Selective Optical Broadcast Component for Reconfigurable Multiprocessor Interconnects

Iñigo Artundo, *Student Member, IEEE*, Lieven Desmet, Wim Heirman, Christof Debaes, *Member, IEEE*, Joni Dambre, *Member, IEEE*, Jan M. Van Campenhout, *Member, IEEE*, and Hugo Thienpont, *Associate Member, IEEE*

Abstract—**Recent advances in the development of optical interconnect technologies suggest the possible emergence of optical interconnects within distributed shared memory (DSM) machines in the near future. Moreover, current developments in wavelength tunable devices could soon allow for the fabrication of low-cost, adaptable interconnection networks with varying switching times. It is the objective of this paper to investigate whether such reconfigurable networks can boost the performance of the DSM machines further. In this respect, we propose a system concept of a passive optical broadcasting component to be used as the scalable key element in such a reconfigurable network. We briefly discuss the necessary opto-electronic components and the limitations they impose on network performance. We show through detailed full-system simulations of benchmark executions, 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 optical broadcast component.**

Index Terms—**Broadcasting, optical interconnections, reconfigurable architectures, shared memory systems.**

I. 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]. Although recent developments of interprocessor communication technologies, such as HyperTransport [2] and Sun Fireplane [3] interconnects, can deliver high data throughput for the current generations of multiprocessor machines, it is widely recognized that replacement technologies will be required in the near future.

In distributed shared memory (DSM) multiprocessor machines all the memory of the system is distributed among its nodes. Nodes can access memory on other nodes in a software transparent way. The interconnection network is thus part of the memory hierarchy, and therefore high network latencies cause a significant performance bottleneck in the program execution. Interprocessor communication is already one of the main bottlenecks in current multiprocessors. This situation will become worse in the future, a result of increasing clock speeds, the rapid growth in instruction level parallelism and the use of multiple cores per die and multiple threads per core [4]. Reconfigurability in this aspect will allow the system to rearrange the interprocessor communication network, in order to avoid congestion and form topologies that are best suited for the particular computing task at hand, thus allowing for a network topology that closely matches the traffic patterns exhibited by the current application.

Optics is a great candidate to introduce fast interconnection networks in the architecture of the multiprocessor systems [5]. Using optical interconnects at the scale of link lengths found in multiprocessor machines (up to a few meters), an increase in connectivity and higher communication bandwidths can be expected, whereas the design of conventional electrical interconnects is limited by the tradeoff between interconnection length and bit rate. The high operating frequency of light tends to virtually eliminate frequency dependent cross-talk, and the inherent voltage isolation is also a highly demanded characteristic that will improve the signal integrity of the high speed communication channels. Finally, a very important aspect of optical interconnects is their inherent ability to switch the light paths easily in a data transparent way. This is paving the way towards adaptable network topologies.

It is the goal of this work to investigate whether these reconfigurable optical interconnect technologies can boost the performance of a DSM machine even further. We specifically target mid-range commercial servers that are often found in corporate IT departments, with large database or application server needs. These are typically shared memory systems with 16–64 processors. We will evaluate through extensive and detailed simulations the possible benefits of a reconfigurable interconnection network, taking in consideration latency, contention, and overall performance of the execution of a set of parallel applications. We will furthermore assess how a practical reconfigurable optical network can be incorporated into DSM systems, presenting a diffractive component that can be used in the proposed optical interconnection design. Finally, we will integrate the limitations that this component would impose into the network simulations, for evaluating its adequacy with the requirements to be met.

Manuscript received August 19, 2005; revised March 28, 2006. This work was supported in part by the GBOU, IWT, in part by the Fund for scientific research in Flanders FWO, in part by the IAP-V 18 PHOTON network sponsored by the Belgian Science Policy Office, in part by the OZR research commission of the Vrije Universiteit Brussel, and in part by the EC 6th FP Network of Excelence on Micro-Optics (NEMO).

I. Artundo, L. Desmet, C. Debaes, and H. Thienpont are with the Department of Applied Physics and Photonics, Vrije Universiteit Brussel, B-1050 Brussels, Belgium (e-mail: iartundo@tona.vub.ac.be; lieven.desmet@vub.ac.be; christof.deabes@vub.ac.be; hthienpo@vub.ac.be).

W. Heirman, J. Dambre, and J. M. Van Campenhout are with the Electronics and Information Systems Department, Ghent University, B-9000 Ghent, Belgium (e-mail: wheirman@elis.ugent.be; jdambre@elis.UGent.be; Jan.VanCampenhout@elis.ugent.be).

Digital Object Identifier 10.1109/JSTQE.2006.876158

The paper is organized as follows. Section II gives an overview of the current issues and technologies for reconfigurable interconnects in DSM systems. In Section III, we propose a design for implementing a broadcast-and-select reconfigurable optical interconnect, based on the concept of a low-cost selective optical broadcasting (SOB) element that allows for a flexible and scalable network clustering. In Section IV, the optical design of the component is described, and finally in Section V, we present the associated speedups such a reconfigurable scheme could allow for.

II. Reconfigurable Optical Interconnect Technologies

A. *Interconnect Requirements Within DSM Systems*

DSM multiprocessor designs include interconnection networks that provide communication between all the nodes in the system [6]. Data and control information—such as memory addresses from caches, or memory blocks to and from the processors—are interchanged at very high speed, using dedicated point-to-point links between pairs of processor nodes. In addition, some control information—like memory read/write requests or invalidations—is usually spread to multiple nodes over the network. This implies that the interconnection topology must be capable of easy data broadcasting while at the same time allowing for heavy data bursts between single node pairs.

Implementing fully interconnected networks is unachievable for a large number of nodes. Different topologies arise (such as hyper-mesh, torus, tree…), balancing high connectivity with acceptable technological complexity. These topologies are hardwired, and as a consequence, the performance can vary greatly for the same architecture depending on the traffic patterns generated by the applications in execution. For this reason, an efficient adaptable scheme would result in a large performance improvement by dynamically adjusting the network topology to match the specific run-time requirements [7]. For example, one could connect the nodes in a three-dimensional (3-D) mesh when running an astrophysical simulation or as a tree when executing a sorting algorithm.

