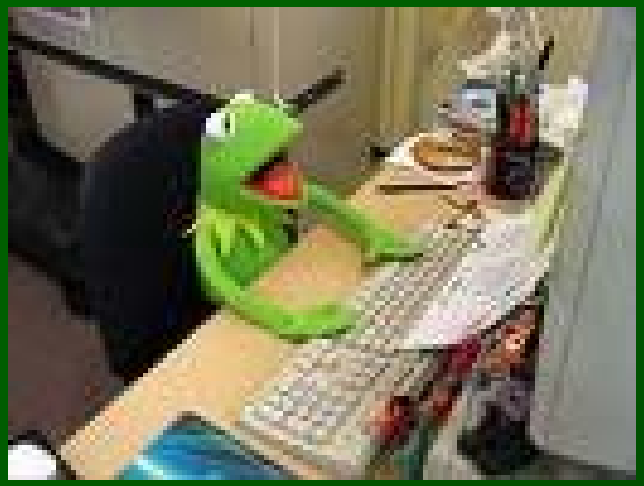
It’s not easy being green…in HPC



Virginia Tech

SCAPE laboratory

Kirk W. Cameron

Scalable Performance Lab

Being Green in HPC

- My Green IT involvement (since 2002)
  - NSF Career Award: HPPAC
  - SPECPower – server benchmarking
  - EnergyStar – US EPA consultant
  - Green500 – co-founder
  - HPPAC Workshop – founding member
  - Uptime Institute – fellow

Economic Impact of HPC Energy Use

- Cost
  - $800,000 per year per megawatt
  - "Power efficient" Roadrunner: 2.3 MW
- Reliability
  - 10 degrees C = 50% reliability
  - Environmental Canada IBM Supercomputer
    - Recycles thermal energy produced by machine
    - Can heat 5-story building (-15 C outside temps)
  - Earth Simulator can heat a domed stadium

Environmental Impact of HPC Energy Use

- Details
  - 1 coal generated kWh = 2 lbs CO2
  - 2,204 pounds = 1 metric ton
  - My auto: 6.6 metric tons annually

1MW ~ 8,000 tons CO2 (1,204 auto/yr)

6 autos/yr

TM CM-5
.005 Megawatts

18 autos/yr

Residential A/C
.015 Megawatts

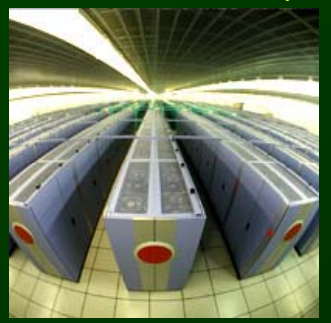
1,023 autos/yr

Intel ASCI Red
.850 Megawatts

12,044 autos/yr

High-speed train
10 Megawatts

14,453 autos/yr



Earth Simulator
12 Megawatts

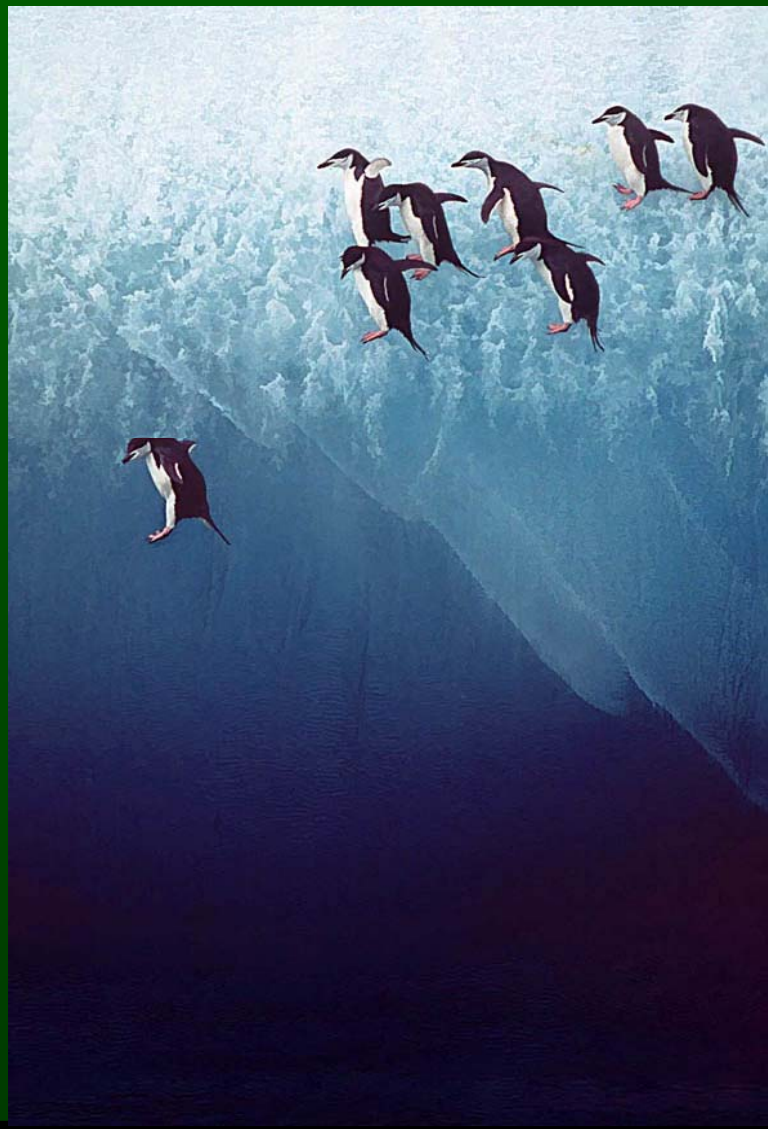
NSF TeraGrid uses ~20MW = 24,088 autos/yr

Our Story…

- My observations circa 2002
  - Power will become disruptive to HPC
  - Laptops outselling PC's
  - Commercial power-aware not appropriate for HPC
  - HPC apps terribly inefficient (<10% peak perf)
- Proposed Solution
  - Exploit inefficiencies
  - Leverage new technologies
  - Create Green HPC SW
- Caveat (It's not easy being Green!)
  - Performance is #1 constraint!!!

HPPAC Emerges 2002

- SCAPE Project
  - High-performance, power-aware computing
  - Two initial goals
    - Measurement tools
    - Power/energy savings
  - Big Goals…no funding (risk all university startup funds)
  - Overall challenge/goal:
    - Same work, less energy!



Intuition confirmed

IT confronts the datacenter power crisis

As energy costs escalate, conserving resources tops the list of challenges for today's IT managers

By Dan Goodin
October 06, 2006

E-mail | Printer Friendly | Reprints | Slashdot It!

