Selective optical broadcasting in reconfigurable multiprocessor interconnects

Iñigo Artundo[*a], Lieven Desmet[a], Wim Heirman[b], Christof Debaes[a], Joni Dambre[b], Jan Van Campenhout[b], Hugo Thienpont[a]

[a]Dept. of Applied Physics and Photonics, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium

[b]Electronics and Information Systems Dept., Ghent University, Sint Pietersnieuwstraat 41, B-9000 Ghent, Belgium

ABSTRACT

Nowadays, multiprocessor systems are reaching their limits due to the large interconnection bottleneck between chips, but recent advances in the development of optical interconnect technologies can allow the use of low cost, scalable and reconfigurable networks to resolve the problem. In this paper, we make an initial evaluation of the performance gain on general network reconfigurability. In a next stage, we propose an optical system concept and describe a passive optical broadcasting component to be used as the key element in a broadcast-and-select reconfigurable network. We also discuss the available opto-electronic components and the restrictions they impose on network performance. Through detailed simulations of benchmark executions, we show that the proposed system architecture can provide a significant speedup for shared-memory machines, even when taking into account the limitations imposed by the opto-electronics and the presented optical broadcast component.

Keywords: Broadcasting, optical interconnections, reconfigurable architectures, shared-memory systems.

1. INTRODUCTION

Currently, metallic connections on printed circuit boards are the standard way to interchange data between processors and memory modules in large-scale multiprocessor machines. These high-speed electrical interconnection networks are running into several physical limitations such as signal attenuation, electromagnetic interference and severe crosstalk [1]. Besides, recent advances in semiconductor technology have set the shared-memory server trend towards multiple cores per die and multiple threads per core [2], driving a huge amount of communication to the interconnection network, and making it one of the most critical elements of the system. In this kind of servers, overall system performance has improved at different rates, leaving behind memory access time (with around 10% performance improvement per year) far from processor performance (more than 40% per year). This leads to a large bottleneck that worries all manufacturers and rises the question about which new technologies and architectures can solve current limitations.

Table 1. Single-Vendor Interconnect Solutions (Source: SGI NUMAlink, whitepaper)

Technology	Vendor	Latency	BW / link
NUMAlink	SGI	1.5-3 μsec	1500 MB/s
QsNet II	Quadrics	1.6 μsec	900 MB/s
ServerNet	HP	3 μsec	125 MB/s
Sun Fire Link	Sun	3-5 μsec	792 MB/s
Myrinet XP2	Myricom	5.5 μsec	495 MB/s

Requirements for the different levels of communication in a multiprocessor system are different depending on the nodes being connected. The links between processors and local memories require the lowest latency and highest performance, while board-to-board links allow for higher latencies but require higher power and bandwidth transmitted, along with card and connector specifications. This way, current point-to-point interconnection technologies include HyperTransport [3], Sun Fireplane [4] and parallel RapidIO [5] for chip-to-chip links, and serial RapidIO and PCI Express [6] for board-to-board links (see table 1).

Reconfigurability at this interconnection level can allow the system to rearrange the interprocessor network creating topologies that are best suited for the particular computing task at hand, allowing for a topology that will closely match the traffic patterns exhibited by the running algorithm. Previous studies have shown that by the use of run-time reconfigurable interconnects, one can speed-up specific communication patterns during execution of any application [7].

Optics is a great candidate to introduce fast interconnection networks in the architecture of multiprocessor systems [8]. Using optical interconnects at the scale of link lengths found in multiprocessor machines (up to a few meters), an increase in connectivity, lower latencies and higher communication bandwidths can be expected, whereas the design of conventional electrical interconnects is limited by the trade-off between interconnection length and bit rate. The high operating frequency of light tends to virtually eliminate any frequency dependent cross-talk, and the inherent voltage isolation will also improve the signal integrity of the high speed communication channels. By using optical interconnects we can furthermore alleviate also the concern of power and thermal budget, and make the system more compact by replacing the bulky cable connectors that are required now with off-the-self electrical interconnects. Finally, a very important aspect of optical interconnects is their inherent ability to switch the light paths easily in a data transparent way, moving towards adaptable network topologies. It is the goal of this work to investigate how a practical reconfigurable optical network can be incorporated into Distributed Shared-Memory (DSM) systems through an optical broadcast-and-select architecture and to assess the resulting performance improvements.

The paper is organized as follows. In section 2, we give a preliminar presentation of the advantages of a reconfigurable network topology for a DSM interconnection system, supported by accurate simulations of a real parallel multiprocessor. Section 3 gives an overview of the current optical technologies that can be used for building a reconfigurable optical interconnect inside multiprocessor machines, and presents an overview of recent reconfigurable demonstrators. In section 4, we propose a low-cost Selective Optical Broadcasting (SOB) element that allows for a flexible reconfiguration method, and finally in section 5 we present the related simulation results with the speed-ups that such a reconfigurable scheme would allow for in a DSM server.

2. RECONFIGURABLE INTERCONNECTIONS

In DSM multiprocessor machines all the memory of the system is physically distributed among its nodes at the hardware level (see Fig. 1). In this type of machines, an interconnection network links all processors together, allowing them to share the single global memory in a transparent way to the developer or user. The coherence between different copies of data on multiple processor caches is maintained by protocols running on the interconnection hardware. The inherent coherency and memory distribution make the programming of such machines relatively easy as opposed to systems relying on a message passing paradigm. This explains the popularity of such systems for mid-range servers. However, the interprocessor network is then very closely integrated with the hardware and is in fact a part of the memory hierarchy. Therefore, it is absolutely necessary to conceive an interconnection network with low latencies which could otherwise cause a significant performance bottleneck in program execution. Communication here is done through data packets being interchanged at very high speed from memory addresses in caches or memory blocks to the CPUs. In addition, some control information like memory requests or invalidations, is usually spread to multiple nodes over the network. This shows that it is advantageous if the interconnection technology is capable of easy data broadcasting while at the same time, allows for heavy data bursts between single node pairs. Broadcasting is typically less complex but requires more bandwidth, whereas point-to-point connections can reduce bandwidth requirements by limiting transmission to only the nodes that are affected by it. Along with bandwidth, end-to-end latency is again extremely important here, since processors must usually wait until coherence-control operations complete before continuing processing.

