COMPUTER & COMPUTATIONAL SCIENCES

Los Alamos National Laboratory
RADIANT
www.lanl.gov/radiant

SUPERCOMPUTING in SMALL SPACES

GREEN DESTINY

A 240-Node Energy-Efficient Supercomputer in Five Square Feet

Wu-chun Feng

feng@lanl.gov

Research & Development in Advanced Network Technology (RADIANT)

Computer & Computational Sciences Division

University of California, Los Alamos National Laboratory

Full Disclosure: Orion Multisystems

IEEE Distinguished Visitors Program; Boise, ID, USA

October 15, 2004

LA-UR-04-7629

NNSA

Los Alamos NATIONAL LABORATORY

Outline

- Motivation
  - Where is Supercomputing?
- Supercomputing in Small Spaces (http://sss.lanl.gov)
  - From MetaBlade to Green Destiny
    - Experimental Results
  - The Evolution of Green Destiny
    - Architectural: Orion Multisystems DT-12
    - Software-Based: CAFfeine Supercomputer
- Past, Present, and Future
- Publications, Awards, Media, etc.
- Conclusion

COMPUTER & COMPUTATIONAL SCIENCES

Where is Supercomputing?

(Pictures: Thomas Sterling, Caltech & NASA JPL, and Wu Feng, LANL)

We have spent decades focusing on performance, performance, performance (and price/performance).

Where is Supercomputing? Top 500 Supercomputer List

- Benchmark
  - LINPACK: Solves a (random) dense system of linear equations in double-precision (64 bits) arithmetic.
    - Introduced by Prof. Jack Dongarra, U. Tennessee
- Evaluation Metric
  - Performance (i.e., Speed)
    - Floating-Operations Per Second (FLOPS)
- Web Site
  - http://www.top500.org
  - Current #1: Japanese Earth Simulator @ 35.9 TFLOPS

COMPUTER & COMPUTATIONAL SCIENCES

Where is Supercomputing? Gordon Bell Awards at SC

- Metrics for Evaluating Supercomputers (or HPC)
  - *Performance (i.e., Speed)*
    - Metric: Floating-Operations Per Second (FLOPS)
    - Example: Japanese Earth Simulator, ASCI Thunder & Q.
  - *Price/Performance → Cost Efficiency*
    - Metric: Acquisition Cost / FLOPS
    - Examples: LANL Space Simulator, VT System X cluster. (In general, Beowulf clusters.)
- Performance & price/performance are important metrics, but …

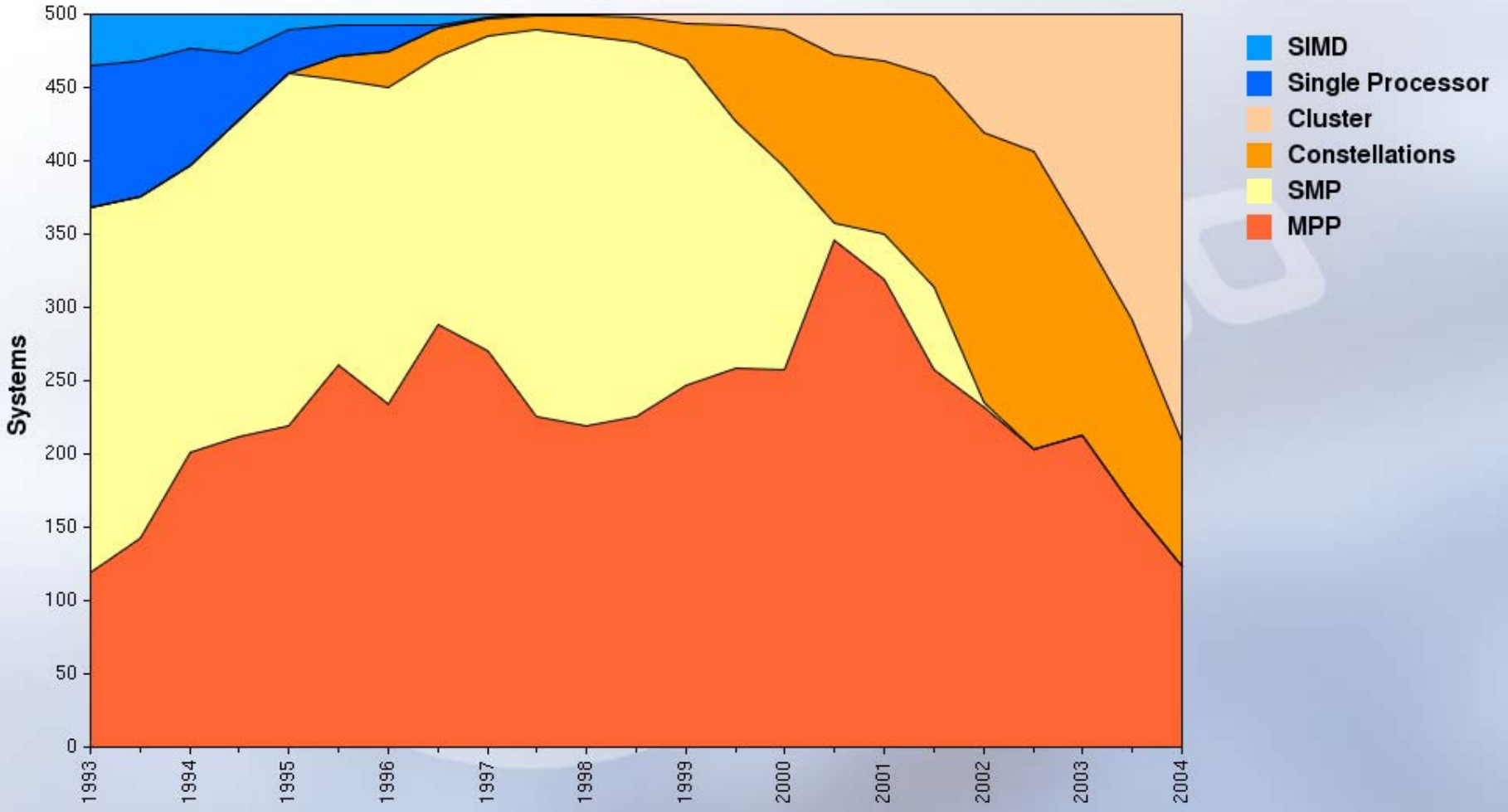
Where is Supercomputing? (Unfortunate) Assumptions

- Humans are infallible.
  - No mistakes made during integration, installation, configuration, maintenance, repair, or upgrade.
- Software will eventually be bug free.
- Hardware MTBF is already very large (~100 years between failures) and will continue to increase.
- Acquisition cost is what matters; maintenance costs are irrelevant.

- The above assumptions are even *more* problematic if one looks at current trends in supercomputing.

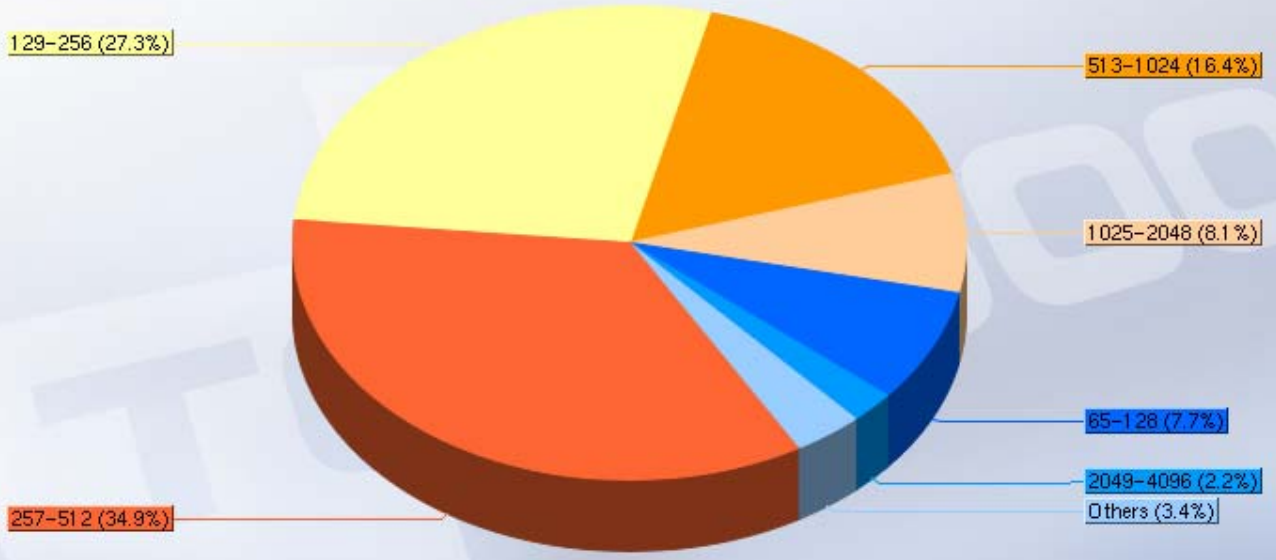
Adapted from David Patterson, UC-Berkeley

Architectures / Systems



TOP500 SUPERCOMPUTER SITES

System Processor Counts / Systems (Jun 2004)



COMPUTER & COMPUTATIONAL SCIENCES

Reliability & Availability of Leading-Edge Supercomputers

Systems	CPUs	Reliability & Availability
ASCI Q	8,192	**MTBI: 6.5 hrs.** 114 unplanned outages/month. ◆ HW outage sources: storage, CPU, memory.
ASCI White	8,192	**MTBF: 5 hrs. (2001) and 40 hrs. (2003).** ◆ HW outage sources: storage, CPU, 3rd-party HW.
NERSC Seaborg	6,656	**MTBI: 14 days. MTTR: 3.3 hrs.** ◆ SW is the main outage source. **Availability: 98.74%.**
PSC Lemieux	3,016	**MTBI: 9.7 hrs.** **Availability: 98.33%.**
Google	~15,000	**20 reboots/day; 2-3% machines replaced/year.** ◆ HW outage sources: storage, memory. **Availability: ~100%.**

MTBI: mean time between interrupts; MTBF: mean time between failures; MTTR: mean time to restore

Efficiency of Leading-Edge Supercomputers

- "Performance" and "Price/Performance" Metrics …
  - Lower efficiency, reliability, and availability.
  - Higher operational costs, e.g., admin, maintenance, etc.
- Examples
  - Computational Efficiency
    - Relative to Peak: Actual Performance/Peak Performance
    - Relative to Space: Performance/Sq. Ft.
    - Relative to Power: Performance/Watt
  - Performance: 2000-fold increase (since the Cray C90).
    - Performance/Sq. Ft.: Only 65-fold increase.
    - Performance/Watt: Only 300-fold increase.
  - Massive construction and operational costs associated with powering and cooling.

