CMP L2 NUCA Cache Power Consumption Reduction Technique

P. Foglia, C.A. Prete, M. Solinas
University of Pisa
Dept. of Information Engineering
via Diotisalvi, 2 56100 Pisa, Italy
{foglia, prete, marco.solinas}@iet.unipi.it

F. Panicucci
IMT Lucca
Institute for Advanced Studies
piazza San Ponziano, 6 55100 Lucca, Italy
f.panicucci@imtlucca.it

Abstract

We analyze how applications use banks in a large shared CMP L2 D-NUCA cache depending on their locality and we define a power consumption model. Then we develop a mechanism to dynamically turn on and off a bankcluster in order to reduce the energy consumption.

1. CMP Way Adaptable

Our system is a large shared L2 D-NUCA cache in CMP environment (Figure 1). In this architecture each bank is accessible apart and the data can migrate within the cache memory to approach toward the CPU which uses it.

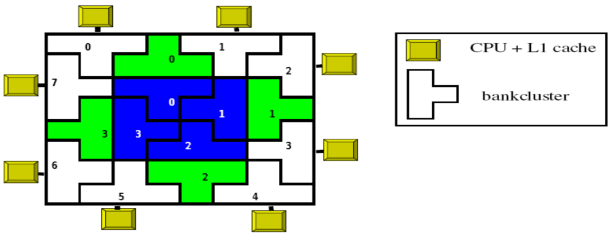

Figure 1. Cache Architecture

To perform our analysis we first developed a power model to evaluate the total energy cost of this system, in order to estimate both the static (leakage) and dynamic component of the energy consumption (Figure 2) and we observed that the leakage is the dominant source of power consumption. Moreover we analyzed how D-NUCA banks and bankclusters are accessed (Figure 3). We noticed that often the most part of the hits occur in bankcluster that are local to the CPUs in which the applications run, whereas some of the other bankcluster present a fewer number of accesses.

Since our objective is to reduce the energy consumption of our system, we plan to develop a mechanism to dynamically turn on and off a bankcluster basing on the frequency of access. We count the number of accesses to each bankcluster within a run interval and we compare this value with the referential value we have before estimated. Then we decide if there is unused memory and it is possible to turn off a bankcluster or if the system needs more cache memory and we have to turn on a bankcluster. This could be considered an extension of the way adaptable technique we presented in [1] for monoprocessor systems.

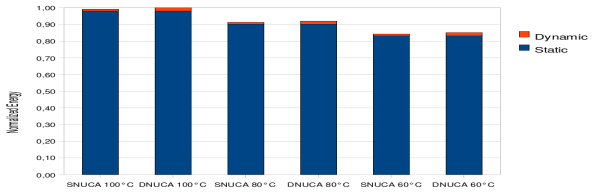

Figure 2. Total Nornalized Energy

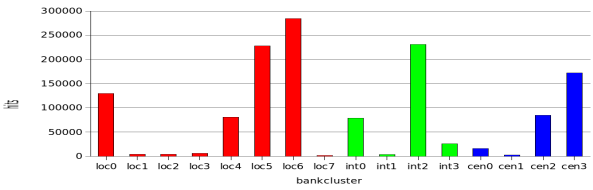

Figure 3. Bankcluster Hits in One Run

By adopting this mechanism we aim to reduce the energy consumption because we use only the cache memory the application needs. Moreover we aim to decrease both network traffic for data search and delay access to bankcluster because there are less banks to visit during the data search and so there are less packets going through the network.

References

[1] A. Bardine, P. Foglia, G. Gabrielli, C. A. Prete, and P. Stenström. Improving power efficiency of d-nuca caches. *SIGARCH Comput. Archit. News*, 35(4):53–58, 2007.