Moreover, the hardware implementation of an optical interconnection network designed for the mid-range servers targeted in our study is restricted to two main factors: simplicity and price. As a consequence, the options for implementing such a design clearly leave behind most of the expensive equipment used commonly in telecom-grade reconfigurable optical networks, and the huge architectures used in massive high performance multiprocessors.

So before we discuss in detail our proposal of a dynamically reconfigurable optical network architecture, we will first give a brief overview of the existing state-of-the-art optical switching technologies that can be adequate for the targeted systems.

B. *Overview of Optical Switching Technologies*

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 active tunable opto-electronic sources, specialty tunable optical modulators, and beam steering micro-optical electro-mechanical systems.

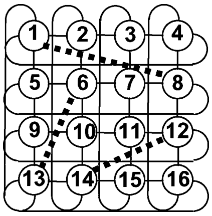

Fig. 1. Torus topology of the base interconnection network with some additional reconfigurable optical links. The numbers correspond to the different processor nodes in the network.

In tunable opto-electronics the characteristics of the emitted light, like the wavelength or polarization state [8], can be tuned by changing the operating conditions such as the temperature, the current injection or the micro-cavity geometry with microelectromechanical (MEMS) membranes. Some of the more remarkable achievements in this field for optical interconnects are the advent of MEMS-based two-dimensional (2-D) tunable vertical-cavity surface-emitting lasers (VCSEL) arrays [9], and the arrayed-waveguide-grating (AWG)-based tunable fiber lasers [10] for the high-end telecom market. The broadcast function can be implemented by a star coupler or beam splitting diffractive optical elements (DOEs) in free-space [11]. Wavelength selection can be obtained with selective resonant cavity photodetectors (RCPDs) [12], [13], tunable optical filters [14], AWGs [15], passive polarization sensitive DOEs [16], [17] or passive wavelength sensitive DOEs [18].

Another method for optical switching is to modulate the propagation of the light path instead of tuning the characteristics of the light sources in the network. In this way, tuning has been achieved by acoustooptic Bragg cells [19], with photorefractive crystals [20] or by using micro-fluids as the tuning element in waveguides [21]. Recently, liquid crystal (LC) components receive a lot of attention and are being used to make adaptive computer generated holograms [22] and switchable gratings [23].

A third approach for optical switching is the use of active free-space laser beam steering. Here, a MEMS micromirror-based device images a 2-D fiber array onto a second one [24], [25]. These devices are compact, consume low power and can be batch fabricated resulting in low cost. Examples of very high speed optical switches and with large port counts have been demonstrated recently [26]–[28].

III. Reconfigurable Optical Interconnect Architecture

A. *Proposal of the Network Architecture*

The proposed interconnection network architecture for the DSM system consists of a fixed-base network, arranged in a torus topology. In addition, a certain number of reconfigurable optical links are provided (see Fig. 1). They can serve as direct point-to-point connections between processor node pairs as reported in [29], or as a shortcut for the ongoing traffic flow between other nodes in the network.

This setup, compared to the case where all links in the network would be used for reconfiguration, has a number of advantages 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.

A broadcast-and-select scheme is used for providing the reconfigurable extra links, because of its simplicity, and because it can be implemented with very low-cost optical and optoelectronic components. These components and the overall physical implementation of the reconfigurable optical interconnect (ROI) network are discussed in Section III-B.

B. Reconfigurable Network Implementation

The ROI enabled part of the inter-processor communication network consists of a tunable optical transmitter per processor node. In this way, each node can transmit data on a fixed number of wavelengths. This signal is then guided to a broadcasting element that divides the data-carrying signal to all (or a selection of) the receiving nodes. Each processor node also incorporates an optical receiver that is sensitive to one wavelength only. Hence, by tuning the wavelength of each transmitter we address the destination, and the newly created links will alter the topology of the resulting network.

It is a challenge for the implementation of the ROI to avoid complicated switching devices or costly optoelectronics. VCSELs are attractive candidate optical sources for optical interconnects in general [30], primarily because of their low-cost mass production and testing on wafer-scale, low-power consumption, easy array integration and the rotationally symmetrical laser beam that can easily be coupled into the optical fiber while offering a small form factor for easy onboard integration. MEMS wavelength tunable VCSELs are an emerging extension to this technology and many technological issues have been solved in recent years. Therefore, we think that this type of devices might offer some opportunities for implementing cost effective coarse wavelength tunability in metro-area networks. This will give the necessary economy of scale to produce them in a cost-effective way. However, the arguments in this work are not constricted to MEMS tunable VCSELs and the use of other types of tunable lasers in the ROI network, such as Fabry–Perot or distributed feedback lasers, are possible.

After the optical broadcasting in the ROI network, the optical signals can be detected by using RCPDs. This type of photodetectors is only sensitive to a very narrow band of optical frequencies since their wavelength selectivity is enhanced by a Fabry–Perot cavity. Transmission of the optical signals in our ROI network will happen by means of single-mode (SM) or multimode (MM) optical fiber.

C. Design Challenges

A number of limitations imposed by the opto-electronic devices affect the reconfigurability of the proposed system.

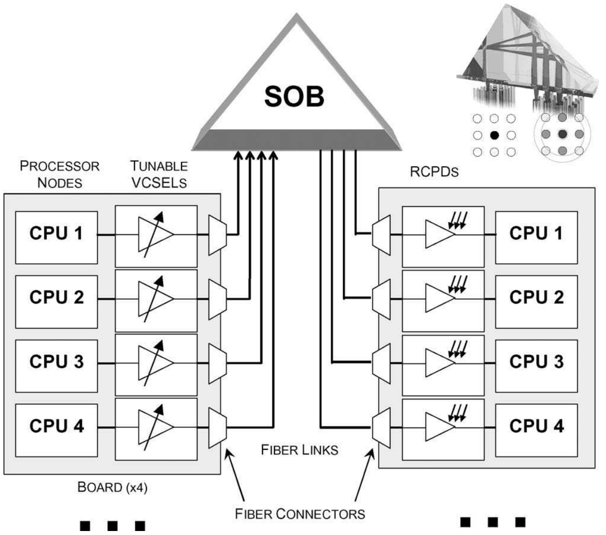

Fig. 2. Schematic representation of the complete reconfigurable optical interconnect with processors divided in groups of four per board. A processor node transmits data on one of nine wavelengths $\lambda1, \ldots, \lambda9$. The optical broadcast element distributes the signal towards nine fellow processor nodes. Since every receiving processor node is sensitive to one wavelength only, the target processor node is selected by emitting at the appropriate wavelength.

