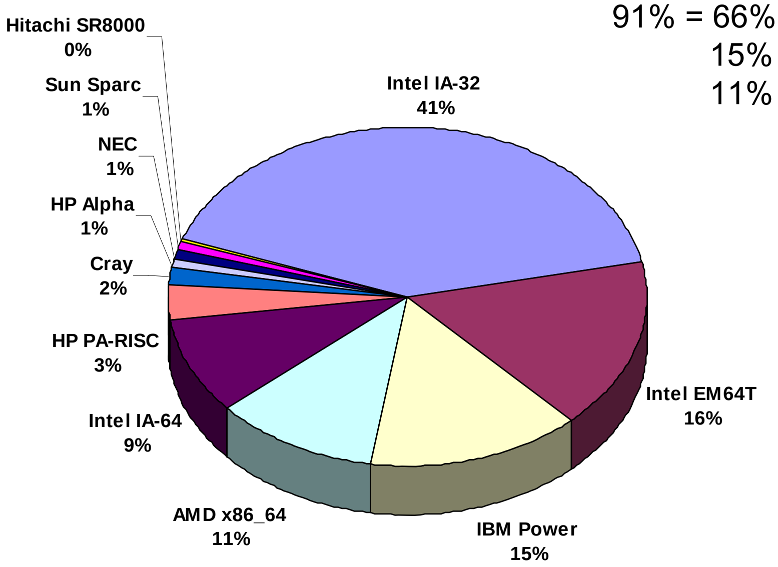
Interconnects / Systems



Processor Types

Processors Used in Each of the 500 Systems

91% = 66% Intel
15% IBM
11% AMD

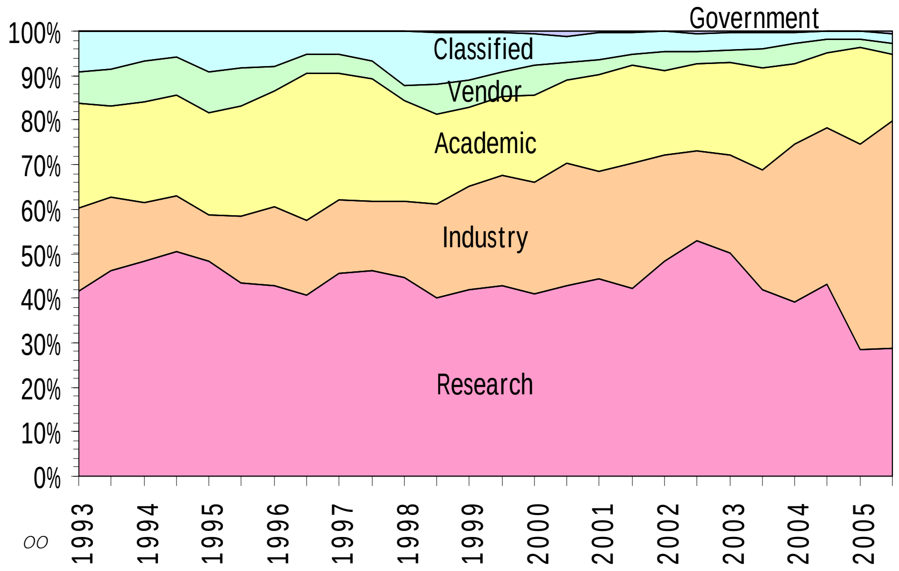


26th List: The TOP10

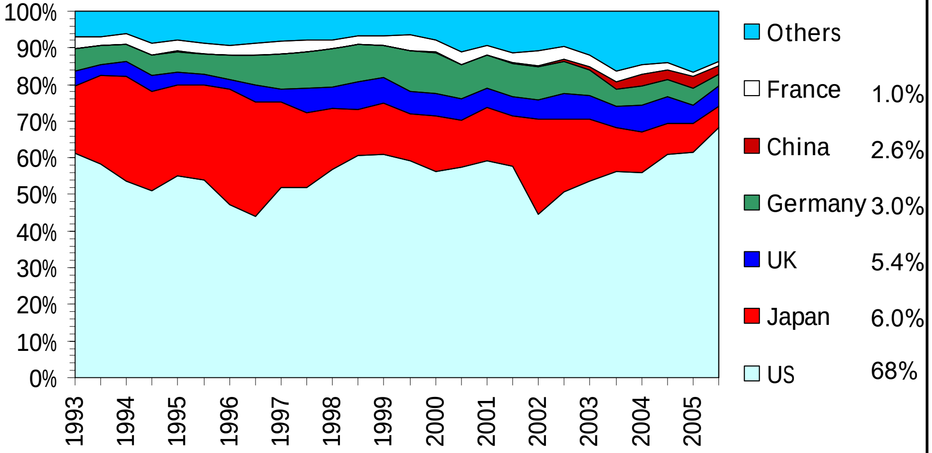
	Manufacturer	Computer	Rmax [TF/s]	Installation Site	Country	Year	#Proc
1	IBM	BlueGene/L eServer Blue Gene	280.6	DOE/NNSA/LLNL	USA	2005	131072
2	IBM	BGW eServer Blue Gene	91.29	IBM Thomas Watson	USA	2005	40960
3	IBM	ASC Purple Power5 p575	63.39	DOE/NNSA/LLNL	USA	2005	10240
4 ~~3~~	SGI	Columbia Altix, Itanium/Infiniband	51.87	NASA Ames	USA	2004	10160
5	Dell	Thunderbird Pentium/Infiniband	38.27	Sandia	USA	2005	8000
6 ~~10~~	Cray	Red Storm Cray XT3 AMD	36.19	Sandia	USA	2005	10880
7 ~~4~~	NEC	Earth-Simulator SX-5	35.86	Earth Simulator Center	Japan	2002	5120
8 ~~5~~	IBM	MareNostrum PPC 970/Myrinet	27.91	Barcelona Supercomputer Center	Spain	2005	4800
9 ~~6~~	IBM	eServer Blue Gene	27.45	ASTRON University Groningen	Netherlands	2005	12288
10	Cray	Jaguar Cray XT3 AMD	20.53	Oak Ridge National Lab	USA	2005	5200