When David Young told his colocation provider late last year that his online applications startup, Joyent, planned to add 10 servers to its 150-system datacenter, he received a rude awakening. The local power utility in Southern California wouldn't be able to provide the additional electricity needed. Joyent's upgrade would have to wait.

In the Data Center, the Heat Is On

Halamka John Today's Top Stories ▸ or Other Servers Stories ▸

October 23, 2006 (Computerworld) -- I recently began a project to consolidate two dat

Data Center Budgets Face Radical Changes

Consortium head says facilities costs are surpassing the price of hardware

Patrick Thibodeau and Patrick Thibodeau Today's Top Stories ▸ or Other IT Management Stories ▸

October 30, 2006 (Computerworld) -- *The business value arising from Moore's Law, which says the number of*

Overview of Our Green HPC Progress…

- Measurement Infrastructure
  - PowerPack (begun 2002)
    - Software/hardware for power measurement
  - Tempest (begun 2005)
    - Software for thermal measurement
- Control Infrastructure (MISER)

  Management Infrastructure for Energy Reduction
  - CPU MISER (begun 2002)
  - Memory MISER (begun 2004)
  - SysteMISER (begun 2007)
  - Heat MISER (begun 2007)

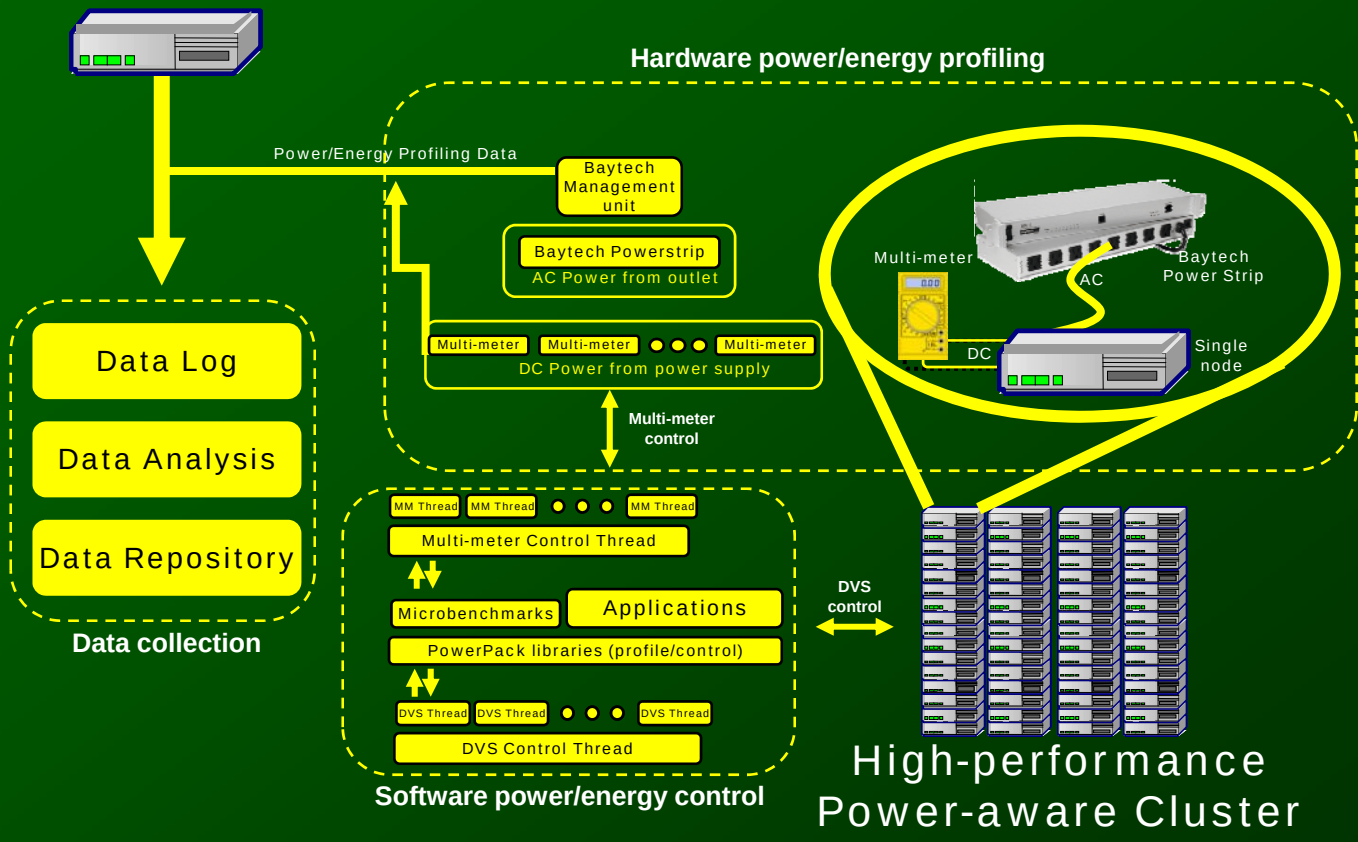
Thanks to our sponsors: NSF (Career, CCF, CNS, CRI), DOE (SC), Intel

AC+DC Measurement Ain't Easy

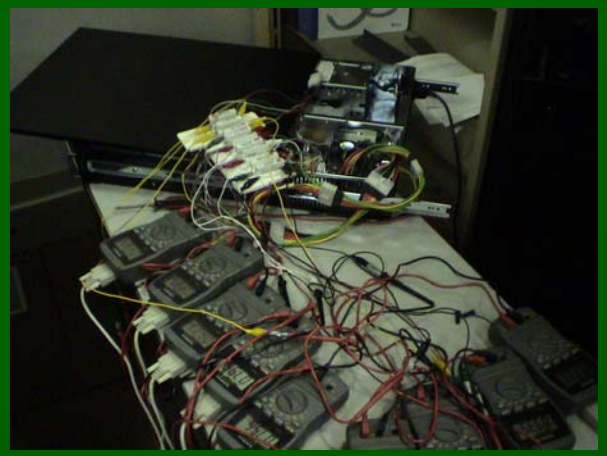
PowerPack I & II

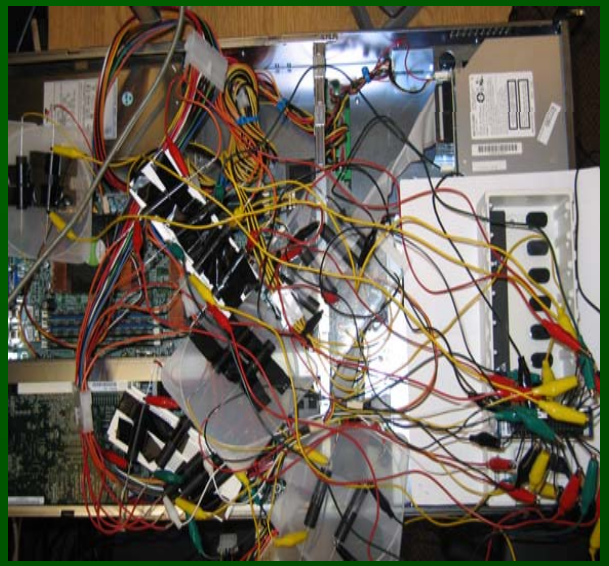
2002 - present

Scalable, synchronized, and accurate.



