28 July 2011

Computer system energy management

Charles Lefurgy

Outline

- A short history of server power management
- POWER7 EnergyScale
- AMESTER power measurement tool
- Challenges ahead

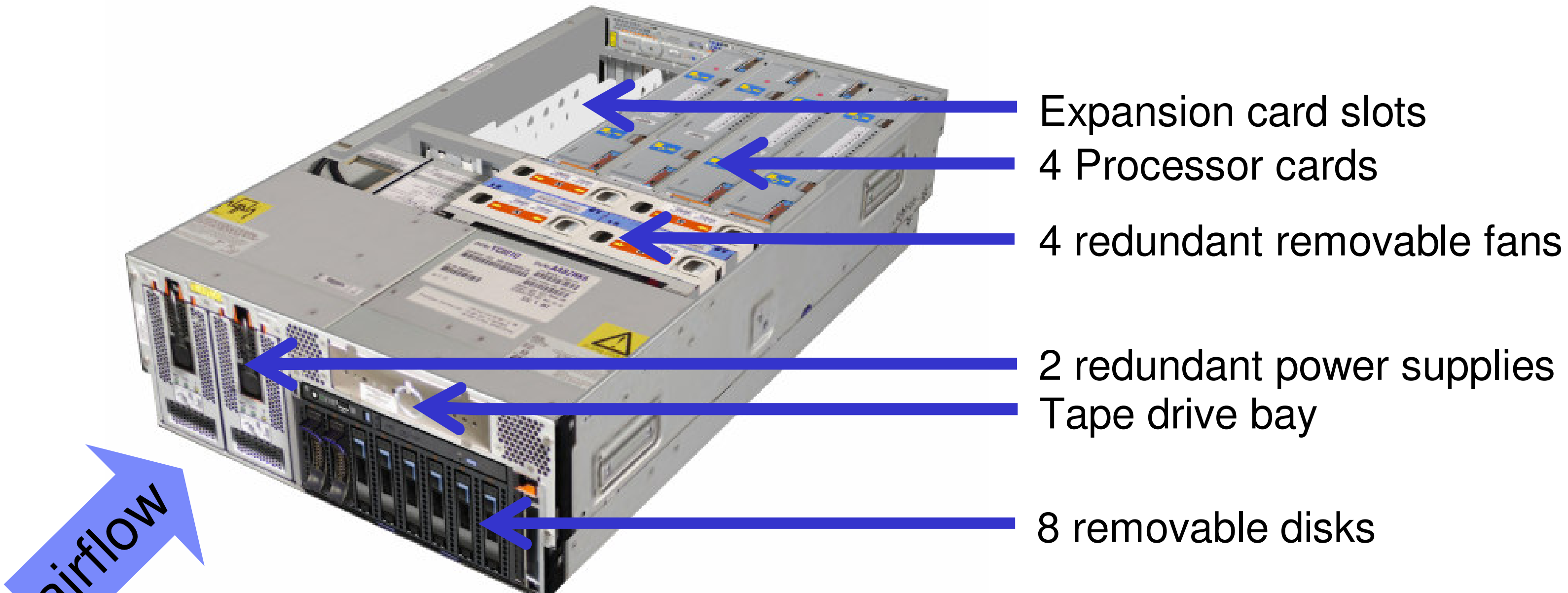
A brief history of server power management in IBM

2005:	Power measurement in servers. **Industry first**	Measure
2006:	IBM PDU+ (power cord-based power measurement)	
2007:	Power capping in servers. **Industry first**	Improve reliability
2008:	POWER6 with DVFS	Save energy
	dynamic DRAM power management	
2010:	POWER7 uses DDR3 self-refresh mode	
	POWER7 with Turbo mode	Improve performance
	Partition-aware power capping	Support virtualization
2011:	Partition-aware DVFS	

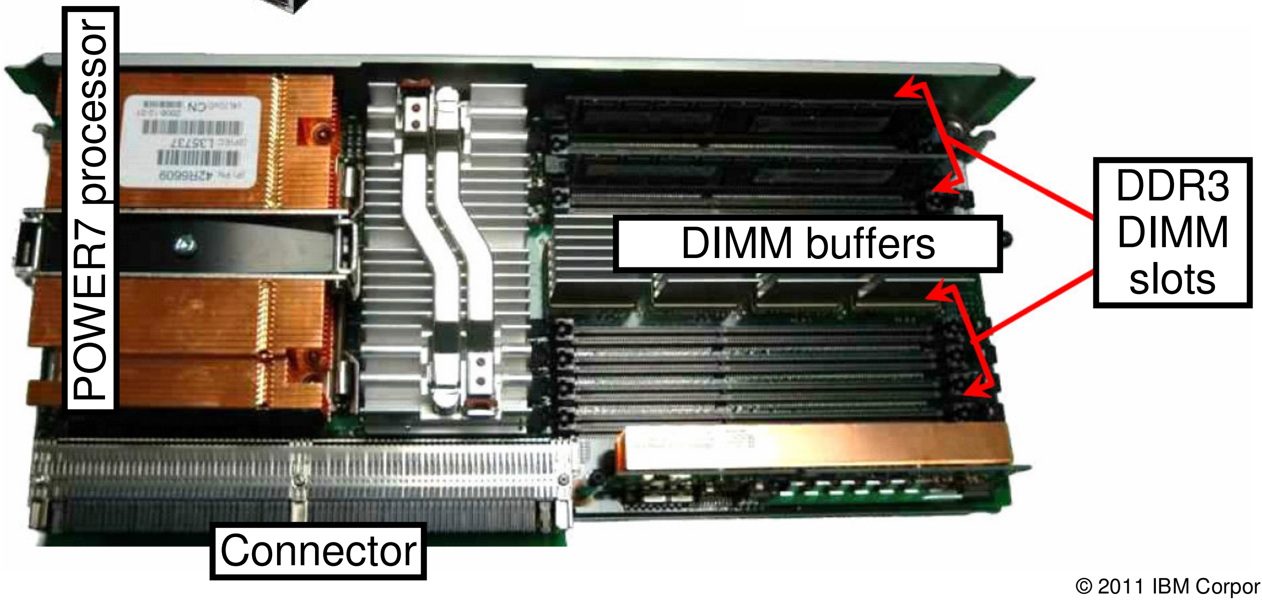
POWER7 energy management

Source: M. Floyd, B. Brock, M. Ware, K. Rajamani, A. Drake, C. Lefurgy and L. Pesantez, “Harnessing the Adaptive Energy Management Features of the POWER7 chip”, Hot Chips 22, 2010.

