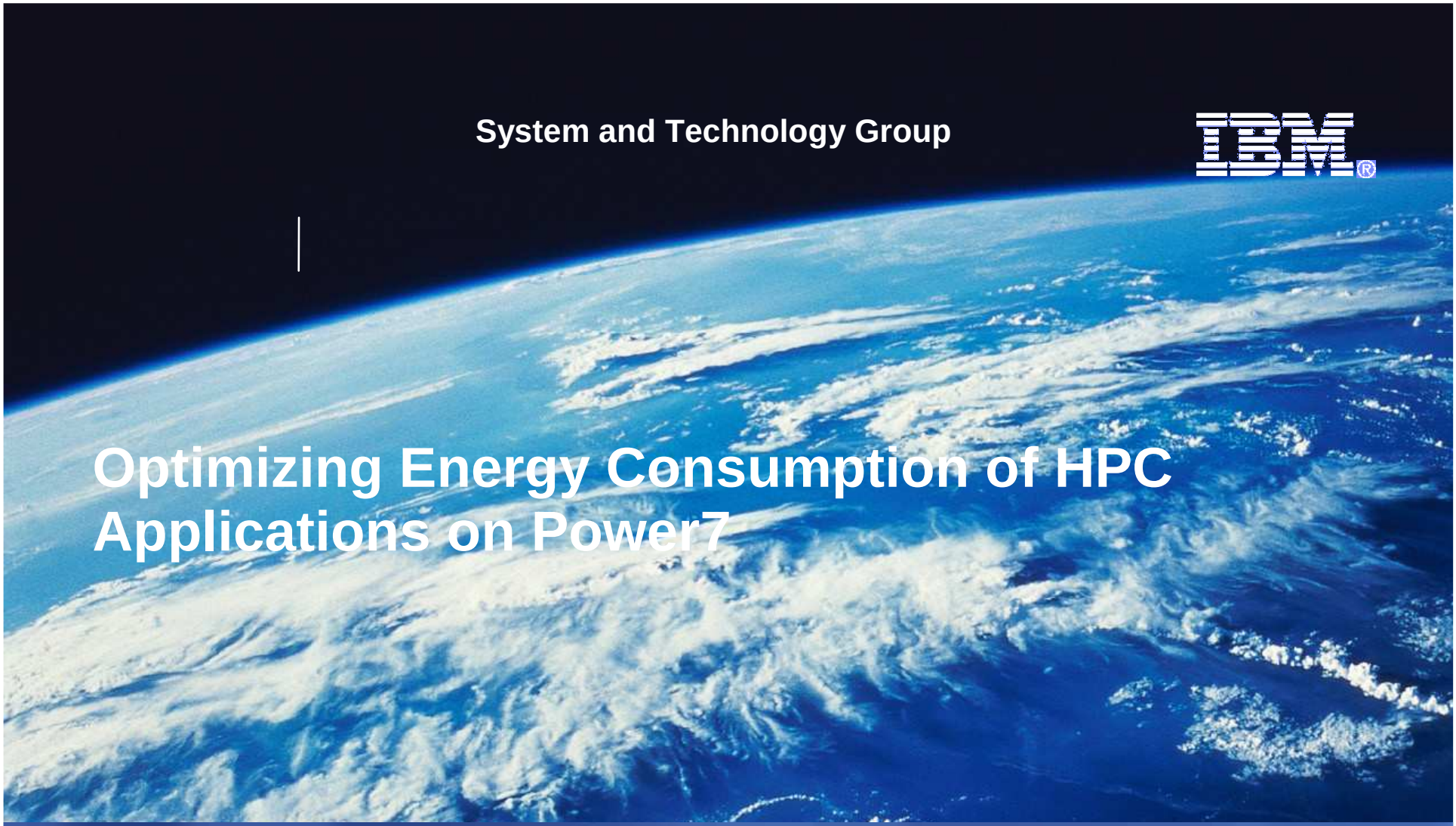
System and Technology Group



Optimizing Energy Consumption of HPC Applications on Power7

Luigi Brochard, Raj Panda, Sid Vemugati, IBM System & Technology Group

ENA-HPC, Hamburg , September 17 2010

Goals

- **Measure fine grain power consumption when running HPC applications workloads**
- **Define correlations between application characteristics and power consumptions**
- **Build a simple method to predict power consumption at different frequencies**

Agenda

- **Tools , Systems and Applications**
- **Measurements**
- **Analysis and Correlation**
- **Power Consumption Model**
- **Examples of Energy Optimization**

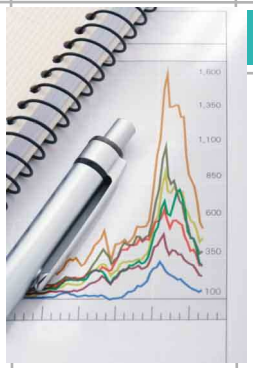
Tools

Tools

- **Active Energy Manager (AEM)**
  - Product available on all platforms
  - Measure power consumptions at the node/blade and frame/chassis level
- **AMESTER**
  - Research tool available on JS22
  - Measure detailed power consumptions at processor and memory level
- **HPM counters**
  - Tool available on System x and System p platforms
  - Provide Hardware Counters