PowerPack AC/DC Power Profiling

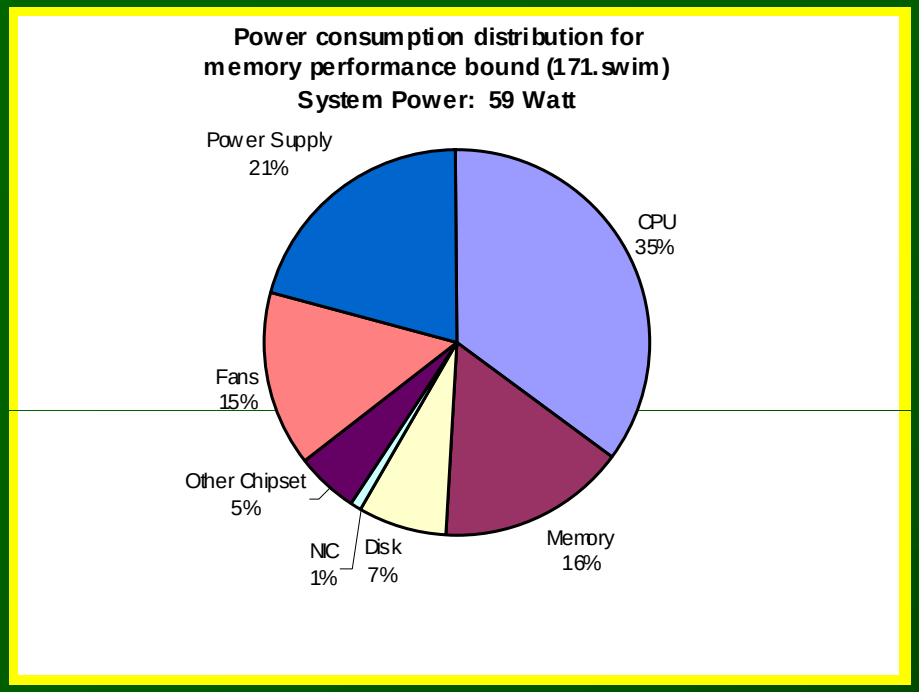


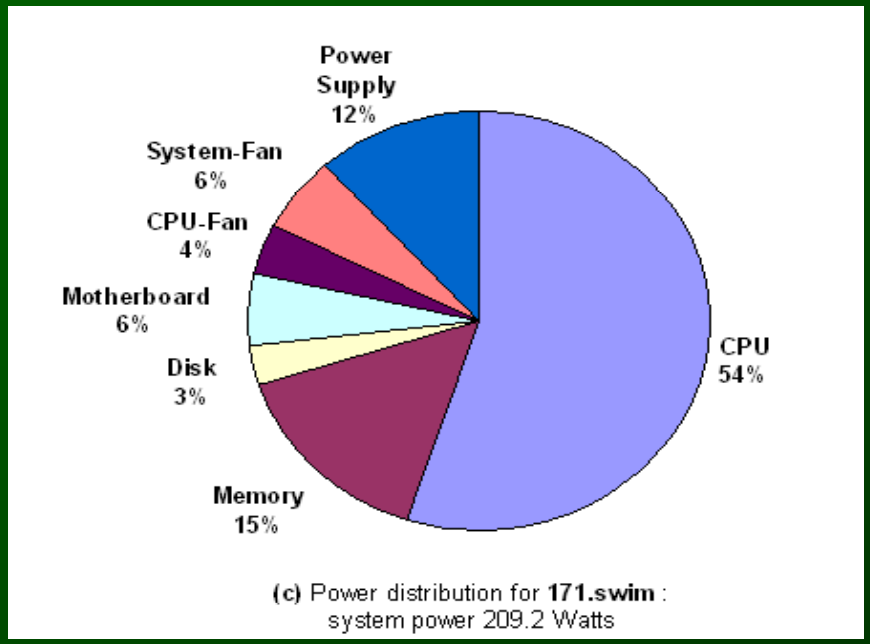


```
If node .eq. root then
        call pmeter_init (xmhost,xmport)
        call pmeter_log (pmlog,NEW_LOG)
endif

<CODE SEGMENT>
```

```
If node .eq. root then
        call pmeter_start_session(pm_label)
endif

<CODE SEGMENT>

If node .eq. root then
        call pmeter_pause()
        call pmeter_log(pmlog,CLOSE_LOG)
        call pmeter_finalize()
endif
```

Power Profiles – Single Node

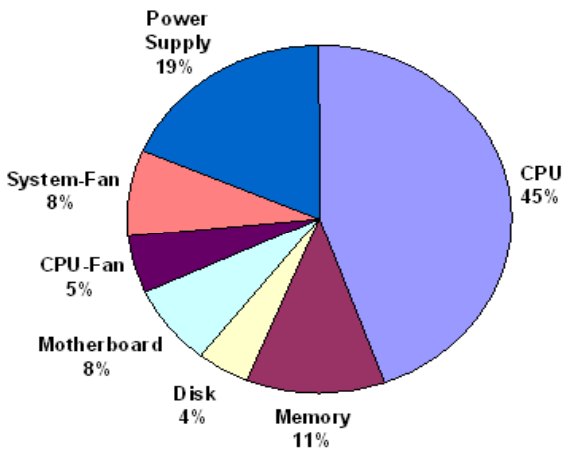




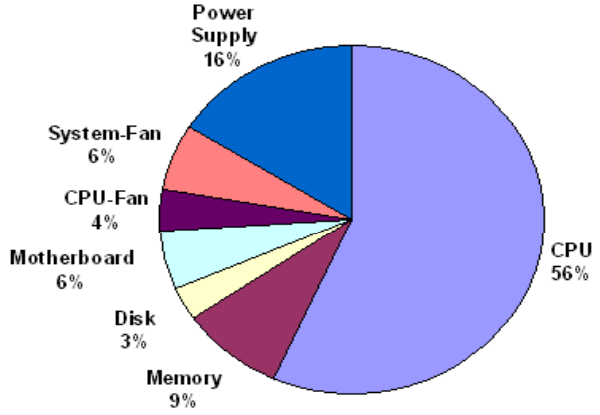
Pentium III (2002) AMD Opteron (2007)

- CPU is largest consumer of power typically (under load)