IBM POWER 750 Express

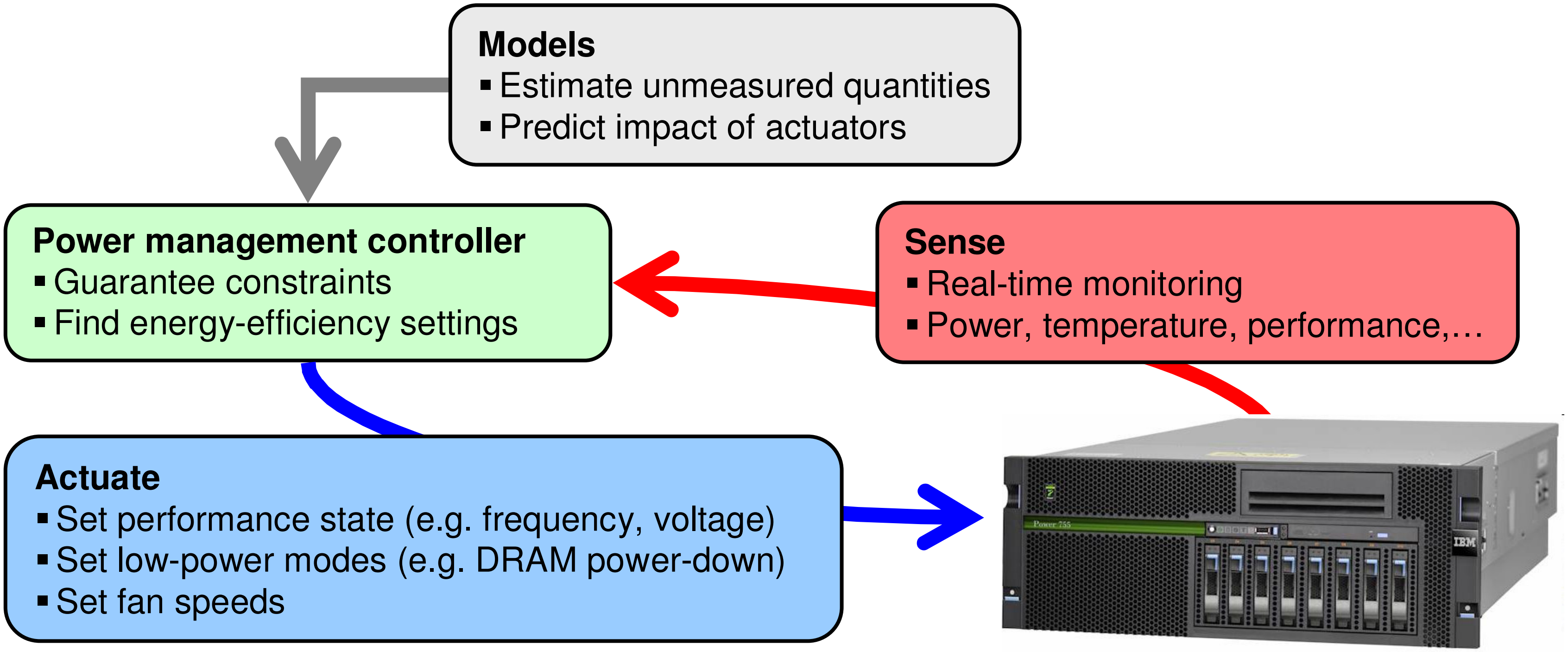


Processor card



Address variability in hardware and operating environment

- Complex environment
  - Installed component count, ambient temperature, component variability, etc.
  - How to guarantee power management constraints across all possibilities?
- Feedback-driven control
  - Capability to adapt to environment, workload, varying user requirements
  - Regulate to desired constraints even with imperfect information



Models
- Estimate unmeasured quantities
- Predict impact of actuators

Power management controller
- Guarantee constraints
- Find energy-efficiency settings

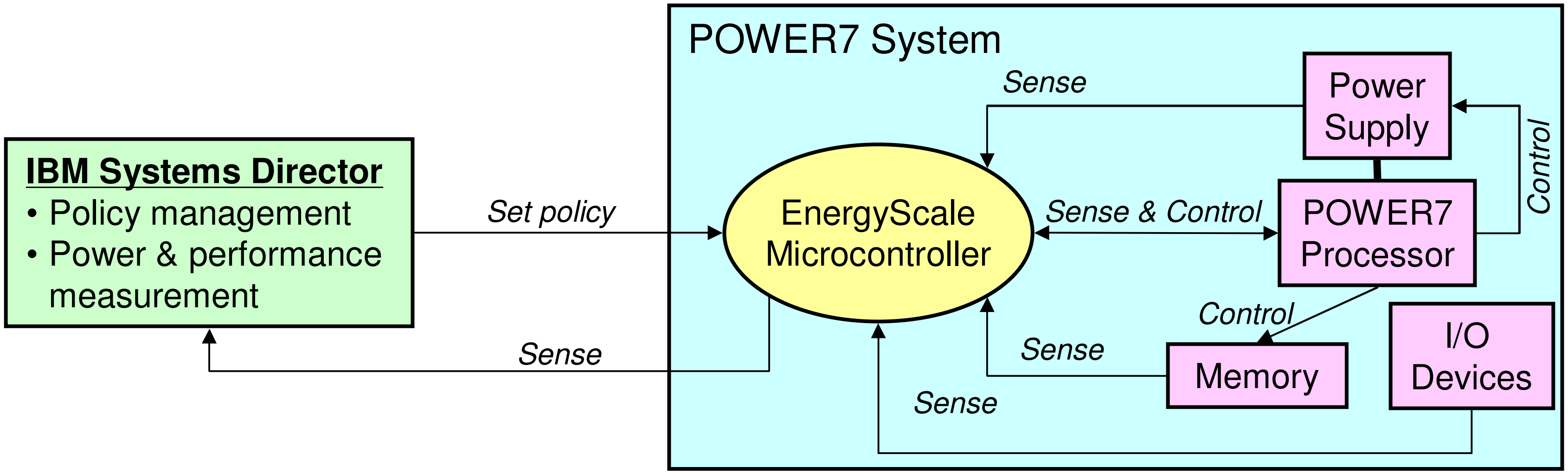
Sense
- Real-time monitoring
- Power, temperature, performance,…

Actuate
- Set performance state (e.g. frequency, voltage)
- Set low-power modes (e.g. DRAM power-down)
- Set fan speeds

EnergyScale

- Cooperative hardware and software solution for power management
  - EnergyScale firmware runs on dedicated microcontroller
    - DVFS, power capping, fan control, etc.
  - POWER7 microprocessor has hardware accelerators for power management
    - Sensor gathering, thermal sensor conversion, power proxy calculation, etc.

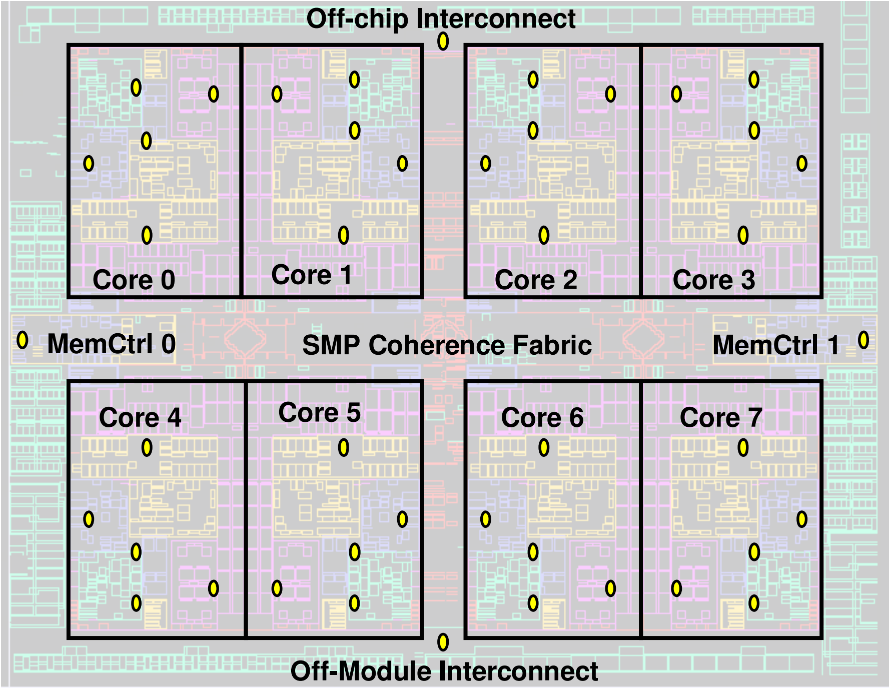
- Goals
  - Increase performance
  - Reduce power at same performance level



