Energy-Efficient E-puting Everywhere

Wu Feng

Dept. of Computer Science and Dept. of Electrical & Computer Engineering

Virginia Bioinformatics Institute

(Dept. of Cancer Biology and Translational Science Institute, Wake Forest U.)

VirginiaTech
Invent the Future

SyNeRG

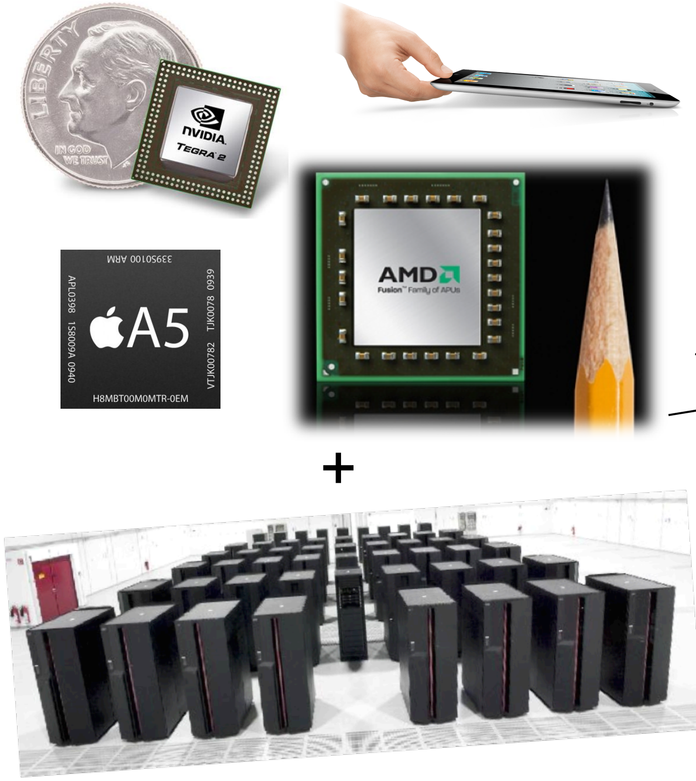
synergy.cs.vt.edu

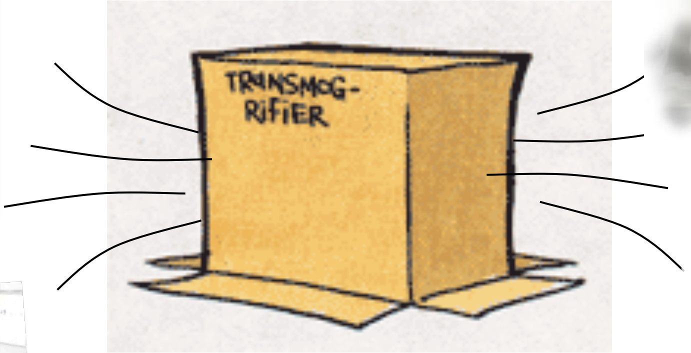
What is E-puting?

- Converged world of consumer ***e***lectronics and supercom***puting***

  … democratizing "supercomputing in small spaces"

  … via "trickle-up" technology





… with apologies to Calvin & Hobbes

What is "Trickle-Up" Technology?

- What is "trickle-down" technology?
  - Technology that initially is so expensive that only a small segment of the population can afford it, *but*

    … it will *trickle-down* the technology chain and

    … become inexpensive enough for the general public to afford
  - "If we can build terascale [petascale/exascale] supercomputers, we will be able to build smaller and more commodity systems that use the same basic technologies for the general public."
- "Trickle-up" technology
  - … will start with smaller and more commodity technology and *"trickle up"* into larger computing systems

SUPERCOMPUTING
in SMALL SPACES

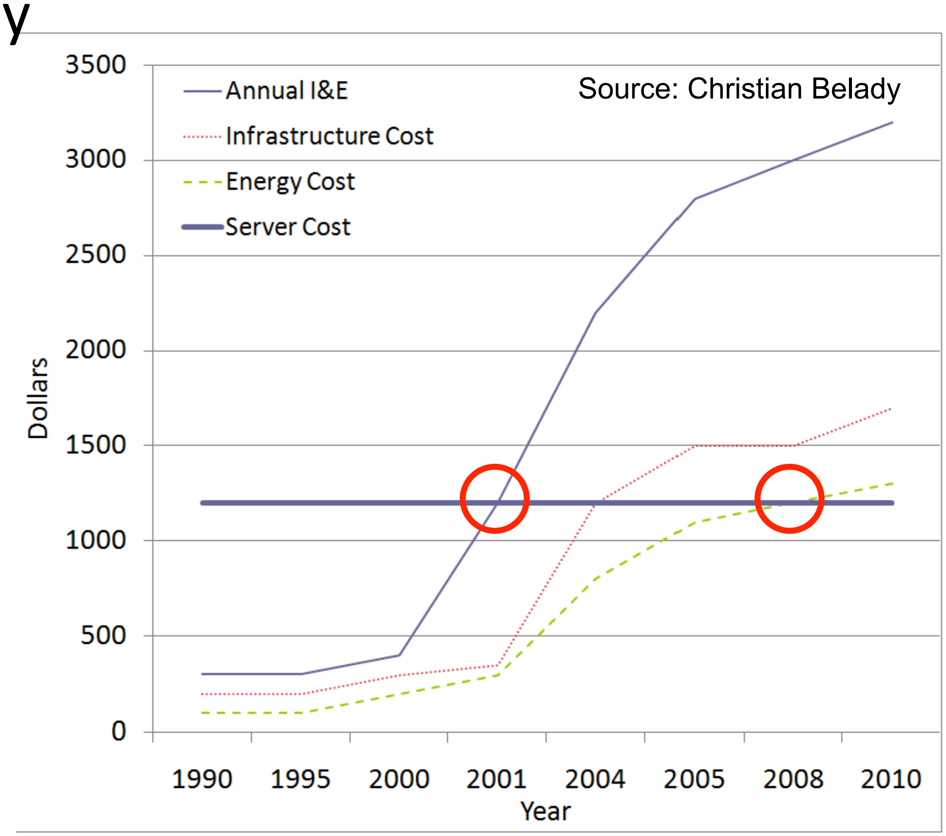
2001

2020

ON THE ROAD TO E-PUTING?

Electrical Power Costs $$$

- In 2001, the annual cost to provision a server
  … in a data center
  … with adequate power/energy
  exceeded the cost of
  the server itself.
- In 2008,
  the annual "Energy" cost for
  an energy-efficient 1U server
  … in a data center
  surpassed its purchase cost.



Source: Christian Belady

Hiding in Plain Sight, Google Seeks More Power

Melanie Conner for The New York Times

Google is building two computing centers, top and left, each the size of a football field, in The Dalles, Ore.

Source: *The New York Times*, June 14, 2006

Too Much Power

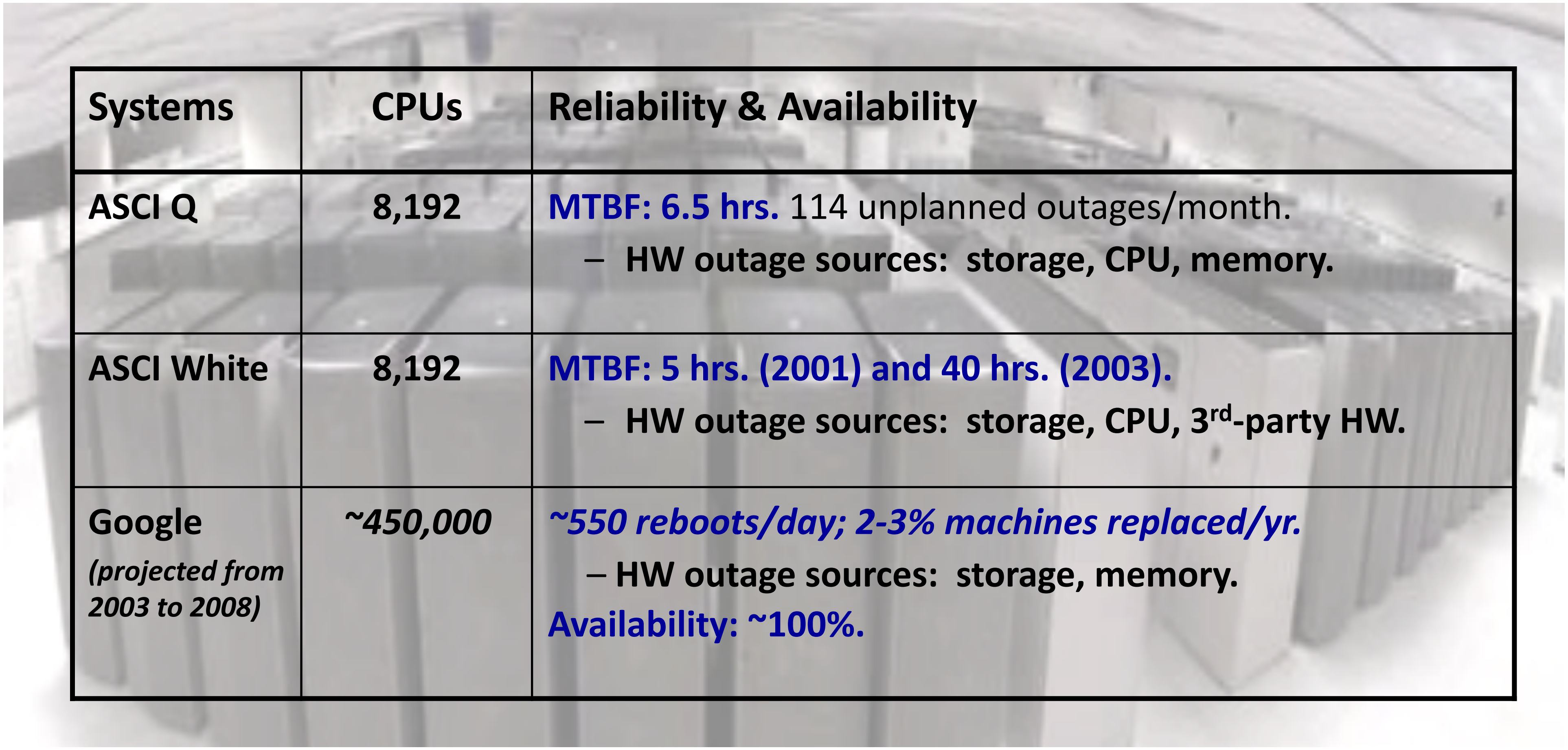
… affects efficiency, reliability, and availability.

- Anecdotal Evidence from a "Machine Room" in 2001 – 2002
  - **Winter**: "Machine Room" Temperature of **70-75° F**
    - Failure approximately ***once*** per week.
  - **Summer**: "Machine Room" Temperature of **85-90° F**
    - Failure approximately ***twice*** per week.
- Arrenhius' Equation (applied to microelectronics)
  - ***For every 10° C (18° F) increase in temperature,***

    … the failure rate of a system doubles.*

* W. Feng, M. Warren, and E. Weigle, "The Bladed Beowulf: A Cost-Effective Alternative to Traditional Beowulfs," *IEEE Cluster*, Sept. 2002.

Too Much Power?



Systems	CPUs	Reliability & Availability
ASCI Q	**8,192**	**MTBF: 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.**
Google ***(projected from 2003 to 2008)***	***~450,000***	***~550 reboots/day; 2-3% machines replaced/yr.*** – **HW outage sources: storage, memory.** **Availability: ~100%.**

Source: Daniel A. Reed

Supercomputing in Small Spaces (SSS)

Efficiency, Reliability, and Availability via Green HPC
(Started in 2001 at Los Alamos Nat'l Lab. Now at Virginia Tech.)

- Goal

  … with respect to space, power, and performance

  Improve efficiency, reliability, and availability (ERA) in supercomputing.

- Analogy

  Traditional Supercomputer vs. Supercomputing in Small Spaces

  - Formula One Race Car: Wins raw performance *but* is energy-inefficient and unreliable, thus requiring frequent "pit stops" and maintenance. Low throughput over the long haul.

  - Nissan 370Z: Loses raw performance but is more energy-efficient and reliable. High throughput (i.e., miles driven → answers/month).

