Green Destiny:

A 240-Node Compute Cluster in One Cubic Meter

Wu Feng

Research & Development in Advanced Network Technology (RADIANT)
Computer & Computational Sciences Division
Los Alamos National Laboratory

LA-UR 02-6127

Disclaimer

- The "Supercomputing in Small Spaces" Project
  - Project Initiation: Oct. 2001.
  - An active research project to determine future directions in efficient high-performance computing.
  - Not a "replacement" for traditional Beowulf clusters or supercomputers (which generally require special infrastructure to house).
  - A complementary solution, particularly for those who are space- or power-constrained.
    - Eric Schmidt, Google: New York Times, Sept. 2002.
      - Low power, NOT speed.
      - DRAM density, NOT speed.

Outline

- Where is Supercomputing?
  - Architectures from the Top 500.
- Evaluating Supercomputers
  - Metrics: Performance & Price/Performance
- An Alternative Flavor of Supercomputing
  - Supercomputing in Small Spaces → Bladed Beowulf
- Architecture of a Bladed Beowulf
- Performance Metrics
- Benchmark Results
- Discussion & Status
- Conclusion
- Acknowledgments

COMPUTER & COMPUTATIONAL SCIENCES

Flavors of Supercomputing

(Picture Source: Thomas Sterling, Caltech & NASA JPL)

Metrics for Evaluating Supercomputers

- *Performance*
  - Metric: Floating-Operations Per Second (FLOPS)
  - Example: Japanese Earth Simulator
- *Price/Performance → Cost Efficiency*
  - Metric: Cost / FLOPS
  - Examples: SuperMike, GRAPE-5, Avalon.
- Performance & price/performance are important metrics, but …

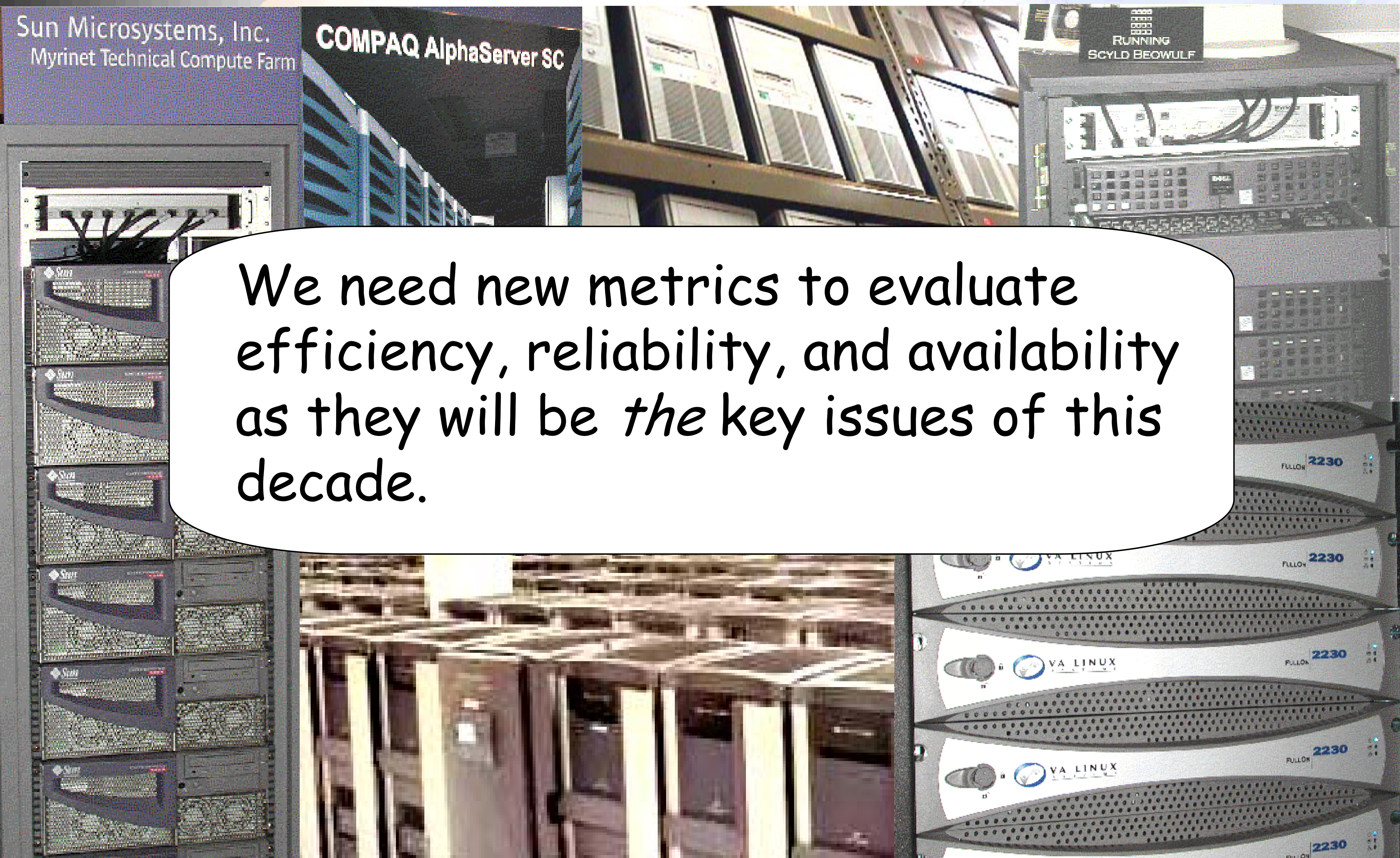
The Need for *New* Supercomputing Metrics