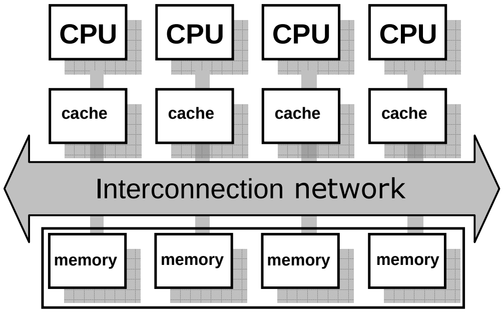


Fig. 1. DSM bus network connecting different processing nodes with the distributed shared memory.

The most common way to interconnect multiple processors with their local memories is to use a single shared bus (a many-to-many configuration) with multiple senders and multiple receivers on a common line. An arbitration mechanism is needed to select the sender that will broadcast packets to multiple receivers over the same physical transmission medium. The shared bus approach has the advantages of simplicity of design and easy scalability, particularly for systems using snooping protocols to maintain cache coherence.

However, the electrical characteristics of a bus limit its useful frequency to the 400 MHz range, and its length to less than tens of cm. As system frequencies move into the GHz range and distances to the backplane length, transmission-line effects such as reflections and matched impedances become critical, and an increasing number of electrical shared buses are being replaced by point-to-point networks that emulate a shared bus behaviour. This eliminates the need for costly arbitration, but relies instead on complex local-state methods for broadcasting data. At the same time, implementing fully interconnected point-to-point networks is unachievable for a large number of nodes. Different topologies arise (such as hyper-mesh, torus, tree, etc...), balancing high connectivity with acceptable technological complexity. These topologies are hard-wired, and as a consequence, the performance can greatly vary for the same architecture depending on the traffic patterns generated by the applications in execution.

Communication between nodes in a parallel system has a dramatic impact on the speedup of the running applications. One can often find certain communication patterns in many parallel algorithms which are regular in time and space (spatial distribution of processes on the nodes), like the one-to-all broadcast, all-to-all broadcast, all-to-one gather, etc. For this reason, a good adaptable reconfiguration scheme would result in a significant performance improvement by dynamically adjusting the network topology to match the specific run-time requirements.

2.1 Reconfigurable point-to-point architecture

The proposed Reconfigurable Optical Interconnect (ROI) network architecture in this paper consists of a fixed electrical or optical base network, arranged in a torus topology. In addition, a certain number of reconfigurable extra optical links are provided (see Fig. 2). These can be used as direct point-to-point connections between processor node pairs or as shortcuts for the ongoing traffic flow between other nodes in the network, as reported in [9]. This setup, compared to the case where all links in the network are available for the reconfiguration, has a number of advantages, mainly because the base network is always available. It is therefore impossible to disconnect parts of the network, greatly reducing complexity in the reconfiguration algorithms. Also, a minimum performance level is guaranteed, since the reconfigurable links may not always result in a shorter path and are even unavailable while the topology is being changed. We will assume equal characteristics for the extra optical links and the base network, yielding this way the same average packet latency for both types of links.

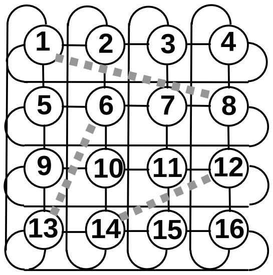


Fig. 2. Torus topology of the base interconnection network with added reconfigurable optical links. The numbers correspond to processor node numbers inside the network.

2.2 Simulation environment

We have established a full-system simulation environment based on the commercially available Simics simulator. A more detailed description of our environment and the workload can be found in [10]. The simulator was configured to mimic a multiprocessor machine based on the Sun Fire 6800 server, with 16 UltraSPARC-III processors at 1 GHz running the Solaris 9 operating system. The interconnection network is a custom extension to Simics, where we modeled a 4x4 torus network with contention and cut-through routing. The SPLASH-2 scientific parallel benchmark suite, as well as the Apache web server along with the SURGE request generator, were chosen as the workload applications for stressing the system under test. There was a certain level of noise (2-5%) on application runtimes, stemming from other scheduled internal tasks of the operating system as well as the initial state of the cache memories. We have partitioned the simulated time in discrete reconfiguration intervals (see Fig. 3) such that the topology changes take place at fixed times. This interval should be sufficiently short to keep pace with the changing demands made by the application, but it must be long enough to amortize the cost of reconfiguration, during which the new links are unusable.

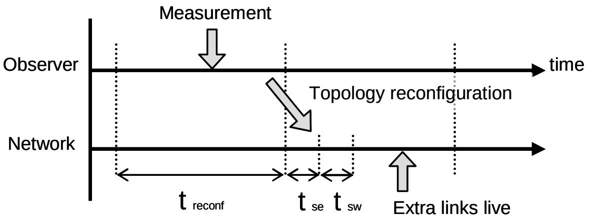


Fig. 3. In every reconfiguration interval, the system is monitoring the traffic flow such that it can adjust the topology to accommodate the

We further remark that this study focuses on switching networks with tuning speeds that are significantly slower than the packet transfer times. Hence, the focus is not on interprocessor networks that will rely on packet level switching. Such networks would require expensive, extremely fast tunable devices. They furthermore need a difficult arbitration system and rely mostly on message queues which increase the transport latency of the packets significantly. In contrast, this proposal makes use of the low-latency wormhole routing that is generally adopted for the mesh networks in such systems. In this design, extra links are assigned every certain time interval, after a network measurement to determine the busiest pairs of nodes, and a selection and switching intervals required to adapt the topology. When a reconfigurable link is in place, it provides a short-cut for packets and lowers their latency. By optimizing the placement of the extra links for the common case on high traffic bursts between pairs of nodes, we are investigating whether reconfigurable networks can provide a significant performance boost, even if slow tunable devices in the millisecond range are adopted.

