AMOS - Analysis and Modelling of an Optically Interconnected Commodity Cluster

J. F. Snowdon, G. A. Russell, K. J. Symington
Department of Physics, Heriot-Watt University, Edinburgh, UK
tel +44 0131 451 3026, fax +44 0131 451 3136, e-mail j.f.snowdon@hw.ac.uk

I. Goulay, P. Dew
School of Computing, University of Leeds, Leeds, UK
tel +44 0113 233 5471, fax +44 0113 233 3136 5432, e-mail iain@scs.leeds.ac.uk

Abstract: An optically interconnected Beowulf cluster has been modelled considering the optical, optoelectronic, electronic and algorithmic behaviour of the system. A Smart Pixel Array (SPA) layer was considered to alleviate the bandwidth mismatch between the commodity IO bus of the PCs and the high bandwidth in the optics. By using the SPA layer to provide load balancing for the system it has been shown that the optical bandwidth can be utilised.

1. Introduction

Free-space, integrated and fibre optics can provide communication bandwidths two to three orders magnitudes greater than that achievable using conventional electronic transmission lines – this being made possible by an increase in both the data-rate of individual channels and the number of channels available. Heriot-Watt has been investigating the construction of free space optical systems for many years in for example, the SCIOS [1] and SPOEC [2] projects. This current project has followed on from the work undertaken in SCIOS/SPOEC on the design and construction of parallel load balancing, routing and sorting algorithms that could utilise a smart pixel interface between the physical optics and silicon. It was recognized that further architectural system research was needed It is worth while pointing out that pervious studies have been generally restricted to comparing the raw bandwidth attainable in optic systems with that obtainable in copper. In this study we have (from the beginning) striven to show how optics can increase performance at system and algorithm levels, as well as at the raw interconnect stage.
The project was concerned with analysis and modelling of hybrid systems at a number of levels of abstraction. An important outcome of the project has been architectural models each with set of interlinked performance parameters representing combinations of the levels of abstraction. The architectural models are: the Abstract Model System View; the Optoelectronic Physical System View; and the Algorithmic View.

2. The Models

2.1 Abstract Model

The architectural model shown in figure 1 is based around a set of processing systems, with communication occurring via the high bandwidth interconnect. In the AMOS project, the processing systems are assumed to be commodity PCs, while the high bandwidth interconnect is implemented by means of optics (generally free-space, planar or integrated in our case studies). A smart-pixel based layer is used to interface between the processing system and the interconnect. This overcomes the mismatch in bandwidth associated with sending data to/from main memory on a PC and the much larger bandwidth associated with communicating data between nodes via the free-space optical interconnect (FSOI). Thus the smart pixel components may be thought of in a loosely analogous manner to impedence matching components in electrical circuits.