Green Destiny Supercomputer

(circa December 2001 – February 2002)

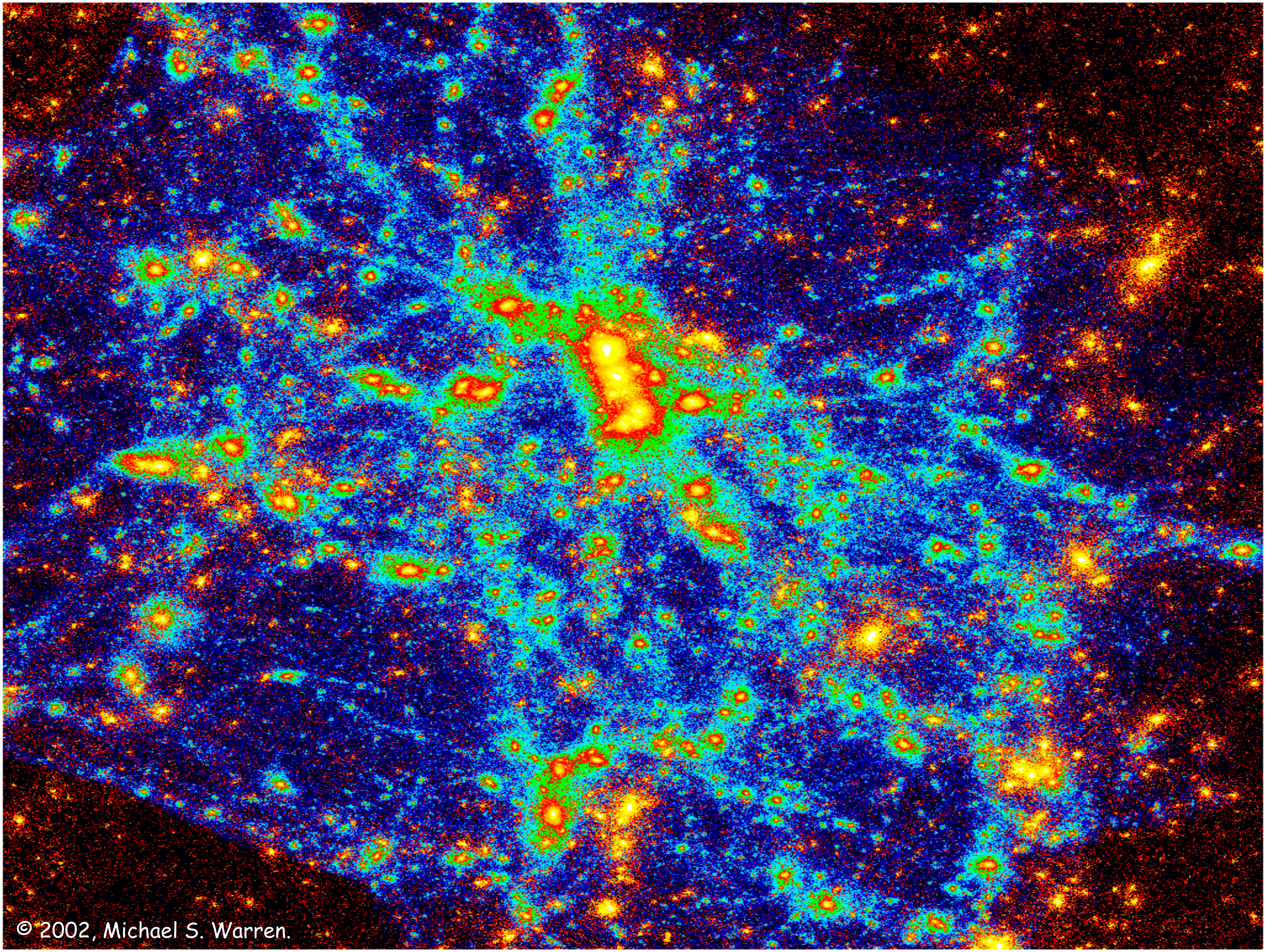
SUPERCOMPUTING in SMALL SPACES

- A 240-Node Cluster in Five Sq. Ft.
- Each Node
  - 1-GHz Transmeta TM5800 CPU w/ *High-Performance Code-Morphing Software* running Linux 2.4.x
    - CPU Power Consumption? **Only 6 watts!**
  - 640-MB RAM, 20-GB hard disk, 100-Mb/s Ethernet
- Total
  - 240 Gflops peak ***(Linpack: 101 Gflops in March 2002.)***
  - ***Power Consumption: Only 3.2 kW (diskless)***
- Reliability & Availability
  - ***No unscheduled downtime in its 24-month lifetime***
    - **Environment: A dusty 85°-90° F warehouse**

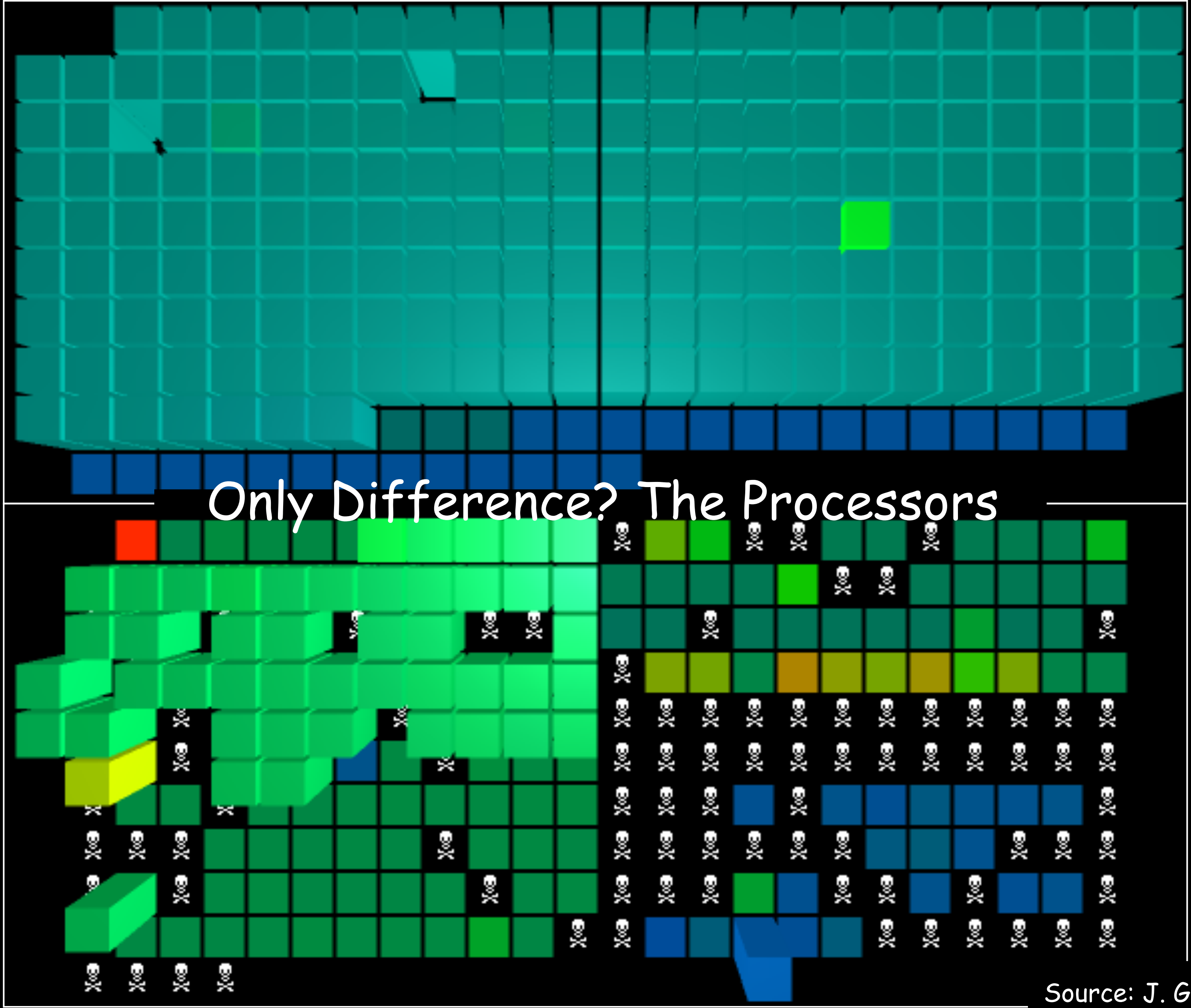
Equivalent Linpack to a 256-CPU SGI Origin 2000 (On TOP500 List at the time)

Featured in *The New York Times*, *BBC News*, and *CNN*. Now in the *Computer History Museum*.

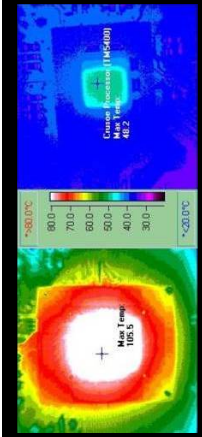
R&D 100



Green Destiny: Low-Power Supercomputer



Only Difference? The Processors



Source: J. Gans, Mar. 2007

Green Destiny "Replica": Traditional Supercomputer

A Perspective from 2001 – 2006:

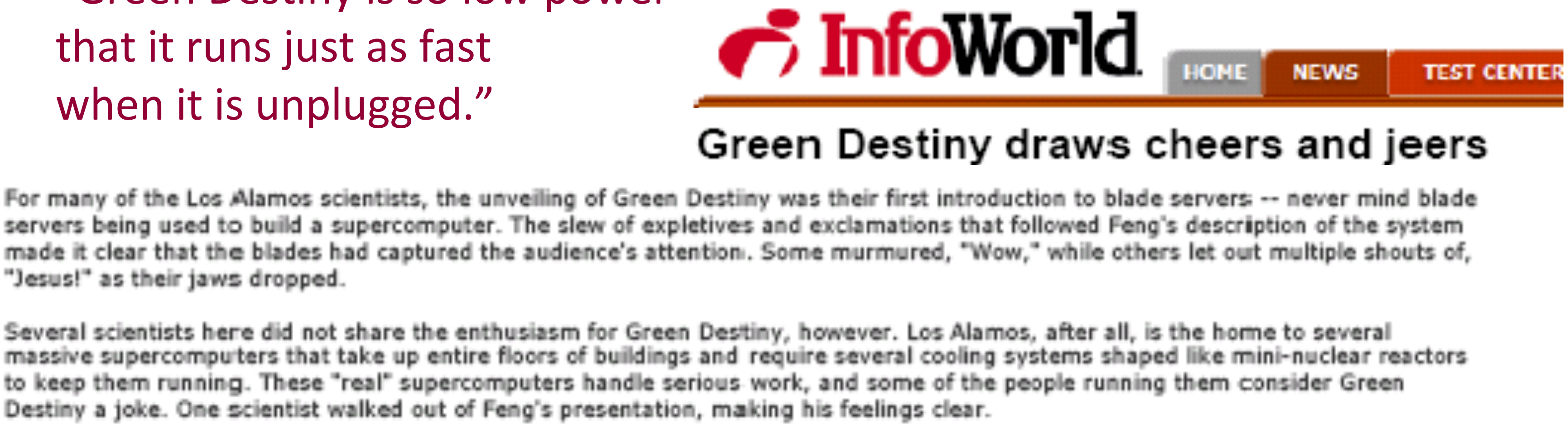
Supercomputing in Small Spaces

SUPERCOMPUTING in SMALL SPACES

- The Way of the Future?

  … or Not?

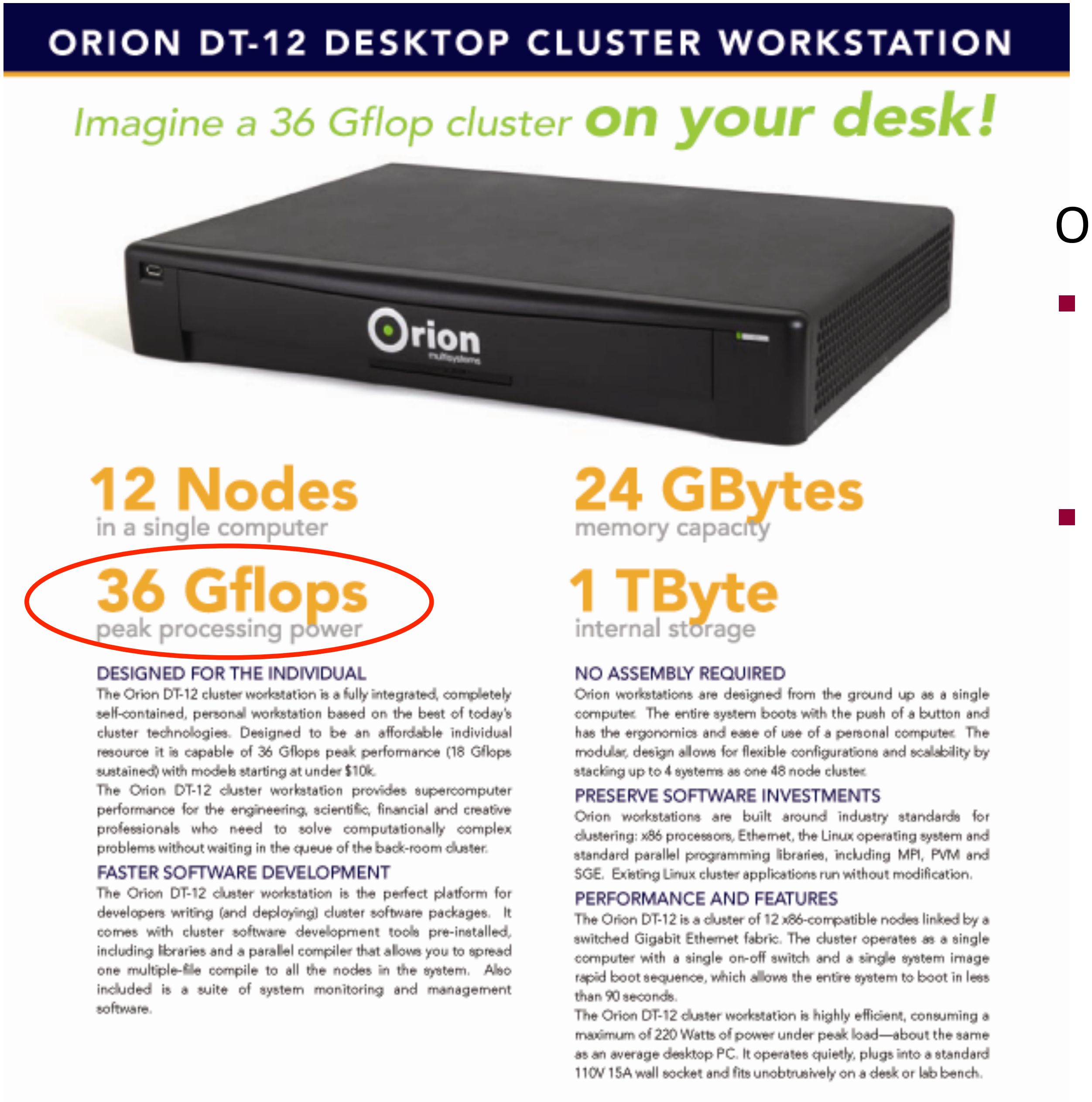
  - “In high-performance computing, no one cares about power & cooling, and no one ever will …”
  - “Moore’s Law for Power will stimulate the economy by creating a new market in cooling technologies.”
  - “Green Destiny is so low power that it runs just as fast when it is unplugged.”



