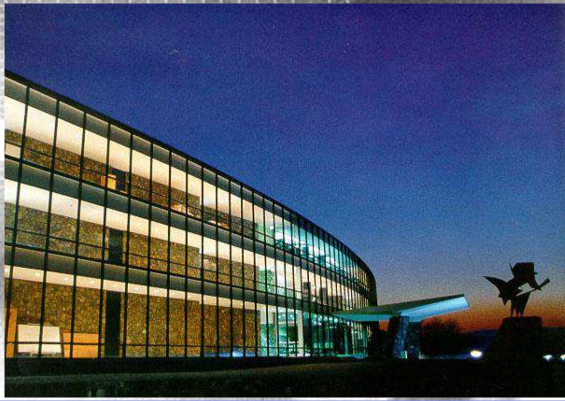
Power-Performance Comparative Evaluation of Alternate Microarchitectures

Rick Eickemeyer, Michael Floyd, John Griswell,
Alex Mericas, Balaram Sinharoy
(IBM Systems and Technology Group)

Pradip Bose, Soraya Ghiasi, Hendrik Hamann,
Hans Jacobson, Tom Keller, Victor Zyuban
(IBM Research Division)



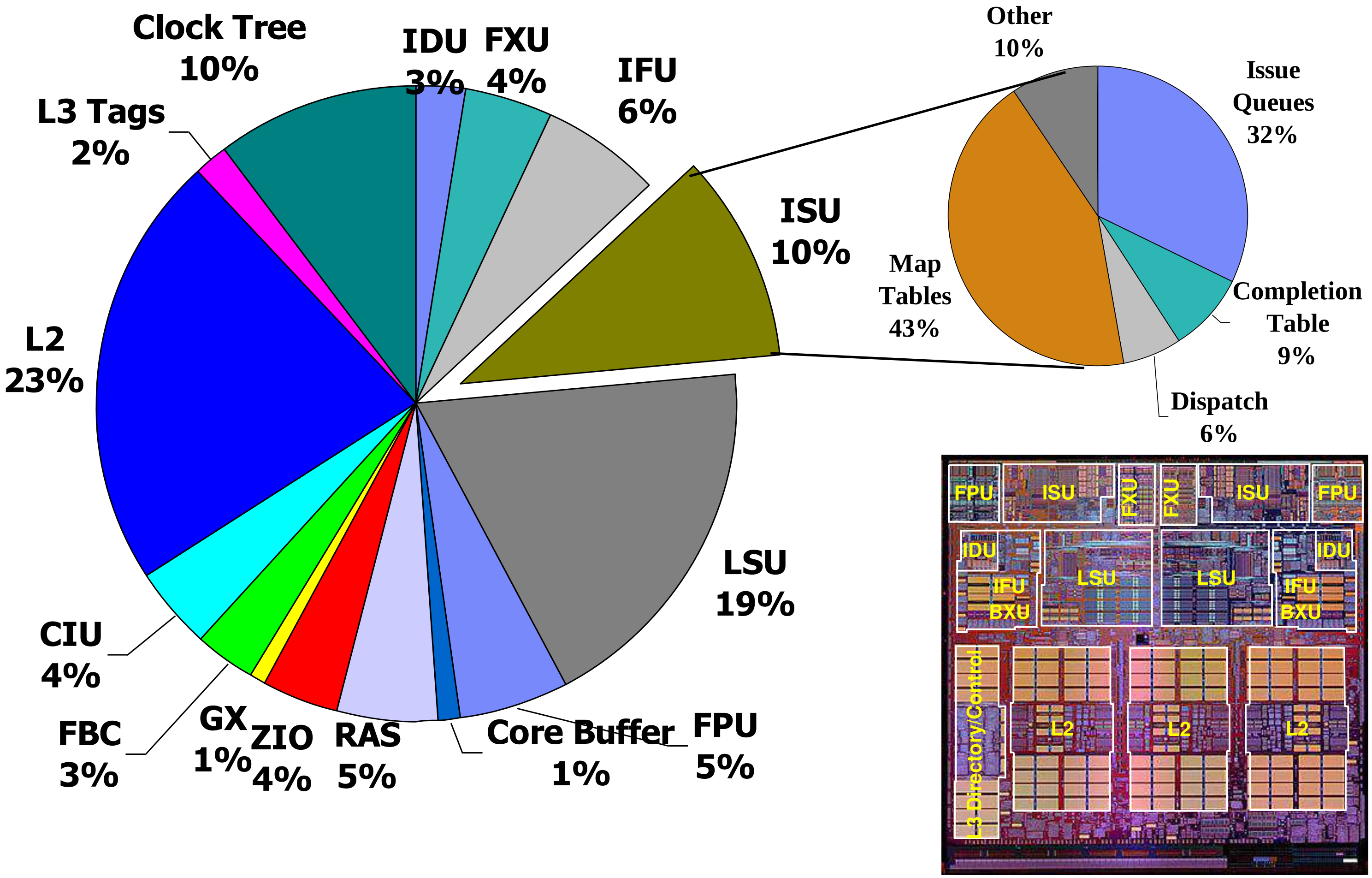
Acknowledgements: Brad McCredie & many others in processor/system design teams within IBM STG

+ members of the power-aware microarchitectures and systems departments within IBM Research

Outline

- Power Dissipation and Efficiency Basics
- POWER4 vs. POWER5
- POWER5 vs. POWER6
- Roadrunner and Blue Gene System Efficiency
- Conclusion
- BACKUP: Looking Ahead: A Few Key Research Issues

Server-Class Processor: Unconstrained Power



Pre-silicon, POWER4-like superscalar design

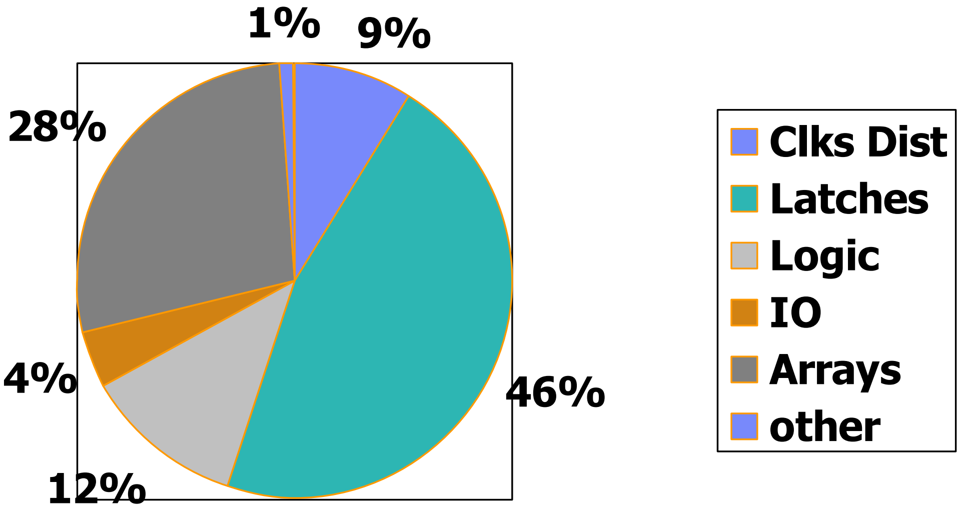
D. Brooks, et. al. MICRO-03 (tutorial)

Processor Power Pie-Chart: Another View

- **High performance processors (prior/current generation) typically burn most of their power in the clocked latches and arrays (registers, caches).**

(taken from: Bose, Martonosi, Brooks: Sigmetrics-2001 Tutorial)

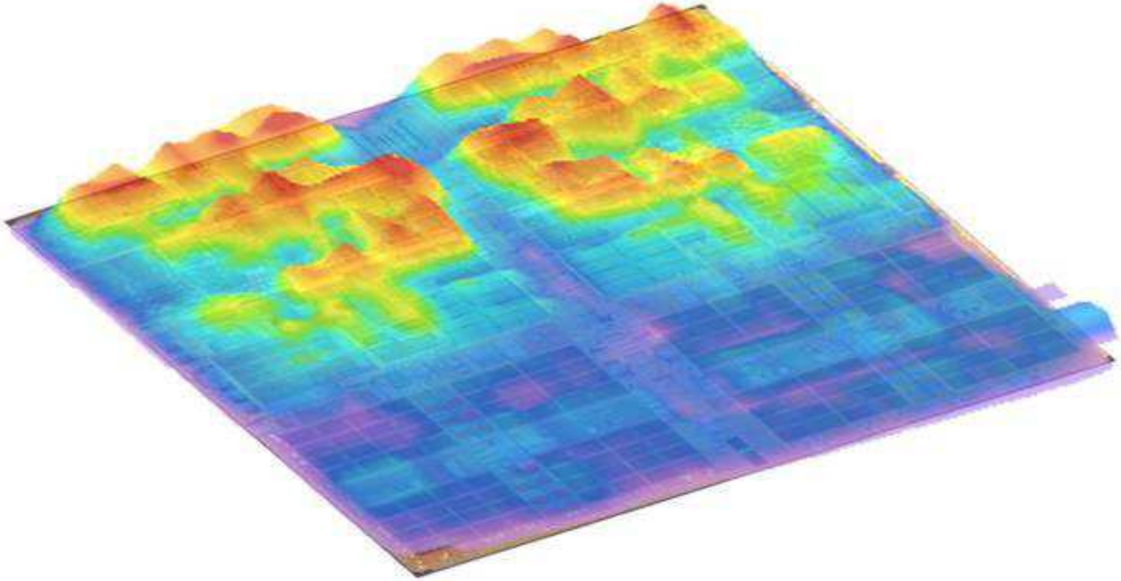
Example data



Pre-silicon ckt-sim based; assumes: no clock-gating

Non-uniform Power Distribution

In addition: Non-uniform power distribution or hotspots aggravate challenges significantly:



Hotspots limit performance, reliability & increase costs

As we go forward (towards decreasing technology nodes):

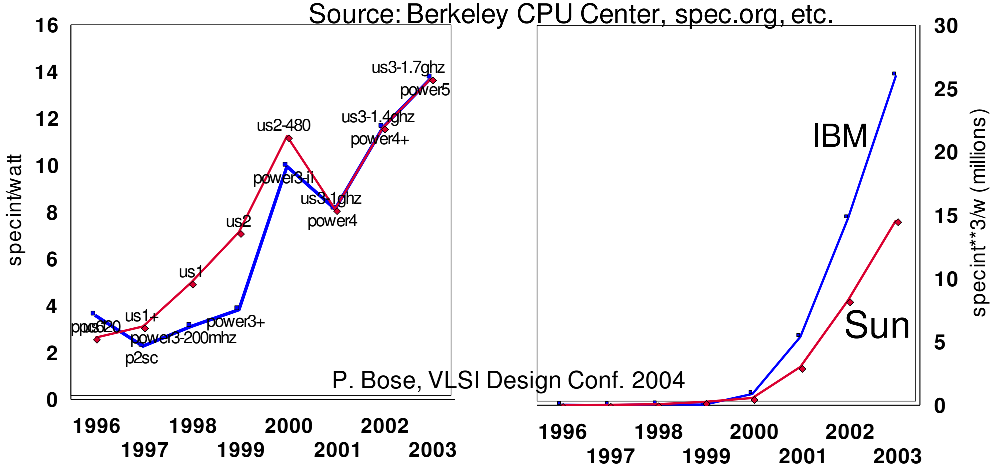
- Hotspot factor (overhead) is likely continue to increase
- predictability of hotspots will be more difficult

(multicore, SoC, power / thermal management, variability etc.)