1) *The tuning speed of the transmitters' 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 beneficial adaptation.
2) *The number of wavelength channels is limited.* Given the need for low-cost devices with a relatively high tuning speed, it is reasonable to believe that the included tuning elements will exhibit a very limited number of wavelength channels, as a tradeoff exists between the tuning speed, the cost and channel count.

In a previous exploration of the influence of the reconfiguration speed on reconfigurable network performance [29], we have found the switching speed requirements on par with recent MEMS-based tunable VCSEL achievements. In Section V, 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 a 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 that broadcasts each channel to only a limited number of outputs. Fig. 2 shows the concept of the proposed ROI network containing such a passive optical broadcast element. The fibers coming from the tunable sources of each processing node are bundled into a fiber array at the ingress of the SOB. The free-space optical component will then fan out each input optical signal to a 3×3 matrix of spots at the output side via a DOE. In that way each processor is capable of connecting to nine different channels. A scalable ROI scheme is thus possible using components with a low number of wavelength channels. The effect on the connectivity and the performance speed-up of such a scheme is measured in Section V.

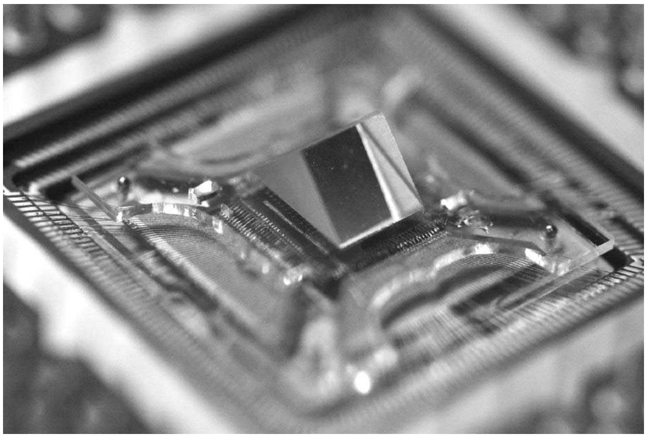

Fig. 3. Photograph showing the capabilities of the DPW technology. The presented optical intra chip interconnection module combines alignment features, refractive microlenses, and a prism.

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. The mapping of the possible connections between the nodes results in what we call the "placement matrix." In Section V, we present computer simulations to find the best solution for this placement matrix.

IV. Design of the Optical Broadcast Component

A. Multichannel Free-Space Optical Interconnect Module as the Basic Design for the Selective Optical Broadcast Component

Over the last few years, several free-space optical modules have been investigated at the research labs in the Applied Physics and Photonics Department at the Vrije Universiteit Brussel. These modules are capable of delivering high aggregate bandwidth through massive parallel free-space point-to-point optical interconnects [31]. Deep proton writing (DPW) is the core technology that is deployed for prototyping these components.

The DPW is prototyping micro-optical technology [32] and involves the irradiation of poly-methylmetacrylate (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. As an example, we show in Fig. 3, a recently prototyped optical interconnect module that promises to bring optical interconnects at the intrachip interconnect level [33]. It consists of a combination of a thick commercially available glass prism with a plastic baseplate containing micro-lens arrays fabricated by DPW. The component images a dense 2-D 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 optical beams through the component. The device can be made low-cost and our DPW technology has been shown to be compatible with current replication technologies [34].

TABLE I
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

In the following paragraphs, we investigate how we can integrate beam shaping and beam splitting diffractive optical elements within this component to enable selective optical broadcasting in an ROI network.

When fiber bundles are used, an in-line-coupling scheme for the module is possible, but in combination with our dense 2-D fiber arrays fabricated by DPW, 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 space occupied by fiber bending is minimized.

The SOB component holds a 5-mm right-angled prism with a base surface of 5 mm × 7 mm. This allows the placement of up to 330 optical nodes at a nodes positional pitch of 220 μm. This number of channels is more than sufficient to wire 16 or 64 processor boards. In case an even larger port count is necessary, the ROI system can be scaled up by using several SOB modules in parallel.

B. Passive Optical Broadcasting by Beam Splitting DOEs

We designed a phase-only DOE to achieve beam collimation and passive optical splitting in the SOB. Every source channel is fanned out towards nine diffracted spots in a 3×3 pattern configuration. The diffractive grating period Λ is calculated with the well-known grating equation (1) in such a way that the diffracted spots exactly fit on the exit fiber node lenses

$$\Lambda \sin \alpha = \frac{m\lambda}{n}. \tag{1}$$

In (1), 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 interconnect 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 the source channel and the exit node channel as given in (2)

$$\alpha = \arctan \frac{P}{\mathrm{OPL}}. \tag{2}$$

The OPL in the optical broadcast component is 8 mm (7 mm of OPL in the micro-prism augmented with two times the 500-μm microlens substrate thickness). 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 by using DPW. Using these values of P and OPL, we summarize in Table I, 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 of 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 of other tunable sources will relax the minimal size of the basic diffractive beamshaping cell.

The basic cell of the beam splitting DOE design (which is called the kinoform and has dimensions of Λ μm × Λ μm) is obtained with the iterative Fourier transform algorithm (IFTA) in the VirtualLab v.1.3 software package of LightTrans GmbH. The software uses the merit functions of the best diffraction efficiency (DE) and the minimal optical signal-to-noise ratio (SNR) to converge to a possible solution for the Fourier spot pattern generator. A good equal power division over the different diffraction spots is obtained.

