NUCA Caches: Analysis of Performance Sensitivity to NoC Parameters

Alessandro Bardine*[1], Manuel Comparetti*[1], Pierfrancesco Foglia*[1], Giacomo Gabrielli*[1], Cosimo Antonio Prete*[1]

* *Dipartimento di Ingegneria dell'Informazione, Università di Pisa*
via Diotisalvi 2, 56122 Pisa, Italy

ABSTRACT

The on-chip network (NoC) is a fundamental component of Non Uniform Cache Architectures and may significantly affect the performance of the overall system. The analysis described in this work evaluates the performance sensitivity of a single processor system adopting a NUCA L2 cache with respect to some NoC parameters, namely the hop latency and the buffering capacity of routers. The results show that the performance sensitivity to the hop latency is very high, thus suggesting that multi-cycle router architectures are not adequate, while the network traffic imposes a moderate load on the buffering resources.

KEYWORDS: Cache memories; Non Uniform Cache Architectures; on-chip networks

1 Introduction

Non Uniform Cache Architectures (NUCA) are a new design paradigm for large last-level on-chip caches and have been introduced to deliver low access latencies in wire-delay dominated environments. Their structure is partitioned into sub-banks and the resulting access latency is a function of the physical position of the requested data. Typically, NUCA caches make use of a switched network to connect the different sub-banks and the cache controller.

While on-chip networks [BDM02, DT01] have been adopted as communication infrastructures in other scenarios, NUCA caches represent an emerging technology and the influence of the network parameters on their performance needs to be investigated. This work analyzes how different parameters for the on-chip network, namely hop latency and buffering capacity of routers, may affect the overall performance of NUCA-based systems for the single processor case, assuming a reference NUCA organization proposed in literature [KBK02, FMP05].

This analysis leads to some important design guidelines: first, the sensitivity of the system to the hop latency is very high and multi-cycle router architectures [PD01, A+07] are not adequate; second, limited buffering capacity is sufficient to achieve a good performance level.

[1]E-mail: {alessandro.bardine,manuel.comparetti,foglia,giacomo.gabrielli,prete}@iet.unipi.it

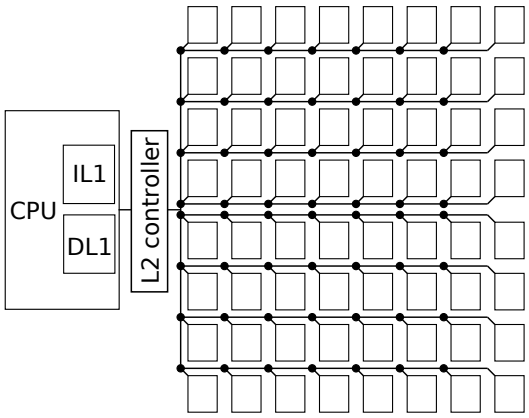

Figure 1: Partial 2D mesh topology. The NUCA structure represented here is made up of 64 banks (8x8). The black circles depict the network routers.

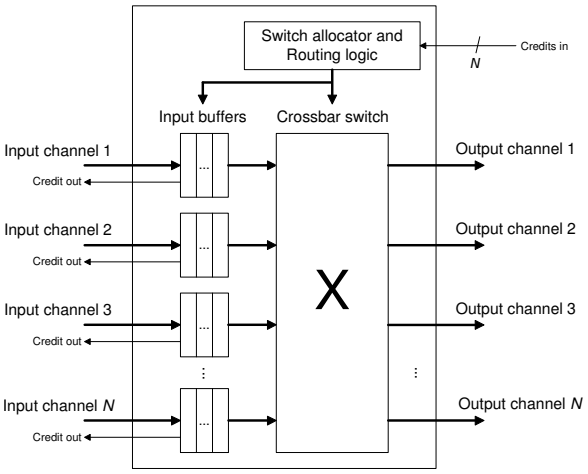

Figure 2: Reference architecture for network routers.

2 Memory hierarchy architectures

This study focuses on a single processor system employing an 8 Mbytes L2 cache. Two different L2 cache architectures, Static NUCA and Dynamic NUCA [KBK02], have been selected, and their performances are compared against a traditional UCA (Uniform Cache Architecture) scheme. The experiments that were conducted led to the following optimal configurations: 4-way set associative UCA; D-NUCA with 64 banks, 8x8, globally behaving like an 8-way set associative cache; S-NUCA with 32 banks, 8x4, with each bank being 4-way set associative.

This analysis assumes a reference NUCA structure which has been derived from previous works [KBK02, FMP05]. The topology of the on-chip network is derived from a 2D mesh, which will be called *partial 2D mesh*, since only a subset of the links of a full 2D mesh are employed in order to reduce the area overhead (Figure 1). Different topology schemes of direct networks, such as toroids, have not been considered for this study because 2D meshes map more effectively onto 2D silicon substrates.

The NUCA on-chip network is based on a wormhole scheme, with routing and flow control policies working on a per-flit basis. The size of a flit is assumed to be equal to the link width. The routing scheme is deterministic, X-Y dimension ordered; a flit is first propagated along the vertical dimension (vertical links in Figure 1), then it is propagated along the horizontal dimension (horizontal links in Figure 1). For D-NUCA caches, since every *bank set* [KBK02] is mapped to a single row of banks, first a flit has to reach the pertaining bank set, then it is propagated to the nodes attached to the banks of its bank set, starting from the nearest one to the cache controller.

The network routers are assumed to be input buffered, and the buffers are managed on a per-flit basis in a FIFO manner. The flow control is credit-based, so each router must keep track of the status of the input queues of its neighbours. This is accomplished by adding two extra credit signals to the link width. Figure 2 represents the reference architecture for the network routers.