H. Hamann, from ISCA08 Tutorial

Metrics Overview: An architect’s View

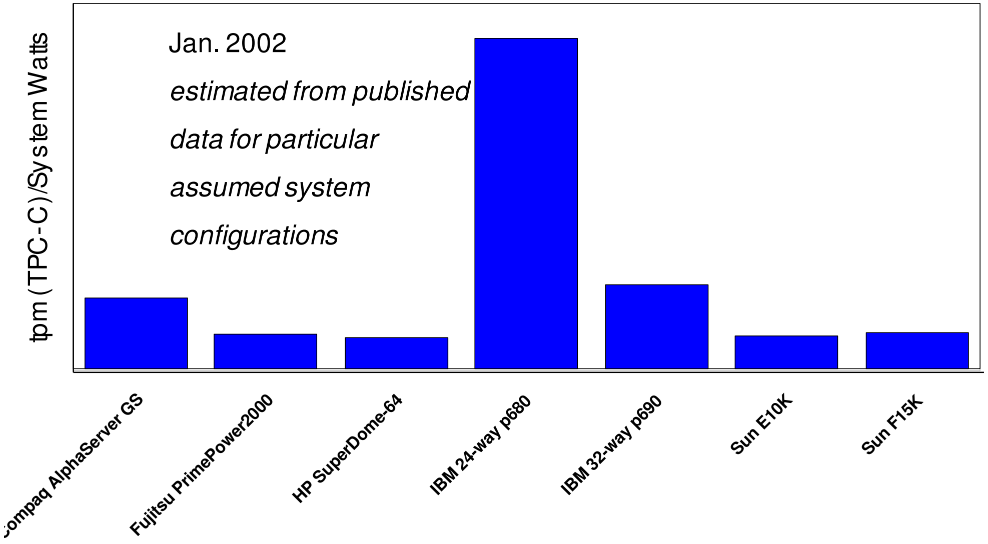
- **Performance metrics:**
  - delay (execution time) per instruction; MIPS
    - CPI (cycles per instr): abstracts out the MHz
    - SPEC (int or fp); TPM: factors in benchmark, MHz
- **energy and power metrics:**
  - joules (J) and watts (W)
- **joint metric possibilities (perf and power or temperature)**
  - watts (W): for ultra LP processors; also, thermal issues
  - MIPS/W or SPEC/W ~ energy per instruction
    - CPI * W: equivalent inverse metric
  - MIPS^2/W or SPEC^2/W ~ energy*delay (EDP)
  - MIPS^3/W or SPEC^3/W ~ energy*(delay)^2 (ED^{2}P)
  - (Peak Temp) * (Execution Time)



Source: Berkeley CPU Center, spec.org, etc.

P. Bose, VLSI Design Conf. 2004

Single-core regime, through start of multi-cores



Jan. 2002

estimated from published data for particular assumed system configurations

System-level perf/watt for commercial OLTP is quite different from processor-level SPECint/watt !

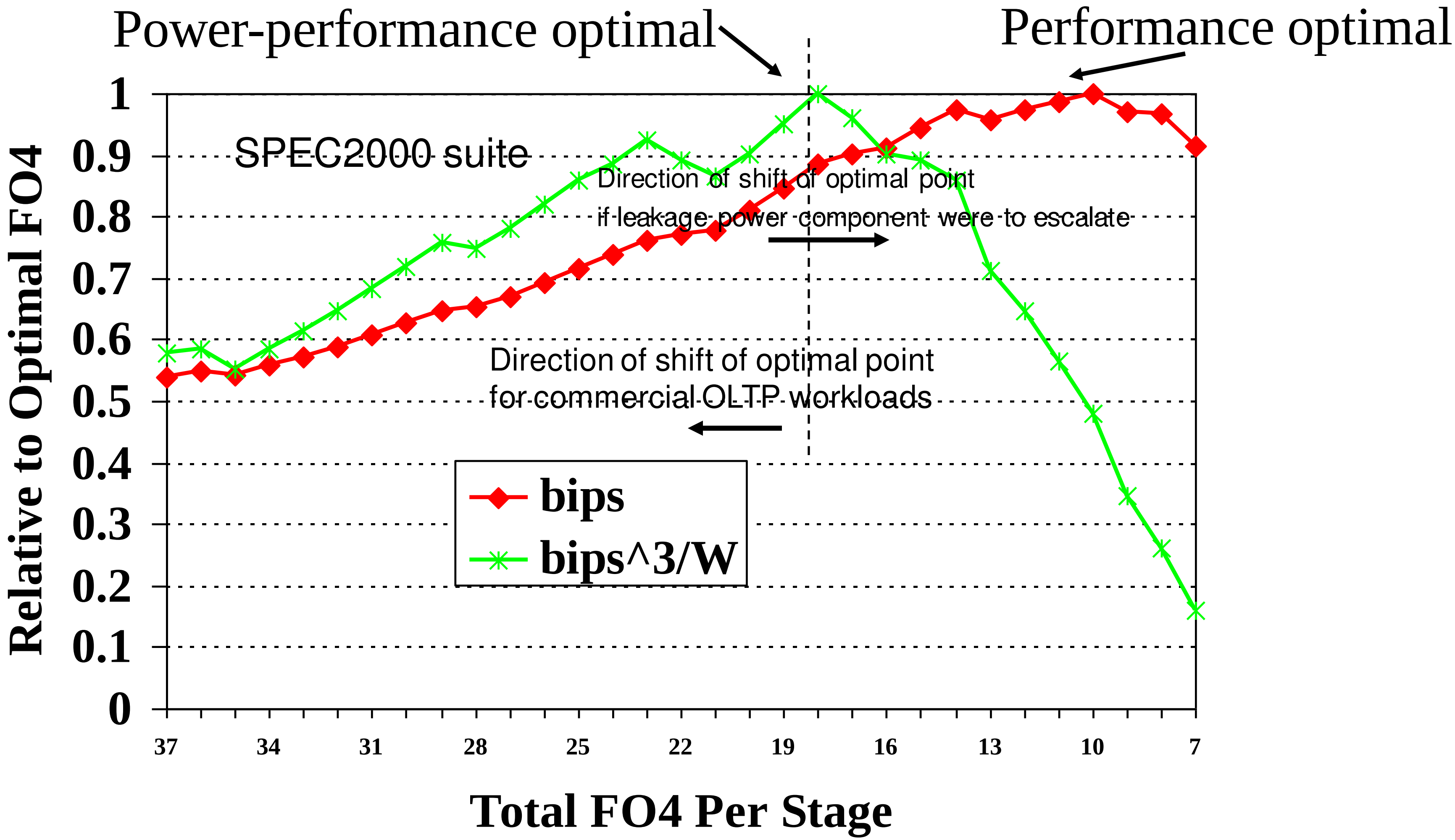
Fundamental Efficiency Determinants

- **Fundamental microarchitectural knobs that determine efficiency**
  - optimal pipeline depth at the core-level
  - optimal core complexity and number of cores
  - type of clock-gating and power-gating (if applicable): coarse-grain vs. fine-grain
  - adaptive microarchitectures: [to control unnecessary energy waste]
  - etc…

- **Fundamental logic/circuit-level efficiency features**
  - support for clock-gating (area and verification-efficient)
  - support for voltage and frequency scaling (performance, reliability and verification-friendly)
  - (Near)-optimal mix of low, medium and high-Vt devices
  - area and power efficient latch design
  - etc..

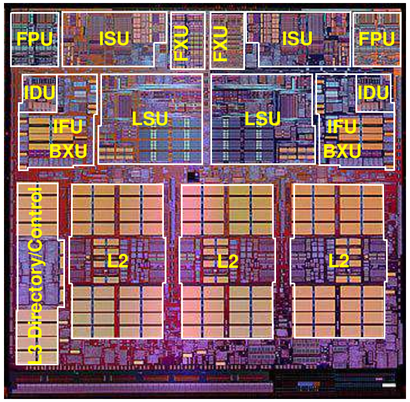
Deducing Optimal Pipe Depths

V. Srinivasan et al., MICRO-35, 2002



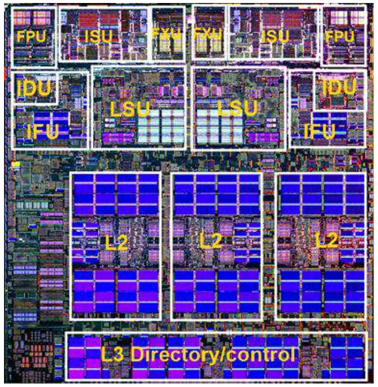
Integrating Multiple Cores on Chip

- **With uniprocessor performance improvements slowing, multiple cores per chip (socket) will help continue the exponential system performance growth**
- **Exploit performance through higher levels of integration in chips, modules, and systems**
- **Invest power in chip-level performance rather than core performance**