IBM Systems Director Active Energy Manager

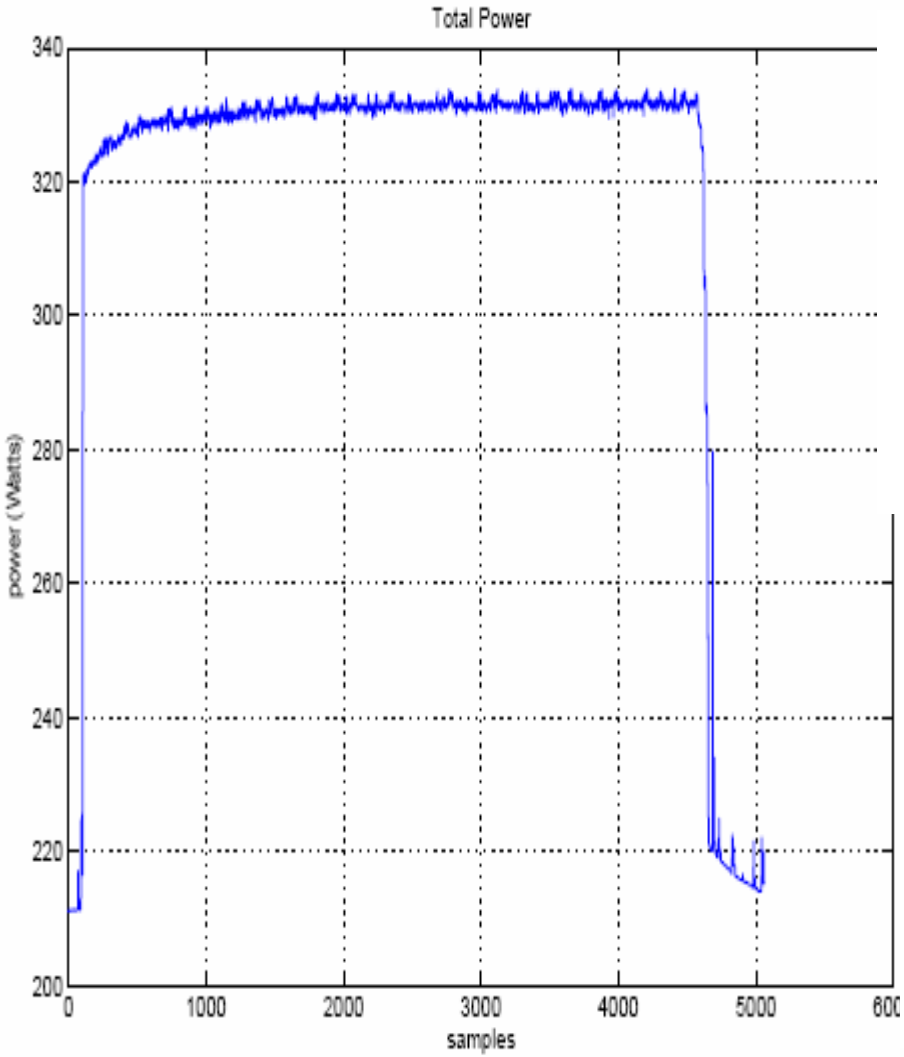
- *Exploit Energy Scale capabilities in POWER servers*
  - **Power Trending, Capping**
  - **Thermal Trending, Effective CPU Trending**
- *Support power savings for POWER 6 and 7 models*
- *Exploit hardware capabilities in IBM System z servers*
  - **Power Trending and Thermal Trending**
  - **Synergistic with "Mainframe Gas Gauge"**
- *Discover and monitor legacy and select non-IBM systems through the intelligent Power Distribution Unit (iPDU)*
  - **Display trending information per load group**
- *Enhancements above PowerExecutive (Windows, xLinux)*
  - **Support for new x86 models: x33850, x3850 M2, x3950 M2, and BC-S 8886**
  - **Cross-system monitoring and management support**
  - **iPDU support**
  - **System polling enhancements**
- *AEM GUI runs on : Windows, Linux/x86, Linux/Power, Linux/System z*