OO

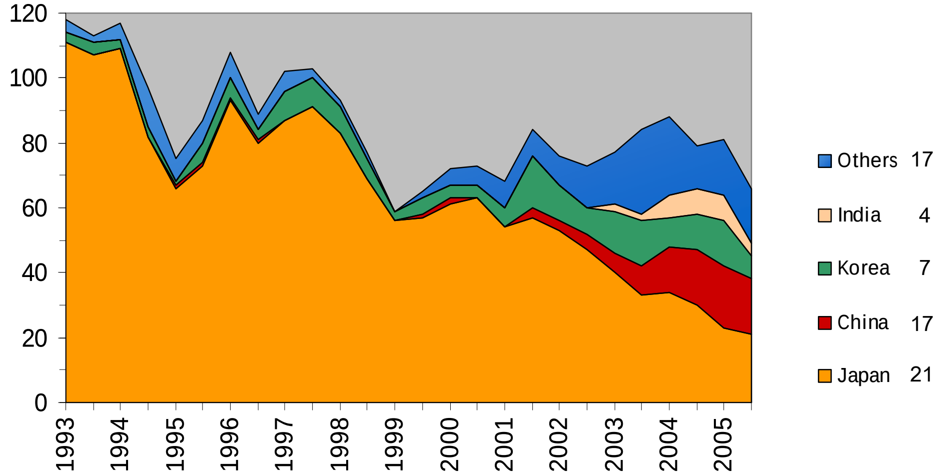
Customer Segments / Performance



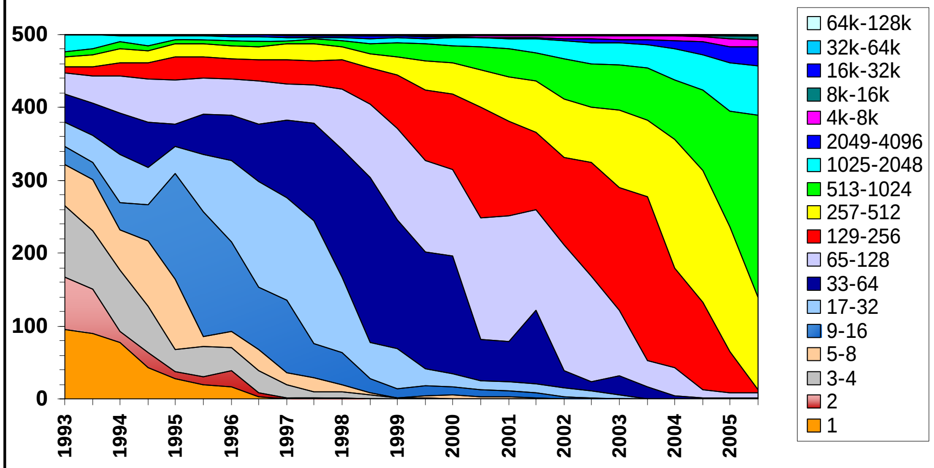
Countries / Performance

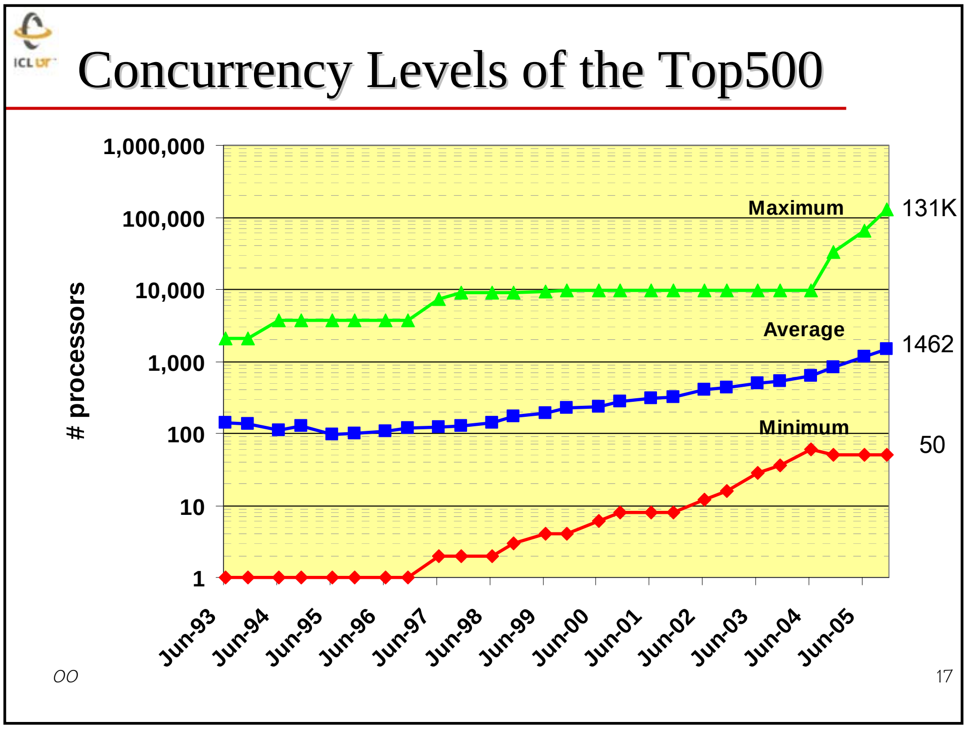


Asian Countries / Systems



Concurrency Levels of the Top500

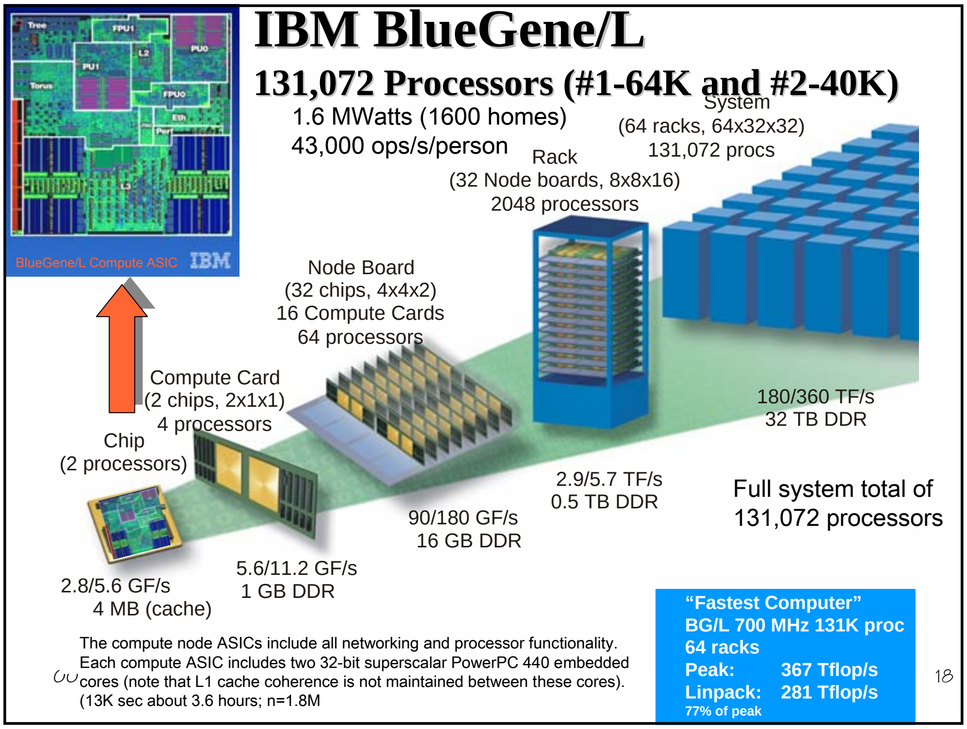




Concurrency Levels of the Top500

00

17



IBM BlueGene/L

131,072 Processors (#1-64K and #2-40K)

1.6 MWatts (1600 homes)
43,000 ops/s/person

Full system total of
131,072 processors

The compute node ASICs include all networking and processor functionality. Each compute ASIC includes two 32-bit superscalar PowerPC 440 embedded cores (note that L1 cache coherence is not maintained between these cores).
(13K sec about 3.6 hours; n=1.8M

"Fastest Computer"
BG/L 700 MHz 131K proc
64 racks

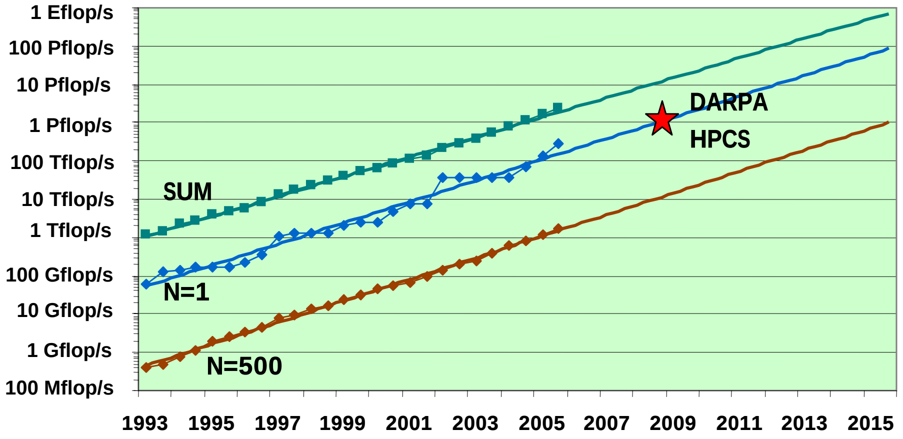
Peak:	367 Tflop/s
Linpack:	281 Tflop/s

77% of peak

00

18

Performance Projection



A PetaFlop Computer by the End of the Decade

- 10 Companies working on a building a Petaflop system by the end of the decade.
  - Cray
  - IBM
  - Sun

