Selective Broadcasting for Reconfigurable Optical Interconnects in DSM systems

I. Artundo (1), L. Desmet (1), W. Heirman (2), C. Debaes (1), J. Dambre (2), J. Van Campenhout (2), H. Thienpont (1)

(1) TW-TONA Department, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium (Tel: +32(0)2/629.34.53)
(2) ELIS Department, Ghent Universiteit, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium (Tel: +32(9)2/64.33.66)

Recent developments of coarse WDM optoelectronic components will allow for the fabrication of low cost, high-speed reconfigurable interconnection networks within large distributed shared-memory (DSM) multiprocessor machines, by using the wavelength tunability in combination with a broadcast-and-select mechanism. We discuss in this work the advantages and the limitations they impose on network performance through detailed simulations of benchmark executions. We furthermore propose a passive selective broadcasting element which allows to scale the reconfiguration scheme to larger DSM multiprocessor networks with the proposed coarse WDM devices that have typically low channel counts.

Introduction

Nowadays, metallic connections on printed circuit boards are the standard way to transmit data between processors and memory modules in large-scale multiprocessor machines. They are however running into several physical limitations and their rate of improvement in comparison to the processor's speed is lacking [1]. Although recent developments of conventional galvanic interprocessor communication technologies can still deliver high data throughputs, it is widely recognized that replacement technologies will be required in the very near future.

Optics is a great candidate to introduce agile interconnection networks in the architecture of multiprocessor systems [2]. By the use of optical interconnects at short lengths, we can expect an increase in connectivity and higher communication bandwidths. Moreover, optics allows for the communication channels to be splitted, combined and reconfigured with suitable optoelectronic devices in a data-transparent way.

We investigate in this work how a practical reconfigurable optical network can be incorporated into Distributed Shared Memory (DSM) servers, and determine the potential speed improvements that can be obtained by rearranging the interprocessor connection topology even when the limitations associated to opto-electronics are included.

Optical Broadcasting and Selection

Data and control messages are constantly routed among processing units and memory modules in multiprocessor systems. The use of a bus-like interconnect architecture, which broadcasts all memory requests to all processor-cache modules, makes it relatively easy to implement simple cache coherence schemes. Previous simulation studies on the traffic patterns of realistic interconnection networks [4], and the current technology trends in multiprocessor networking [1], seem to indicate broadcasting to be an appropiate technique for DSM

interconnection communication.

Broadcasting generally is implemented electrically with a bus, over lengths of a few tens of centimeters. For high system clock speeds (more than 500 MHz) and long distances (backplane length 30+ cm), non desirable transmission-line effects appear. Above certain speeds, only point-to-point links can be reliably used for conventional electrical lines. The broadcasting must therefore be implemented as a switched multicast, with data being replicated as it traverses the multiple nodes of the network.

However, by adopting optical communications, we completely disregard these complications. It is indeed relatively easy and efficient to split optical signals even when they carry high data rates. The restrictions now come from power limitations, as a signal splitted among N receivers can provide maximally 1/N or less of the power available to every single receiver, and the non-zero off-state of the laser sources when sharing one channel over more than one emmitter, the so-called optical squelch. It must be suppressed, or wavelength division be used, for not having interferences in the active channel.

Our solution then is using WDM techniques to optically broadcast information in the network by means of a prototyped optical module capable of delivering massive parallelism through free-space point-to-point connections. A more detailed description on this module can be found in [5]. Reconfiguration is achieved by the selection process done at the receivers, which are sensitive to only one wavelength. This selection provides a power-efficient, simple and low-cost solution in comparison with traditional optical broadcasting techniques implemented with star-couplers.

Reconfigurable optical interconnect (ROI) architecture

The proposed interconnection network architecture consists of a fixed (optical or electrical) base network, arranged in a torus topology, to connect all nodes of the distributed system (processors and memories). In addition, a certain number of reconfigurable optical links are provided during runtime, which can be used as direct point-to-point connections between processor node pairs or as a shortcut for the ongoing traffic flow between other nodes in the network.

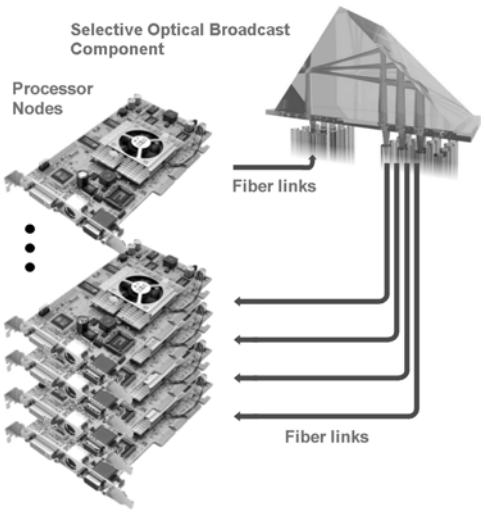

Fig 1. Schematic representation of the broadcast from one processor node in the proposed reconfigurable interconnection design (component sizes are not on scale).

The physical implementation of the ROI network would consist of a single tunable VCSEL per processor board, because of their easy optoelectronic integration, inexpensive batch production and low power consumption. Each processor node also incorporates an optical receiver which is sensitive to one wavelength only (e. g. by using resonant cavity photodetectors). Hence, by tuning the wavelength of each transmitter, the topology of the resulting network can be altered when the optical signals are broadcasted over the network. The target processor is thus selected by emitting at the appropriate wavelength. The low-cost tunable VCSELs will however only be capable of tuning to a limited number of channels (8-10).