COMPUTER & COMPUTATIONAL SCIENCES

Another Perspective: "Commodity-Use Supercomputers"

- Requirement: Near-100% *availability* with *efficient* and *reliable* resource usage.
  - E-commerce, enterprise apps, online services, ISPs.

- Problems

Source: David Patterson, UC-Berkeley

  - Frequency of Service Outages
    - 65% of IT managers report that their websites were unavailable to customers over a 6-month period.
  - Cost of Service Outages
    - NYC stockbroker: **$ 6,500,000/hr**
    - Ebay (22 hours): **$ 225,000/hr**
    - Amazon.com: **$ 180,000/hr**
    - Social Effects: negative press, loss of customers who "click over" to competitor (e.g., TeraGrid, Akamai/Google)

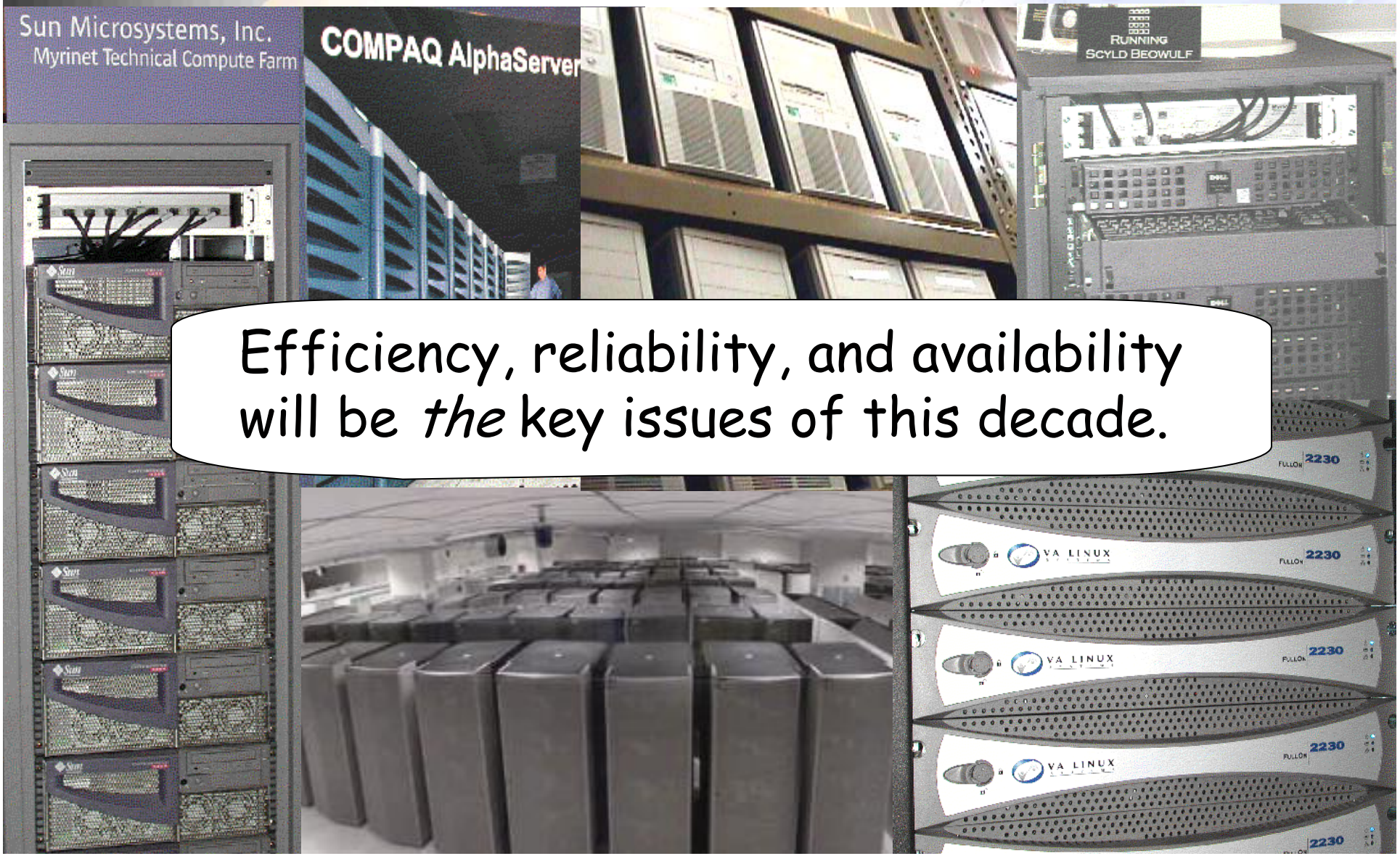
Another Perspective: "Commodity-Use Supercomputers"

- Pharmaceutical, financial, actuarial, retail, aerospace, automotive, science and engineering, data centers.
- Sampling of Consumer Requirements of HPC Systems
  - Myself, LANL (high-performance network simulations)
    - Traditional cluster fails weekly, oftentimes more frequently.
    - [1] Reliability, [2] Space, [3] Performance.
  - Peter Bradley, Pratt & Whitney (CFD, composite modeling)
    - [1] Reliability, [2] Transparency, [3] Resource Management
  - Eric Schmidt, Google (instantaneous search)
    - Low power, NOT speed.
    - DRAM density, NOT speed.
    - Availability and reliability, NOT speed.

COMPUTER & COMPUTATIONAL SCIENCES

Where is Supercomputing?

(Pictures: Thomas Sterling, Caltech & NASA JPL and Wu Feng, LANL)



Efficiency, reliability, and availability will be *the* key issues of this decade.

COMPUTER & COMPUTATIONAL SCIENCES

Where is Supercomputing? Supercomputing in Our Lab

- Operating Environment
  - 85-90°F warehouse at 7,400 feet above sea level.
  - No air conditioning, no air filtration, no raised floor, and no humidifier/dehumidifier.
- Computing Requirement
  - Parallel computer to enable high-performance network research in simulation and implementation.
- Old Solution: Traditional Supercomputing Cluster
  - 100-processor cluster computer that failed weekly in the above operating environment.
- New Solution: Low-Power, Reliable Supercomputing Cluster
  - A 240-processor cluster in five square feet → Green Destiny
- How did we come to the above solution?

Outline

- Motivation
  - Where is Supercomputing?
- Supercomputing in Small Spaces
  - From MetaBlade to Green Destiny
    - Experimental Results
  - The Evolution of Green Destiny
    - Architectural: Orion Multisystems DT-12
    - Software-Based: CAFfeine Supercomputer
- Past, Present, and Future
- Publications, Awards, Media, etc.
- Conclusion

Supercomputing in Small Spaces
http://sss.lanl.gov

- Goal
  - Improve efficiency, reliability, and availability (ERA) in large-scale computing systems.
    - Sacrifice a little bit of raw performance.
    - Improve overall system throughput as the system will "always" be available, i.e., effectively no downtime, no HW failures, etc.
  - Reduce the total cost of ownership (TCO). Another talk …
- Crude Analogy
  - Formula One Race Car: Wins raw performance but reliability is so poor that it requires frequent maintenance. Throughput low.
  - Honda S2000: Loses raw performance but high reliability results in high throughput (i.e., miles driven → answers/month).

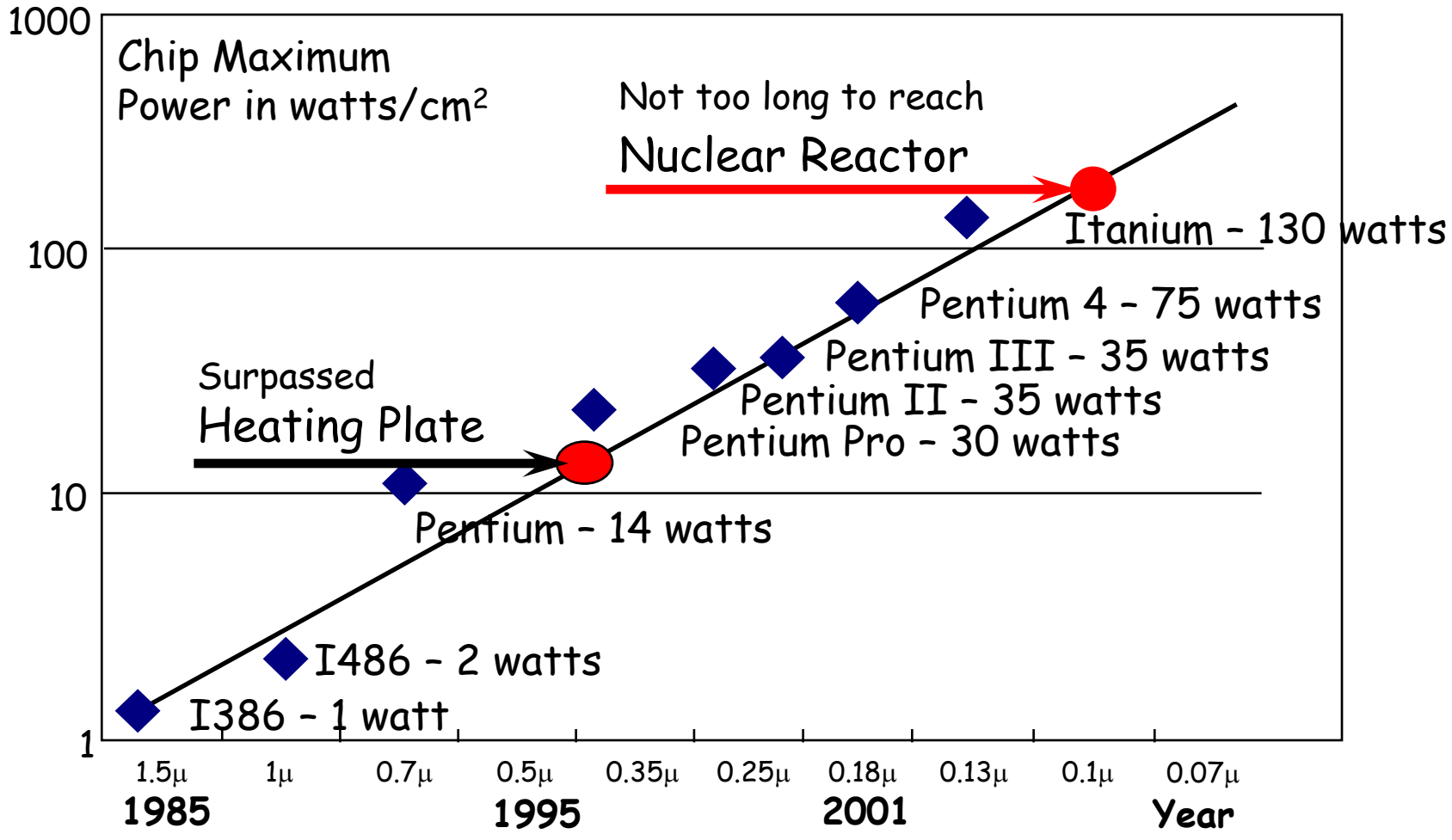
COMPUTER & COMPUTATIONAL SCIENCES

How to Improve Efficiency, Reliability & Availability?