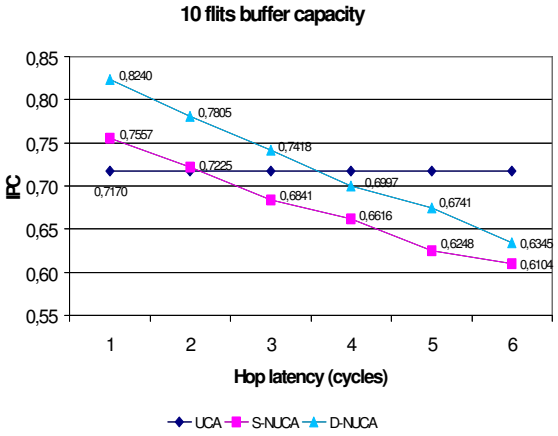

Figure 3: IPC vs. hop latency.

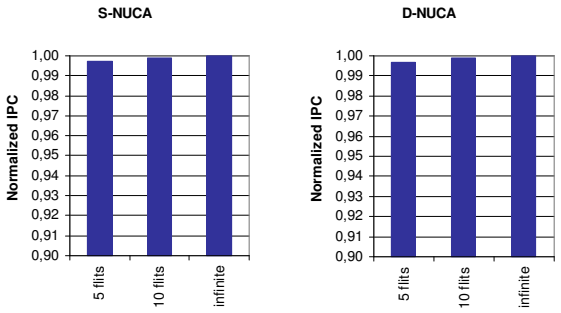

Figure 4: Performance degradation due to limited buffering capacity for S-NUCA and D-NUCA respectively.

3 Methodology

The evaluation of the different architectures have been performed with the execution-driven simulator *sim-alpha*, which was extended to accurately model the NUCA on-chip network. We selected a set of applications from the SPEC CPU2000 and NAS Parallel Benchmarks suites, and for each benchmark we simulated a representative phase of the entire execution.

We derived the timing parameters for the access time and cycle time of cache banks and for the transmissions on the network links with the CACTI 5.1 tool [TMJ07], assuming a 65nm technology node and a 16 FO4 (fanout of four) clock cycle time. The links were modeled assuming delay-optimal repeated, semi-global wires.

The level of detail of the NUCA architecture model which was developed to perform this analysis allows to specify different parameters for the on-chip network, including the latency of transmissions on links (differentiated between vertical and horizontal links, since the aspect ratio of NUCA cache banks may differ from unity), and the buffer capacity and latency of routers. The behavior of the modeled network routers reflects the characteristics of their reference architecture described in the previous section.

4 Results

Figure 3 shows the average IPC (Instructions Per Cycle) for the entire workload as the hop latency varies from 1 to 6 cycles for 10 flits buffering capacity. We can highlight that the overall system performance for NUCA is highly sensitive to the hop latency. While D-NUCA always outperforms S-NUCA, the performance of NUCA-based architectures rapidly decreases from a simulation node to the next. For 3 cycles hop latency, S-NUCA is less performing than UCA, while the benefits of employing a D-NUCA are poor (only 3.5% improvement over UCA). This high sensitivity witnesses that the delay introduced by the on-chip network has strong effects on the overall system performance, while the latency of bank accesses becomes less influential as we move towards higher latencies for hops.

Focusing on a single value for hop latency, e.g. 2 cycles, it is possible to quantitatively evaluate the performance degradation due to limited buffering capacity with respect to the ideal router case (infinite buffering capacity), for both S-NUCA and D-NUCA. Figure 4 high-

lights this degradation, reporting the normalized IPC with respect to the ideal router case with infinite buffering capacity. The resulting performance degradation is negligible even for the 5 flits buffering capacity; for both S-NUCA and D-NUCA the degradation is less than 0.1%.

These results show that the latency introduced by the on-chip network has a very high impact on the performance of NUCA caches, thus introducing strong constraints on the design of the network routers: throughput-oriented architectures are not adequate for NUCA and latency should be the main design goal. These results also suggest that limited buffering capabilities do not jeopardize the performance improvements introduced by NUCA structures.

References

[A+07] Mani Azimi et al. Integration challenges and tradeoffs for tera-scale architectures. *Intel Technology Journal*, 11(3):173–184, 2007.

[BDM02] Luca Benini and Giovanni De Micheli. Networks on chips: a new SoC paradigm. *IEEE Computer*, 35(1):70–78, 2002.

[DT01] William J. Dally and B. Towles. Route packets, not wires: on-chip interconnection networks. In *Proceedings of the 38th Conference on Design Automation (DAC)*, pages 684–689, 2001.

[FMP05] Pierfrancesco Foglia, Daniele Mangano, and Cosimo Antonio Prete. A NUCA model for embedded systems cache design. In *Proceedings of the 3rd Workshop on Embedded Systems for Real-Time Multimedia (ESTIMedia)*, pages 41–46, 2005.

[KBK02] Changkyu Kim, Doug Burger, and Stephen W. Keckler. An adaptive, non-uniform cache structure for wire-delay dominated on-chip caches. In *Proceedings of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 211–222, 2002.

[PD01] Li-Shiuan Peh and William J. Dally. A delay model and speculative architecture for pipelined routers. In *Proceedings of the 7th International Symposium on High-Performance Computer Architecture (HPCA)*, pages 255–266, 2001.

[TMJ07] Shyamkumar Thoziyoor, Naveen Muralimanohar, and Norman P. Jouppi. CACTI 5.0. Technical report, HP, 2007.