- But which metric?!
- Analogy: Buying a car. Which metric to use?
  - Raw Performance: Ferrari 550.
  - Price/Performance: Ford Mustang GTO.
  - Fuel Efficiency: Honda Insight.
  - Reliability: Toyota Camry.
  - Storage: Honda Odyssey.
  - All-Around: Volvo XC90.
- So many metrics to evaluate a car …
  why not to evaluate a supercomputer?
- But which metrics?

COMPUTER & COMPUTATIONAL SCIENCES

Flavors of Supercomputing

(Picture Source: Thomas Sterling, Caltech & NASA JPL)



We need new metrics to evaluate efficiency, reliability, and availability as they will be *the* key issues of this decade.

Why Efficiency, Reliability, and Availability (ERA)?

- Requirement: Near-100% availability.
  - E-commerce, enterprise apps, online services, ISPs.
- Problems
  - Frequency of Service Outages
    - 65% of IT managers report that their websites were unavailable to customers over a 6-month period
      - 25%: 3 or more outages
  - 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.

Source: InternetWeek 4/3/2000

Why Efficiency, Reliability, and Availability?

- A Different Perspective …
  - "Performance" and "Price/Performance" Metrics …
    - Lower efficiency, reliability, and availability, e.g., Cray C90 to Q.
    - Higher operational costs, e.g., admin, maintenance, etc.
  - Consumers of Cluster Cycles
    - Pharmaceutical, financial, actuarial, retail, aerospace, data centers for web-server farms.
  - Requirements of Clusters
    - Wu-chun Feng, LANL. *CNN, 5/2002; NY Times, 6/2002*
      - [1] Power → Reliability, [2] Space, [3] Performance → Speed.
    - Peter Bradley, Pratt & Whitney. *IEEE Cluster, 9/2002*
      - [1] Reliability, [2] Transparency, [3] Resource Management (Avail.)
    - Eric Schmidt, Google. *IEEE Hot Chips & NY Times, 9/2002*
      - Low power, NOT speed.
      - DRAM density, NOT speed.
      - Reliability, NOT speed.

An Alternative Flavor of Supercomputing

- Supercomputing in Small Spaces (http://sss.lanl.gov)
  - First instantiation: *Bladed Beowulf*
    - MetaBlade (24), MetaBlade2 (24), and Green Destiny (240).
- 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 hardware failures, etc.
  - Reduce the *total cost of ownership* (TCO).
- Analogy
  - Ferrari 550: Wins raw performance but reliability is poor so it spends its time in the shop. Throughput low.
  - Toyota Camry: Loses raw performance but high reliability results in high throughput (i.e., miles driven).

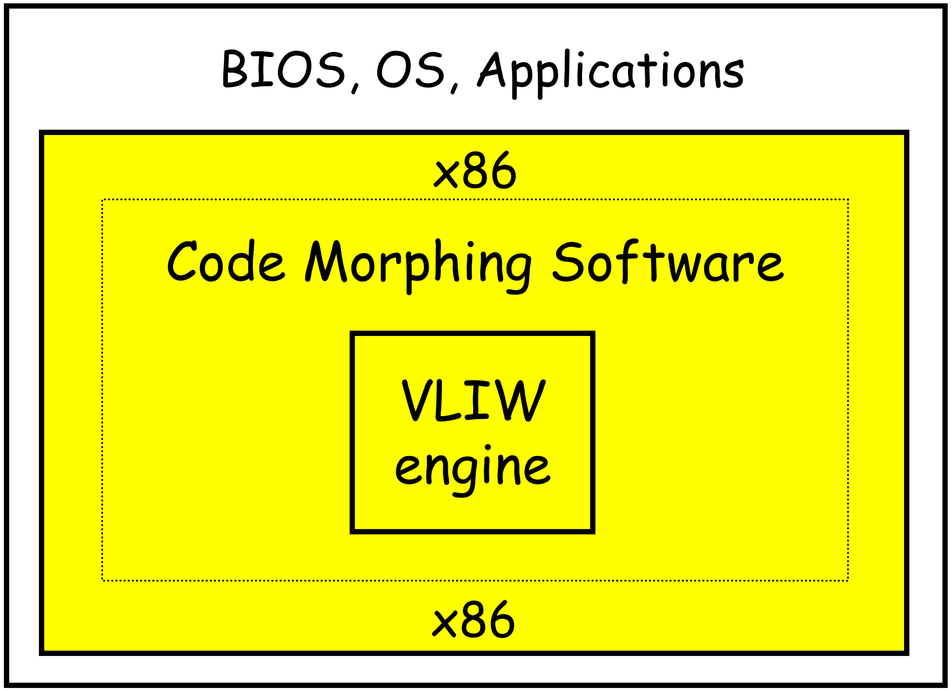
Architecture of a Bladed Beowulf

A Fundamentally Different Approach to High-Performance Computing

COMPUTER & COMPUTATIONAL SCIENCES
NNSA
Los Alamos NATIONAL LABORATORY

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 → ~ 6x-7x less power dissipation
  - Less power → lower "on-die" temp. → better reliability & availability

Transmeta TM5x00 Comparison

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

- Current-generation Transmeta TM5800 performs comparably to an Intel PIII over iterative scientific codes on a clock-for-clock-cycle basis.
- Next-generation Transmeta CPU rectifies the above mismatch in functional units.

