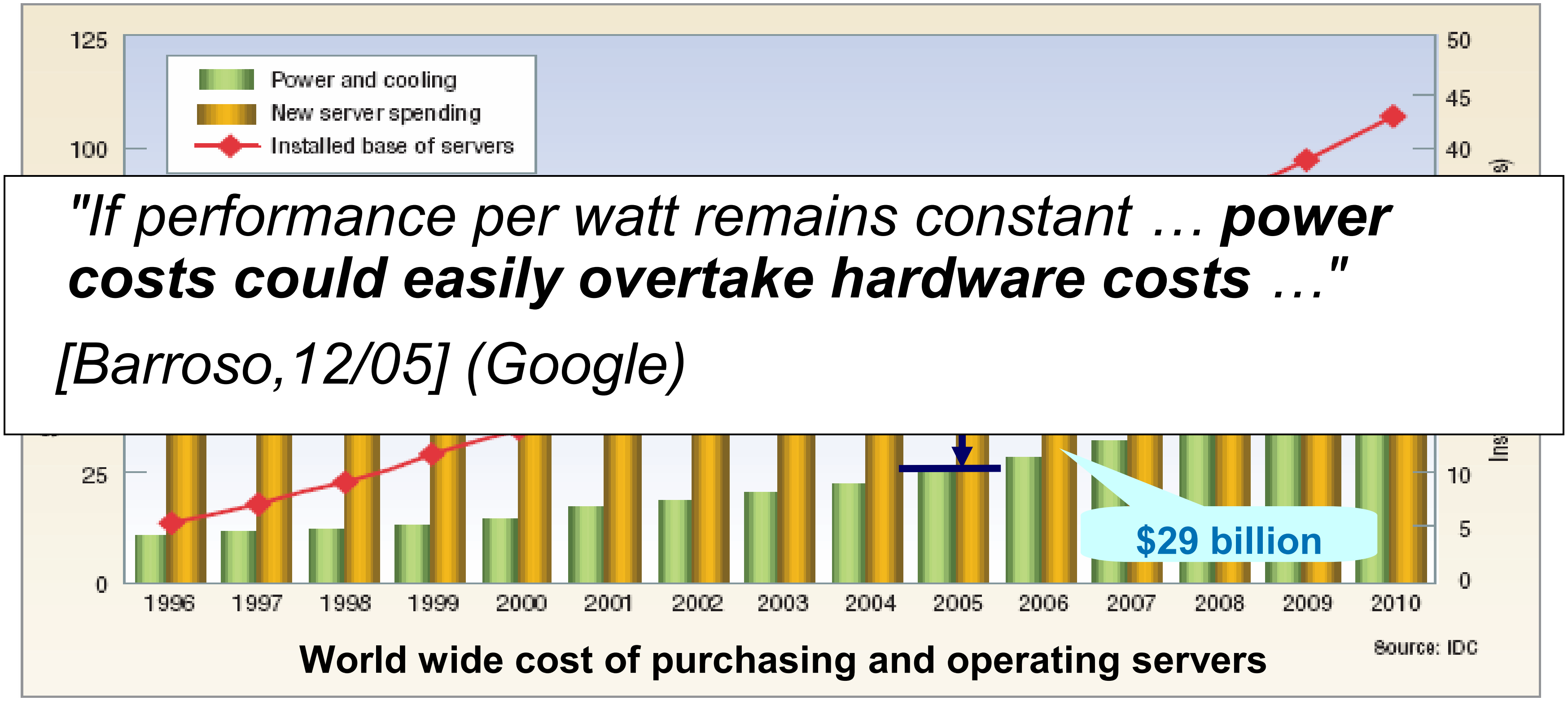
JouleSort: A Balanced Energy-Efficiency Benchmark

**Suzanne Rivoire (Stanford), Mehul Shah (HP Labs),
Partha Ranganathan (HP Labs), Christos Kozyrakis (Stanford)**

Energy Use is Important (1 of 2)

- From data centers to mobile devices
- Data center: power and cooling



*"If performance per watt remains constant … **power costs could easily overtake hardware costs** …"*

[Barroso,12/05] (Google)

World wide cost of purchasing and operating servers

Source: IDC

Energy Use is Important (2 of 2)

- Data center: power and cooling
  - Implications on reliability, density, and scalability
  - Pollution – 4M tons CO_2 [*C. Patel et al., 2006*]
  - Load on utilities
- Desktops: electricity costs
- Mobile devices: battery life affects usability

Benchmarks

- Inspire energy-efficiency improvements

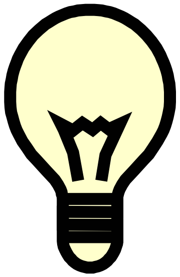
the green grid™ ENERGY STAR spec

- Current efforts
  - E.g., MIPS/Watt, SPECint/Watt, SWaP, …
  - E.g., Ongoing activity in Green Grid, EPA, SPEC Power, …
- But often …
  - Focused on specific component
  - Under-specified or “under construction”
  - Application specific: realistic but complex

No simple holistic benchmark

JouleSort: Simple and Holistic

- Primarily meant for system designers
  - Simple: easy to setup and experiment
  - Evaluate disruptive technology, gain insights
  - Technology bellwether: anticipate trends



- Measure whole-system energy-efficiency
- Workload, metric, and guidelines
- Based on external sort

Our Contributions

I: JouleSort: Holistic energy-efficiency benchmark

- Design: workload, metrics, guidelines
- Rationale and pitfalls

II: Energy-efficient system design: 2007 “winner”

- 3.5X better than previous estimated best
- Insights on future designs

Why External Sort? (1 of 2)

- Simple, balanced workload
  - Exercises all core components
  - CPU, memory, disk, I/O, OS, filesystem
- Applies to systems small and large
  - PDAs, Laptops, Desktop, Supercomputers
- Representative of sequential I/O tasks
  - Data warehousing, Business analytics, etc.