C. DOE Integration in the Optical Broadcast Component

The diffractive kinoform covers the source channel side of the baseplate underneath the micro-prism. The refractive microlenses 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 150 μm and the front focal length (FFL) is 335 μm. The distance between the broadcast component 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 minimizing the power clipping loss in the first lens aperture. Using standard Gaussian beam propagation, the clipping loss at the ingress lens aperture was only 0.2% using SM fiber with a numerical aperture (NA) of 0.12 (this corresponds to a source FWHM of 8°). The clipping loss will increase to 16.5% if MM ingress fibers with a NA of 0.22 (this corresponds to a source FWHM of 15°) would be deployed instead.

D. Wave Optical Simulations

With the use of the wave optical simulator, we modeled the SOB component from the ingress fiber facet towards the plane where the receiving fiber facets are located. The optical field was propagated with the Rayleigh–Sommerfeld method. The demonstrator component would require the addition of a reflective metal coating on the deflecting sidewalls of the prism to satisfy the TIR condition for all of the diffracted beams. The total sidewall reflection power penalty is estimated to be 0.1 dB, since metal coatings exhibit reflectivities up to 99%. The simulated efficiencies given in Table II, however, are obtained under lossless reflection conditions. The simulation sampling size was 500 nm.

Simulations show that the focused diffraction spots of neighboring channels in the fiber facet plane enter the receiving fiber off-center. The oblique incidence of the higher order diffracted beams results in a slight lateral focal shift of 9 μm. The focal shift δx observed in the simulations is in good agreement with the estimated value, according to (3)

$$\delta x = \mathrm{FFL}_{\mathrm{microlens}} \tan\alpha. \tag{3}$$

Since the nine 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 62.5 μm at the exit side of the SOB.

Wave optical simulations show that, when using SM fiber at the ingress side of the SOB, the total optical power in the diffracted spots is 93%. The focal spots are 6 μm in diameter. When deploying MM fiber instead, the efficiency of the SOB drops to 75.4%. In these idealized simulations, the diffractive designs are represented by their ideal continuous relief profile. However such complicated, asymmetrical surface designs can at present only be commercially fabricated at reasonable cost through binary lithographic and binary etching processes. State-of-the-art fabrication can involve a set of four-phase etching masks containing pixels with minimum feature sizes down to 1–2 μm. Using four-phase masks results in a 16-level discretisation of the height profile of the optical relief. The fabricated DOE will, therefore, only approximately come close to the idealized diffractive continuous relief profile design. With subsequent wave optical simulations we show that given these fabrication constraints, the efficiency of the SOB drops to 90.7% (SM fiber) and 73.9% (MM fiber) when 16-level phase quantization is taken into account. Additionally the SOB efficiency drops in the case of SM ingress fiber, towards 86.5% and 66.9% when 2- and 1-μm lateral pixelation of the designs is addressed. In case of using MM fiber, the SOB efficiencies are respectively, 67.9% and 48.4% (see Table II).

We have furthermore performed some first order simulations on the thermal behavior of the SOB component, in which we assumed an isotropic heating of the component. The thermo-optic effect on the index of refraction of PMMA is -3.1×10^{-4}/K and the linear expansion coefficient is 7×10^{-5}/K. When the working temperature was elevated from 20 °C to 50 °C, we found only a small efficiency penalty of 0.2% in comparison with the values presented in Table II. However, a worst-case scenario including asymmetrical and nonisotropic heating of the component would demand for a full thermo-opto-mechanical tolerance study to be performed.

V. Architecture Simulation

To evaluate the performance gain with 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.

A. Optimal Link Placement

At first, we have investigated how the fibers should be plugged into the SOB component. We have determined an optimal placement, based on the following considerations. The latency of a packet across the network is a function of the *hop distance*, i.e., the minimal number of intermediate nodes (hops)

TABLE II
OPTICAL BROADCAST COMPONENT EFFICIENCY RESULTS BY WAVE OPTICAL SIMULATIONS

Source node type		*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

the packet must traverse before reaching its destination. On average, it is most beneficial for packet latency to create extra optical links between node pairs that are far apart in the base network topology. An optimal topology is hence one that has a minimal hop distance among all node pairs. With the proposed selective broadcasting scheme, nine additional nodes can be potentially reached in a single hop for every extra link to the SOB. At runtime however, only one of these nine nodes can actually be addressed since the transmitter on this node will be tuned to the frequency of only one of them. This means that the hop *distance* for the reconfigurable network will vary over time. Therefore, we have 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 the different placements.

We have implemented a simulated annealing algorithm to determine the best possible placement matrix in terms of total potential hop distance (summed over all N^2 node pairs). For the 16 nodes torus network case and the broadcast configuration towards nine neighboring nodes, our best placement 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, with an asymptotic value of 60% for very large networks.

B. Simulation Environment

We have established a full-system simulation environment based on Simics [35], a commercially available execution-driven multiprocessor simulator, that is capable of simulating complete computing systems, including the behavior of the operating system and the user loads. A more detailed description of our environment can be found in [36]. The simulator was configured to model a multiprocessor machine based on the Sun Fire 6800 server [3], with 16 UltraSPARC III processors at 1-GHz running the Solaris 9 operating system. We extended the Simics simulator with a module that simulates the interconnection layer of a 4 × 4 fixed network and the reconfigurable links. The models use contention and cut-through routing. The SPLASH-2 scientific parallel benchmark suite [37], as well as the Apache web server along with the SURGE request generator [38], were chosen as the workload applications for stressing the system under test. Since the proposed ROI scheme with the SOB scales well with the number of processing nodes, our simulation results would benefit from higher processor counts. Unfortunately, due to the extremely long simulation times, it is not feasible to perform simulations for the moment with more than 16 parallel processors.