2.3 Preliminar results on ideal reconfigurability

For a first set of experiments, the simulated architecture is one in which 16 extra links can be added to the network. Here no limits are imposed on which node pairs the extra reconfigurable links are connected to. Therefore, the 16 busiest node pairs for every time interval can be directly connected by extra links, which alleviates heavy traffic loads in them. After adding these links, latency is greatly reduced for a large percentage of the traffic (see Fig. 4), and the base network is relieved so that less congestion occurs. The relative improvement of the computation time of different benchmark applications is shown in the first column of table 2 when a reconfiguration interval (t_{reconf}) of 100 µs is used. For now, we have not considered in the simulations any fixed down-time (t_{se}+t_{sw}) that occurs during network readjustments to keep the performance study independent of the chosen switching technology.

Table 2. 16 Extra links configuration speed-ups

Application	**∞ links/node**	**2 links/node**	**1 link/node**
Apache	7 %	4 %	4 %
Lu	7 %	5.1 %	3.5 %
Cholesky	9 %	7 %	1.2 %
Radix	17 %	7.2 %	6.9 %
Ocean	38 %	23.6 %	21.9 %
FFT	42 %	26.8 %	17.5 %
Radiosity	43 %	12.5 %	3.1 %
AVERAGE	**23.3 %**	**12.3%**	**8.3 %**

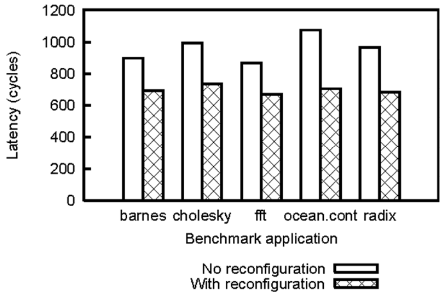


Fig. 4. Average memory access latency improvement (16 links)

The aggregated time of all transactions, point-to-point as well as broadcast, is reduced by a larger amount than the total execution time, showing the reconfiguration saving in communication time (without any associated overhead). This solution is however far from a physical implementation, as it would require an optical transceiver for every link that potentially terminates in each processing node. In the worst case, 15 extra links would end at the same node and therefore each node would need 15 optical transceivers.

We next impose the limit on the system such that at most one or two reconfigurable links can terminate in each node (see last columns of table 2). This can be implemented using a fixed number of transceivers in each node (one or two), together with an ideal star coupler to spread the optical signal. This of course reduces the number of optical devices located in every processor node, but it also means that sometimes, when a node is receiving a burst of traffic from different destinations, reconfiguration can not fully accommodate a star-shaped traffic pattern by having all extra links terminated at the aggregating node. This effect decreases the efficiency of the extra links, as the new channels are not always situated where they would be needed, but resembles more a real interconnection scheme.

3. OVERVIEW OF AVAILABLE OPTICAL TECHNOLOGIES

In the previous section, we have proposed a design for a reconfigurable interconnection and the benefits of an ideally adaptable topology. In a next step, it is necessary to relate this solution to the building blocks and components that are currently available for setting up such an optical interconnection, i.e. light sources and detectors, switches, etc. In the next sections we will give an overview of current technologies being developed and the techniques for optical reconfiguration that can allow the construction of the interconnect in a technically-feasible and cost-effective way.

3.1 Optical switching technologies

For optical interconnection links, Vertical-Cavity Surface-Emitting Lasers (VCSELs) in the 850 nm wavelength are generally chosen, primarily because of their low-cost mass production on wafer-scale, high-reliability, low power consumption, easy array integration and their rotationally symmetrical laser beam that can easily be coupled into optical fiber while offering a small form factor for easy onboard integration. Other lasers, such as Fabry-Perot and distributed feedback lasers, can be installed in linear array configurations but the overall process is generally more expensive because of yield and testing issues. The lightpaths emmitted by the sources can be then redirected in different ways by optical means, leading us to many new opportunities. Optical reconfiguration has recently gained much interest, driving the development of a large variety of physical tuning mechanisms. The three major approaches used for optical reconfiguration are the active tunable opto-electronics, the specialty tunable optical components, and the beam steering micro-optical electro-mechanical systems.

In tunable opto-electronics the characteristics of the emitted light, like the wavelength or polarisation state [11] can be tuned by changing the operating conditions such as the temperature, the current injection or the micro-cavity geometry with micro-electro-mechanical (MEMS) membranes. Some of the more remarkable achievements in this field for optical interconnects are the advent of MEMS-based 2D tunable VCSEL arrays [12], and the arrayed waveguide grating (AWG) -based tunable fiber lasers [13] for the high-end telecom market. Another method for optical switching is to tune the functionality of the optical components themselves instead of tuning the characteristics of the light sources in the network. In this way, tuning has been achieved by acousto-optic Bragg cells [14], with photorefractive crystals [15], by heating liquids in thermal bubble switches [16] or by using micro-fluids as the tuning element in waveguides [17]. Recently, Liquid Crystal (LC) components receive a lot of attention and are being used to make adaptive computer generated holograms [18] and switchable gratings [19]. A third approach for optical switching is the use of active free-space laser beam steering. Here, a MEMS micromirror-based device images a 2D fiber array onto a second one [20]-[22]. These devices are compact, consume low power and can be batch fabricated resulting in low cost. However, optical MEMS crossconnects have yet to overcome design challenges in reliable opto-mechanical packaging, mirror fabrication and electronic control algorithms.

3.2 Optical broadcast-and-select

As already explained before, broadcasting capabilities in a multiprocessor network are critical, but the usual electrical bus implementations are running into severe problems as we increase the performance demands and the number of nodes of the network. However, by optically implementing the bus-like broadcast function we can still retain the architectural advantage, while getting rid of most of the electrical problems. For example, splitting an optical signal to multiple receivers can be done without impedance discontinuities that affect the achievable bit rate. But optical broadcasting still has some limitations, like power splitting, wide bus transmission and multiple-user issues [23].

