

Fig. 2. POWER6 floorplan and thermal map [8].

cores, clock and power gating, adaptive microarchitectures, voltage/frequency scaling, mixed threshold-voltage designs, and area and power efficient latch design.

Similar trend and techniques can also be seen in POWER processor family. POWER4 was a performance-centric processor with minimal clock gating, as power consumption was not as critical as it is nowadays. POWER5 featured dual cores and simultaneous multithreading (SMT) to boost throughput, and fine grain dynamic clock gating to manage switching power. Additionally, the usage of low-threshold-voltage transistors was reduced to limit the effect of leakage [10]. POWER6 used a very high frequency as a mean to continue increasing performance, and aggressive fine grain clock gating to manage power and thermals. In addition to that, several actuators, such as processor pipeline throttling, dynamic voltage and frequency scaling (DVFS), low-power modes and memory controller throttling were introduced [12], [9], [5]. POWER7, the latest generation of the family, further boosted performance with 8 cores and 4-way SMT and continued on aggressive clock gating. POWER7 also includes new low-power modes capable of significantly reducing the power consumption when the system is idle [14], [13]. Overall, POWER series microprocessors had progressive improvements both in performance and power with POWER7 achieving the best performance and power efficiency.

B. Early Stage Power/Performance Tradeoff Analysis

Power/performance tradeoff analysis is an integral part of early-stage definition of microprocessors. In general, pre-silicon power-performance modeling and validation methodology is a key investment that will prevent post-silicon surprises. In addition, early stage analysis eliminates fundamental design decision errors that can lead to post-silicon power overruns and performance shortfalls. Furthermore, power analysis and tuning must percolate through all stages of design with closed-loop feedback to higher levels.

One of the areas where early stage analysis takes place is in temperature modeling. Temperature modeling is important because it discovers non-uniform power distribution and resulting hotspots that aggravate thermal challenges significantly.

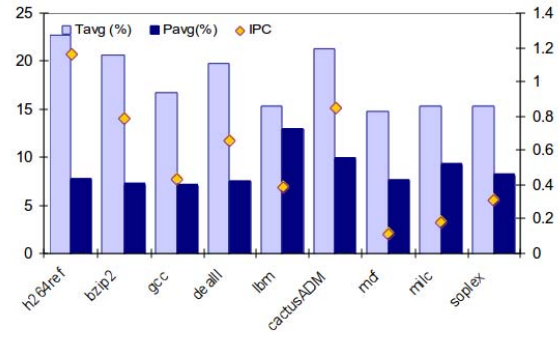


Fig. 3. SPEC CPU2006 temperature/power results when executing 1 thread [6]. Power and temperature are relative to the measured values when the system is idle.

Hotspots limit performance, reliability and increase costs. Predicting hotspots has become more difficult due to variability, multi-core architectures, and power/thermal management. In designing microprocessors, extensive thermal modeling for different workloads and ambient conditions are performed to analyze the thermal characteristics [14], [3], [2] to achieve maximum efficiency. For example, Figure 2 shows the floorplan and thermal profile for a POWER6. In general, while core areas have higher power density than the rest of the chip, careful design planning and analysis resulted in balanced thermal profiles across the chip.

III. POWER/THERMAL AWARE DYNAMIC MANAGEMENT

Power aware adaptive management is a key lever in future power reduction at the chip and system level. Such management requires a careful co-design with circuit/technology and software stack. This section discusses two motivating examples. First example shows how potential benefits of adaptive power management depends on dynamic characteristics of workloads. Second example shows how thread placement in a scheduler can help to adaptively manage power [6], [7].

A. Effect of Workload Characteristics on Power and Thermal Behavior

It is well established that power and thermal behavior of computing systems strongly depend on dynamic characteristics of the running workloads. While, in general, power and thermal behavior change with the amount of activity in the system, there is not a single characteristic factor that directly reflects the power consumption of the system. It is rather a combination of application features such as its IPC and memory intensity. This section presents measured power and thermal characteristics of SPEC CPU2006 benchmarks for a POWER6 system in Figure 3. The figure shows measured average temperature, T_{avg} (percentage over the baseline ¹),

¹We consider the baseline for both power and temperature measurements, the readings obtained when the system is idle (i.e., no processes are available to run and therefore the cores are into low-power mode).

average system power, P_{avg} (percentage over the baseline), and IPC for each benchmark. The results presented in Figure 3 show the strong influence of different workload characteristics on power and thermal behavior. We observe strong deviations among benchmarks in terms of their power and thermal behavior and their associated performance metrics. Here we discuss specific benchmark categories and derive the relations between major workload features and their impact on power and temperature.