Why External Sort? (2 of 2)

- Hard to cheat
  - Measure system while doing useful work
- Technology trend bellwether
  - E.g. supercomputers to clusters, GPU?
- Holistic measure of improvement

Existing Sort Benchmarks

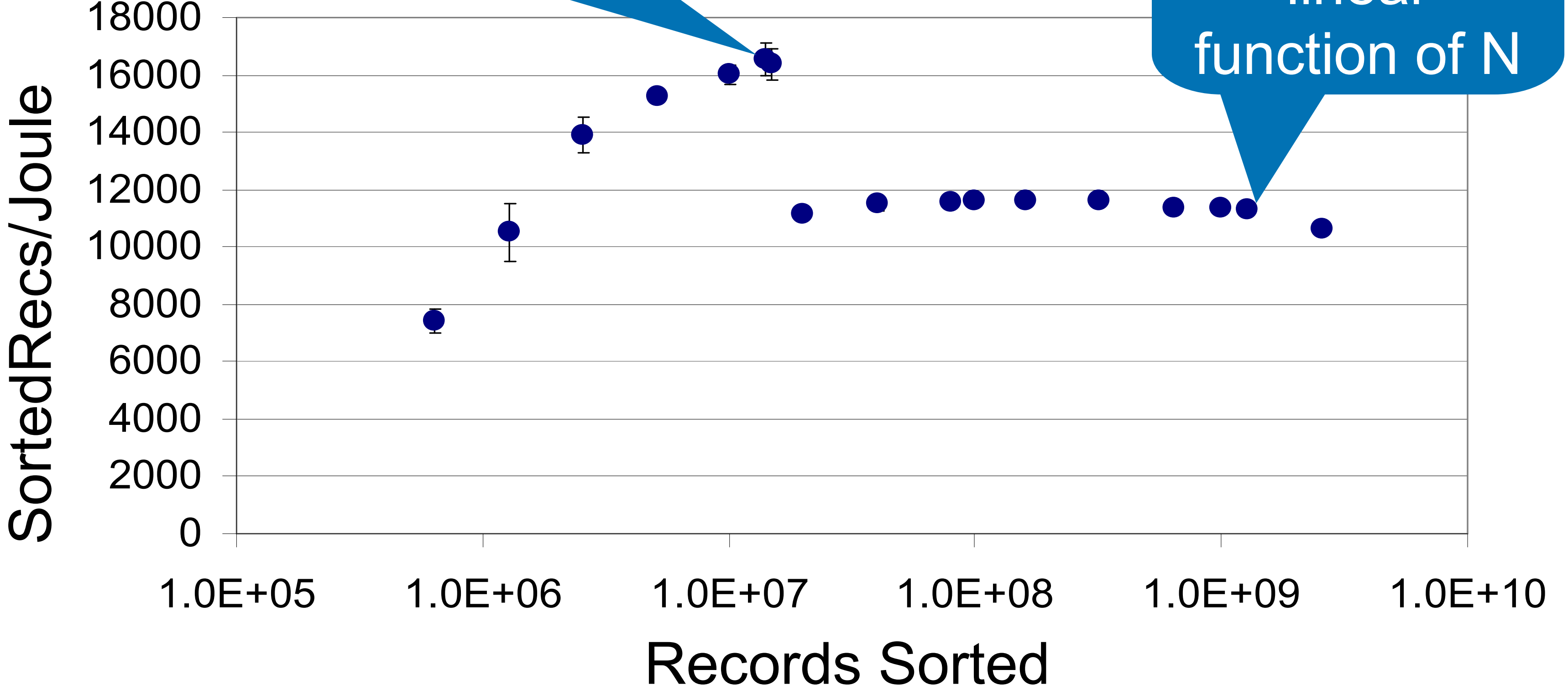
- Pure performance
  - MinuteSort: How much can you sort in 1 min ?
  - TeraByte: How fast can you sort 1 TB ?
- Cost efficient
  - PennySort: How much can you sort for 1 penny ?
  - Performance-Price: Maximum SRecs/$ in 1 min ?

Initial

Our ^ JouleSort Proposal

- *Workload*
  - Sort 100-byte records with 10-byte keys
  - From file on non-volatile store to file on non-volatile store
- *Metric?*
  - Energy (Joules) = Power (Watts)* Time (secs)
  - Fixed time budget (like MinuteSort, Price-Perf Sort)
    - 1 minute budget
    - Measure records sorted and Joules
    - Winner: max SortedRecs/Joule?