We will try to address all these concerns with our proposed architecture, considering that broadcasting can also be done by splitting wavelengths instead of beams of light, via wavelength division multiplexing (WDM). This way, bandwidth and connectivity can be greatly increased, exceeding the possibilities of any electrical interconnect, and what is more important, easy reconfiguration can be obtained by adjusting the wavelength emmitted or received by every node (or set of nodes). This will be the solution adopted on this paper and presented more in detail on the next section.

Tunable opto-electronic components can be used to build a passive optical broadcast-and-select scheme where the wavelength is selecting the destination. The broadcast function can be achieved through passive couplers in a guided wave approach [24], or by beam splitting diffractive optical elements (DOEs) in the free-space case [25]. The selection mechanism can be implemented by the use of resonant cavity photodetectors (RCPDs) [26]-[27], tunable optical filters [28], AWGs [29], passive polarisation sensitive DOEs [30]-[31] or passive wavelength sensitive DOEs [32]. In general, these special DOEs generate different fan-out patterns for different wavelength or polarization conditions of the incident light. The most recent examples of wavelength selective devices are microring resonator waveguides [33]-[34], being very good candidates for photonic integration since high compactness and low-cost can be achieved in the silicon on insulator technology.

For the multiplexing technique, one can use here a coarse WDM (CWDM) solution which exhibits a lower number of channels and wider spacing as opposed to dense WDM (DWDM), mostly used in telecom applications and with very tight channel integration. CWDM is designed for short distances, where amplification is not required, and it uses a wide range of frequencies spreading the wavelengths apart and allowing this way to use more flexible and simpler components at the cost of wavelength accuracy. As no cooling is required, CWDM devices are considerably smaller and a lot less expensive than DWDM, allowing the use of 850 nm tunable VCSEL arrays (with tuning ranges around 40 nm) and MM fiber for the transmission layer of the interconnect.

3.3 Overview of reconfigurable optical interconnects demonstrators

There has been lots of proposals for reconfigurable optical architectures on the past, but only a few of them have been accomplished in the form of demonstrators. During the recent years, some of them have achieved remarkable results by implementing the reconfiguration in very different ways. For example we have the OCULAR-II system [35], which is a two-layer pipelined prototype in which the processing elements, with VCSEL outputs and photodetector input arrays, are connected via modular, compactly stacked boards. Between each of the layers there is a free-space optical interconnection system, and by changing the phase pattern displayed on a phase-modulating parallel-aligned Spatial Light Modulator (SLM), the light paths between the nodes can be dynamically altered.

Another reconfigurable architecture constructed was the Optical Highway [36], a free space design which interconnects multiple nodes through a series of relays used to add/drop thousands of channels at a time. The architecture considered here was a network-based distributed memory system (cluster style), with a 670 nm laser diode as a transmitter and a diffractive optical element to produce a fan-out simulating a laser array. Polarising optics defined a fixed network topology, and a polarising beam splitter deflected channels of a specific polarisation to the corresponding node, with each channel's polarisation state determined by patterned half wave plates. It can be made reconfigurable using also a SLM, allowing to switch the beam-path of a single channel from an electronic control signal, and route it to only one of three detectors.

The SELMOS system [37] was designed to be a board-level reconfigurable optical interconnect, whose core was built from a 3D microoptical switching system and a self-organized lightwave network. Here, the reconfiguration process was done with 2x2 Waveguide Prism Deflector micro-optical switches (WPD-MOS) in a 1024x1024 Banyan network. Self-organizative network formation worked by arranging first the optoelectronic devices with waveguides in a designed configuration, stacking them to create a 3D structure, and then introducing some excitation to this structure, creating a self-aligned wiring coupling several waveguides. However, only simulations and partial experiments have been realized and a full working demonstrator is still to be constructed.

Finally, there has been some work also on reconfigurable buses, being the Linear Array with a Reconfigurable Pipelined Bus System (LARPBS) [38] the best example of a complete architecture, although again a complete implementation has not been realised yet. It is a fiber-based optical parallel bus model that uses three folded waveguides, one for message passing and the other two for addressing via the coincident pulse technique. The reconfigurability in this model is provided by pairs of 2x2 bus-partition optical switches located between each processor that can partition the system into two subsystems with the same characteristics at any of these pairs of switches, by introducing some conditional delay.

4. SELECTIVE OPTICAL BROADCASTING ARCHITECTURE

The biggest challenge for the implementation of a ROI is to avoid complicated switching devices or costly opto-electronics. Driven by the progress in metro-access networks, low-cost tunable devices such as MEMS-based tunable VCSELs are becoming readily available. Thus, we believe that the system which relies on coarse wavelength tunability presented in the previous section is a viable approach for introducing reconfiguration at reasonable costs. The ROI enabled part of the inter-processor communication network would have a single tunable VCSEL as optical transmitter per processor board. In this way, each board would be able to transmit data on a fixed number of wavelengths. This signal is then guided through MM or SM fiber to a broadcasting element which divides the data-carrying signal to all (or a selection of) the receiving nodes. Each processor node also incorporates an optical receiver which is sensible to one wavelength only. Hence, by tuning the wavelength of each transmitter, the topology of the resulting network can be altered. At the receiving part of the processor node, a RCPD could be used.

4.1 Design challenges

A number of limitations imposed by the optoelectronic devices will affect the reconfigurability of the proposed system:

- The tuning speed of the transmitter sources is limited. This will set the reconfiguration rate. Since no signal can be transmitted during the switch, one needs to allow a minimal time interval between each topology change in order to obtain a profitable reconfiguration scheme.
- The number of wavelength channels is limited. Given the need for a low-cost device with a reasonably high tuning speed, the included tuning element will exhibit a very limited number of wavelength channels, as a trade-off exists between the tuning speed, the cost and the channel count.

