RETROSPECTIVE: Wattch: A Framework for Architectural-level Power Analysis and Optimizations

David Brooks
Harvard University

Vivek Tiwari
Intel

Margaret Martonosi
Princeton University

I. The Origins of Power-Aware Computing

Circa mid-1990s - there are signs of trouble brewing on the microprocessor roadmap horizon. Traditional Dennard scaling is running out of steam. CPU architects can no longer count on process technologists and circuit designers to give them power reduction for free. Terms like switched capacitance, sub-threshold leakage and decoupling capacitors were entering the lexicon of architects. But they have no easy way to relate these low-level micro phenomena to high-level macro architectural decisions. This was the environment at the time when Wattch appeared on the scene, introducing a much needed quantitative approach to power to the CPU architecture research community.

Power considerations were already prominent for researchers in circuits and design automation topic areas since the early 1990's. By the late 1990's, they were beginning to also be on the minds of computer architects but mostly in the context of low power designs for battery driven mobile or embedded systems. This was not yet a problem that high-performance CPU architects felt they needed to tackle as a primary constraint since thermal dissipation and power delivery costs were still manageable. Process technology scaling had not yet hit the "leakage barrier" that was soon going to prevent both voltage and gate oxide thickness reductions. Cross-chip communication (the finite speed of light!) and the limits of instruction-level parallelism were the hot topics in the architecture community. The special issue of IEEE Computer in September 1997 [1] that resulted as a follow-up to a vigorous debate at ISCA'96, was focused on options for Billion Transistor CPUs. Other than a brief mention of power in the introductory editorial, the rest of the articles did not explicitly address the impending reality that power was going to be the primary limiter for performance for general purpose CPUs. But this was changing [2] and the awareness of power issues among CPU architects was ramping up rapidly. In fact the first Workshop on Power-Driven Microarchitecture was held in conjunction with ISCA'98 in Barcelona (in a small room with a small but highly engaged audience).

We were excited about the opportunities to demonstrate how architectural techniques could mitigate power dissipation challenges and optimize power-performance tradeoffs, but a key hurdle was how to offer quantitative results on the promise and potential of different ideas. Early work in power-aware architecture would use "proxy metrics" to quantify benefits. For example, an earlier paper from Brooks and Martonosi considered narrow-bitwidth operations and offered results on how frequently such optimizations could be applied [3]. What was missing, however, was a holistic architecture-level power model that could be used to run simulations just as instruction-level simulators were commonly used for quantifying the performance benefits of architectural proposals.

II. What is Wattch?

The development of Wattch was motivated by a few key goals:

- To assess the accuracy and promise of architecture-level power modeling approaches.
- To give ourselves the tools we needed to do the research we wanted to do on power optimizations.
- To help broaden the access of the architecture research community to models and tools for power-aware research.
- To bridge the gap between proprietary models used within industry and what was available to the broader research community.

Thus, Wattch was an architecture-level power model that offered chip-level aggregates from module-level estimates of power dissipation for a benchmark run through the simulator. A key goal was to be able to leverage existing tools in broad use. The work would not have been possible if we could not leverage existing tools like SimpleScalar, to model the processor core and collect activity statistics, and CACTI, which we used for optimizing the cache hierarchy. An important part of Wattch's popularity was the tight integration with SimpleScalar, the most widely used architectural simulator of the 1990s and early 2000s [4].

SimpleScalar scaled-up microarchitectural research since it allowed new CPU configurations to be generated through simply choosing a key parameters (out-of-order parameters, number of execution unit, cache size, cache policy etc.). And then the effect of a change being studied could be simulated on actual workloads, allowing a "full-chip" and "full workload" view. But at that time there was no available way to assign a power cost with those parameter changes. Similarly, the version of CACTI available when Wattch was developed could generate timing parameters for cache configurations but not power [5].

A key contribution of Wattch was development of parameterized power models such that the power estimates of the configurable blocks in SimpleScalar were automatically created. These models were developed for the major elements of out-of-order microprocessor cores, including RAM, CAM,

functional units, and clock trees. Thus researchers who were already using SimpleScalar for performance optimization studies now could get power estimates at the same time. Further since this was a simulator, the power estimates could be obtained as a function of time. The power waveform over time also provided the basic information needed for building thermal maps as a function of time and workload.

Having power models available during the early stages of the design process would offer considerable potential leverage for architects considering different organizational and implementation options in their system. A key question was whether such architecture-level power models could be accurate enough at such early design stages. Our work with Wattch helped demonstrate the opportunity for early-stage architecture-level power modeling.

III. Legacy of Wattch

Architectural tools are developed as a means to quantify important metrics and demonstrate the merit of architectural innovations. Power models are no different. We started the research with the goal of exploring power optimizations in CPU microarchitecture and many CPU design innovations have been enabled by Wattch and other similar models.