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.
- Modules for CMS
  - Interpreter
    - Interprets x86 instructions (*a la* Java).
    - Filters infrequently executed code from being optimized.
    - Collects run-time statistical information.
  - Translator
    - Re-compiles x86 instructions into optimized VLIW instructions (*a la* JIT compiler).

RLX ServerBlade™ 633 (circa 2000)

Code Morphing Software (CMS), 1 MB

Public NIC
33 MHz PCI

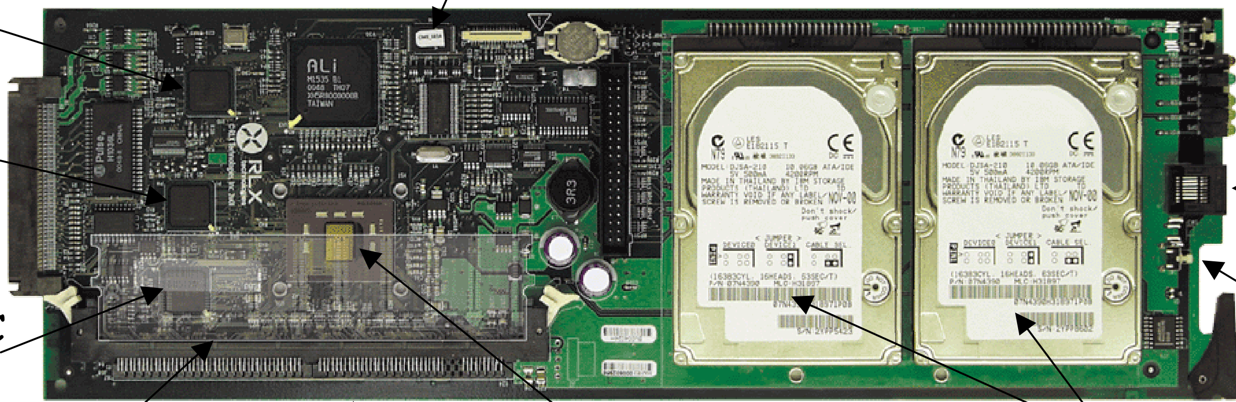
Private NIC
33 MHz PCI

Management NIC
33 MHz PCI

Status LEDs

Serial RJ-45 debug port

Reset Switch



128MB, 256MB, 512MB
DIMM SDRAM
PC-133

512KB
Flash ROM

Transmeta™
TM5600 633 MHz

Crusoe™

ATA 66
0 or 1 or 2 - 2.5" HDD
10 or 30 GB each

RLX ServerBlade™ 1000T
$999

128KB L1 cache, 512KB L2 cache
LongRun, Northbridge, x86 compatible

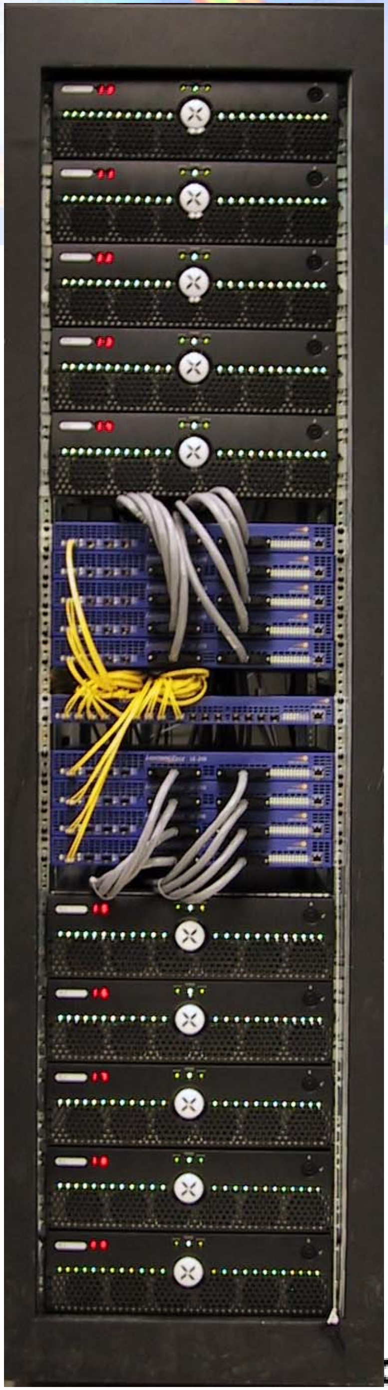
RLX System™ 324

3U chassis that houses 24 blades

- 3U vertical space
  - 5.25" x 17.25" x 25.2"
- Two hot-pluggable 450W power supplies
  - Load balancing
  - Auto-sensing fault tolerance
- System midplane
  - Integration of system power, management, and network signals.
  - Elimination of internal system cables.
  - Enabling efficient hot-pluggable blades.
- Network cards
  - Hub-based management.
  - Two 24-port interfaces.