In a previous exploration of the influence of the tuning speed on reconfigurable network performance, we have found the switching speed requirements to be on par with recent MEMS-based tunable VCSEL designs [38]. In section 5.2, we will furthermore address the influence of the reconfiguration time on the proposed system. The restriction on the channel count, however, prohibits the use of the broadcast-and-select scheme in which all processing nodes (say over 64 nodes) are connected together via a single star-coupling element. We therefore propose a selective broadcasting component which broadcasts each channel to only a limited number of outputs, but can scale to larger configurations.

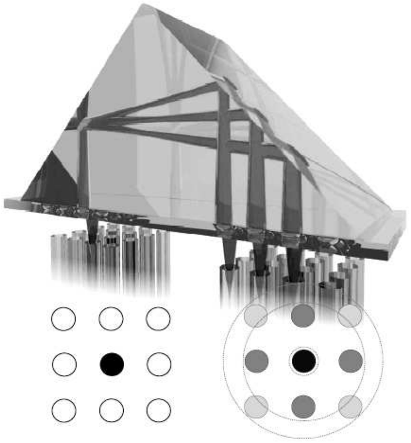


Fig. 5. Schematic representation of the SOB component for the proposed ROI architecture. A processor node can transmit data on a set of 9 wavelengths and the SOB element distributes the signal towards 9 fellow nodes. Since every receiver is sensitive to one wavelength only, the target node is selected by emitting at the appropriate wavelength.

Fig. 5 shows the concept of the proposed network containing such a passive optical broadcast element. The fibers coming from the tunable sources in each processing node are bundled into a fiber array at the ingress of the SOB. The free-space optical component will fan-out the signal to a 3x3 matrix of spots at the output side via a DOE. This way, each processor is capable of connecting only to 9 different nodes. It is important to note that the mapping of the source node connections and the receiving nodes on the broadcasting component is now critical, because it directly determines the possible addressable nodes for every transmitting processor. A scalable ROI design is thus possible using components with low channel counts; we will measure the effect on the restricted connectivity and the performance speed-up in section 5.2.

4.2 Selective Optical Broadcasting (SOB) element

On the previous years, an optical module has been prototyped at our labs. This module is capable of delivering massive parallelism through free-space point-to-point optical interconnects. We have used Deep Proton Writing (DPW) for prototyping this component [40]-[41]. DPW is our in-house technology and consists of the irradiation of polymethyl metacrylate (PMMA) with a pencil-like proton beam followed by a selective etching or swelling of the irradiated zones. Selective etching results in high quality optical surfaces, or micro-hole arrays in case of proton beam point irradiations.

We can also swell the point irradiations, resulting in large arrays of microlenses with dedicated focal numbers. In Fig. 6 we show the prototyped optical interconnect module. It consists of a combination of a commercially available thick glass prism with a plastic baseplate containing micro-lens arrays fabricated by DPW. The component images a dense 2D array of source channels onto the corresponding channel array at the exit side via total internal reflection (TIR) on the prism facets. Microlens arrays are used to collimate and focus the beams through the prism.

An in-line-coupling scheme for the module is also possible, but in combination with our dense 2D fiber arrays fabricated by DPW too, it is more favorable to have the source and exit node channels in the same plane at one side of the optical component. This approach results in a smaller form factor since the necessary space for fiber bending can be kept at one side. For the prism design, we have used a right-angled glass prism with an edge of 5 mm. The prism's base surface of 5x7 mm allows us to place 15x22 channels on both exit and source channel sides in the case of an inter-channel distance of 220 μm. This number of channels is more than enough to wire 64 processor nodes.

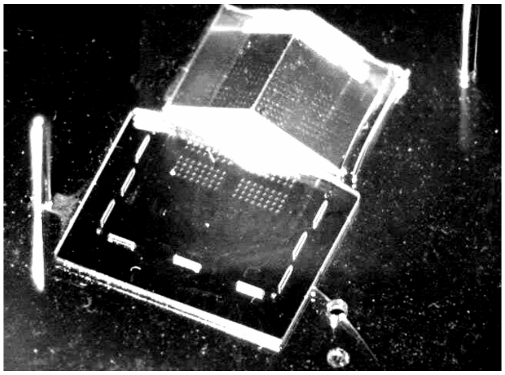
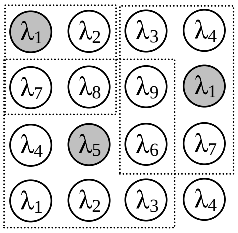


Fig. 6. Picture of the prototyped free-space optical interconnect module. The baseplate contains four arrays of 4x8 microlenses with a diameter of 123 μm and a channel spacing (pitch) of 250 μm. The glass prism has a side facet of 5 mm. The cylindrical holes in the base plate hold mechanical MT-ferrule guiding pins.

A phase-only DOE creates the passive optical beam splitting in the selective optical broadcast component. The DOE broadcasts the light from every source channel towards 9 diffracted spots in a 3x3 pattern configuration. The diffractive grating period Λ is calculated with the well-known grating equation (1) so that the diffracted spots exactly fit on the exit fiber node lenses. In this equation the vacuum wavelength is denoted by λ, the index of refraction of the component material by *n*, the diffraction order by *m*, and the diffraction angle by α. The lateral distance between the diffractive spots in the generated pattern is called the pitch *P*, which is equal to the receiving fiber channel pitch in the overall broadcast component. *P* is connected to the grating diffraction angle α through the optical pathway length (*OPL*) between source channel and exit node channel as given in (2). The OPL of all the component is 8 mm. (7 mm of *OPL* in the micro-prism plus 2 times 500 μm thickness of the microlens baseplate). Exclusive wavelength selection on the receivers is done according to the distribution shown in Fig. 7.