POWER7 sensors

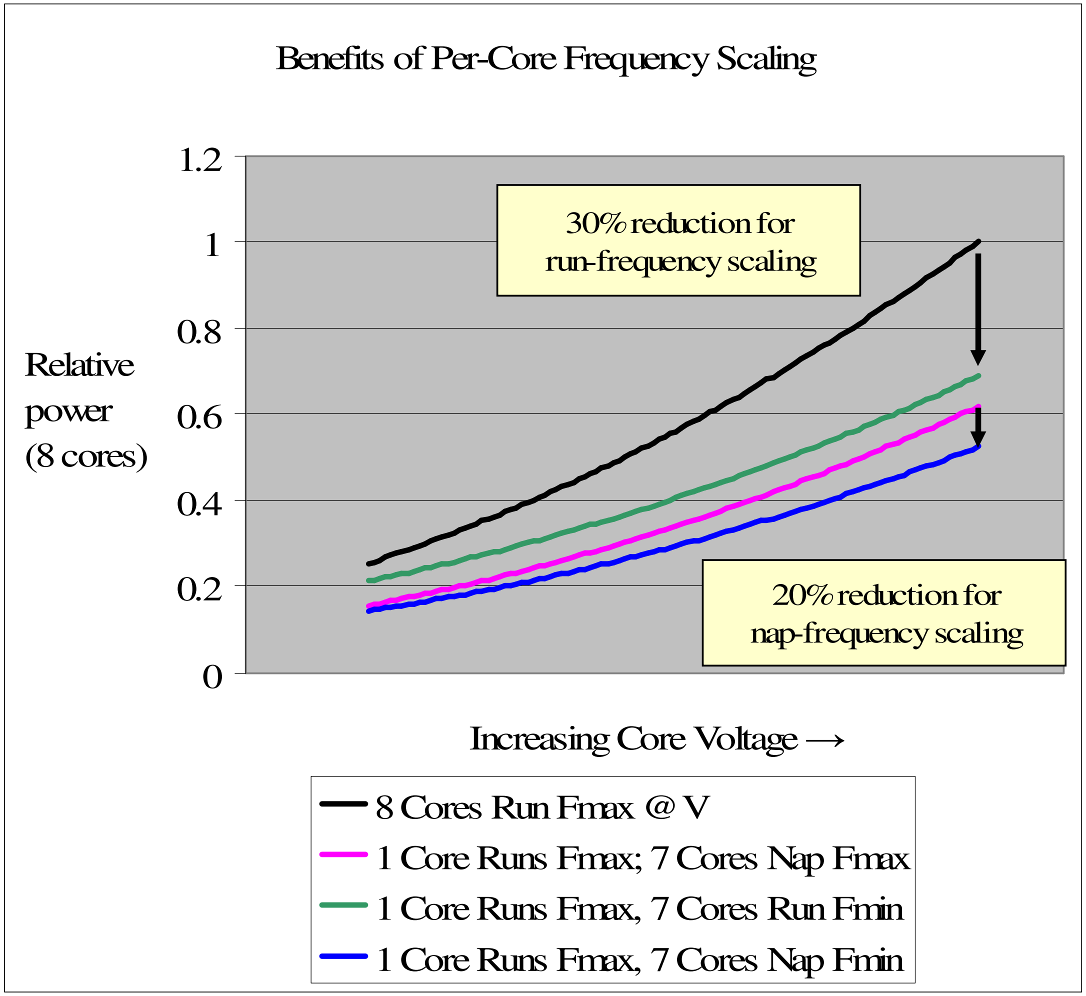
- POWER7 microarchitecture activity & event counters
  - Processor core, memory hierarchy, and main memory access
  - Provide performance, utilization, and activity measurements
  - Used to direct power/performance tradeoff decisions & techniques
- POWER7 Digital Thermal Sensors
  - 44 on-chip sense points
- POWER7 Critical Path Monitor
  - Detects circuit timing margin
- System sensors
  - Fan speed
  - Power by voltage domain
  - Temperature by component

Physical Locations of Thermal Sensors



Performance control

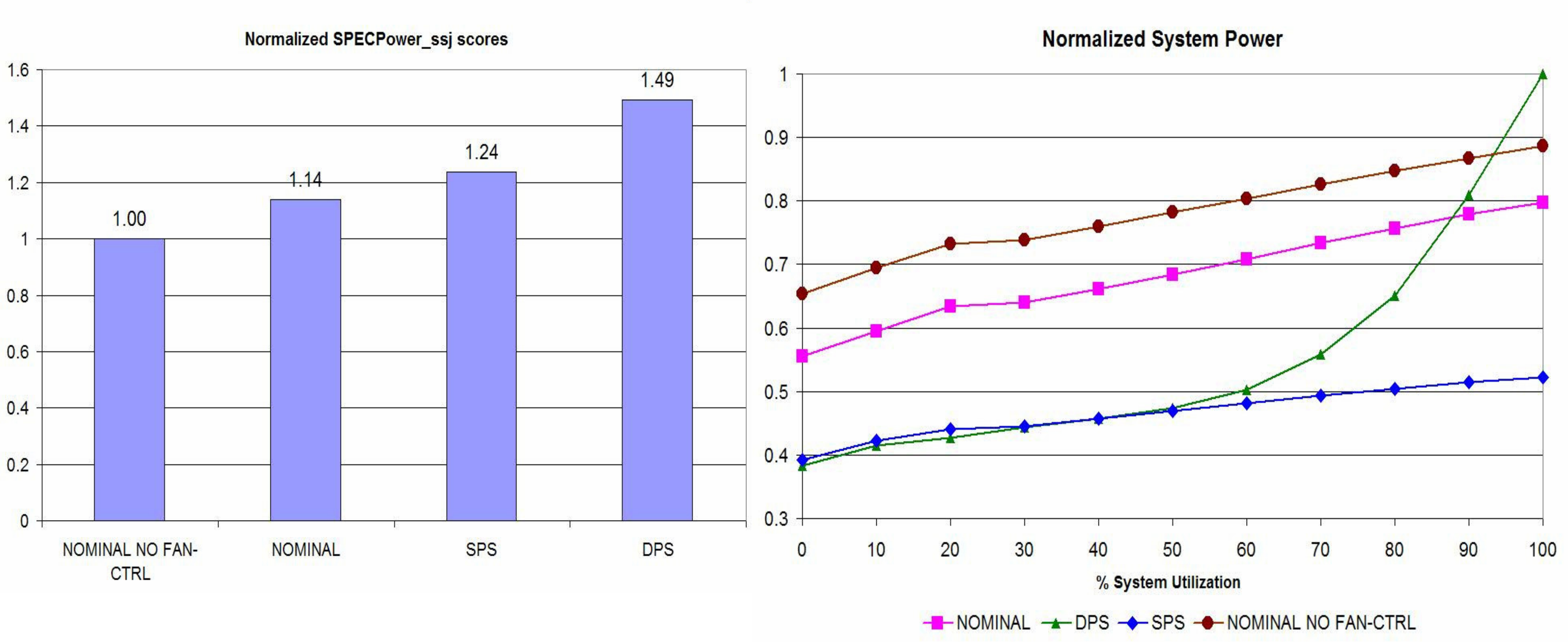
- Per-core frequency control
  - Digital PLL (DPLL) clock source supports -50% to +10% of nominal frequency
  - 25Mhz resolution
  - Automated fast frequency slew in excess of 50Mhz per us
- Supports energy optimization in partitioned system configurations
- Each partition can run under different energy-savings policy
  - Less-utilized partitions can run at lower frequencies
  - Heavily utilized partitions maintain peak performance
- EnergyScale Dynamic Power Savings algorithm looks for workload slack in time and across cores



Note: highest frequency core determines the required voltage

Dynamic power savings

- SPECPower_ssj2008 running on IBM Power 750 Express system**
- SPS (Static Power Save): fixed, low-power operating point = improved score almost 25%
- DPS (Dynamic Power Savings): DVFS with Turbo mode = improved score almost 50%