POWER 4: 2001
180 nm, Cu, SOI
2 cores / chip

POWER 4+: 130 nm



POWER 5: 2004
130 nm, Cu, SOI
2 cores / chip
2 way SMT / core

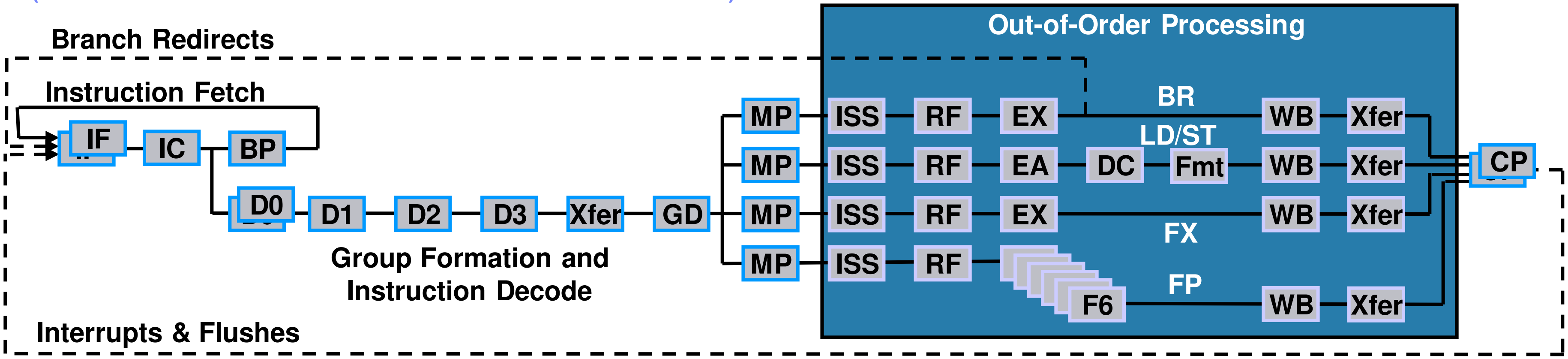
POWER5+: 90nm

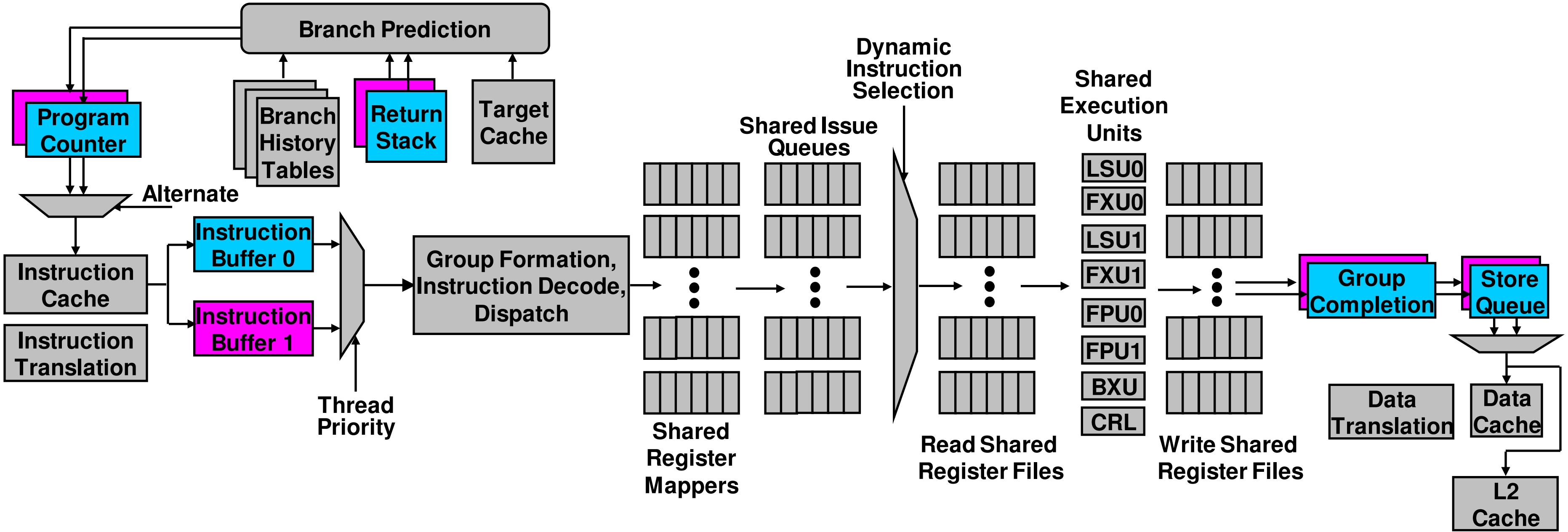
Adaptive Microarchitecture Principles

- **Basic concepts:**
  - use (i.e. power or clock) a storage, compute or interconnect (e.g. bus) resource only to the extent needed: adapt or reconfigure dynamically in tune with workload resources
    - Predictive power-gating to reduce leakage
    - Dynamic resizing of queues, buffers, caches
  - dynamically change a bandwidth parameter to conserve power
    - Adaptive fetch to minimize speculative waste
    - Adaptive prefetch to conserve bus bandwidth and prefetch logic usage; reduce speculative waste (cache pollution)
    - etc.…
- **Issues that prevent widespread adoption in high-end processors:**
  - complexity (verification cost, overhead area/power)
  - relatively small power savings, if performance loss is not tolerable

In general, dynamic voltage-frequency scaling (DVFS) offers the most efficient knob for power management

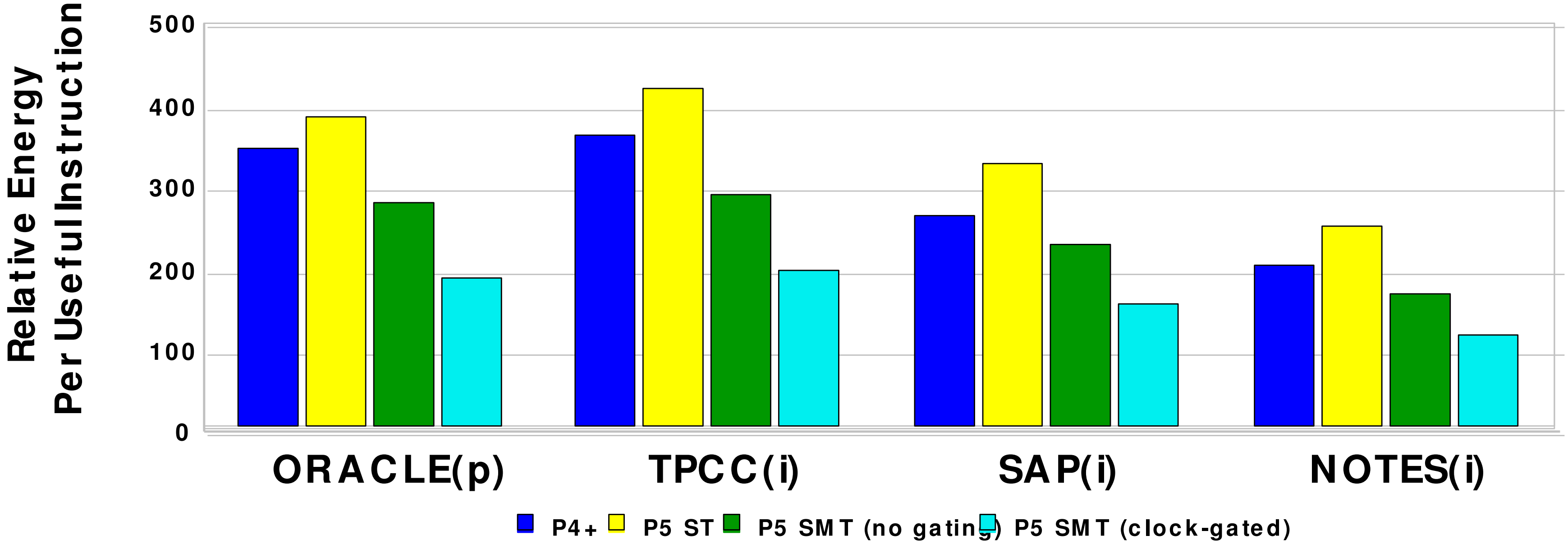
Multithreaded Instruction Flow in Processor Pipeline
(transition from POWER4 to POWER5)



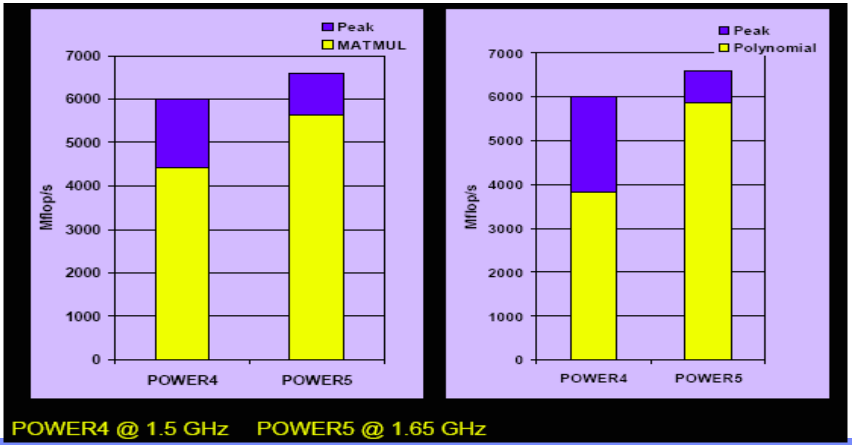


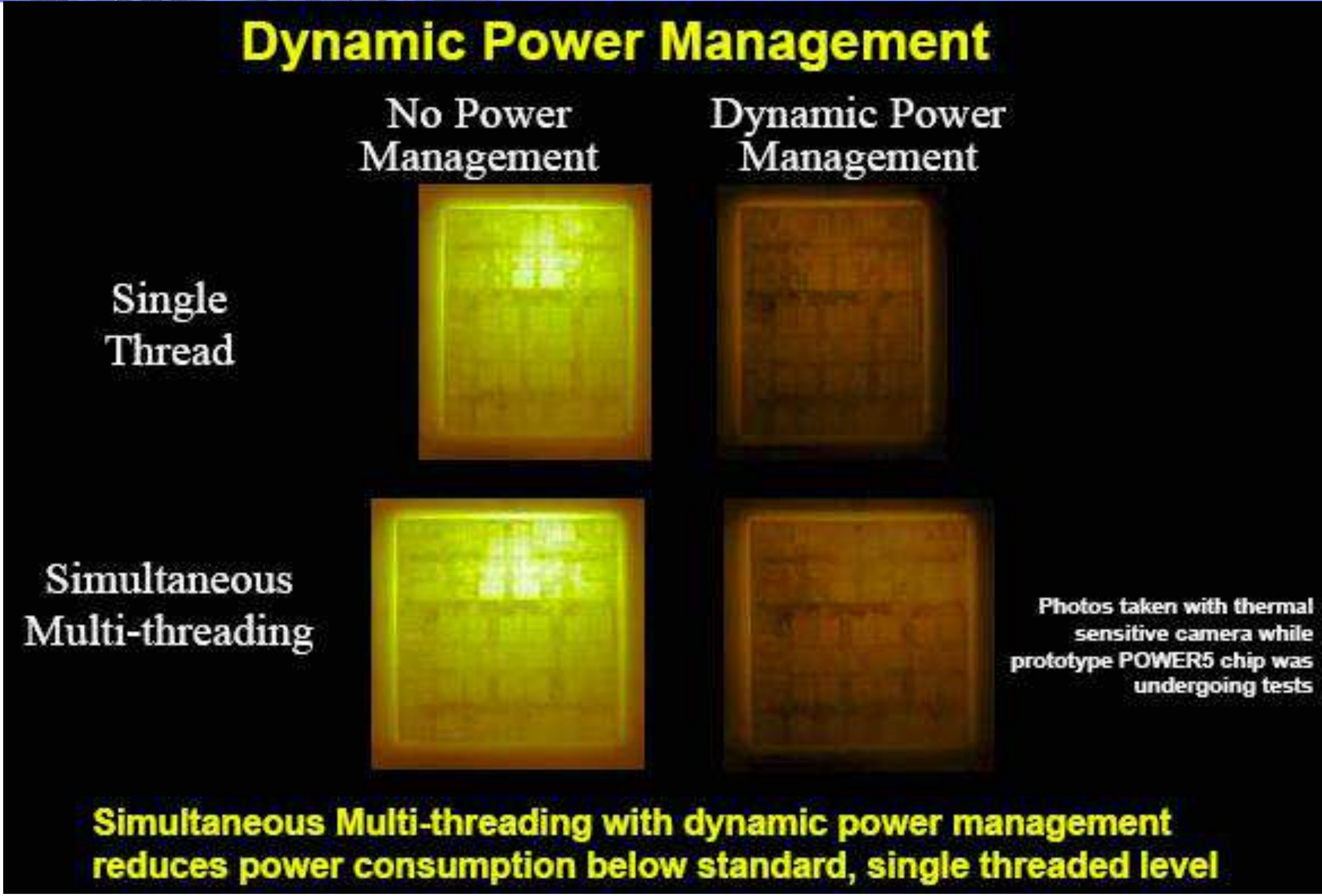
▭ Shared by two threads 🟦 Resource used by thread 0 🟪 Resource used by thread 1