  HPCS

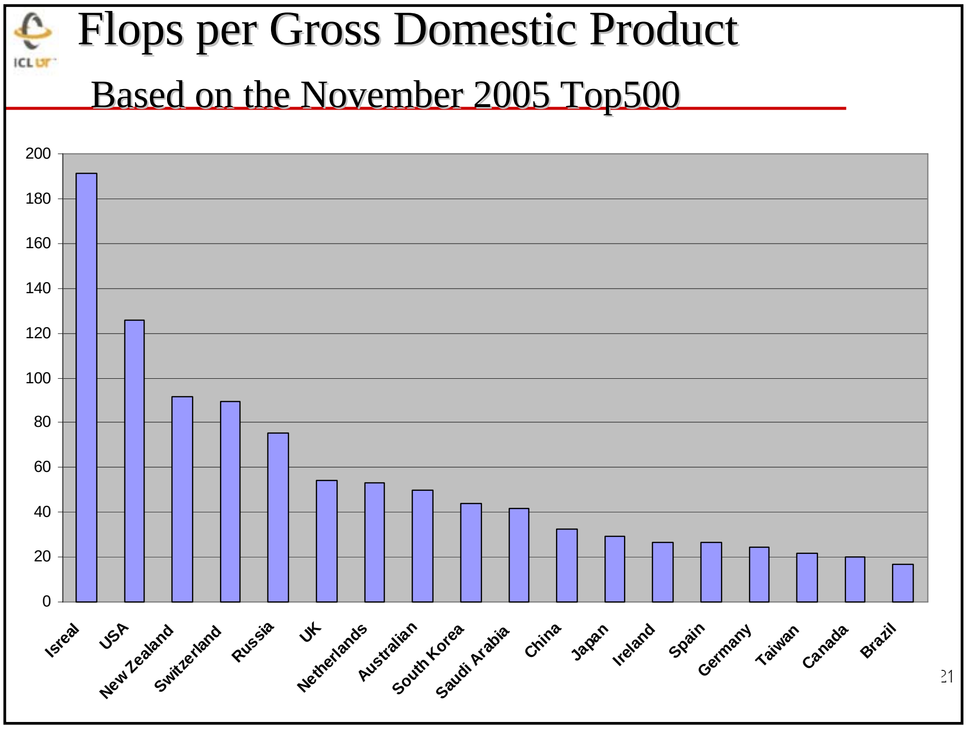
  - Dawning
  - Galactic
  - Lenovo

  Chinese Companies

  - Hitachi
  - NEC
  - Fujitsu

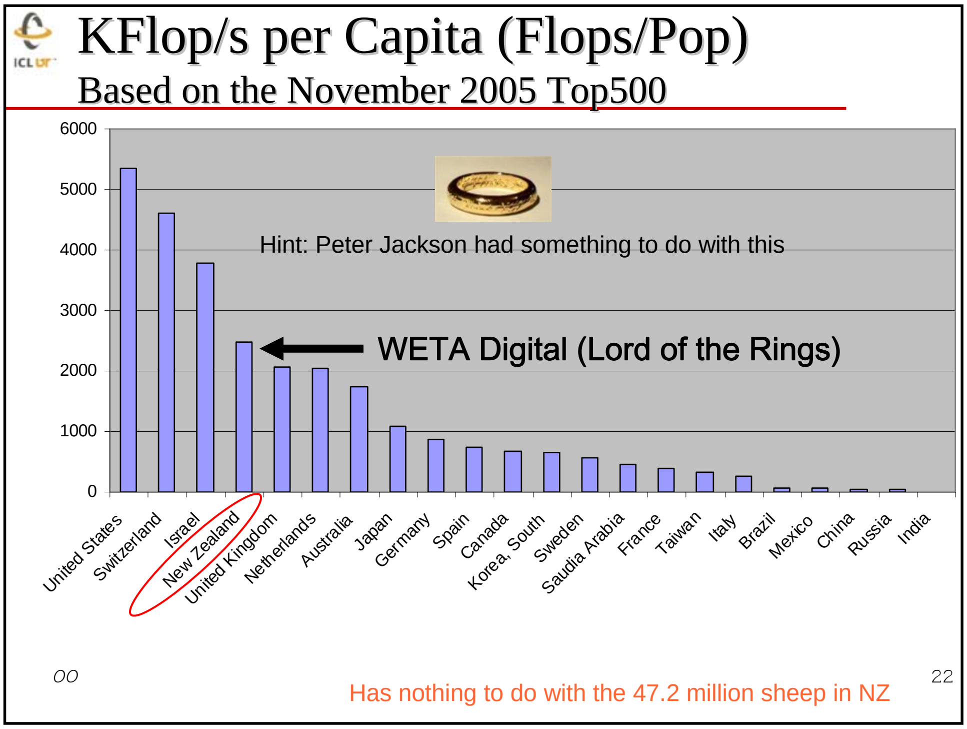
  Japanese
  “Life Simulator” (10 Pflop/s)

  - Bull



Flops per Gross Domestic Product

Based on the November 2005 Top500



KFlop/s per Capita (Flops/Pop)

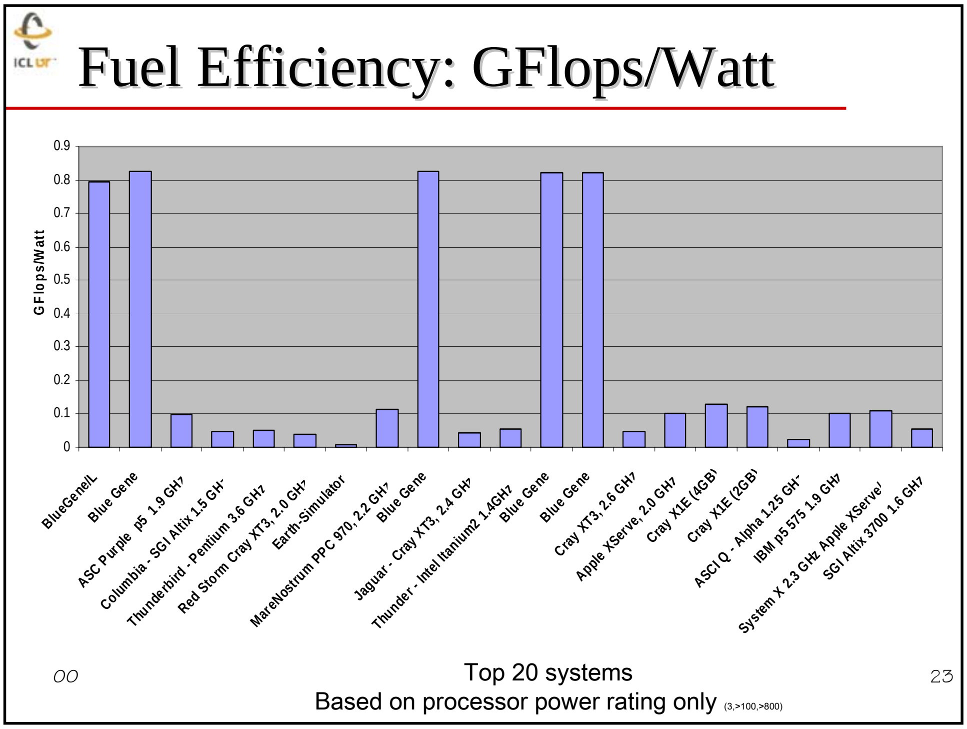
Based on the November 2005 Top500

Hint: Peter Jackson had something to do with this

WETA Digital (Lord of the Rings)

