

Evaluating Power Consumption of D-NUCA caches

A. Bardine, P. Foglia, G. Gabrielli, C.A. Prete

*Dept. of Information Engineering, University of Pisa, via Diotisalvi 2, 56122 Pisa, Italy
E-mail: {giacomo.gabrielli, alessandro.bardine, foglia, prete}@iet.unipi.it*

Members of HiPEAC, the European Network of Excellence
on High Performance Embedded Architecture and Compilation

ABSTRACT

D-NUCA caches are on-chip cache memories characterized by multi-bank partitioning and data migration. They exhibit high hit rates while keeping the access latency low. As counterpart, such caches are affected by high static and dynamic power consumption. In this work we present a preliminary power consumption evaluation of a D-NUCA cache. Results show the existing balance among static and dynamic contributions to total power budget.

KEYWORDS: Cache; NUCA; Dynamic-NUCA; Wire Delay; Power Consumption; Leakage

1 Introduction

Technology trends are leading to the use of large, on-chip, level-two (L2) and level-three (L3) cache memories. For high frequency systems, the latencies of such caches are dominated by wire delay [1]: in order to reduce the effects of high access latencies, NUCA Caches (Non-Uniform Cache Architectures) [2][3] have been proposed. In a NUCA design the cache is partitioned into many independent banks usually interconnected by an on-chip switched network; in this model banks' access latency is proportional to the physical distance from the cache controller. The mapping between cache lines and banks can either be Static or Dynamic (namely S-NUCA and D-NUCA); in the latter model most frequently accessed cache lines are allowed to dynamically migrate toward the controller. Modern CMOS processes (70 nm and below) are characterized by high static power consumption due to leakage currents [4], and big SRAM structures, like the one employed in a NUCA-based system, are responsible for a big portion of total leakage power budget. D-NUCA caches also introduce extra dynamic power consumption due to increased bank accesses and network traffic (for data search and migration) with respect to a traditional cache. As the reduction of the total power consumption is a big issue in modern and future designs [4], understanding the relative contribution of static and dynamic components for a D-NUCA cache is the very first step to be performed in order to plan future research efforts. In this work we present the evaluation of static and dynamic power consumption of a L2 D-NUCA cache for some SPEC CPU2000 benchmarks [5].

2 D-NUCA Power Consumption Evaluation

The basic elements of a D-NUCA cache are memory banks, network switches and network links. For each of such elements, we derived the energy parameters that should be taken into account for its contribution to static and dynamic power consumption.

Cache bank dynamic	41.5 pJ (read), 17.9 pJ (write)
Cache bank static	235.6 mW (100°C), 137.4 mW (80°C), 70.7 mW (60°C), 35.3 mW (40°C)
NoC flit transmission	6.0 pJ (vert. link), 1.8 pJ (horiz. link)
NoC switch dynamic	135 pJ
NoC switch static	23.1 mW (100°C), 13.5 mW (80°C), 6.93 mW (60°C), 3.46 mW (40°C)
Off-chip access	12.2 nJ

Table 1 - Energy parameters for an 8 MB D-NUCA cache, 16x8 64 KB banks, targeted at 70 nm technology.

We derived the energy parameters for the SRAM banks from CACTI 4.2 tool [6]; we modeled each bank as a whole to obtain energy consumption per access, static power and physical dimensions.

For network links, we calculated the energy required for each transmission adopting a simple RC model [7]; we referred to the Berkeley Predictive Technology Model [8] to derive wire resistance and capacitance per unit length assuming an intermediate wiring layer; the length of each link, and thus its total resistance and capacitance, was determined according to the physical dimensions of cache banks derived from CACTI.

For network switches, the energy parameters were determined according to Muralimanohar *et al.* [9]; in that work a 3-stage switch architecture for worm-hole routed networks is shown. We obtained the following data: energy consumption per flit traversal, static power.