Energy Per Useful Instruction: POWER4+ vs. POWER5



- Steady growth in single-thread performance:
  - POWER4 → POWER4+ → POWER5
- Efficient throughput increase in POWER5:
  - typical OLTP:
    - 40 % IPC growth at 20 % more power





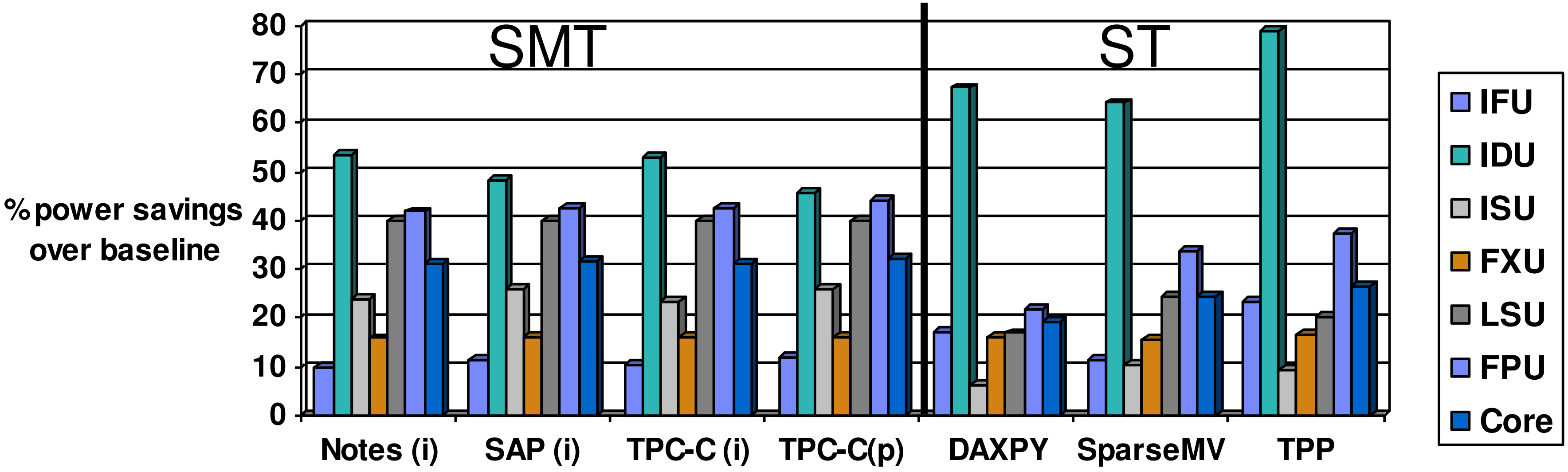
Dynamic Power Management

	No Power Management	Dynamic Power Management
Single Thread		
Simultaneous Multi-threading		

Photos taken with thermal sensitive camera while prototype POWER5 chip was undergoing tests

Simultaneous Multi-threading with dynamic power management reduces power consumption below standard, single threaded level

Active Power Savings from Clock-Gating (% over baseline) (POWER5: pre-silicon projections)



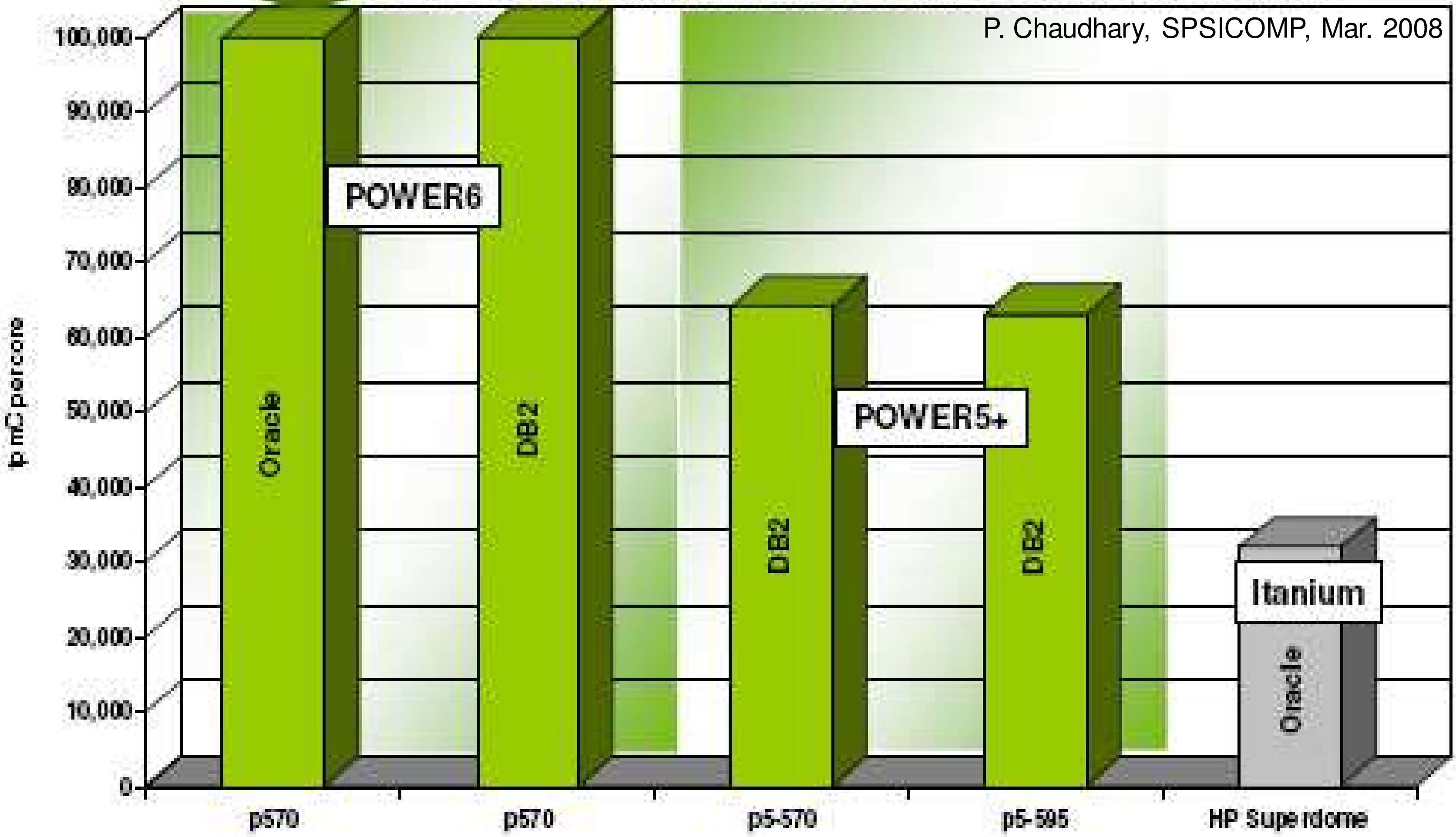
Note: post-silicon hardware-based analysis shows good agreement at the chip level

POWER6 p570 scores big on tpmC per core

Transaction Performance - Single System tpmC per core

New! New Oracle Benchmark demonstrates linear scaling and performance

P. Chaudhary, SPSICOMP, Mar. 2008



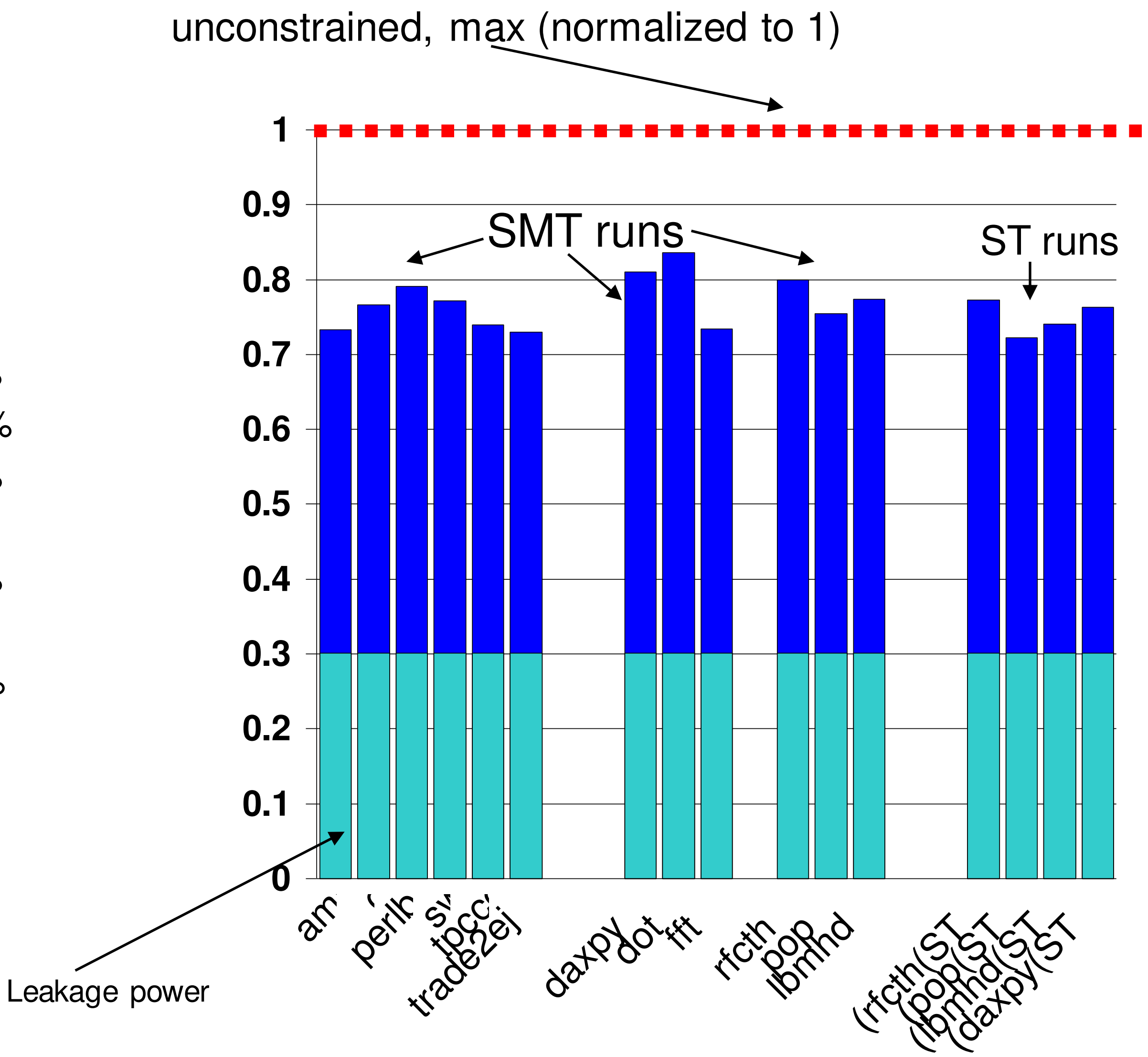
Best results listed for single systems capable of being configured with at least 16 cores for IBM POWER6, IBM POWER5+, and HP Integrity.
Source: http://www.tpc.org as of 10/22/07. Not all results listed.

