ENERGY EFFICIENCY IMPROVEMENT FOR 3-D CHIP MULTIPROCESSORS WITH NUCA ARCHITECTURE

D.Priya, J.Arunarasi, S.Leo Pauline, J.Jenisha

Address for Correspondence
Department of ECE, Vel Tech Multitech, Avadi, Chennai-54, TamilNadu, India;

ABSTRACT
The (CMPs) chip multi processors is considered by uniform cache access (UCA) architecture. As cache become larger and also partitioned into multiple banks. Increasing wire delay makes it difficult to provide uniform access latencies to all L2 cache banks. In order to reduce the cache memories and wire delay the NUCA architecture has been implemented. By monitoring the NUCA architecture, we can allow nearer cache banks to have lower access latencies than further banks, which analogously increases the memory and time consumption based on both SRAM and MRAM memory and also power consumption criteria. NUCA architecture was initially proposed for uniprocessor system (1 large L2 cache). In this project, Xilinx software was simulated and results were obtained for S-NUCA-1 vs S-NUCA-2 and D-NUCA. And comparison results shows between S-NUCA-1 and D-NUCA architecture.

KEYWORDS: Chip Multi-processors, SRAM-MRAM memory, L2 cache.

1. INTRODUCTION
Non-volatile random access memory such as Static RAM (SRAM) and Magnetic RAM (MRAM). Static Random Access Memory offers power dissipation is very small, small leakage current and operate at lower power supply voltage. Static RAM stores the data can be retained as long as the power supply is ON without any refreshing operation. Magnetic RAM offers higher density and it is stored in the magnetic storage elements, few nanosecond, low power consumption and unlimited write cycle, high speed for read and write without using power supply. Memory technologies such as SRAM and MRAM arranged in the silicon vias enables together onto chip-multiprocessors (CMPs). 3D integration and parallelism of CMP, it has high performance and power efficient memory. The proposed method combines dynamic cache management to remove the high temperature such as resource re-located, power gating and data migration. This paper describes NUCA(Non-Uniform Cache Access) architecture is used in multiple processor, where the memory right to use the time which depends on the memory location. It solves the problem of SRAM-MRAM (Hybrid memory) in wire delay. These design combines a network in the cache memory, it allows data migrate, grouping the cache region in the processor. Figure 1 shows types of cache architecture organization. There are five types of cache architecture are UCA, ML-UCA, S-NUCA-1,S-NUCA-2 and D-NUCA. UCA (Uniform cache architecture) shows in figure 1(a). This model implies a poor perform in the cache architecture in the interior wire delay and less number of banks. ML-UCA (Multi-level Uniform cache architecture) shows in figure 1(b). ML-UCA has L2 and L3 cache. It indicates two duplicate cache and consumes more space. S-NUCA-1 (Static- Non Uniform cache architecture- 1) shows in figure 1(c). The bank has two-way pipelined transmission channel. First pipeline is address channel and second pipeline is data channel and it maps the data in banks based on the block. S-NUCA-2(Static-Non Uniform cache architecture- 2) shows in figure 1(d). In previous architecture does not allow the number of banks. The address channel adds a more wires on the banks to avoid a number of banks. It design with two-dimensional, switched design network and has large number of banks. D-NUCA (Dynamic-Non Uniform cache architecture) shows in figure 1(e). The data is mapped from one bank to other bank to migrate on it. It uses a switched network. There are three types of D-NUCA methods are:
(1)Mapping
   - Simple mapping
   - Fair mapping
   - Shared mapping
(2)Searching
(3)Movement

2. EXISTING SYSTEM
2.1 UNIFORM CACHE ACCESS (UCA)
By sub-divided the large current into various sub-banks, the access time can be minimized to a greater extent. In the Figure 2 it is shown that the route of the signals drives from predecoders to the local decoders in the sub-banks. The information can enter into the sub-banks and then returned to the output drivers after going through muxes. The necessary line is joined together at muxes and passed to the cache regulator. In this wire-delay dominated process, large caches of sub-banking group work poorly. It is basically due to immense delay in receiving the portion of a line from the slowest of the sub-banks.

Figure 1: Types of cache architecture organization

Figure 2: S-NUCA-1 Architecture
The considerably high unloaded access latencies in the UCA cache create a serious problem. Since the increase in area will enhance the loaded access times considerably, the multiport cells are the poor way of solving the problem of over lapping access in large caches. For larger caches, the most apt overall cache size is 2MB. The increase in latency can’t be controlled by the continuous reduction in L2 misses. Since the UCA organization is not suitable for large,
wire-dominated caches, it serves as a standard for measuring the performance enhancement of more sophisticated cache group.