- Complementary Approaches
  - Via HW design & manufacturing (e.g., IBM Blue Gene)
  - Via a software reliability layer that assumes underlying hardware unreliability *a la* the Internet (e.g., Google).
  - Via systems design & integration (e.g., Green Destiny)
- Observation
  - High power density α high temperature α low reliability
  - Arrhenius' Equation*
    (circa 1890s in chemistry → circa 1980s in computer & defense industries)
    - As temperature increases by 10° C …
      - The failure rate of a system doubles.
    - Twenty years of unpublished empirical data .

* The time to failure is a function of $e^{-Ea/kT}$ where Ea = activation energy of the failure mechanism being accelerated, k = Boltzmann's constant, and T = absolute temperature

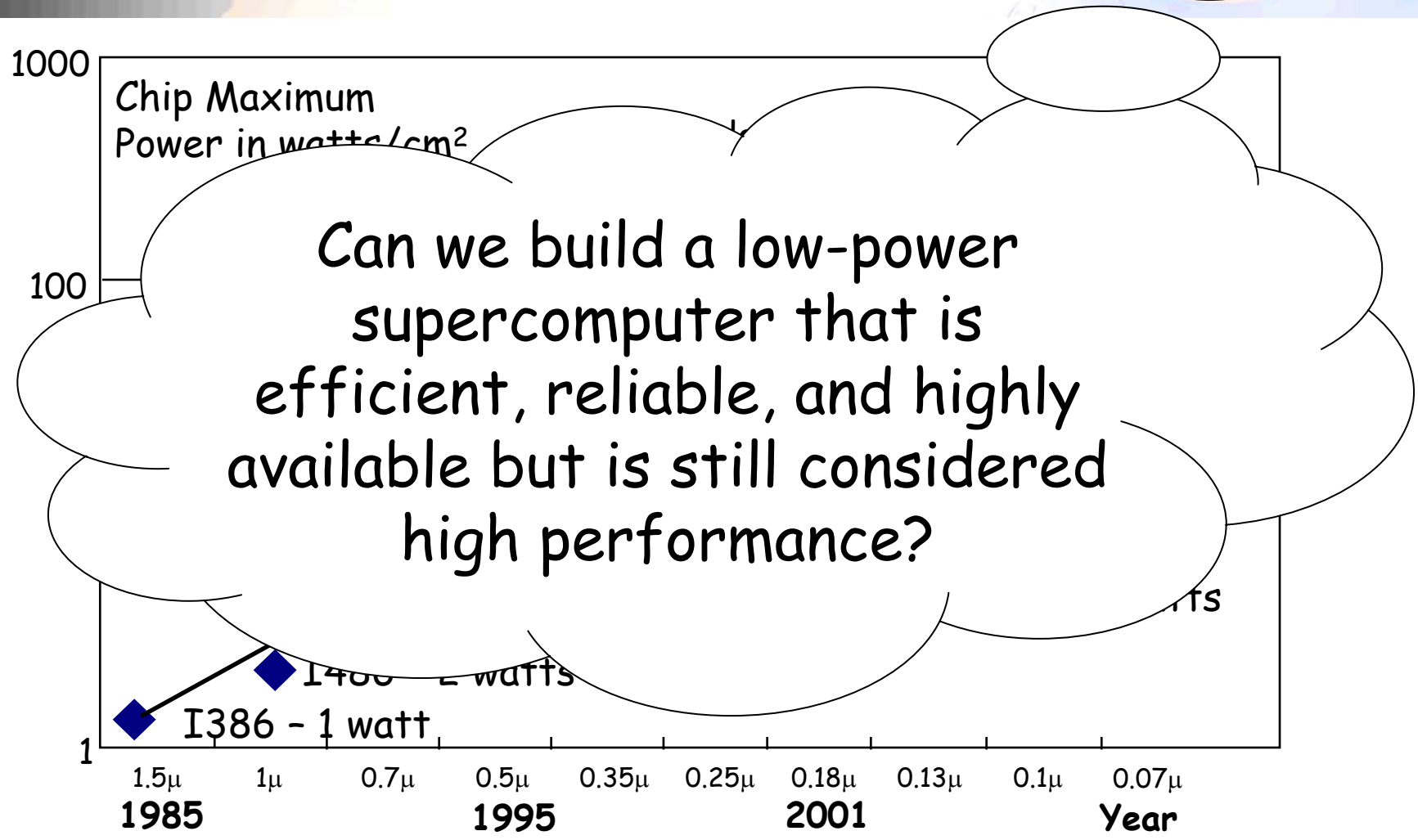
Moore's Law for Power



Source: Fred Pollack, Intel. New Microprocessor Challenges in the Coming Generations of CMOS Technologies, MICRO32 and Transmeta

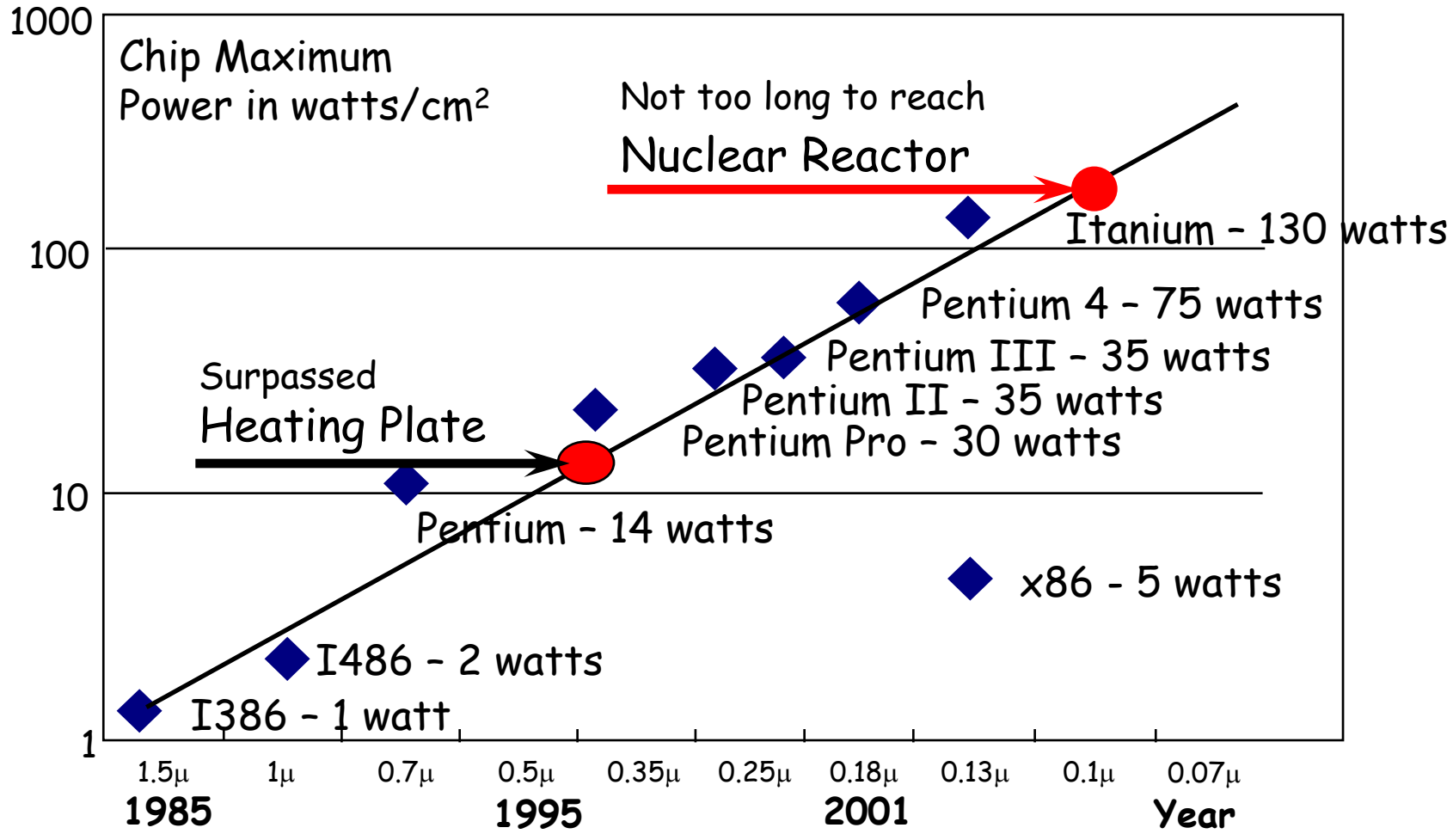
Wu Feng
feng@lanl.gov

Moore's Law for Power



Source: Fred Pollack, Intel. New Microprocessor Challenges in the Coming Generations of CMOS Technologies, MICRO32 and Transmeta

Moore's Law for Power



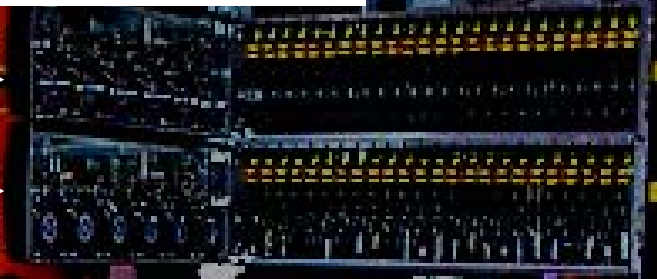
Source: Fred Pollack, Intel. New Microprocessor Challenges in the Coming Generations of CMOS Technologies, MICRO32 and Transmeta

COMPUTER & COMPUTATIONAL SCIENCES

MetaBlade: The Origin of Green Destiny

- Project Conception: Sept. 28, 2001.
  - On a winding drive home through Los Alamos Canyon … the need for *reliable* compute cycles.
    - Leverage RLX web-hosting servers with Transmeta CPUs.
- Project Implementation: Oct. 9, 2001.
  - Received the "bare" hardware components.
  - Two man-hours later …
    - Completed construction of a 24-CPU RLX System 324 (dubbed *MetaBlade*) and installation of system software.
  - One man-hour later …
    - Successfully executing a 10-million N-body simulation of a galaxy formation
- Public Demonstration: Nov. 12, 2001 at SC 2001.