Benefit of Fine-Grain Clock Gating in POWER6 pre-silicon simulation

- **Power reduction due to clock-gating (average)**

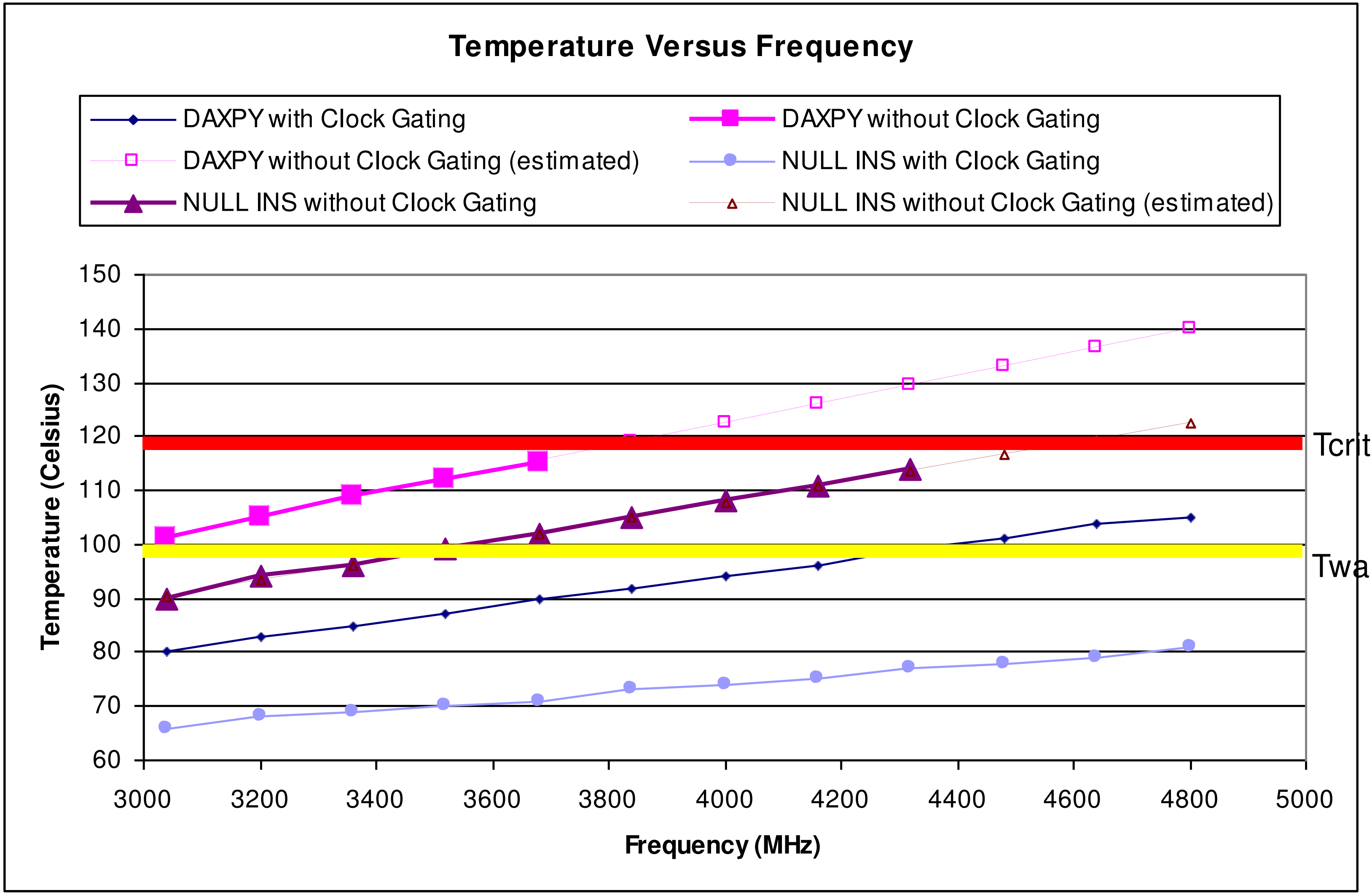
– Floating point kernels	19%
– SPEC2K	24%
– Commercial	26%
– RFCTH	20%
– POP	25%
– LBMHD	23 %

RFCTH is as "hot" as daxpy; POP is similar To SPEC2K average



Daxpy clock-gating factors – validated via direct post-silicon measurements

Clock Gating – Temperature Benefit



Prototype hardware, both cores good, real h/w measurements (POWER6)

Comparative Summary on Clock-Gating Efficiency

- Clock gating benefit
  - POWER4: performance-centric, with minimal clock-gating
  - POWER5: SMT throughput boost, matched with fine-grain clock-gating to manage power
  - POWER6: High frequency performance boost with aggressive, fine-grain clock-gating to manage power and thermals

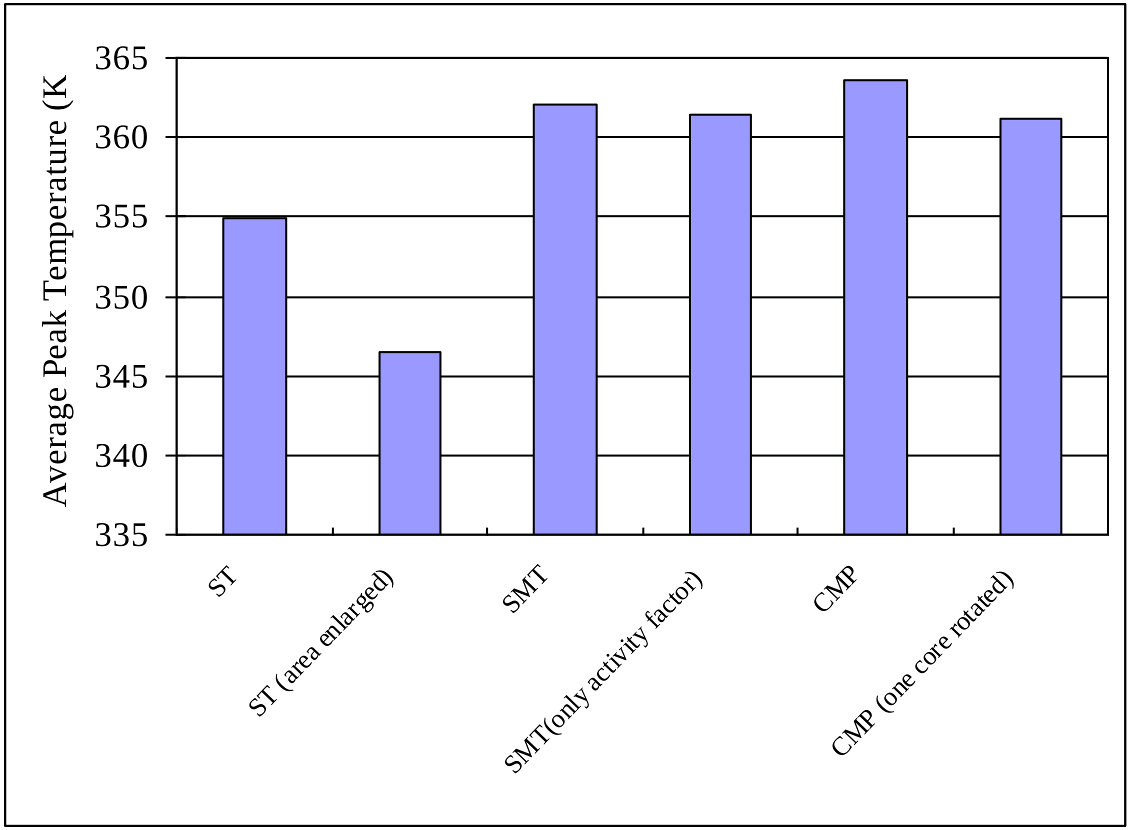
- Net: **progressive improvement with POWER6 being the best**