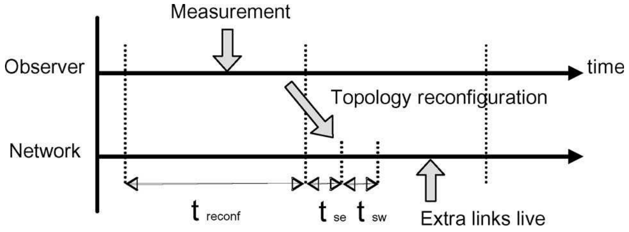

Fig. 4. In every reconfiguration interval, the system is monitoring the traffic flow, such that it can adjust the topology to accommodate the actual communication needs in the next reconfiguration interval.

We have partitioned the simulated time in discrete reconfiguration intervals such that the topology changes take place at fixed times (Fig. 4). This interval, t_{reconf}, should be sufficiently short to keep pace with the changing demands made by the application, but it must be long enough to amortize on the cost of reconfiguration, during which the extra links are unusable. For now, we have not considered any downtime that occurs during network readjustments (selection process and optical switching, $t_{\text{se}} + t_{\text{sw}}$), to keep the performance study independent of the chosen tuning technology.

The congestion model is derived by viewing the network as an interconnected set of queues and servers, routed in a virtual cut-through way, with some small packet buffering at each entrance of every processing node to avoid deadlocks. Adding waiting times over all links on the path of a packet gives us the total waiting time. We finally compute the average expected memory access latency, which is used as the performance indicator for the network [39].

We have furthermore assumed equal speed characteristics for the fixed and the extra reconfigurable optical links, yielding the same average per hop packet latency for both types of links.

C. Simulation Results

The impact of adding optical reconfiguration to a heavily stressed multiprocessor machine was measured. We have simulated different architectural scenarios, ranging from a situation in which no limits are imposed on the placement of the extra links, to our specific architecture using the SOB component. With these different simulations, we quantified the impact of each physical limitation our implementation is imposing. The metrics used are the performance speedup of the complete

TABLE III
RECONFIGURABLE ARCHITECTURE SIMULATION SPEEDUPS
(RECONFIGURATION INTERVAL $t_{\mathrm{reconf}} = 100\ \mu$s)

	16 links Unlimited connectivity			16 links SOB	
Application	**N = ∞**	**N = 2**	**N = 1**	**N = 2**	**N = 1**
Radix	17 %	7.2 %	6.9 %	5 %	5 %
Lu	7 %	5.1 %	3.5 %	4 %	3 %
Cholesky	9 %	7 %	1.2 %	8 %	3 %
Radiosity	43 %	12.5 %	3.1 %	0.5 %	0.5 %
FFT	42 %	26.8 %	17.5 %	17 %	18 %
Ocean	38 %	23.6 %	21.9 %	13 %	11 %
Apache	7 %	4 %	4 %	6 %	5 %
AVERAGE	**23.3 %**	**12.3 %**	**8.3 %**	**7.6 %**	**6.5 %**

execution of the applications, and the latency savings on the packet communications. There is a certain level of noise (2%–5%) on the application runtimes, stemming from the initial state of the cache memories as well as other scheduled internal tasks of the operating system at the beginning of the simulations.

The first simulated architecture is one in which 16 extra links can be added to the network. No limits are imposed on the choice on which of the 16 node pairs are connected at each time. Therefore, the 16 busiest node pairs can be directly connected by extra links and there is no limit on the number of extra links N that can end at the same node. After adding these links, latency is greatly reduced for a large percentage of the traffic, 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 III, using a reconfiguration interval of 100 μs.

This solution is however far from a physical implementation, as it requires an extra optical transceiver for every link that potentially terminates in each processing node. In the worst case 15 extra links end at the same node, and therefore each node must have 15 optical transceivers.

In the next stage, we gradually move from the ideal situation to something that more closely resembles the proposed architecture. In the second and third columns of Table III, we impose the limit on the system such that at most one or two reconfigurable links can terminate in each node ($N = 1$ or 2). This can be implemented using a small number of transceivers in each node, together with one or two broadcasting devices that can reach any destination in the network, such as using optical passive star couplers. Introducing this limitation greatly reduces the number of transceivers located in every processor node. It also means that sometimes, when a node is receiving a burst of traffic from different destinations, reconfiguration cannot fully accommodate the "star"-shaped traffic pattern by having all extra links terminate at the single aggregating node. This effect decreases the efficiency of the extra links, as the new channels are not always situated where they would be needed.

As explained in Section III-C, such a solution is not scalable as the wavelength tunable sources will likely to have only a limited number of wavelength channels available. This is why we use a SOB element that only allows each link to connect to nine different destinations. Although our solution is easily scalable, it further restricts the number of node pairs that can be connected together. Light from one node can now only be broadcasted to one of a limited subset of other nodes. This effectively clusters our nodes into (nondistinct) subsets. A direct connection between an arbitrary, highly communicating node pair may no longer be possible in all cases. However, if we choose these subsets of nodes carefully (by using an optimal placement matrix), we may still obtain a performance gain that is close to the situation using the star-couplers. The last two columns of Table III show the performance speedups when using the proposed scalable architecture, with one and two SOB element respectively, and nine 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 predicted speedup with ideal passive star-couplers is maintained.

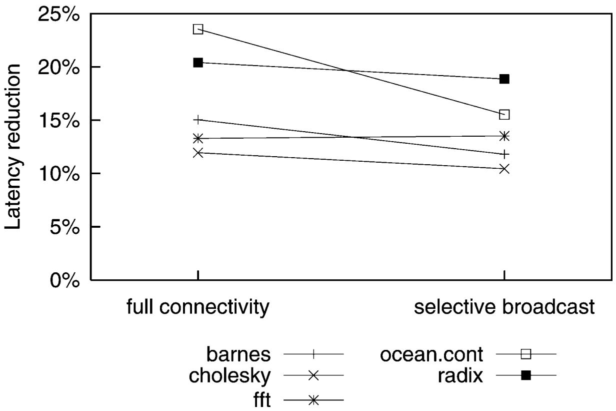

Fig. 5. Reduction of the network access latency for the case where two links can simultaneously end in one node (SOB, $N = 2$).