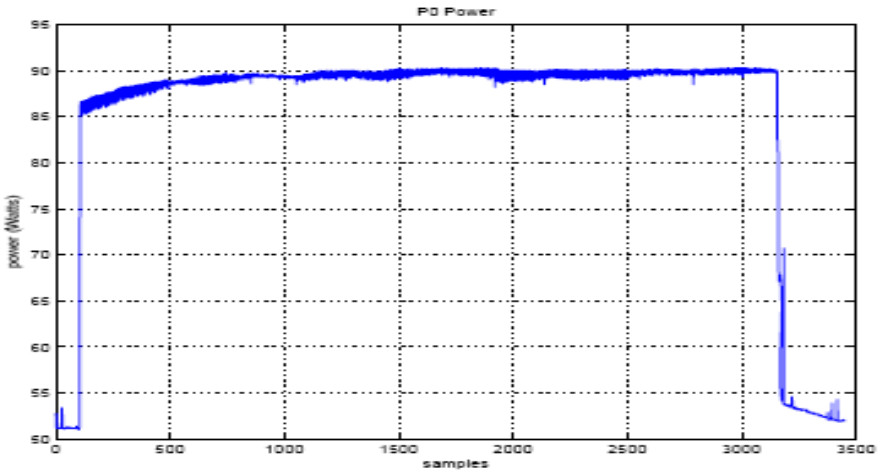


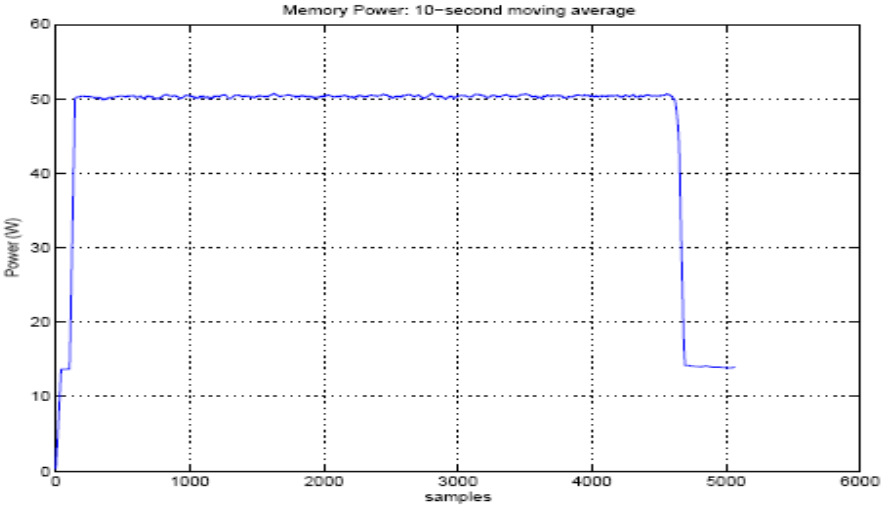
AEM and Amester Comparison

	Amester	Active Energy Manager
Use	A tool for research	A product for customers
Systems with feature	HS20, JS21, JS22, p755,…	many systems
Requirements	Windows or Linux (x86 only)	IBM Director
Testing	None	Yes
Scaling limits	Not tested	Thousands of servers
Power measurement resolution	0.1 W	1 W
Power measurement timescale	1 millisecond	1 minute for all systems and 10 minutes for blades
Documentation	Yes	Yes
Programmable	Excellent	Poor