SC 2001: The First Bladed Beowulf



MetaBlade: 24 ServerBlade 633s

MetaBlade2: 24 ServerBlade 800s

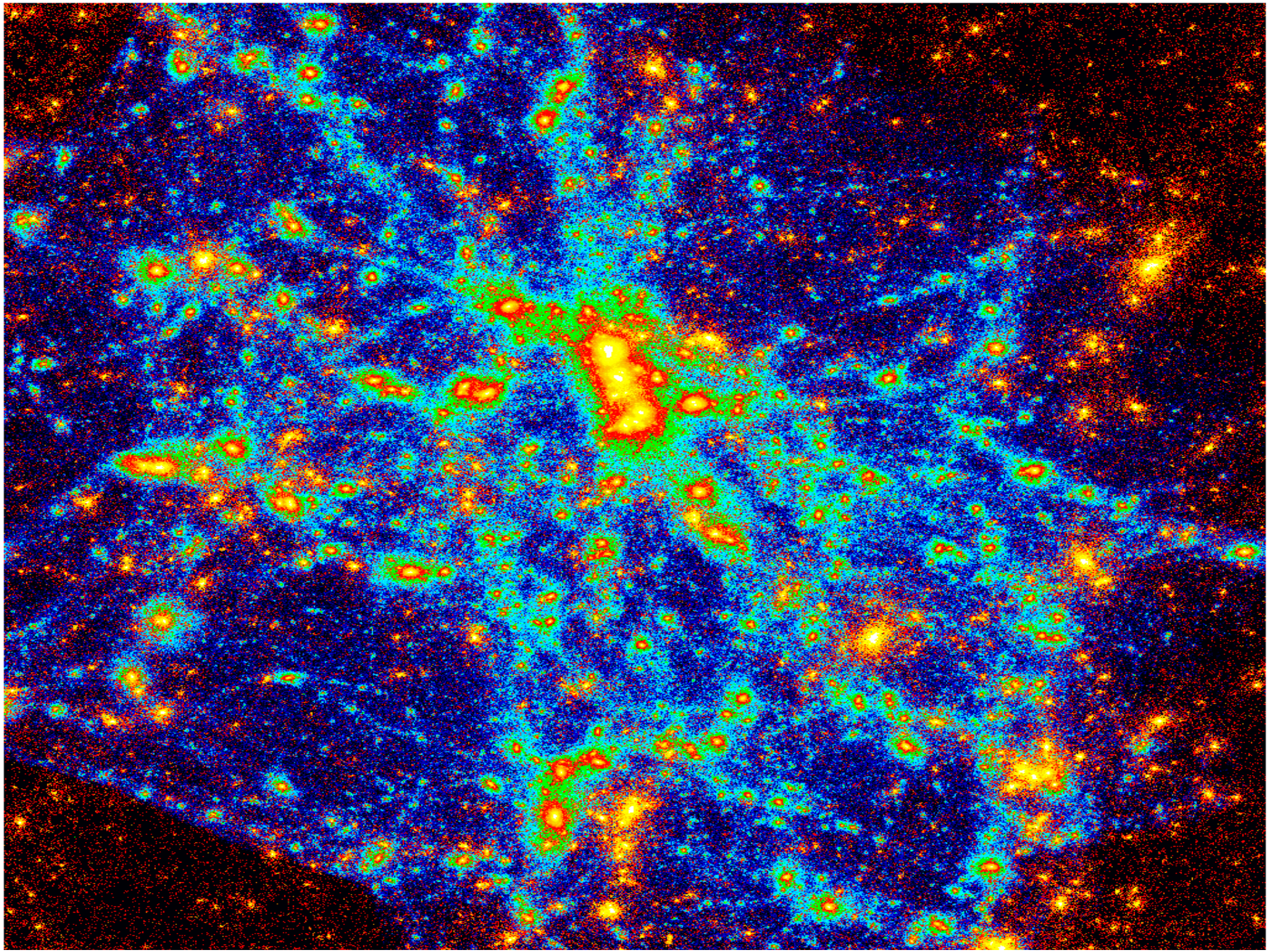
(On-loan from RLX for SC 2001)

- MetaBlade Node
  - 633-MHz Transmeta TM5600
  - 512-KB cache, 256-MB RAM
  - 100-MHz front-side bus
  - 3 x 100-Mb/s Ethernet
- MetaBlade2 Node
  - 800-MHz Transmeta TM5800
  - 512-KB cache, 384-MB RAM (128-MB on-board DDR + 256-MB SDR DIMM)
  - 133-MHz front-side bus
  - 3 x 100-Mb/s Ethernet

Performance of an N-body Simulation of Galaxy Formation

- MetaBlade: 2.1 Gflops; MetaBlade2: 3.3 Gflops

No failures in its lifetime despite no cooling facilities.

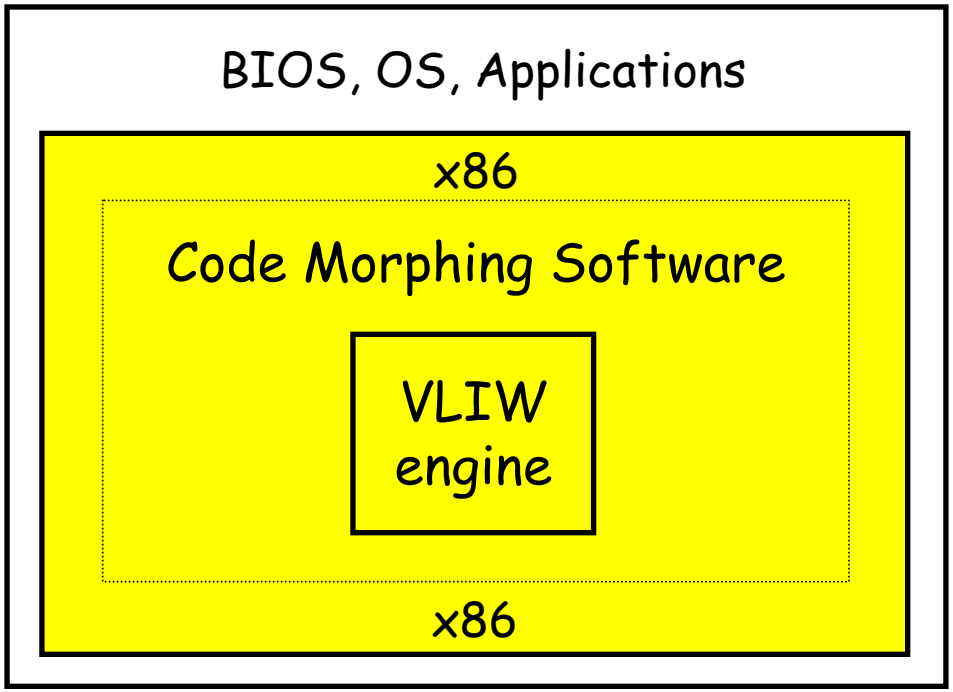


Scaling *MetaBlade* …

- Interest in MetaBlade and MetaBlade2 ?
  - Continual crowds over the three days of SC 2001.
- Inspiration
  - Build a full 42U rack of MetaBlade clusters.
    - Scale up performance/space to 3500 Mflop/sq. ft.
  - Problem: In 2001, performance per node on MetaBlade was *three* times worse than the fastest processor at the time (i.e., 2.0-GHz Intel P4).
  - Can we improve performance while maintaining low power? Yes via Transmeta's code-morphing software, which is part of the Transmeta CPU.
    - What is code-morphing software?

Transmeta TM5600 CPU: VLIW + CMS

- VLIW Engine
  - Up to four-way issue
    - In-order execution only.
  - Two integer units
  - Floating-point unit
  - Memory unit
  - Branch unit



- VLIW Transistor Count ("Anti-Moore's Law")
  - ~$\frac{1}{4}$ of Intel PIII → ~ 7x less power consumption
  - Less power → lower "on-die" temp. → better reliability & availability

Transmeta TM5x00 CMS

- Code-Morphing Software (CMS)
  - Provides compatibility by dynamically "morphing" x86 instructions into simple VLIW instructions.
  - Learns and improves with time, i.e., iterative execution.

- High-Performance Code-Morphing Software (HP-CMS)
  - Results (circa 2001)
    - *Optimized to improve floating-pt. performance by 50%.*
    - *1-GHz Transmeta performs as well as a 1.2-GHz PIII-M.*
  - How?

Transmeta TM5x00 Comparison

Intel P4	MEM	MEM	2xALU	2xALU	FPU	SSE	SSE	Br
Transmeta TM5x00	MEM		2xALU		FPU			Br

- Previous-generation Transmeta TM5800 + HP-CMS
  - Performs better than an Intel PIII over iterative scientific codes on a clock-for-clock-cycle basis.
  - Performs only *twice* as slow as the fastest CPU (at the time) rather than three times as slow.
- Efficeon, the current-generation CPU from Transmeta, rectifies the above mismatch in functional units.

COMPUTER & COMPUTATIONAL SCIENCES

Low-Power Network Switches

- WWP LE-410: 16 ports of Gigabit Ethernet
- WWP LE-210: 24 ports of Fast Ethernet via RJ-21s
- (Avg.) Power Dissipation / Port: A few watts.

Wu Feng
feng@lanl.gov

http://www.lanl.gov/radiant
http://sss.lanl.gov

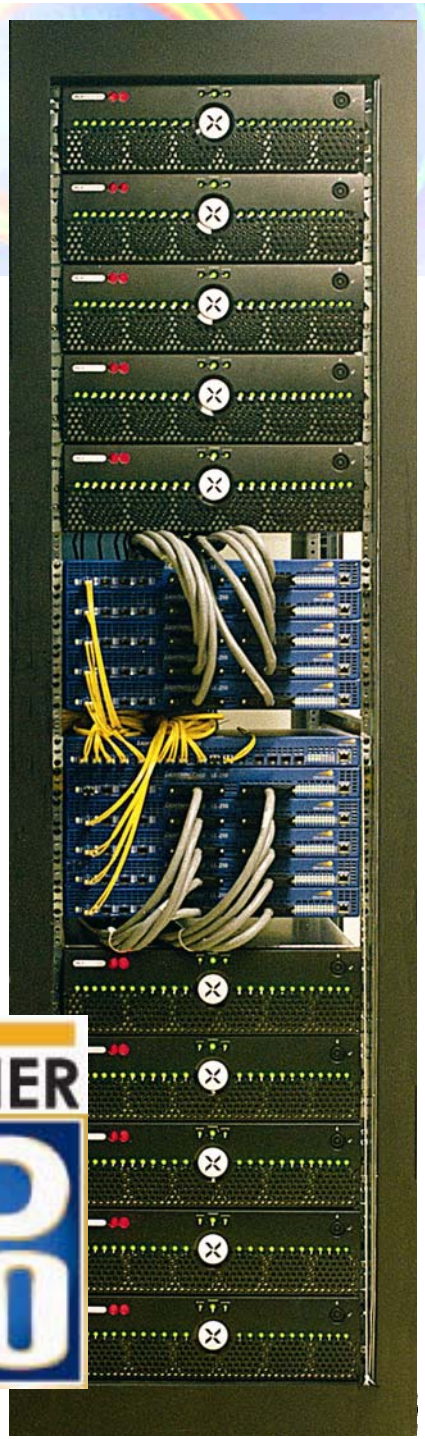
Los Alamos NATIONAL LABORATORY

COMPUTER & COMPUTATIONAL

"Green Destiny" Bladed Beowulf

(circa April 2002)