CPU-bound benchmarks: We see that high-IPC and highly-CPU-bound benchmarks generally lead to higher core temperatures. Within SPEC CPU2006, the benchmarks that cause higher core temperatures are *h264ref*, *bzip2* and *cactusADM*. These three benchmarks also present the highest IPC among the SPEC CPU2006 benchmarks used in this work.

Memory-bound benchmarks: While CPU-bound benchmarks achieve higher core temperatures, they do not consume the most *total system* power. As Figure 3 shows, memory intensive benchmarks generally consume more power. This is because of the accesses to main memory, which carry significant power cost. For the SPEC CPU2006 benchmarks, memory-intensive ones like *milc* and, especially, *lbm* consume more power than the rest of the benchmarks. For instance, relative to the baseline, *lbm* consumes 5.3% more than *h264ref*, with significantly lower temperature in comparison. The core temperatures are generally low for memory-intensive benchmarks as they spend most of the time waiting for data from the main memory. *mcf* is a low-IPC benchmark with a considerable amount of L2 cache misses per kilo-cycle and with similar characteristics to *milc*. However, the power consumption of *mcf* is considerably smaller (1.7% less). The most significant difference between them is the number of L2 store misses per kilo-cycle, which is 10X higher for *milc*. This is because accessing main memory for a store operation leads to a higher power consumption. Accordingly, *lbm*, which has the highest number of L2 store misses, also shows the highest power consumption among the evaluated benchmarks. In general, metrics such as cache misses per cycle are good indicators of the power consumption in the memory subsystem.

B. Thread Placement

With the arrival of SMT and multi-core architectures, job scheduling techniques have been used in order to reduce power consumption. For instance, Linux provides a setting, `sched_mc_power_savings`, that attempts to save power consumption by grouping several processes into a single chip, therefore leaving other chips idle. An analogous setting, `sched_smt_power_savings`, exists to consolidate several processes into a single core [11]. Other works such as [4] study this problem as well.

In this section we study the effect of thread placement on the power consumption. Given a set of processes, there are different possible ways of assigning them to hardware threads, considerably varying the impact on power and performance. A scheduler that is aware of the workload characteristics can

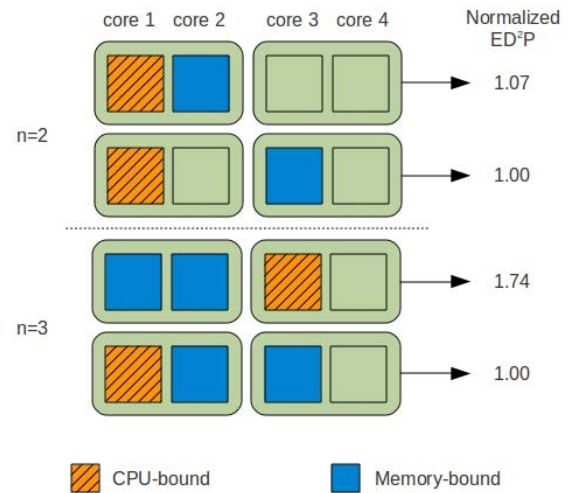


Fig. 4. Thread assignment effect on energy consumption [6]. The test system is a POWER6-based dual-chip, having four cores in total. Each core has a private 4MB L2 cache. Both cores in a chip are connected to memory through a single memory controller.

use this information to increase the system performance and/or reduce the power consumption. Figure 4 displays several thread assignments for the case of two and three threads. Threads colored with orange are CPU-bound while blue ones are memory-bound. In the case of two threads there is no much difference between placing both threads into a single chip or each one into a different chip. However, for three threads, placing both memory-bound workloads on the same chip limits their performance, without decreasing the total system power consumption. Thus, by co-scheduling the high-IPC and the memory-intensive workloads on the same chip we can reduce the interference between them, boosting the performance and reducing the energy consumption (1.7X improvement in the ED^2P).

The availability of thermal and power sensors in systems enables workload schedulers such as Linux OS or system administrators to consider workload characteristics in their scheduling decisions. The results in Figure 4 show that a scheduler may take advantage of the sensors to analyze workload characteristics and interaction in order to achieve optimal performance with minimum energy consumption.