λ_1 λ_2 λ_3 λ_4
λ_7 λ_8 λ_9 λ_1
λ_4 λ_5 λ_6 λ_7
λ_1 λ_2 λ_3 λ_4

Fig. 7. Wavelength distribution and example destinations for 3 nodes.

$$\Lambda \sin\alpha = \frac{m\lambda}{n} \qquad (1)$$

$$\alpha = Bgtg\,\frac{P}{OPL} \qquad (2)$$

Table 3. Grating constant and index of refraction in function of design wavelength

Wavelength λ (nm)	**Grating constant Λ (μm)**	**Index of refraction n (PMMA)***
850	20.82	1.4849
980	24.03	1.4834
1310	32.17	1.4812
1550	38.09	1.4804

* According to the Schott dispersion formula

The system channels are spaced at a distance *P* of 220 μm. We regard this as the highest channel density at which a mechanically stable fiber holder can be fabricated with our technology. With these values of *P* and *OPL*, we show in table 3 the calculated value of Λ for the standard telecommunication wavelengths. It is also readily seen in (1) that Λ scales with the wavelength. Our system design wavelength with 850 nm thus imposes the most stringent demands for the fabrication of the DOE design. Adapting the system to the emitted wavelength of 1550 nm for other MEMS tunable VCSEL sources will relax the minimal size of the basic diffractive cell.

The diffractive kinoform covers the source channel side of the baseplate underneath the micro-prism. The refractive micro-lenses there have been replaced by their diffractive counterparts, thus embedding the optical fan-out functionality. Also, the refractive microlenses at the exit channel side are replaced by diffractive ones but without beam splitting functionality. The source and exit node microlens parameters are equal for symmetry reasons. The microlens diameter is 140 µm and the front focal length is 335 µm. The distance between the broadcast prism and the fiber arrays (called the working distance d_0) is 350 µm. The value of these parameters was determined by maximizing the efficiency of initial point-to-point interconnect simulations and maximizing the ratio of the optical power received within the first lens aperture. Using standard Gaussian beam propagation, the clipping loss at the first lens aperture was only 0.2% using single mode fiber with a numerical aperture (NA) of 0.12 at the input (corresponds to 8° FWHM). The clipping loss increases to 16.5% if multimode fibers with a NA of 0.22 (15° FWHM) would be deployed instead.

We used the Rayleigh-Sommerfeld propagator to propagate the optical field from the entrance source channel towards the receiving fibers. We did not include the reflections of the diffracted beams at the sidewall of the micro-prism into the simulation model, because they cannot be modeled in the wave optical simulator we have been using. With basic geometrical calculations we can show that the TIR condition inside the micro-prism is not always fulfilled for all the diffracted beams. Therefore, in a practical implementation the prism needs a reflective gold coating, reducing somewhat the simulated component efficiencies, presented in Table 4. The simulation sampling size was 500 nm.

Table 4. Optical broadcast component efficiency results by wave optical simulations

Source node		**Kinoforms**	**Phase quantization in 16 phase levels**	**Lateral pixel quantization**	
				1 µm	**2 µm**
SM fiber	*Total optical power in diffracted focal spots*	93.0 %	90.7 %	86.0 %	66.9 %
	Focal spot diameter	5 µm	6 µm	6 µm	6 µm
MM fiber	*Total optical power in diffracted focal spots*	75.4 %	73.9 %	67.9 %	48.4 %
	Focal spot diameter	9 µm	10 µm	10 µm	10 µm

Wave optical simulations show that the focused diffraction spots of neighboring channels in the fiber facet plane do not enter the receiving fiber exactly on its optical axis. The oblique incidence of the higher order diffracted beams results in a slight lateral focal shift of 9 µm. Since the 9 foci of the surrounding channels have to be coupled into the same receiver fiber core, we need to use MM fiber with a sufficiently large core diameter of 62.5µm at the exit side of the component. When using SM fiber, the total optical power in the diffracted spots is 93%. Due to the necessary 16-level phase quantization and pixel quantization to 1x1 or 2x2 µm pixels in the fabrication process, this power ratio drops to 90.7%, 86.5% and 66.9% respectively, with a focal spot diameter of 6 µm. For MM fiber, the optical power in the diffractive spots decreases in case no quantization happens, and the power ratios in case of phase and lateral pixel quantization decrease too, being the focal spot bigger in this case.

5. ARCHITECTURE SIMULATION

To evaluate the SOB device plugged into a real DSM multiprocessor server architecture, we have performed practical computer simulations by augmenting a commercially available system simulator with our reconfigurable optical interconnect design, as presented before in section 2.2. For stablishing first the optimal placement maping between the nodes and the SOB component, we calculated the potential hop distance (the minimal hop distance over all reconfiguration possibilities) for every node pair as a metric for evaluating the relative performance of different placements. By using simulated annealing calculations, it results in a potential hop distance improvement of 40.8% compared to the average of 1000 random placements. This improvement increases as the number of nodes in the network increases. Performing the same optimization for larger torus networks resulted in more pronounced potential hop distance reductions as logically expected, with an asymptotic value of 60% for very large networks.

5.1 Simulation results

Finally, the impact of adding optical reconfiguration to a heavily stressed multiprocessor machine was measured, resulting in a speedup of the applications executed. We have simulated different architectural scenarios, ranging from the first ideal situation presented in section 2.3 in which no limits are imposed on the placement of the extra links, to our specific SOB architecture, quantifying the impact of each physical limitation that this implementation imposes.

As explained in section 3.2, a full connectivity solution with star-couplers is not scalable as the tunable sources will likely have only a limited number of wavelength channels available. By using the SOB element, we only allow each link to connect to 9 different destinations, making our solution easily scalable but also reducing even further the number of node pairs that can be connected by extra links. Light from one node can now only be broadcasted to one of a limited subset of nodes, effectively clustering our nodes into subsets. A direct connection between a highly communicating node pair may no longer be possible in all cases. However, if we choose the subsets of nodes carefully by using an optimal placement matrix, we may still obtain a performance gain that is close to the situation using ideal star-couplers. Having a second set of destinations with another SOB increases the number of nodes that can be reached, but full connectivity still can not be reached. Table 5 and Fig. 8 show the performance gain in execution time when using the proposed scalable architecture with one and two SOB elements respectively, and only 9 possible destinations per link. As expected, the performance of the reconfiguration architecture is moderated, compared to the ideal case. Still, on average, our simulations show that with the SOB, as much as 70% of the previously predicted speedup with ideal passive star-couplers is maintained. The gain in latency is higher because it reflects the reconfiguration improvements in the communication parts and not the overall execution.