Proble Time Budg



- Biased toward systems that sort fewer records
- Better efficiency with 1-pass sort and sleep
  - System not doing useful work

Revised

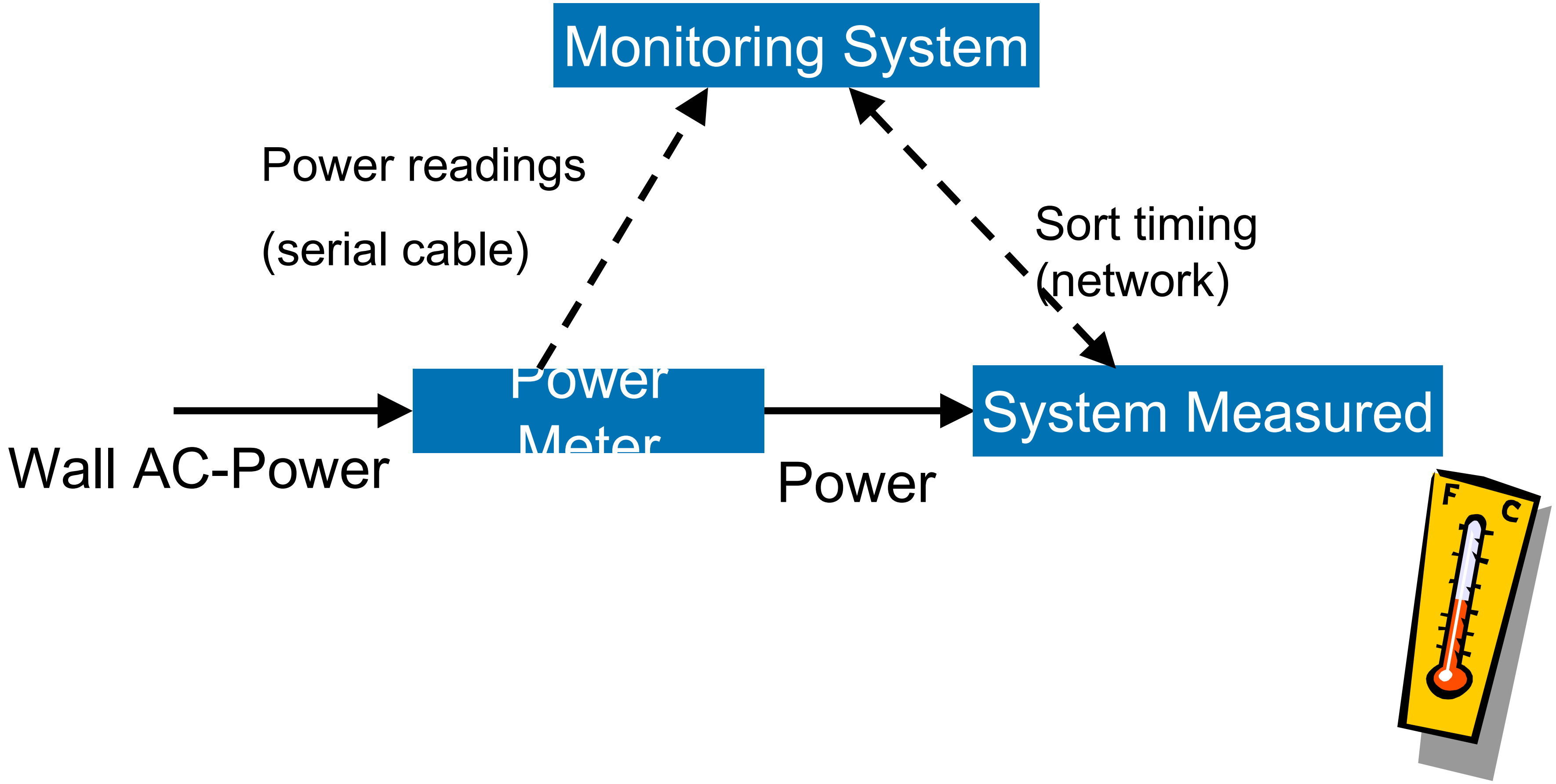
Our ^ JouleSort Proposal

- Fixed input size (like TeraByte)
  - Three classes: 10GB, **100GB**, 1TB
  - Winner: minimum energy
  - Report SortedRecs/Joule (like MPG for cars)

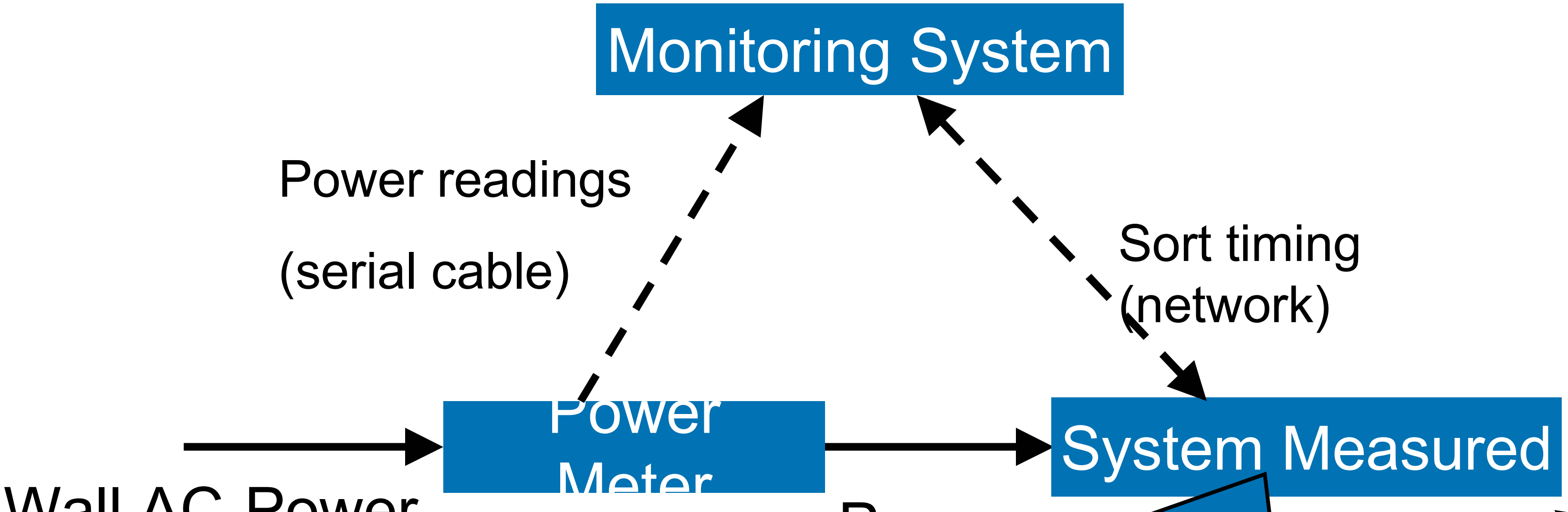
  - Inter-class comparisons imperfect
  - Adjust classes as technology improves