Green Destiny draws cheers and jeers

For many of the Los Alamos scientists, the unveiling of Green Destiny was their first introduction to blade servers -- never mind blade servers being used to build a supercomputer. The slew of expletives and exclamations that followed Feng's description of the system made it clear that the blades had captured the audience's attention. Some murmured, "Wow," while others let out multiple shouts of, "Jesus!" as their jaws dropped.

Several scientists here did not share the enthusiasm for Green Destiny, however. Los Alamos, after all, is the home to several massive supercomputers that take up entire floors of buildings and require several cooling systems shaped like mini-nuclear reactors to keep them running. These "real" supercomputers handle serious work, and some of the people running them consider Green Destiny a joke. One scientist walked out of Feng's presentation, making his feelings clear.



ORION DT-12 DESKTOP CLUSTER WORKSTATION

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

Orion
multisystems

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.

(Circa 2004)

Orion DT-12

- Footprint
  - 3 sq. ft. (24" x 18")
  - 1 cu. ft. (24" x 4" x 18")
- Power Consumption
  - 170 watts at load

👍 Power Efficient

✋ Performance/Core

👍 Price

👎 Proprietary Hardware (Limited Trickle-Down)

SiCortex SC 648 and SC 5832

(Circa 2006)

Sources: SiCortex, Google, and BigNComputing

CPU Power: 0.6 W

SiCortex SC 648 (648 Gflops peak)

- 2 kW for 648-CPU system

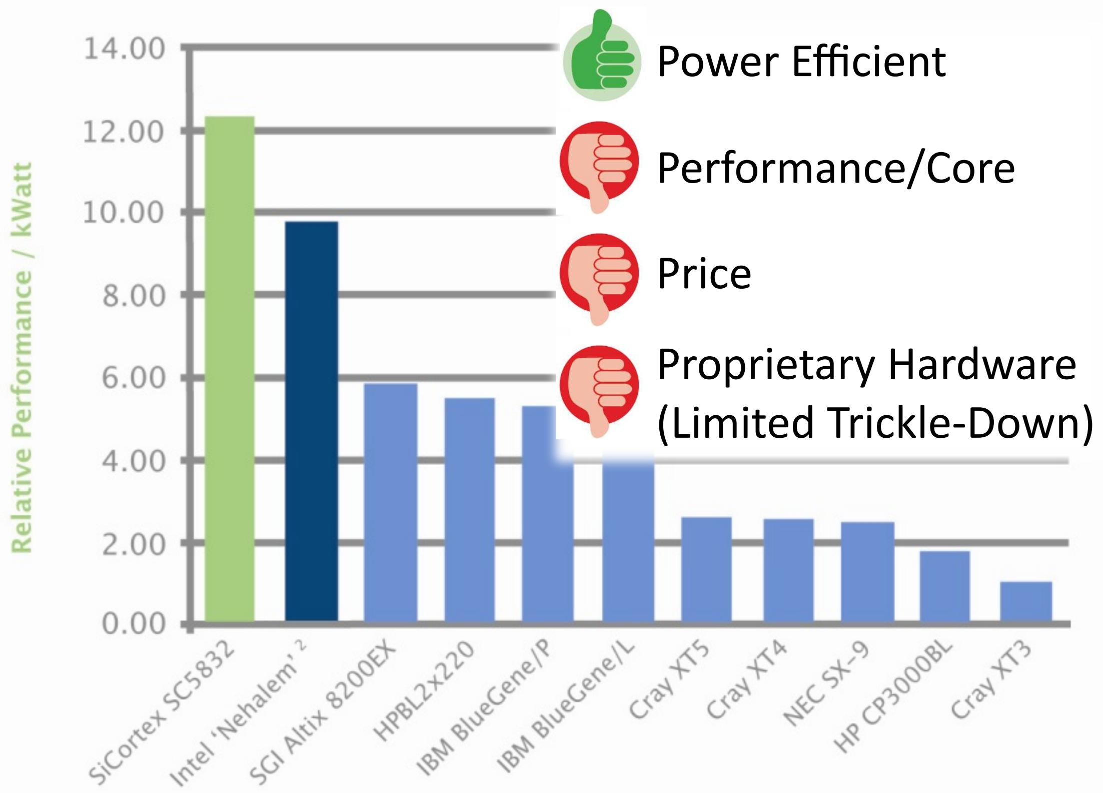
SiCortex SC 5832 (5.8 Tflops peak)

- 18 kW for 5832-CPU system

Green Computing Performance Index (GCPI)

Sample Comparisons of Leading HPC Systems

GCPI = (System Performance) / kWatt[1]



👍 Power Efficient

👎 Performance/Core

👎 Price

👎 Proprietary Hardware (Limited Trickle-Down)

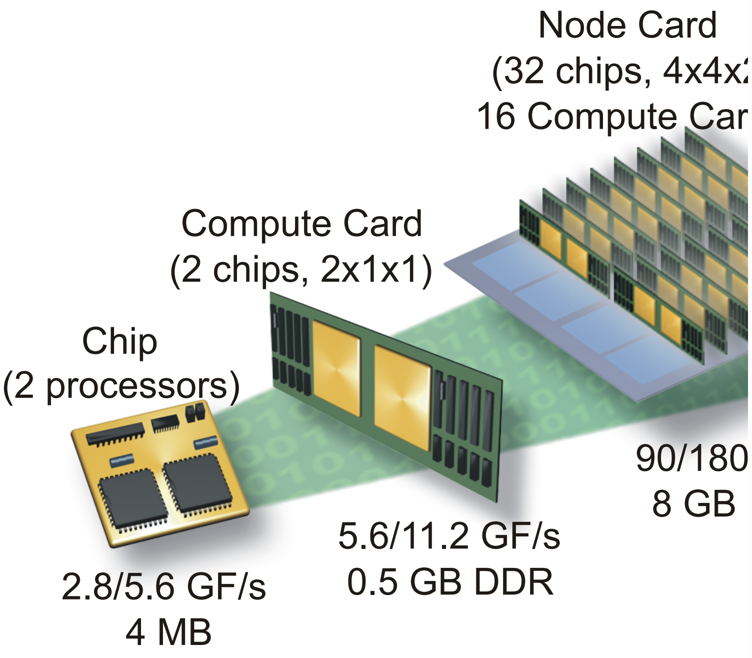
[1] GCPI = n(HPCC results)/kWatt, where n = results normalized to Cray XT3 reference system. HPCC is an industry-standard benchmark suite comprising 7 tests and a total of 28 benchmarks.

[2] Intel 'Nehalem' GCPI results estimated from posted HPCC benchmark results (30 March 09) for the Intel Endeavor Xeon 5560 and derived energy consumption.

IBM Blue Gene/L

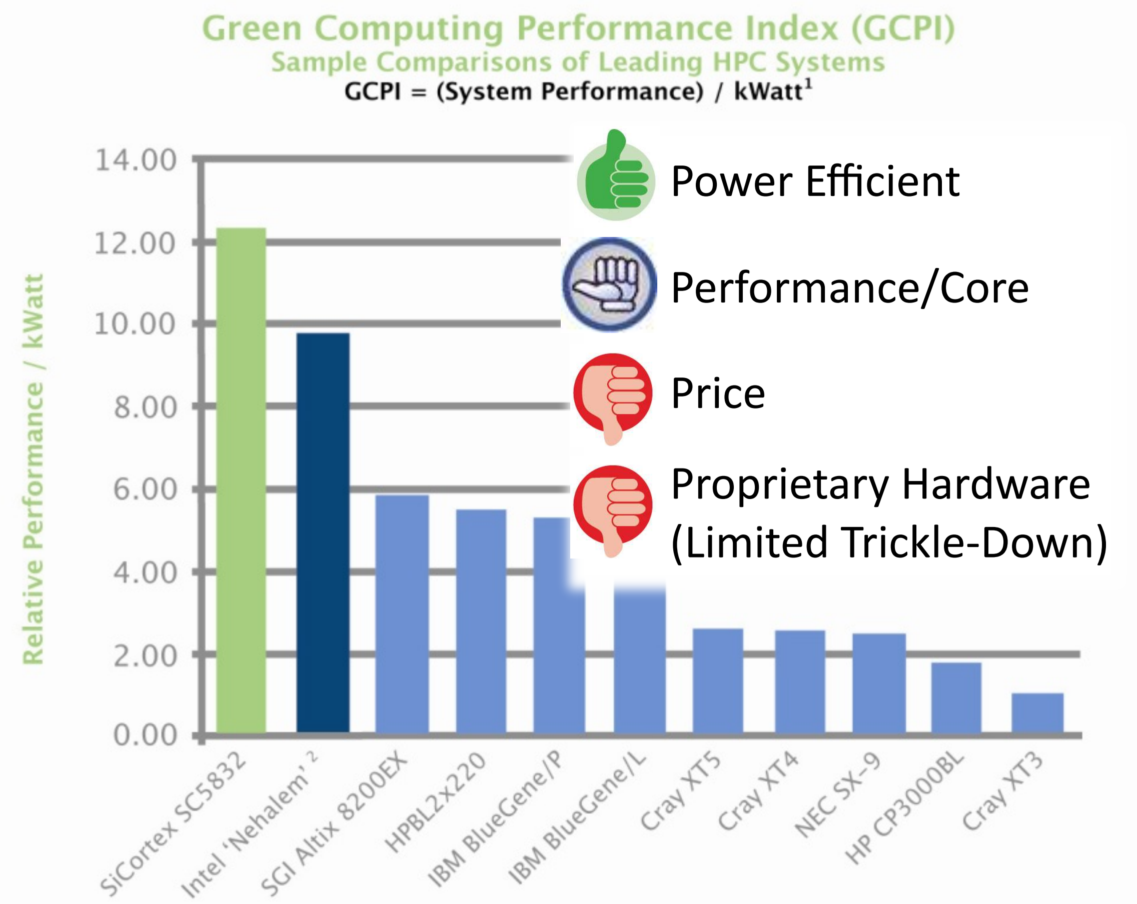
(Circa 2004)

Debuts in TOP500 List, November 2004



Each processor consumes 15 watts.

© 2004 IBM Corporation



👍 Power Efficient

Performance/Core

👎 Price

👎 Proprietary Hardware (Limited Trickle-Down)

[1] GCPI = n(HPCC results)/kWatt, where n = results normalized to Cray XT3 reference system. HPCC is an industry-standard benchmark suite comprising 7 tests and a total of 28 benchmarks.

[2] Intel 'Nehalem' GCPI results estimated from posted HPCC benchmark results (30 March 09) for the Intel Endeavor Xeon 5560 and derived energy consumption.