Table 5. Restricted SOB connectivity speed-ups

Application	2 Links/node	1 Link/node
Radiosity	0.5 %	0.5 %
Lu	4 %	3 %
Radix	5 %	5 %
Apache	6 %	5 %
Cholesky	8 %	3 %
Ocean	13 %	11 %
FFT	17 %	18 %
AVERAGE	**7.6 %**	**6.5 %**

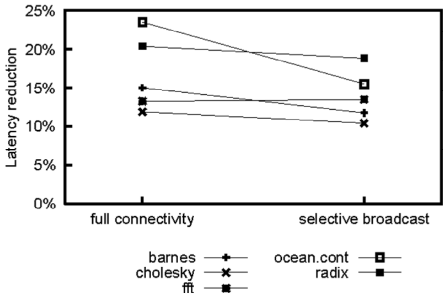


Fig. 8. Latency improvement with limited connectivity.

Finally, we have explored the influence of the reconfiguration interval, i.e., the time step between topology changes. Our study indicates that a reconfiguration interval of 10 ms should be fast enough to follow most traffic bursts and still profit from the addition of the extra links (see Table 6 and Fig. 9). The speedup for our set of benchmark applications does not changed significantly for reconfiguration intervals between 100 µs and 10 ms, and latency does not improve neither in a drastic way. Hence, we believe that it is possible to use the proposed reconfiguration scheme with current and inexpensive slow-switching photonics, although further simulations should address the impact of different selection and switching downtimes for the extra optical links.

Table 6. Impact of the reconfiguration interval

Application	100 µs	1 ms	10 ms
Radiosity	0.5 %	0.5 %	0.1 %
Lu	3 %	1 %	3 %
Cholesky	3 %	4 %	3 %
Radix	5 %	5 %	5 %
Ocean	11 %	9 %	9 %
FFT	18 %	14 %	16 %
AVERAGE	**6.8 %**	**5.6 %**	**6 %**

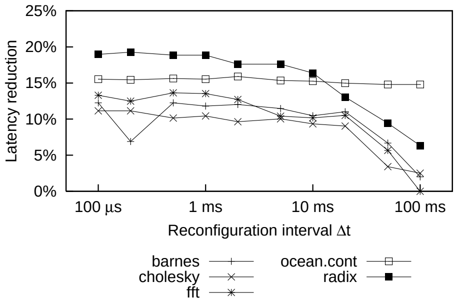


Fig. 9. Latency improvement with limited tuning speed.

6. CONCLUSIONS

We conclude that it can be possible to build a DSM machine augmented with reconfigurable optical interconnects. In order to be able to use low-cost optoelectronic devices that can follow long lasting communication patterns, a system with MEMS-based wavelength tunable VCSELs was proposed. Given the constraints on the tuning speed, such devices will likely not have large channel counts available for a full connectivity WDM scheme. Therefore we have also proposed a broadcast-and-select scheme that makes use of a dedicated SOB element. This element has been prototyped with the presented DPW technology and has potential for mass production with current replication technologies.

The use of the passive SOB element and tunable optoelectronics allows for a scalable scheme with reconfigurable optical interconnects. Using extensive full system simulations, we predict an improvement on the performance with an average of 6.5% and 7.6% speed-ups over all execution time when SOB is combined with tunable transceivers in every node. The parallel behavior of every application affects the performance of the reconfiguration, as some are more dependent on network connectivity. In future multiprocessor designs with a higher number of nodes in the system or with multicore processors in every board, the obtained performance gain can be expected to be more pronounced.

ACKNOWLEDGEMENTS

The authors thank all members of the PMMA group at the Dept. of Applied Physics and Photonics in the Vrije Universiteit Brussel (VUB) for collaboration and research in recent years on the SOB component and for providing the picture shown in Fig. 6. This work is financed by the EC 6th FP Network of Excellence on Micro-Optics (NEMO), the FWO, IWT-GBOU, VUB-GOA, the DWTC IAP Photon network and the OZR of the VUB.

REFERENCES

1. E. Mohammed et al., “Optical interconnect system integration for ultra-short-reach application”, *Intel Technology Journal*, Vol. 8, 2004.
2. K. Krewell, “Best servers of 2004: where multicore is the norm”, *Insider Guide to Microprocessor hardware*, 2005.
3. HyperTransport Consortium (www.hypertransport.org).
4. A. Carlesworth, “The Sun Fireplane system interconnect”, Sun Microsystem, Inc., 2001.
5. RapidIO Trade Association, (www.rapidio.org).
6. PCI Special Interest Group, (www.pcisig.com).
7. P. Krishnamurthy, “Reconfigurability of the interconnect architecture for chip multiprocessors”, *Proc. of the 4th International Symposium on Information and Communication Technologies*, pp. 136–141, 2005.
8. J. H. Collet et al., “Architectural approach to the role of optics in monoprocessor and multiprocessor machines”, *Applied Optics*, Vol. 39, pp. 671-682, 2000.
9. 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”, *Proc. ECOC*, Vol. 502, pp. 527-528, 2005.
10. W. Heirman et al., “Prediction model for evaluation of reconfigurable interconnects in Distributed Shared-Memory systems”, *Proc. of the International Workshop on System Level Interconnect Prediction*, pp. 51-58, 2005.
11. G. Verschaffelt et al., “Polarisation switching in vertical cavity surface-emitting lasers: from experimental observations to applications”, *Opto-electronics Review*, Vol. 9, pp. 257-268, 2001.
12. C. J. Chang-Hasnain, “Tunable VCSEL”, *IEEE Journal on Selected Topics in Quantum Electronics*, Vol. 6, pp. 978-987, 2000.
13. D. Van Thourhout, L. Zhang, W. Yang, B.I. Miller, N.J. Sauer, C.R. Doerr, “Compact digitally tunable laser”, *IEEE Photonics Technology Letters*, Vol. 15, pp. 182-184, 2003.
14. Dong Il Yeom et al., “Tunable narrow-bandwidth optical filter based on acoustically modulated fiber Bragg grating”, *IEEE Photonics Technology Letters*, Vol. 16, pp. 1313–1315, 2004.
15. A.E. Chiou, P. Yeh, “2 x 8 photorefractive reconfigurable interconnect with laser diodes”, *Applied Optics*, Vol. 31, pp. 5536-5541, 1992.