2.2 STATIC NUCA IMPLEMENTATIONS
A lot of functioning is lost due to the requirement of worst-case uniform access in a wire delay dominated cache. If each bank can be accessed at different speeds in relation to the distance of the bank from the cache controller, the multiple banks can lessen those losses. Each bank is addressable independently in the case of banked cache models and it can be sized and divided into a locally favorable physical sub-bank organization. Earlier, after a thorough examination of the design space, the number and physical organization of banks and subbanks were selected to maximize overall IPC. With the help of low-order bits of the index determining the bank, the information is statically mapped into banks. Each bank we simulate is four-way set associative. As described elsewhere, these static, nonuniform cache systems (S-NUCA) have two advantages than the UCA organization. Firstly, accesses to banks closer to the cache controller gain only lower latency. Second, the contention may reduce if the accesses to different banks may proceed in parallel. Since the mappings of data to banks are static, we call these caches as SNUCA caches. Further the banks have non-uniform access times.

2.2(A) S-NUCA-1(PRIVATE CHANNELS)
Each addressable bank in the S-NUCA-1 group as shown in figure 2 has two private, per-bank 128-bit channels of which one going in each direction. This type of plan is used in the IBM power4 level-2 cache. Each bank can be accessed independently as all the banks have private channels. The smaller banks would facilitate more concurrency and a greater reliability of non-uniform access whereas the several per-bank channels add area Tag Array Data Bus Addr ss Bus Bank k Subbank Prede coder Sense amplifier Wordline driver and decoder overhead to the array that restricts the number of banks. The cross-cache routing at 16MB delays required by the cache causes the hit latencies to overwhelm the performance benefit of the reduced misses. The optimal number of banks does not increase further after 4MB owing to the area excess of each-bank channels the cache size increases and thus makes each bank larger and slower. Due to this the S-NUCA-1 organization has been prevented from exploiting the potential access fidelity of small, fast banks.

3. PROPOSED SYSTEM
3.1 S-NUCA-2(SWITCHED CHANNELS)
The assemblage as illustrated in figure 3 removes most of the wires ensuing from per-bank channels. This sort of assemblage is known as called S-NUCA-2 which embeds a lightweight, wormhole routed 2-D mesh with point-to-point links in the cache by placing simple switches at each bank. Each link has There are two dedicated 128-bit channels for bidirectional routing. For our performance evaluation, the delays from a detailed circuit simulator in our cycle-accurate model of the switched network have been used. It has been illustrated under this caption the process of utilizing the future cache access inconsistency by inserting frequently accessed data in closer banks and the lesser significant cached-data in far away banks. It has to be evaluate that a number of hardware policies that transfer the data among the banks.

![Figure 3: S-NUCA-2 Architecture](image)

There are three important questions about the management of data in the cache has to be answered for these strategies. They are: (1) mapping: the process of mapping the data to the banks, and also the detail of the banks in which a datum can exist in (2) search: the system of finding the set of possible locations in order to find a line, (3) movement: the probable conditions under which the data should be transferred from one bank to another.

3.2.1 MAPPING
There are many banks that can facilitate considerable flexibility for mapping lines to banks. One of the farthest choices is the S-NUCA strategy. With the support of it the mapping a line of data to a single statically determined bank can be done. The placement flexibility can be made maximum in the latter approach, whereas the overhead of locating the line may be too large. The midway strategy known spread sets is explored wherein the multi-banked cache is treated as a set-associative structure. In this scheme each set is spread across multiple banks, and each bank holds one ways of the set. The bank set is actually stand for the collection of banks used to implement this associatively. The number of banks in the set is corresponding to the associatively. The three methods of allocating banks to bank sets and ways, called simple mapping has been appraised in this article.

![Figure 4: Simple mapping](image)

The problems of the simple mapping policy at the cost of additional intricacy have been reasonably dealt in the fair mapping policy. By means of arrows and shading the mapping of sets to the physical banks has been shown in the diagram.

![Figure 5: Fair mapping](image)
3.2.2 SEARCH
It is necessary to have certain devices to locate a data block. One of the methods to explore the block is the Incremental manner, i.e., one bank after another bank. The other way is the multicast search operation, in which the request can be sent to all banks simultaneously. Although the latter approach will yield higher performance, it also required higher power. There is one more scheme known as Smart Search mechanism. In the method the partial tag (six bits) for each block is stored at the cache controller. By browsing this partial tag structure we can identify a small subset of banks that probable have the requested data and search can be made only in those banks.

4. RESULTS AND DISCUSSION
The figure 7(a) and 7(b) shows the experimental results of S-NUCA-2 architecture. The input values are given as in1, in2, a, b, e, a2, b2, e2, a4, b4, e4, a3, b3, e3 and output values is obtained out1, v01, v_out1, out2, v02, v_out2, c, c2, c4, c3 and finally output screenshot is shown below.

The table 1 shows the parameters of no. slice used in the core. No. of slices FF, No. of 4 input LUTs, No. of bonded IOBs, No. of GCLKs, minimum period, Min input arrival time before clock, Max output required time after clock between the S-NUCA-2 vs D-NUCA.

5. CONCLUSION
The performance of proposed approach was studied based on the area, power and timing details. A NUCA architecture was designed by using dynamic and Static were simulated and studied in detail. The simulation results showed that the proposed approach performed on both the temperature and energy.

REFERENCES