IV. CONCLUSIONS

Power and temperature characteristics have become primary goals for design-time and run-time optimization. While the total power dissipation and power density increases over generations, more sophisticated power management techniques compensate for the aggravated power trends. Such schemes involve detailed pre-silicon modeling to account for variations in workload behavior as well as various levels of process variation. The pre-silicon models are correlated with data from extensive post-silicon characterization. Run-time monitoring and adaptation is another key element to achieve maximum

efficiency. A collection of processor, temperature and performance sensors/counters are placed on chip to provide key information to run-time management mechanisms. Peak performance in next generation microprocessor architectures will be constrained by power dissipation more than any other design criteria. As a result, static and dynamic co-optimization of power performance characteristics are both essential in achieving green processing.

ACKNOWLEDGMENTS

This work was supported by a Collaboration Agreement between IBM and BSC with funds from IBM Research and IBM Deep Computing organizations. It has also been supported by the Ministry of Science and Technology of Spain under contract TIN-2007-60625 and grants AP-2005-3776 and AP-2005-3318, and by the HiPEAC Network of Excellence (IST-004408).

REFERENCES

- [1] http://www.spec.org/cgi-bin/osgresults?conf=power_ssj2008&op=fetch&field=COMPANY&pattern=IBM.
- [2] P. Bose, A. Buyuktosunoglu, C-Y. Cher, J. A. Darringer, M. S. Gupta, H. Hamann, H. Jacobson, P. N. Kudva, E. Kursun, N. Madan, I. Nair, J. A. Rivers, J. Shin, A. J. Weger, and V. Zyuban. Power-efficient, reliable microprocessor architectures: modeling and design methods. 2010.
- [3] J. Casazza et al. Intel Core i7-800 Processor Series and the Intel core i5-700 Processor Series Based on Intel Microarchitecture (Nehalem). 2009.
- [4] Gaurav Dhiman, Giacomo Marchetti, and Tajana Rosing. vGreen: a system for energy efficient computing in virtualized environments. 2009.
- [5] M. S. Floyd, S. Ghiasi, T. W. Keller, K. Rajamani, F. L. Rawson, J. C. Rubio, and M. S. Ware. System power management support in the IBM POWER6 microprocessor. *IBM J. Res. Dev.*, 51(6), 2007.
- [6] V. Jiménez, C. Boneti, F.J. Cazorla, R. Gioiosa, E. Kursun, C-Y. Cher, C. Isci, A. Buyuktosunoglu, P. Bose, and M. Valero. Power and Thermal Characterization of POWER6 System. *International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 2010.
- [7] E. Kursun, C-Y. Cher, A. Buyuktosunoglu, and P. Bose. Investigating the effects of task scheduling on thermal behavior. 2006.
- [8] E. Kursun, J. Wakil, and M. Iyengar. Analysis of the Spatial and Temporal Behavior of Three dimensional Multi-Core Architectures towards Run-Time Thermal Management. *ITherm*, 2010.
- [9] H. Q. Le, W. J. Starke, J. S. Fields, F. P. O'Connell, D. Q. Nguyen, B. J. Ronchetti, W. M. Sauer, E. M. Schwarz, and M. T. Vaden. IBM POWER6 microarchitecture. *IBM J. Res. Dev.*, 51(6), 2007.
- [10] B. Sinharoy, R. N. Kalla, J. M. Tendler, R. J. Eickemeyer, and J. B. Joyner. POWER5 system microarchitecture. *IBM J. Res. Dev.*, 49(4/5), 2005.
- [11] V. Srinivasan, G. R. Shenoy, S. Vaddagiri, D. Sarma, and V. Pallipadi. Energy-Aware Task and Interrupt Management in Linux. *Linux Symposium*, 2, August 2008.
- [12] B. Stolt, Y. Mittlefehldt, S. Dubey, G. Mittal, M. Lee, J. Friedrich, and E. Fluhr. Design and Implementation of the POWER6 Microprocessor. *International Symposium on Solid-State Circuits (ISSCC)*, 2008.
- [13] M. Ware, K. Rajamani, M. Floyd, B. Brock, J.C. Rubio, F. Rawson, and J.B. Carter. Architecting for power management: The IBM® POWER7™ approach. *International Symposium on High-Performance Computer Architecture (HPCA)*, 2010.
- [14] D. Wendel, R. Kalla, R. Cargoni, J. Clables, J. Friedrich, R. French, J. Kahle, B. Sinharoy, W. Starke, S. Taylor, S. Weitzel, S.G. Chu, S. Islam, and V. Zyuban. The Implementation of POWER7: A Highly Parallel and Scalable Multi-Core High-End Server Processor. *International Symposium on Solid-State Circuits (ISSCC)*, 2010.