These resulting 7.6% and 6.5% improvement can be perceived as low and thus not worth of the effort for implementing the architectural change. By only changing the networking capabilities, it is not possible to improve the overall performance by a large amount. This is due to the fact that many of the scientific benchmark applications spend considerable time in processing local data that do not require network transportation. Furthermore, a large part of the network access latency is also taken up by the SDRAM latency (100 cycles) at the memory banks, which we assumed to be constant in order to present a fair comparison. As a result, by even considering an ideal network with no contention and zero delay, you would find resulting performance gains that are in the same order of magnitude.

However, as the size of the network and the processing speed increases, the bottleneck is shifted more towards reducing the network access latency. As there are no cost-effective improvements on the horizon to reduce the memory access time, the best we can do is to reduce the network delays as much as possible. We have furthermore included the reduction in the total network access latency in Figs. 5 and 6.

Finally, we have explored the influence of the reconfiguration interval, i.e., the time between topology changes. Our study indicates that a reconfiguration interval of 10 ms is fast enough to follow most longstanding communication patterns (see Table IV), which is in accordance with previous studies [29]. The speedup for our set of benchmark applications does not change significantly for reconfiguration intervals between 100 μs and

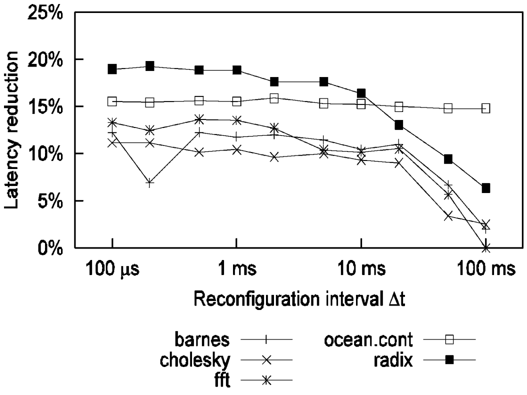

Fig. 6. Performance improvement of the network access latency for different reconfiguration intervals (16 SOB links, $N = 1$).

TABLE IV
IMPACT OF THE RECONFIGURATION INTERVAL (t_{reconf})
(16 SOB LINKS, $N = 1$ LINK PER NODE)

Application	100 µs interval	1 ms interval	10 ms interval
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 %**

10 ms. Hence, we believe that it is possible to use the proposed reconfiguration scheme with slow (millisecond range) tunable photonics technology.

VI. CONCLUSION

We conclude that it is sensible to build a DSM machine with optical reconfigurable interconnects with switching times that are much larger than the processor clock speed. These devices exhibit constraints in the tuning speed and the number of wavelength channels available. Therefore, we have proposed a broadcast-and-select reconfigurable interconnection system with a dedicated SOB element. The device can be made with the DPW technology and can be replicated at low cost.

The combination of the passive SOB element with tunable opto-electronics allows for a scalable scheme for reconfigurable optical interconnects. Using extensive full system simulations for evaluating the performance of the proposed interconnect design, we predict an improvement of the global execution time with an average of 6.5% speedup (for a 16 processor machine) when a single SOB is combined with one reconfigurable transceiver per node. If two SOBs are used in parallel, we can expect a gain of 7.6%. Latency reductions of about 20% clearly show the reduction of congestion obtained by the addition of the extra optical links.

A study of the effect of the reconfiguration time shows that the proposed scheme is already practical with current MEMS-based tunable VCSELs. The obtained speed improvements will be more pronounced for larger number of processing nodes.

ACKNOWLEDGMENT

The authors would like to thank all recent and past members of the PMMA group at the Department of Applied Physics and Photonics at the Vrije Universiteit Brussel for collaboration and research in the recent years on the optical broadcast component and for providing the photograph shown in Fig. 3.

REFERENCES

[1] E. Mohammed *et al.* (2004, May). Optical interconnect system integration for ultra-short-reach application. *Intell. Technol. J.* [Online]. *8(2)*. Available: http://developer.intel.com/technology/itj/2004/volume08issue02/

[2] C. N. Keltcher, K. J. McGrath, A. Ahmed, and P. Conway. The AMD opteron processor for multiprocessor servers. IEEE Micro. [Online]. *23(2)*. pp. 66-76. Available: http://www. hypertransport.org

[3] A. Charlesworth, "The sun fireplane system interconnect," in *Proc. 2001 ACM/IEEE Conf. Supercomput.*, p. 2.

[4] K. Krewell, (2005, Jan. 18). Best servers of 2004: Where multicore is the norm. *Microprocessor Rep., In-Stat.* [Online]. Available: www.mdronline.com/mpr/h/2005/0118/190301.html

[5] J. H. Collet *et al.*, "Architectural approach to the role of optics in monoprocessor and multiprocessor machines," *Appl. Opt.*, vol. 39, no. 5, pp. 671–682, Feb. 2000.

[6] J. Duato, S. Yalamanchili, and L. Ni, *Interconnection Networks*. San Mateo, CA: Morgan Kaufmann, 2002.

[7] P. Krishnamurthy, "Reconfigurability of the interconnect architecture for chip multiprocessors," in *Proc. 4th Int. Symp. Inform. Commun. Technol.*, 2005, pp. 136–141.

[8] G. Verschaffelt *et al.*, "Polarisation switching in vertical cavity surface-emitting lasers: From experimental observations to applications," *Optoelectron. Rev.*, vol. 9, pp. 257–268, 2001.

[9] C. J. Chang-Hasnain, "Tunable VCSEL," *IEEE J. Sel. Topics Quantum Electron.*, vol. 6, no. 6, pp. 978–987, Nov.–Dec. 2000.

[10] D. Van Thourhout, L. Zhang, W. Yang, B. I. Miller, N. J. Sauer, and C. R. Doerr, "Compact digitally tunable laser," *IEEE Photon. Technol. Lett.*, vol. 15, no. 2, pp. 182–184, Feb. 2003.

[11] B. E. Lemoff, "Demonstration of a compact low-power 250-Gb/s parallel-WDM optical interconnect," *IEEE Photon. Technol. Lett.*, vol. 17, no. 1, pp. 220–222, Jan. 2005.