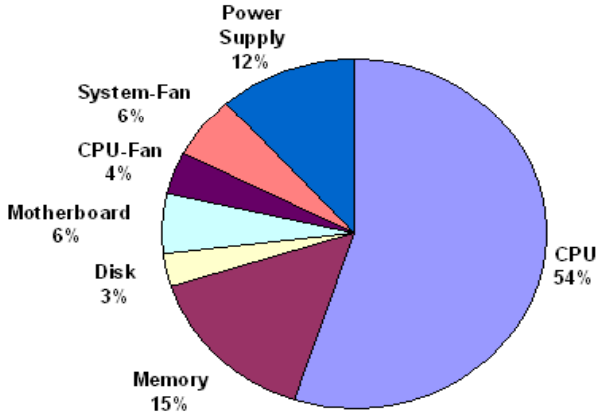
Power Profiles – Single Node



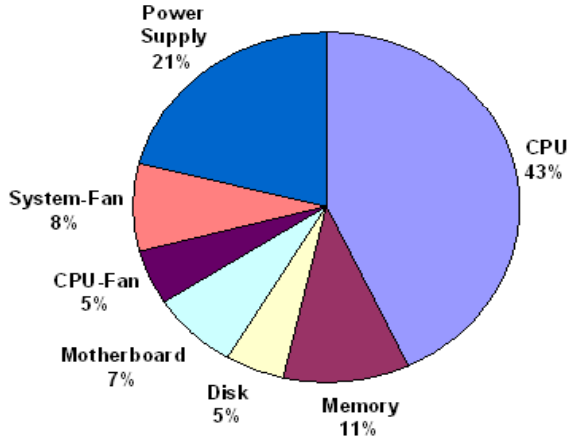
(a) Power distribution for **system idle**: system power 152.5 Watts



(b) Power distribution for **164.gzip**: system power 206.5 Watts



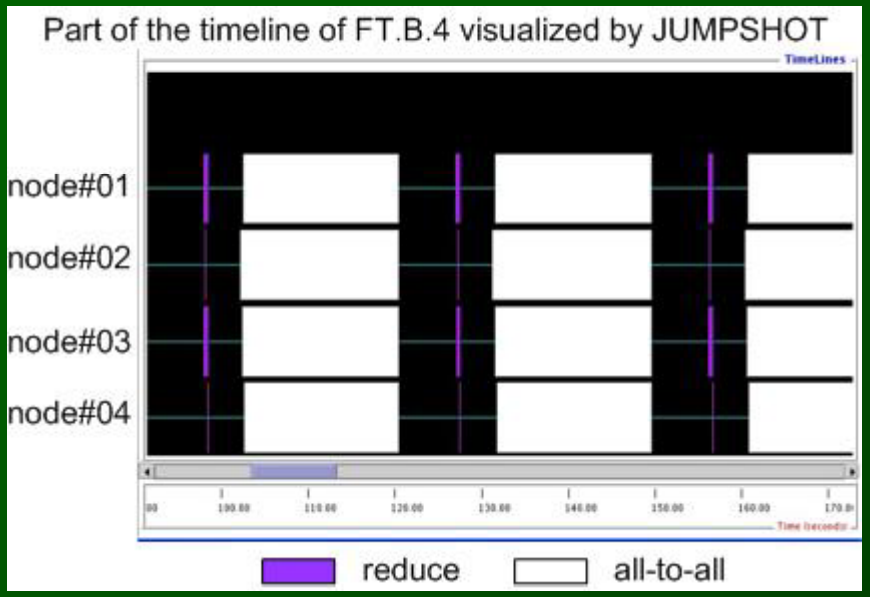
(c) Power distribution for **171.swim** : system power 209.2 Watts

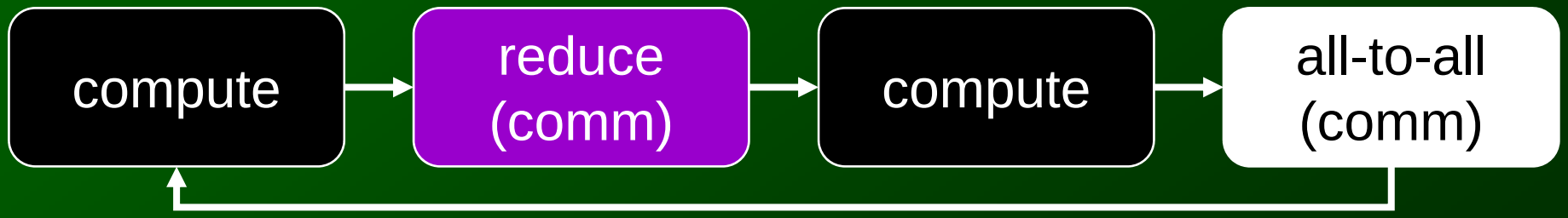


(d) Power distribution for **cp**: system power 165.2 Watts

Fig. 5. Power distribution for a single node under different workloads: (a) zero workload (system is in idle state); (b) CPU bounded workload; (c) memory bounded workload; (d) disk bounded workload.

NAS PB FT - Performance Profiling





About 50% time spent in communications.