leslie3d power consumption on JS22 with Amester

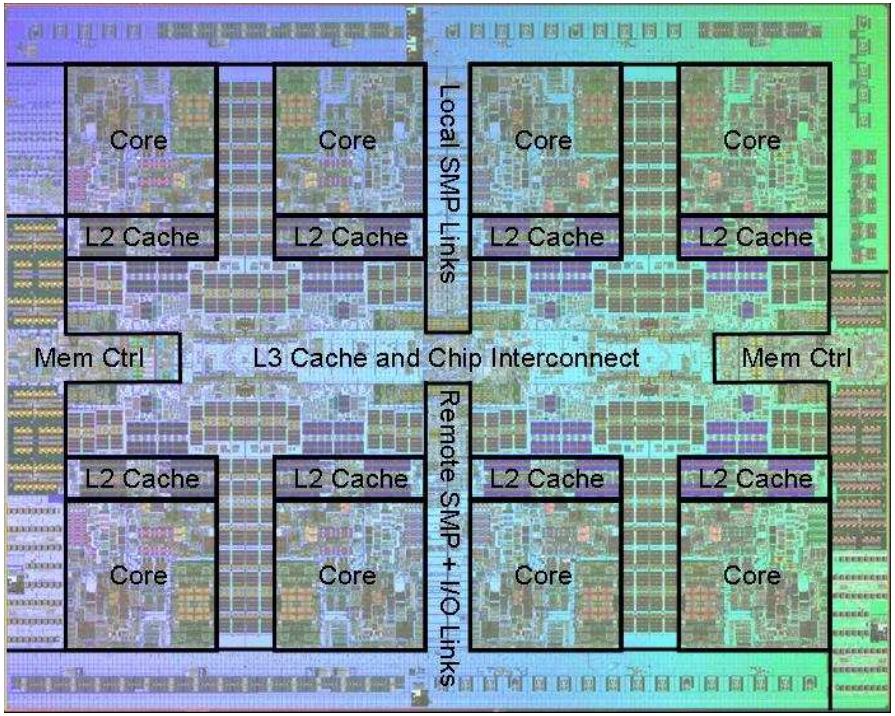






POWER7 Processor

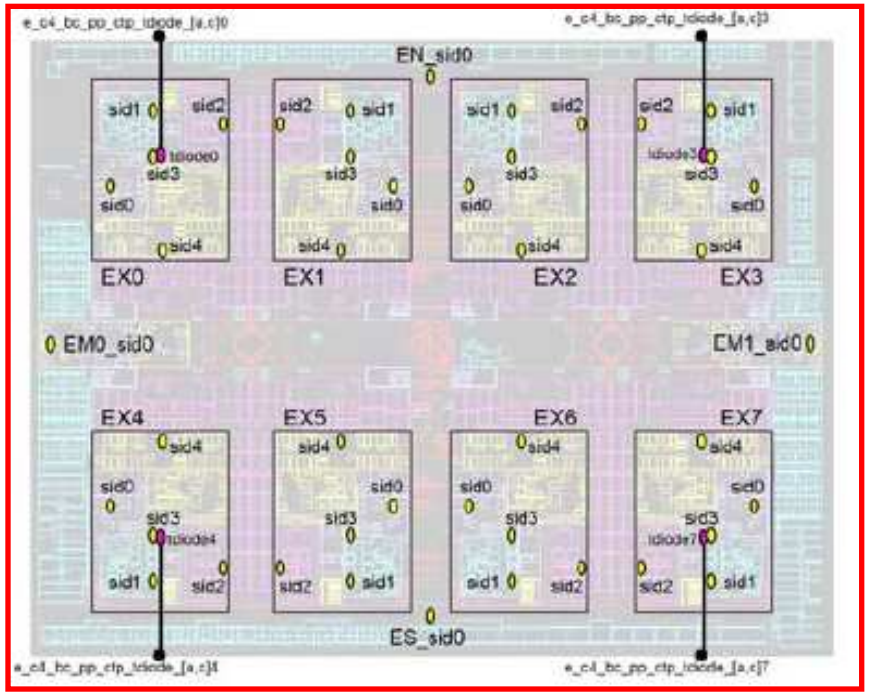
- IBM's 45nm SOI process
- 567 mm2, 1.2B transistors
- 8 out-of-order cores, 4-way SMT
- 32KB L1 D/I, 256KB L2 per core, 32MB shared L3 in IBM's eDRAM process
- 2 on-chip memory controllers, 2 pairs of buffered memory channels each
- Designed for blades, commercial SMPs, supercomputers



4X cores in similar power envelope

Designed for energy-efficiency and effective power management.

Thermal, Power and Activity Sensors



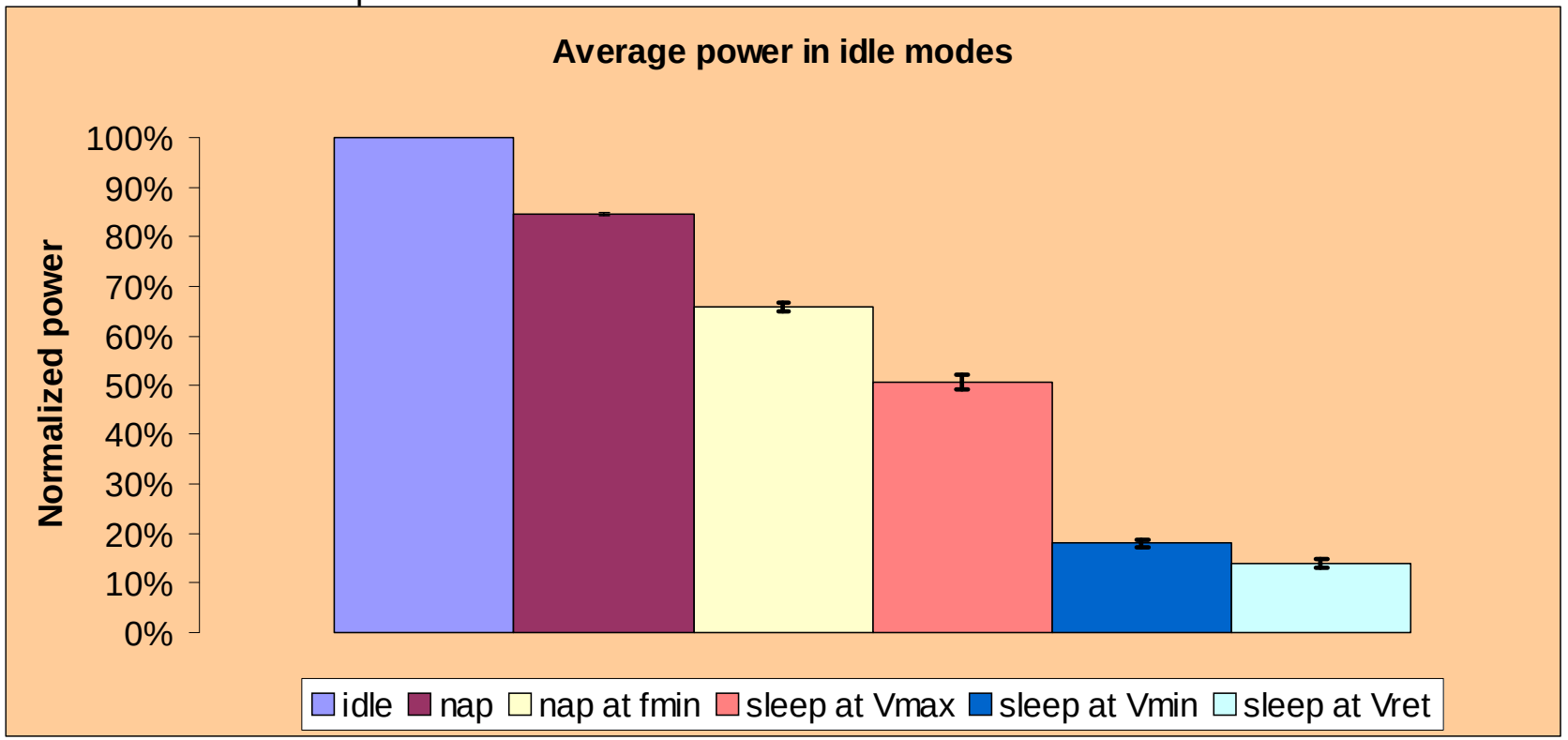
- 44 digital thermal sensors (5 per chiplet, 4 extra-chiplet) on chip; Max chiplet thermal sensor(s) also directly available to firmware.
- On-board ambient temperature sensor, memory buffer/DIMM thermal sensors and VRM thermal-trip logic.
- On-board measurement circuits and A/D channels for
  - full system,
  - processor socket,
  - memory sub-system, I/O sub-system and fan power measurements
- Performance/activity sensors
  - Core-level usage with active cycle counts, instruction throughput counts
  - Core-level memory hierarchy usage – event-based programmable weight counters for frequency impact at high loads
  - Memory controller-level activity – requests and power-mode usage stats