Peak Temperature: SMT vs. CMP

3 heat-up mechanisms

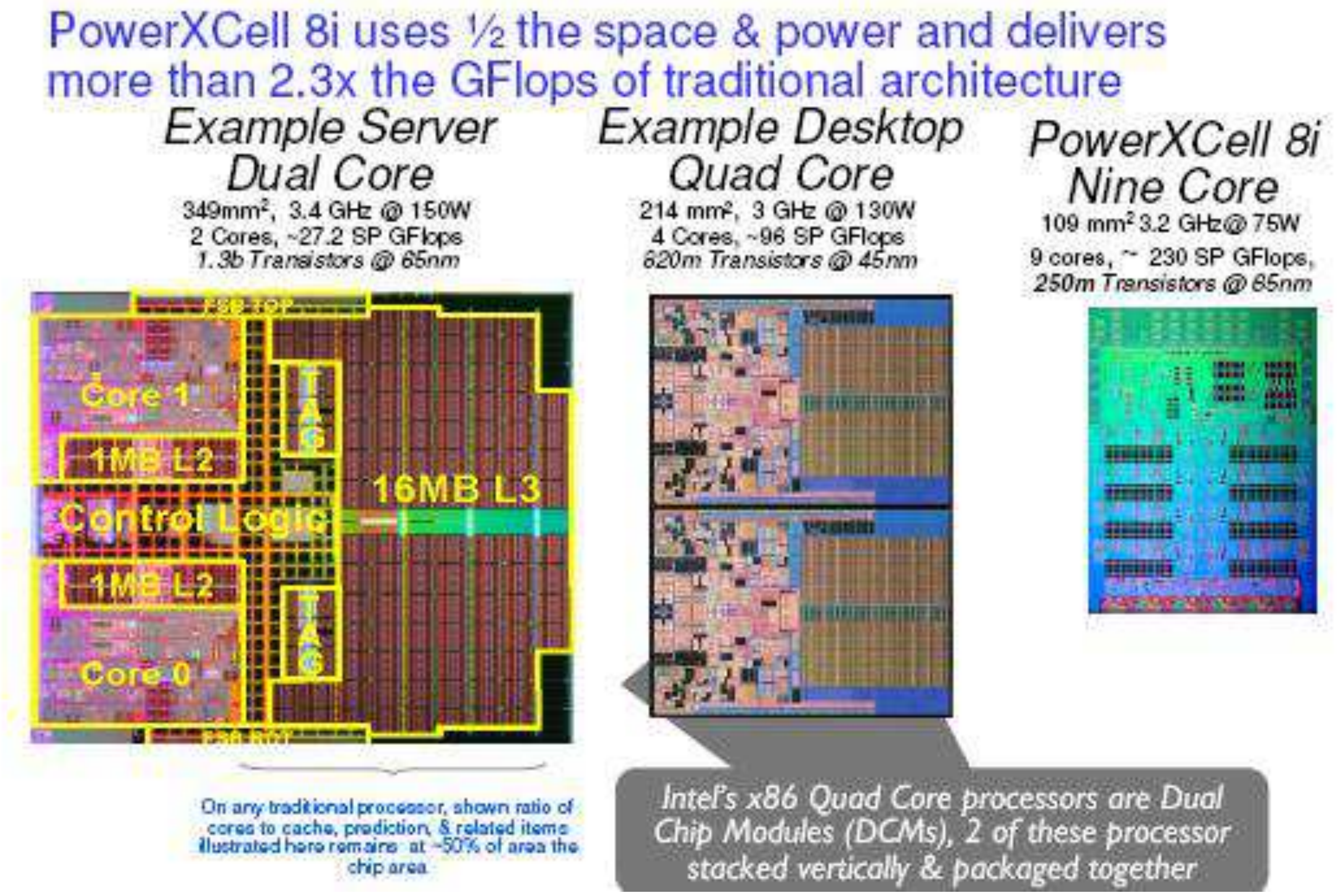
- Unit self heating determined by the power density of the unit
- Lateral thermal coupling between neighboring units
- Global heating through TIM (thermal interface material), heat spreader, and heat sink

SMT: area-efficient, thermally-efficient



P. Bose, VLSI Design 2005, quoted from Y. Li, Z. Hu et al. 2004

A brief look now at a different system product space ….



PowerXCell 8i uses ½ the space & power and delivers more than 2.3x the GFlops of traditional architecture

Example Server Dual Core

349mm², 3.4 GHz @ 150W
2 Cores, ~27.2 SP GFlops
1.3b Transistors @ 65nm

On any traditional processor, shown ratio of cores to cache, prediction, & related items illustrated here remains at ~50% of area the chip area

Example Desktop Quad Core

214 mm², 3 GHz @ 130W
4 Cores, ~96 SP GFlops
620m Transistors @ 45nm

Intel's x86 Quad Core processors are Dual Chip Modules (DCMs), 2 of these processor stacked vertically & packaged together

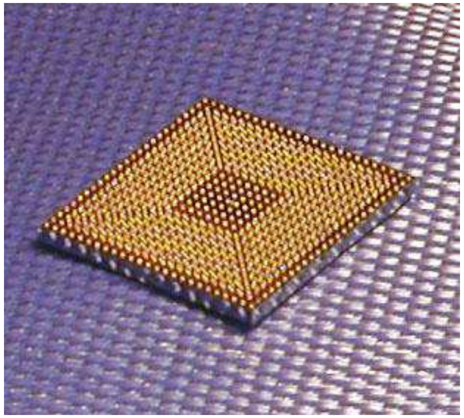
PowerXCell 8i Nine Core

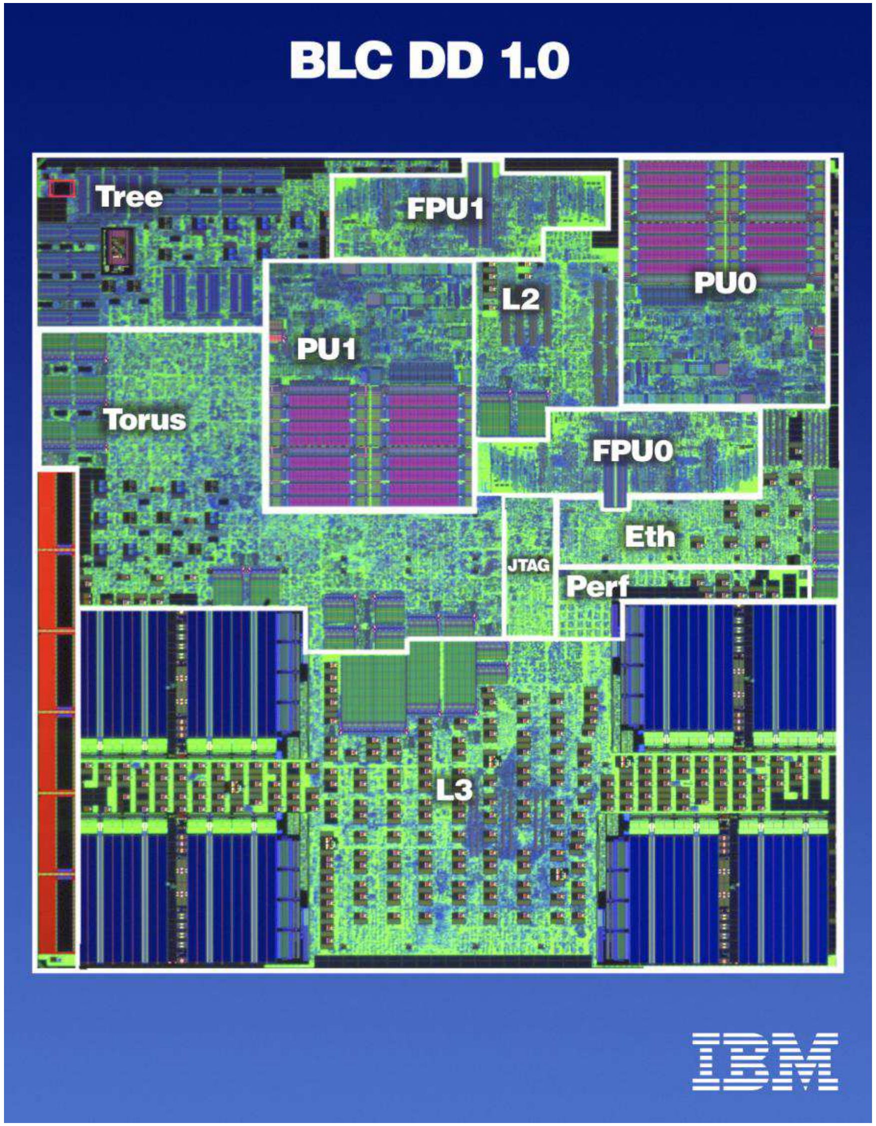
109 mm² 3.2 GHz@ 75W
9 cores, ~ 230 SP GFlops,
250m Transistors @ 65nm

D. Grice, SCICOMP-14, March 2008; http://www.spscicomp.org/

Hardware Integration in BlueGene/L: System-on-a-Chip ASIC

- IBM CU-11, 0.13 µm
- 11 x 11 mm die size
- 25 x 32 mm CBGA
- 474 pins, 328 signal
- 1.5/2.5 Volt





Integrated functionality

- Two PPC 440 cores
- Two “double FPUs”
- L2 and L3 caches
- Torus network
- Tree network
- JTAG
- Performance counters
- EDRAM

The Green500 Top 10 (http://www.green500.org)

Green500 Rank	MFLOPS/Watt	Computer (all IBM)	Total Power (kW)	Top 500 Rank
1	488.14	Roadrunner BladeCenter QS2 – PowerXCell 8i	22.76	324
1	488.14	Roadrunner	18.97	464
3	437.43	Roadrunner	2345.50	1
4	371.75	Blue Gene/P	31.50	304
4	371.75	BG/P	31.50	305
4	371.75	BG/P	94.50	306
7	371.67	BG/P	63.00	52
7	371.67	BG/P	94.50	75
7	371.67	BG/P	126.00	51
7	371.67	BG/P	63.00	37

Concluding Remarks [Based on POWER Systems experiences so far]

- **Power-performance tradeoff analysis must be integral part of early-stage definition of microprocessors**
  - Fundamental design decision errors can lead to post-silicon power overruns and/or performance shortfalls
  - Pre-silicon power-performance modeling and validation methodology: key investment needed to prevent post-silicon surprises
  - Power analysis and tuning must percolate through all stages of design, with closed loop feedback to higher levels.
  - Temperature-aware vs. power-aware: needs careful balance

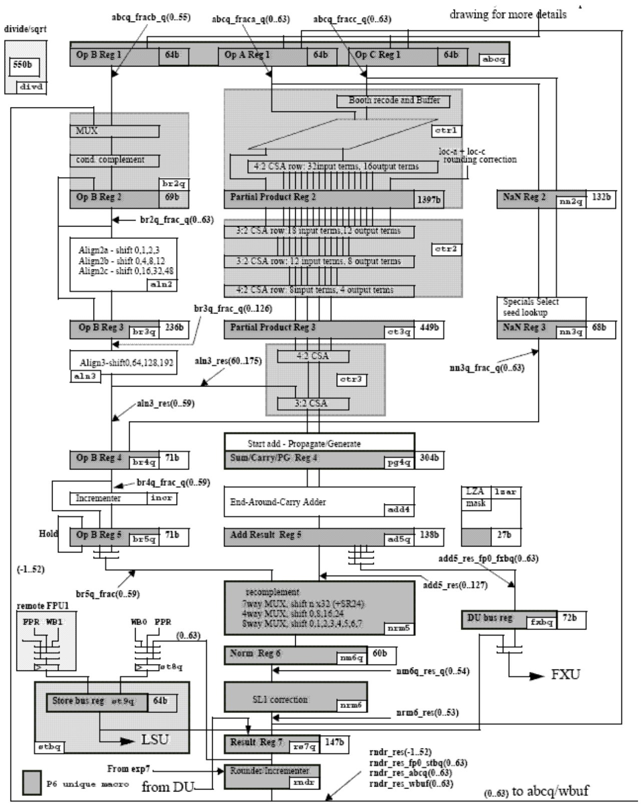
- **Power-aware microarchitecture techniques: can be a key lever in future power reduction at the chip and system level**
  - But, co-design with circuit/technology and software groups is key

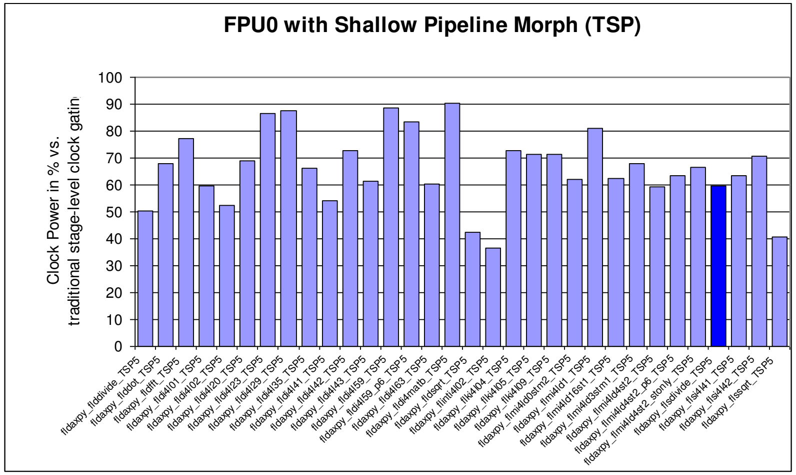
- **Power "optimization" in server-class, high-end systems can be quite different from that in embedded systems**
  - System-level power-performance efficiency requires careful separation of emphasis on efficiency at the processor, memory and system sub-components
  - IBM's POWER Systems microprocessors have been designed with system-level efficiencies in mind and have proven to be very successful offerings in the Green Computing Era.