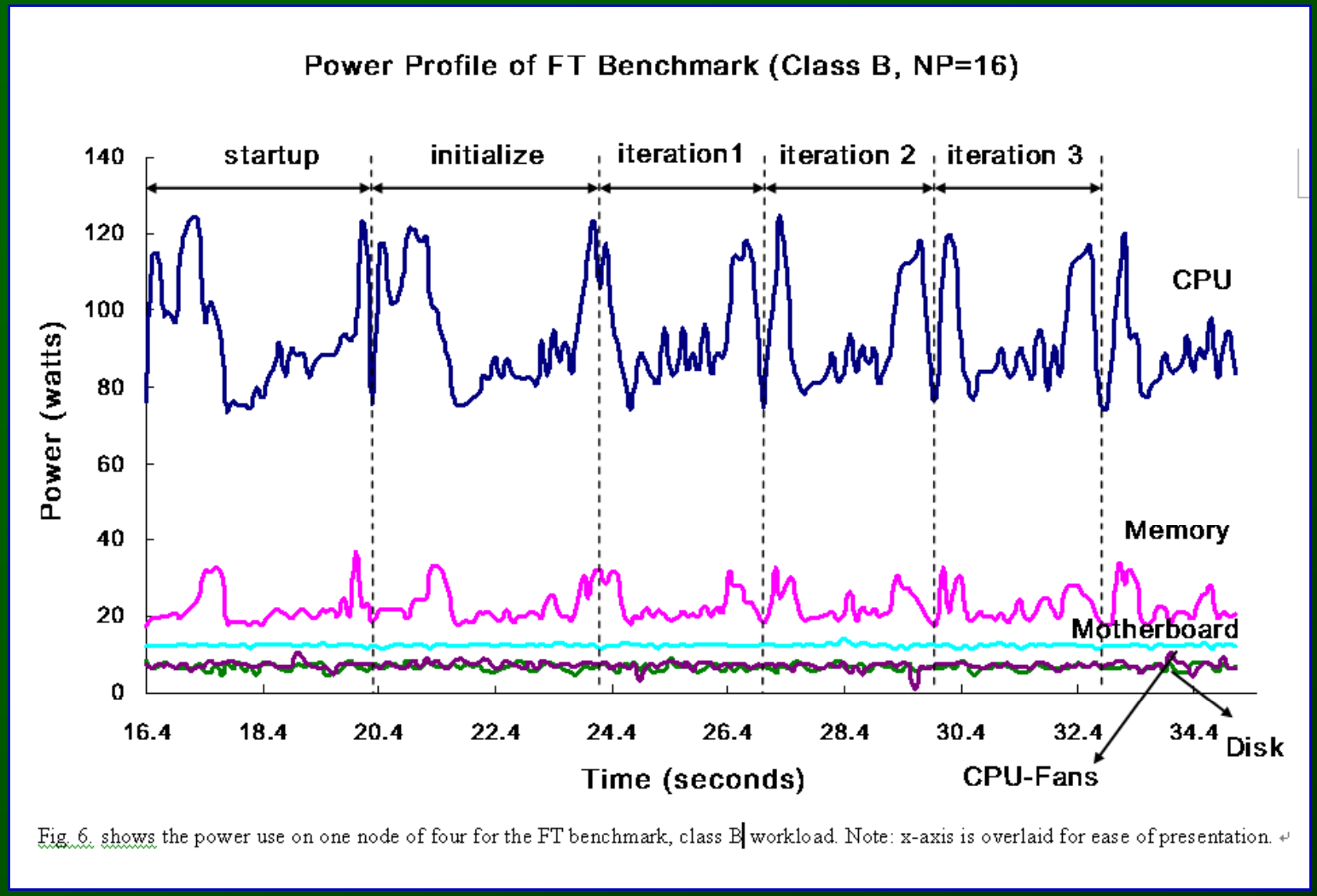


Fig. 6. shows the power use on one node of four for the FT benchmark, class B workload. Note: x-axis is overlaid for ease of presentation.

Power profiles reflect performance profiles.

One FFT Iteration (PowerPack)