OO

Has nothing to do with the 47.2 million sheep in NZ



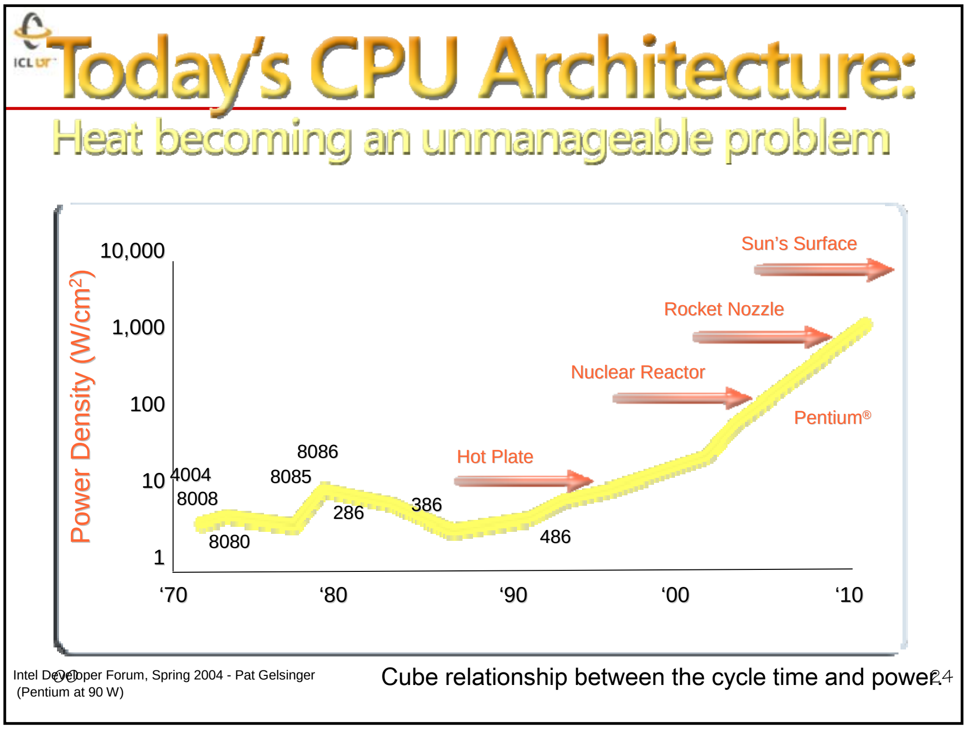
Fuel Efficiency: GFlops/Watt

00

Top 20 systems

Based on processor power rating only (3,>100,>800)

23



Today's CPU Architecture:

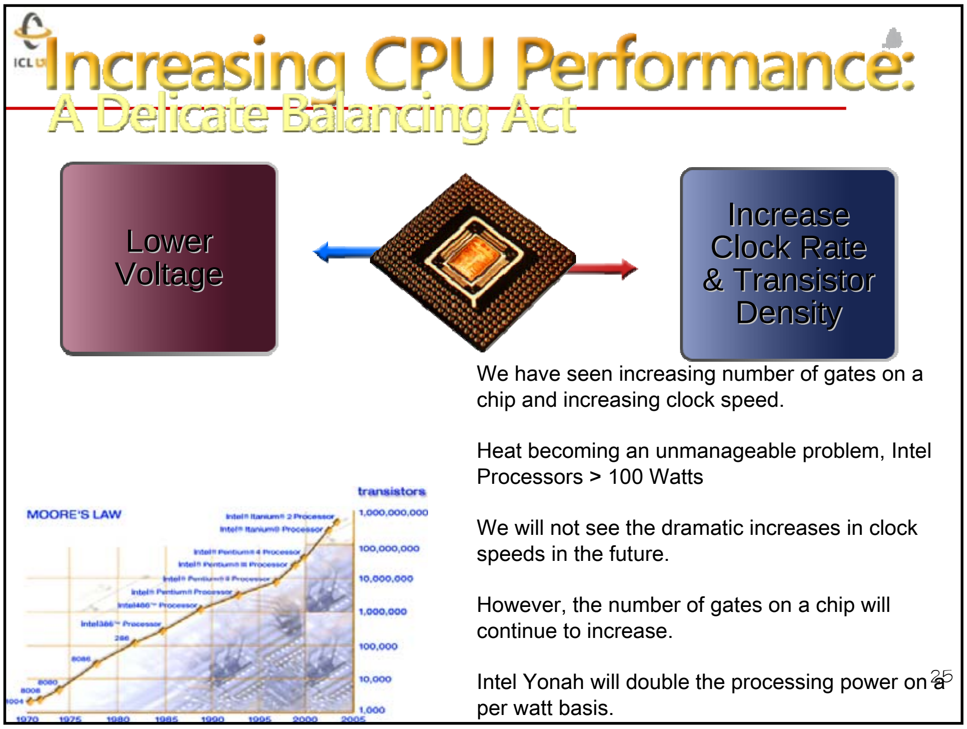
Heat becoming an unmanageable problem

Intel Developer Forum, Spring 2004 - Pat Gelsinger
(Pentium at 90 W)

00

Cube relationship between the cycle time and power.

24



Increasing CPU Performance: A Delicate Balancing Act

Lower Voltage

Increase Clock Rate & Transistor Density

We have seen increasing number of gates on a chip and increasing clock speed.

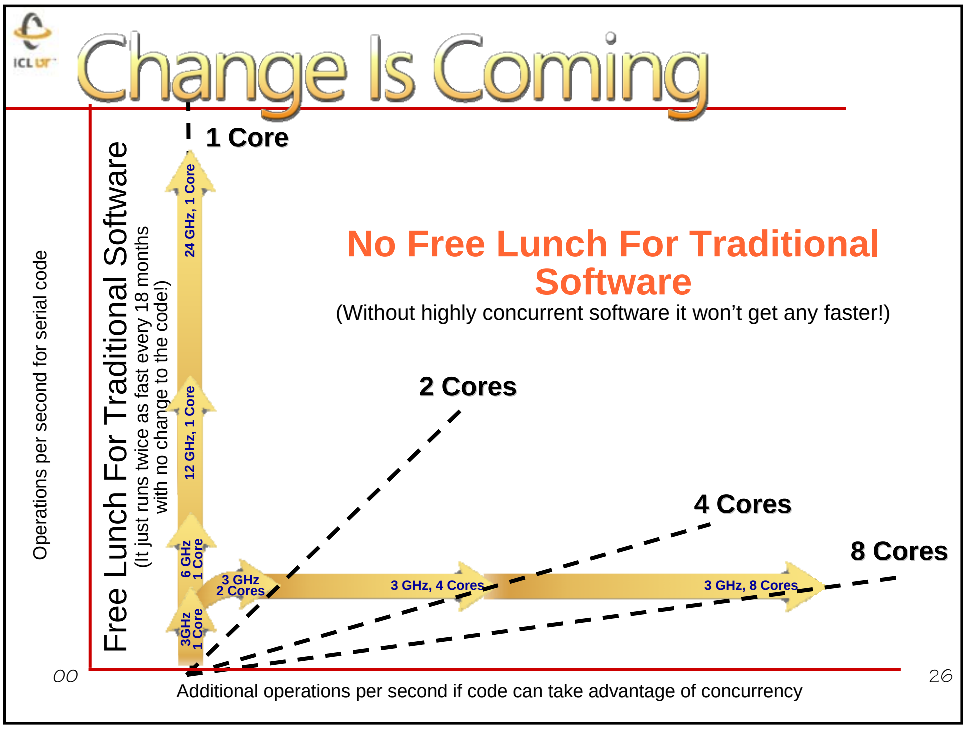
Heat becoming an unmanageable problem, Intel Processors > 100 Watts

We will not see the dramatic increases in clock speeds in the future.

However, the number of gates on a chip will continue to increase.

Intel Yonah will double the processing power on a per watt basis.