BACKUP: Some Key Research Issues of the Future

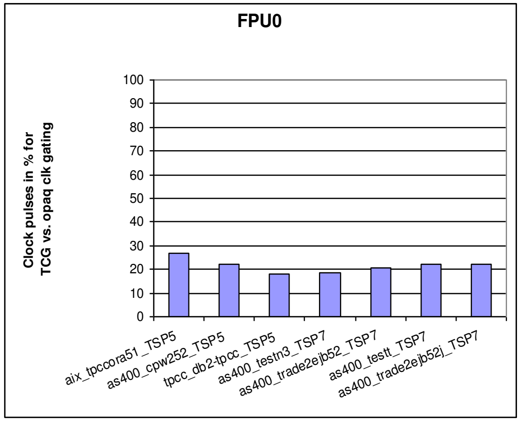
Advancing the State-of-the-Art in Clock Gating

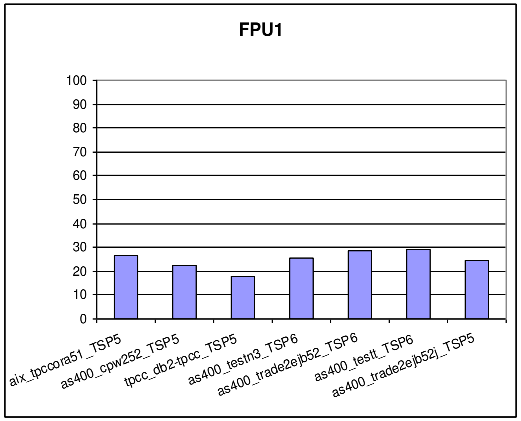
- **M1-level simulation (FPU)**
  - Transparent clock gated pipeline





FLTLOOPS



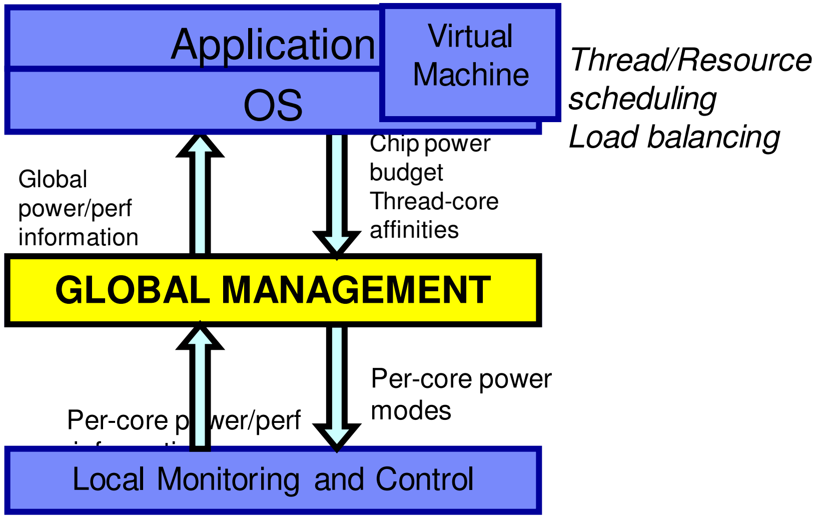


Commercial / TPC-C

Ref: H. Jacobson et al., ISLPED04, HPCA-05

Dynamic mgmt of power, temperature, noise, reliability, performance….

Across-die monitored variability (in perf, power, temp, Vdd, …) will increase in the multi/many-core area. Effective control and management will require integrated, hierarchical, closed-loop feedback control mechanisms



Application

Virtual Machine

OS

Thread/Resource scheduling
Load balancing

Global power/perf information

Chip power budget
Thread-core affinities

GLOBAL MANAGEMENT

Per-core p wer/perf information

Per-core power modes

Local Monitoring and Control

•On-chip controller can also serve as static (v,f) setting device for effective yield and good baseline performance

•Per-core DFVS: costly and requires async interfaces to bus/fabric; also further exacerbates soft error rates

•Control loop stability issues must be analyzed (pre-silicon)

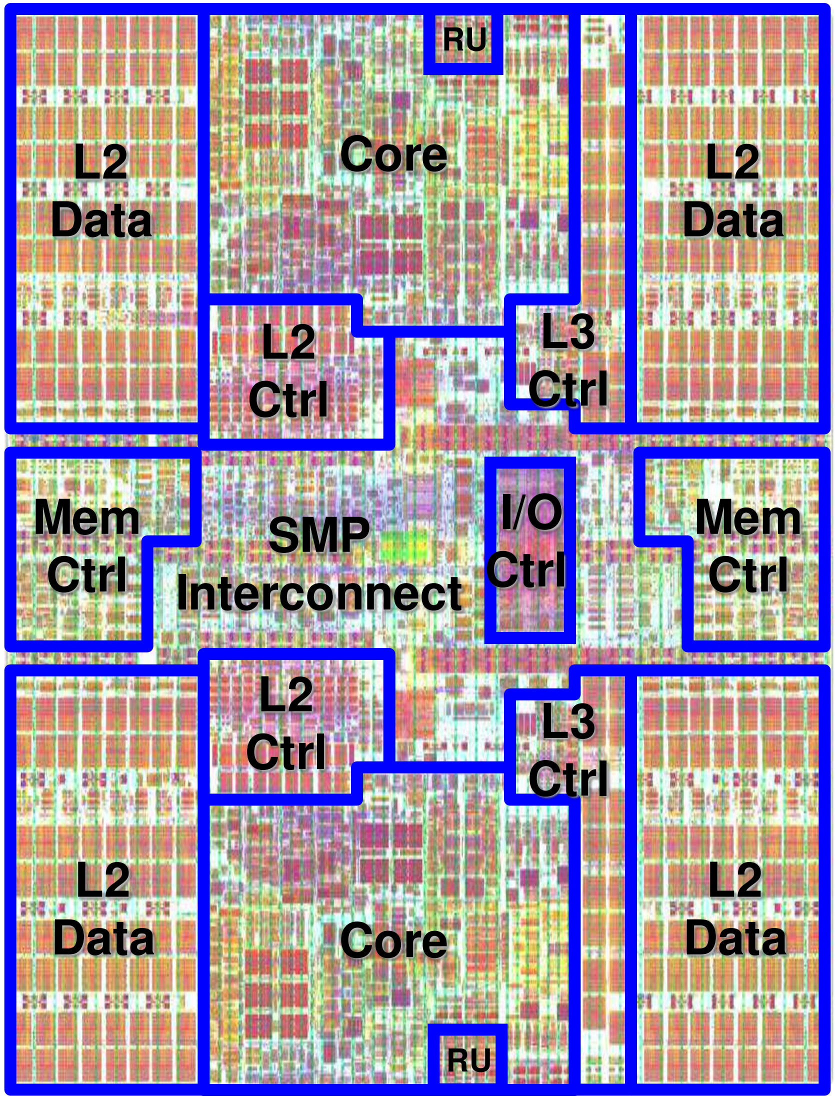
•Simple, scalable global DVFS control algorithms: optimize perf for given power budget

C. Isci, A. Buyuktosunoglu, C-Y-Cher, P. Bose, M. Martonosi, MICRO-39, 2006

K. Reick et al. Hot Chips-2007

POWER6 Chip Overview

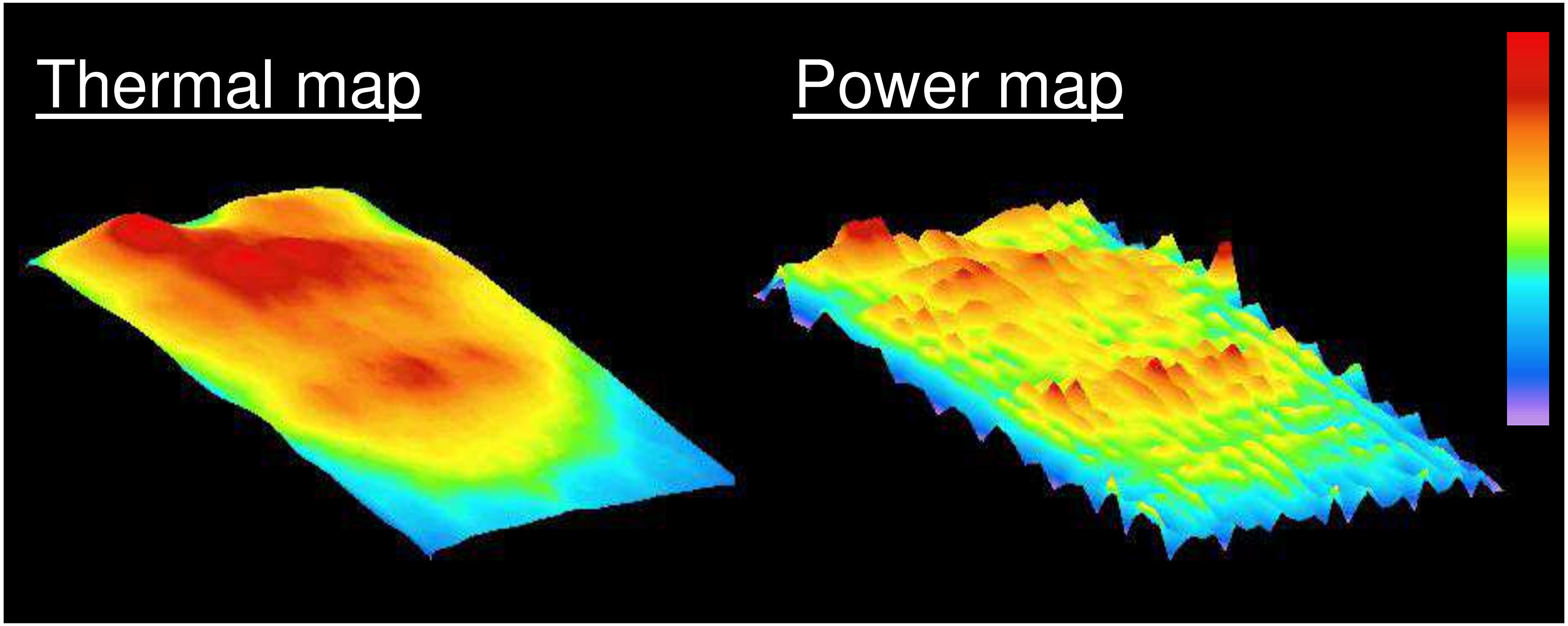
- **Ultra-high frequency dual-core chip**
  - 7-way superscalar, 2-way SMT core
  - 9 execution units
    - 2LS, 2FP, 2FX, 1BR, 1VMX,1DFU
  - 790M transistors
  - Up to 64-core SMP systems
  - 2x4MB on-chip L2
  - 32MB On-chip L3 directory and controller
  - Two memory controllers on-chip
  - Recovery Unit
- **Technology**
  - CMOS 65nm lithography, SOI
- **High-speed elastic bus interface at 2:1 freq**
  - I/Os: 1953 signal, 5399 Power/Gnd