[12] M. K. Emsley, "Silicon based resonant cavity enhanced photodetectors for optical interconnects," in *Proc. 17th Annu. Meeting IEEE Lasers Electro-Opt. Soc.* Nov. 7–11, 2004, vol. 1, pp. 146–147.

[13] I. L. Chung, "A method to tune the cavity-mode wavelength of resonant cavity-enhanced photodetectors for bidirectional optical interconnects," *IEEE Photon. Technol. Lett.*, vol. 18, no. 1, pp. 46–48, Jan. 2006.

[14] S. Matsuo, Y. Yoshikuni, T. Segawa, Y. Ohiso, and H. Okamoto, "A widely tunable optical filter using ladder-type structure," *IEEE Photon. Technol. Lett.*, vol. 15, no. 8, pp. 1114–1116, Aug. 2003.

[15] Y. Doi *et al.*, "Flat and high responsivity CWDM photoreceiver using silica-based AWG with multimode output waveguides," *Electron. Lett.*, vol. 39, no. 22, pp. 1603–1604, Oct. 2003.

[16] A. Goulet *et al.*, "Polarization-based reconfigurable optical interconnects in free-space optical processing modules," *IEEE Photon. Technol. Lett.*, vol. 10, no. 3, pp. 367–369, Mar. 1998.

[17] D. M. Marom, P. E. Shames, F. Xu, and Y. Fainman, "Folded free-space polarization-controlled multistage interconnection network," *Appl. Opt.*, vol. 37, pp. 6884–6891, 1998.

[18] I. M. Barton, P. Blair, and M. R. Taghizadeh, "Dual-wavelength operation diffractive phase elements for pattern formation," *Opt. Exp.*, vol. 1, pp. 54–59, 1997.

[19] D. I. Yeom *et al.*, "Tunable narrow-bandwidth optical filter based on acoustically modulated fiber Bragg grating," *IEEE Photon. Technol. Lett.*, vol. 16, no. 5, pp. 1313–1315, May 2004.

[20] A. E. Chiou and P. Yeh, "2 × 8 photorefractive reconfigurable interconnect with laser diodes," *Appl. Opt.*, vol. 31, pp. 5536–5541, 1992.

[21] V. Lien, Y. Berdichevsky, and Y. H. Lo, "A prealigned process of integrating optical waveguides with microfluidic devices," *IEEE Photon. Technol. Lett.*, vol. 16, no. 6, pp. 1525–1527, Jun. 2004.

[22] T. H. Barnes *et al.*, "Reconfigurable free-space optical interconnections with a phase-only liquid-crystal spatial light modulator," *Appl. Opt.*, vol. 31, pp. 5527–5535, 1992.

[23] I. Fujieda, O. Mikami, and A. Ozawa, "Active optical interconnect based on liquid-crystal grating," *Appl. Opt.*, vol. 42, pp. 1520–1525, 2003.

[24] M. K. Gruber, "Planar-integrated free-space optical fan-out module for MT-connected fiber ribbons," *J. Lightw. Technol.*, vol. 22, no. 9, pp. 2218–2222, Sep. 2004.

[25] V. Argueta-Diaz and B. L. Anderson, "Reconfigurable photonic switch based on a binary system using the White cell and micromirror arrays," *IEEE J. Sel. Topics Quantum Electron.*, vol. 9, no. 2, pp. 594–602, Mar.–Apr. 2003.

[26] T. Yamamoto *et al.*, "A three-dimensional MEMS optical switching module having 100 input and 100 output ports," *IEEE Photon. Technol. Lett.*, vol. 15, no. 10, pp. 1360–1362, Oct. 2003.

[27] R. Ryf *et al.*, "1296-port MEMS transparent optical crossconnect with 2.07 petabit/s switch capacity," in *Proc. Optic Fiber Commun. Conf.*, 2001, vol. 4, pp. PD28-1–PD28-3.

[28] K. J. Barker *et al.*, "On the feasibility of optical circuit switching for high performance computing systems," in *Proc. ACM/IEEE Conf. Supercomput.*, 2005, p. 16.

[29] W. Heirman *et al.*, "Wavelength tunable reconfigurable optical interconnection network for shared-memory machines," in *Proc. 31st IEE Eur. Conf. Opt. Commun.*, 2005, vol. 3, pp. 527–528.

[30] R. Michalzik, J. M. Ostermann, M. Riedl, F. Rinaldi, H. Roscher, and M. Stach, "Novel VCSEL designs for optical interconnect applications," presented at the 10th OptoElectron. Commun. Conf., Seoul, Korea, 2005.

[31] C. Debaes *et al.*, "Low-cost microoptical modules for MCM level optical interconnections," *IEEE J. Sel. Topics Quantum Electron.*, vol. 9, no. 2, pp. 518–530, Mar.–Apr. 2003.

[32] B. Volckaerts *et al.*, "Deep lithography with protons: a generic fabrication technology for refractive micro-optical components and modules," *Asian J. Phys.*, vol. 10, no. 2, pp. 195–214, 2001.

[33] M. Vervaeke *et al.*, "Opto-mechanical Monte Carlo tolerancing study of a packaged free-space intra-MCM optical interconnect system," *J. Sel. Topics Quantum Electron.*, vol. 42, no. 7/8, 2006.

[34] P. Tuteleers *et al.*, "Replication of refractive micro opto-mechanical components made with deep lithography with protons," in *Proc. SPIE*, 2001, vol. 4408, pp. 329–337.

[35] P. S. Magnusson, M. Christensson, J. Eskilson, D. Forsgren, G. Hâllberg, J. Högberg, F. Larsson, A. Moestedt, and B. Werner, "Simics: A full system simulation platform," *Computer*, pp. 50–58, Feb. 2002.

[36] W. Heirman, J. Dambre, D. Stroobandt, C. Debaes, H. Thienpont, and J. Van Campenhout, "Prediction model for evaluation of reconfigurable interconnects in distributed shared-memory systems," in *Proc. Int. Workshop Syst. Level Interconnect Predict.*, 2005, pp. 51–58.