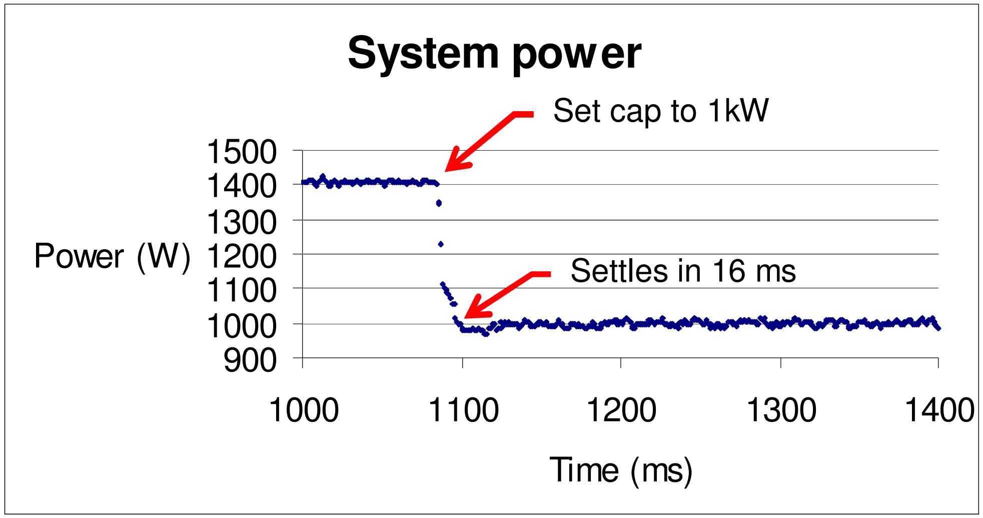


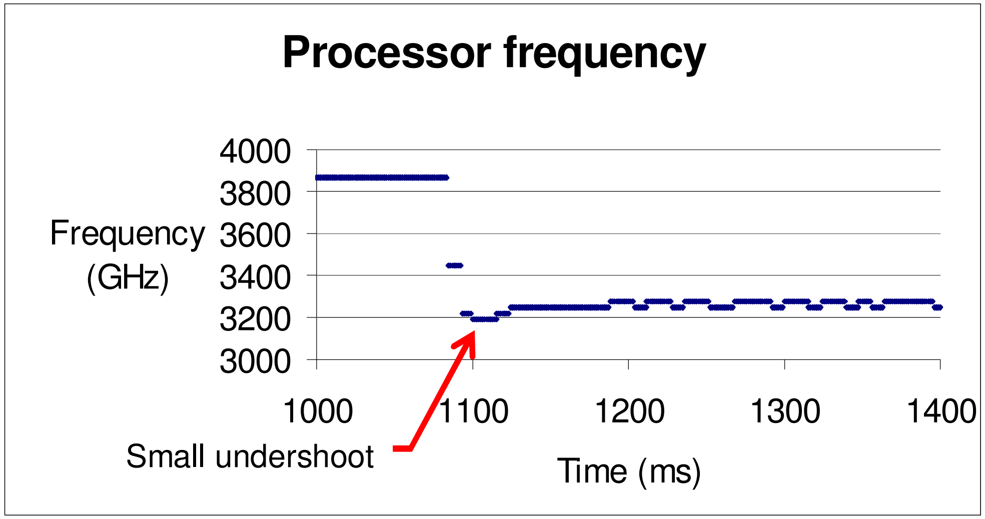
Source: Heather Hanson, IBM research

* Results shown on our prototype system, should not be construed as committed capability for a shipping IBM Server.
* SPEC and the benchmark name SPECpower_ssj are trademarks of the Standard Performance Evaluation Corporation

Power capping controller

- Caps when redundant power supply fails or customer sets power cap target
- Every 8 ms, measure system power and adjust processor voltage and frequency to meet power cap
- Partition-aware: take down frequency of Dynamic Power Saving partitions first.
- Precision measurement desired
  - Measurement error translates to lost performance
  - More guardband in power cap target
- Opportunity for improvement
  - On-line modeling
  - Relationship of frequency to power changes over time





Summary

- POWER7 builds upon initial POWER6™ EnergyScale features by including automated on-chip functions and accelerators to assist the off-chip microcontroller firmware

- POWER7 energy management features combined with new energy-saving algorithms show a **50%** improvement in SPECpower score over baseline operation

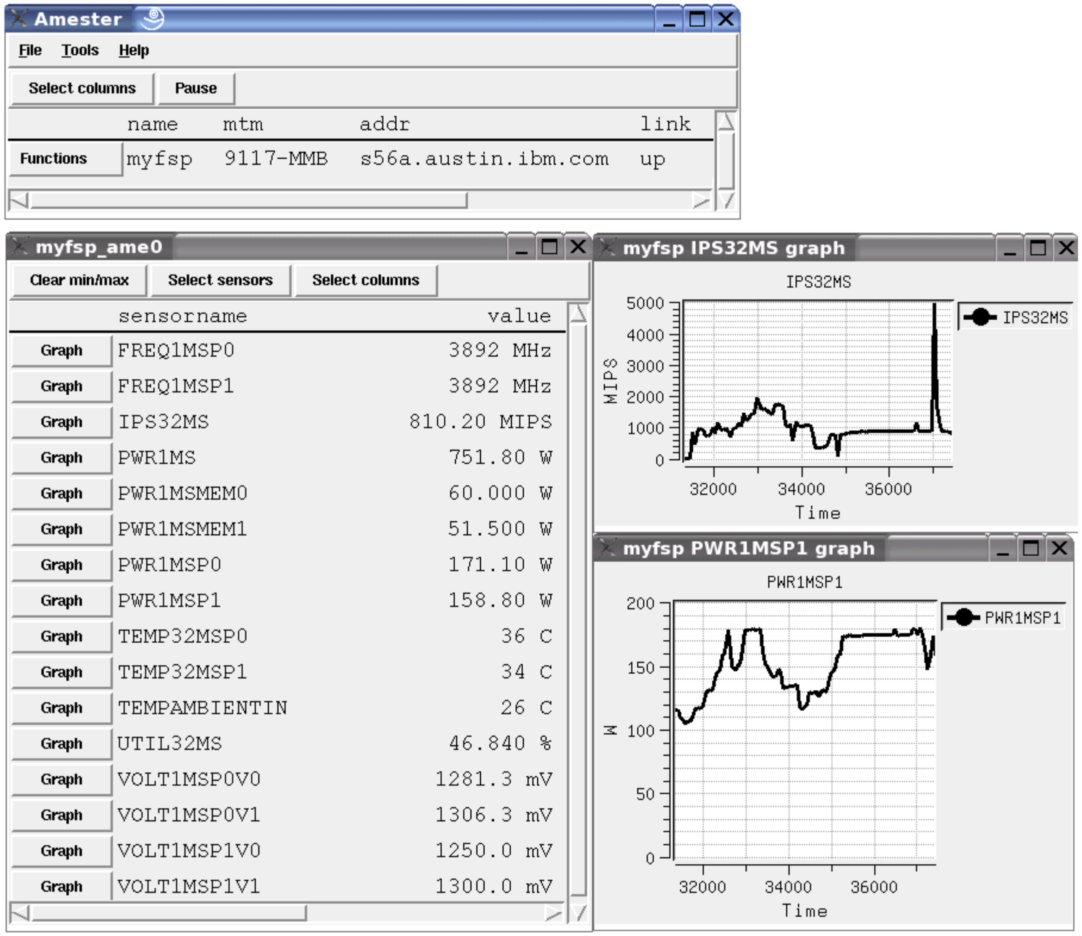
- Customers can select the best EnergyScale policy to match their needs, relying on the system to balance power consumption and performance accordingly

* Results shown on our prototype system, should not be construed as committed capability for a shipping IBM Server.
* SPEC and the benchmark name SPECpower_ssj are trademarks of the Standard Performance Evaluation Corporation

Measuring power in POWER7

AMESTER: Automated Measurement of Systems for Energy and Temperature Reporting

- A research tool for detailed monitoring and control of power consumption on a single IBM server
  - Non-intrusive, remote measurement
  - Interacts with server firmware
  - Scriptable for rapid prototyping
- Related product: IBM Systems Director Active Energy Manager
  - Monitors entire data centers