P7 Power Save States

Power Save States	Freq (Max)
Pstate0	**1.1**
Pstate1	**1.0**
Pstate2	**0.9**
Pstate3	**0.8**
Pstate4	**0.7**
Pstate5	**0.6**
Pstate6	**Fmax@vmin**
Pstate7	**0.50**

Architected Idle Modes (OS+hypervisor managed)

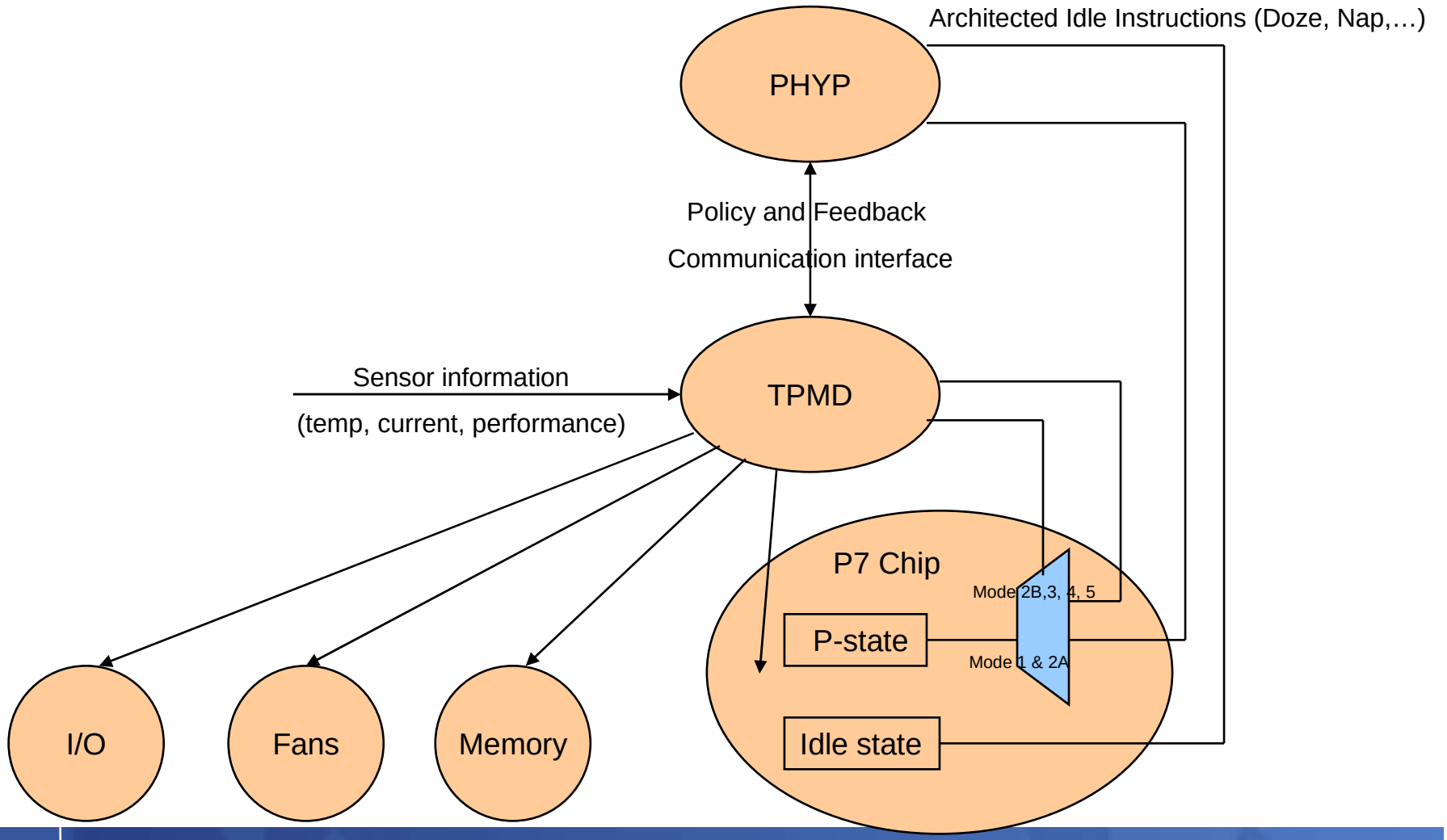
- *Nap* - clocks off for execution units and L1 caches within core
  - Optional *auto* frequency drop support for additional power reduction
- *Sleep* – clocks off for entire chiplet, caches flushed prior to entry.
  - Optional *auto* voltage drop to *latch-state-retention level* for additional power reduction when all cores sleep.