16. J. Yang, Q. Zhou, R. T. Chen, “Polyimide-waveguide-based thermal optical switch using total-internal-reflection effect”, *Applied Physics Letters*, Vol. 81, pp. 2947-2949, 2002.
17. V. Lien, Y. Berdichevsky, Y-H. Lo, “A prealigned process of integrating optical waveguides with microfluidic devices”, *IEEE Photonics Technology Letters*, Vol. 16, pp. 1525-1527, 2004.
18. T.H. Barnes et al., “Reconfigurable free-space optical interconnections with a phase-only liquid-crystal spatial light modulator”, *Applied Optics*, Vol. 31, pp. 5527-5535, 1992.
19. I. Fujieda, O. Mikami, A. Ozawa, “Active optical interconnect based on liquid-crystal grating”, *Applied Optics*, Vol. 42, pp. 1520-1525, 2003.
20. T. Yamamoto et al., “A three-dimensional MEMS optical switching module having 100 input and 100 output ports”, *IEEE Photonics Technology Letters*, Vol. 15, pp. 1360-1362, 2003.
21. V.A. Argueta-Diaz, “Reconfigurable photonic switch based on a binary system using the White cell and micromirror arrays”, *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 9, pp. 594–602, 2003.
22. M.K. Gruber, “Planar-integrated free-space optical fan-out module for MT-connected fiber ribbons”, *Journal of Lightwave Technology*, Vol. 22, pp. 2218–2222, 2004.
23. A.F. Benner, “Exploitation of optical interconnects in future server architectures”, *IBM Journal of Research and Development*, 2005.
24. S.S. Cho, “Optical signal distribution on silicon using a thin film photodetector array embedded in a multimode interference (MMI) coupler”, *Digest of the LEOS Summer Topical Meetings*, pp. 17–18, 2004.
25. B.A. Lemoff, “Demonstration of a compact low-power 250-Gbs parallel-WDM optical interconnect”, *IEEE Photonics Technology Letters*, Vol. 17, pp. 220–222, 2005.
26. M. Emsley, “Silicon based resonant cavity enhanced photodetectors for optical interconnects”, *Proc. 17th Annual Meeting of the IEEE Lasers and Electro-Optics Society*, Vol. 1, pp. 146–147, 2004.
27. I.L. Chung, “A Method to Tune the Cavity-Mode Wavelength of Resonant Cavity-Enhanced Photodetectors for Bidirectional Optical Interconnects”, *IEEE Photonics Technology Letters*, Vol. 18, pp. 46–48, 2006.
28. S. Matsuo, Y. Yoshikuni, T. Segawa, Y. Ohiso, H. Okamoto, “A widely tunable optical filter using ladder-type structure”, *IEEE Photonics Technology Letters*, Vol. 15, pp. 1114-1116, 2003.
29. Y. Doi et al. “Flat and high responsivity CWDM photoreceiver using silica-based AWG with multimode output waveguides”, *IEE Electronics Letters*, Vol. 39, pp. 1603-, 2003.
30. A. Goulet et al., “Polarization-based reconfigurable optical interconnects in free-space optical processing modules”, *IEEE Photonics Technology Letters*, Vol. 10, pp. 367-369, 1998.
31. D. M. Marom, P. E. Shames, F. Xu, Y. Fainman, “Folded free-space polarization-controlled multistage interconnection network”, *Applied Optics*, Vol. 37, pp. 6884-6891, 1998.
32. I. M. Barton, P. Blair, M. R. Taghizadeh, “Dual-wavelength operation diffractive phase elements for pattern formation”, *Optics Express*, Vol. 1, pp. 54-59, 1997.
33. B. E. Little, S. T. Chu, W. Pan, Y. Kokubun, “Microring resonator arrays for VLSI photonics”, *IEEE Photonics Technology Letters*, Vol. 12, pp. 323-325, 2000.
34. R. M. Kubacki, “Microresonator fabrication and integration for high density chip to chip optical interconnect”, Proc. *OSA Optics in Computing technical digest*, 2003.
35. M. Ishikawa, M. Naruse, A. Goulet, H. Toyoda, Y. Kobayashi, “Reconfigurable free-space optical interconnection module for pipelined optoelectronic parallel processing”, *Proceedings of SPIE*, Vol. 4457, pp. 82-87, 2001.
36. G. A. Russell, “Analysis and Modelling of Optically Interconnected Computing Systems”, *PhD. Thesis*, Heriot-Watt University of Edinburgh, 2004.
37. T.O. Yoshimura, “Three-dimensional self-organized microoptoelectronic systems for board-level reconfigurable optical interconnects: Performance, Modelling and simulation”, *IEEE Journal of STQE*, Vol. 9, pp. 492- 511, 2003.
38. R. Roldan, B.J. d'Auriol, “A preliminary feasibility study of the LARPBS optical bus parallel model”, *Proc, of the 17th International Symposium on High Performance Computing Systems and Applications*, pp. 181-188, 2003.
39. W. Heirman, J. Dambre, J. Van Campenhout, C. Debaes, H. Thienpont, “Traffic Temporal Analysis for Reconfigurable Interconnects in Shared-Memory Systems”, *Proc. IEEE Computer Society IPDPS*, pp. 150-, 2005.
40. B. Volckaerts et al., “Deep lithography with protons: a generic fabrication technology for refractive micro-optical components and modules”, *Asian Journal of Physics*, Vol. 10, pp. 195-214, 2001.
41. C. Debaes, et al., “Low-Cost Micro-Optical Modules for MCM Level Optical Interconnections”, *IEEE Journal of Selected Topics in Quantum Electronics, Special issue on Optical Interconnects*, Vol. 9, pp. 518-530, 2003.