- Categories
  - **Daytona** “street-car”: sold and supported
  - Indy “no-holds-barred”

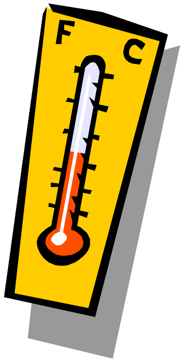
Energy Measurement



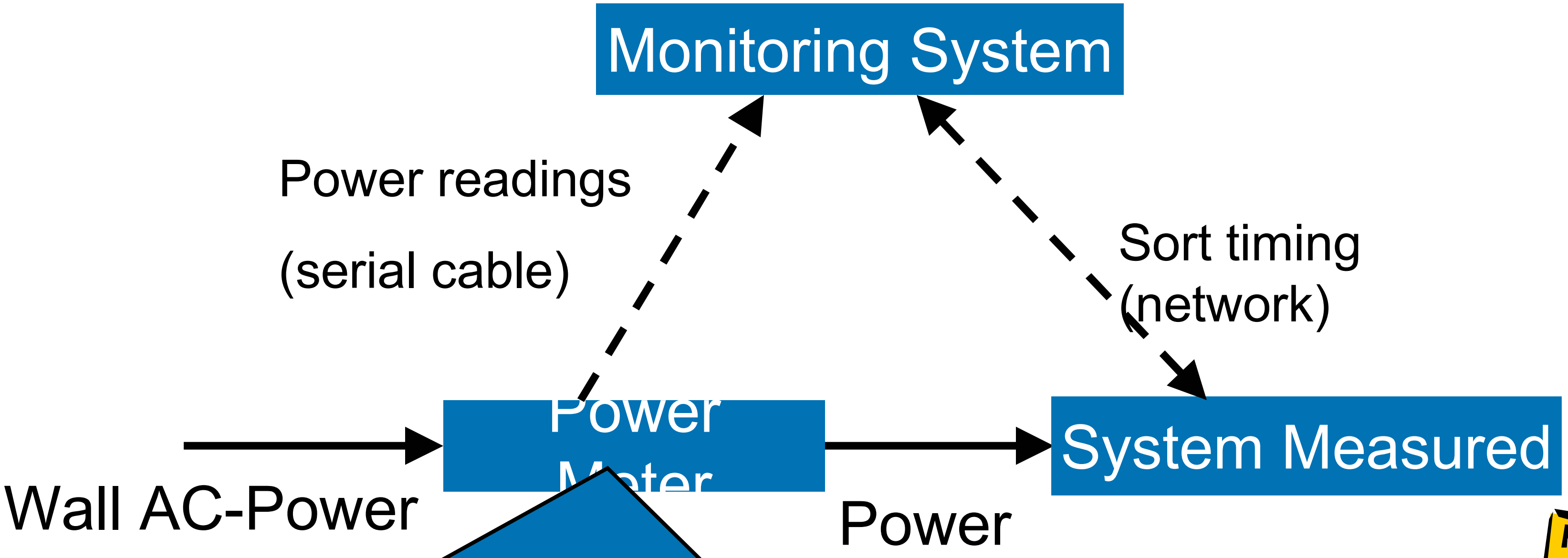
Energy Measurement



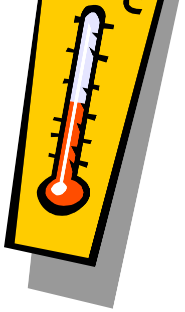
- Measure energy of all components
  - No unaccounted potential energy
  - Cooling devices attached to system
- 20-25 C at inlet or within 1 foot of device