Soon after developing Wattch, Brooks and Martonosi explored using the tool to evaluate dynamic thermal management schemes [6]. Tools like HotSpot and HotLeakage were later developed as valuable companions to Wattch and versions of such tools are still in use today [7].

A few years after Wattch was developed, it became clear that researchers needed to explore issues beyond just the CPU core – tools like Orion [8] built upon the Wattch-style of power modeling and moved research in this direction. Over time, researchers developed similar power models for heterogeneous system components like GPUs [8] and custom accelerators [9]. At the same time, modern architectural power models like McPAT are extremely useful tools for the architecture community building on this power modeling lineage [10].

In the last few years, there has also been an increased interest in exploring the environmental sustainability of computing systems, treating carbon emissions as a distinct metric from energy efficiency [11]. Inspired by the long history of power and energy modeling, architectural-level carbon modeling tools have recently been developed for both operational and embodied carbon [12].

Part of the impact of Wattch was the facilitation of research efforts in the 2000s that enabled power to be considered as a primary metric alongside performance, and nearly all architectural proposals consider the energy impact of designs. In the early 2000's, many papers provided quantified energy/power/thermal results graphs in their publications. More recently, over the past ten years, the field has shifted a bit. While power remains a fundamental design constraint for essentially all systems, the shift towards large-scale, heterogeneous parallel systems has often meant less focus on power implications of design approaches within a single core. For example, datacenters designers must optimize for thermal and power delivery issues at the scale of individual racks and buildings.

Finally, this paper is an example of how industry and academia can work together. Vivek's PhD thesis work on power while at Princeton benefited from his own internship at Intel and he joined Intel right after graduation. His initial projects were related to building architectural power models but these were proprietary and could not be shared externally. It was conversations with Vivek that encouraged us to work on power in the first place. And very specific to this paper, David's internship at Intel with Vivek was extremely important for understanding implementation details of modern CPUs and for performing validation against lower-level circuit models. This also allowed certain practical and simplifying assumptions that benefited from access to internal models where the relative importance of these assumptions could be assessed. Without these detailed views into real industry designs, the paper and the tool would not have had the credibility nor the impact they were able to have.

References

[1] D. Burger and J. Goodman, "Billion-transistor architectures," *Computer*, vol. 30, no. 9, pp. 46–48, 1997.

[2] V. Tiwari, D. Singh, S. Rajgopal, G. Mehta, R. Patel, and F. Baez, "Reducing power in high-performance microprocessors," in *Proceedings of the 35th annual Design Automation conference*, pp. 732–737, 1998.

[3] D. Brooks and M. Martonosi, "Dynamically exploiting narrow width operands to improve processor power and performance," in *Proceedings Fifth International Symposium on High-Performance Computer Architecture*, pp. 13–22, 1999.

[4] D. Burger and T. M. Austin, "The simplescalar tool set, version 2.0," *SIGARCH Comput. Archit. News*, vol. 25, p. 13–25, jun 1997.

[5] S. Wilton and N. Jouppi, "Cacti: an enhanced cache access and cycle time model," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 5, pp. 677–688, 1996.

[6] D. Brooks and M. Martonosi, "Dynamic thermal management for high-performance microprocessors," in *Proceedings HPCA Seventh International Symposium on High-Performance Computer Architecture*, pp. 171–182, 2001.

[7] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. Stan, "Hotspot: a compact thermal modeling methodology for early-stage vlsi design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14, no. 5, pp. 501–513, 2006.

[8] H.-S. Wang, X. Zhu, L.-S. Peh, and S. Malik, "Orion: A power-performance simulator for interconnection networks," in *Proceedings of the 35th Annual ACM/IEEE International Symposium on Microarchitecture*, MICRO 35, (Washington, DC, USA), p. 294–305, IEEE Computer Society Press, 2002.

[9] Y. S. Shao, B. Reagen, G.-Y. Wei, and D. Brooks, "Aladdin: A Pre-RTL, Power-Performance Accelerator Simulator Enabling Large Design Space Exploration of Customized Architectures," in *ISCA*, 2014.

[10] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "Mcpat: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *2009 42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 469–480, 2009.

[11] U. Gupta, Y. G. Kim, S. Lee, J. Tse, H.-H. S. Lee, G.-Y. Wei, D. Brooks, and C.-J. Wu, "Chasing carbon: The elusive environmental footprint of computing," *IEEE Micro*, vol. 42, p. 37–47, jul 2022.

[12] U. Gupta, M. Elgamal, G. Hills, G.-Y. Wei, H.-H. S. Lee, D. Brooks, and C.-J. Wu, "Act: Designing sustainable computer systems with an architectural carbon modeling tool," in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ISCA '22, (New York, NY, USA), p. 784–799, Association for Computing Machinery, 2022.