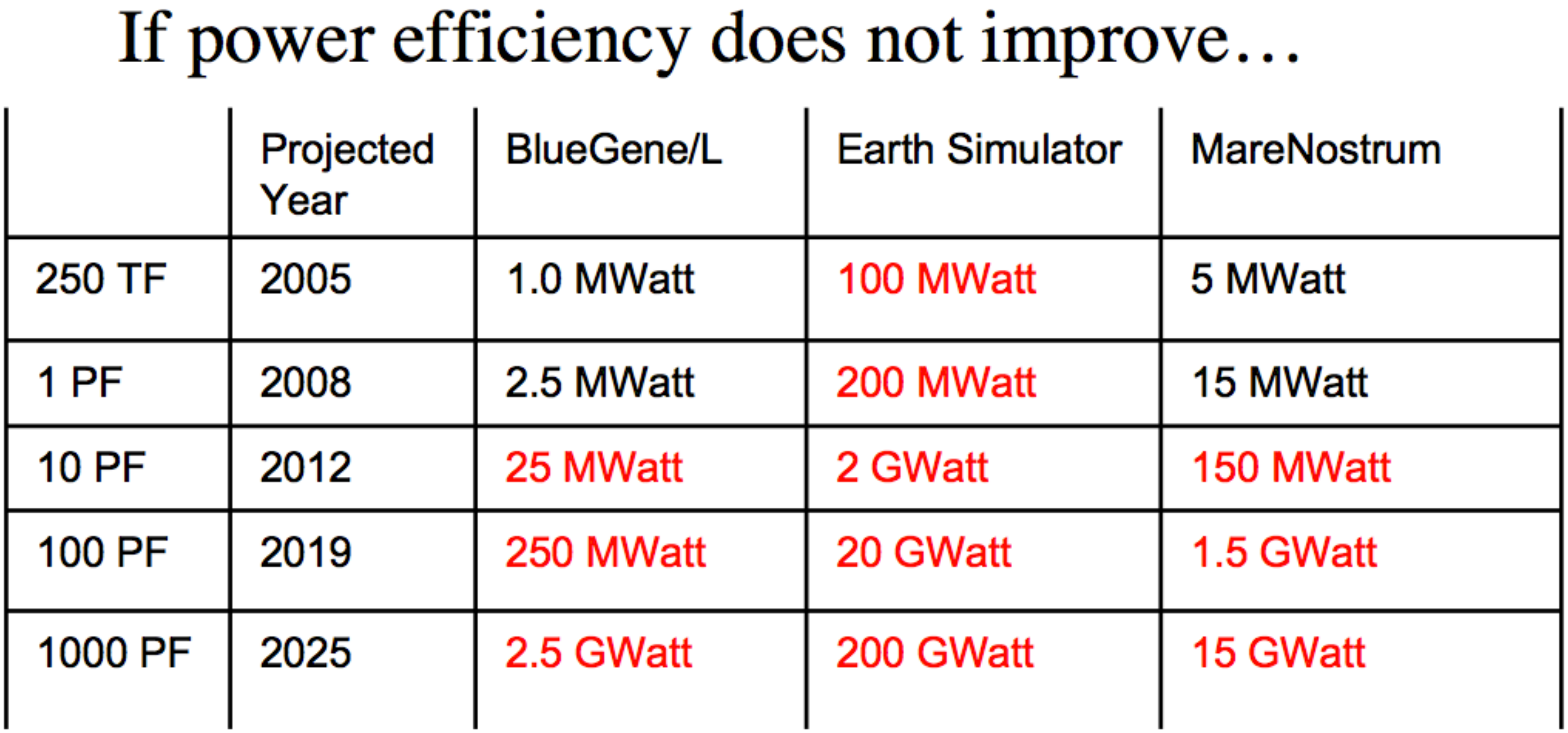
Outline

- Motivation
- "Supercomputing in Small Spaces" Project
- The Future of Energy-Efficient Computing?
- Energy-Efficient E-puting
  - Convergence of consumer electronics and supercomputing
  - "Trickle-up" technology
  - Multi-dimensional optimization
- Conclusion

SUPERCOMPUTING in SMALL SPACES

The Future of Energy-Efficient Computing?

- Is the need to be green enough?



If power efficiency does not improve…

	Projected Year	BlueGene/L	Earth Simulator	MareNostrum
250 TF	2005	1.0 MWatt	100 MWatt	5 MWatt
1 PF	2008	2.5 MWatt	200 MWatt	15 MWatt
10 PF	2012	25 MWatt	2 GWatt	150 MWatt
100 PF	2019	250 MWatt	20 GWatt	1.5 GWatt
1000 PF	2025	2.5 GWatt	200 GWatt	15 GWatt

Source: Alan Gara, “Blue Gene: The Next Generation Supercomputer,” 2007.

The Future of Energy-Efficient Computing?

- Are the efforts below enough?

Orion Multisystems	SiCortex	IBM BlueGene
👍 Power Efficient	👍 Power Efficient	👍 Power Efficient
✋ Performance/Core	👎 Performance/Core	✋ Performance/Core
👍 Price	👎 Price	👎 Price
👎 Proprietary Hardware (Limited Trickle-Down)	👎 Proprietary Hardware (Limited Trickle-Down)	👎 Proprietary Hardware (Limited Trickle-Down)

No.

… not economically sustainable

… the trickle-down effect is limited

Charting a New Course:
Energy-Efficient E-puting Everywhere

- Features
  - Convergence of consumer ***e***lectronics and supercom***puting***
  - "Trickle-up" technology
  - Multi-dimensional optimization
    - Performance → Sequential & Parallel
    - Power
    - Proprietariness → Commoditization
    - Portability
    - Processor Heterogeneity
    - Price

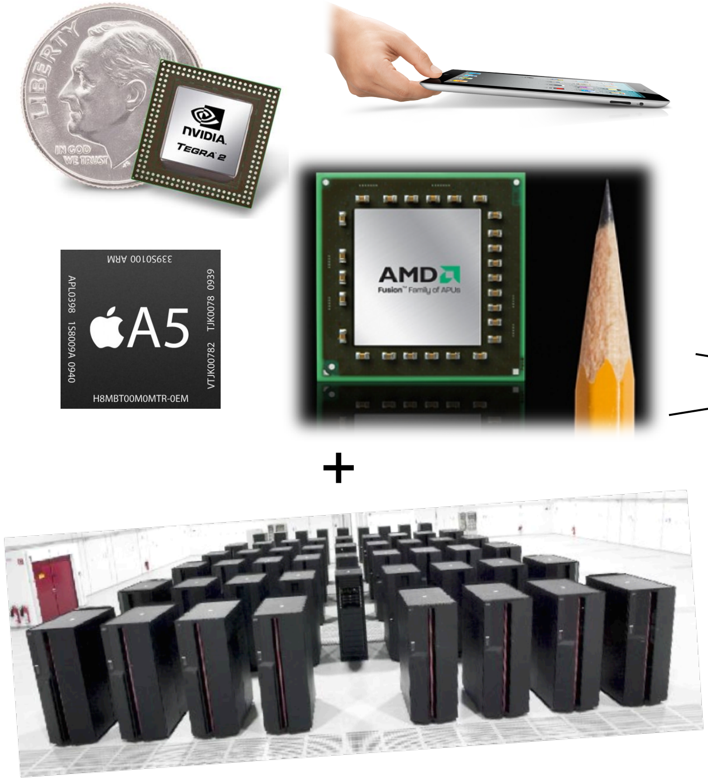
Hardware-Software
Co-Design

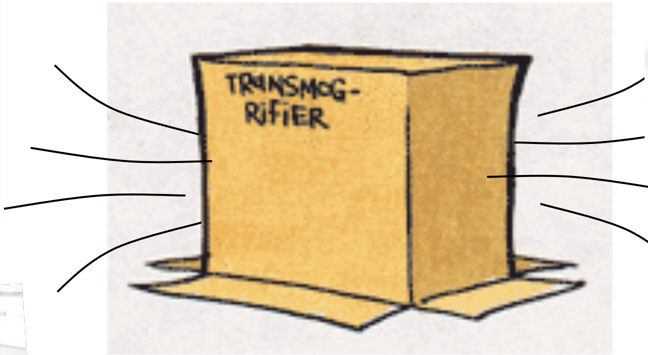
E-puting is a …

- Converged world of consumer ***e***lectronics and supercom***puting***

  … democratizing “supercomputing in small spaces”

  … via “trickle-up” technology





… with apologies to Calvin & Hobbes

What is "Trickle-Up" Technology?

- What is "trickle-down" technology?
  - Technology that initially is so expensive that only a small segment of the population can afford it, *but*

    … it will *trickle-down* the technology chain and

    … become inexpensive enough for the general public to afford
  - "If we can build terascale [petascale/exascale] supercomputers, we will be able to build smaller and more commodity systems that use the same basic technologies for the general public."
- "Trickle-up" technology
  - … will start with smaller and more commodity technology and *"trickle up"* into larger computing systems

Multi-Dimensional Optimization

Hardware-Software
Co-Design

Example: Sequential Performance and Power

Linear Programming for an Energy-Optimal DVFS Schedule

(DVFS = Dynamic Voltage & Frequency Scaling)

- Definitions
  - A DVFS system exports *n* { (f_i, P_i) } settings.
  - T_i : total execution time of a program running at setting *i*
- Given a program with deadline *D*, find a DVFS schedule (t_1*, …, t_n*) such that
  - If the program is executed for t_i seconds at setting *i*, the total energy usage E is minimized, the deadline D is met, and the required work is completed

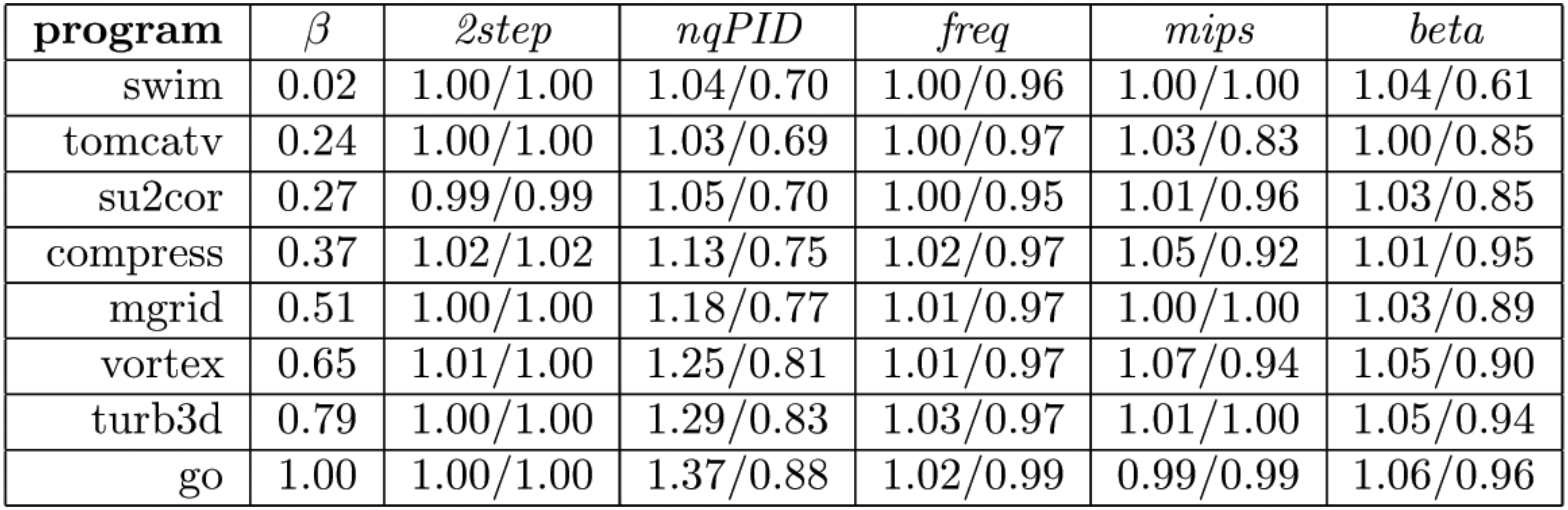
$$\min E = \sum_i P_i \cdot t_i$$

subject to

$$\sum_i t_i \le D$$
$$\sum_i t_i / T_i = 1$$
$$t_i \ge 0$$

Embrace the power wall
… select the *right* setting
… at the *right* time
for the workload at hand

β – Adaptation with Sequential Codes (SPEC CPU)



program	β	*2step*	*nqPID*	*freq*	*mips*	*beta*
swim	0.02	1.00/1.00	1.04/0.70	1.00/0.96	1.00/1.00	1.04/0.61
tomcatv	0.24	1.00/1.00	1.03/0.69	1.00/0.97	1.03/0.83	1.00/0.85
su2cor	0.27	0.99/0.99	1.05/0.70	1.00/0.95	1.01/0.96	1.03/0.85
compress	0.37	1.02/1.02	1.13/0.75	1.02/0.97	1.05/0.92	1.01/0.95
mgrid	0.51	1.00/1.00	1.18/0.77	1.01/0.97	1.00/1.00	1.03/0.89
vortex	0.65	1.01/1.00	1.25/0.81	1.01/0.97	1.07/0.94	1.05/0.90
turb3d	0.79	1.00/1.00	1.29/0.83	1.03/0.97	1.01/1.00	1.05/0.94
go	1.00	1.00/1.00	1.37/0.88	1.02/0.99	0.99/0.99	1.06/0.96

relative time / relative energy
with respect to total execution time and system energy usage

SMALLER numbers are BETTER.