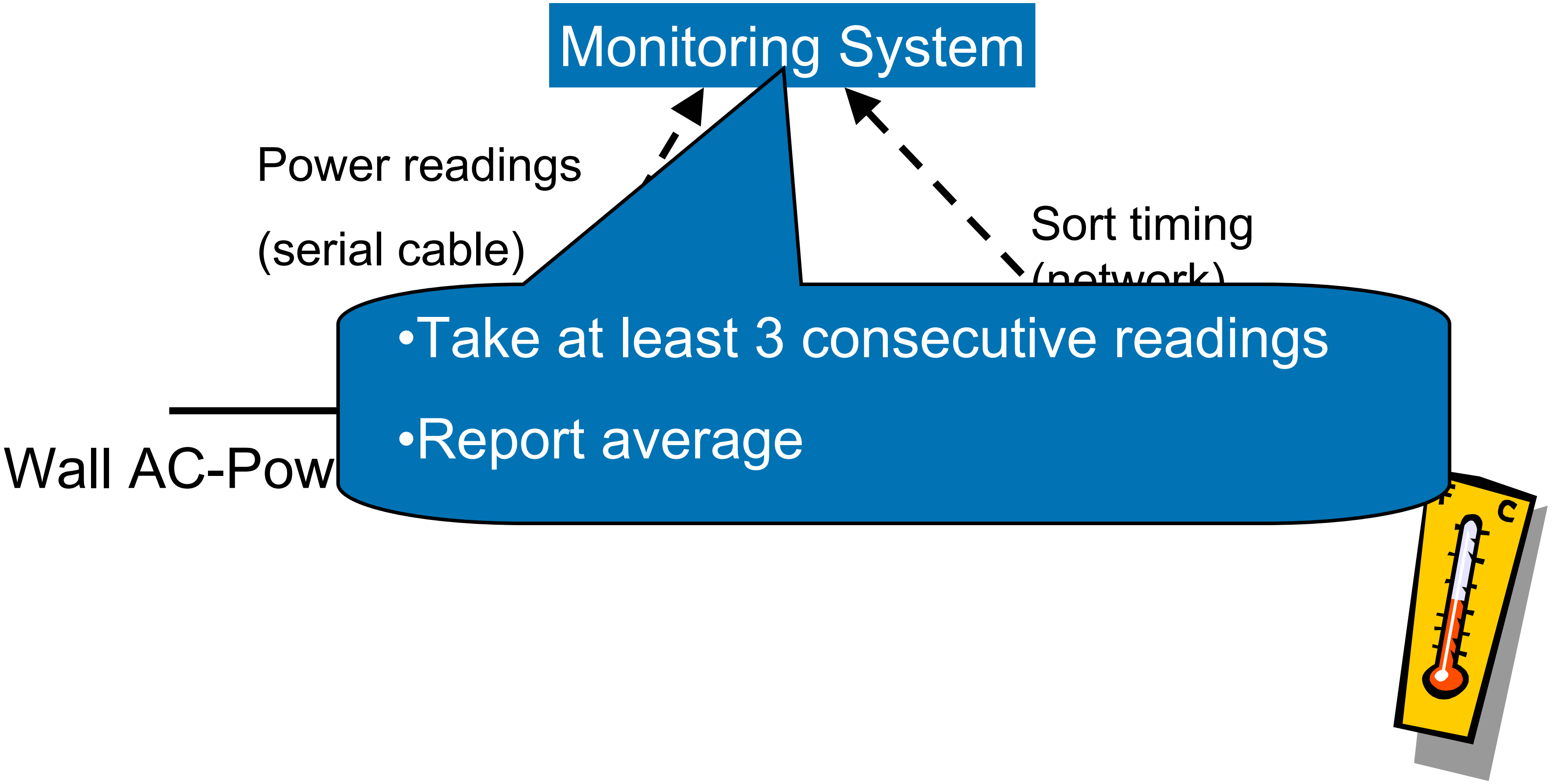
Energy Measurement



- Measure true power from wall
  - Applies to AC and DC
  - Report power-factor
- Leverage SPEC-Power/EPA specifications



Energy Measurement



Monitoring System

Power readings

(serial cable)

Sort timing

(network)

Wall AC-Pow

- Take at least 3 consecutive readings
- Report average

Road Map

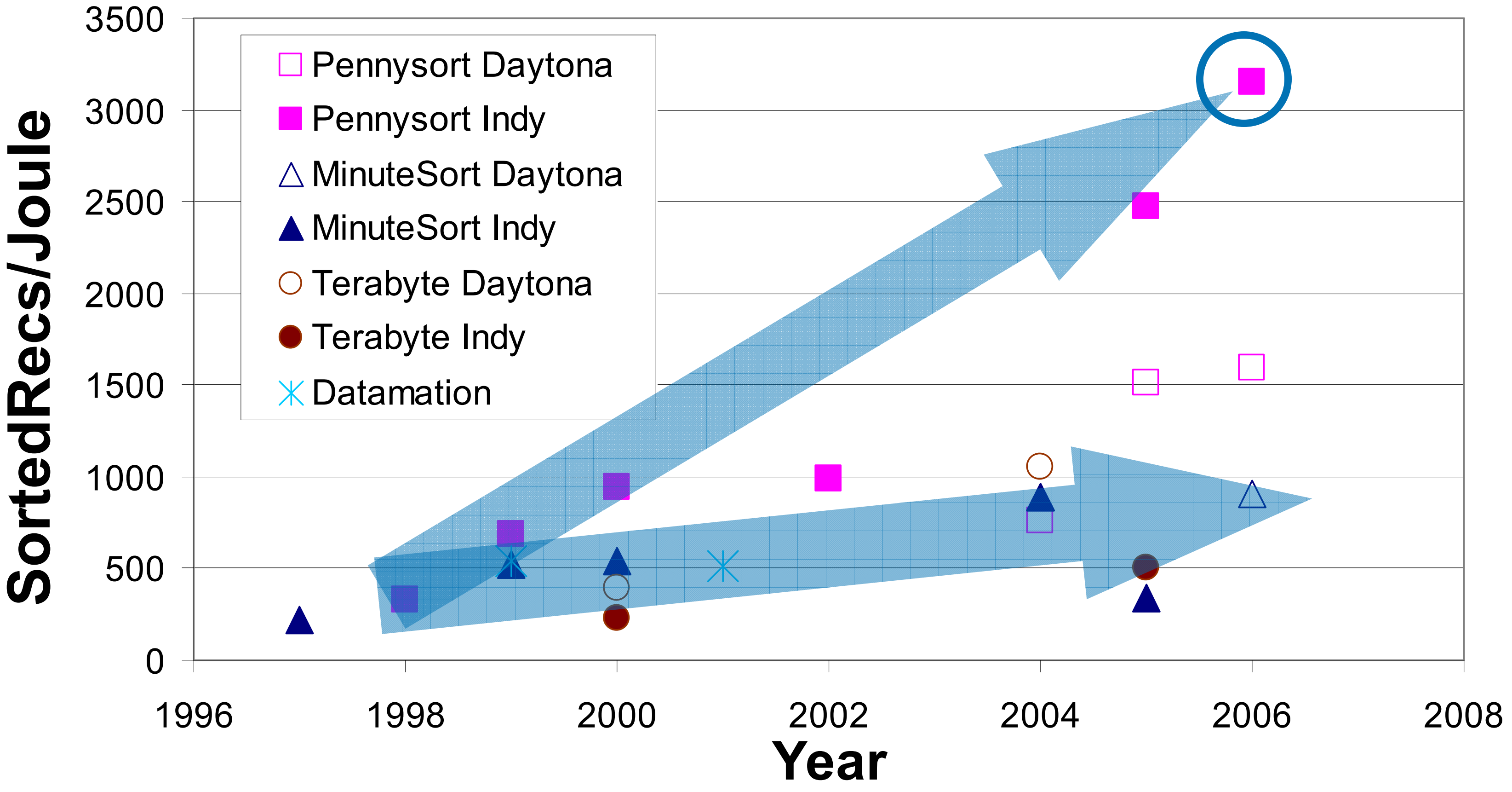
I: JouleSort: Holistic energy-efficiency benchmark