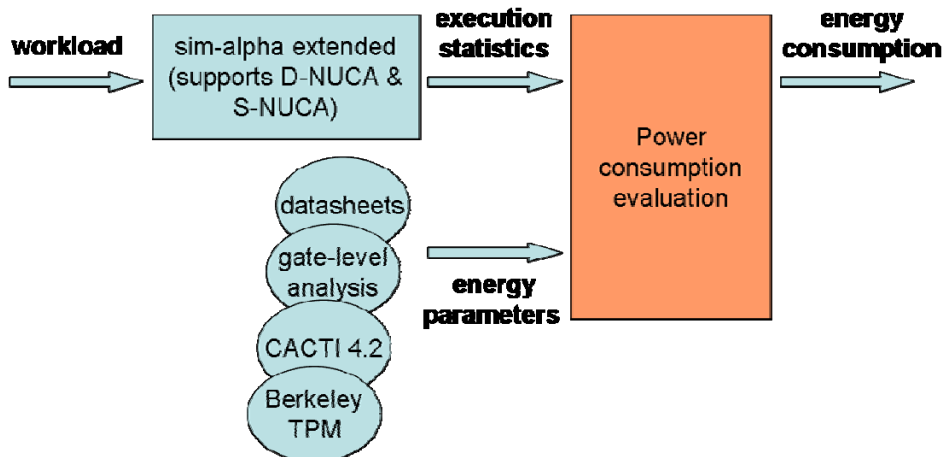


Figure 1 - Evaluation methodology.

To take into account the energy required on each off-chip access caused by a cache miss, we derived the energy dissipated on each main memory access from DRAM module datasheets [10], assuming a modern DDR2 system.

Since leakage power greatly depends on temperature, we scaled the static power values assuming 4 selected operating values: 100°C, 80°C, 60°C and 40 °C. We performed temperature scaling according to the model implemented in HotLeakage tool [11].

3 Experimental Results

The methodology employed to calculate energy consumption is synthesized in Figure 1: the workload is fed to *sim-alpha* simulator [12], which was extended to support NUCA caches, and the resulting execution statistics and the energy parameters are collected to derive total energy consumption. We conducted the experiments with an 8 MByte D-NUCA cache built up of 16x8 64 KByte banks and targeted at 70 nm technology. System clock is assumed to be 5 GHz. Table 1 lists the energy parameters that characterize our evaluation.

In Figure 2 the energy consumption for two different SPEC CPU2000 benchmarks are shown; normalized energy is used to better highlight the different contributions to total power budget, grouped into static energy, dynamic energy and off-chip accesses. It is worth noting that, for both the benchmarks, the static component largely exceeds the dynamic and off-chip components for each temperature value.

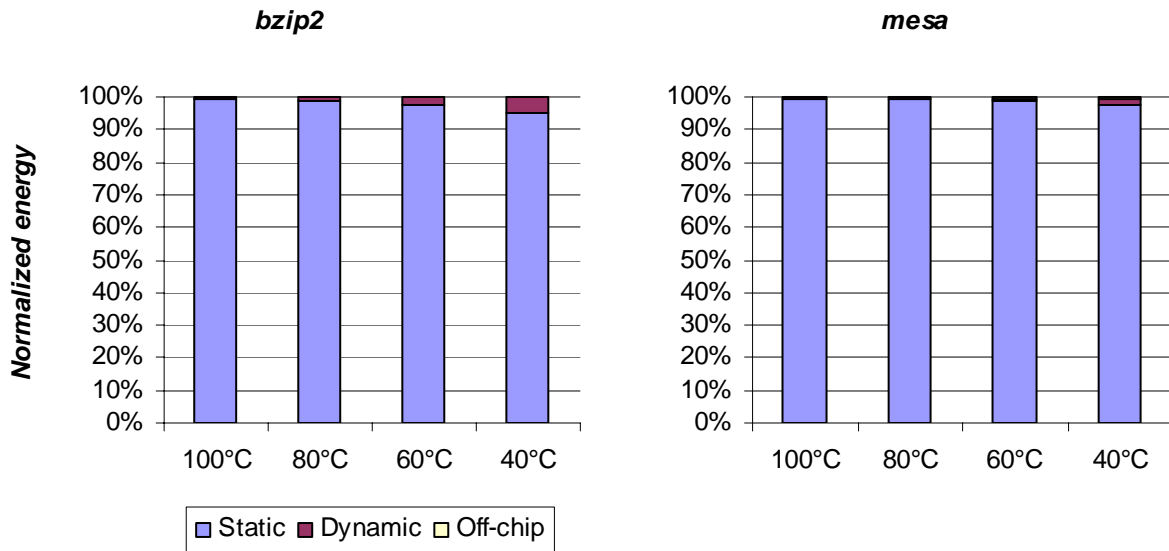


Figure 2 - Energy consumption for two different SPEC CPU2000 benchmarks.

