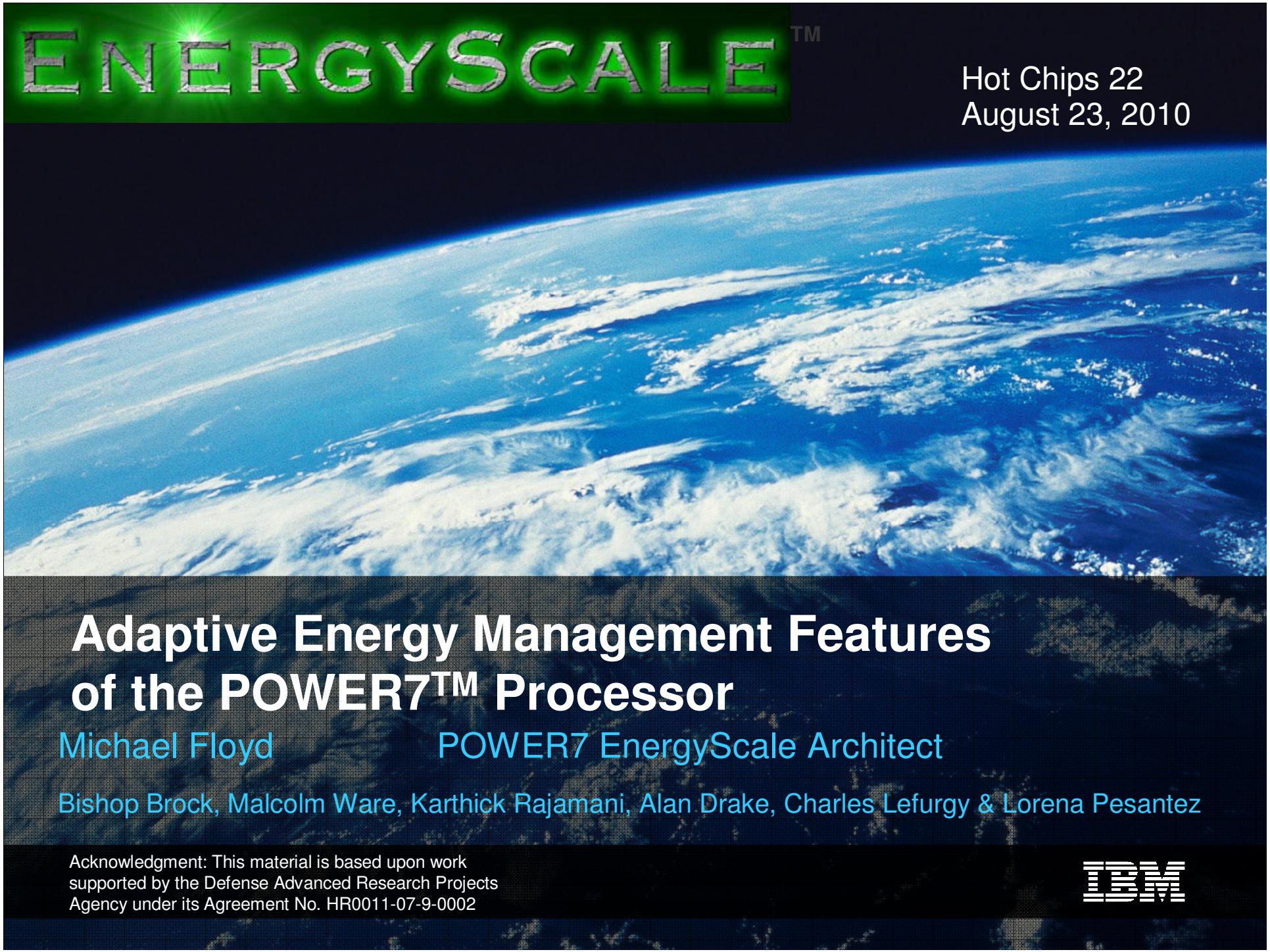


Hot Chips 22
August 23, 2010

Adaptive Energy Management Features of the POWER7™ Processor

Michael Floyd POWER7 EnergyScale Architect

Bishop Brock, Malcolm Ware, Karthick Rajamani, Alan Drake, Charles Lefurgy & Lorena Pesantez

Acknowledgment: This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0002

IBM

Outline

- EnergyScale™ Overview
- POWER7 Energy Management Features
- New POWER7 Autonomic Mechanisms