- Design: workload, metrics, guidelines
- Rationale and pitfalls

II: Energy-efficient system design: 2007 “winner”

- 3.5X better than previous estimated best
- Insights on future designs

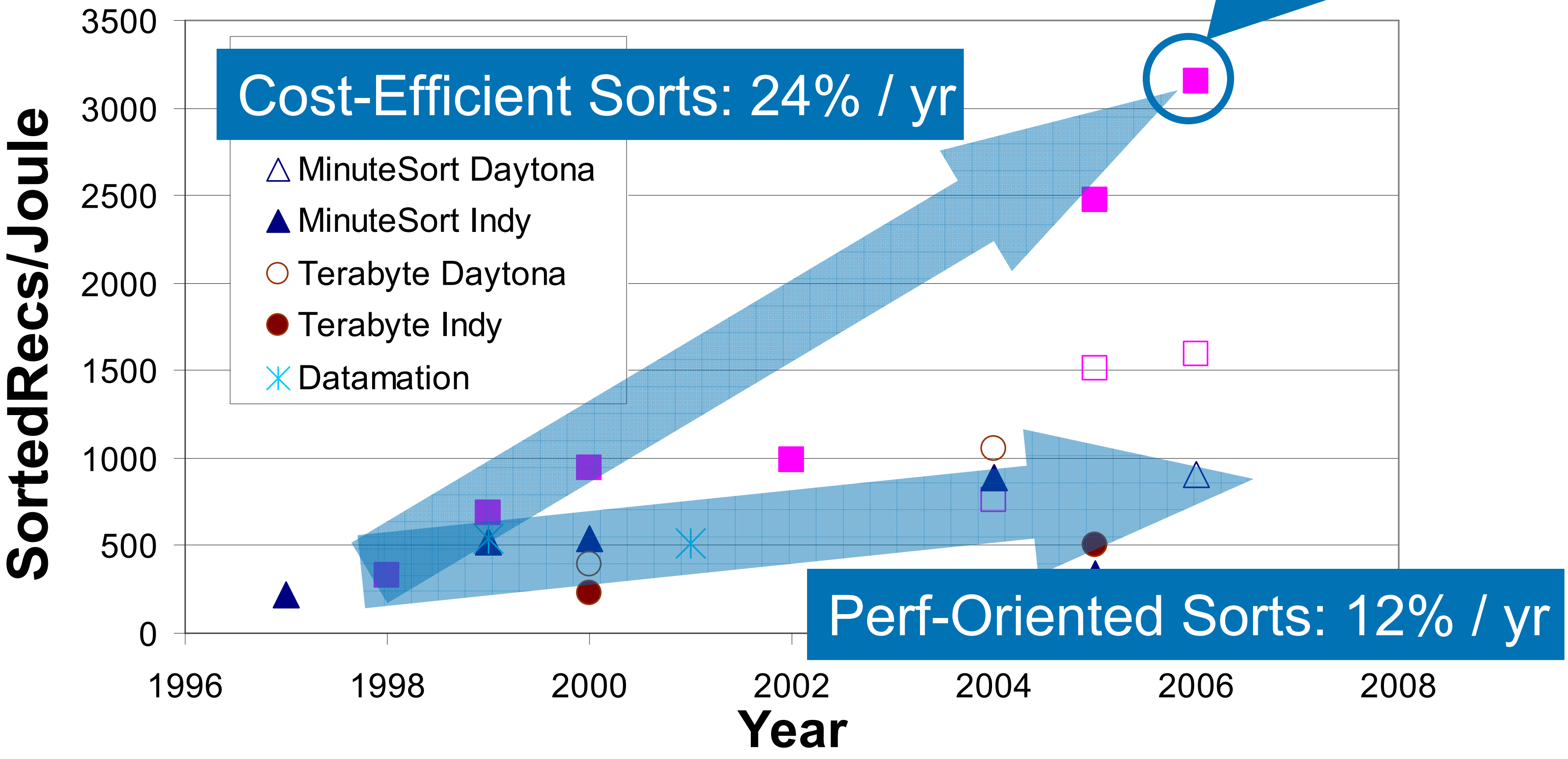
Historical Analysis (Estimate)