IBM EnergyScale functions

- **Power / Thermal Trending**
  - Collect and report power consumption, inlet and exhaust temp
- **Power Capping**
  - **Guaranteed (Hard Cap)**
    - Enforces a power cap via Dynamic Frequency and Voltage Slewing
  - **Soft Power Cap**
    - Attempted lower cap, but not guaranteed.
- **Energy Management Modes – Enhanced for P7**
  - **Static Power Save (SPS)**
    - Save power via a fixed voltage and frequency drop – as much as 30% down for P7
  - **Dynamic Power Save (DPS)**
    - Optimize power vs performance using Dynamic Voltage and Frequency Slewing
    - Will provide performance boost at very high utilization
    - Will save power at most utilizations
  - **Dynamic Power Save - Favor Performance (DPS-FP)**
    - Will provide performance boost at most utilizations
    - Will save power only at very low utilization

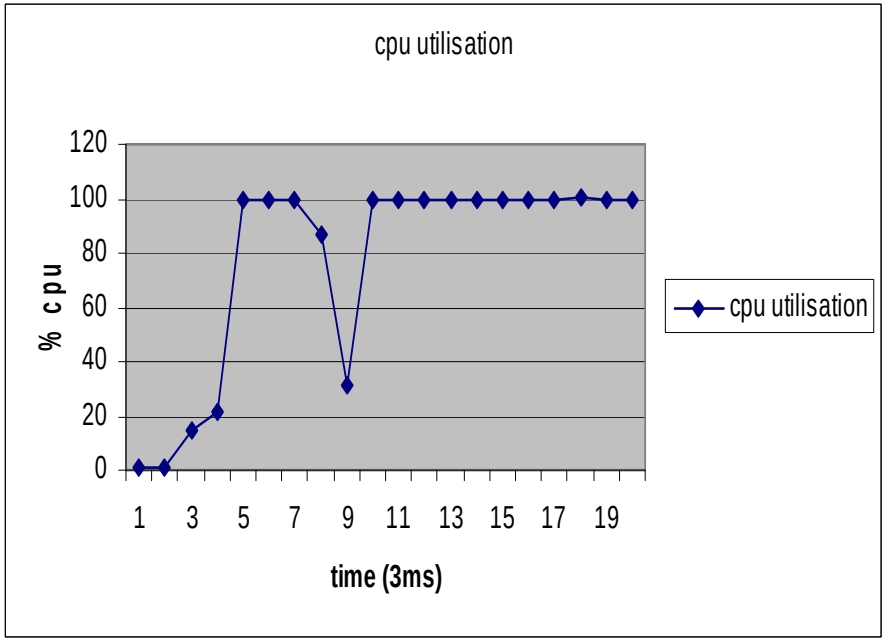
High Level System Power Control View

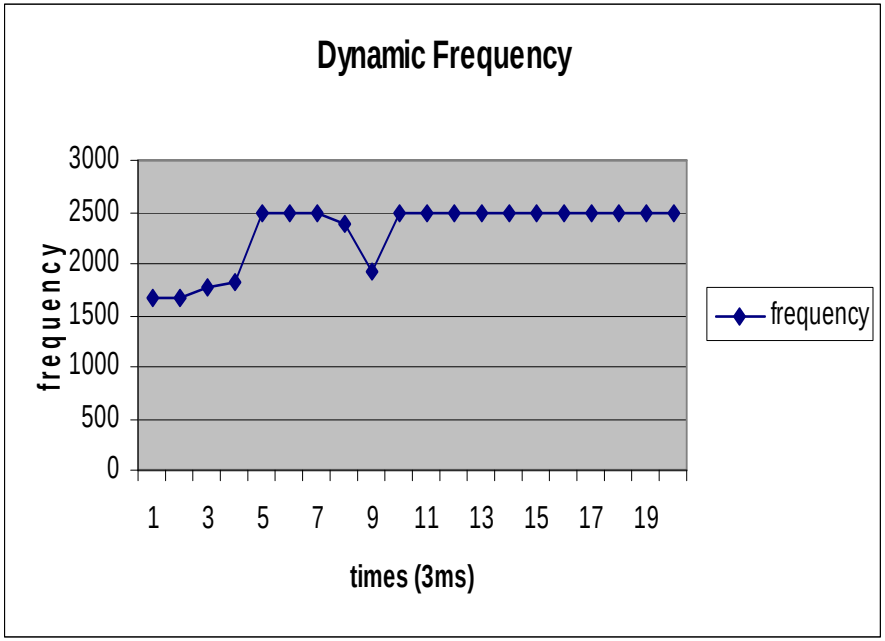


p750 characteristics

System	Processor	Nominal /Power Save Frequency (GHz)	Memory / speed (MHz)	Cores
p750	IBM POWER7	3.55/2.50	32 x 4GB 1066	32

SPS on p750





Applications

Application	Area
416.gamess	Quantum Chemistry
433.milc	Physics
435.gromacs	Molecular Dynamics
437.leslie3d	Fluid Dynamics
444.namd	Molecular Dynamics
454.calculix	Structural Analysis
459.GemsFDTD	Electromagnetics
481.wrf	Weather Forecasting