- A 240-Node Beowulf in Five Square Feet
- Each Node
  - 667-MHz Transmeta TM5600 CPU w/ Linux 2.4.x
    - Upgraded to 1-GHz Transmeta TM5800 CPUs
  - 640-MB RAM, 20-GB hard disk, 100-Mb/s Ethernet (up to 3 interfaces)
- Total
  - 160 Gflops peak (240 Gflops with upgrade)
    - LINPACK: 101 Gflops in March 2003.
  - 150 GB of RAM (expandable to 276 GB)
  - 4.8 TB of storage (expandable to 38.4 TB)
  - Power Consumption: Only 3.2 kW.
- Reliability & Availability
  - *No unscheduled failures in 24 months.*



2003 WINNER R&D 100

Wu Feng
feng@lanl.gov

http://www.lanl.gov/radiant
http://sss.lanl.gov

COMPUTER & COMPUTATIONAL SCIENCES

Experimental Results

NNSA

Los Alamos NATIONAL LABORATORY

Gravitational Microkernel on Transmeta CPUs

- Gravitational Microkernel Benchmark (circa June 2002)

Processor	Math `sqrt`	Karp `sqrt`
500-MHz Intel PIII	87.6	137.5
533-MHz Compaq Alpha EV56	76.2	178.5
633-MHz Transmeta TM5600	115.0	144.6
800-MHz Transmeta TM5800	174.1	296.6
375-MHz IBM Power3	298.5	379.1
1200-MHz AMD Athlon MP	350.7	452.5

Units are in Mflops.

Source: Michael S. Warren, Theoretical Astrophysics Group at Los Alamos National Laboratory.

COMPUTER & COMPUTATIONAL SCIENCES

Treecode Benchmark within n-Body Galaxy Formation

Year	Site	Machine	CPUs	Gflops	Mflops/CPU
2003	LANL	ASCI QB	3600	2793	775.8
2003	LANL	Space Simulator	288	179.7	623.9
2002	NERSC	IBM SP-3	256	57.70	225.0
2000	LANL	SGI O2K	64	13.10	205.0
2002	LANL	Green Destiny	212	38.90	183.5
2001	SC'01	MetaBlade2	24	3.30	138.0
1998	LANL	Avalon	128	16.16	126.0
1996	LANL	Loki	16	1.28	80.0
1996	SC '96	Loki+Hyglac	32	2.19	68.4
1996	Sandia	ASCI Red	6800	464.90	68.4
1995	JPL	Cray T3D	256	7.94	31.0

Source: Michael S. Warren, Theoretical Astrophysics Group at Los Alamos National Laboratory.

COMPUTER & COMPUTATIONAL SCIENCES

Treecode Benchmark within n-Body Galaxy Formation

Year	Site	Machine	CPUs	Gflops	Mflops/CPU
2003	LANL	ASCI QB	3600	2793	775.8
2003	LANL	Space Simulator	288	179.7	623.9
2002	NERSC	IBM SP-3	256	57.70	225.0
2000	LANL	SGI O2K	64	13.10	205.0
2002	LANL	Green Destiny	212	38.90	183.5
2001	SC'01	Meta	24	3.30	138.0
1998	LANL				126.0
19					0
19					.4
1996	Sandia	ASCI Red	6800	464.90	68.4
1995	JPL	Cray T3D	256	7.94	31.0

Upgraded "Green Destiny" (Dec. 2002)
58 Gflops → **274 Mflops/CPU**

Source: Michael S. Warren, Theoretical Astrophysics Group at Los Alamos National Laboratory.

Wu Feng
feng@lanl.gov

Los Alamos NATIONAL LABORATORY


COMPUTER & COMPUTATIONAL
SCIENCES
Experimental Results
(relative to efficiency, reliability, availability)
NNSA
Los Alamos
NATIONAL LABORATORY

Performance Metrics for …

- Efficiency, Reliability, and Availability (ERA)
  - Total Cost of Ownership. Another talk …
  - Computational Efficiency
    - Relative to Space: Performance/Sq. Ft.
    - Relative to Power: Performance/Watt.
  - Reliability
    - MTBF: Mean Time Between Failures.
  - Availability
    - Percentage of time that resources are available for HPC.

- Continue to use *n-body galaxy formation* application as benchmark.

Parallel Computing Platforms ("Apples-to-Oranges" Comparison)

- Avalon (1996)
  - 140-CPU *Traditional Beowulf Cluster*
- ASCI Red (1996)
  - 9632-CPU *MPP*
- ASCI White (2000)
  - 512-Node (8192-CPU) *Cluster of SMPs*
- Green Destiny (2002)
  - 240-CPU *Bladed Beowulf Cluster*

COMPUTER & COMPUTATIONAL SCIENCES

Parallel Computing Platforms Running the N-body Code

Machine	Avalon Beowulf	ASCI Red	ASCI White	Green Destiny
Year	1996	1996	2000	2002
Performance (Gflops)	18	600	2500	39
Area (ft^2)	120	1600	9920	6
Power (kW)	18	1200	2000	5
DRAM (GB)	36	585	6200	150
Disk (TB)	0.4	2.0	160.0	4.8
DRAM density (MB/ft^2)	300	366	625	25000
Disk density (GB/ft^2)	3.3	1.3	16.1	800.0
Perf/Space ($Mflops/ft^2$)	150	375	252	6500
Perf/Power (Mflops/watt)	1.0	0.5	1.3	7.5

COMPUTER & COMPUTATIONAL SCIENCES

Parallel Computing Platforms Running the N-body Code

Machine	Avalon Beowulf	ASCI Red	ASCI White	Green Destiny
Year	1996	1996	2000	2002
Performance (Gflops)	18	600	2500	39
Area (ft^2)	120	1600	9920	6
Power (kW)	18	1200	2000	5
DRAM (GB)	36	585	6200	150
Disk (TB)	0.4	2.0	160.0	4.8
DRAM density (MB/ft^2)	300	366	625	25000
Disk density (GB/ft^2)	3.3	1.3	16.1	800.0
Perf/Space ($Mflops/ft^2$)	150	375	252	6500
Perf/Power (Mflops/watt)	1.0	0.5	1.3	7.5

Wu Feng
feng@lanl.gov

COMPUTER & COMPUTATIONAL SCIENCES

Parallel Computing Platforms Running the N-body Code

Machine	Avalon Beowulf	ASCI Red	ASCI White	Green Destiny+
Year	1996	1996	2000	2002
Performance (Gflops)	18	600	2500	*58*
Area (ft^2)	120	1600	9920	6
Power (kW)	18	1200	2000	5
DRAM (GB)	36	585	6200	150
Disk (TB)	0.4	2.0	160.0	4.8
DRAM density (MB/ft^2)	300	366	625	25000
Disk density (GB/ft^2)	3.3	1.3	16.1	800.0
Perf/Space (Mflops/ft^2)	150	375	252	9667
Perf/Power (Mflops/watt)	1.0	0.5	1.3	11.6

Wu Feng
feng@lanl.gov

Green Destiny vs. Earth Simulator: LINPACK

Machine	Green Destiny+	Earth Simulator
Year	2002	2002
LINPACK Performance (Gflops)	101	35,860
Area (ft^2)	6	17,222 * 2
Power (kW)	5	7,000
Cost efficiency ($/Mflop)	3.35	11.15
Perf/Space (Mflops/ft^2)	16,833	1,041
Perf/Power (Mflops/watt)	20.20	5.13

Disclaimer: This is not a fair comparison. Why?
(1) Price and the use of area and power do *not* scale linearly.
(2) Goals of the two machines are different.

Efficiency, Reliability, and Availability for …

- Green Destiny
  - Computational Efficiency
    - Relative to Space: Performance/Sq. Ft.

      Up to 60x better.
    - Relative to Power: Performance/Watt

      Up to 20x better.
  - Reliability
    - MTBF: Mean Time Between Failures

      "Infinite"
  - Availability
    - Percentage of time that resources are available for HPC.

      Nearly 100%.

Outline

- Motivation
  - Where is Supercomputing?
- Supercomputing in Small Spaces
  - From MetaBlade to Green Destiny
    - Experimental Results
  - The Evolution of Green Destiny
    - Architectural: Orion Multisystems DT-12
    - Software-Based: CAFfeine Supercomputer
- Past, Present, and Future
- Publications, Awards, Media, etc.
- Conclusion

COMPUTER & COMPUTATIONAL SCIENCES

http://www.orionmultisystems.com

- LINPACK Performance
  - 16.97 Gflops
- Footprint
  - 3 sq. ft.
  - 1 cu. ft.
- Power Consumption
  - < 150 watts at load
- How does this compare with a traditional desktop?

ORION DT-12 DESKTOP CLUSTER WORKSTATION

Imagine a 36 Gflop cluster **on your desk!**

12 Nodes
in a single computer

24 GBytes
memory capacity

36 Gflops
peak processing power

1 TByte
internal storage

DESIGNED FOR THE INDIVIDUAL

The Orion DT-12 cluster workstation is a fully integrated, completely self-contained, personal workstation based on the best of today's cluster technologies. Designed to be an affordable individual resource it is capable of 36 Gflops peak performance (18 Gflops sustained) with models starting at under $10k.

The Orion DT-12 cluster workstation provides supercomputer performance for the engineering, scientific, financial and creative professionals who need to solve computationally complex problems without waiting in the queue of the back-room cluster.

FASTER SOFTWARE DEVELOPMENT

The Orion DT-12 cluster workstation is the perfect platform for developers writing (and deploying) cluster software packages. It comes with cluster software development tools pre-installed, including libraries and a parallel compiler that allows you to spread one multiple-file compile to all the nodes in the system. Also included is a suite of system monitoring and management software.

NO ASSEMBLY REQUIRED

Orion workstations are designed from the ground up as a single computer. The entire system boots with the push of a button and has the ergonomics and ease of use of a personal computer. The modular, design allows for flexible configurations and scalability by stacking up to 4 systems as one 48 node cluster.

PRESERVE SOFTWARE INVESTMENTS

Orion workstations are built around industry standards for clustering: x86 processors, Ethernet, the Linux operating system and standard parallel programming libraries, including MPI, PVM and SGE. Existing Linux cluster applications run without modification.

PERFORMANCE AND FEATURES

The Orion DT-12 is a cluster of 12 x86-compatible nodes linked by a switched Gigabit Ethernet fabric. The cluster operates as a single computer with a single on-off switch and a single system image rapid boot sequence, which allows the entire system to boot in less than 90 seconds.