Research Issue: power-efficient RAS microarchitecture

in the face of increasing SER and other sources of unreliability

POWER5 Hotspot Patterns



Thermal map

Power map

-50 different workloads for POWER5 imaged & analyzed

- •HotGen microbenchmark generator tool

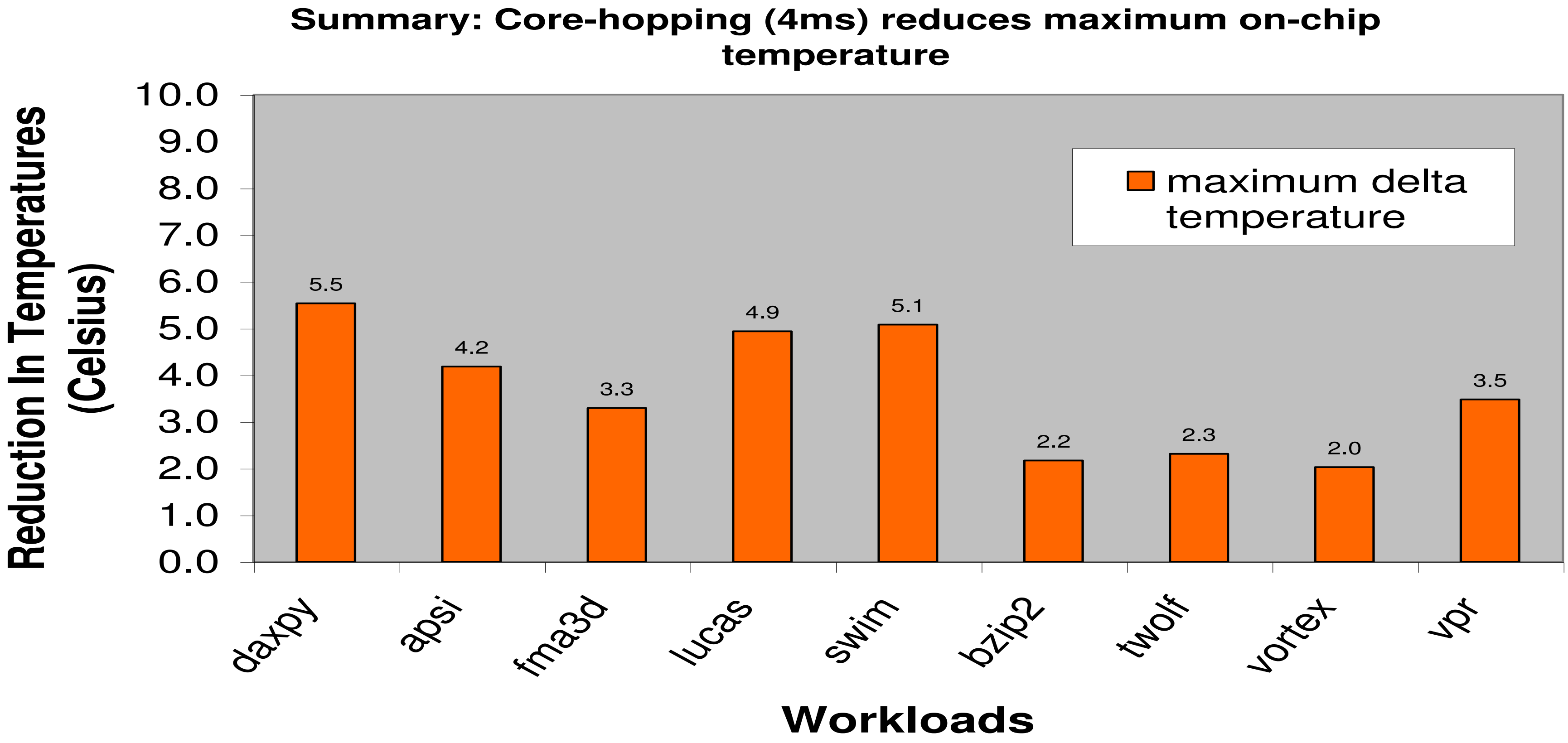
- observed significant differences in circuit utilization

(H. Hamann et al., ISSCC-2006)

Leveraging Spatial Heat Slack

Activity Migration reduces Hotspots

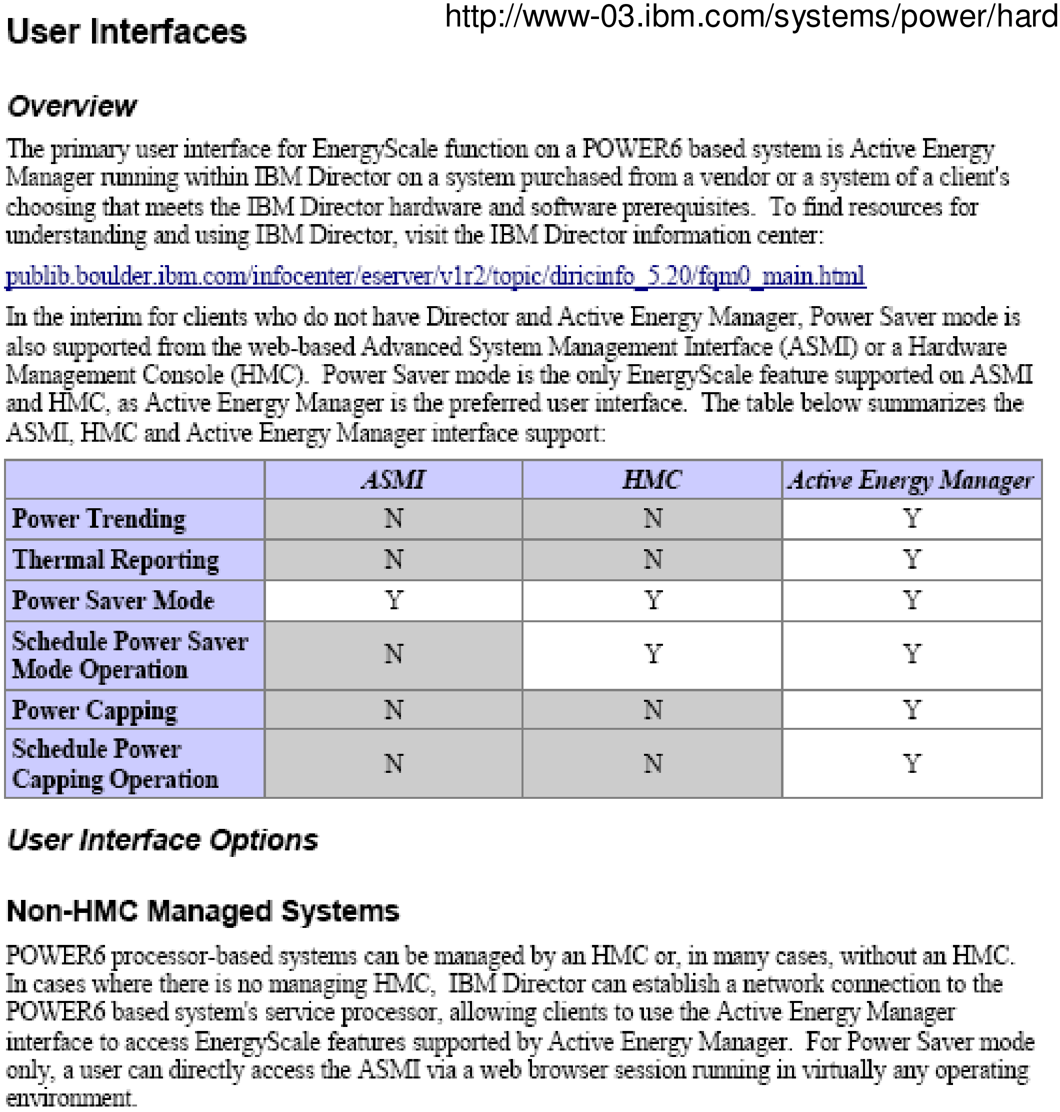
J. Choi, C-Y, Cher et al., ISLPED07



% slow down	0.1	-1.1	-0.5	0.4	1.0	1.1	1.6	0.9	2.5

A Page from IBM EnergyScale for POWER6 Systems

http://www-03.ibm.com/systems/power/hardware/whitepapers/energyscale.html



User Interfaces

Overview

The primary user interface for EnergyScale function on a POWER6 based system is Active Energy Manager running within IBM Director on a system purchased from a vendor or a system of a client's choosing that meets the IBM Director hardware and software prerequisites. To find resources for understanding and using IBM Director, visit the IBM Director information center:

publib.boulder.ibm.com/infocenter/eserver/v1r2/topic/diricinfo_5.20/fqm0_main.html

In the interim for clients who do not have Director and Active Energy Manager, Power Saver mode is also supported from the web-based Advanced System Management Interface (ASMI) or a Hardware Management Console (HMC). Power Saver mode is the only EnergyScale feature supported on ASMI and HMC, as Active Energy Manager is the preferred user interface. The table below summarizes the ASMI, HMC and Active Energy Manager interface support:

	ASMI	*HMC*	*Active Energy Manager*
Power Trending	N	N	Y
Thermal Reporting	N	N	Y
Power Saver Mode	Y	Y	Y
Schedule Power Saver Mode Operation	N	Y	Y
Power Capping	N	N	Y
Schedule Power Capping Operation	N	N	Y

User Interface Options

Non-HMC Managed Systems

POWER6 processor-based systems can be managed by an HMC or, in many cases, without an HMC. In cases where there is no managing HMC, IBM Director can establish a network connection to the POWER6 based system's service processor, allowing clients to use the Active Energy Manager interface to access EnergyScale features supported by Active Energy Manager. For Power Saver mode only, a user can directly access the ASMI via a web browser session running in virtually any operating environment.