The broadcast-and-select scheme is used to provide reconfigurability because of its simplicity and the low price of the components that integrate it, which are the main two obstacles to confront when fabricating a novel

interconnection system. However, a number of limitations imposed by the intermediate optoelectronic devices affect the performance of the system. Firstly, the tuning speed of the transmitter' sources is limited, and since no signal can be transmitted during a topology switch, one needs to allow a minimal time interval between each change in order to obtain a performance gain. We have studied the impact on this tuning time on the actual reconfigurations as shown in the next section. Secondly, the number of wavelength channels of the tunable VCSELs is limited, effectively restricting the number processor nodes to address in the DSM system. To make the solution scalable to larger networks (>10 processor nodes), we have introduced a new broadcast-and-select scheme relying on a Selective Optical Broadcasting (SOB) element. This SOB component will broadcast the transmitted light only to a subset of predetermined processor nodes. Smart partitioning of the target processors allows then for a scalable solution where there is still enough reconfiguration freedom to reduce the hop distance between most transmitter-receivers pairs.

Architecture simulation

To evaluate the performance gain by using our architecture in a real DSM multiprocessor server, we have performed several computer simulations, establishing a full-system simulation environment based on the *Simics* simulator. The simulator was configured to model a multiprocessor machine with 16 *UltraSPARC III* processors at 1 GHz running the *Solaris 9* operating system. The interconnection network is a custom extension modelling a 4x4 torus network with contention and cut-through routing. A more detailed description of the environment can be found in [3].

The simulated time was partitioned in discrete reconfiguration intervals such that topology changes take place at fixed times. We have performed several simulations with interval lengths of 100us, 1ms and 10ms. This interval should be sufficiently short to keep pace with the changing demands made by the applications, but at the same time long enough to amortize on the cost of reconfiguration, during which the extra links are unusable.

The first simulated architecture was one in which 16 extra links can be added to the network without any connectivity limitations. As a result, latency was greatly reduced for a large percentage of the traffic, and the base network was relieved so that less congestion occurs. Using SPLASH-2, a standard parallel computing benchmark, the reconfiguration of the extra channels leads to an average of 23% speed up in execution performance. This is a theoretical solution only, as each node would require potentially as many receivers as the number of nodes in the machine[1]. It serves however as an upper limit of the reconfiguration improvement at the different reconfiguration intervals

Next, we imposed the physical limit on the system such that at most one (or two) reconfigurable links can start in each node. This can be implemented only using one (or two) transceivers in each node. For now, the tunable VCSELs have no limits on the channel count and thus each transceivers is broadcasting on its proper wavelength to all other nodes. This scheme can be implemented using conventional broadcasting devices such as passive optical star couplers. The obtained speed ups are now lower as reconfiguration can not fully accommodate some traffic patterns by having all extra links terminate at the single aggregating node, decreasing actually the efficiency of the extra links. It results in a performance gain of 8.3% or 12.3% when respectively one or two reconfigurable links per node.

1 When all heavy traffic is directed to a single node for example.

Finally we used a SOB element that only allows each link to connect to a subset of 9 different destinations. Although our solution is easily scalable, it furthermore reduces the number of node pairs that can be connected by the extra links. The simulations predict an average performance speed-up of 7.9% or 8.6% depending on the use of one or two parallel links in each node. Still, on average, our simulations showed that with the SOB, as much as 82% of the predicted speedup was maintained compared to the case with the unlimited channel counts and conventional passive star-couplers.

For completion, we explored the influence of the reconfiguration interval, observing that a reconfiguration interval of 10ms is fast enough to follow most traffic bursts. Hence, we believe that it is possible to use the proposed reconfiguration scheme within the timescales of current photonic technology.

Conclusions

We can conclude that it is possible to build a DSM machine with optical reconfigurable interconnects by use of a low-cost dedicated Selective Optical Broadcast (SOB) element. Using extensive full-system simulations of a 16 processors server, we predict an average improvement in performance of 8% for different configurations. A study on the effect of the reconfiguration times shows that the proposed scheme is already implementable with current tunable optoelectronic devices. In future parallel computer designs, with a higher number of processors and with multicore processors within every board, we can expect the obtained performance gain to be more pronounced.

References

[1] A. F. Benner, M. Ignatowski, J. A. Kash, D. M. Kuchta, M. B. Ritter "Exploitation of optical interconnects in future server architectures" IBM Journal of research and development, Volume 49, Number 4/5, Page 755, 2005

[2] 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.

[3] W. Heirman, J. Dambre, D. Stroobandt, C. Debaes, H. Thienpont, J. Van Campenhout, "Prediction model for evaluation of reconfigurable interconnects in Distributed Shared-Memory systems," *Proceedings of the 2005 International Workshop on System Level Interconnect Prediction*, San Francisco, USA, April 2005.

[4] C. Debaes, I. Artundo, W. Heirman, J. Dambre, K. Bui Viet, H. Thienpont, "Architectural study of the opportunities for reconfigurable optical interconnects in distributed shared memory systems". Proceedings Symposium pp. 275-278. IEEE/LEOS Benelux Chapter 2004.

[5] W. Heirman, I. Artundo, D. Carvajal, L. Desmet, J. Dambre, C. Debaes, H. Thienpont, J. Van Campenhout. "Wavelength Tuneable Reconfigurable Optical Interconnection Network for Shared-Memory Machines". Proceedings of the 31st European Conference on Optical Communication. The Institution of Electrical Engineers. Vol. 3. 2005. pp. 527-528, Glasgow, Scotland, Sept 2005.