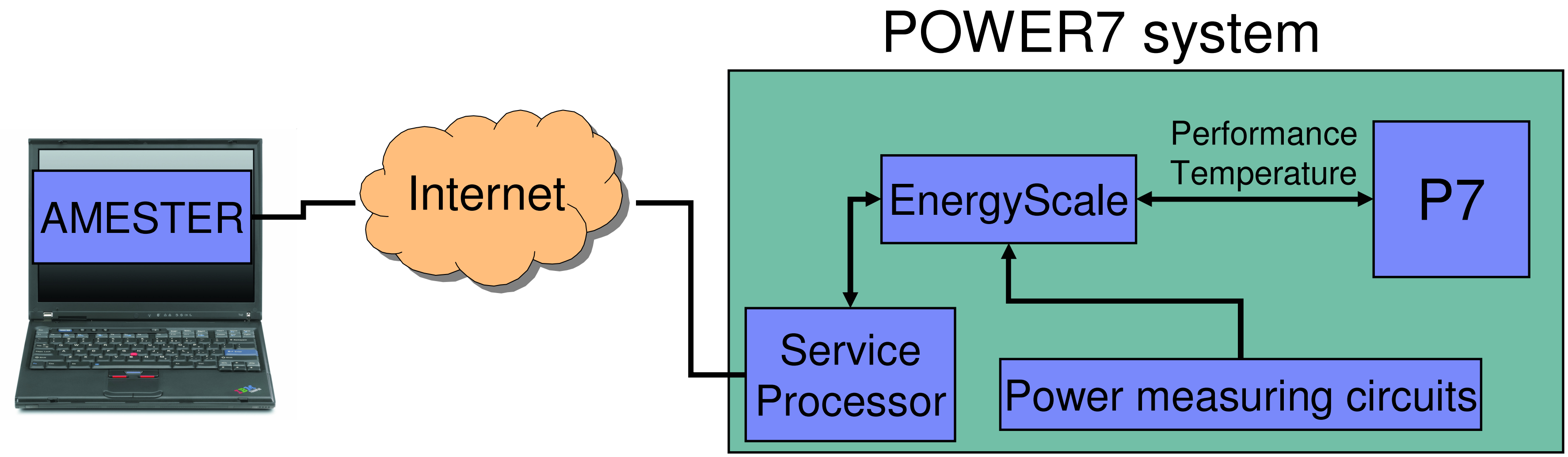


	name	mtm	addr	link
Functions	myfsp	9117-MMB	s56a.austin.ibm.com	up

	sensorname	value
Graph	FREQ1MSP0	3892 MHz
Graph	FREQ1MSP1	3892 MHz
Graph	IPS32MS	810.20 MIPS
Graph	PWR1MS	751.80 W
Graph	PWR1MSMEM0	60.000 W
Graph	PWR1MSMEM1	51.500 W
Graph	PWR1MSP0	171.10 W
Graph	PWR1MSP1	158.80 W
Graph	TEMP32MSP0	36 C
Graph	TEMP32MSP1	34 C
Graph	TEMPAMBIENTIN	26 C
Graph	UTIL32MS	46.840 %
Graph	VOLT1MSP0V0	1281.3 mV
Graph	VOLT1MSP0V1	1306.3 mV
Graph	VOLT1MSP1V0	1250.0 mV
Graph	VOLT1MSP1V1	1300.0 mV

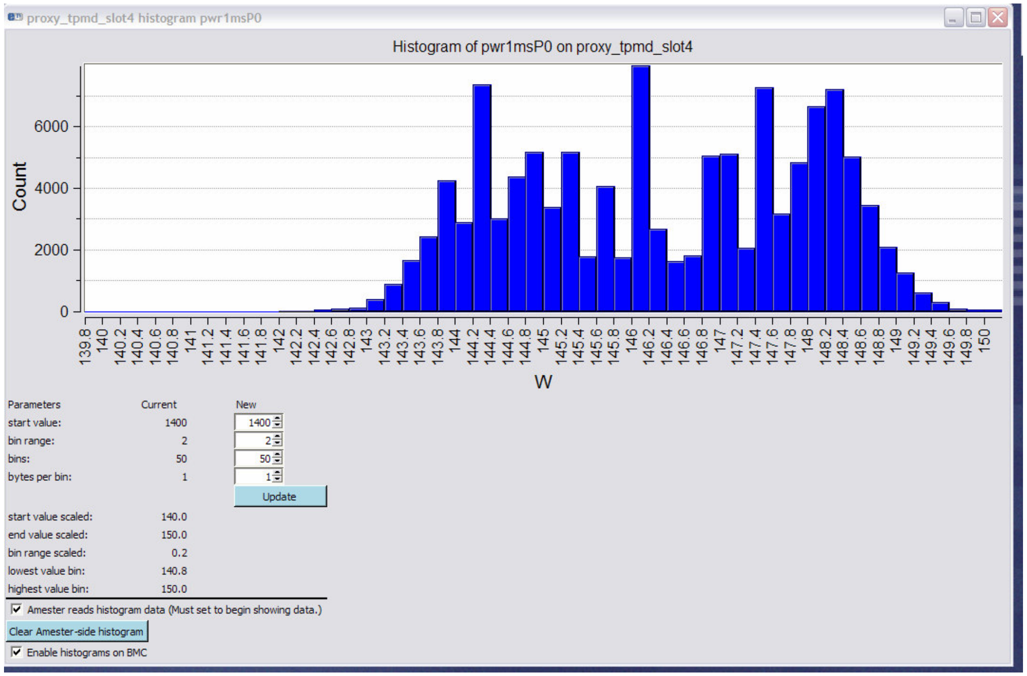
How it works

- AMESTER runs on a laptop or server (Windows/Linux)
- Connects to remote system to measure
- EnergyScale microcontroller
  - Firmware for power management
  - Implements AMESTER command set
- Out-of-band data collection (no OS support required)



Basic functions included in AMESTER

- Sensor data collection
  - Whole system power measurement
    - Component power on POWER: CPU, DIMMs, fan, etc.
  - CPU temperature, CPU frequency, CPU utilization, voltage, instructions per second, etc.
  - Histograms
- High-resolution tracing
  - 1ms for sensors
- Power capping
- Scripting
  - Tcl command line
  - Job management library to run workloads remotely
  - Data library to collect and graph user data



Parameters	Current	New
start value:	1400	1400
bin range:	2	2
bins:	50	50
bytes per bin:	1	1

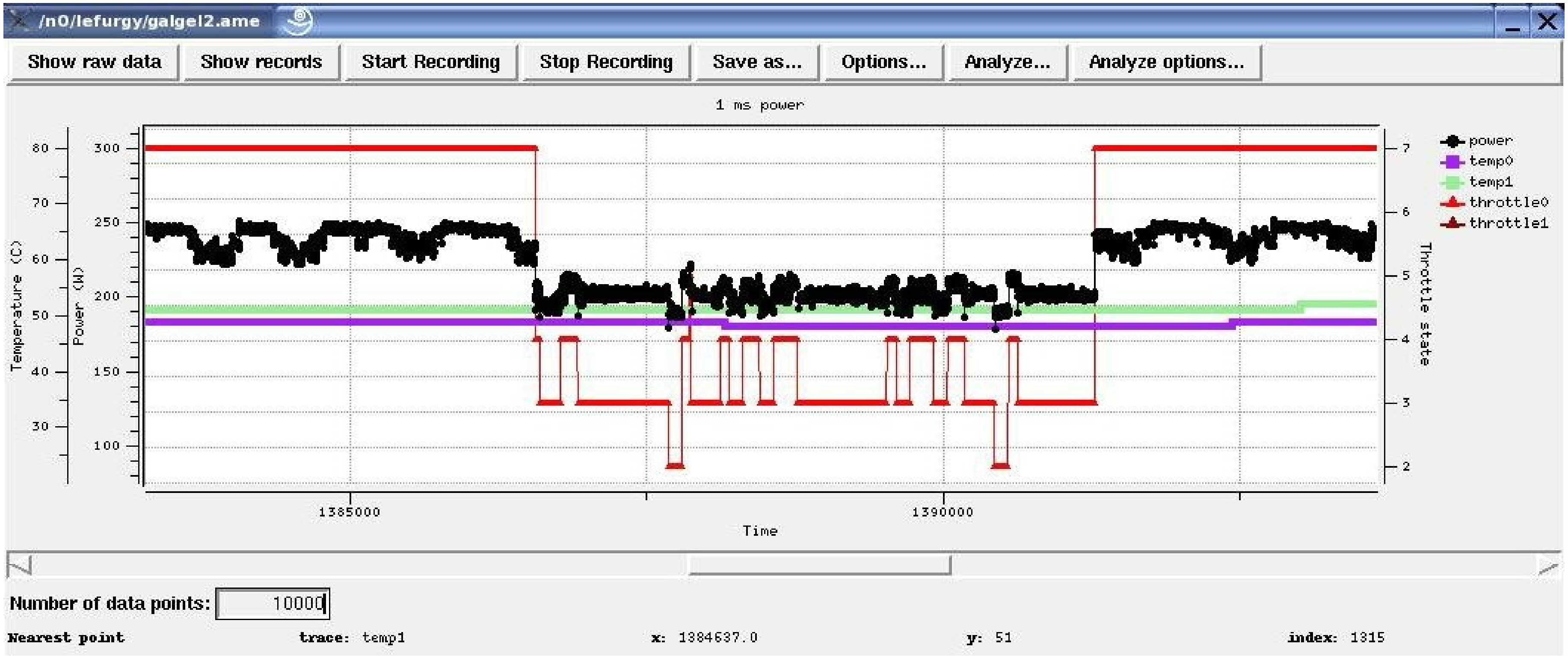
start value scaled:	140.0
end value scaled:	150.0
bin range scaled:	0.2
lowest value bin:	140.8
highest value bin:	150.0

☑ Amester reads histogram data (Must set to begin showing data.)