RLX System™ 300ex
- Interchangeable blades
  - Intel, Transmeta, or both.
- Switched-based management

"Green Destiny" Bladed Beowulf



- A 240-Node Beowulf in One Cubic Meter
- Each Node
  - 667-MHz Transmeta TM5600 CPU
    - Upgrade to 933-MHz Transmeta TM5800 CPUs
  - 640-MB RAM
  - 20-GB hard disk
  - 100-Mb/s Ethernet (up to 3 interfaces)
- Total
  - 160 Gflops peak (224-240 Gflops with upgrade)
  - 240 nodes
  - 150 GB of RAM (expandable to 276 GB)
  - 4.8 TB of storage (expandable to 38.4 TB)

Who Cares? So What? It's a Smaller Beowulf …

- Goal
  - Improve *efficiency*, *reliability*, and *availability* (ERA) in large-scale computing systems.
  - Reduce the *total cost of ownership* (TCO).
- How to quantify ERA?
- What exactly is TCO?
  - Can it be concretely quantified?
  - Or is it a "foofy" metric?

Performance Metrics

What is TCO?

- Cost of Acquisition ← Fixed, one-time cost
  - $$$ to buy the supercomputer.
- Cost of Operation ← Variable, recurring cost
  - Administration
    - $$$ to build, integrate, configure, maintain, and upgrade the supercomputer over its lifetime.
  - Power & Cooling
    - $$$ in electrical power and cooling that is needed to maintain the operation of the supercomputer.
  - Downtime
    - $$$ lost due to the downtime (unreliability) of the system.
  - Space
    - $$$ spent to house the system.

Total Price-Performance Ratio

- Price-Performance Ratio
  - *Price = Cost of Acquisition*
  - Performance = Floating-Point Operations Per Second
- Total Price-Performance Ratio (ToPPeR)
  - *Total Price = Total Cost of Ownership (TCO)*
  - Performance = Floating-Point Operations Per Second
- Using "Flops" as a performance metric is problematic as well … another talk, another time …

Quantifying TCO?

- Why is TCO hard to quantify?
  - Components
    - Acquisition + Administration + Power + Downtime + Space

Quantifying TCO?

- Why is TCO hard to quantify?
  - Components
    - Acquisition + Administration + Power + Downtime + Space

Too Many Hidden Costs

Institution-Specific

Quantifying TCO?

- Why is TCO hard to quantify?
  - Components
    - Acquisition + Administration + Power + Downtime + Space

Too Many Hidden Costs

Institution-Specific

  - Traditional Focus: Acquisition (i.e., equipment cost)
    - Cost Efficiency: Price/Performance Ratio

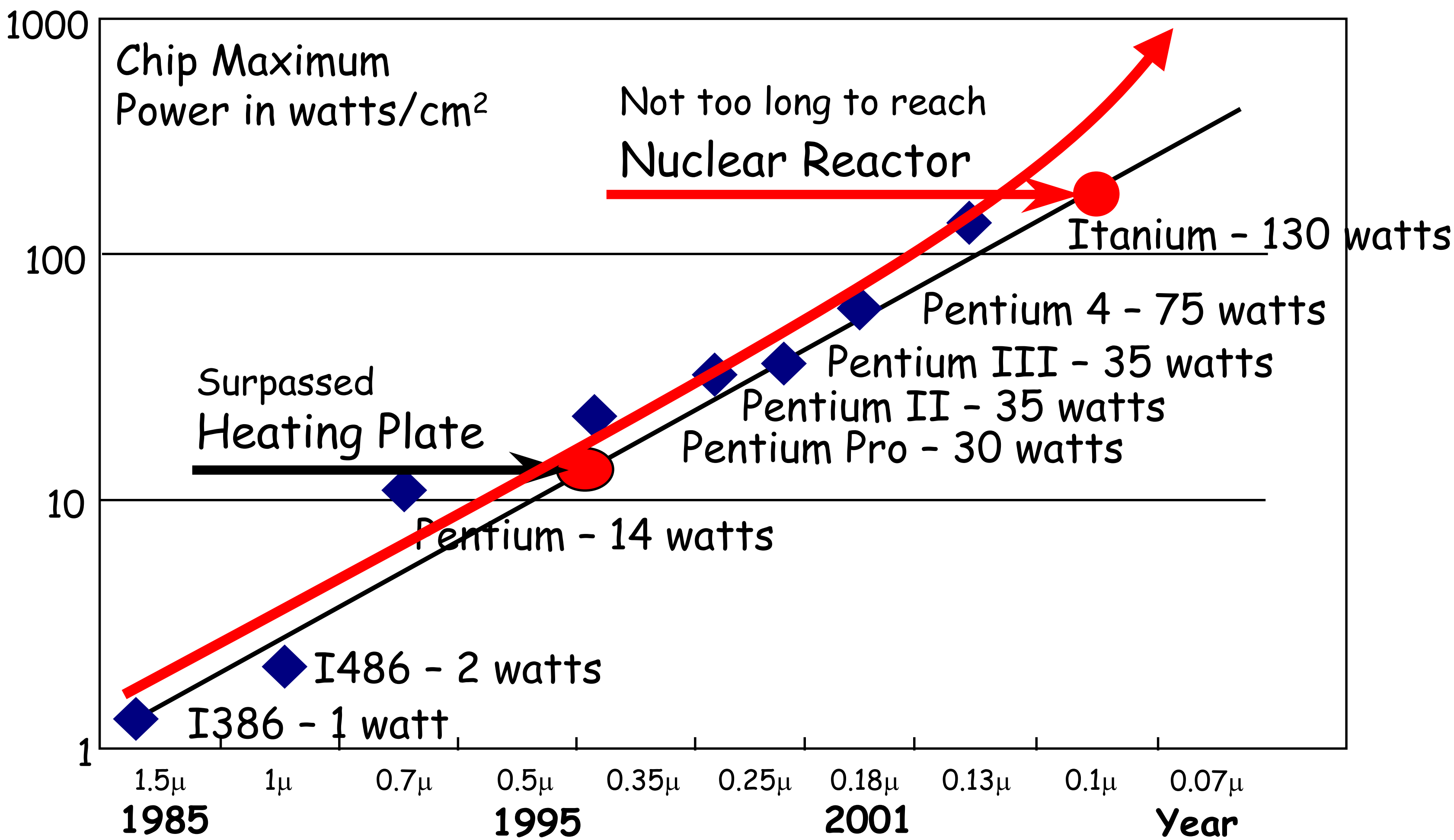
Quantifying TCO?

