Performance of large-scale reconfigurable optical interconnection networks in DSM systems

I. Artundo[1], W. Heirman[2], C. Debaes[1], J. Dambre[2], J. Van Campenhout[2], H. Thienpont[1]

1: Department of Applied Physics and Photonics, Vrije Universiteit Brussel,
Pleinlaan 2, B-1050 Brussels, Belgium

2: Electronics and Information Systems Department, Universiteit Gent,
Sint Pietersnieuwstraat 41, B-9000 Ghent, Belgium

We present how a custom reconfigurable optical network can be incorporated into Distributed Shared-memory (DSM) multiprocessor machines, and show the potential speed improvement of the interprocessor communication, even when the limits associated to opto-electronics are included. We find that for 32 processors connected in a torus topology, slowly reconfiguring interconnects can provide up to 30% reduction in communication delay. For larger 64-node networks, the expected gain can rise up to 40%. We also introduce the elements for a possible reconfigurable optical network implementation: a selective broadcasting system using focusing-splitting diffractive lenses is described.

Introduction

Communication has always been a limiting factor in making efficient computing architectures with large processor counts [1]. The traffic patterns found on an interprocessor communication network are far from uniform. This makes the load over the different network links vary greatly across individual links, as well as over time, when different applications are executed on a system or one application goes through different phases. Most fixed-topology networks are therefore a suboptimal match for realistic network loads. These problems magnify as the network size increases since even a single-link congestion can quickly spread to slow down the whole network.

One solution to this problem is to employ a reconfigurable network, which has a topology that can be changed at runtime. This way the traffic pattern can constantly be accommodated in the most efficient way (i.e., with the highest performance, the lowest power consumption, or some tradeoff between them). We have previously introduced a generalized architecture in which a fixed base network with regular topology is augmented with reconfigurable extra links that can be placed between arbitrary node pairs [2]. While the network traffic changes, the extra links are repositioned to locations were contention on the base network is most significant.

In this paper we extend our previous work in reconfigurable network evaluation to larger size networks (from 16 to 32 and 64 nodes) and observe the improvement in communication performance as we scale up the shared-memory multiprocessor system. Besides, we give more details on the proposed optical implementation using broadcasting diffractive optics and tunable lasers suggested in [3].

Reconfigurable interconnection network

Our network architecture starts from a base network with fixed topology. In addition, we provide a second overlapped network that can realize a limited number of connections between arbitrary node pairs – these will be referred to as extra links or elinks. Network packets are able to use a combination of base network links and at most one elink on their path from source to destination. We reduce the number of hops for most of the traffic by minimizing a cost function that expresses the total number of network nodes traversed by all bytes being transferred.

The elinks are placed such that most of the traffic has a short path from source to destination, ending up with a correspondingly low (uncongested) latency, and heavy traffic is not spread out over a large number of intermediate links. After each execution time interval of length Δt (the reconfiguration interval), a new optimum topology is computed using the traffic pattern measured in the previous interval, and the elinks are repositioned. The reconfiguration technology must have a switching time much shorter than the reconfiguration interval, which we assume realistically here to be so.

Optical Implementation

The physical implementation of the reconfigurable optical network we envision, the Selective Optical Broadcast (SOB) system, can be realized with a broadcast-and-select scheme, using low-cost tunable laser sources (tunable Vertical Cavity Surface Emitting Lasers – VCSELs – with a range of 40 nm and switching speeds in the order of 100 µs to 10 ms), an optical prism, diffractive microlenses with a 3×3 splitting grating for the broadcasting, and wavelength selective receivers (Resonant Cavity Photodetectors, RCPD) on every node, using single mode/multimode fibers to connect ports, as shown in Fig. 1. By tuning the laser source, the correct destination is addressed.

The prism's base will consist on a 500 µm thick fused silica plate, measuring 10×5 mm, where sets of 10×10 diffractive micro-lenses are arranged in blocks for transmission and reception. The diameter of the microlenses is 150 µm and the initial focusing distance is 335 µm, for a transmission wavelength of 850 nm. At the input, a diffractive grating is added in phase to get a 3×3 beam splitting function. The pitch between lenses is 250 µm. The glass prism is right-angled, with catheti measuring 5 mm. More information about this proposed implementation can be found in [3].