25



Change Is Coming

No Free Lunch For Traditional Software

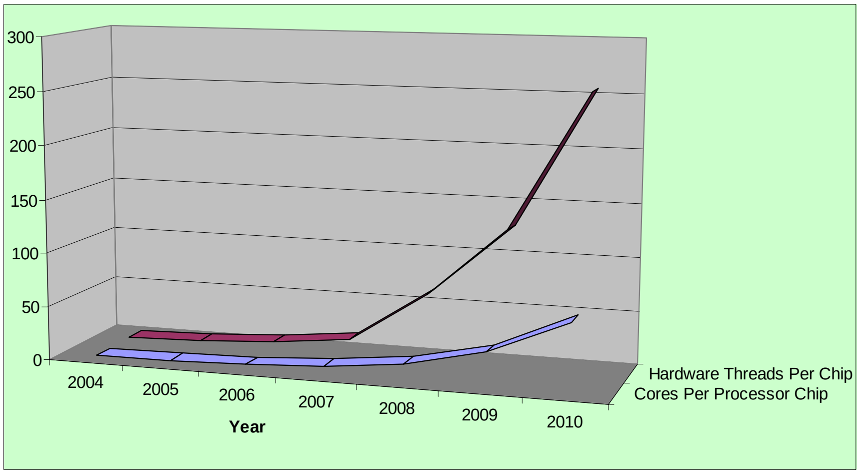
(Without highly concurrent software it won't get any faster!)

1 Core
2 Cores
4 Cores
8 Cores
24 GHz, 1 Core
12 GHz, 1 Core
6 GHz 1 Core
3GHz 1 Core
3 GHz 2 Cores
3 GHz, 4 Cores
3 GHz, 8 Cores
Operations per second for serial code
Free Lunch For Traditional Software
(It just runs twice as fast every 18 months with no change to the code!)
Additional operations per second if code can take advantage of concurrency
00

26

CPU Desktop Trends 2004-2010

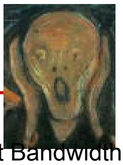
- Relative processing power will continue to double every 18 months
- 256 logical processors per chip in late 2010



27

Commodity Processor Trends

Bandwidth/Latency is the Critical Issue, not FLOPS

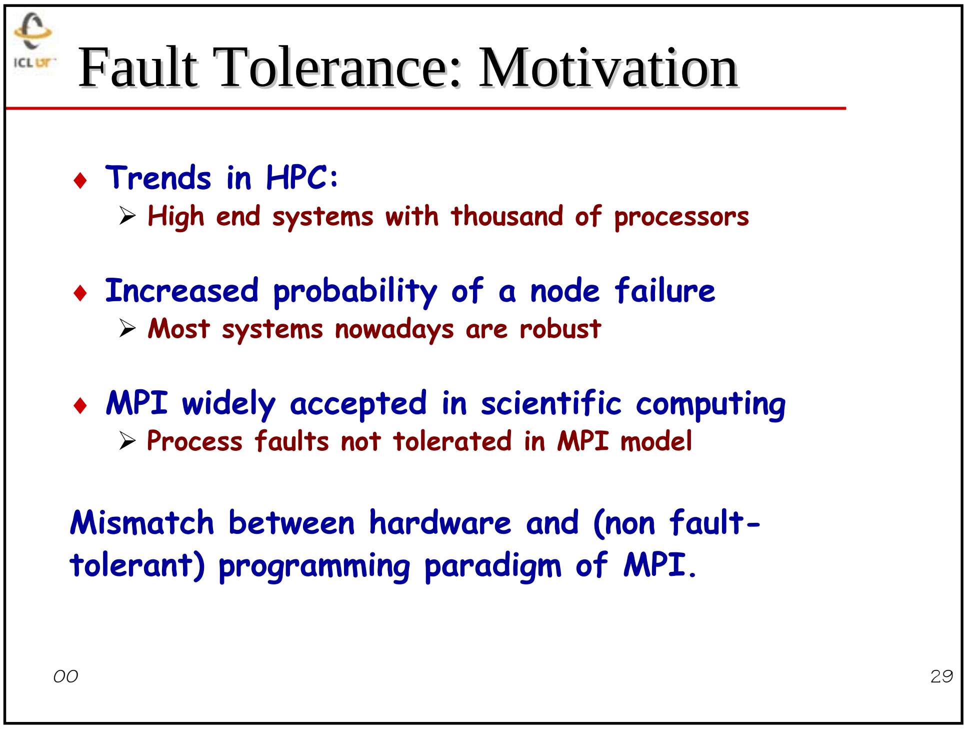


Got Bandwidth?

	Annual increase	Typical value in 2005
Single-chip floating-point performance	59%	4 GFLOP/s
Front-side bus bandwidth	23%	1 GWord/s = 0.25 word/flop
DRAM latency	(5.5%)	70 ns = 280 FP ops = 70 loads

00 Source: *Getting Up to Speed: The Future of Supercomputing*, National Research Council, 222 pages, 2004, National Academies Press, Washington DC, ISBN 0-309-09502-6.

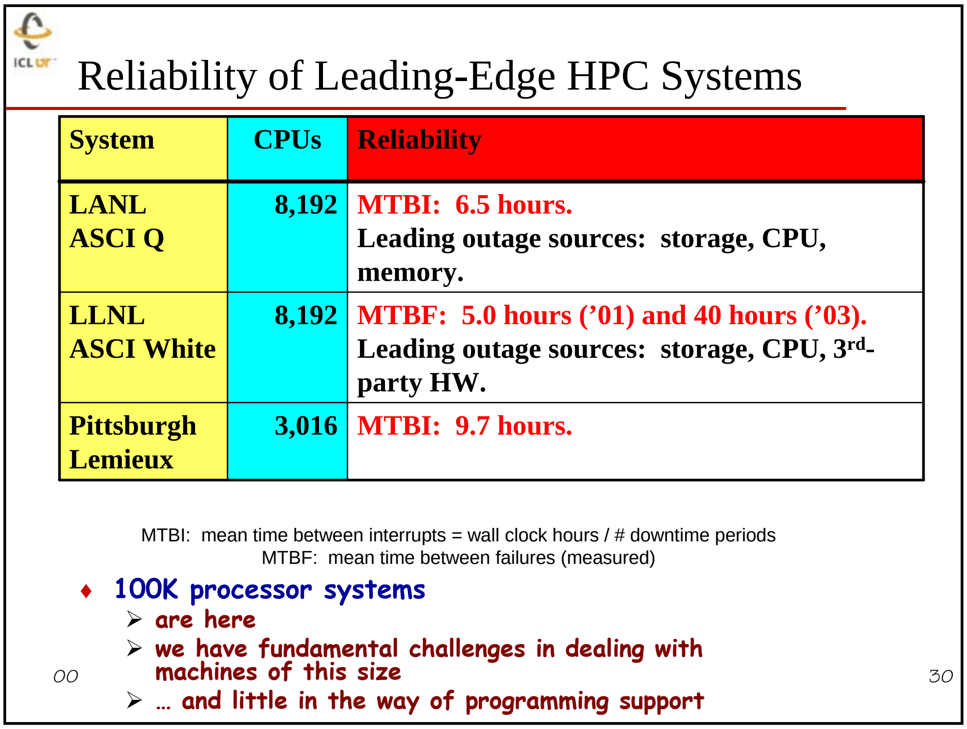
28



Fault Tolerance: Motivation

- Trends in HPC:
  - High end systems with thousand of processors
- Increased probability of a node failure
  - Most systems nowadays are robust
- MPI widely accepted in scientific computing
  - Process faults not tolerated in MPI model

Mismatch between hardware and (non fault-tolerant) programming paradigm of MPI.



Reliability of Leading-Edge HPC Systems

System	CPUs	Reliability
LANL ASCI Q	8,192	MTBI: 6.5 hours. Leading outage sources: storage, CPU, memory.
LLNL ASCI White	8,192	MTBF: 5.0 hours ('01) and 40 hours ('03). Leading outage sources: storage, CPU, 3rd-party HW.
Pittsburgh Lemieux	3,016	MTBI: 9.7 hours.

