The University of Jordan

Faculty of Engineering

Computer Engineering Department

Simulation Report

Evaluating Cache Power Dissipation using CACTI 5.3 simulator

Dr. Gheith Abandah

Prepared By

Shereen Ismael

JAN 2010

1. Abstract

Power dissipation has become a major concern to those designing processors for high performance desktops, servers, and battery-operated portable devices. With this trend, designing cache modules focusing singularly on performance is insufficient. Power dissipation has become a top priority for today's microprocessors. What was previously a concern mainly for mobile devices has also become of paramount importance for general-purpose and even high-performance microprocessors, especially with the recent industry emphasis on processor "Performance-per-Watt."

Dynamic power dissipation due to signal transitions is the main power dissipation source nowadays, but static power will become increasingly significant in upcoming processors. While dynamic power is directly related to the activity of the circuits, static power depends on the amount of powered-on transistors and their physical characteristics. Thus, large circuits are usually the main source of static power. Caches are normally the largest structures in the processor, so they are the most important sources of the static power dissipation. It is also known that increasing cache associativity and/or size to reduce the miss ratio and increase performance has an impact on static power and access time. On the other hand, a low access time is desired for performance. Power and performance also depend on the number of ports [6].

In this report, CACTI 5.3 simulator is used to evaluate cache power dissipation combined with nanometer models of various cache configurations. Cache power dissipation is focused on to show how much power cache consumes, and what fraction can be attributed to dynamic (switching) and static (leakage) currents by exploring a three-dimensional cache design space by studying caches with different sizes (32kB to 256kB), associativities (direct-mapped to 16-way) and process technologies (90nm, 65nm, 45nm and 32nm).

2. Background

Higher energy dissipation requires more expensive packaging and cooling technology, which in turn increases cost and decreases system reliability. There are fundamentally two ways in which power can be dissipated as mentioned before: either dynamically (due to switching activity), or statically (which is mainly due to leakage in the gates). Cache power dissipation has typically been significant, especially the increase in static power since most of its transistors are inactive

(dissipating no dynamic power, only static) during any given access. It is therefore essential to properly account for these leakage effects during the cache design process.

One of the most effective ways of reducing the dynamic energy dissipation is to scale down the transistor supply voltage. To maintain high switching speed under reduced voltages, the threshold voltage must be also scaled down accordingly. As the threshold voltage drops, it is easier for current to leak through the transistor resulting in significant leakage energy dissipation.

Static power is dissipated by leakage currents that flow even when the device is inactive. Figure 1 shows a typical 6-transistor memory cell (6TMC) and the typical leakage currents involved for the memory cell idle state (i.e. word line is off, one storage node is "0" and the other is "1"). Although a sizeable number of transistors in a cache are active for any given access, the majority of memory cells are in this inactive state, dissipating static power.

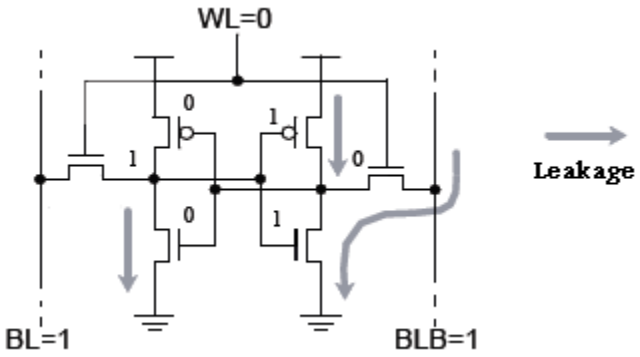

Figure 1: Memory cell leakage currents. An inactive six-transistor memory cell (6TMC) showing the leakage currents flowing across the devices

3. Simulator

CACTI [1] is a tool that has become widely used in the computer architecture community by architects either directly for modeling caches and plain memories, or indirectly through other tools such as Wattch [2]. CACTI 5.3 was a major revision of CACTI overcoming several limitations of earlier versions. First, the technology models of earlier versions were largely based on linear scaling of a 0.8μm process. However, as scaling has progressed to nanometer dimensions, scaling has become much less linear. To better model this, the base technology modeling in CACTI 5.3 was changed from simple linear scaling of the original CACTI 0.8μm technology to models based on the ITRS roadmap [4].

CACTI is an integrated cache and memory access time, cycle time, area, leakage, and dynamic power model. By integrating all these models together, users can have confidence that tradeoffs between time, power, and area are all based on the same assumptions and, hence, are mutually consistent. CACTI is available in two forms: a web-based version and a C++ source code version.