Figure 3 shows the breakdown of dynamic energy components due to on-chip network activity (including switching and transmitting flits), cache bank accesses and off-chip accesses. For both the benchmarks, the contribution due to bank accesses is minimal with respect to the contribution due to NoC activity while the contribution due to off-chip accesses greatly varies between the two benchmarks because of the different miss-rates.

4 Conclusions

The preliminary results of our evaluation show that D-NUCA caches are likely to be dominated by leakage power, even at low temperatures, while the dynamic contribution to total power budget is negligible. For such reasons research efforts should go in the way of reducing static power, and the performance improvements could be pursued at cost of an eventual slight increase in dynamic energy requirements; in fact these latter could be

compensated by the resulting static energy savings thank to the reduced overall run-time. Acting on saving dynamic energy will bring limited results, and efforts should be concentrated in the reduction of energy required for NoC activity.

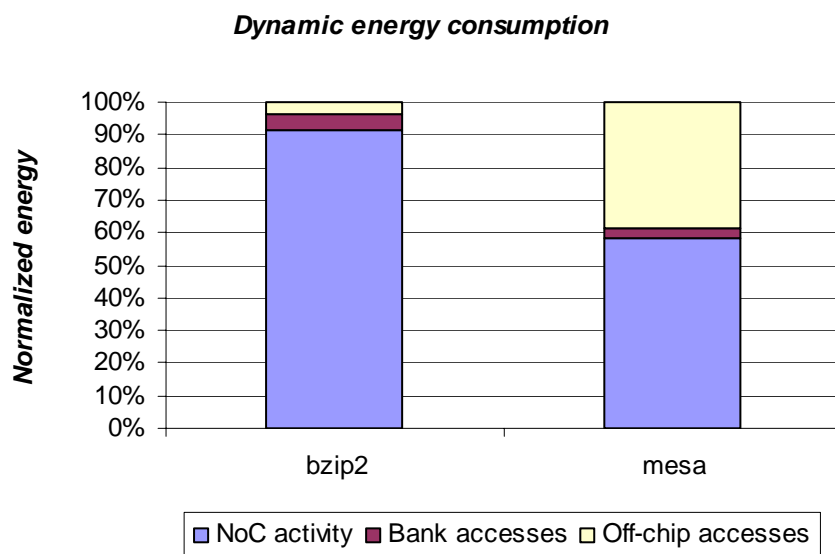


Figure 3 - Dynamic energy consumption for two different SPEC CPU2000 benchmarks.

5 Acknowledgements

We wish to thank Prof. Massimo Macucci who furnished us with the Alpha platform that we used to compile the SPEC CPU2000 benchmarks.

This work is partially supported by the SARC project founded by the European Union under the contract no. 27648.

6 References

- [1] D. Matzke, "Will physical scalability sabotage performance gains?," *IEEE Computer*, 30, Sept. 1997.
- [2] C. Kim, *et al.*, "An Adaptive, Non-Uniform Cache Structure for Wire-Delay Dominated On-Chip Caches," *Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, October 2002.
- [3] C. Kim, *et al.*, "Nonuniform Cache Architectures for Wire-Delay Dominated On-Chip Caches," *IEEE Micro*, 23:6, November/December 2003.
- [4] N. S. Kim, *et al.*, "Quantitative Analysis and Optimization Techniques for On-Chip Cache Leakage Power," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 13:10, October 2005.
- [5] Standard Performance Evaluation Corporation [Online]. Available: <http://www.spec.org/cpu2000/>.
- [6] D. Tarjan, *et al.*, "CACTI 4.0," HP Technical Report, HPL-2006-86, June 2006.
- [7] Neil Weste, and David Harris, "CMOS VLSI Design, A Circuits and Systems Perspective," 3rd edition, Addison Wesley, New York, 2004.
- [8] Predictive Technology Model [Online]. Available: <http://www.eas.asu.edu/~ptm/>.
- [9] N. Muralimanohar, and R. Balasubramonian, Interconnect Design Considerations for Large NUCA Caches, *34th International Symposium on Computer Architecture (ISCA-34)*, San Diego, June 2007.
- [10] Micron 1 GB DDR2 SDRAM Module Datasheet [Online]. Available: <http://www.micron.com>.
- [11] Y. Zhang, *et al.*, "HotLeakage: A Temperature-Aware Model of Subthreshold and Gate Leakage for Architects," Technical Report, CS-2003-05, March 2003.
- [12] R. Desikan, *et al.*, "Sim-alpha: a Validated, Execution-Driven Alpha 21264 Simulator," Technical Report, TR-01-23, 2001.