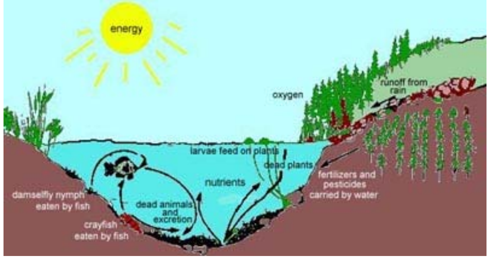
MTBI: mean time between interrupts = wall clock hours / # downtime periods
MTBF: mean time between failures (measured)

- 100K processor systems
  - are here
  - we have fundamental challenges in dealing with machines of this size
  - … and little in the way of programming support

Future Challenge: Developing the Ecosystem for HPC

From the NRC Report on “The Future of Supercomputing”:

- ♦ Hardware, software, algorithms, tools, networks, institutions, applications, and people who solve supercomputing applications can be thought of collectively as an ecosystem
- ♦ Research investment in HPC should be informed by the ecosystem point of view - progress must come on a broad front of interrelated technologies, rather than in the form of individual breakthroughs.


energy
oxygen
runoff from rain
larvae feed on plants
dead plants
nutrients
fertilizers and pesticides carried by water
damselfly nymph eaten by fish
dead animals and excretion
crayfish eaten by fish

A supercomputer ecosystem is a continuum of computing platforms, system software, algorithms, tools, networks, and the people who know how to exploit them to solve computational science applications.

Real Crisis With HPC Is With The Software

- ♦ **Our ability to configure a hardware system capable of 1 PetaFlop (10^{15} ops/s) is without question just a matter of time and $$.**
- ♦ **A supercomputer application and software are usually much more long-lived than a hardware**
  - ➢ **Hardware life typically five years at most…. Apps 20-30 years**
  - ➢ **Fortran and C are the main programming models (still!!)**
- ♦ **The REAL CHALLENGE is Software**
  - ➢ **Programming hasn’t changed since the 70’s**
  - ➢ **HUGE manpower investment**
    - ➢ **MPI… is that all there is?**
  - ➢ **Often requires HERO programming**
  - ➢ **Investments in the entire software stack is required (OS, libs, etc.)**
- ♦ **Software is a major cost component of modern technologies.**
  - ➢ **The tradition in HPC system procurement is to assume that the software is free… SOFTWARE COSTS (over and over)**

Summary of Current Unmet Needs

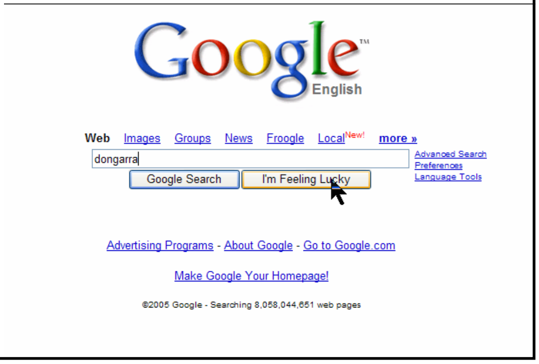
- Performance / Portability
- Fault tolerance
- Memory bandwidth/Latency
- Adaptability: Some degree of autonomy to self optimize, test, or monitor.
  - Able to change mode of operation: static or dynamic
- Better programming models
  - Global shared address space
  - Visible locality
- Maybe coming soon (incremental, yet offering real benefits):
  - Global Address Space (GAS) languages: UPC, Co-Array Fortran, Titanium, Chapel)
    - "Minor" extensions to existing languages
    - More convenient than MPI
    - Have performance transparency via explicit remote memory references

Collaborators / Support

- Top500 Team
  - Erich Strohmaier, NERSC
  - Hans Meuer, Mannheim
  - Horst Simon, NERSC

Office of Science
U.S. DEPARTMENT OF ENERGY

http://www.top500.org/



dongarra

Advertising Programs - About Google - Go to Google.com

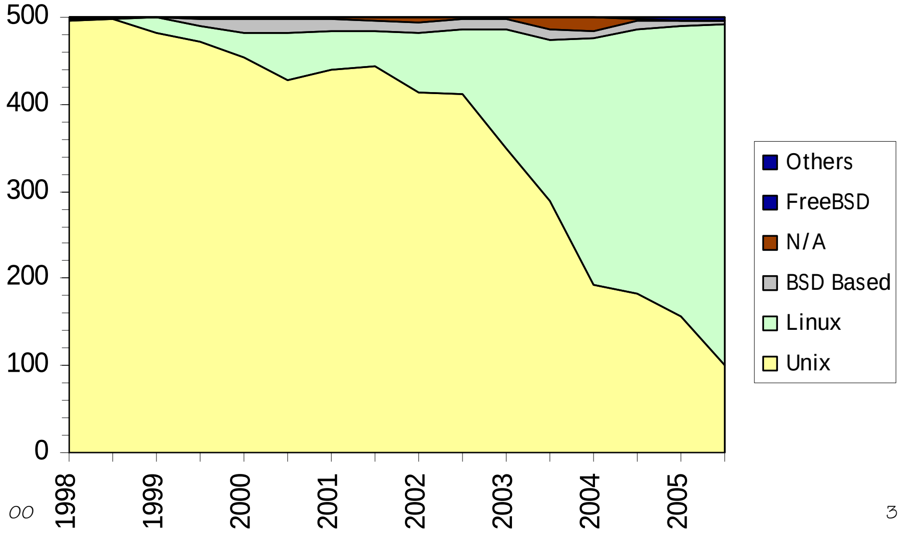
Make Google Your Homepage!