β – Adaptation with Sequential Codes (SPECjbb)

Power Management	Watts	% Power Reduction
None	264	0%
Cpuspeed	257	3%
Ondemand	253	4%
β	196	25%

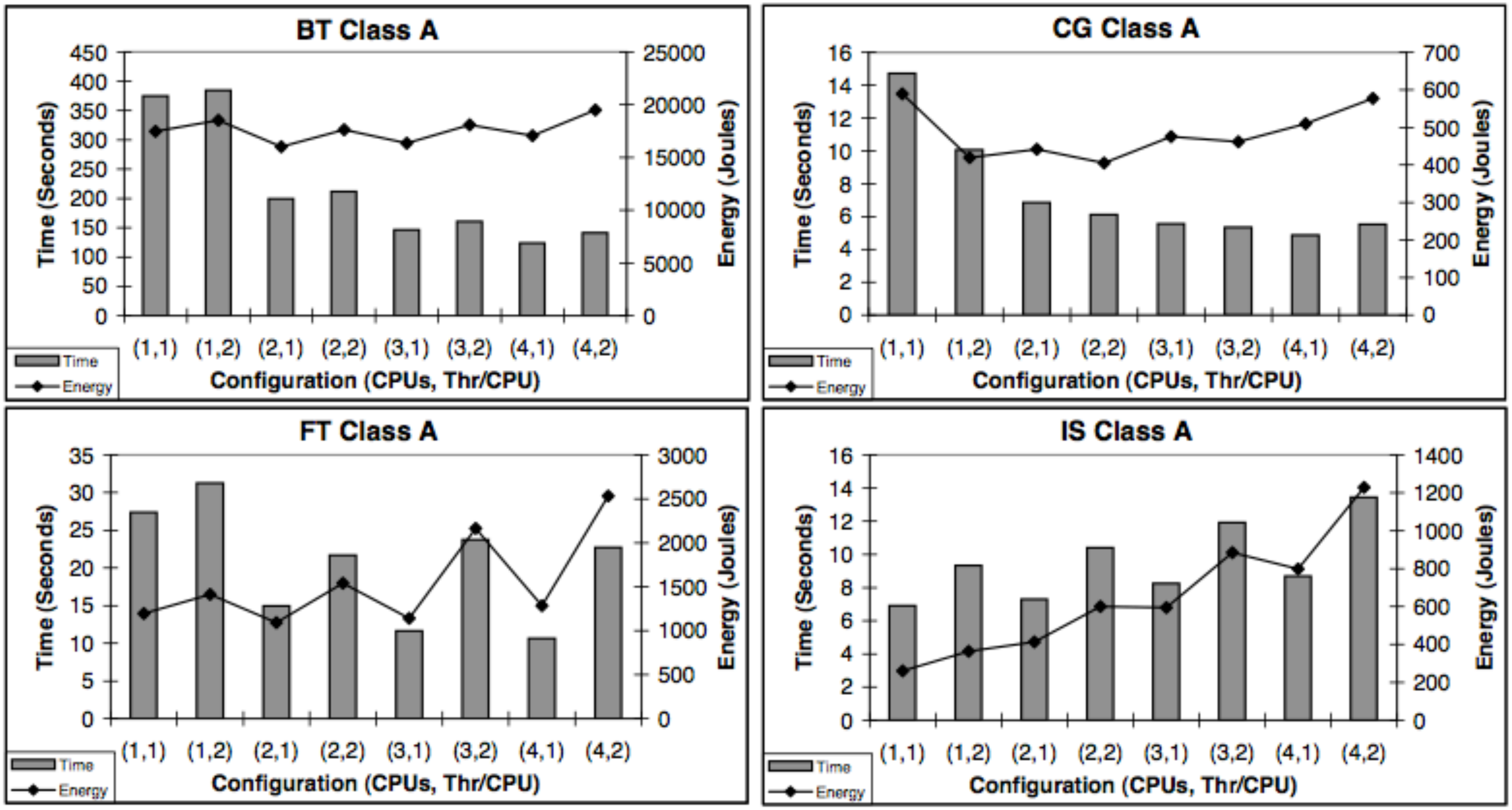
Power Management	bops/watt
None	100.00%
Cpuspeed	102.56%
Ondemand	104.37%
β	123.70%

Preliminary Results in the Commercial Sector:

β − Adaptation with Embedded Mobile Device

- Workload
  - Interactive (internal)
- Performance
  - < 2% slowdown
- Power
  - 50% reduction
- Performance/Power
  - 2x improvement

Need for Hardware-Software Co-Design (*a la* β)



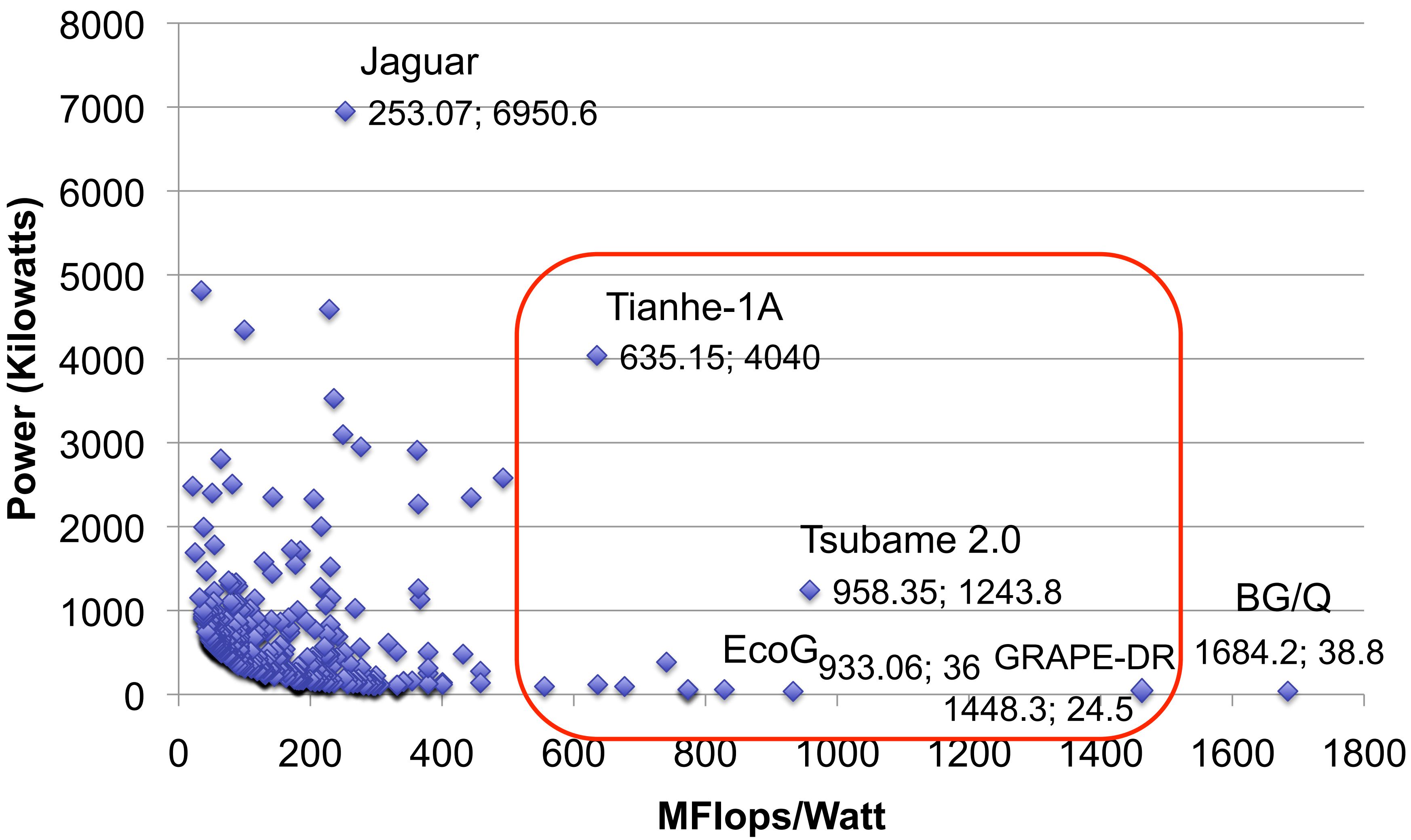
Source: Virginia Tech

Charting a New Course:
Energy-Efficient E-puting Everywhere

- Features
  - Convergence of consumer ***e***lectronics and supercom***puting***
  - “Trickle-up” technology
  - Multi-dimensional optimization
    - Performance → Sequential & Parallel
    - Power
    - Proprietariness → Commoditization
    - Portability
    - Processor Heterogeneity
    - Price

Hardware-Software
Co-Design

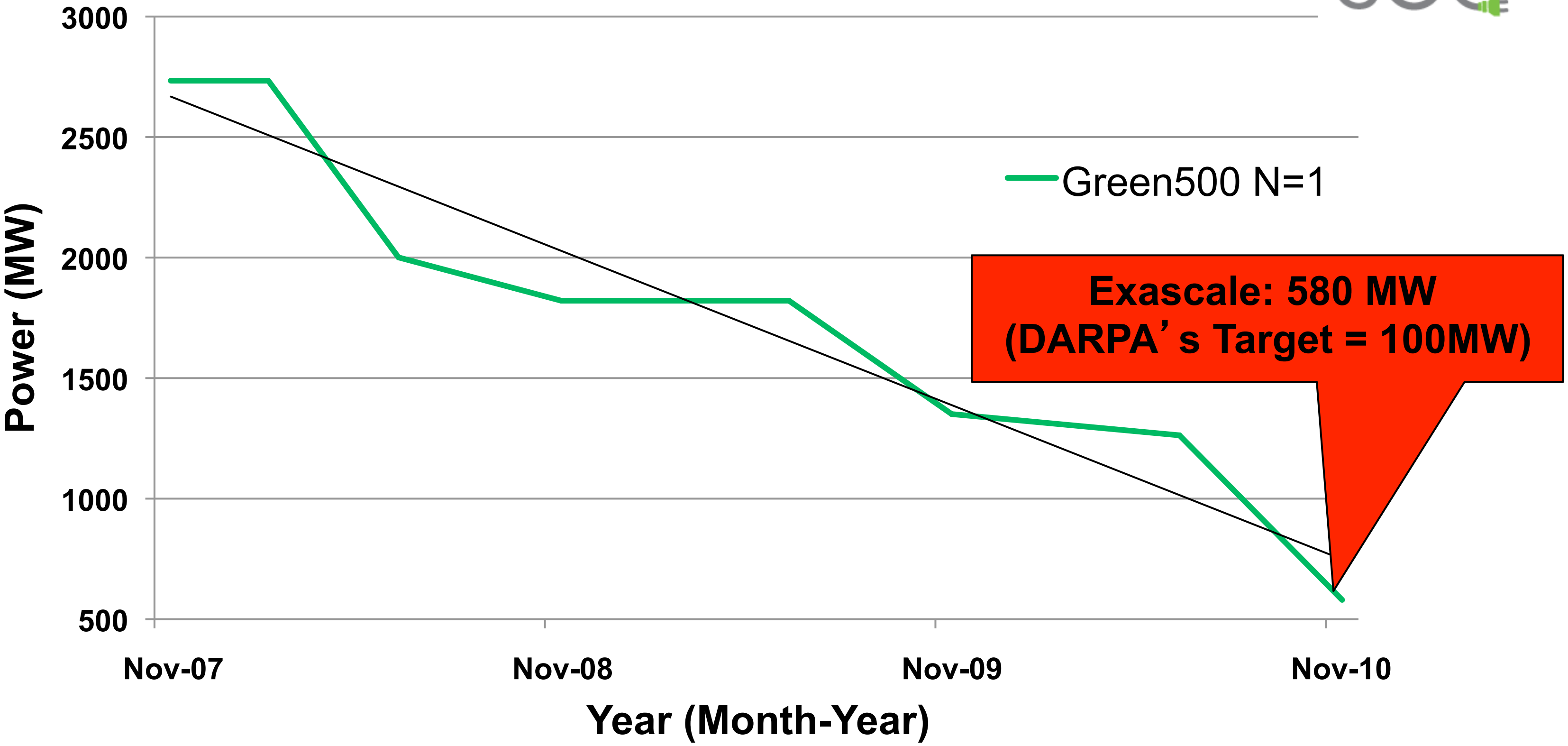
Towards Energy-Efficient E-puting in HPC



Towards Energy-Efficient E-puting in HPC

- What would the power consumption be of the greenest supercomputer extrapolated to an exascale machine?

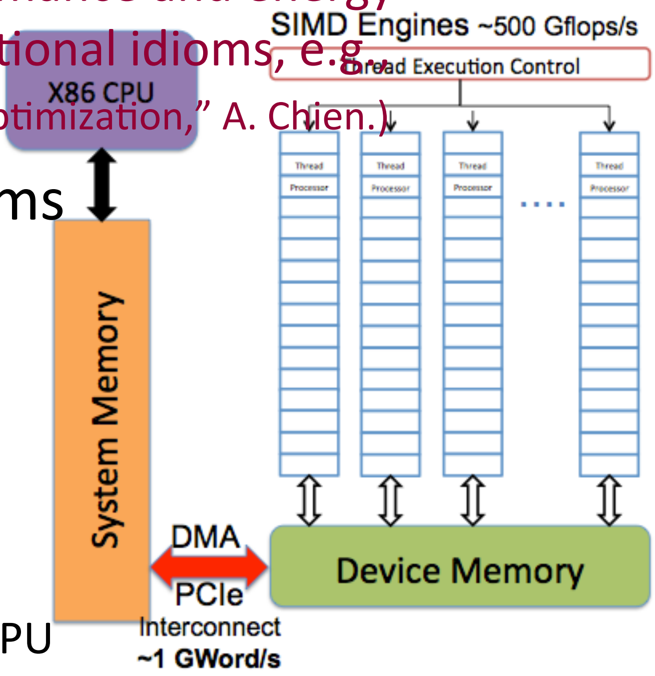
Source: THE GREEN 500



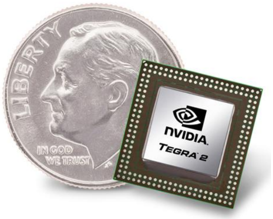
Exascale: 580 MW
(DARPA's Target = 100MW)

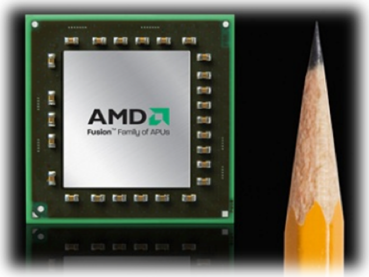
Charting a New Course:
Energy-Efficient E-puting Everywhere