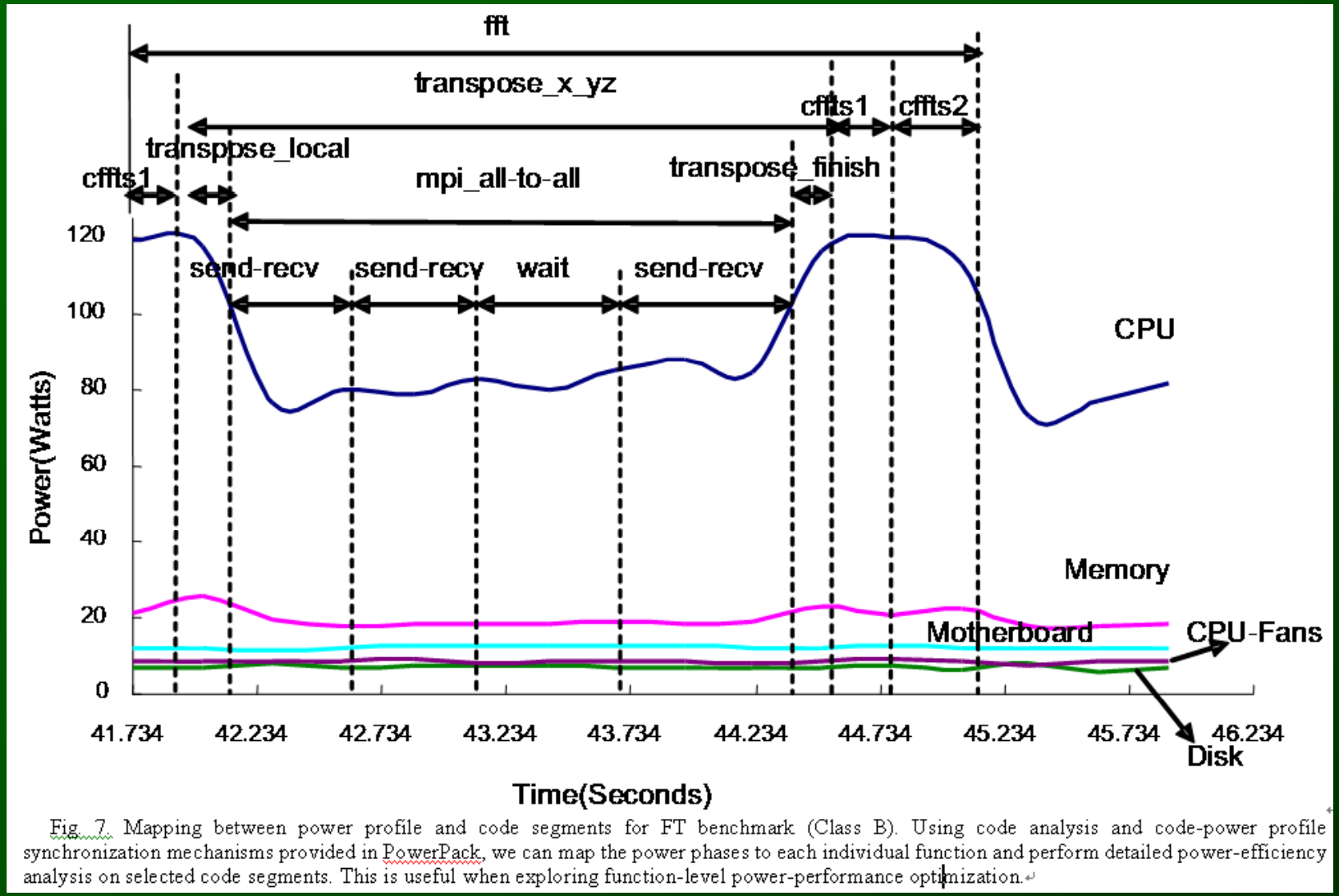
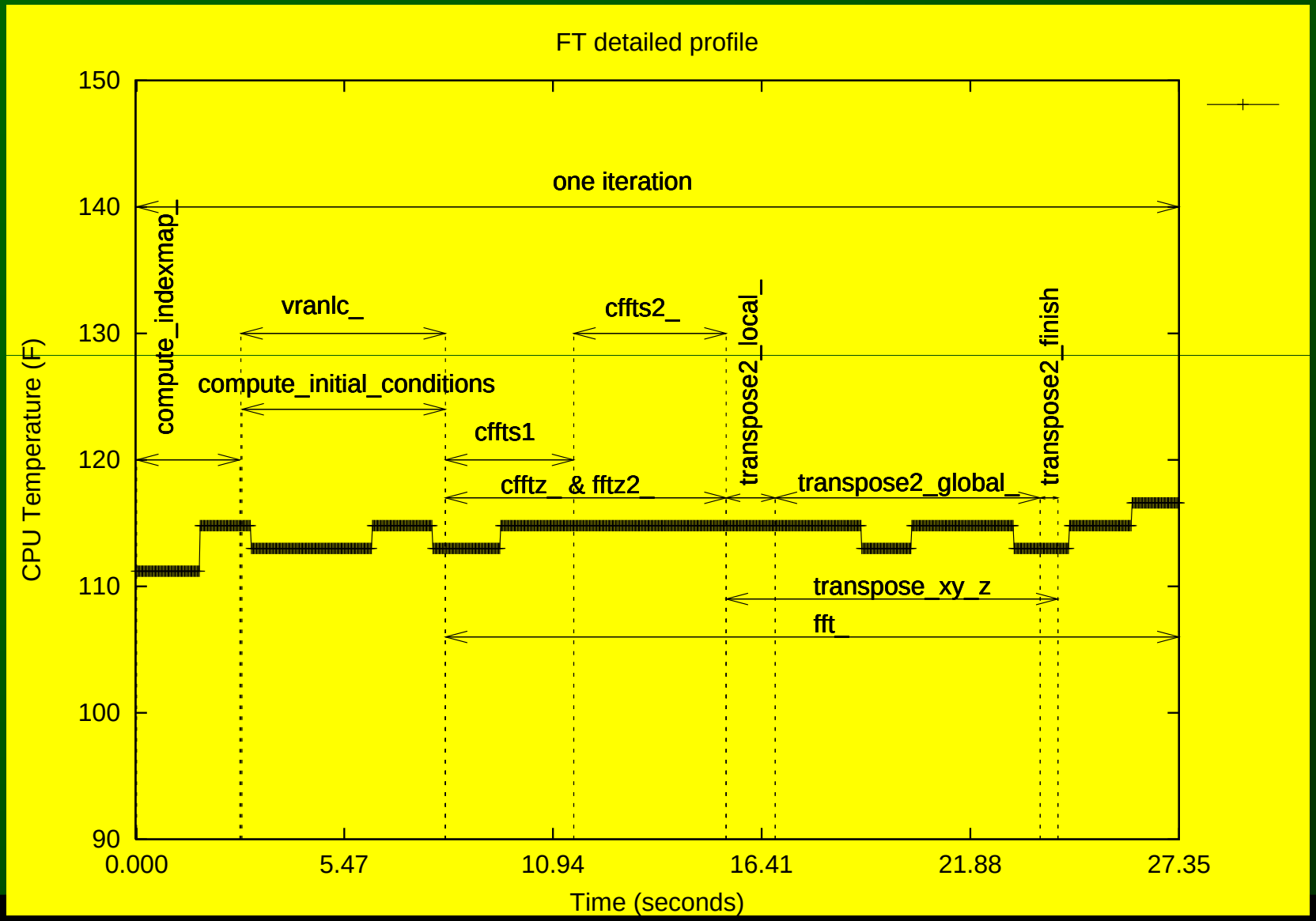
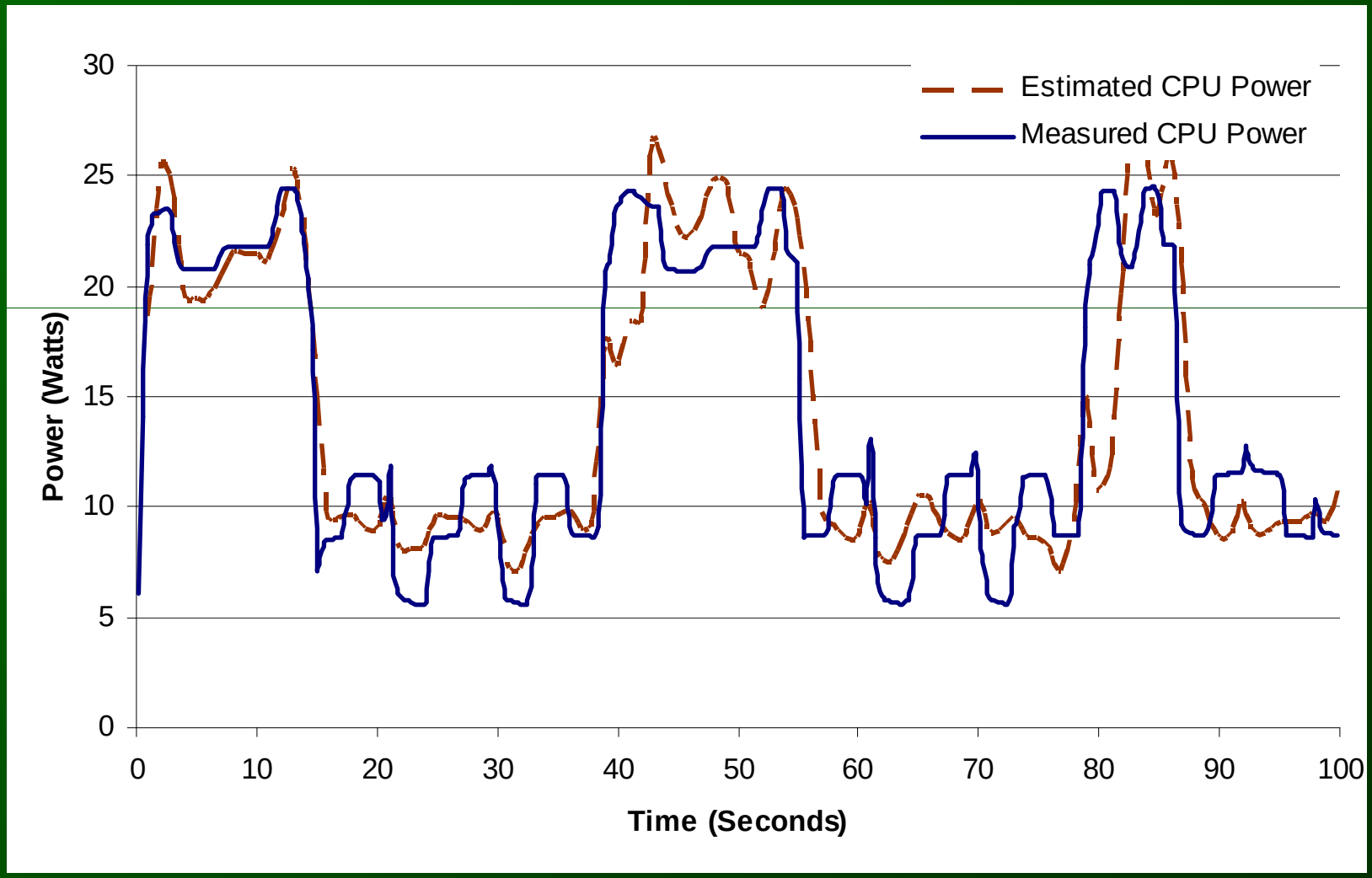


Fig. 7. Mapping between power profile and code segments for FT benchmark (Class B). Using code analysis and code-power profile synchronization mechanisms provided in PowerPack, we can map the power phases to each individual function and perform detailed power-efficiency analysis on selected code segments. This is useful when exploring function-level power-performance optimization.