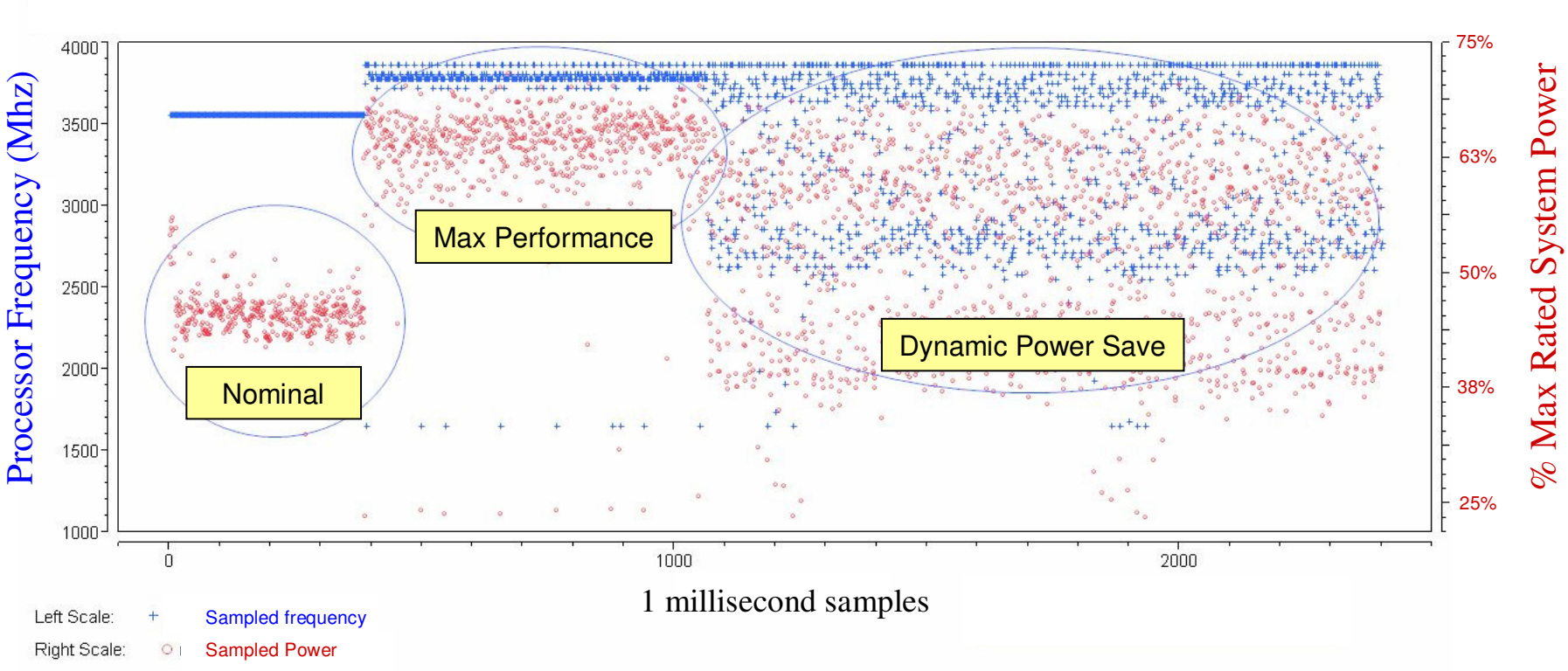
* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

POWER7 EnergyScale™ Goals

- **System-level performance-aware energy management**
  - Implement customer-selected energy management policy
  - Directly measure:
    performance, utilization, power consumption, temperature
- **Take advantage of workload and environment**
  - Save energy when not fully utilized
  - Optimize frequency and voltage to match
    - *needs of workload*
    - *limits of environment*
- **Apply benefits to either:**
  - Increased performance OR
  - Reduced power at the same performance level

* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

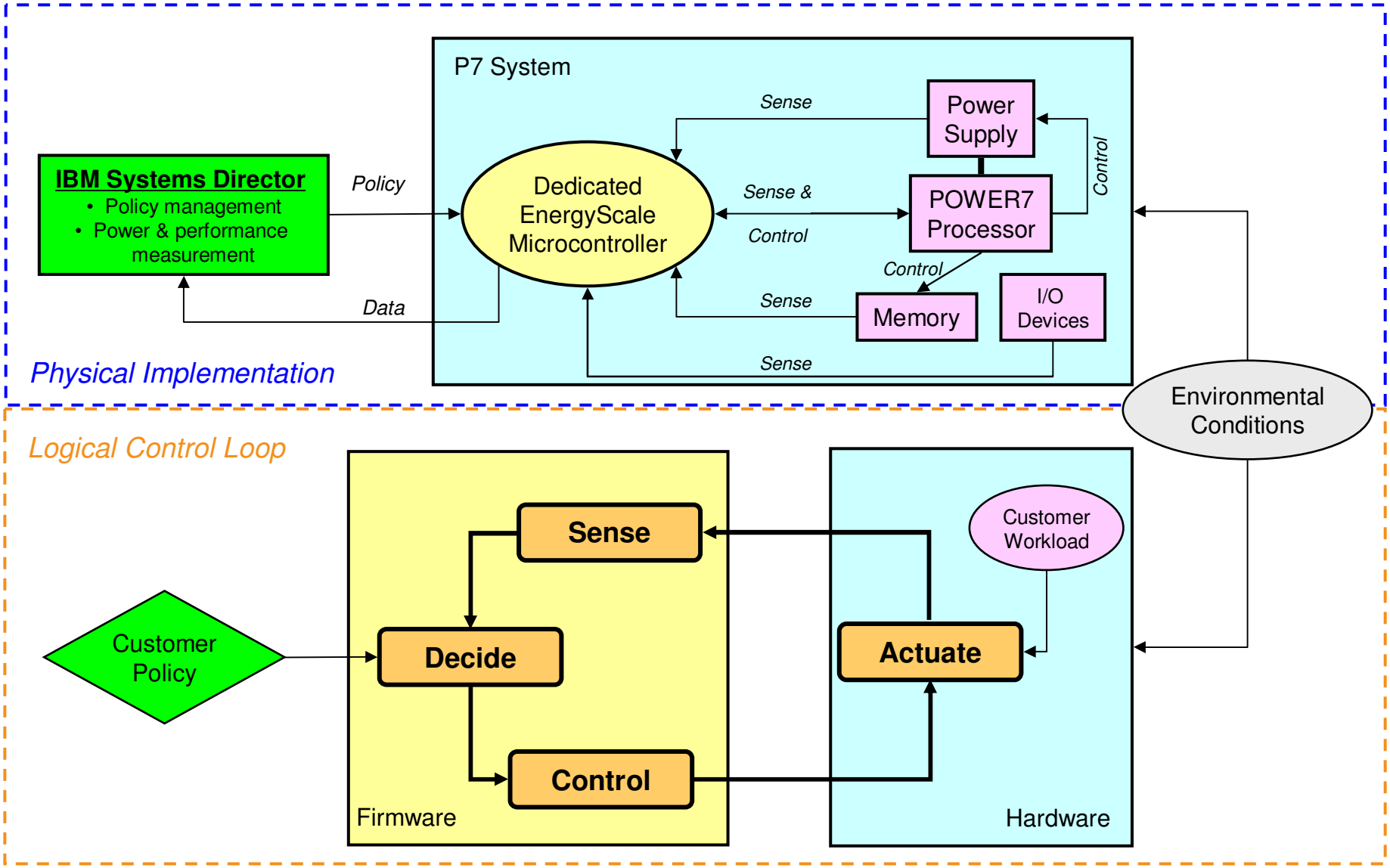
EnergyScale Primary Policies In Action



- Shipping EnergyScale policies with representative usage (real customer code)
- **IBM Power 750 Express Server** (not fully populated)
- Highly utilized scientific application with varying workload profile

* energy savings vary by workload, system configuration, environment, etc.