©2005 Google - Searching 8,058,044,651 web pages

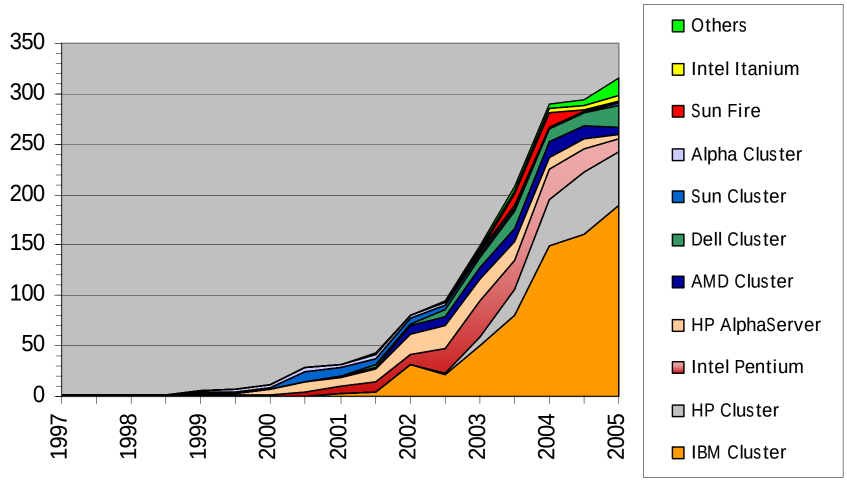
Next Steps

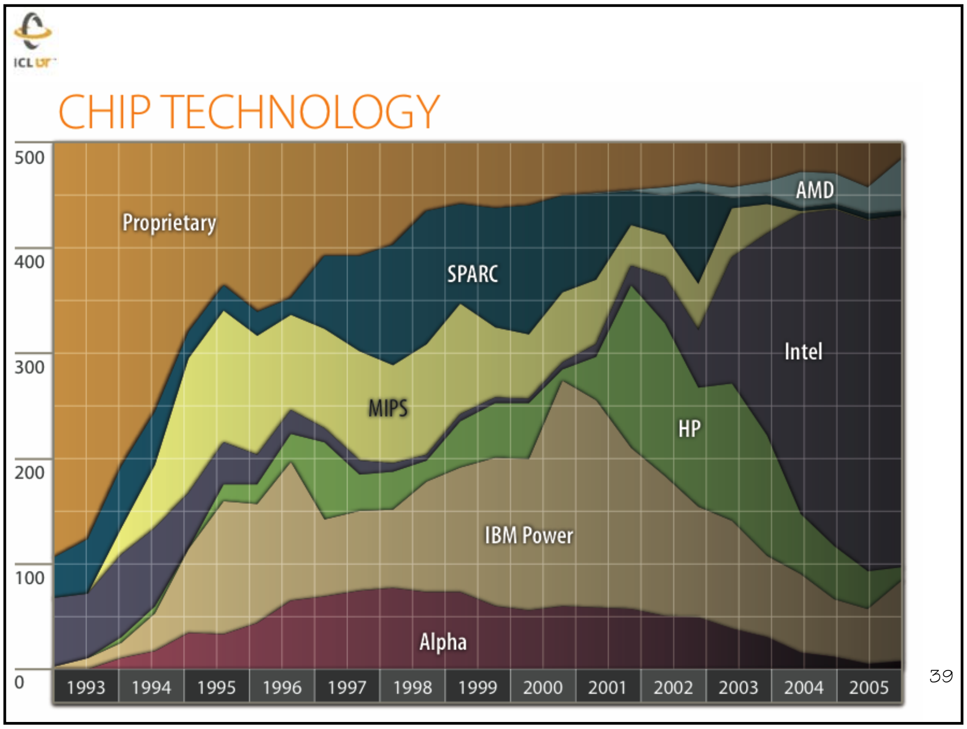
- Software to determine the checkpointing interval and number of checkpoint processors from the machine characteristics.
  - Perhaps use historical information.
  - Monitoring
  - Migration of task if potential problem
- Local checkpoint and restart algorithm.
  - Coordination of local checkpoints.
  - Processors hold backups of neighbors.
- Have the checkpoint processes participate in the computation and do data rearrangement when a failure occurs.
  - Use p processors for the computation and have k of them hold checkpoint.
- Generalize the ideas to provide a library of routines to do the diskless check pointing.
- Look at "real applications" and investigate "Lossy" algorithms.

Operating Systems / Systems



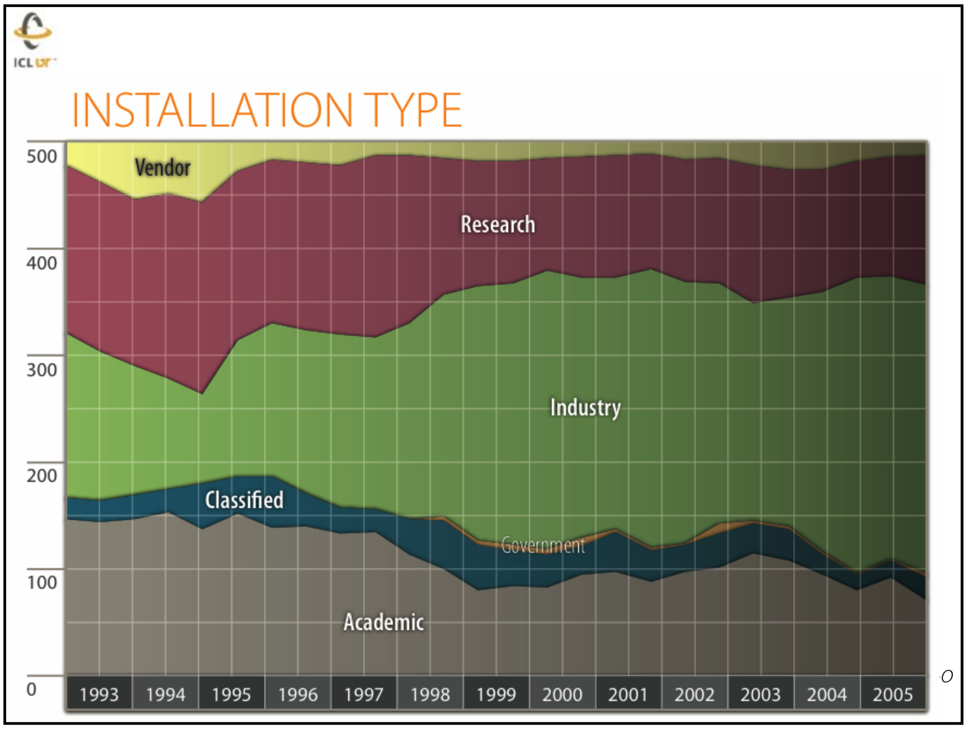
Clusters / Systems





CHIP TECHNOLOGY

39

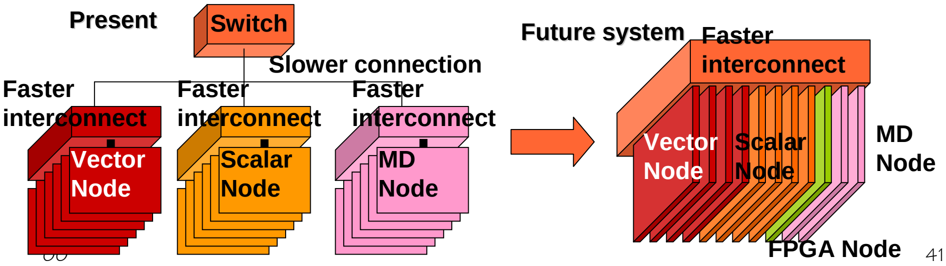


INSTALLATION TYPE

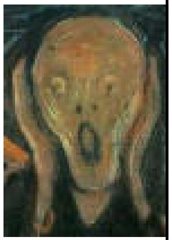
0

Japanese:
Tightly-Coupled Heterogeneous System

- Would like to get to 10 PetaFlop/s by 2011
- Scalable, fits any computer center
  - Size, cost, ratio of components
- Easy and low-cost to develop new component
- Scale merit of components

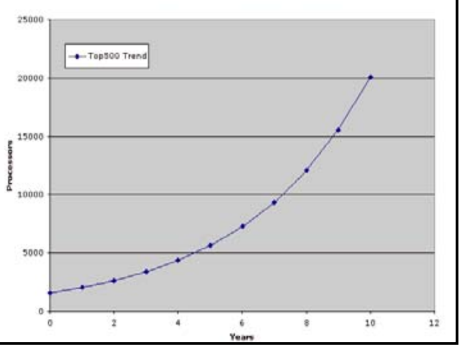


How Big Is Big?



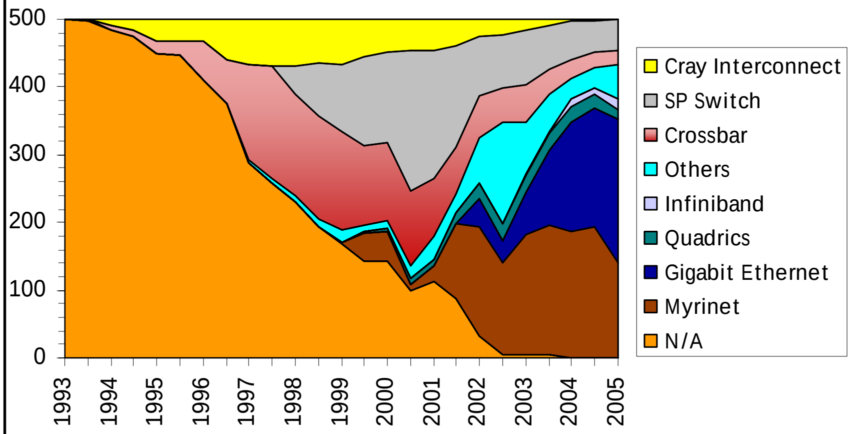
- Every 10X brings new challenges
  - 64 processors was once considered large
    - it hasn't been "large" for quite a while
  - 1024 processors is today's "medium" size
  - 8096 processors is today's "large"
    - we're struggling even here

2004-2014 System Size Trends



- 100K processor systems
  - are in construction
  - we have fundamental challenges in dealing with machines of this size
  - … and little in the way of programming support