Note: All applications are from SPEC CPUFP2006 suite

Measurements

Components of power consumption on p750, 3.55 GHz

	Average Power			
	Total	Core	DIMM	Static
416.gamess	984	724	110	151
433.milc	1084	646	268	171
435.gromacs	1002	705	144	153
437.leslie3d	1141	699	266	176
444.namd	1001	717	135	149
454.calculix	1041	747	142	151
459.GemsFDTD	1070	645	264	161
481.wrf	1147	750	234	163
idle power	768	517	98	153

“Core” includes 2 processor cores,L1 cache + NEST (memory, fabric, L2 and L3 controllers,..)
“Staticr” includes, L2 cache, Nova chip, P5IOC2, VRM losses, etc.

Correlation of Power and Performance

POWER PMU – Events and Metrics Used for Analysis

- Events used
  - PM_RUN_CYC
    - Non Idle Cycles
  - PM_RUN_INST_CMPL
    - Non Idle Instructions Completed
  - PM_MEM_RQ_DISP
    - Event counted at memory controller to count Memory Read Dispatches.
  - PM_MEM_WR_DISP
    - Event counted at memory controller to count Memory Write Dispatches
- Metrics used
  - CPI = [PM_RUN_CYC/ PM_RUN_INST_CMPL]
  - Read GB/s = (128* (PM_MEM0_RQ_DISP/2))/(PM_CYC/Frequency)
  - Write GB/s = (128* PM_MEM0_WR_DISP/2))/(PM_CYC/Frequency)
  - Total BW = Read GB/s + Write GB/s
  - GIBS = 32 * Frequency / CPI
  - Where clk_ratio = Core to nest clock ratio = 2

Components of power consumption on p750

			Average Power			
	CPI	GB/s	Total	Core	DIMM	Static
416.gamess	0.59	0.06	984	724	110	151
433.milc	2.74	40.00	1084	646	268	171
435.gromacs	0.75	0.98	1002	705	144	153
437.leslie3d	0.91	39.42	1141	699	266	176
444.namd	0.72	0.36	1001	717	135	149
454.Calculix	0.53	3.07	1041	747	142	151
459.GemsFDTD	1.96	38.54	1070	645	264	161
481.wrf	0.58	29.15	1147	750	234	163
idle power			768	517	98	153

“Core” includes 2 processor cores,L1 cache + NEST (memory, fabric, L2 and L3 controllers,..)
“Staticr” includes, L2 cache, Nova chip, P5IOC2, VRM losses, etc.

Components of power consumption on p750

			Average Power			
	CPI	GB/s	Total	Core	DIMM	Static
416.gamess	0.59	0.06	984	724	110	151
433.milc	2.74	40.00	1084	646	268	171
435.gromacs	0.75	0.98	1002	705	144	153
437.leslie3d	0.91	39.42	1141	699	266	176
444.namd	0.72	0.36	1001	717	135	149
454.Calculix	0.53	3.07	1041	747	142	151
459.GemsFDTD	1.96	38.54	1070	645	264	161
481.wrf	0.58	29.15	1147	750	234	163
idle power			768	517	98	153

“Core” includes 2 processor cores,L1 cache + NEST (memory, fabric, L2 and L3 controllers,..)
“Staticr” includes, L2 cache, Nova chip, P5IOC2, VRM losses, etc.

Components of power consumption on p750

			Average Power			
	CPI	GB/s	Total	Core	DIMM	Static
416.gamess	0.59	0.06	984	724	110	151
433.milc	2.74	40.00	1084	646	268	171
435.gromacs	0.75	0.98	1002	705	144	153
437.leslie3d	0.91	39.42	1141	699	266	176
444.namd	0.72	0.36	1001	717	135	149
454.Calculix	0.53	3.07	1041	747	142	151
459.GemsFDTD	1.96	38.54	1070	645	264	161
481.wrf	0.58	29.15	1147	750	234	163
idle power			768	517	98	153

“Core” includes 2 processor cores,L1 cache + NEST (memory, fabric, L2 and L3 controllers,..)
“Staticr” includes, L2 cache, Nova chip, P5IOC2, VRM losses, etc.

Power consumption on p 750

– on p750 (normal= 3.55 GHz, power save = 2.5 GHz

th				Average Power			
	CPI	GIPS	GBS	Total	Core	DIMM	Static
416.games	0.59	6.05	0.06	984	724	110	151
433.milc	2.74	1.30	40.00	1084	646	268	171
435.groma	0.75	4.75	0.98	1002	705	144	153
437.leslie3	0.91	3.90	39.42	1141	699	266	176
444.namd	0.72	4.95	0.36	1001	717	135	149
454.calculi	0.53	6.74	3.07	1041	747	142	151
459.GemsF	1.96	1.82	38.54	1070	645	264	161
481.wrf	0.58	6.15	29.15	1147	750	234	163
				16.6%	18.4%	13.8%	2.4%
idle power				768	517	98	153

th				Average Power			
	CPI	GIPS	GBS	Total	Core	DIMM	Static
416.games	0.61	4.06	0.06	665	423	103	138
433.milc	1.94	1.28	39.11	841	416	265	160
435.groma	0.75	3.34	0.69	704	417	143	144
437.leslie3	0.65	3.84	38.78	874	453	264	157
444.namd	0.72	3.48	0.25	695	423	131	141
454.calculi	1.07	2.33	2.18	721	441	138	142
459.GemsF	1.44	1.73	37.58	832	417	261	154
481.wrf	0.55	4.51	21.30	811	449	210	152
				31.6%	8.6%	18.5%	2.5%
idle power				541	304	98	139