CACTI 5.3 is the latest version of CACTI to be released includes a number of features such as: it takes as input the cache capacity, associativity, cache line size the number of read_write ports and the feature size. It uses analytical models to compute the access time of the cache and the energy consumed by it for different configurations.

3.1 Technology Modeling

CACTI 5.3 makes use of technology projections from the ITRS. In this report it will be focused on four ITRS technology nodes – 90, 65, 45, and 32 nm – which cover years 2004 to 2013 in the ITRS [4].

As mentioned before, improving fabrication process technologies cause transistors to become steadily smaller so device leakage currents are expected to significantly contribute to processor power dissipation.

4. Results and Discussion

Table 1 depicts results produced by CACTI 5.3 for 32kB 2-way set-associative caches with 32-byte cache lines, using high performance transistors. It will give us an overview about the values of energy dissipated in cache.

technology [nm] number of banks	45 1	45 2	65 1	65 2
Access time [nsec]	0.828	0.750	1.392	1.261
Dynamic read energy[nJ]	0.061	0.048	0.107	0.084

Table 1: Results produced by CACTI 5.3 for 2-way set-associative caches with 32-byte cache lines, using 90 nm technology high performance transistors.

$$Dynamic\ Power\ (W) = \frac{Dynamic\ Energy\ \ (nJ)}{Cycle\ Time\ (nS)}$$

Figure 2 shows the power dissipation of the different cache configurations as a function of process technology. Each column of plots represents a specific process technology, while each row represents a specific cache size. Each plot shows the dynamic, leakage, and total power as a function of associativity for the given cache size and technology node.

$$Total\ Power = Dynamic\ Power + Leakage\ Power$$

The most basic observation here is that total power is dominated by the dynamic power in the larger technology nodes but is dominated by static power in the deep nanometer nodes (with the exception of very highly-associative small to medium caches).

This can be seen from the plots for the 90nm and 65nm nodes, where the dynamic power comprises the majority of the total power.

In the 45nm node, the leakage power is significant enough to become dominant in some configurations.

For small caches of any associativity, dynamic power typically dominates because there are fewer leaking devices that contribute to leakage power. But as cache sizes increase, the number of idle transistors that dissipate leakage power also increases, making the leakage power dominant component in total power except for the configurations with high associativity (which requires more operations to be done in parallel, resulting in dissipation of more dynamic power).

In the 32nm node, the leakage power is already comparable to the dynamic power even for small caches (of any associativity), and starts to become really dominant as cache sizes increase (even at high associativities where we expect the cache to burn more dynamic power).

Some of the plots in Figure 2 (e.g. the 256k and 512k caches for the 45nm and 32nm node) shows that increasing cache associativity from direct-mapped to 2-way or 4- way does not automatically cause an increase in power dissipation, as the internal organization may allow a more power optimal implementation of set-associative caches compared to direct-mapped caches, especially for medium to large-sized caches.

5. Conclusion

In this report, power dissipation in nanometer cache configurations in a three-dimensional design space consisting of different cache sizes, associativities and process technologies. There are two ways in which power can be dissipated, either dynamically or statically.

Static power dissipation due to leakage has being dominant source for power consumption as technology scales down below 100nm. Techniques need to be studied that will limit power consumption like turn off cache lines that are not used for a long time or, use low supply voltage to save power and other ideas that will guide future power management in cache designs.

6. References

[1] http://www.hpl.hp.com/research/cacti/

[2] D. Brooks, V. Tiwari, and M. Martonosi. Wattch: A framework for architectural-level power analysis and optimizations. In *ISCA*, Jun 2000.

[3] S. Rodriguez and B. Jacob.Energy/Power Breakdown of Pipelined Nanometer Caches (90nm/65nm/45nm/32nm).In *ISPLED*, Oct 2006.

[4] Semiconductor Industries Association, "International Technology Roadmap for Semiconductors," 2005, http://www.itrs.net/.

[5] Shyamkumar Thoziyoor, Naveen Muralimanohar, Jung Ho Ahn, and Norman P. Jouppi HP Laboratories, Palo Alto. HPL-2008-20,CACTI 5 technical report, April 2, 2008

[6] Jaume Abella*, Antonio Gonzلlez*+, Power Efficient Data Cache Designs, Proceedings of the 21st International Conference on Computer Design (ICCD'03) 2003 IEEE.