☑ Enable histograms on BMC

Insights

- Visualization is key to rapid prototyping and problem solving
  - Understand how power capping controller reacts to workload changes
- Correlation of power with other metrics
  - Study DVFS algorithm
  - High-resolution collection of core utilization, clock frequency, and performance
  - Example debugging: small blips in an otherwise steady-state behavior



Number of data points: 10000

Nearest point trace: temp1 x: 1384637.0 y: 51 index: 1315

Available for academic collaborations

- Current collaborations
  - National Center for Supercomputing Applications
  - Barcelona Supercomputing Center
  - Forschungszentrum Jülich

- Contact Charles Lefurgy (lefurgy@us.ibm.com)

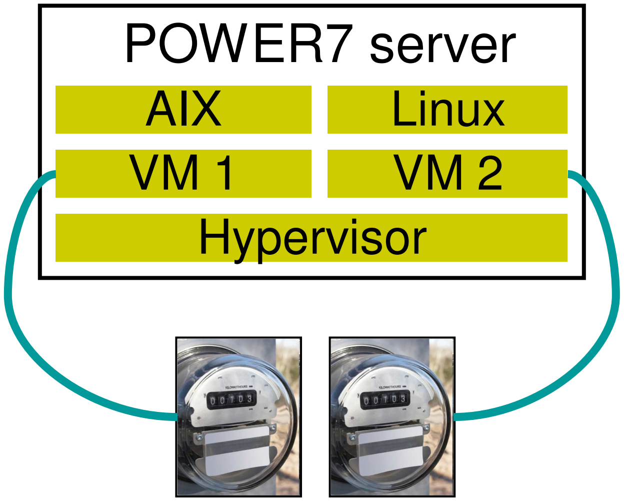
Challenges

May you live in interesting times

- Power management is a first class design consideration (from circuits to full systems)
- “Dark silicon” power limitations predicted to severely impact multi-core chip performance
  - H. Esmaeilzadeh,E. Blem, R. St Amant, K. Sankaralingam, and D. Burger, “Dark Silicon and the End of Multicore Scaling”, ISCA 2011.
  - W. Huang, K. Rajamani, M. R. Stan, and K. Skadron. "Scaling with Design Constraints – Predicting the Future of Big Chips." *IEEE Micro* special issue on Big Chips, July/Aug. 2011.
- Power and thermal management becomes more important for future technology nodes
- Opportunities:
  - Virtualization of power management (give end-user greater role)
  - Guardband reduction (save energy, increase performance)
  - Combining power measurement with other measurements for insights (save energy)
  - On-line modeling to improve all of the above

Measuring virtual machine power

- Clouds are managed on a virtual machine partition basis
  - Traditional platform power management is insufficient
- Measurement:
  - Today: Power measured at the power supply
  - How to bill each VM user for energy cost?
- Provide server owner with energy meter for each VM
  - Billing
  - Provisioning
  - Insight
- Use modeling to cover gaps in power measurement
  - Core-level power models based on “power proxies”
  - Allocate energy according to VM-to-hardware mapping
  - Per-socket measurements provide ability to learn and correct on-line models
- Unanswered questions
  - Fairness: How to pay for fan power? Splitting the bill is not fair for low-power VM.
  - Fairness: If a VM on core A, heats up core B, should B pay for the extra leakage power?
  - Infrastructure: How to charge for remote device power (network storage)?

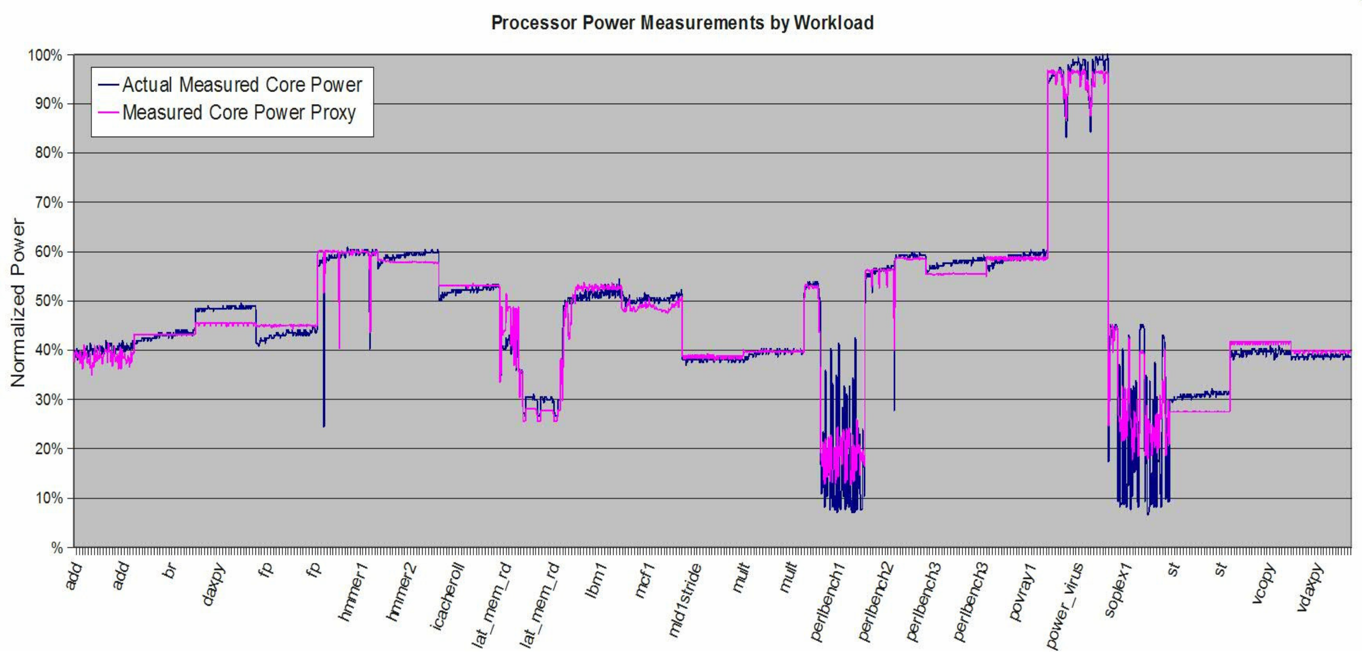


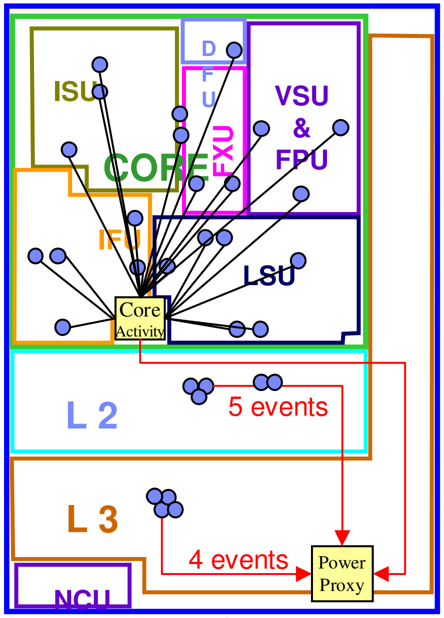
Processor Core Power Proxy

- Estimate per-core chiplet active power
- For each functional unit, pick subset of activities
- Weight each activity counter relative to power it consumes
- Sum weighted counter, clock grid power, and constant offset