EnergyScale = Cooperative Hardware & Firmware Solution



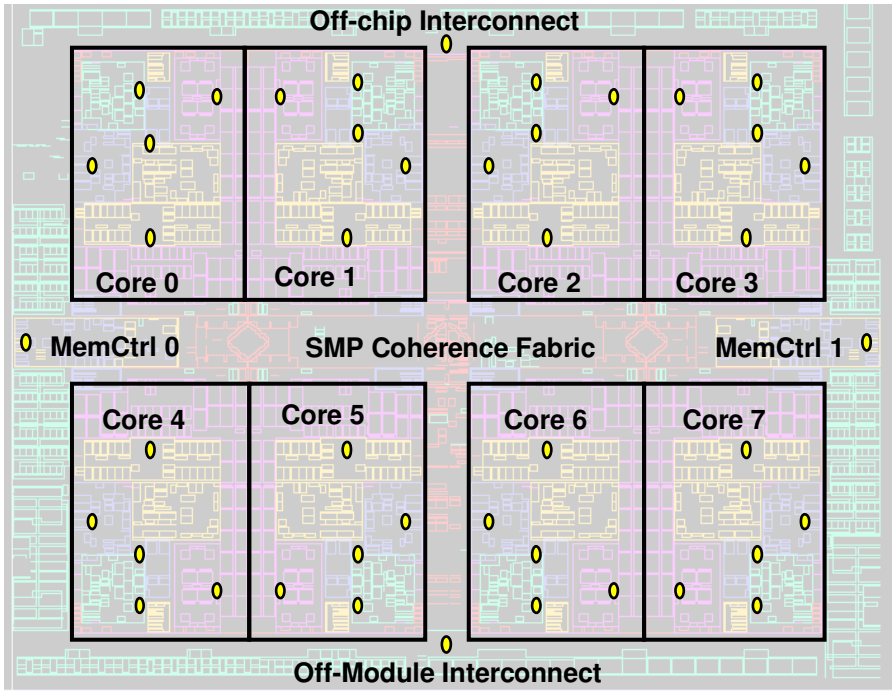
POWER7 Features — Sense

- **Dedicated microarchitectural 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

- **Digital Thermal Sensor (DTS)**
  - 44 on-chip sense points
  - 5 per core chiplet
  - Emergency self-protect thermal throttling

- **Critical Path Monitor (CPM)**
  - Detects circuit timing margin
  - Assists in choosing optimal frequency & voltage

Physical Locations of Thermal Sensors



* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

POWER7 Features — Decide

- **Dedicated off-chip EnergyScale microcontroller**
  - Runs real-time firmware whose sole purpose is to manage system energy
  - Power7 Chip provides dedicated I2C Slave communication port

- **POWER7 accelerators for off-chip microcontroller decisions**
  - **Reducing communication bandwidth need = Faster control loop response time**
    - Sensor packing — to reduce number of read operations
    - Multicast function — to reduce number of write operations
    - Automated on-chip transaction table — to stream out sensor data via single I2C command
  - **Offload & automate compute-intensive chores from EnergyScale microcontroller**
    - Thermal Sensor Conversion to degrees C using quadratic curve fit
    - Chiplet Power Proxy Calculation
    - Automated Voltage change sequencer *(hardware state machine slew assist)*

* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

POWER7 Features Control

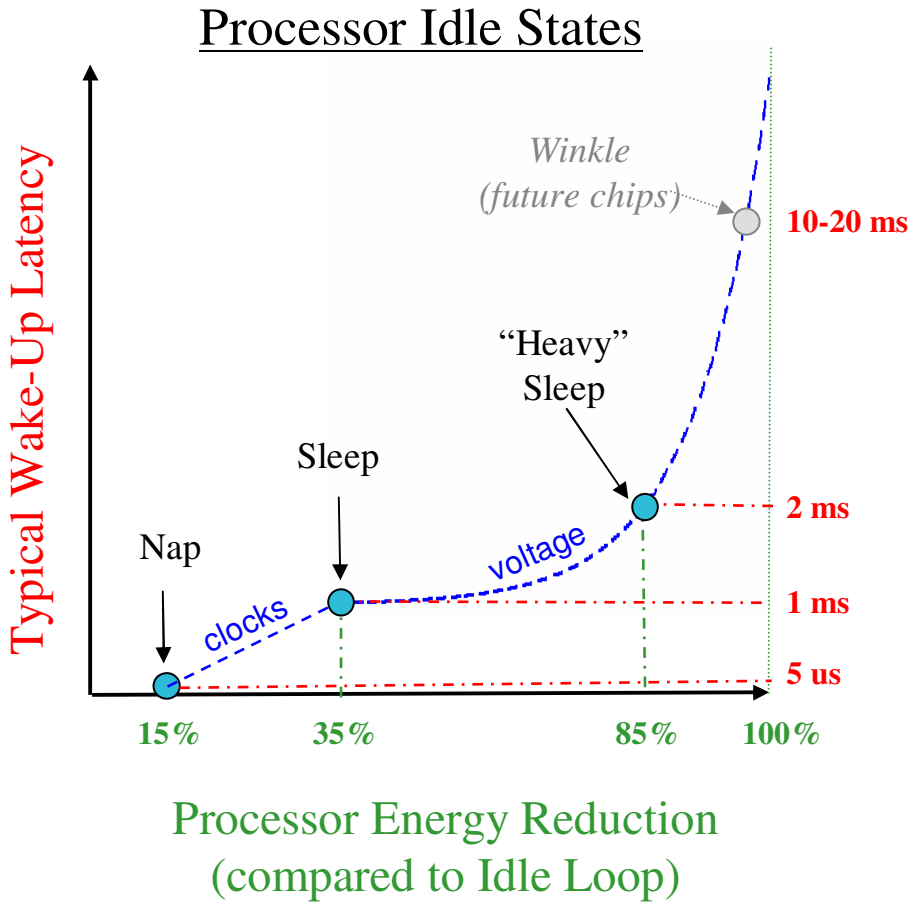
- **Per-core frequency control**
  - Digital PLL (DPLL) clock source supports full EnergyScale dynamic range of: -50% to +10% of nominal frequency with 25Mhz resolution
  - Automated fast frequency slew in excess of 50Mhz per us
- **On-chip support for Off-chip voltage control**
  - Industry-standard Parallel VID interface for low-end to midrange systems' VRM control
  - Serial Voltage command interface to automate multi-step I2C transactions to power supply
    - Necessary to support high-end systems' RAS and power delivery requirements
- **Memory (DIMM) power management**
  - Power-down and reduced access rate modes
  - Channel-pair level memory activity control
- **Changing Coherence Interconnect Command Rate**
  - Incorporating asynchronous core support into the SMP interconnect design was not trivial
  - Coherence command rate limits amount of frequency reduction possible
  - Ability to change coherence command rate on the fly to tune processor versus SMP bandwidth
  - Firmware can "learn" ideal command rate for workload by setting utilization thresholds

* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

Actuate Save Energy When Idle

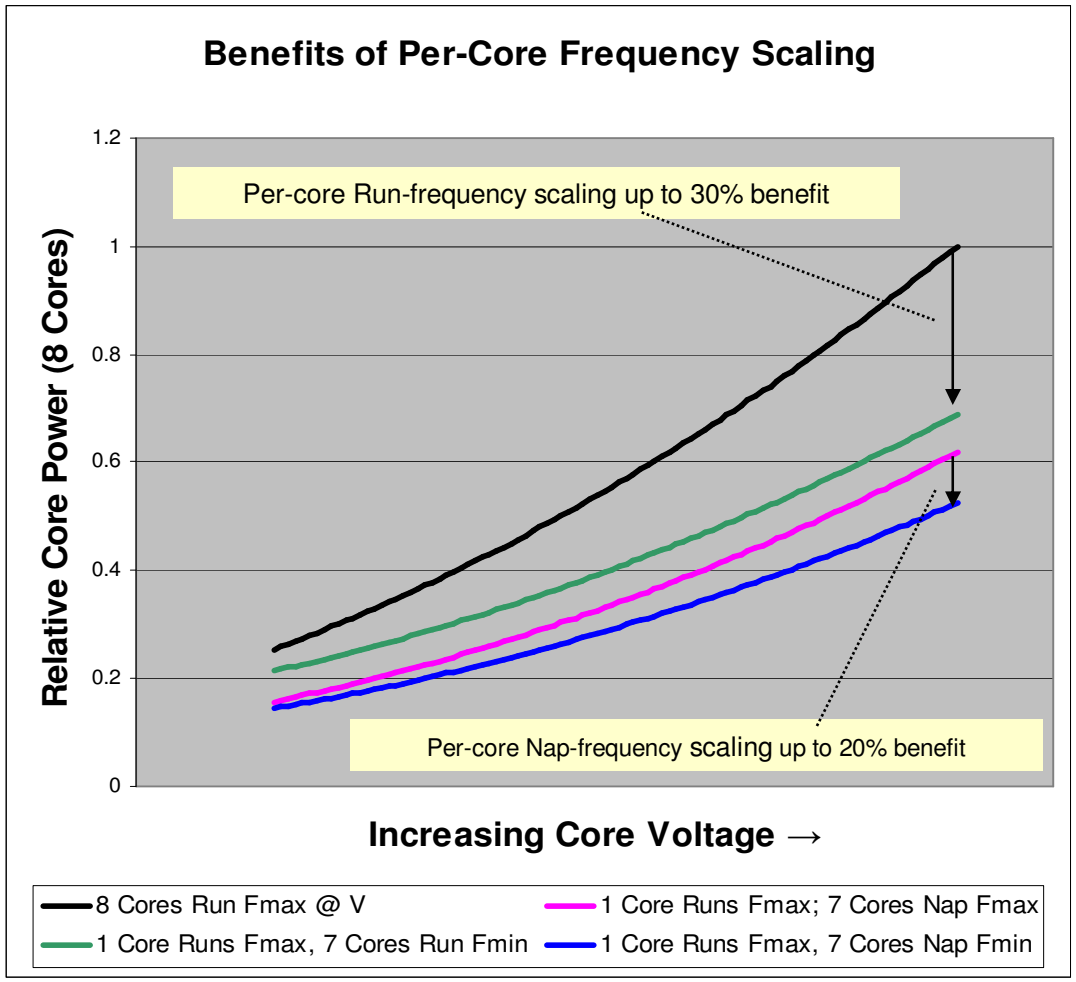
Three idle states were implemented to optimize power vs. latency

- **Nap**
  - Optimized for wake-up time
  - Turn off clocks to execution units
  - Caches remain coherent
- **Sleep**
  - More savings at higher latency
  - Purge and clock off core plus caches
- **"Heavy" Sleep**
  - All cores sleep mode
  - Reduce voltage of all cores to retention
  - Voltage ramps automatically on wake-up
  - No hardware re-initialization required



* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

Actuate Per-Core Frequency Scaling



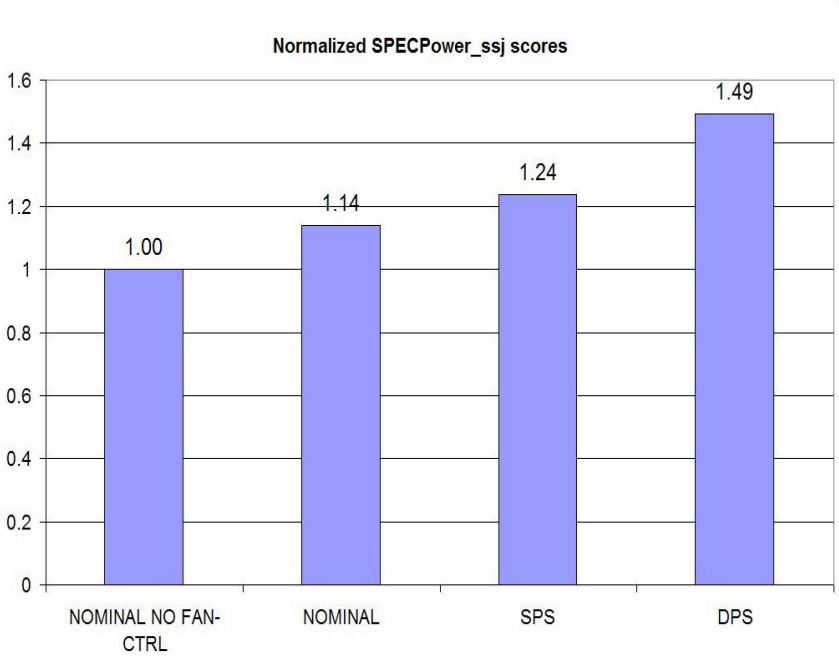
- Allows tuning within a partition for non-homogenous workloads
  (different on each processor core)
- Supports energy optimization in partitioned system configurations
  - Less-utilized partitions can run at lower frequencies
  - Heavily utilized partitions maintain peak performance
- Each partition can run under different energy-savings policy

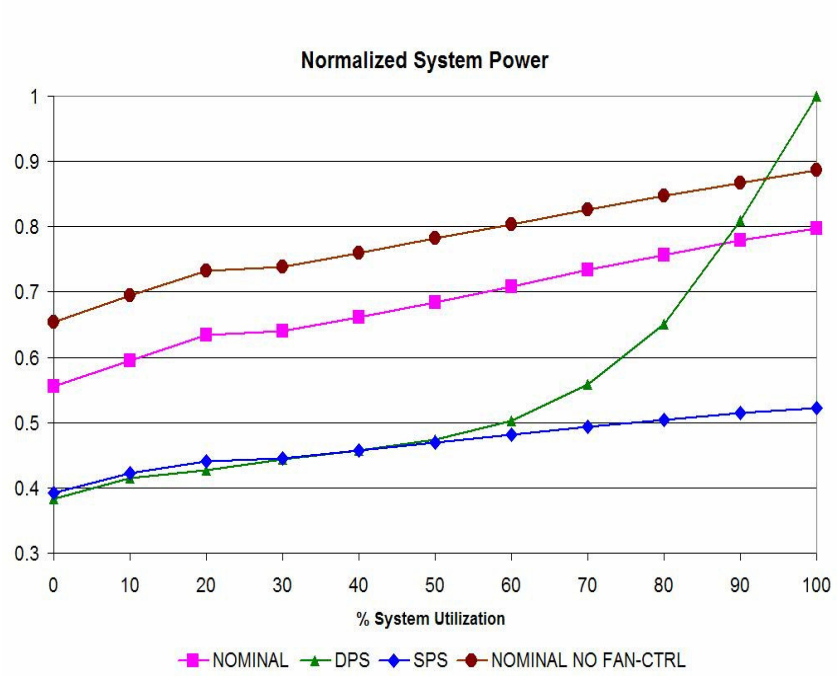
Note: highest frequency core determines the required voltage

* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

Result EnergyScale Impact with POWER7

- SPECPower_ssj2008 runs on a **IBM Power 750 Express** system**





Source: Heather Hanson, IBM research

- **Nominal** adding dynamic fan speed control = 14% higher score
- Two EnergyScale energy-saving modes
  - **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%**

* 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
* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