[37] S. Woo, M. Ohara, E. Torrie, J. Singh, and A. Gupta, "The SPLASH-2 programs: Characterization and methodological considerations," in *Proc. 22nd Annu. Int. Symp. Comput. Architect.*, 1995, pp. 24–36.

[38] P. Barford and M. Crovella, "Generating representative web workloads for network and server peformance evaluation," in *Proc. ACM SIGMETRICS*, 1998, pp. 151–160.

[39] W. Heirman, J. Dambre, and J. Van Campenhout, "Congestion modeling for reconfigurable inter-processor networks," in *Proc. Int. Workshop Syst. Level Interconnect Predict.*, 2006, pp. 59–66.

Iñigo Artundo (S'04) was born in Pamplona, Spain, on October 21, 1979. He received the Master degree in telecommunication engineering from the Public University of Navarra, Navarra, Spain, in 2004. Currently, he is working toward the Ph.D. degree in the field of reconfigurable optical interconnects architectures in the Department of Applied Physics and Photonics, Vrije Universiteit Brussel, Brussels, Belgium.

His current research interests are reconfigurable architectures, optical interconnection networks, and distributed shared-memory systems.

Lieven Desmet was born in Kortrijk, Belgium, on August 25, 1975. He received the Master degree in industrial engineering with majors in micro-electronics with distinction from the Polytechnical School BME-CTL, Ghent, Belgium, in 1997. He received the Master degree in civil engineering with majors in photonics with distinction from the Vrije Universiteit Brussel (VUB), Brussels, Belgium, in 2000. He is currently working toward the Ph.D. degree in reconfigurable free-space optical interconnects in the Department of Applied Physics and Photonics, Vrije Universiteit Brussel.

His current research interests are diffractive optics, free-space optics, and free-space laser links.

Wim Heirman was born in Temse, Belgium, on November 28, 1980. He received the M.Sc. degree in computer engineering from Ghent University, Ghent, Belgium, in 2003. He is currently working toward the Ph.D. degree in the Department of Electronics and Information Systems, Ghent University.

His current research interests include parallel computing systems, reconfigurable architectures, and interconnection networks.

Christof Debaes (S'99–M'04) was born in Geraardsbergen, Belgium, in 1975. He received the graduate degree in electrotechnical engineering from the Vrije Universiteit Brussel (VUB), Brussels, Belgium, in 1998, and the Ph.D. degree from the Applied Physics and Photonics Department, VUB, in collaboration with the Ginzton Laboratory, Stanford University, Stanford, CA, in 2003.

He is currently with the VUB on a postdoctoral fellowship from the Flemish Fund for Scientific Research (FWO-Vlaanderen). His research activities include optical interconnects, covering a wide range of subjects such as optical clock injection, opportunities for reconfigurable optical interconnect and the use of Deep Proton Writing for micro-optical components.

Joni Dambre (S'99–M'04) was born in Ghent, Belgium, in 1973. She received the M.Sc. degree in electrotechnical engineering and the Ph.D.degree in computer engineering from Ghent University, Ghent, Belgium, in 1996 and 2003, respectively.

She is currently a Postdoctoral Researcher in the Department of Electronics and Information Systems, Ghent University. Her research interests include early evaluation of new interconnect techniques in digital systems.

Dr. Dambre is a member of ACM.

Jan M. Van Campenhout (M'95) was born in Vilvoorde, Belgium, on August 9, 1949. He received the degree in electromechanical engineering from the University of Ghent, Ghent, Belgium, in 1972, and the M.S.E.E. and Ph.D. degrees from Stanford University, Stanford, CA, in 1975 and 1978, respectively.

He is the Head of the Electronics and Information Systems Department (ELIS) in the Faculty of Engineering and serves on the Board of Directors of Ghent University. His research interests include the study and implementation of various forms of parallelism in information processing systems, currently focused upon the modeling and design of short-range parallel optical interconnects from a systems perspective. He is a Member of the PARIS research team of the ELIS department.

Prof. Van Campenhout is a member of Sigma Xi and K.VIV.

Hugo Thienpont (M'97–A'97) was born in Ninove, Belgium, in 1961. He received the graduate degree in electrotechnical engineering in 1984 and the Ph.D. degree in applied sciences in 1990 from the Vrije Universiteit Brussel (VUB), Brussels, Belgium.

He became a Professor with the Faculty of Applied Sciences, VUB, with teaching responsibilities in photonics, in 1994. He became Research Director of the Department of Applied Physics and Photonics at the VUB in 2000, and coordinates the activities of 35 researchers in the field of micro-optics and micro-photonics, where he became the Chair in 2004. Currently, he is a Coordinator of several basic research and networking projects such as the European Network of Excellence on Micro-optics "NEMO." Besides academic-oriented research projects, he manages micro-photonics-related industrial projects with companies like Barco, Agfa-Gevaert, Tyco, and Umicore. He has authored 70 SCI-stated journal papers and more than 250 publications in international conference proceedings. He has edited 15 conference proceedings and authored five chapters in books. He was the Invited Speaker at 40 international conferences and is co-inventor of ten patents.

Dr. Thienpont is a member of SPIE, EOS, IEEE-LEOS, and the OSA. He was the Guest Editor of several special issues of *Optics in Computing* and of *Optical Interconnects* for *Applied Optics* and the IEEE JOURNAL OF SELECTED TOPICS IN QUANTUM ELECTRONICS serves in technical and scientific program committees of photonics-related conferences, organized by international societies like SPIE, IEEE, OSA, EOS, and ICO. He is the General Chair of the SPIE Photonics Europe Conference to be held in Strasbourg, France, in 2006. He received the International Commission for Optics Prize ICO'99 and the Ernst Abbe Medal from Carl Zeiss in 1999. He was awarded the title of "IEEE LEOS Distinguished Lecturer" for serving as international lecturer from 2001 to 2003 on the theme "Optical Interconnects to Silicon Chips." He received the SPIE President's Award for dedicated service to the European Community, in 2005.