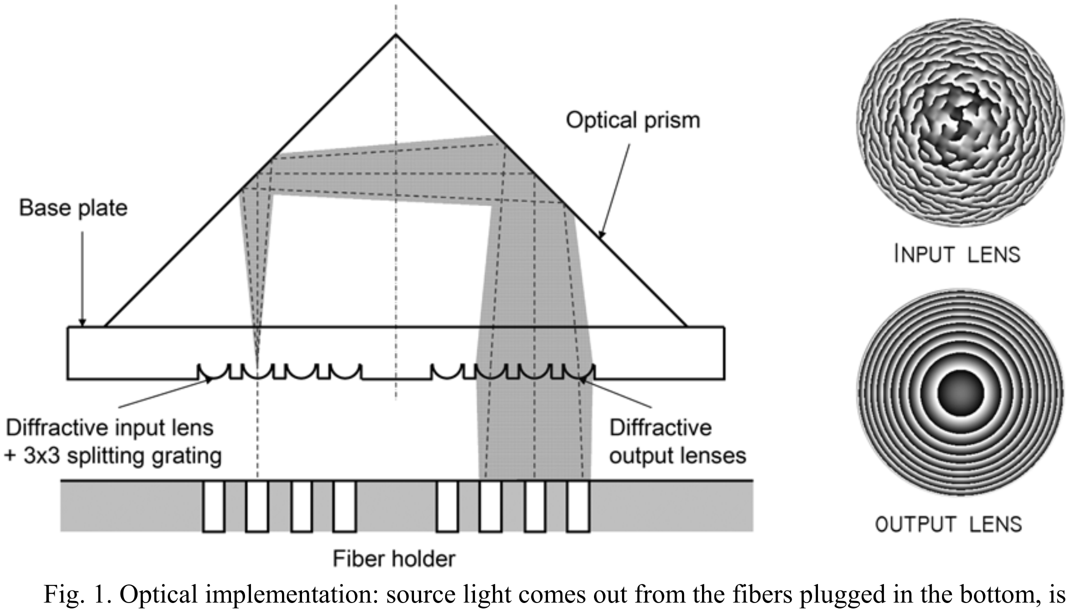

Fig. 1. Optical implementation: source light comes out from the fibers plugged in the bottom, is collimated and splitted in the lens plate into the prism, and reflected to 9 output lenses.

Simulation environment

We have based our simulation platform on the commercially available Simics full-system simulator. It was configured to simulate a multiprocessor machine based on the Sun Fire 6800 server, with 32 or 64 UltraSPARC III processors clocked at 1 GHz. The network models a corresponding a 4×8 or 8×8 torus with traffic contention and cut-through routing. The SPLASH-2 benchmark suite was chosen as the workload. More information about the environment can be found in [2].

A number of extra point-to-point (optical) links, referred as 'elinks', can be added to the torus topology at any point in the simulation, with characteristics equal to those in the base network. Two physical constraints are made on the set of elinks that are active at the same time:

- the maximum number n of elinks that can be active concurrently.
- the fan-out f, meaning the number of elinks that terminate at a single node.

The network used in the Selective Optical Broadcast (SOB) system can be modeled using $f = 1$, n equal to the number of processors, and an additional constraint on which destinations can be reached through an elink from each source node. In our case, only the 8 surrounding neighbors of the destination node through the prism will be available as destinations, limiting significantly the connectivity of the design. This way, the mapping of the source and receiving nodes will be crucial, as it directly determines the possible destinations for every transmitting node through the prism. This mapping results in a placement matrix, which is later on optimized through simulated annealing such that the resulting inter-node distances are minimized.

Results

Here we present the results for larger size networks as compared to similar results presented previously for 16 processor networks (see [3]). Each graph plots the percentage of improvement of the network latency relative to a 'base network'-only implementation, averaged over the execution of all benchmark applications.

Figures 2 and 3 show the effects of faster reconfiguration, while we vary Δt between 1 μs and 10 ms, realistic values for current optical technologies. As a general trend, it is clear that the shorter this time is, the more accurate is the topology adaptation, and the communication improvement will therefore be larger. For the extra links added to the network, in Figure 2 we used $f = 2$ and $n = \#CPUs$, as a fan-out of 2 was shown to be a good balance between performance improvement and complexity/price of the implementation. Figure 3 is for the SOB implementation, which can be expressed as $f = 1$, $n = \#CPUs$ and an extra limitation on which nodes are reachable from each processor with the elinks through the prism.