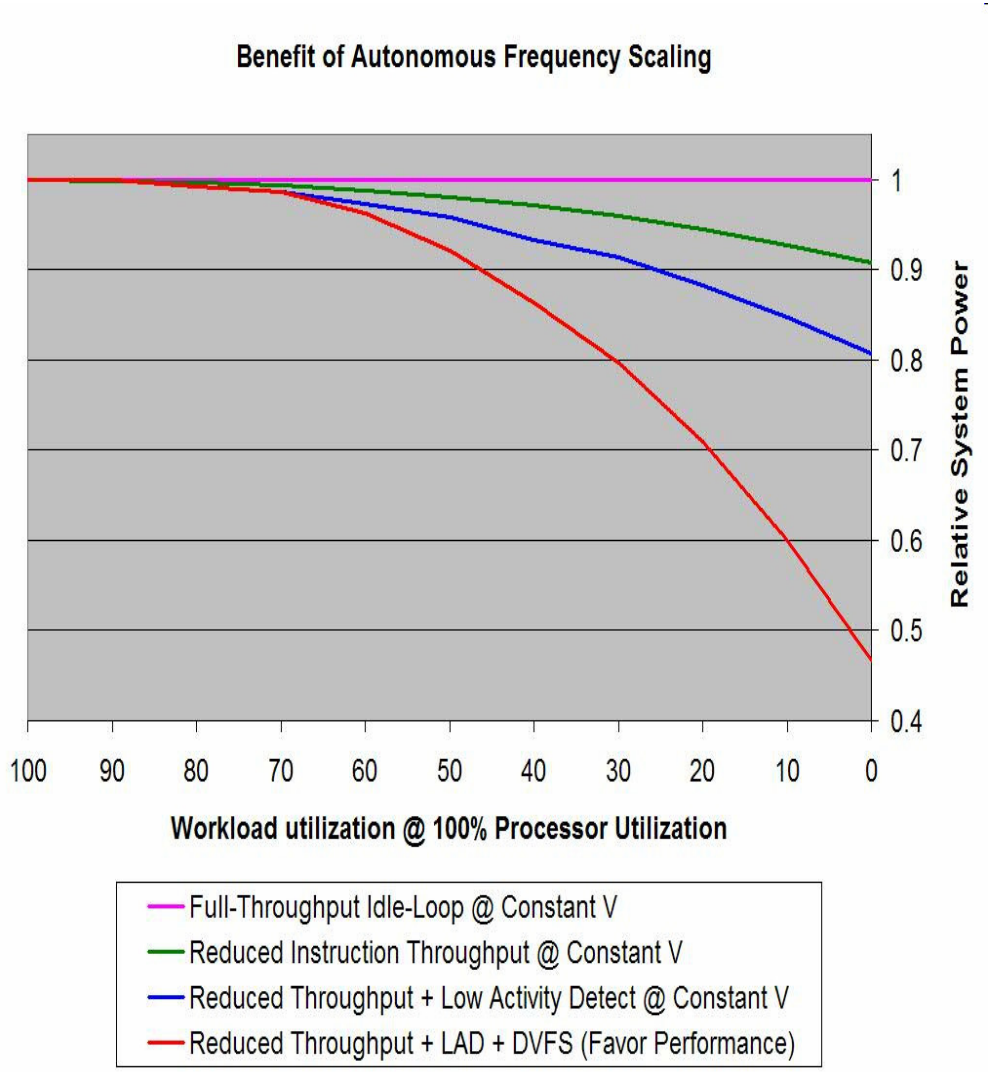
New Autonomic EnergyScale Features

- **Exploratory mechanisms available in POWER7 hardware**
  - Low Activity Detect
  - Power Proxy
  - Autonomic circuit timing margin feedback control

* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

Autonomic Frequency Reduction During Low Activity

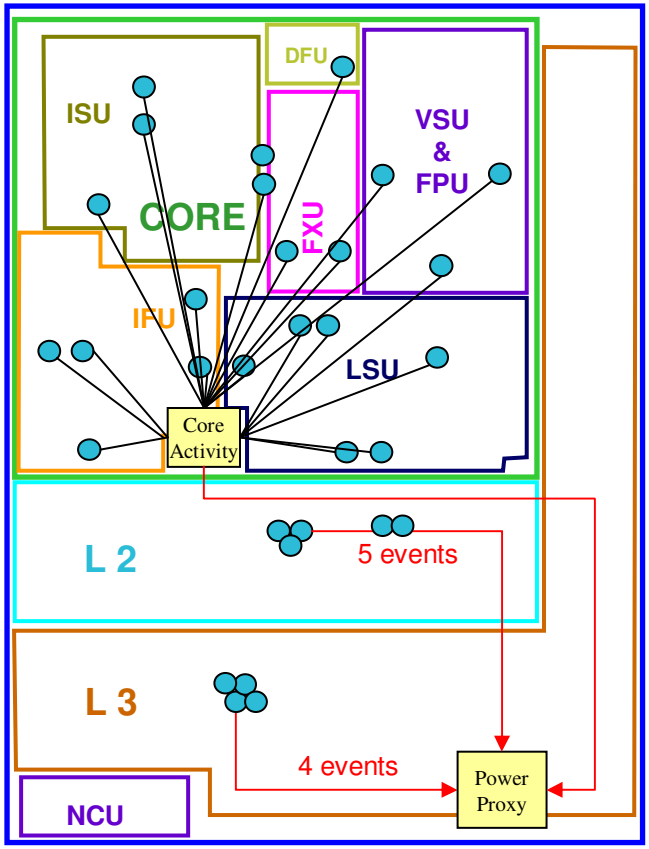
- **Memory bound workloads:** do not need full processor compute frequency while waiting on data.
- Systems **100% utilized by traditional metrics** may actually be **100% idle** in terms of servicing real applications, e.g. while polling for work to arrive.
- **Solution = Low Activity Detect (LAD)**
  1. Hardware reduces processor frequency in response to drop in instruction throughput
  2. EnergyScale algorithms actuate in response to autonomous frequency reduction
- **Minimize service latency impact on work arrival:** Hardware is still running instructions and can rapidly restore full frequency using DPLL slewing.
- **Green Polling**
  Software "artificially" drops instruction throughput to engage autonomic hardware mechanism
- **Useful for:**
  - **Traditional Idle loops:** prefer to avoid overhead of idle state context switch and retain fast reaction time to work arriving in the queue
  - **Message-passing work queuing:** Idle state not possible. Goal is to react to a change in memory location as fast as possible.



* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

Sense Processor Core Power Proxy

Processor Core Chiplet



● = Activity Sense point

Goal:

Estimate per-core chiplet power that we cannot directly measure

Method:

- For each functional unit, pick small subset of activities to infer power consumption (*e.g. cache & regfile reads & writes, execution pipeline issue)*
- Weight each activity to represent how much relative power it consumes
- Combine weighted Core, L2, and L3 activity, then add constant offset plus clock grid power to form:

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

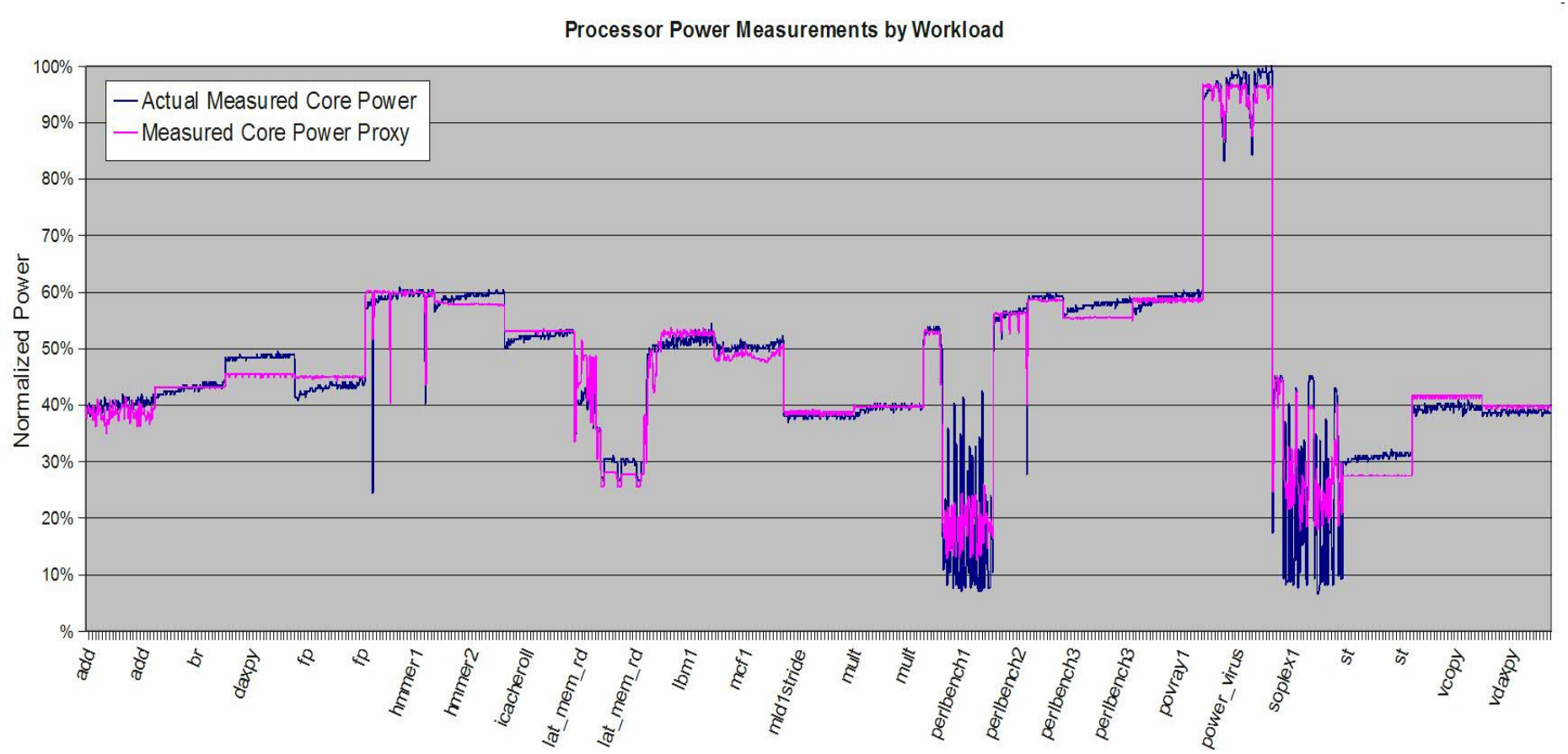
Result:

➢ EnergyScale Firmware adjusts this value for effects of leakage, temperature, and voltage

* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

Result Power Proxy Measurements

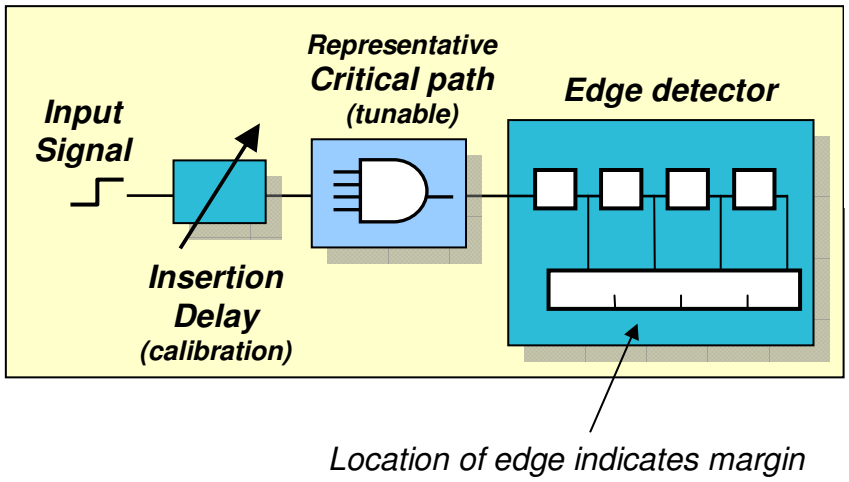
- EnergyScale firmware budgets power across multiple processors and memory, used to:
  - Shift power to cores or other components (e.g. memory) that need it the most
    (Especially important to achieve higher overall performance under a power cap)
  - Enable Server Partition power accounting