The Orion DT-12 cluster workstation is highly efficient, consuming a maximum of 220 Watts of power under peak load—about the same as an average desktop PC. It operates quietly, plugs into a standard 110V 15A wall socket and fits unobtrusively on a desk or lab bench.

COMPUTER & COMPUTATIONAL SCIENCES

http://www.orionmultisystems.com

ORION DS-96 DESKSIDE CLUSTER WORKSTATION

Imagine a 300 Gflop cluster...
under your desk.

96 Nodes
in a single computer

300 Gflops
peak processing power

192 GBytes
memory capacity

9.6 TBytes
internal storage

INCREASE YOUR PRODUCTIVITY

The Orion DS-96 cluster workstation is the highest performance general-purpose computing platform that can be plugged into a standard wall outlet and operated in an office or laboratory environment.

PRESERVE SOFTWARE INVESTMENTS

Orion workstations are built around industry standards for clustering: x86 processors, the Linux operating system and standard parallel programming libraries, including MPI, PVM and SGE. Your existing Linux cluster software applications can run without modification.

NO ASSEMBLY REQUIRED

Orion workstations are designed from the ground up as a single computer. The entire system boots with the push of a button and has the ergonomics and ease of use of a personal computer. Modular, solid state design allows for flexible configurations and scalability.

PERFORMANCE AND FEATURES

The Orion DS-96 cluster workstation is a fully integrated, completely self-contained personal workstation based on the best of today's cluster technologies and commodity components. Designed to be an individual or departmental resource, it is capable of 300 Gflops peak performance (150 Gflops sustained). The DS-96 is also highly efficient, consuming a maximum of 1500 Watts of power under peak load. It operates quietly, plugs into a standard 110V 15A wall socket, and fits unobtrusively beneath a desk or lab bench.

The DS-96 is a cluster of 96 x86-compatible nodes linked by an integrated Gigabit Ethernet fabric. The cluster operates as a single computer, with a single on-off switch, and a single-system-image rapid boot sequence which allows the entire system to boot in less than 2 minutes. The DS-96 comes with standard Linux and drivers pre-installed, including an optimized MPI message-passing library. Also included is a suite of cluster software development tools, system monitoring and system management software.

Wu Feng
feng@lanl.gov

http://www.lanl.gov/radiant
http://sss.lanl.gov

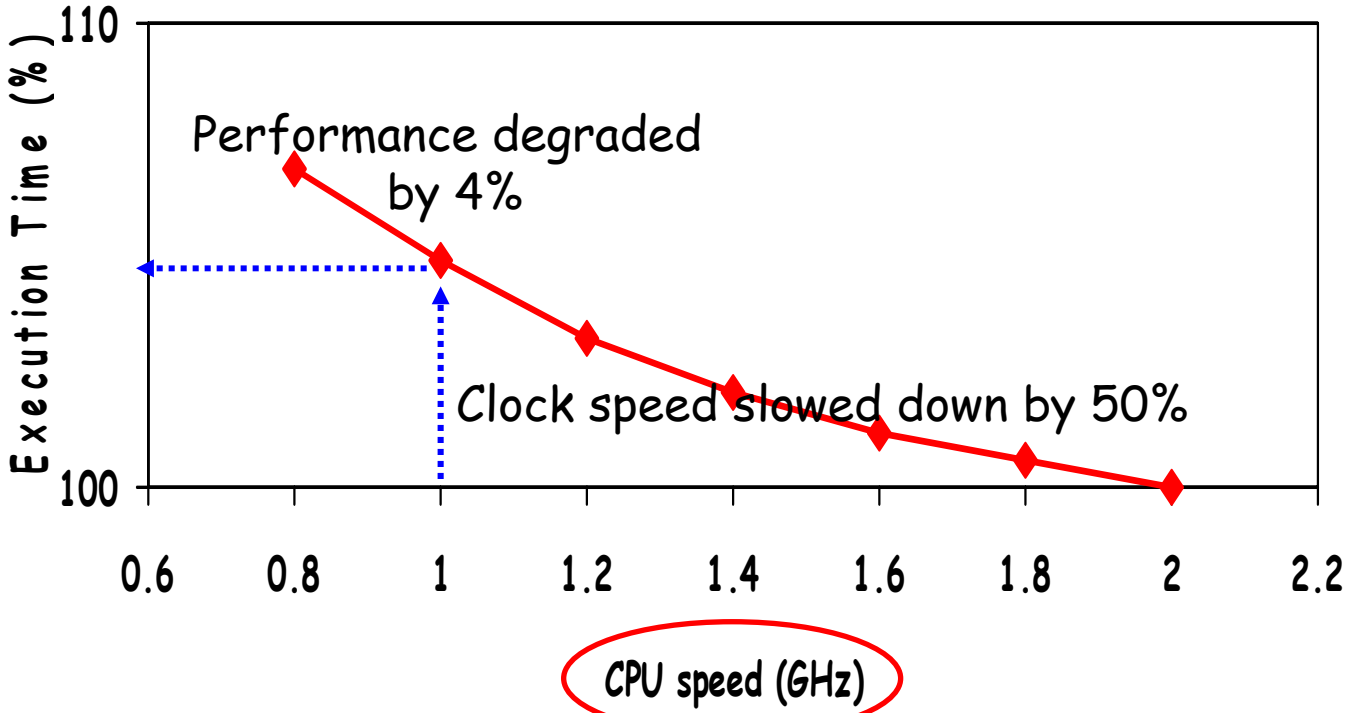
Los Alamos NATIONAL LABORATORY

Software-Based Evolution of GD: Dynamic Voltage Scaling (DVS)

- DVS Mechanism
  - Trades CPU performance for power reduction by allowing the CPU supply voltage and/or frequency to be adjusted at run-time.
- Why is DVS important?
  - Recall: Moore's Law for Power.
  - CPU power consumption is directly proportional to the *square of the supply voltage* and to *frequency*.
- DVS Algorithm
  - Determines *when* to adjust the current frequency-voltage setting and *what* the new frequency-voltage setting should be.

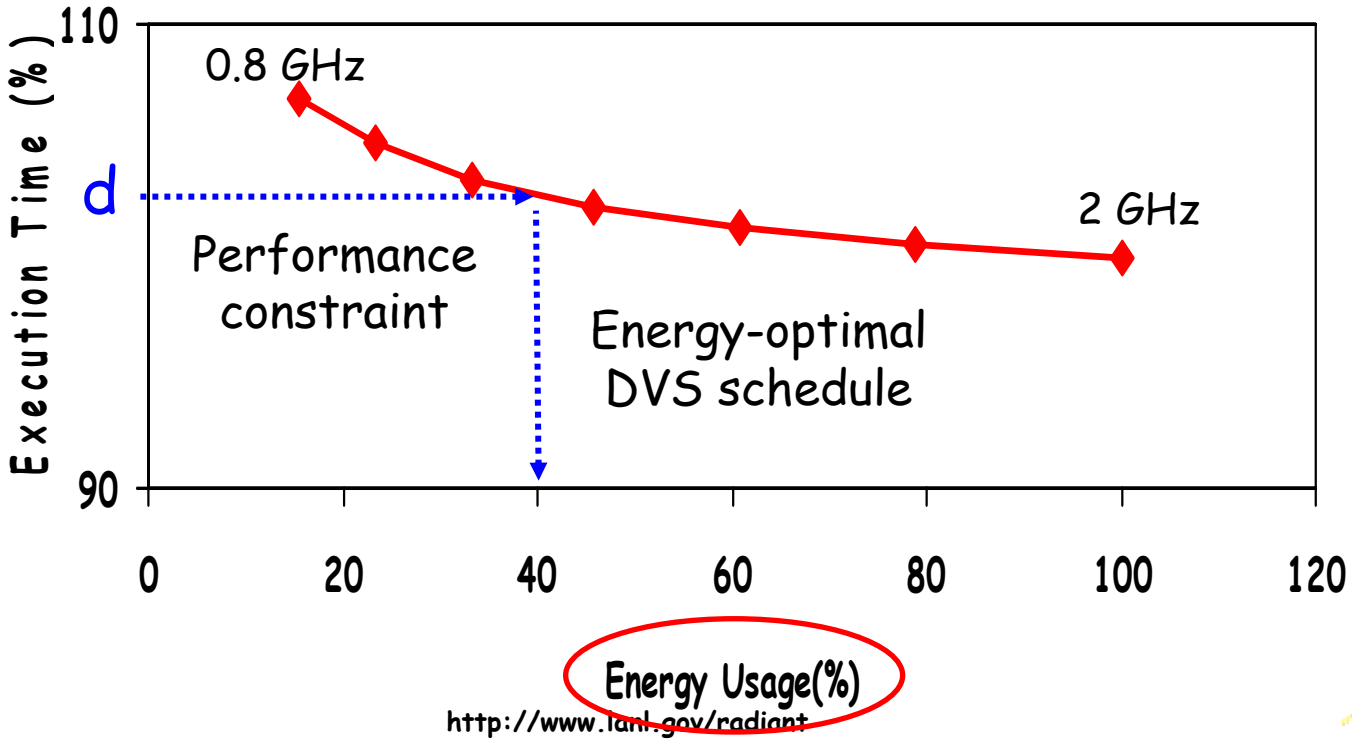
Motivation for Real-Time Constraint-Based DVS

- Key Observation
  - The execution time of many programs are insensitive to the CPU speed change. e.g., NAS IS benchmark.



Approach to Real-Time Constraint-Based DVS

- Key Idea
  - Applying DVS to these programs will result in significant power and energy savings at a minimal performance impact.