Historical Analysis (

Target: GPUTeraSort

~3200 SortedRecs/Joule



Cost-Efficient Sorts: 24% / yr

Perf-Oriented Sorts: 12% / yr

A Look at Existing Systems

	# Disks	CPU %	Input Size	Power (Watt)	SortedRecs per Joule
GPUTeraSort (estimated)	9	n/a	59GB	290	~3200
Low-power Blade	1	11%	5GB	90	~300
Low-end server DL360G3	2	26%	10GB	140	~1200
Modern Laptop	1	1%	10GB	22	~3400
Sort-balanced Fileserver	12	90%+	10GB	406	~3800

A Look at Existing Systems

	# Disks	CPU %	Input Size	Power (Watt)	SortedRecs per Joule
GPUTeraSort (estimated)	9	n/a	[illegible]	[illegible]	[illegible]
Low-power Blade	1	11%	[illegible]	[illegible]	[illegible]
Low-end server DL360G3	2	26%	10G[illegible]	[illegible]	~1200
Modern Laptop	[illegible]	[illegible]	[illegible]	2[illegible]	~3400
Sort-balanced Fileserver	12	90%+	10GB	406	~3800

DL360G5 server: 180W

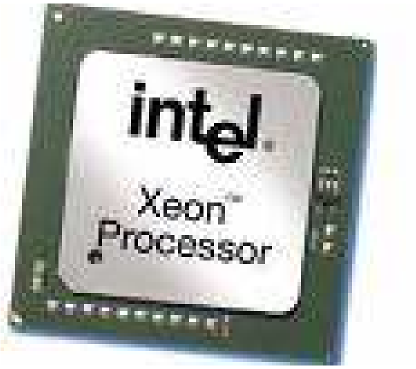
Disk trays + disks: 226W

Active Idle: 370W

Optimizing for Energy-Efficiency: Step 1

Lower power components w/o equal perf. loss

Fileserver



Sort BW: 313 MB/s

65W (peak)

75% perf

→

52% power

Our winner

Sort BW: 236 MB/s

34W (peak)

Optimizing for Energy-Efficiency: Step 1

Lower power components w/o equal perf. loss

Fileserver

Seagate Barracuda

Seq. BW: 80MB/s

13W

50% perf

→

15% power

Our winner

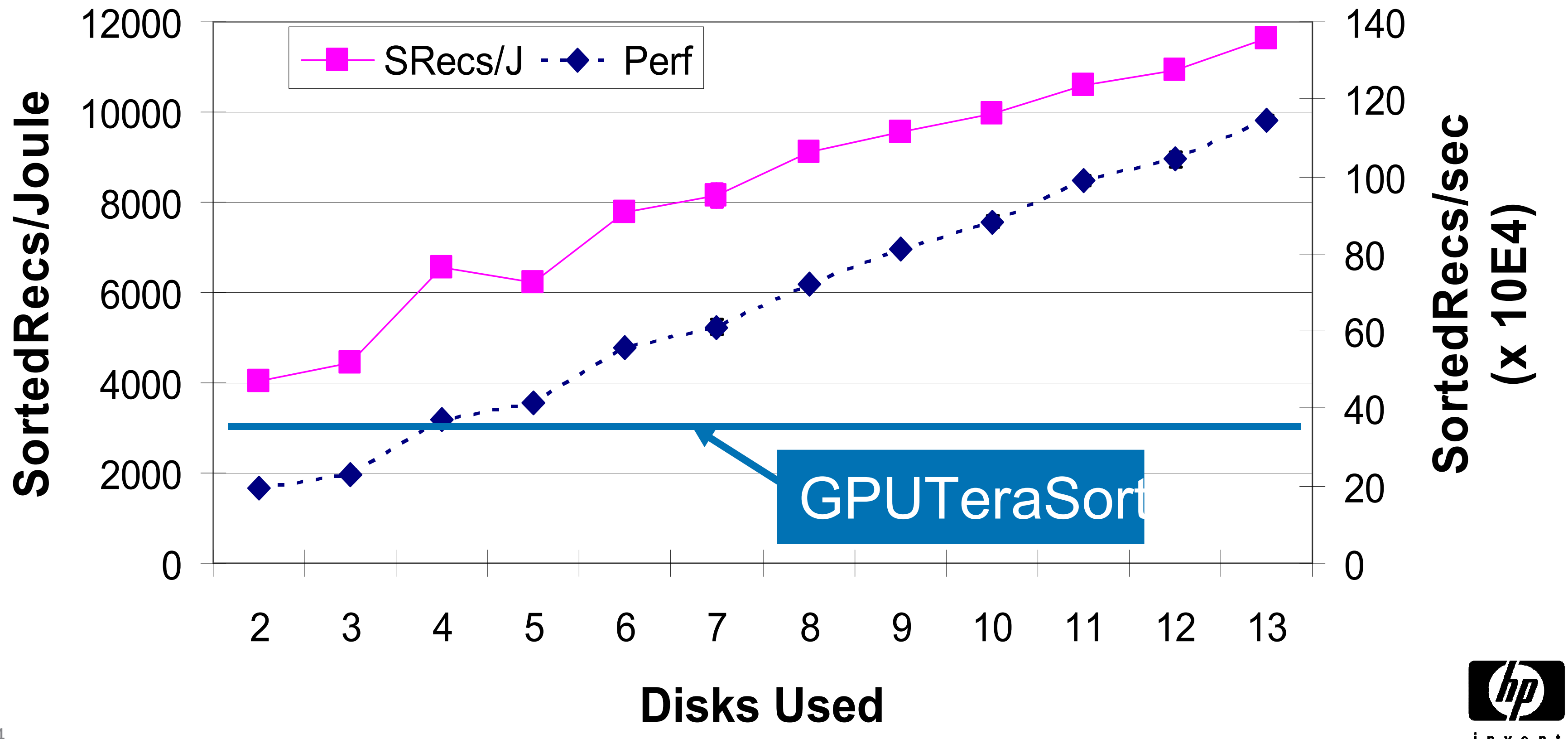
Hitachi Travelstar

Seq. BW: 40MB/s

2W

Optimizing for Energy Efficiency: Step 2

- Maximize performance
  - Balanced sort: enough disks to fully utilize CPU
  - Disks running near peak BW