- Why is TCO hard to quantify?
  - Components
    - Acquisition + Administration + Power + Downtime + Space

      Institution-Specific

      Too Many Hidden Costs
  - Traditional Focus: Acquisition (i.e., equipment cost)
    - Cost Efficiency: Price/Performance Ratio
  - New *Quantifiable* Efficiency Metrics
    - Power Efficiency: Performance/Power Ratio
    - Space Efficiency: Performance/Space Ratio

Moore's Law for Power Dissipation



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

Power, Temperature, Reliability

- What's wrong with high power?
  - Costs $$$ to power such a system; costs $$$ to cool it.
  - Causes reliability problems. Why?
    - Higher power implies higher temperatures.

- Arrhenius' Equation (circa 1980s)
  - As temperature increases by 10° C …
    - The failure rate of a system *doubles*.
    - The reliability of a system is cut in *half*.
  - Twenty years of unpublished empirical data .

Empirical Data on Temperature

- From off to system boot-up, after 25 seconds:

Processor	Clock Freq.	Voltage	Peak Temp.*
Intel Pentium III-M	500 MHz	1.6 V	252° F (122° C)
Transmeta Crusoe TM5600	600 MHz	1.6 V	147° F (64° C)

*Peak temperature measured with *no* cooling.

- Recall: Arrehenius' Equation
  - Every 10° C increase *doubles* the failure rate.

Summary of Performance Metrics

- Total Price/Performance Ratio (ToPPeR)
  - Price is more than the *cost of acquisition.*
  - Operational costs: sys admin, power & cooling, space, downtime.
- Performance/Power Ratio → "Power Efficiency"
  - How efficiently does a computing system use energy?
  - How does this affect reliability and availability?
    - Higher Power Dissipation α Higher Temperature α Higher Failure Rate
- Performance/Space Ratio → "Space Efficiency"
  - How efficiently does a computing system use space?
  - Performance has increased by 2000 since the Cray C90; performance/sq. ft. has only increased by 65.

Benchmark Results

LINPACK Benchmark

- LANL Tests
  - 933-MHz RLX ServerBlade 1000t — 600 Mflops
    (but with a special high-performance CMS)
  - 1200-MHz RLX ServerBlade 1200i — 625 Mflops
  - 933-MHz RLX ServerBlade 1000t — (est.) 410 Mflops
    (but with "plain" CMS)
- RLX Tests
  - 800-MHz RLX ServerBlade 800i — 349 Mflops
  - 1200-MHz RLX ServerBlade 1200i — 523 Mflops

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.

Memory Bandwidth for Transmetas (via STREAMS): 350 MB/s

Treecode Benchmark for n-Body

Site	Machine	CPUs	Gflops	Mflops/CPU
NERSC	IBM SP-3	256	57.70	225.0
LANL	SGI O2K	64	13.10	205.0
LANL	Green Destiny	212	38.90	183.5
SC'01	MetaBlade2	24	3.30	138.0
LANL	Avalon	128	16.16	126.0
LANL	Loki	16	1.28	80.0
NASA	IBM SP-2	128	9.52	74.4
SC'96	Loki+Hyglac	32	2.19	68.4
Sandia	ASCI Red	6800	464.90	68.4
CalTech	Naegling	96	5.67	59.1
NRL	TMC CM-5E	256	11.57	45.2