- For a given task, select the right core (i.e., “tool”) at the right settings (e.g., degree of parallelism, voltage & frequency) at the right time.
  - Each core could be designed for high performance and energy efficiency for each of the different computational idioms, e.g., Berkeley dwarfs. (Example: “The Rise of 10x10 Optimization,” A. Chien.)
- Hints of the above with CPU+GPU systems
  - General-purpose cores → CPU
  - Data-parallel/task-parallel cores → GPU
    - Reduced overhead
    - Explicitly hidden memory latency
    - Simplified control
    - Problem: Data movement between CPU & GPU



Simpler CPU Cores & GPU Cores





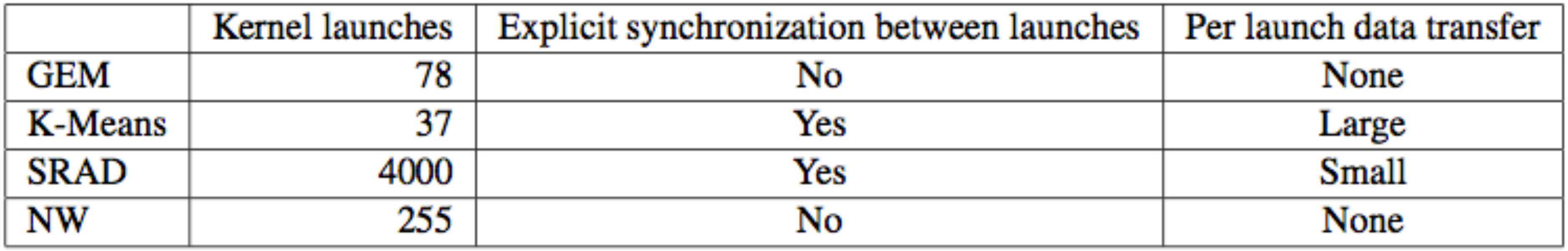
- Simpler cores enable use of slower clock rates, resulting in cubic drop in power due to $V^2 * f$

- Simpler cores use less area and produce lower leakage power

- Simpler cores place more burden on the programmer
  - Need better languages and tools to *express* massive parallelism.
  - Need better system software and run-time systems to *manage* massive parallelism.

Machines and Workload

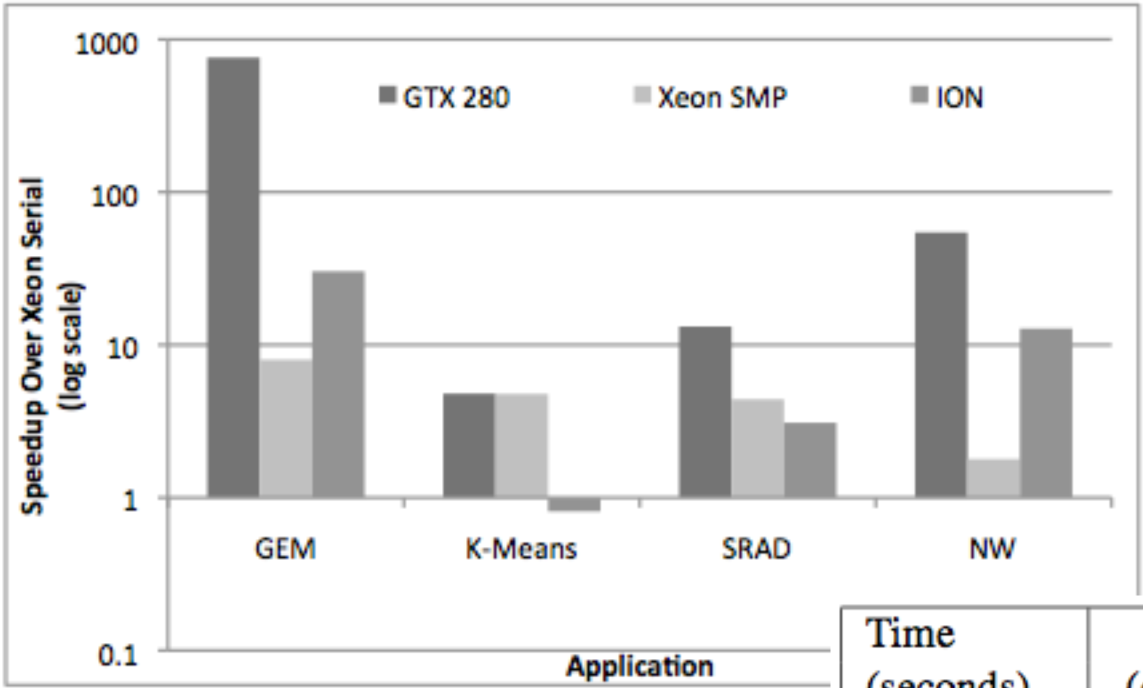
- Two 2.0-GHz Intel Xeon E5405 quad-core CPUs w/ 4GB RAM
  - NVIDIA *GTX 280* GPU
- One 1.6-GHz Intel Atom 230 dual-core CPU, 3GB RAM, and NVIDIA MCP79 chipset w/ an integrated *ION* graphics chip with 256MB of graphics memory and 2 multiprocessors (16 stream cores) at compute capability 1.1 and a clock rate of 1.1 GHz.



	Kernel launches	Explicit synchronization between launches	Per launch data transfer
GEM	78	No	None
K-Means	37	Yes	Large
SRAD	4000	Yes	Small
NW	255	No	None

Performance:

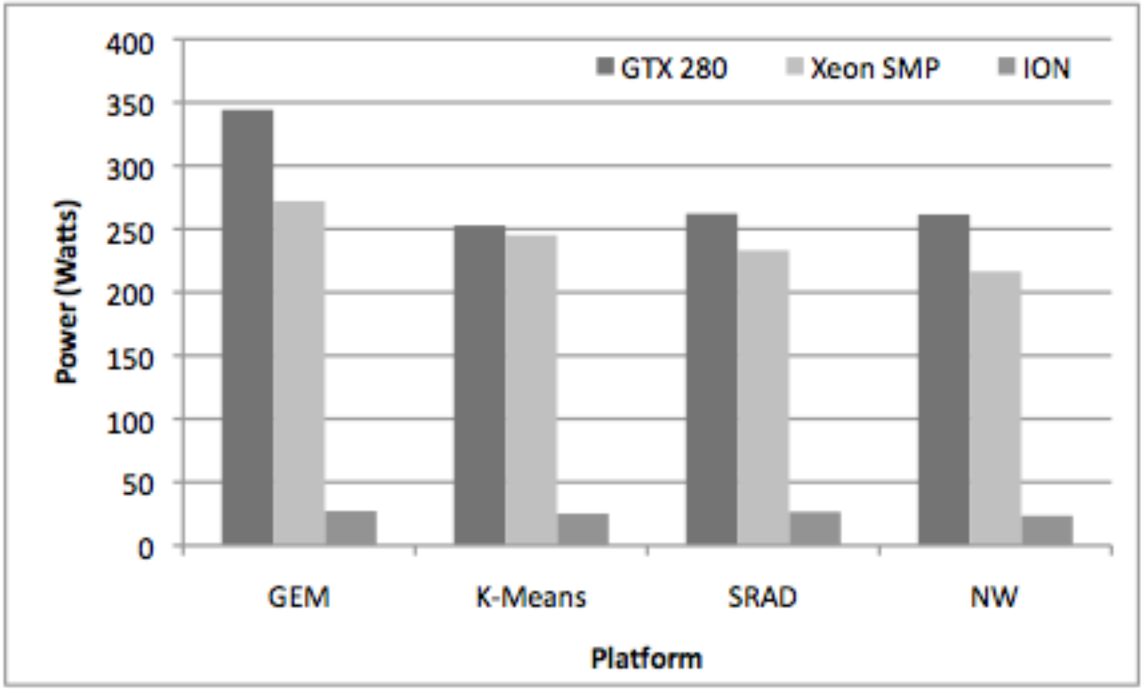
Integrated “Low-Power CPU + GPU” MCP



Time (seconds)	GEM (capsid)	K-Means	SRAD	NW
Xeon serial	63,029.5	7.9	788.5	377.0
Xeon SMP	7,878.7	1.7	179.0	210.5
GTX 280	82.9	1.7	59.8	6.9
ION	1,998.5	9.7	254.9	29.5

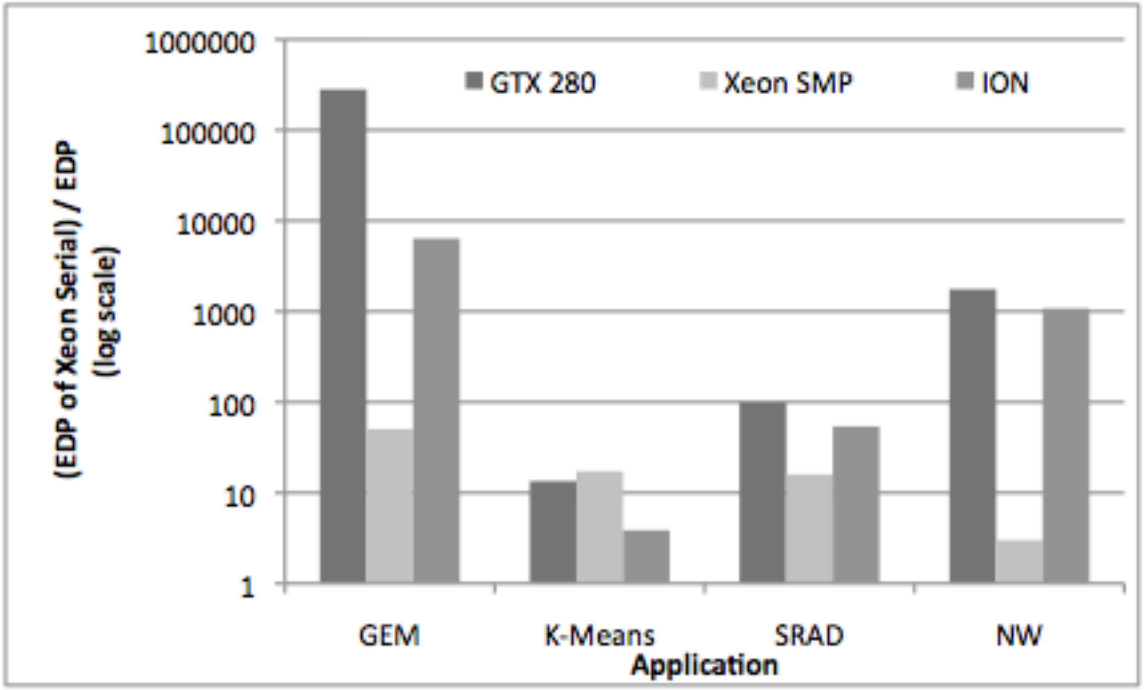
Power:

Integrated “Low-Power CPU + GPU” MCP

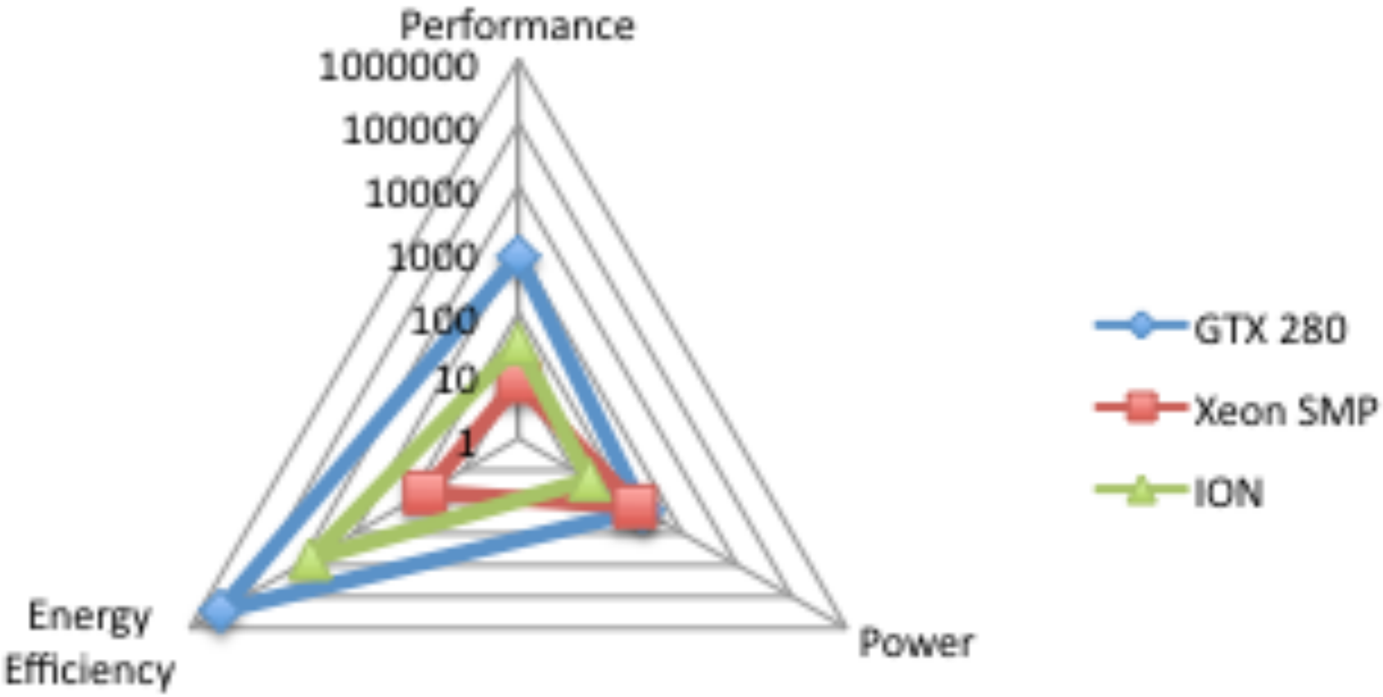


Energy Efficiency:

Integrated "Low-Power CPU + GPU"



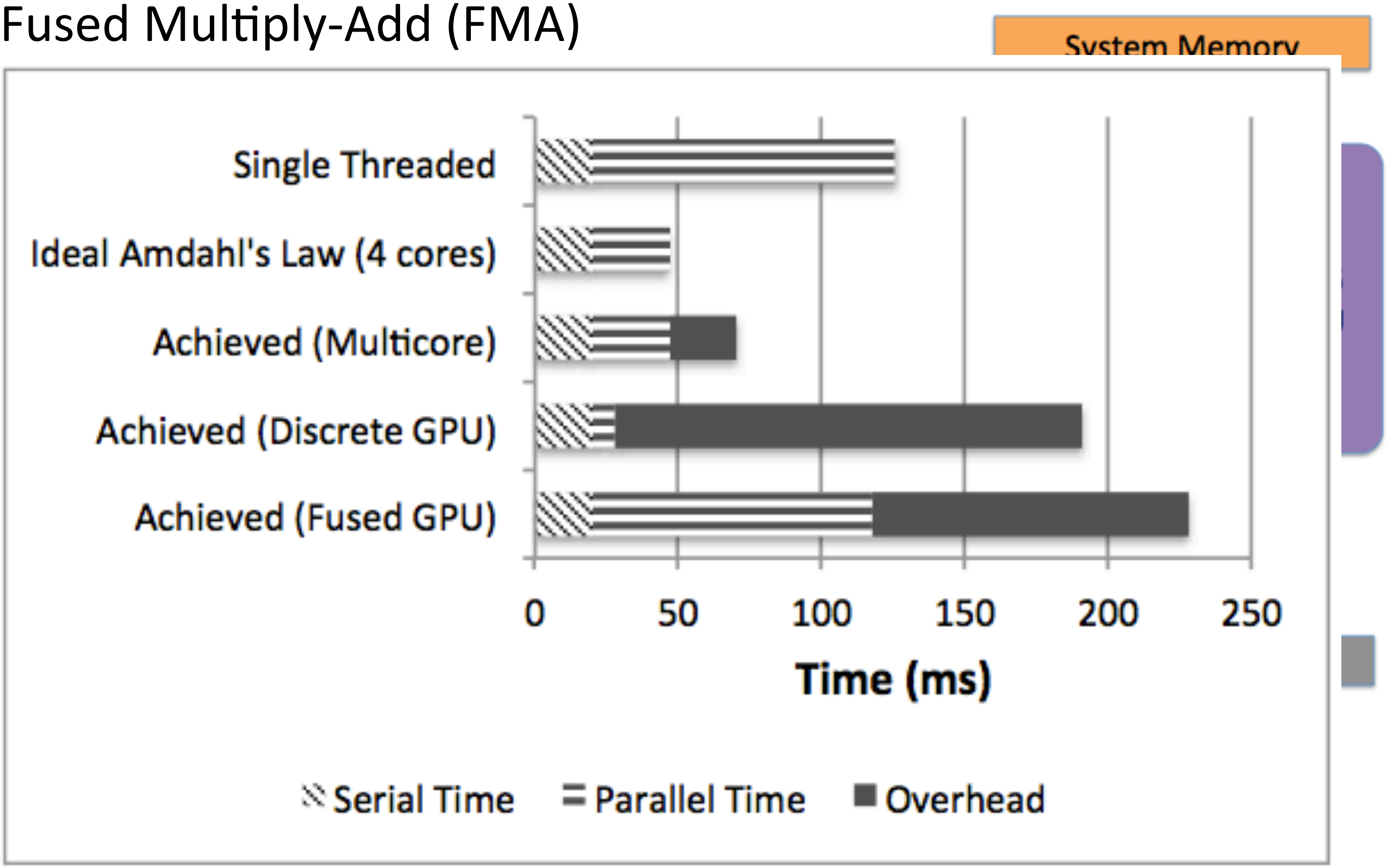
Kiviat Diagram:
Performance, Power, and Energy Efficiency



Early Results:

Energy-Efficient E-puting on Fused CPU+GPU Laptop

- Fused Multiply-Add (FMA)



Re-visiting Amdahl’s Law

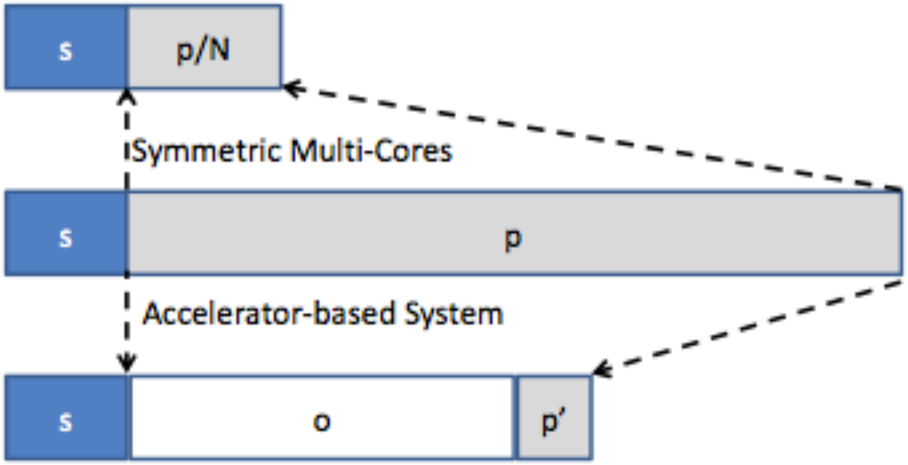
$$S = \frac{1}{s + p/N}$$

⇩

$$S' = \frac{1}{s + o + p'}$$

$$\text{where, } o = \text{parallel overhead}$$

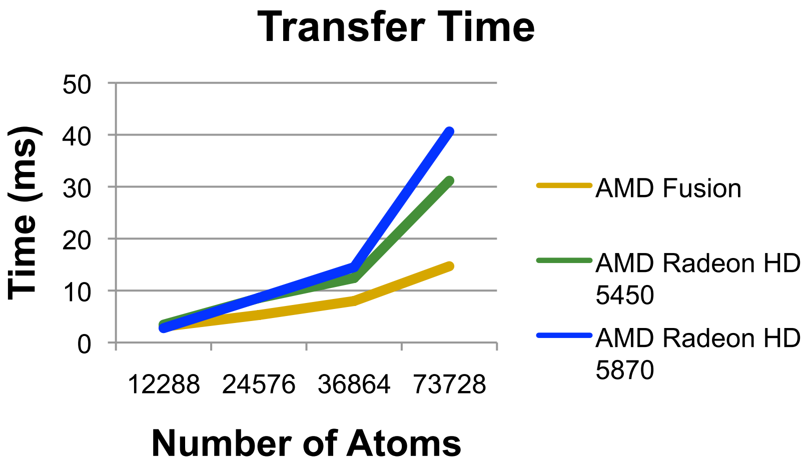
$$p' = \text{accelerated parallel fraction}$$

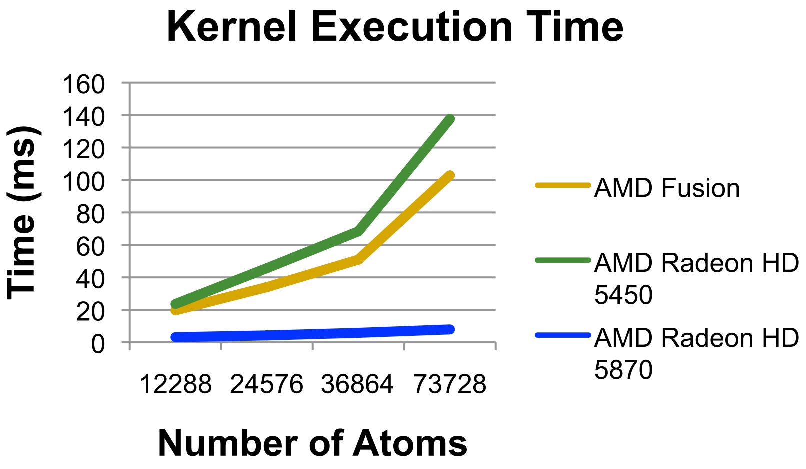


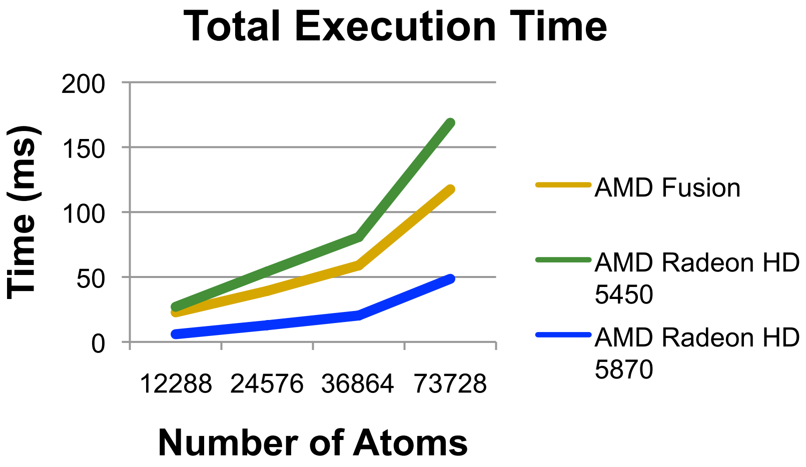
Source: M. Daga, A. Aji, W. Feng, “On the Efficacy of a Fused CPU+GPU Processor for Parallel Computing,” SAAHPC ’11, July 2011 (to appear)

Performance: Molecular Dynamics (N-Body)

Compute-bound



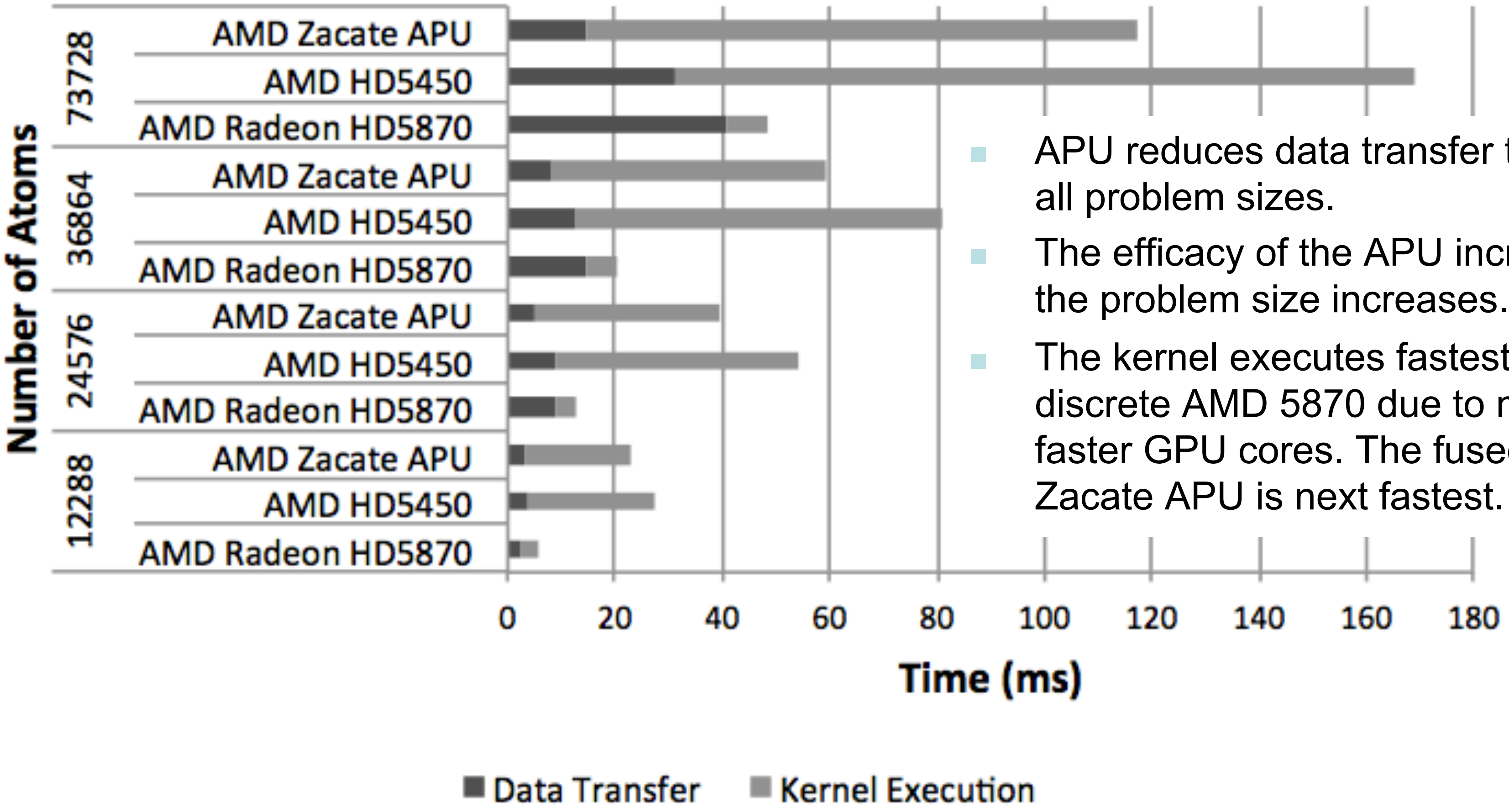




- APU reduces data transfer times for all problem sizes.
- The efficacy of the APU increases as the problem size increases.
- The kernel executes fastest on discrete AMD 5870 due to more and faster GPU cores. The fused Zacate APU is next fastest.