Power Consumption Model

Model for the power consumption

$$PWR(f_n) = A_n*GIPS(f_0) + B_n*GBS(f_0) + C_n$$

GIPS(f0) and GBS(f0) are application characteristics measured at the nominal frequency (f0). An, Bn and Cn are measured for the given platform at all possible frequencies.

This model hides the dependency of GIPS and GBS of a given workload with the clock frequency.

Power consumption model: Coefficients

$$PWR(fn) = An*GIPS(f0) + Bn*GBS(f0) + Cn$$

Platform	Clock	An	Bn	Cn
p750 ST	2.5	8.4	4.3	666.3
p750 ST	3.55	22.3	4.3	877.1

Power consumption on p 750

– on p750 (normal= 3.55 GHz, power save = 2.5 GHz

	CPI	GIPS	GBS		Average Power				core + DIMM based	
					Total	Core	DIMM	Static	Projection	Error
416.games	0.59		6.05	0.06	984	724	110	151	1012	2.85%
433.milc	2.74		1.30	40.00	1084	646	268	171	1080	0.41%
435.groma	0.75		4.75	0.98	1002	705	144	153	987	1.52%
437.leslie3	0.91		3.90	39.42	1141	699	266	176	1135	0.47%
444.namd	0.72		4.95	0.36	1001	717	135	149	989	1.19%
454.calculi	0.53		6.74	3.07	1041	747	142	151	1041	0.01%
459.GemsF	1.96		1.82	38.54	1070	645	264	161	1085	1.43%
481.wrf	0.58		6.15	29.15	1147	750	234	163	1141	0.54%
					16.6%	18.4%	13.8%	2.4%	Avg.	1.05%

	CPI	GIPS	GBS		Average Power				core + DIMM based	
					Total	Core	DIMM	Static	Projection	Error
416.games	0.61		4.06	0.06	665	423	103	138	701	5.41%
433.milc	1.94		1.28	39.11	841	416	265	160	847	0.67%
435.groma	0.75		3.34	0.69	704	417	143	144	697	1.01%
437.leslie3	0.65		3.84	38.78	874	453	264	157	867	0.88%
444.namd	0.72		3.48	0.25	695	423	131	141	697	0.19%
454.calculi	1.07		2.33	2.18	721	441	138	142	695	3.53%
459.GemsF	1.44		1.73	37.58	832	417	261	154	844	1.43%
481.wrf	0.55		4.51	21.30	811	449	210	152	796	1.80%
					31.6%	8.6%	18.5%	2.5%	Avg.	1.87%

Performance and Power trade-off on p750

– on p750 using power save

- Memory bound workloads have ~3% performance degradation while power and energy saving is ~20%. mem bound
- Cpu bound workloads suffer 30 to 40% performance degradation leading to no/little energy saving cpu bound

p750 , from 3.55 to 2.5 GHz (downclocking)

	2.5 GHz	2.5 GHz	2.5 GHz	2.5 GHz	2.5 GHz	3.55GHz	3.55GHz	3.55GHz	3.55GHz	3.55GHz	Delta	Saving	Saving
workload	**CPI**	**BW(GB/s)**	**Runtime**	**Power**	**Energy**	**CPI**	**BW(GB/s)**	**Runtime**	**Power**	**Energy**	**Perf**	**Power**	**Energy**
416.gamess	0.61	0.06	1348.00	664.60	3.9	0.59	0.06	956.0	984.3	4.1	-41.0%	32.5%	4.8%
433.milc	1.94	39.11	575.00	840.96	2.1	2.74	40.00	563.0	1084.2	2.6	-2.1%	22.4%	20.8%
435.gromacs	0.75	0.69	738.00	704.45	2.3	0.75	0.98	517.0	1002.5	2.2	-42.7%	29.7%	-0.3%
437.leslie3d	0.65	38.78	487.00	874.41	1.8	0.91	39.42	480.0	1140.7	2.4	-1.5%	23.3%	22.2%
444.namd	0.72	0.25	521.00	695.27	1.6	0.72	0.36	365.0	1001.1	1.6	-42.7%	30.5%	0.9%
454.calculix	1.07	2.18	625.00	720.68	2.0	0.53	3.07	442.0	1041.0	2.0	-41.4%	30.8%	2.1%
459.GemsFDTD	1.44	37.58	799.00	831.81	2.9	1.96	38.54	776.0	1069.6	3.6	-3.0%	22.2%	19.9%
481.wrf	0.55	21.30	641.00	811.06	2.3	0.58	29.15	474.0	1147.3	2.4	-35.2%	29.3%	4.4%

Conclusion

- **We don’t need a power meter !!!**
- **We can predict the power consumption at different frequencies with reasonable accuracy**
  - **=> Power & Energy Aware Scheduling**