On-Chip Networks: Impact on the Performance of NUCA Caches*

Alessandro Bardine, Manuel Comparetti, Pierfrancesco Foglia, Giacomo Gabrielli, Cosimo Antonio Prete †
Dipartimento di Ingegneria dell'Informazione, Università di Pisa
Via Diotisalvi 2, 56122 Pisa, Italy
{alessandro.bardine, manuel.comparetti, foglia, giacomo.gabrielli, prete}@iet.unipi.it

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 [2, 3] 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 [5, 4]. This analysis shows that the sensitivity of the system to the hop latency is very high, thus suggesting that multi-cycle router architectures [6, 1] are not adequate; moreover, limited buffering capacity is sufficient to achieve a good performance level.

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 [5], 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 [5, 4]. 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.

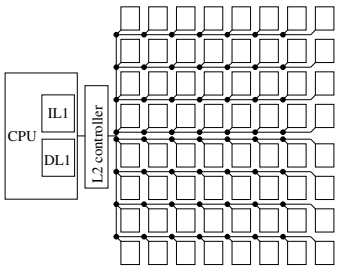

Figure 1. Partial 2D mesh topology. The NUCA structure represented here is made up of 64 banks (8x8). The black circles depict the network routers.

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). 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. In order to guarantee routing fairness, a round-robin scheme is used when multiple transmission requests for the same channel occur.

* This work is partially supported by the SARC project funded by the European Union under contract no. 27648.
† Members of the HiPEAC EU Network of Excellence.

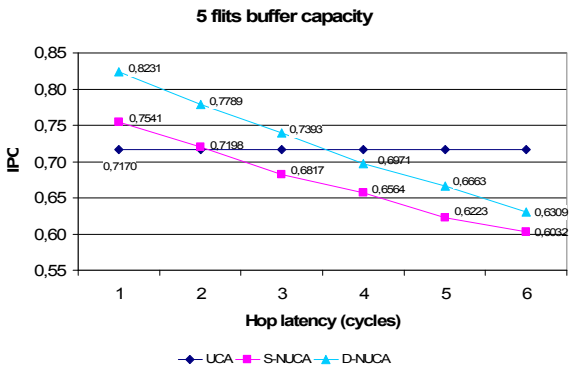

Figure 3. IPC vs. hop latency.

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 [7], 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.

4. Results and future work

Figure 2 shows the average IPC (Instructions Per Cycle) for the entire workload as the hop latency varies from 1 to 6 cycles for 5 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.

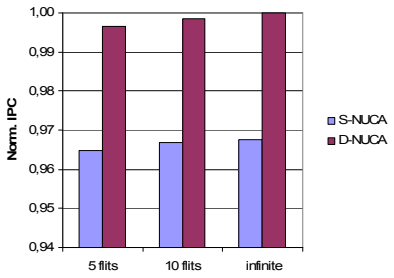

Figure 2. Performance sensitivity to buffering capacity of routers (2 cycles hop latency).

Figure 3 highlights this degradation, reporting the normalized IPC with respect to the ideal router case. The resulting performance degradation is negligible even for 5 flits buffering capacity; for both architectures the degradation is less than 0.1%. These results show that the latency of the on-chip network has a high impact on the performance of NUCA caches and introduces strong constraints on design of routers: multi-cycle architectures are not adequate and latency should be the main design goal. These results also suggest that limited buffering capabilities do not jeopardize the performance improvements of NUCA structures.

Future work will be focused on the extension of this analysis to CMP (Chip Multiprocessor) systems and on the identification of suitable techniques to reduce the NUCA on-chip network latency.

References

[1] M. Azimi et al. Integration challenges and tradeoffs for terascale architectures. *Intel Technology Journal*, 11(3):173–184, 2007.

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

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

[4] P. Foglia, D. Mangano, and C. A. Prete. A NUCA model for embedded systems cache design. In *Proceedings of the 3rd Workshop on Embedded Systems for Real-Time Multimedia*, pages 41–46, 2005.

[5] C. Kim, D. Burger, and S. 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*, pages 211–222, 2002.

[6] L.-S. Peh and W. J. Dally. A delay model and speculative architecture for pipelined routers. In *Proceedings of the 7th International Symposium on High-Performance Computer Architecture*, pages 255–266, 2001.

[7] S. Thoziyoor, N. Muralimanohar, and N. P. Jouppi. CACTI 5.0. Technical report, HP, 2007.