Treecode Benchmark for n-Body

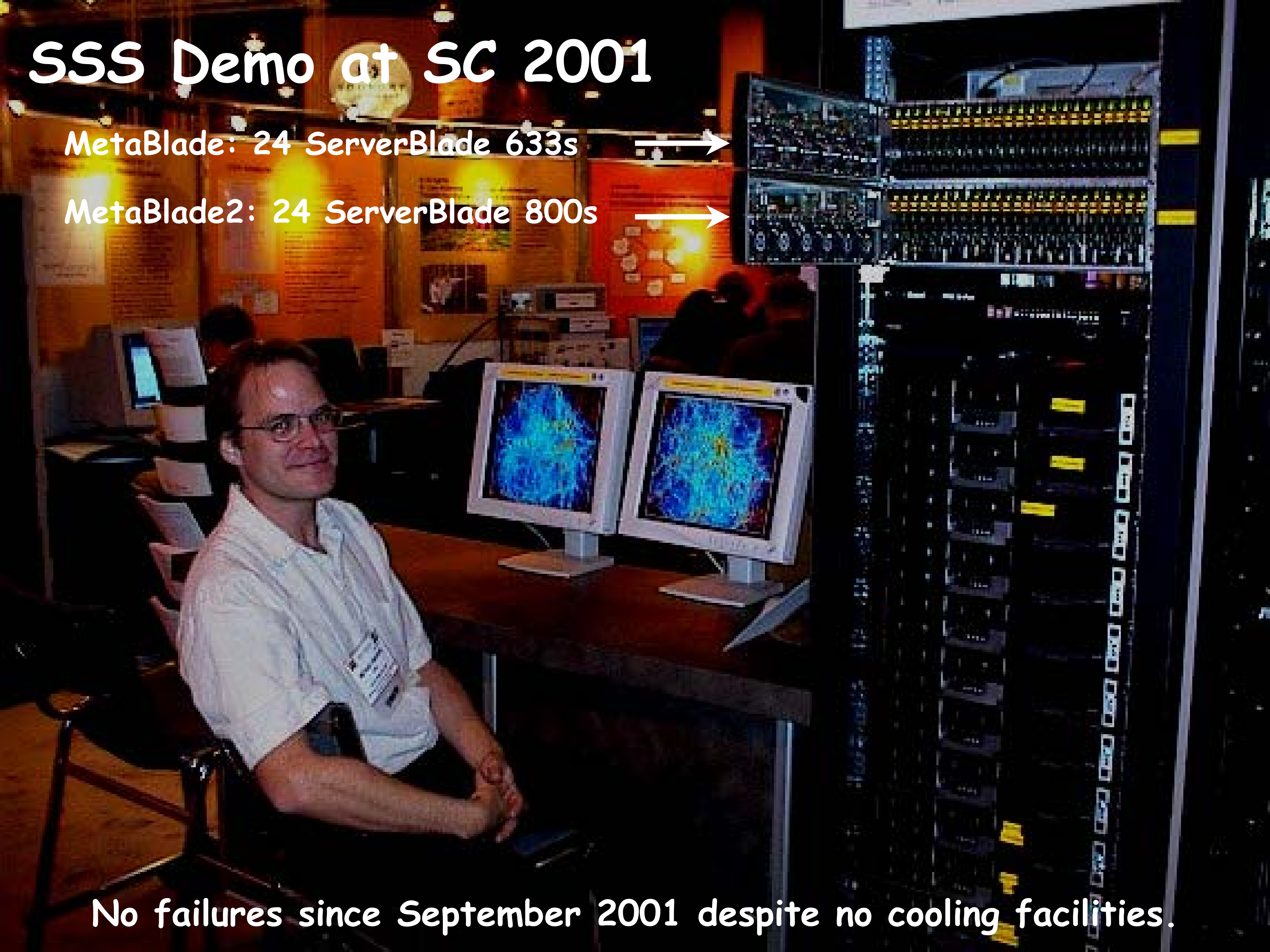
Site	Machine	CPUs	Gflops	Mflops/CPU
NERSC	IBM SP-3	256	57.70	225.0
LANL	SGI O2K	64	13.10	205.0
LANL	Green Destiny	212	38.90	183.5
SC'01	MetaB[illegible]	24	3.30	138.0
LANL	Avalon	[illegible]	[illegible]	126.0
LAN[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
NAS[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
SC'96	Loki+Hyglac	[illegible]	[illegible]	68.4
Sandia	ASCI Red	6800	464.90	68.4
CalTech	Naegling	96	5.67	59.1
NRL	TMC CM-5E	256	11.57	45.2