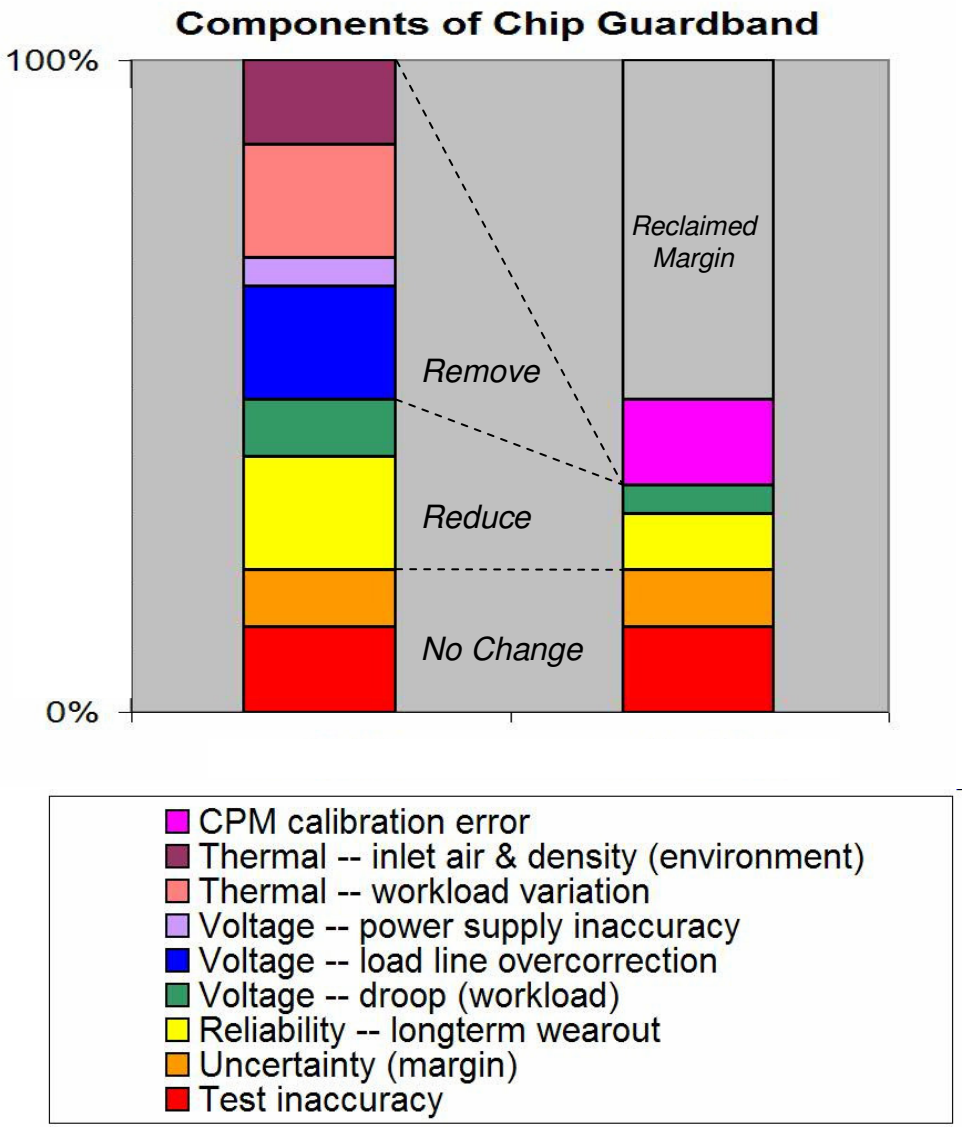


* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

Reducing "wasteful" guardband

- **Conventional guardband**
  - Static, conservative voltage margins for potential worst-case conditions
  - Causes unnecessary loss of energy efficiency during typical server usage
- **Critical Path Monitor (CPM)**
  - Dynamic detection of available circuit timing margin



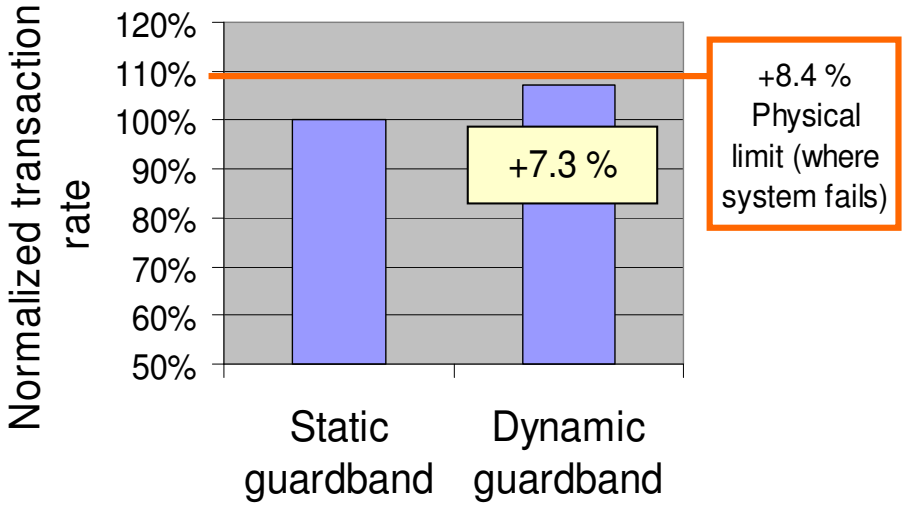


* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

Results Using CPMs for dynamic guardbanding

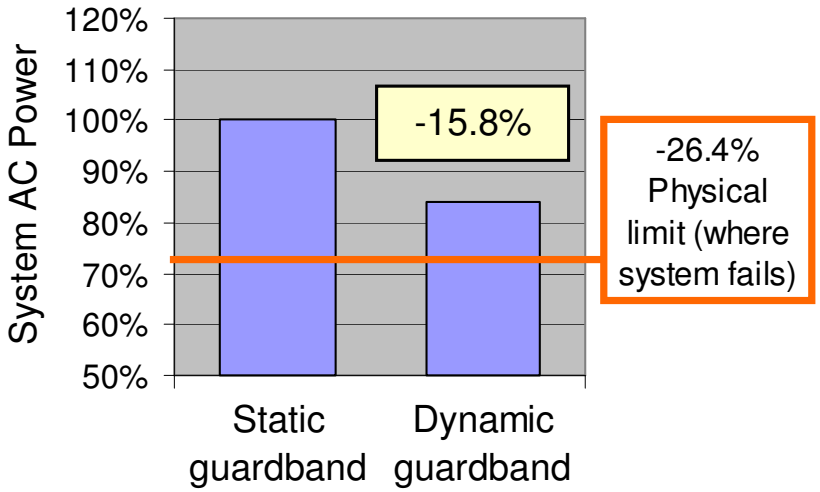
- **Static guardband:** Traditional guardband selection
- **Dynamic guardband**: use CPM feedback to optimize frequency or voltage
- Workload: SPECPower_ssj 100% load level (EnergyScale DPS-FP policy)
- Running on **IBM Power 750 Express Server** (32 cores, 64GB @ 22C Ambient)

Over-clocking: Improve performance



- All cases use same "turbo" voltage level
- Only frequency is adjusted

Under-volting: Save energy



- All cases same frequency & performance
- Only voltage is adjusted

* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability

Summary

- POWER7 builds upon initial POWER6™ EnergyScale features by including automated on-chip functions and accelerators to assist the off-chip microcontroller firmware
- New energy management features in POWER7 have shown 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
- Autonomous low activity, power proxy, and circuit margin feedback functions provide even more opportunity for energy efficient operation

* 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
* Statements regarding EnergyScale features do not imply that IBM will introduce a system with this capability