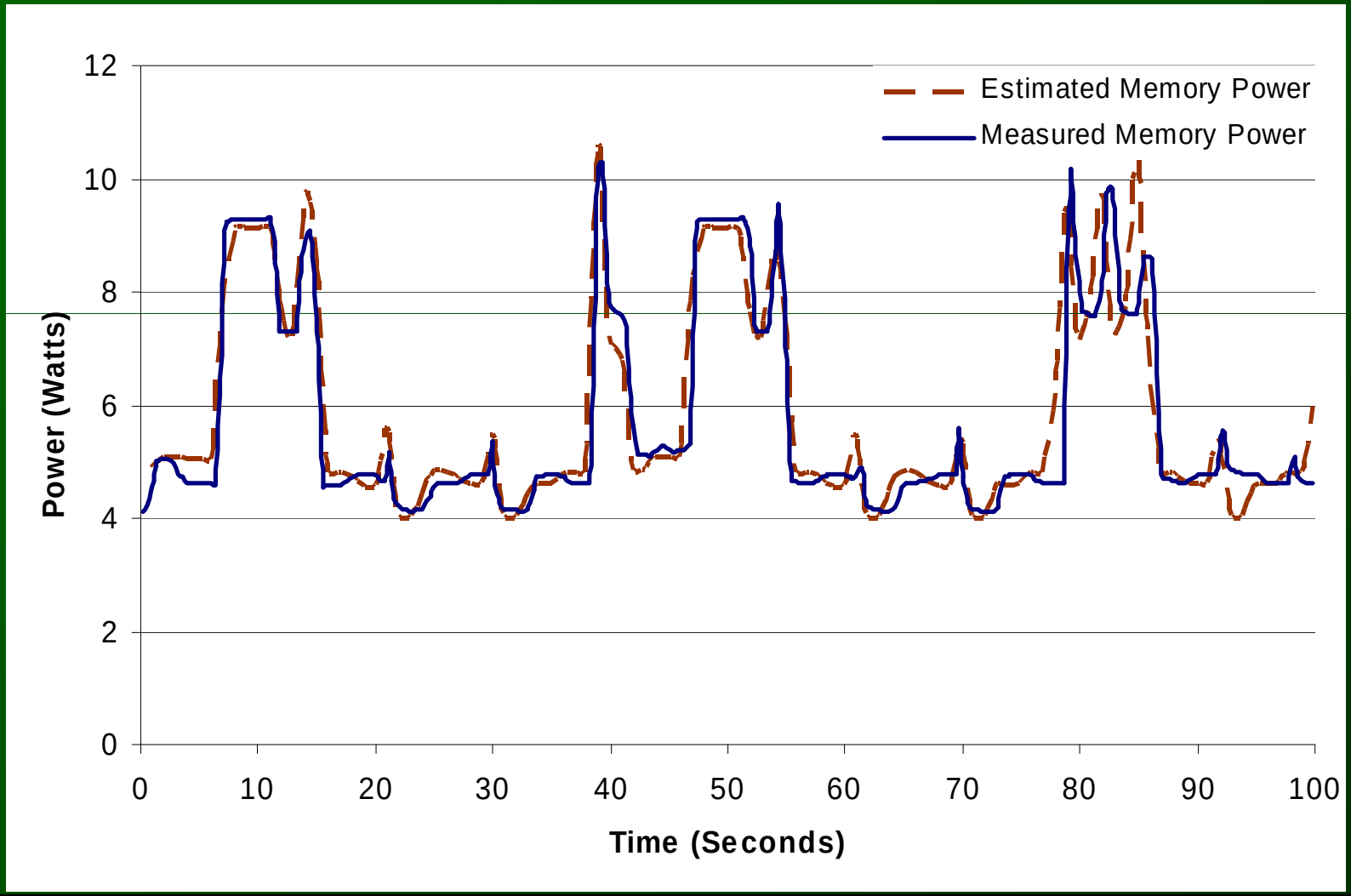
One FFT Iteration (Tempest)



Avoid Power Meas: Predict CPU Power



Avoid Power Meas: Predict Memory Power



Green HPC (using smart DVFS scheduling)

CPUSPEED Daemon

```
[example]$ start_cpuspeed
[example]$ mpirun –np 16 ft.B.16
```

Internal scheduling

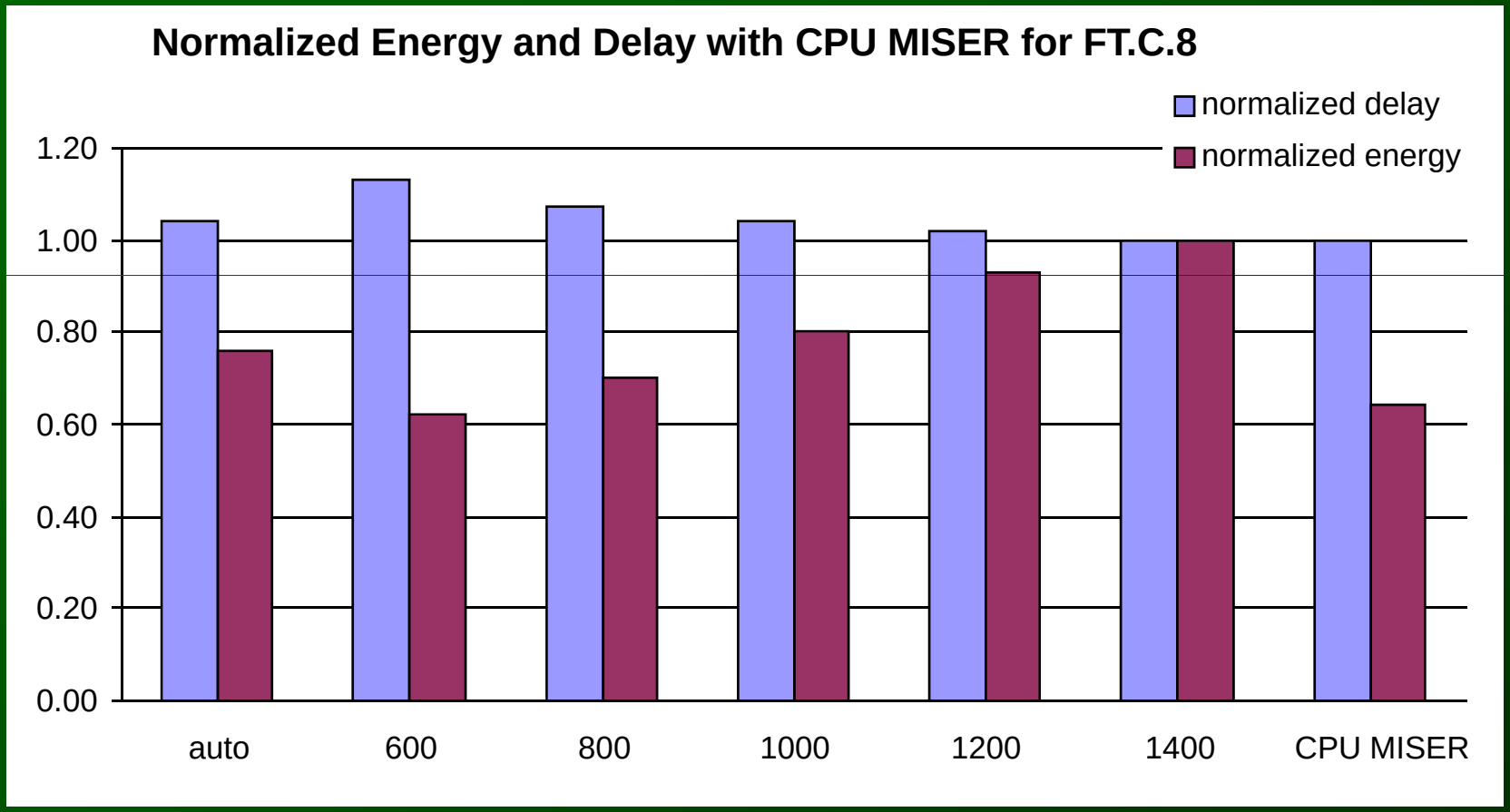
```
MPI_Init();
<CODE SEGMENT>
setspeed(600);
<CODE SEGMENT>
setspeed(1400);
<CODE SEGMENT>
MPI_Finalize();
```