Interconnects / Systems



24%

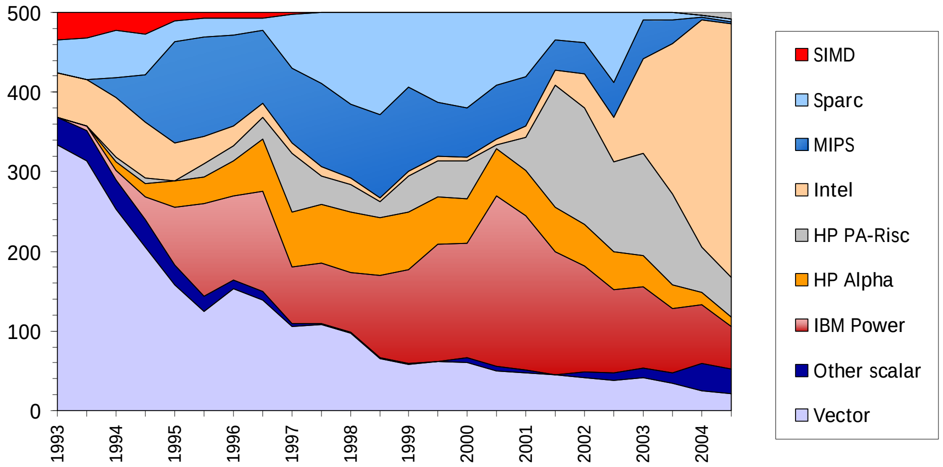
28%

Myrinet and GigE > 50% of market

Real Crisis With HPC Is With The Software

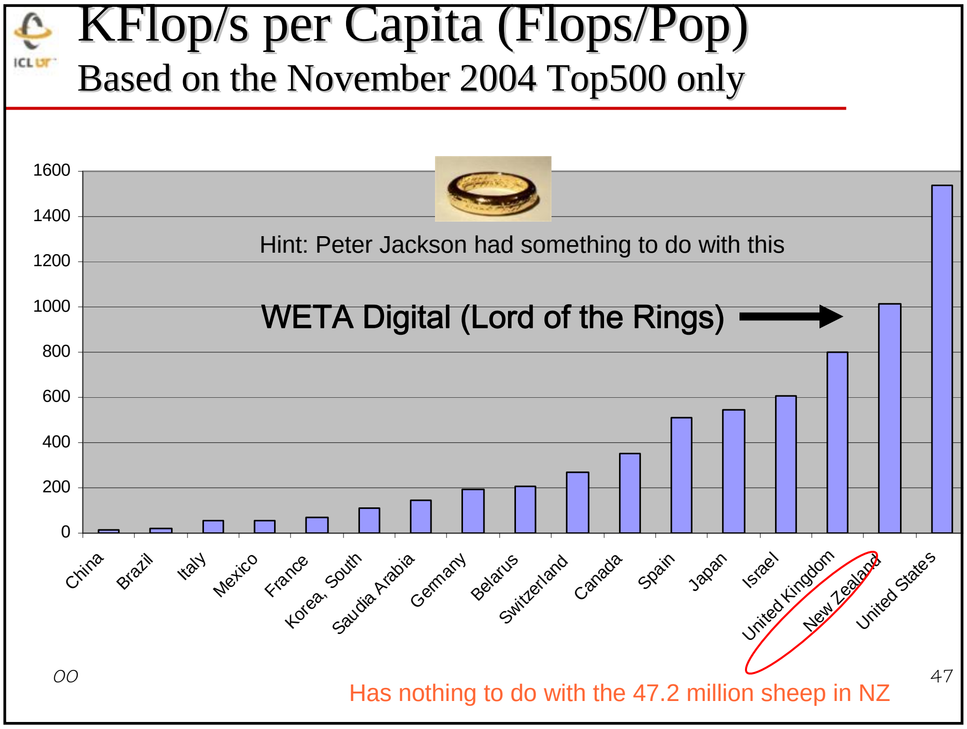
- Programming is stuck
  - Arguably hasn't changed since the 60's
- It's time for a change
  - Complexity is rising dramatically
    - highly parallel and distributed systems
      - From 10 to 100 to 1000 to 10000 to 100000 of processors!!
    - multidisciplinary applications
- A supercomputer application and software are usually much more long-lived than a hardware
  - Hardware life typically five years at most.
  - Fortran and C are the main programming models
- Software is a major cost component of modern technologies.
  - The tradition in HPC system procurement is to assume that the software is free.
- We have too few ideas about how to solve this problem.

Processor Types



Today’s Processors

- pipelining (superscalar, OOO, VLIW, branch prediction, predication)
- simultaneous multithreading (SMT, Hyper-Threading, multi-core)
- SIMD vector instructions (VIS, MMX/SSE, AltiVec)
- caches and the memory hierarchy
- Intel added 36 instructions per year to IA-32, or 3 instructions per month!



KFlop/s per Capita (Flops/Pop)

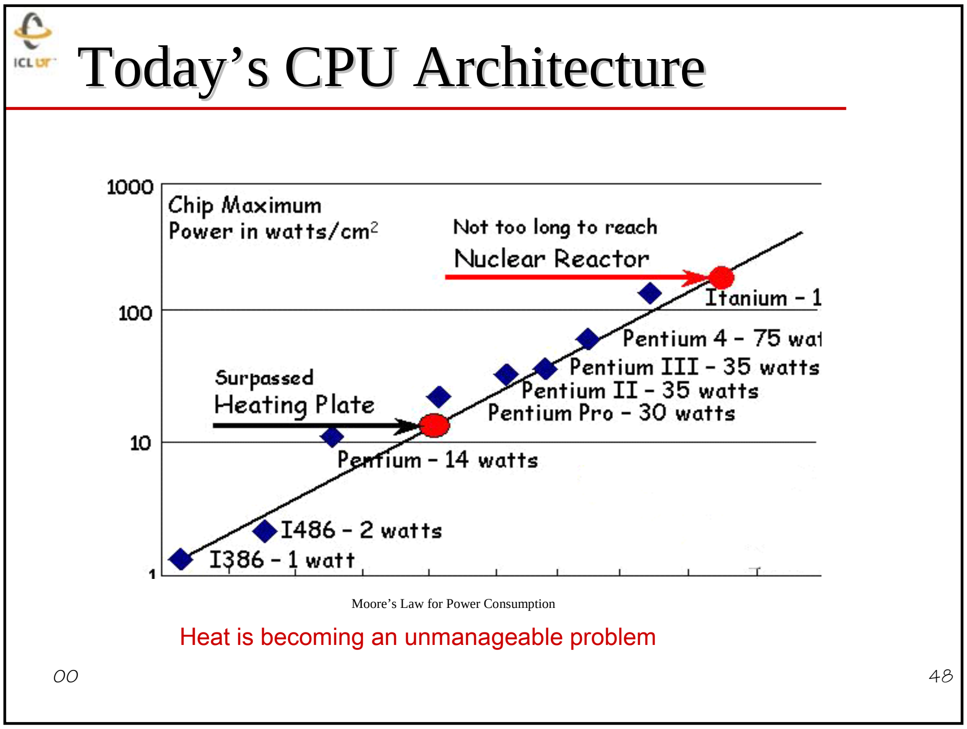
Based on the November 2004 Top500 only

Hint: Peter Jackson had something to do with this

WETA Digital (Lord of the Rings)

Has nothing to do with the 47.2 million sheep in NZ

00 47



Today’s CPU Architecture

Moore’s Law for Power Consumption

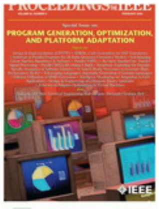
Heat is becoming an unmanageable problem

00 48

Self Adapting Numerical Software

- ♦ The process of arriving at an efficient solution involves many decisions by an expert.
  - ➢ Algorithm decisions
  - ➢ Data decisions
  - ➢ Management of the computing environment
  - ➢ Processor specific tuning

Proceedings of the IEEE,
V: 93 #: 2 Feb. 2005
Issue on Program
Generation,
Optimization, and
Platform Adaptation



Complex set of interaction between

Users' applications
Algorithm
Programming language
Compiler
Machine instruction
Hardware

Many layers of translation from the application to the hardware. Changing with each generation of hardware.