$$\text{Chiplet Active Power} = \sum (W_i * A_i) + K^*f + C$$

- +/-10% error for 90% of samples





● = Activity Sense point

Processor Core Chiplet

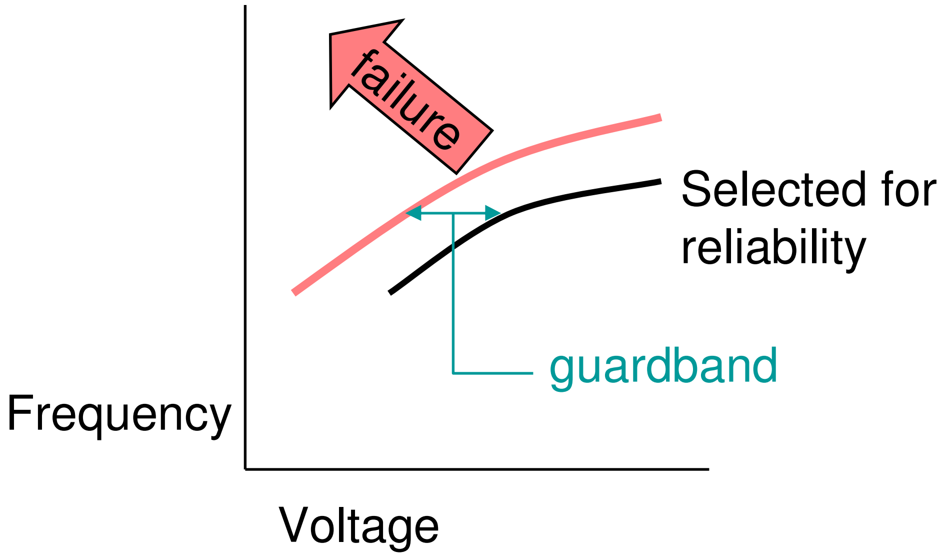
Guardband reduction

- The voltage used on a microprocessor is conservative to provide a safety cushion in case workload spikes causing noise or voltage droops

- **Guardband** is the difference between the operating voltage and the voltage at which the microprocessor fails

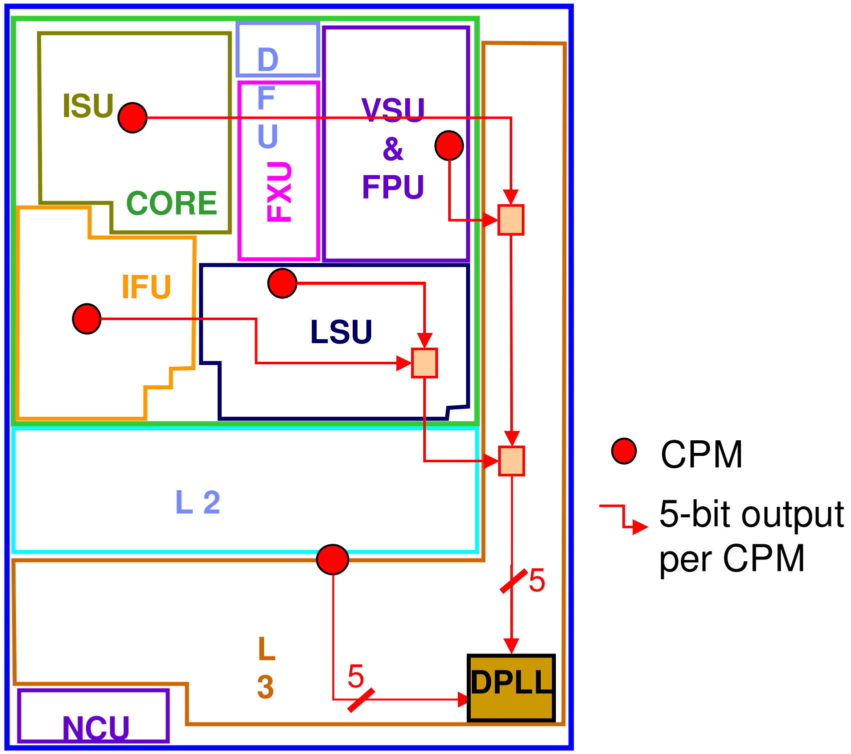
- **Concern:** Energy-efficiency is reduced to guarantee reliability.

Microprocessor operating points



Timing margin sensor

- Direct measurement of remaining timing margin with Critical Path Monitor (CPM)



POWER7 core chiplet

CPM output

“11111” = large margin
“11110” = some margin
“11100” = ideal margin
“11000” = margin too small
“10000” = not enough margin

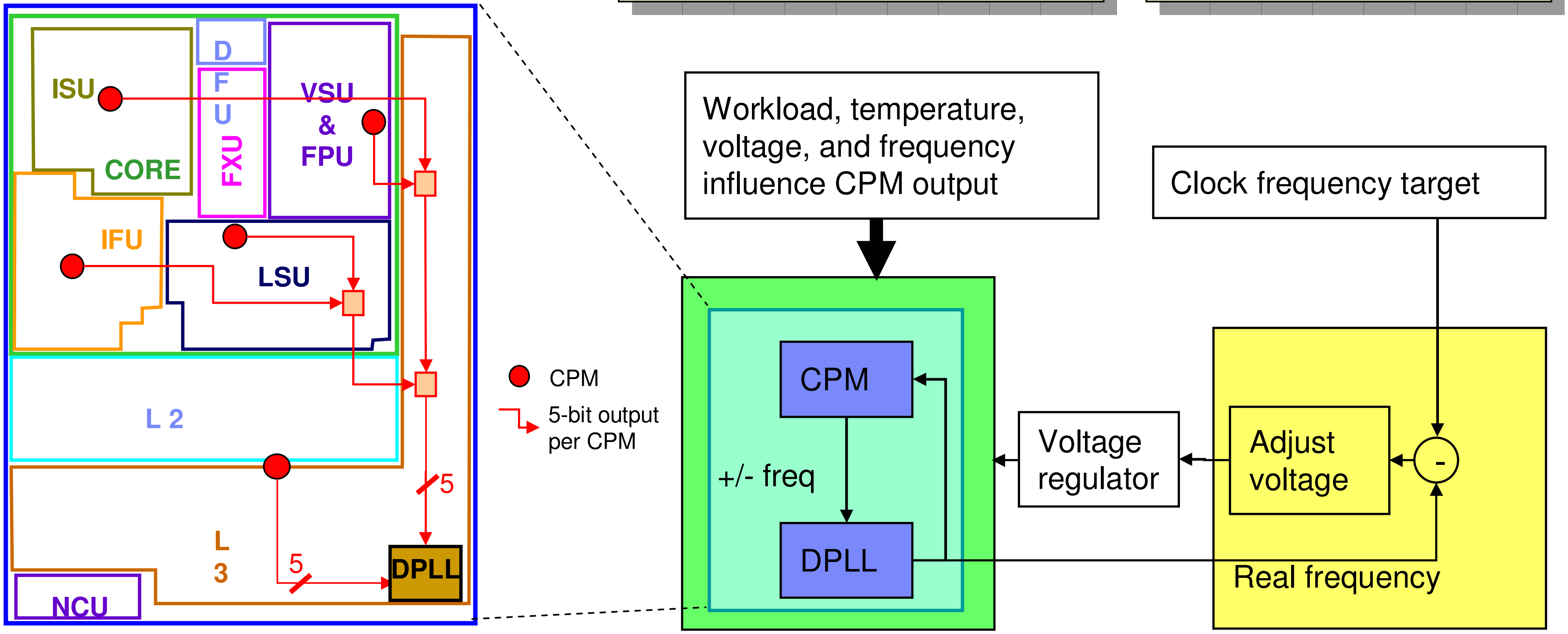
Undervolting solution

Protect

Timing margin controller responds to changing operating conditions by setting highest, safe frequency that avoids timing failures.

Optimize

Performance controller adjusts voltage to meet desired clock frequency target.