Upgraded "Green Destiny"
58 Gflops → 274 Mflops/CPU

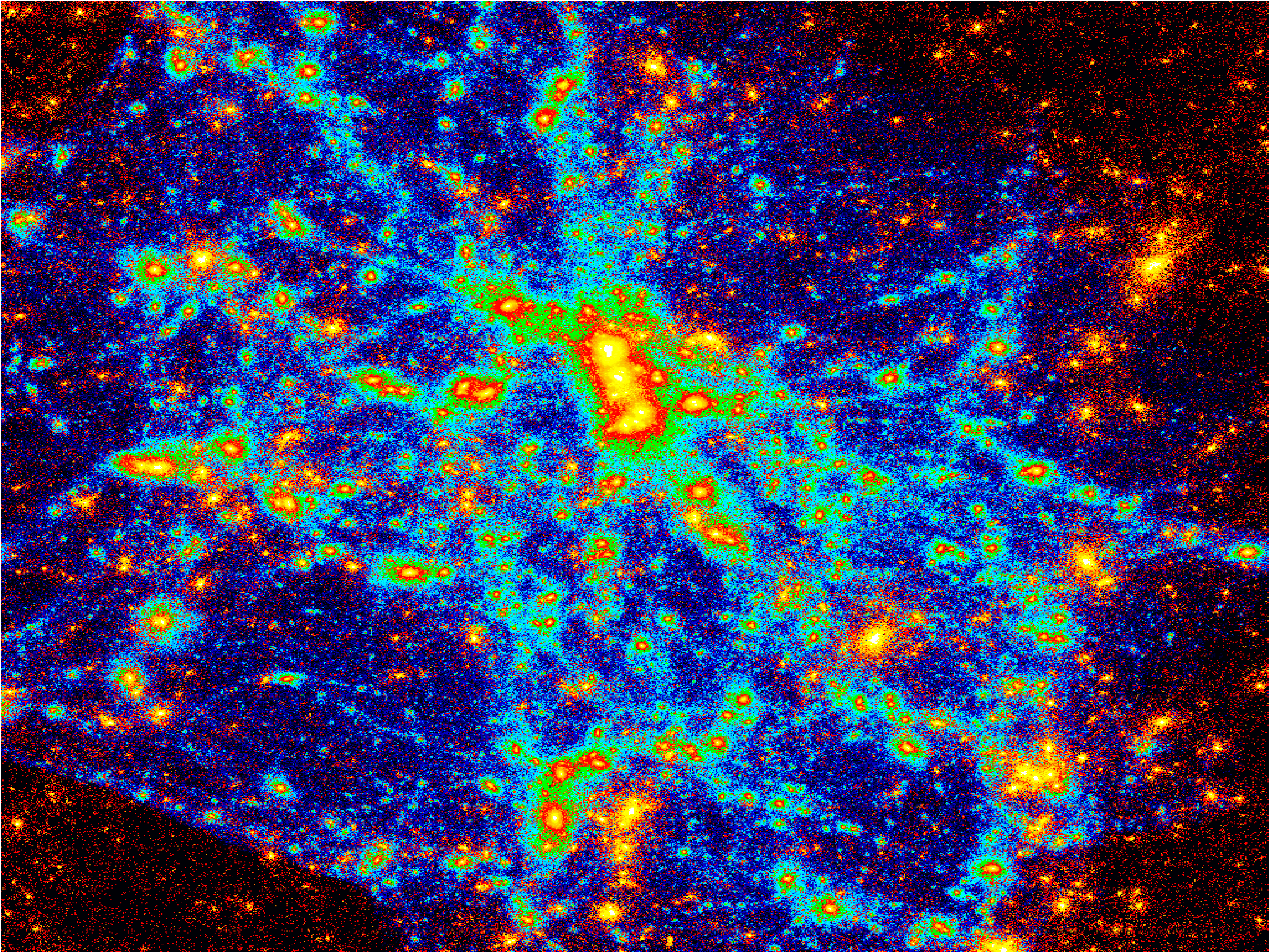
SSS Demo at SC 2001

MetaBlade: 24 ServerBlade 633s

MetaBlade2: 24 ServerBlade 800s



No failures since September 2001 despite no cooling facilities.



"Cost Efficiency" Metrics

- Price-Performance Ratio
  - *Price = Cost of Acquisition*
  - ✓ Performance = Floating-Point Operations Per Second
- Total Price-Performance Ratio (ToPPeR)
  - *Total Price = Total Cost of Ownership (TCO)*
  - ✓ Performance = Floating-Point Operations Per Second

ToPPeR Metric

- TCO: Total Cost of Ownership
(over the lifetime of a 24-node cluster in a 80° F environment)

Cost Parameter	Alpha	Athlon	PIII	P4	TM5600
Acquisition	\$17K	\$15K	\$16K	\$17K	\$26K
System Admin	\$60K	\$60K	\$60K	\$60K	\$5K
Power & Cooling	\$12K	\$27K	\$12K	\$22K	\$2K
Space	\$8K	\$8K	\$8K	\$8K	\$2K
Downtime	\$12K	\$12K	\$12K	\$12K	\$1K
TCO (four yrs)	\$109K	\$122K	\$108K	\$119K	\$36K

- ToPPeR is upwards of 2.0x better for Green Destiny…
- *Problem: Too many hidden costs & institution-specific.*

Parallel Computing Platforms (for "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*

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
Space efficiency ($Mflops/ft^2$)	150	375	252	6500
Power efficiency (Mflops/watt)	1.0	0.5	1.3	7.5

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
Space efficiency ($Mflops/ft^2$)	150	375	252	6500
Power efficiency (Mflops/watt)	1.0	0.5	1.3	7.5

Green Destiny vs. Japanese Earth Simulator

Machine	Green Destiny+	Earth Simulator
Year	2002	2002
LINPACK Performance (Gflops)	144 (est.)	35,860
Area (ft^2)	6	17,222
Power (kW)	5	7,000
Cost efficiency ($/Mflop)	2.33	11.15
Space efficiency (Mflops/ft^2)	24,000	2,085
Power efficiency (Mflops/watt)	28.8	5.13

Disclaimer: This is not exactly a fair comparison. Why?