Winner 100GB Category

- 11300 SortedRecs/Joule
  - 3.5x better than GPUTeraSort
  - Average Power: 100W
  - Ordinal Technology’s NSort (thanks Chris Nyberg)

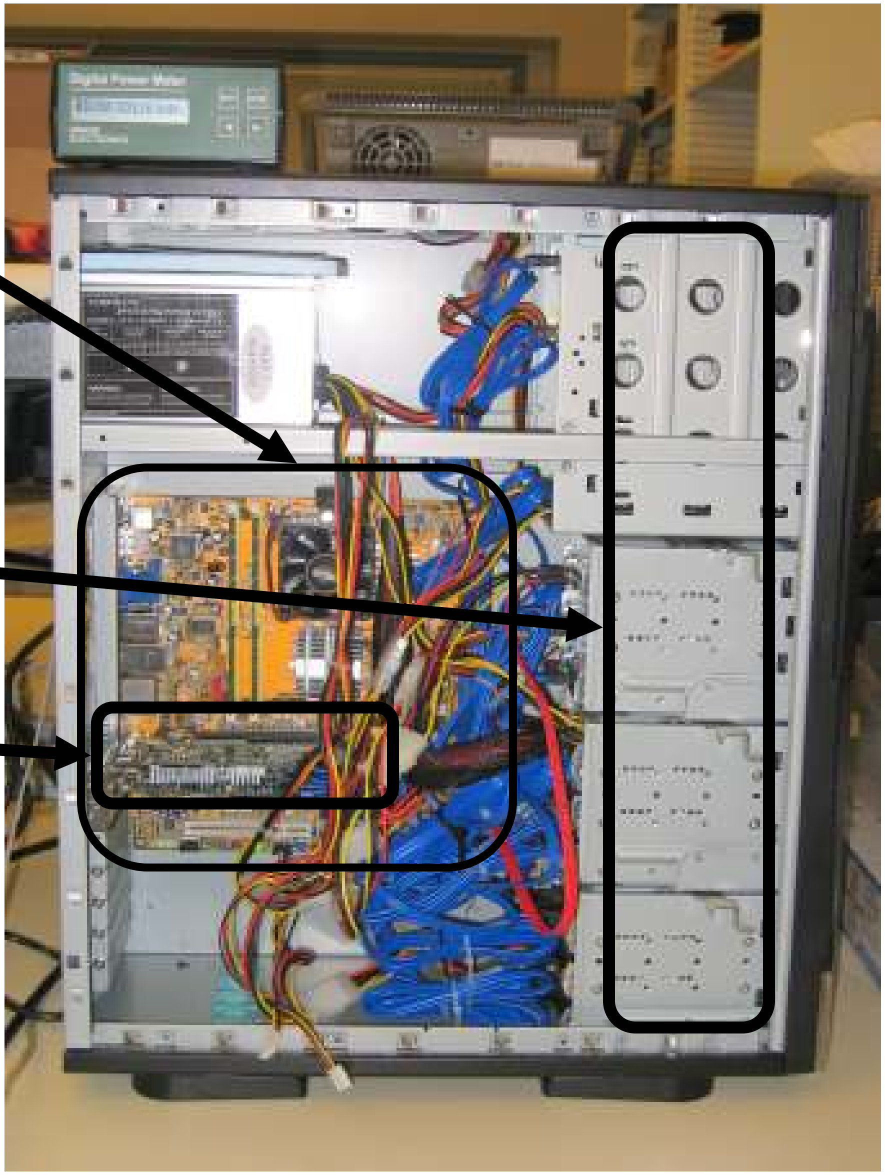
Winner 100GB Category

Asus motherboard:

Mobile CPU + 2 PCI-e slots

13 Hitachi TravelStar 160GB

RocketRAID Disk Controllers



Detailed SW/HW sensitivity experiments in paper

Insights for Future Designs

- All components matter
  - CPU, Disks, Memory, …
  - Low hanging fruit: use low-power HW

- Current technology
  - Limited dynamic range
  - For fixed HW: peak efficiency = peak performance

- Want “scale-down efficiency”
  - 1TB → 100GB and give best of both

Other Issues

- Benchmark design
  - Data-center cooling and control
  - Display power, GPUs, etc.
  - Total cost of ownership

- System design
  - Flash is becoming practical
  - Cheaper, faster, and lower power

Conclusion

- Energy-use is important
  - From data centers to handhelds
- JouleSort
  - Simple, holistic energy-efficiency benchmark
- Built energy-efficient sorting system
  - 3.5x better than 2006 estimated winner (GPUTeraSort)
  - Insights: low-power HW, limited dynamic range
- Part of Sort Benchmark suite
  - Entries welcome for 2008
  - http://joulesort.stanford.edu