Wu Feng
feng@lanl.gov

http://www.lanl.gov/radiant
http://sss.lanl.gov

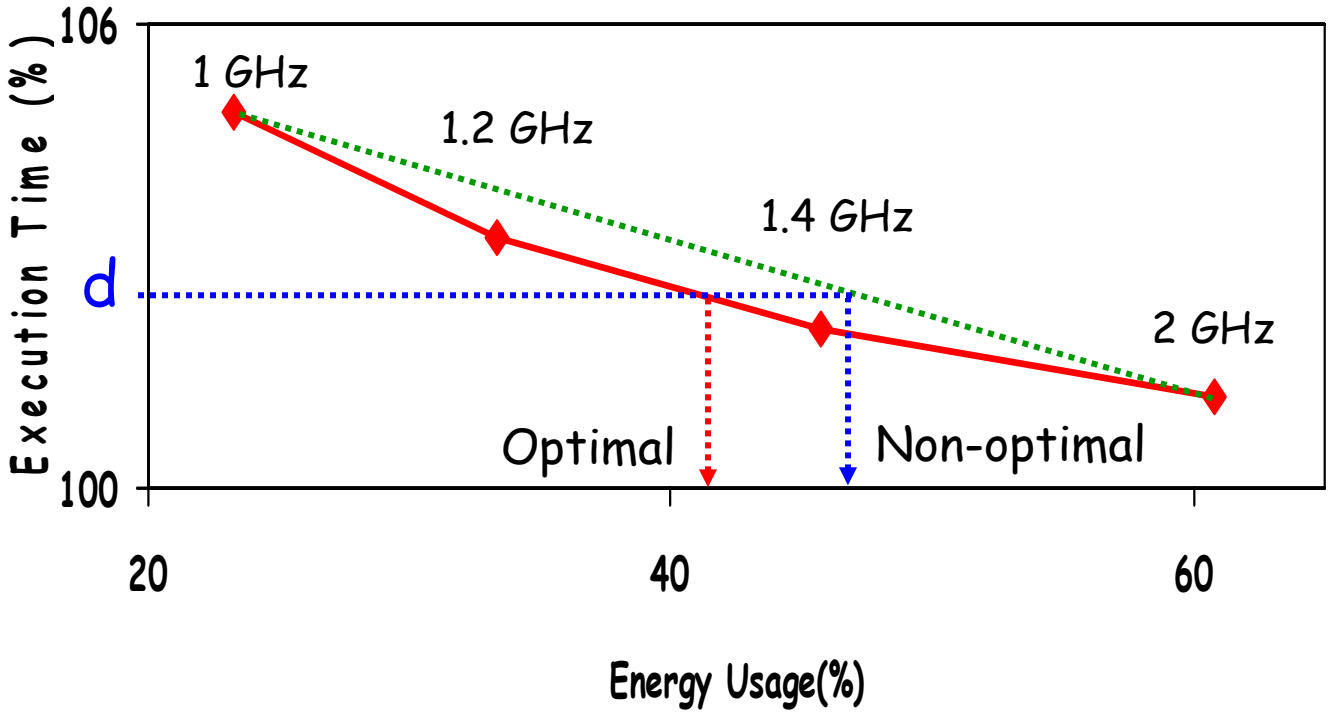
Problem Formulation for Real-Time Constraint-Based DVS

- Key Challenge
  - Find a performance-constrained, energy-optimal DVS schedule on a *realistic processor in real time*.
- Previous Related Work — Targeted at Embedded Systems …
  - $P \alpha V^2 f$
    1. $P \alpha f^3$ [assumes $V \alpha f$]
    2. Discretize *V*. Use continuous mapping function, e.g., $f = g(V)$, to get discrete *f*, e.g., 512 MHz, 894 MHz. Solve as ILP (offline) problem.
    3. Discretize *V* and *f*, e.g., AMD frequency-voltage table.
  - Simulation vs. Real Implementation
    - Problem with Simulation: Simplified Power Model
      - Does not account for leakage power.
      - Assumes zero-time switching overhead between *(f, V)* settings.
      - Assumes zero-time to construct a DVS schedule.
      - Does not assume realistic CPU support.

Creating an Energy-Optimal DVS Schedule

- Solve the following constraint-based problem:

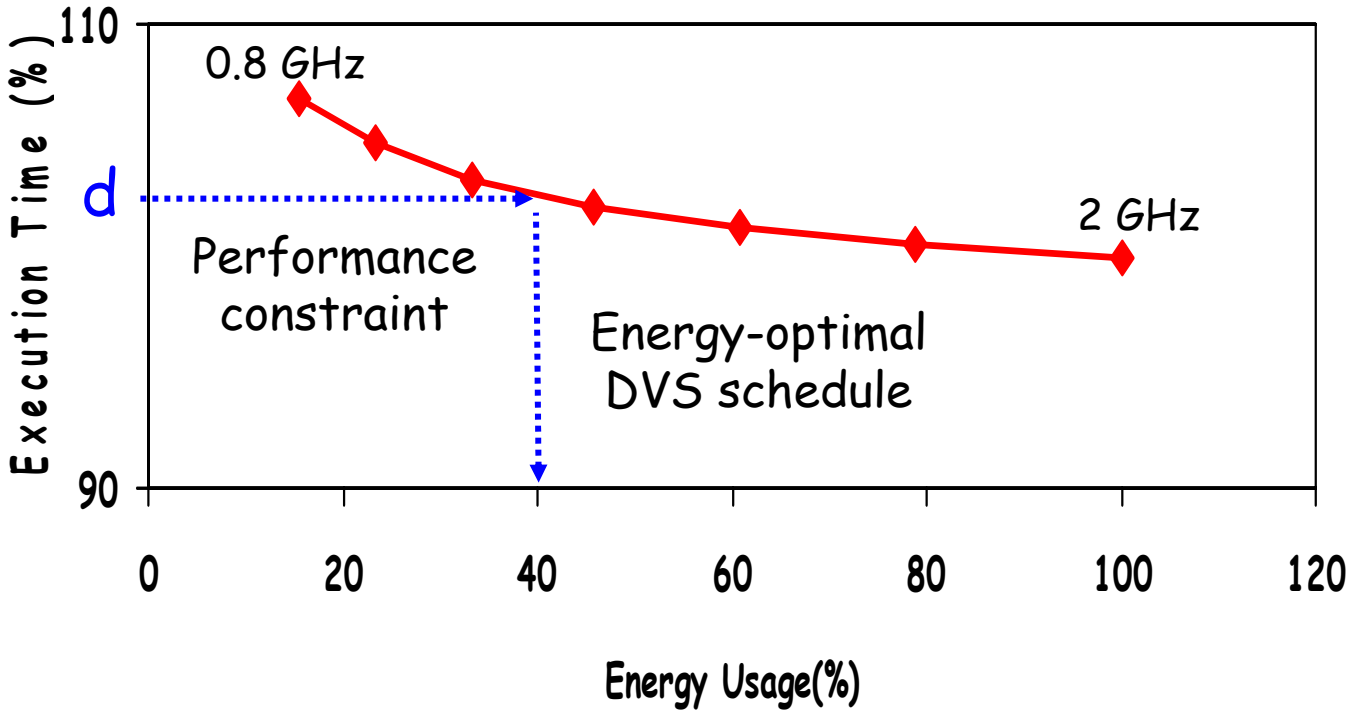
$$E = \min\{\sum_i r_i \cdot E_i : \sum_i r_i \cdot T_i \le d, \sum_i r_i = 1, r_i \ge 0\}$$



Wu Feng
feng@lanl.gov

Theorem for Real-Time Constraint-Based DVS

- If the execution-time vs. energy curve is convex, then the *energy-optimal DVS schedule* can be constructed in constant time.

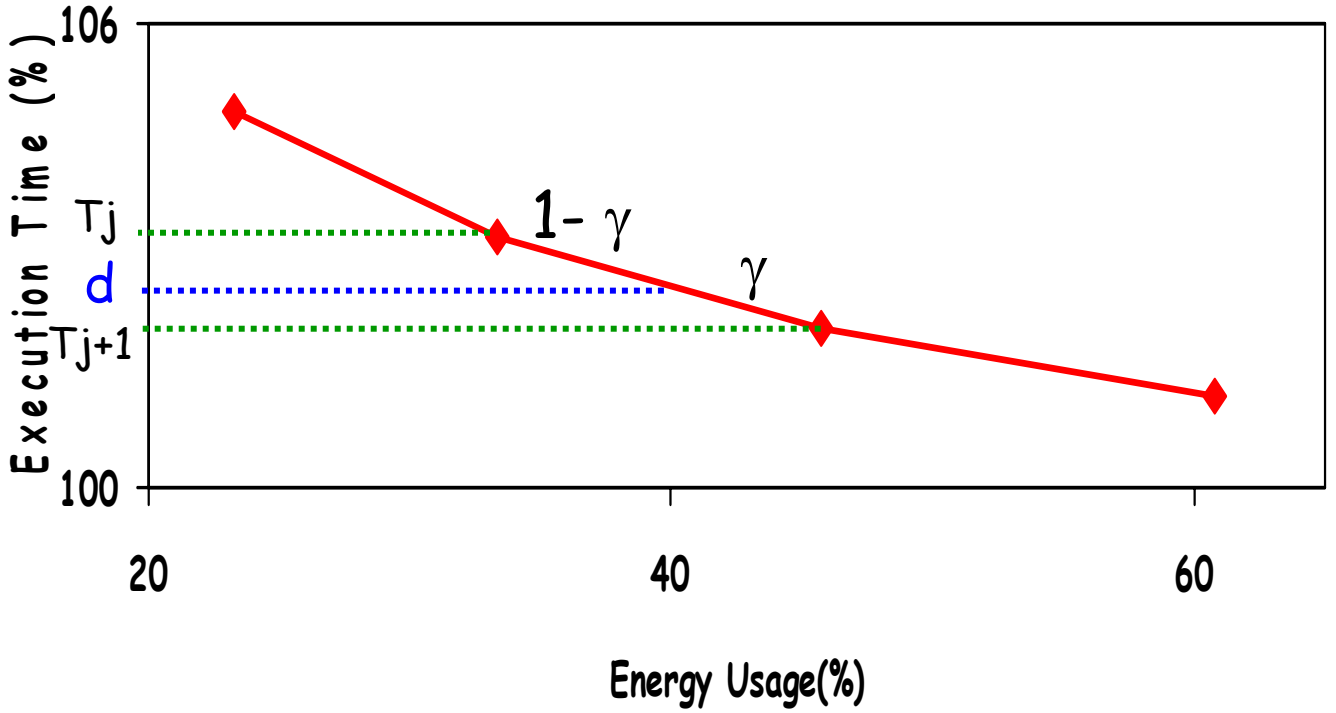


Wu Feng
feng@lanl.gov

Emulating Frequencies for an Energy-Optimal DVS Schedule

$$E = \gamma \cdot E_j + (1-\gamma) \cdot E_{j+1} \text{ where}$$

$$\gamma = \frac{d - T_{j+1}}{T_j - T_{j+1}} \quad \text{and} \quad T_{j+1} < d \leq T_j$$



Wu Feng
feng@lanl.gov

DVS Scheduling Algorithm

Input: deadline d and performance model $T(f)$

Output: deadline-constrained energy-optimal DVS schedule

Algorithm:

1. Figure out f_j and f_{j+1}.

$$T(f_{j+1}) < d \leq T(f_j)$$

2. Compute the ratio γ.

$$\gamma = \frac{d - T_{j+1}}{T_j - T_{j+1}}$$

3. Execute for γ percent of time at f_j
4. Execute for $1 - \gamma$ percent of time at f_{j+1}.

DVS Scheduling Algorithm

- Many programs can be modeled as follows:

$$\frac{T(f)}{T(f_{max})} = \beta \cdot \frac{f_{max}}{f} + (1 - \beta)$$

- To guarantee the execution-time vs. energy curve is convex, the following theorem is useful:

Theorem. If the above performance model holds and

$$\frac{P_1 - 0}{f_1 - 0} \leq \frac{P_2 - P_1}{f_2 - f_1} \leq \frac{P_3 - P_2}{f_3 - f_2} \leq \cdots \leq \frac{P_n - P_{n-1}}{f_n - f_{n-1}}$$

then

$$0 \geq \frac{E_2 - E_1}{T_2 - T_1} \geq \frac{E_3 - E_2}{T_3 - T_2} \geq \cdots \geq \frac{E_n - E_{n-1}}{T_n - T_{n-1}}$$

Initial Experimental Results

- Tested on a mobile AMD Athlon XP system with 5 settings
- Measured through Yokogawa WT210 digital power meter
- $\beta \in [0, 1]$ indicates performance sensitivity to changes in CPU speed, with 1 being most sensitive.

program	β	T_{rel}/E_{rel}
swim	0.02	1.02/0.46
tomcatv	0.24	1.01/0.80
su2cor	0.27	1.02/0.81
compress	0.37	1.05/0.80
mgrid	0.51	1.04/0.84
vortex	0.65	1.06/0.85
turb3d	0.79	1.04/0.92
go	1.00	1.05/0.93

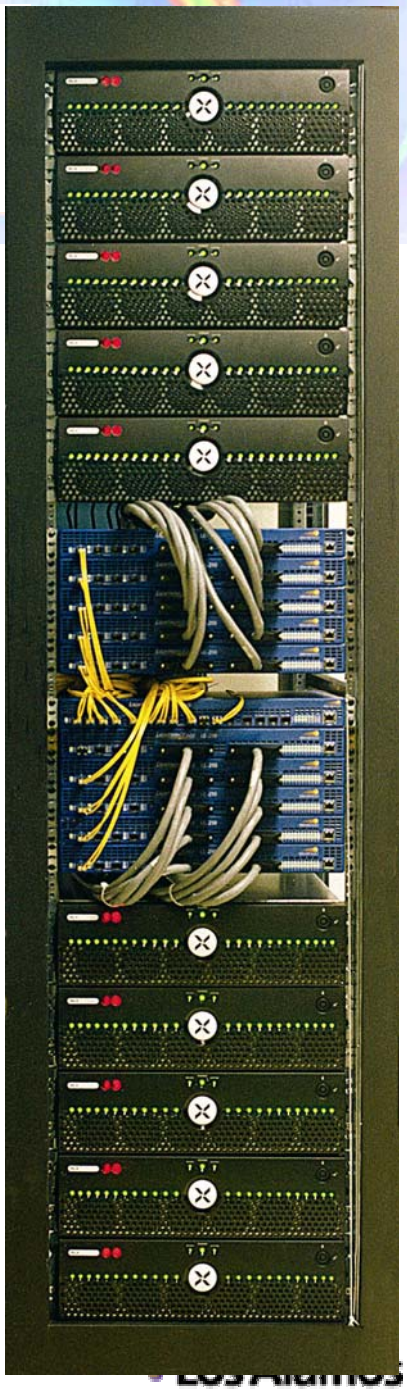
CAFfeine Supercomputer

- To debut by the end of the calendar year …
  - Imagine the results on the previous slide applied to a high-end compute node.
- Stay tuned …

Outline

- Motivation
  - Where is Supercomputing?
- Supercomputing in Small Spaces
  - From MetaBlade to Green Destiny
    - Experimental Results
  - The Evolution of Green Destiny
    - Architectural: Orion Multisystems DT-12
    - Software-Based: CAFfeine Supercomputer
- Past, Present, and Future
- Publications, Awards, Media, etc.
- Conclusion

mpiBLAST (http://mpiblast.lanl.gov) Performance on Green Destiny



BLAST Run Time for 300-kB Query against nt

Nodes	Runtime (s)	Speedup over 1 node
1	80774.93	1.00
4	8751.97	9.23
8	4547.83	17.76
16	2436.60	33.15
32	1349.92	59.84
64	850.75	94.95
128	473.79	170.49

The Bottom Line

mpiBLAST reduces search time from 1346 minutes (or 22.4 hours) to under 8 minutes.

Wu Feng
feng@lanl.gov

COMPUTER & COMPUTATIONAL SCIENCES

Past, Present, and Future for Green Destiny

- Application Codes
  - Astrophysics: Galaxy formation and 3-D supernova simulations
  - Large-scale molecular dynamics
  - mpiBLAST: Parallel genetic search tool
  - LINPACK: A collection of subroutines that analyze and solve linear equations and linear least-squares problems
- System Reliability
  - No unscheduled failures in 24 months despite the absence of special infrastructure such as cooling.
- Green Destiny R.I.P. and Re-Birth
  - Dismantled, re-configured, and expanded to 360 nodes in ten square feet. April 2004 – present.

Publications

- W. Feng and C. Hsu, "The Origin and Evolution of Green Destiny," *IEEE Cool Chips VII: An International Symposium on Low-Power and High-Speed Chips*, April 2004.
- W. Feng, "Making a Case for Efficient Supercomputing," *ACM Queue*, Oct. 2003. (Invited Paper)
- W. Feng, "Green Destiny + mpiBLAST = Bioinfomagic," *10th International Conference on Parallel Computing (ParCo'03)*, Sept. 2003.
- M. Warren, E. Weigle, and W. Feng, "High-Density Computing: A 240-Processor Beowulf in One Cubic Meter," *SC 2002*, Nov. 2002.
- W. Feng, M. Warren, and E. Weigle, "Honey, I Shrunk the Beowulf!," *International Conference on Parallel Processing*, Aug. 2002.

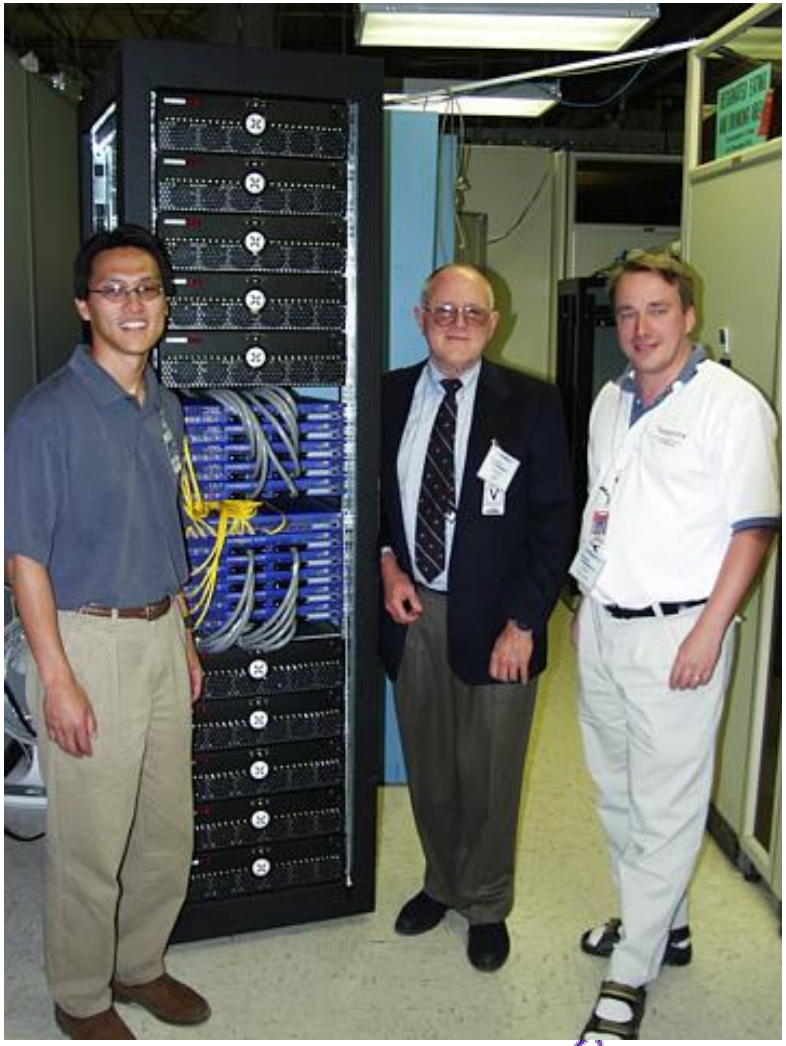
Awards

- Green Destiny
  - R&D 100 Award, Oct. 2003.
  - … and Its Evolving Parts
    - Innovative Supercomputer Architecture Award at the 2004 International Supercomputer Conference, where the Top 500 Supercomputer List is announced every June.

COMPUTER & COMPUTATIONAL SCIENCES

Sampling of Media Overexposure

- "Green Destiny: A 'Cool' 240-Node Supercomputer in a Telephone Booth," *BBC News*, Aug. 2003.
- "Servers on the Edge: Blades Promise Efficiency and Cost Savings," *CIO Magazine*, Mar. 2003.
- "LANL Researchers Outfit the 'Toyota Camry' of Supercomputing for Bioinformatics Tasks," *BioInform / GenomeWeb*, Feb. 2003.
- "Developments to Watch: Innovations," *BusinessWeek*, Dec. 2002.
- "Craig Venter Goes Shopping for Bioinformatics to Fill His New Sequencing Center," *GenomeWeb*, Oct. 2002.
- "Not Your Average Supercomputer," *Communications of the ACM*, Aug. 2002.
- "At Los Alamos, Two Visions of Supercomputing," *The New York Times*, Jun. 25, 2002.
- "Researchers Deliver Supercomputing in Smaller Package," *Yahoo! Finance*, Jun. 2002.
- "Supercomputing Coming to a Closet Near You?" *PCWorld.com*, May 2002.
- "Bell, Torvalds Usher Next Wave of Supercomputing," *CNN*, May 2002.



Wu Feng
feng@lanl.gov

http://www.lanl.gov/radiant
http://sss.lanl.gov

Los Alamos NATIONAL LABORATORY

Conclusion

- Efficiency, reliability, and availability will be *the* key issues of this decade.

- Performance Metrics for Green Destiny (circa 2002)
  - Performance: 2x to 2.5x worse than fastest AMD/Intel.
  - Price/Performance: 2x to 2.5x worse.
  - Overall Efficiency (Total Price-Performance Ratio)
    - 1.5x to 2.0x better. See ACM Queue, Oct. 2003.
  - Power Efficiency (Perf/Power): 10x to 20x better.
  - Space Efficiency (Perf/Space): 20x to 60x better.
  - Reliability: "Infinite"
  - Availability: Nearly 100%.

SUPERCOMPUTING
in SMALL SPACES

http://sss.lanl.gov

Research And Development In Advanced Network Technology

http://www.lanl.gov/radiant

Wu-chun (Wu) Feng

feng@lanl.gov