Performance: Molecular Dynamics (N-Body)

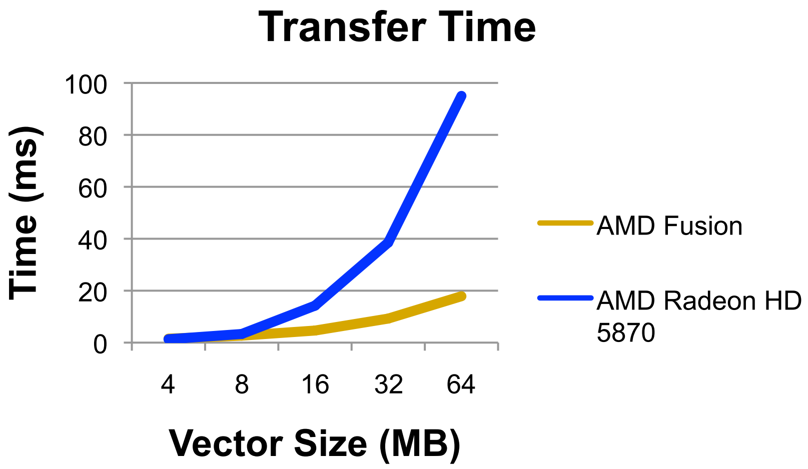
Compute-bound

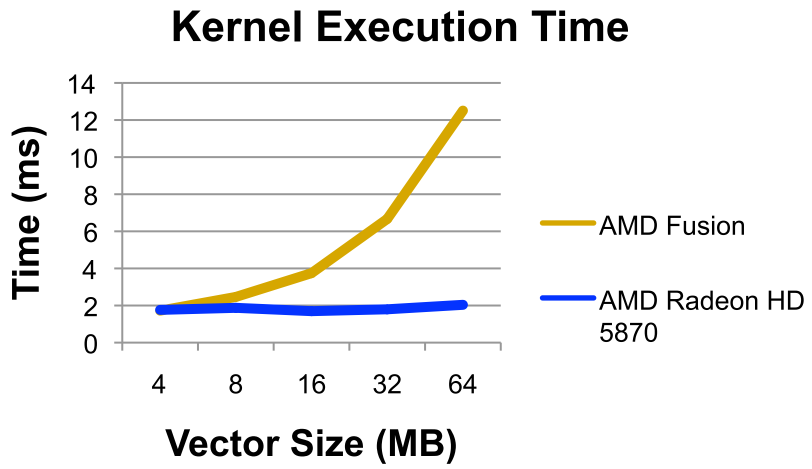


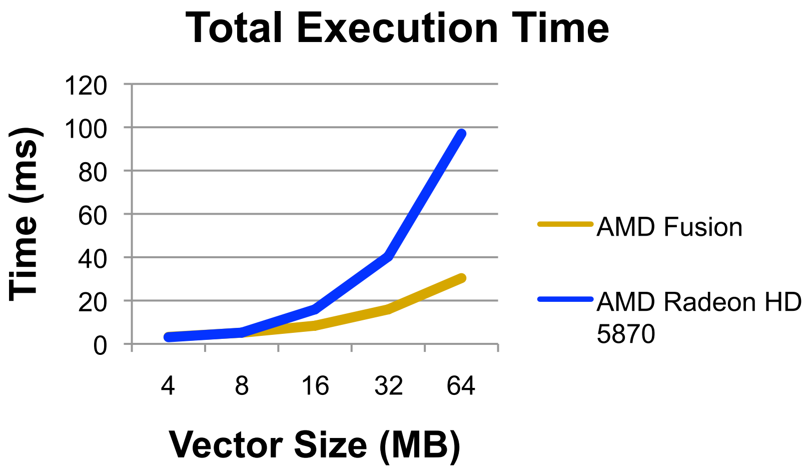
- APU reduces data transfer times for all problem sizes.
- The efficacy of the APU increases as the problem size increases.
- The kernel executes fastest on discrete AMD 5870 due to more and faster GPU cores. The fused AMD Zacate APU is next fastest.

Performance: Reduction (Dense Linear Algebra)

I/O-bound



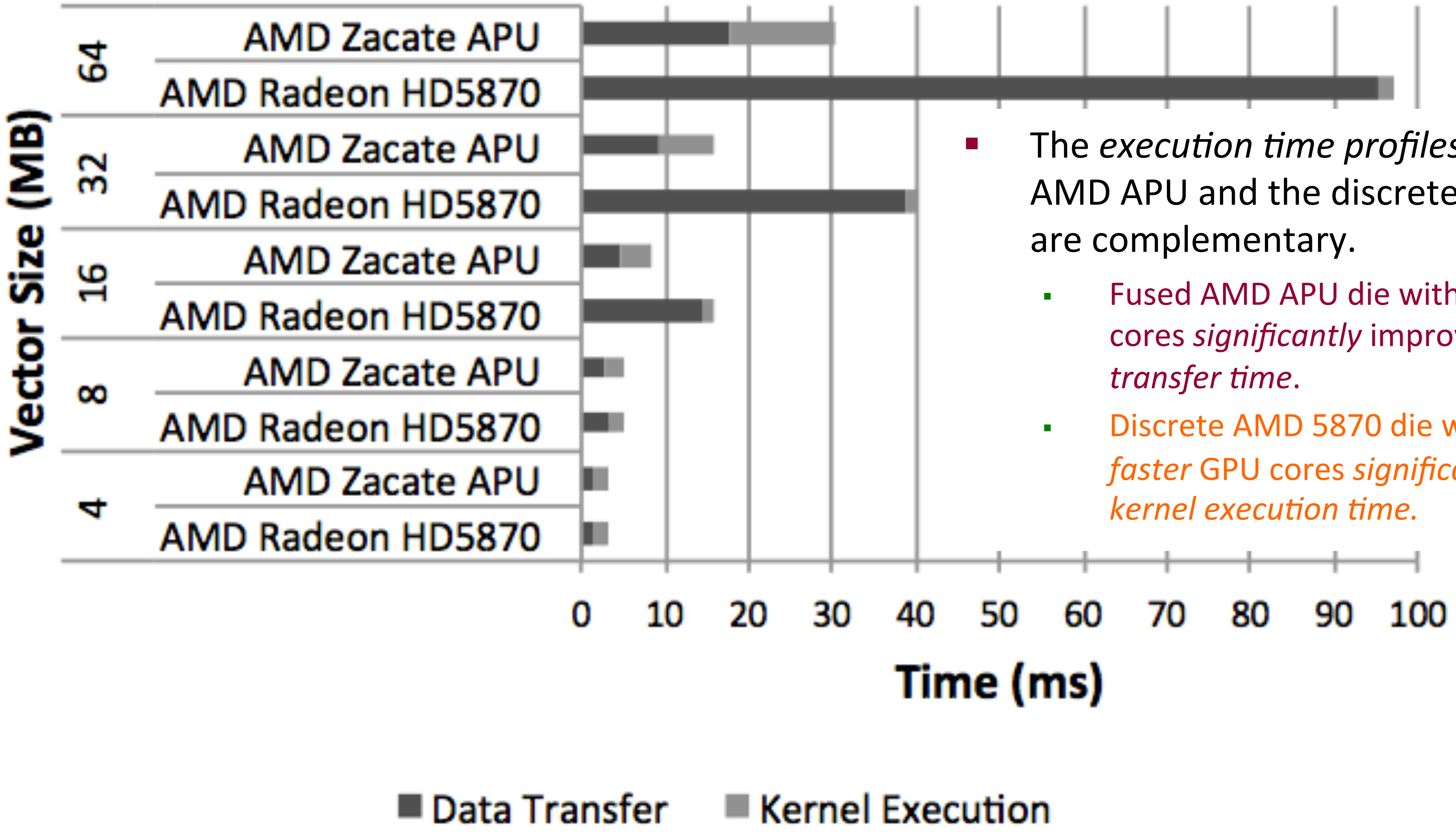




- The *execution time profiles* of the fused AMD APU and the discrete AMD 5870 are complementary.
  - Fused AMD APU die with only 80 GPU cores *significantly* improves *data transfer time*.
  - Discrete AMD 5870 die with 1600 *faster* GPU cores *significantly* improves *kernel execution time.*

Performance: Reduction (Dense Linear Algebra)

I/O-bound



- The *execution time profiles* of the fused AMD APU and the discrete AMD 5870 are complementary.
  - Fused AMD APU die with only 80 GPU cores *significantly* improves *data transfer time*.
  - Discrete AMD 5870 die with 1600 *faster* GPU cores *significantly* improves *kernel execution time.*

Power Consumption of Fused vs. Discrete GPU

- AMD Zacate APU Machine
  - Idle: 11 watts
  - Load: 17-20 watts
- AMD Radeon HD 5870 Machine w/ 2-GHz Intel Xeon E5405
  - Idle: 188 watts
  - Load: 260 watts

Energy Efficiency of Fused vs. Discrete GPU

- AMD Zacate APU
  - SpMV: ~5x (on avg)
  - N-body: ~3x (on avg)
  - FFT: ~1.5x (on avg)
  - Scan: ~8x (on avg)
  - Reduce: ~25x (on avg)

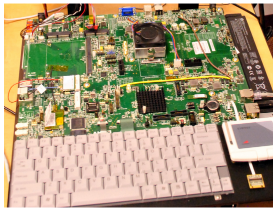
  better than the AMD Radeon HD 5870

Energy-Efficient E-puting Everywhere

- Converged world of consumer ***e***lectronics and com***puting***

  … democratizing “supercomputing in small spaces”

… via “trickle-up” technology



System 0.10 cluster

Acknowledgements

NSF CHREC NSF Center for High-Performance Reconfigurable Computing NVIDIA VirginiaTech AMD IBM

AMD Los Alamos NATIONAL LABORATORY HARRIS XILINX AIR FORCE RESEARCH LABORATORY

- Worldwide Collaborators
- Staff: Mark Gardner, Heshan Lin, and the Green500 Staff
- Students
  - Mayank Daga (VT → AMD)
  - Thomas Scogland
  - Balaji Subramaniam

HokieSpeed, a 500-Tflop GPU-accelerated supercomputer

Wu Feng, wfeng@vt.edu, 540-231-1192

SyNeRG

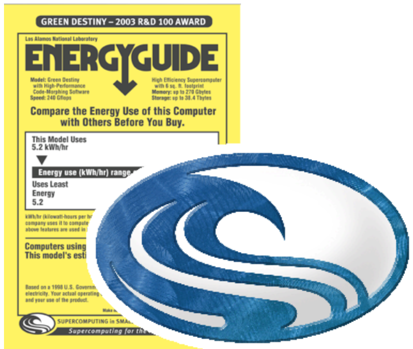
http://synergy.cs.vt.edu/

CHREC NSF Center for High-Performance Reconfigurable Computing

http://www.chrec.org/

mpiBLAST

http://www.mpiblast.org/



GREEN DESTINY – 2003 R&D 100 AWARD

Los Alamos National Laboratory

ENERGYGUIDE

Model: Green Destiny with High-Performance Code-Morphing Software
Speed: 240 Gflops

High Efficiency Supercomputer with 6 sq. ft. footprint
Memory: up to 270 Gbytes
Storage: up to 38.4 Tbytes

Compare the Energy Use of this Computer with Others Before You Buy.

This Model Uses 5.2 kWh/hr

Energy use (kWh/hr) range

Uses Least Energy 5.2

kWh/hr (kilowatt-hours per h
company uses it to compute
above features are used in

Computers using
This model's esti

Based on a 1998 U.S. Governm
electricity. Your actual operating
and your use of the product.

SUPERCOMPUTING in SMAL

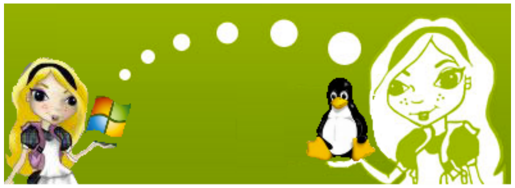
Supercomputing for the

SUPERCOMPUTING in SMALL SPACES

http://sss.cs.vt.edu/

THE GREEN 500

http://www.green500.org/



http://myvice.cs.vt.edu/

"Accelerators 'R Us"

http://accel.cs.vt.edu/