We have plotted the different benchmarks separately, indicating that the behavior of the program being executed on the multiprocessor machine highly influences the performance improvement obtained. An extreme case for instance is the Radix application running on 64 nodes, which suffers much less performance degradation when moving from an $f = 2$ network to a SOB implementation, whereas the limited connectivity of the second prism-limited network causes a more significant drop in performance for all other benchmarks. However, even in this case, the average latency improvement is around 30% for the 64 nodes case, with a ceiling of 41% in the best of the cases. If we put no restrictions on the fan-out of every node (in this case we used the worst case scenario, 10 ms reconfiguration interval), we get an average memory latency

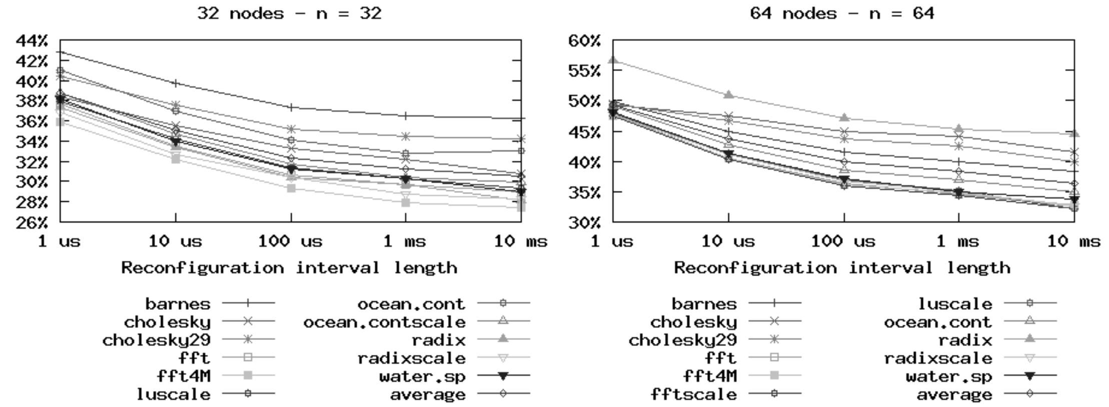

Fig. 2. Performance trends for varying reconfiguration intervals, all with *f = 2* and *n = #CPUs*.

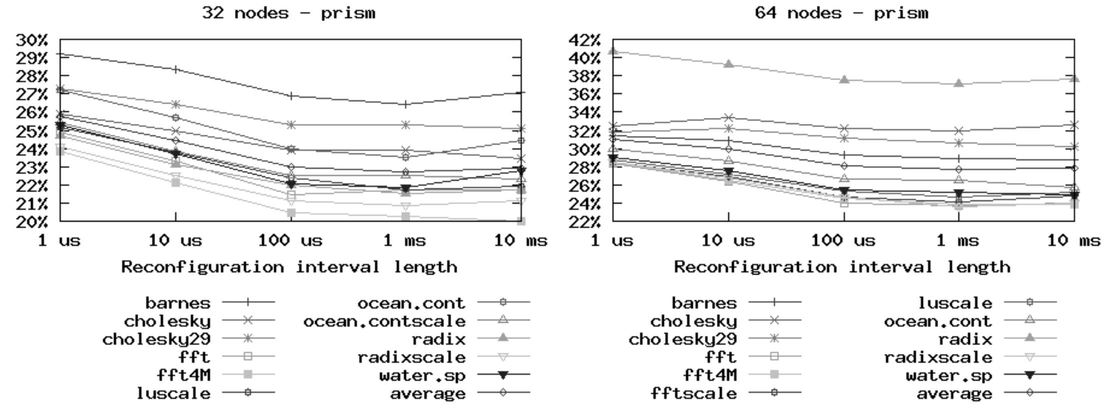

Fig. 3. Performance trends for varying reconfiguration intervals, SOB implementation.

improvement of 44.5% and 49.7% for the 32/64 nodes and 32/64 elinks respectively. This shows that fanout restriction is indeed limiting the performance but not in a drastic way. Doubling the number of elinks raises the performance to 49.1% and 56.1%.

Conclusions

As processor count increases in DSM systems, the interconnection network becomes a performance bottleneck, and by reconfiguring its topology, communication can be improved. We have characterized the speedup for large networks, and found how it depends on the application running on the system, the network size, and the topological constraints of the reconfigurable design, reaching values as high as 56.1% improvement in communication latency. Our proposed optical design has been validated, and further results can be extracted from this model towards its implementation.

References

[1] J. H. Collet et al., "Architectural approach to the role of optics in monoprocessor and multiprocessor machines," Applied Optics, Vol. 39, No. 5, February 2000.

[2] W. Heirman, J. Dambre, I. Artundo, C. Debaes, H. Thienpont, D. Stroobandt, J. Van Campenhout. "Predicting the Performance of Reconfigurable Optical Interconnects in Distributed Shared-Memory Systems." Photonic Network Communications. 2008. *To appear.*

[3] I. Artundo, L. Desmet, W. Heirman, C. Debaes, J. Dambre, J. Van Campenhout and H. Thienpont. "Selective Optical Broadcast Component for Reconfigurable Multiprocessor Interconnects." Journal of Selected Topics in Quantum Electronics, Vol. 12, pp. 828-837, 2006.