External Scheduling

```
[example]$ psetcpuspeed 600
[example]$ mpirun –np 16 ft.B.16
```

NEMO & PowerPack Framework for saving energy

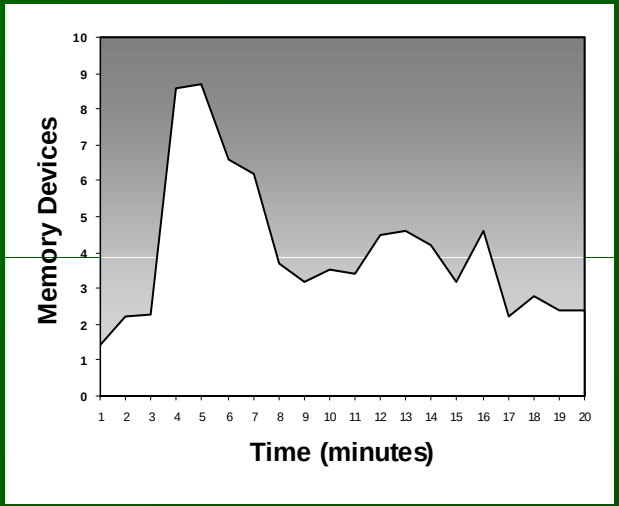
CPU MISER Scheduling (FT)



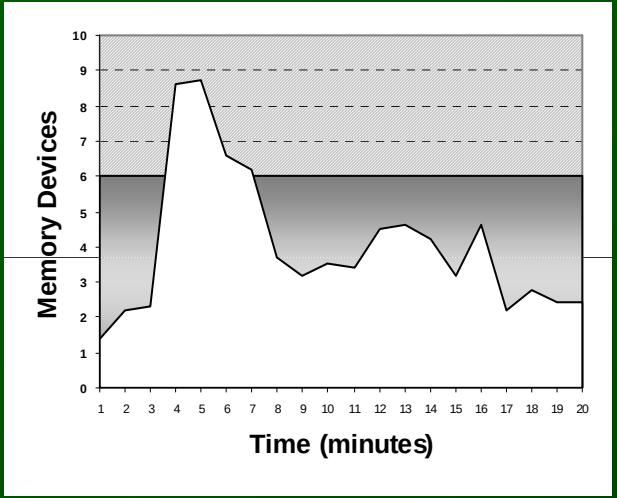
36% energy savings, less than 1% performance loss

See *SC2004, SC2005* publications.

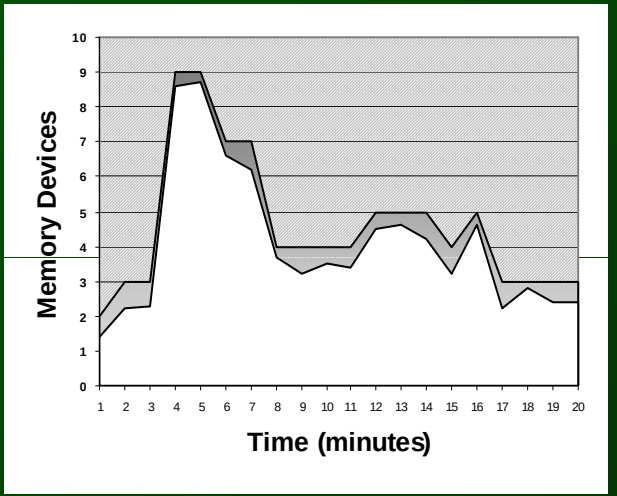
Green HPC (using smart memory management)



Default



Static



Dynamic

Memory MISER =

Page Allocation Shaping + Allocation Prediction + Dynamic Control

The future…my POV

- Will Moore's Law continue?
  - Absolutely. Until at least 2012. Power still a problem
- Will there be a silver bullet for Green HPC?
  - No. Exotic solutions succumb to commodity market.
  - Vendors will sell silver bullets like magic beans.
  - Look for solutions across the stack
    - New challenge: integration

The future…what's next?

- Information-driven control
  - Streaming sensor data from any source
  - Power & thermal sensors on board, PDUs, wireless, etc.
- Look for policy (EPA) to impact commodity → HPC
  - Example: power supply efficiencies, Energy Star ratings
  - How will you build a cluster from non-Energy Star equip
- HPC will *eventually* "turn on" power saving software
  - May not have a choice (policy, commodities, etc)
  - We already do at the micro-architecture level
  - Job scheduling easy place to start, virtualization will help
  - All techniques I discussed are easily integrated in OS

Thanks for listening. cameron@cs.vt.edu

See http://green500.org

See http://www.spec.org/specpower/

See http://www.energystar.gov/

See http://www.uptimeinstitute.org/

Thanks to our sponsors: NSF (Career, CCF, CNS, CRI), DOE (SC), Intel

Kirk W. Cameron Virginia Tech