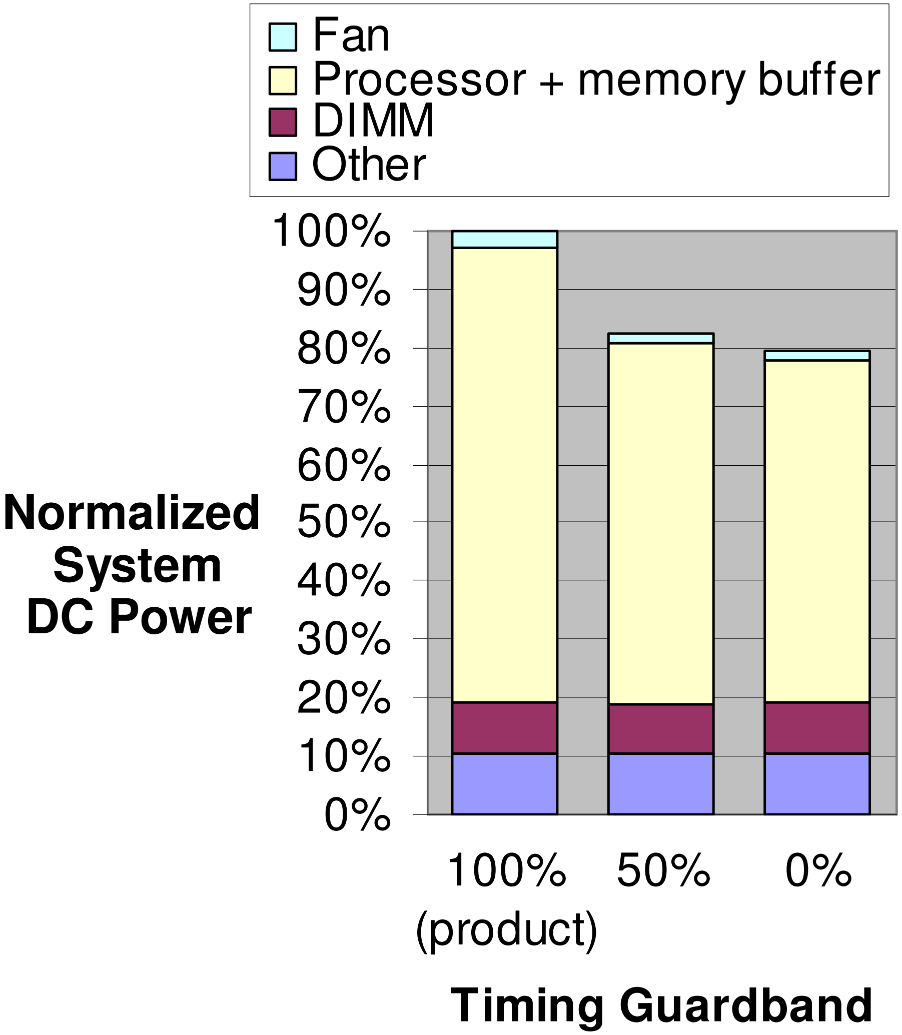
POWER7 core chiplet

POWER7

Microcontroller

Undervolting results

- Prototype POWER 750 Express server
  - Run SPEC CPU 2006 workloads
- 20% power reduction of POWER7 processor + memory buffer
- 18% power reduction in system power
  - Fan power is reduced by 50%
- No change in performance



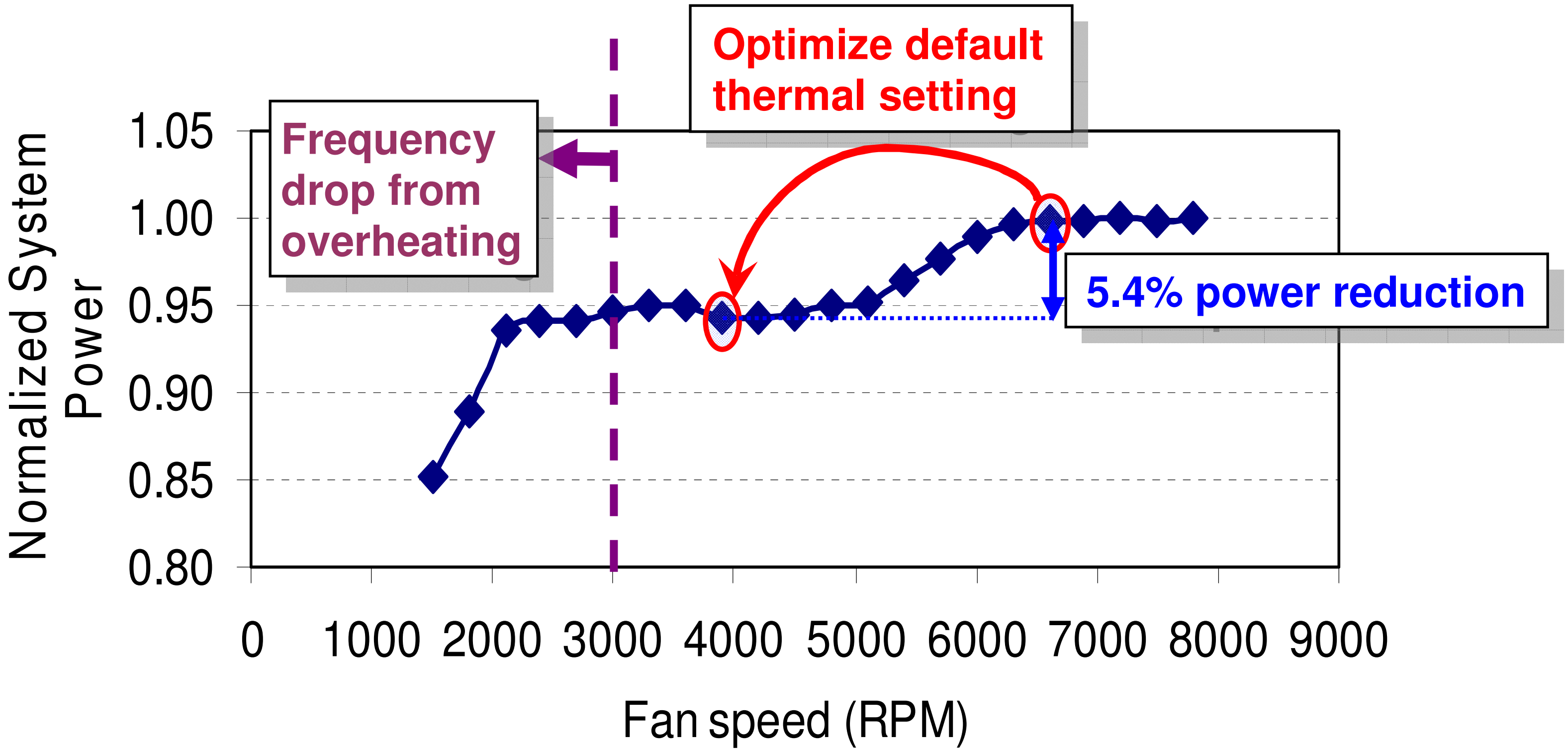
Guardband reduction to save energy in POWER7 prototype server

Click box to play video

Undervolting Video

Using power measurement with other metrics

- TAPO: Thermal-aware power optimization
  - Minimize total of server fan power and microprocessor leakage power
  - Reduces server power 5% at peak Turbo performance in P7 server prototype
  - No performance loss



Source: W. Huang, M. Allen-Ware, J. Carter, E. Elnozahy, H. Hamann, T. Keller, C. Lefurgy, J. Li, K. Rajamani, and J. Rubio, “TAPO: Thermal-Aware Power Optimization Techniques for Servers and Data Centers”, IGCC, 2011.

Summary

- Server power management has made significant progress in just a few years
  - Extending for virtual machines
  - Extending for reliability

- Power measurement is the basis for server power management

- Many improvements yet to come
  - Wider coverage of components
  - Shorter timescale measurements and correlation to other metrics
  - Self-tuning on-line modeling

- Measurement improvements lead to guardband reduction
  - Improve performance, save energy, lower cost, improve reliability

End