(1) LINPACK performance is extrapolated for Green Destiny+.

(2) Use of area and power does *not* scale linearly.

(3) Goals of the two machines are different.

Discussion: Interesting Tidbits

- DARPA Contributes $2M to IBM's Low Power Center in Aug. 2002
  - http://www.computerworld.com/industrytopics/defense/story/0,10801,73289,00.html
- Transmeta performance (with high-performance CMS) on N-body code can match Intel performance on a clock-for-clock-cycle basis, e.g., LINPACK.
  - Problem: Fastest Transmeta? Fastest Intel?
- Low component count on blade server enhances reliability.
  - 105 parts per RLX node vs. 800-1000 parts per typical node.
- Intel-based Bladed Beowulf: 18 nodes in 3U
  - Room at 80° F: "Silent" failure on LINPACK.
1/3 of nodes inaccessible.
  - Room at 65° F: ~20% better LINPACK perf. vs. 933-MHz Transmeta.
- Why 10/100? GigE has been available for two years now.
  - In 2000-01, GigE ~12-15 W. Now, GigE ~6-8W?
- Systems community vs. applications community.

Status

- Recent Work
  - April 2002: Assembled and integrated a 240-node Beowulf in one cubic meter called *Green Destiny*.
  - July 2002: Worked with Transmeta to demonstrate comparable performance to similarly-clocked Intels.
  - July 2002: Worked with DOE SciDAC-funded *3-D Supernova* project to demo "base code" on Green Destiny. (A vertically-integrated solution from hardware on up to the application.)
  - August 2002: Completed mpiBLAST code. Presented at IEEE Bioinformatics. Demonstrated super-linear speed-up.
- Future Work
  - Work with additional code teams, e.g., climate modeling, computational fluid dynamics, large-scale molecular dynamics.
  - Complete upgrade of Green Destiny processors from 667 MHz to 933 MHz.

Conclusion

- New Performance Metrics
  - Overall Efficiency
    - ToPPeR: Total Price-Performance Ratio
  - Power Efficiency
    - Performance-Power Ratio
  - Space Efficiency
    - Performance-Space Ratio

Conclusion

- Performance Metrics for Green Destiny
  - Performance
    - 2x to 2.5x worse than fastest Intel/AMD processor.
  - Price/Performance
    - 2x to 2.5x worse.
  - Overall Efficiency: ToPPeR
    - 1.2x to 1.7x better.
  - Power Efficiency: Performance-Power Ratio
    - 7x – 8x better.
  - Space Efficiency: Performance-Space Ratio
    - 20x – 30x better.

Conclusion

- Keeping It In Perspective
  - The "Supercomputing in Small Spaces" project (http://sss.lanl.gov) is *not* meant to replace today's large supercomputers.
    - Focus on metrics related to efficiency, reliability, and availability (ERA) rather than raw performance.
      - i.e., SSS = "Toyota Camry" of supercomputing.
    - Works particularly well as a departmental cluster (or even institutional cluster if there exists power and space constraints).

Acknowledgments

- Technical Co-Leads
  - Mike Warren and Eric Weigle
- Contributions
  - Mark Gardner, Adam Engelhart, Gus Hurwitz
- Enablers
  - Chris Hipp (Founder & Former CTO of RLX)
  - J. Thorp, A. White, R. Oldehoeft, and D. Lora (LACSI)
  - W. Feiereisen and S. Lee (CCS Division Office)
- Funding Agencies
  - LACSI
  - IA-Linux
- Encouragement & Support
  - Gordon Bell, Chris Hipp, Linus Torvalds

The "Hype": A Sampling of Press Coverage

- "Not Your Average Supercomputer," *Communications of the ACM*, 8/02.
- "At Los Alamos, Two Visions of Supercomputing," *The New York Times*, 6/25/02.
- "Two Directions for the Future of Supercomputing," *slashdot.org*, 6/25/02.
- "Researchers Deliver Supercomputing in Smaller Package," *Yahoo! Finance*, 6/4/02.
- "Computer World Faces Heat Wave," *Santa Fe New Mexican*, 6/3/02.
- "Supercomputing Coming to a Closet Near You?" *HPCwire*, 5/31/02.
- "Smaller, Slower Supercomputers May Someday May Win The Race," *HPCwire*, 5/31/02.
- "Supercomputing Coming to a Closet Near You?" *PCworld.com*, 5/27/02.
- "Bell, Torvalds Usher Next Wave of Supercomputing," *HPCwire*, 5/24/02.
- "Supercomputing Cut Down to Size," *Personal Computer World*, 5/22/02.
- "Bell, Torvalds Usher Next Wave of Supercomputing," *CNN.com*, 5/21/02.
- "Transmeta's Low Power Finds Place in Supercomputers," *ZDNet News*, 5/20/02.
- "Transmeta Blades Power Landmark Supercomputer Breakthrough," *The Register*, 5/20/02.

SUPERCOMPUTING
in SMALL SPACES

http://sss.lanl.gov

Wu-chun (Wu) Feng
Research and Development in Advanced Network Technology

RADIANT

http://www.lanl.gov/radiant