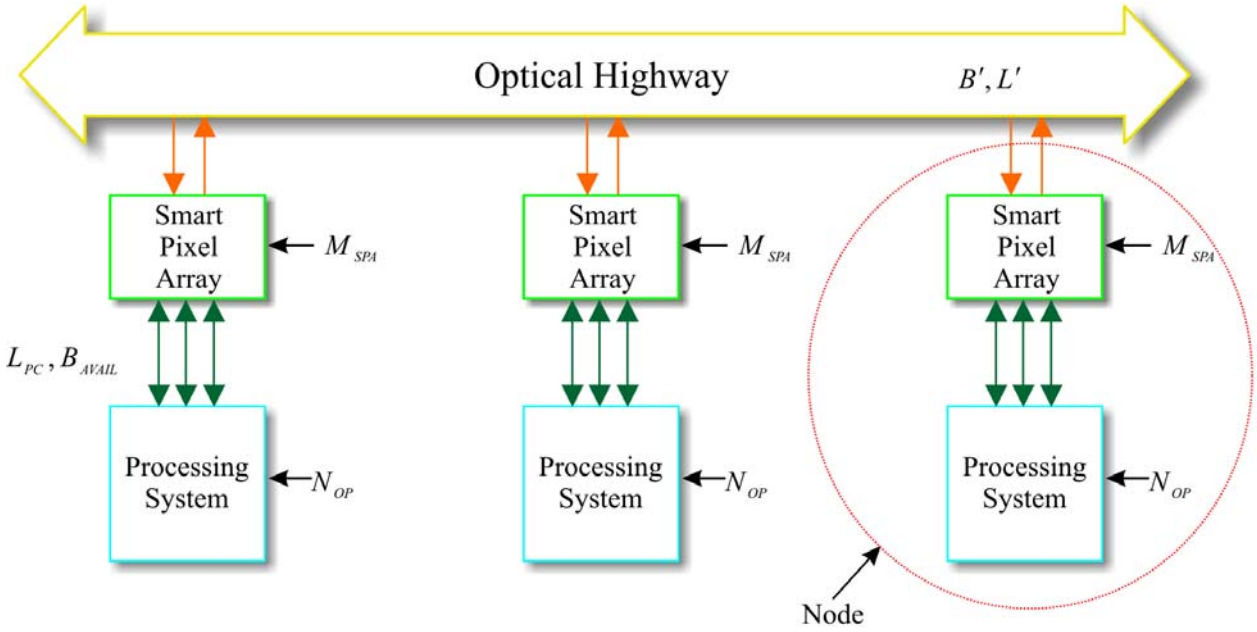

Figure 1: Optical Highway Abstract Model. Six parameters are used to define the optical highway. This diagram indicates to which logical part of the model each parameter is associated. Typically the High Bandwidth Interconnect is assumed to be capable of carrying more tham 100,000 channels of information. These channels are distributed at a reduced width of around 10,000 channels to the individual Smart Pixel Arrays associated with each node of the system.

Each PC can communicate with an associated smart-pixel array which has simple computational functionality and is capable of handling various tasks associated with the management of communication and (e.g.) load balancing.

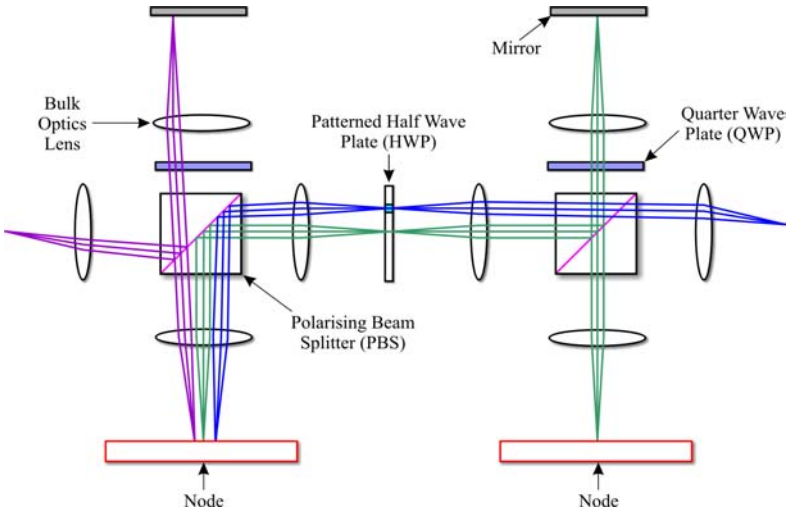

Figure 2: One method of implementing a freespace optical highway - a single stage is shown. In this PBS Optical Highway the polarising optics define a fixed network topology. Depending on the optoelectronic devices deployed (VCSELs, modulators, detectors etc.) the mirrors and/or the QWP must be patterned. Note that for some devices and geometries, additional beam delivery and combination optics may be required It is possible to replace the QWP with programmable polarisation devices such as liquid crystals, to make the system fully reconfigurable from software.

2.2 Optoelectronic Physical System View

Parameter values in the abstract model described above, are obtained by low-level analysis of the system components. An implementation known as the optical highway [3,4] is shown in figure 2. Polarisation is used to control which beams are routed into or out of each node. The characteristics of any optical component used to construct an optical highway places constraints on maximum achievable interconnectivity. These characteristics are *optical power*, *aberration* and *device densities* on the optoelectronic chip, the most significant of the three providing the final limit.

The maximum bandwidth of an emitter-detector pair is limited by the amount of optical power received at the detector, which is dependent on the efficiencies of the optics used.

As the number of stages increase, the spot size grows due to aberrations. Primary aberration is Spherical Aberration (SA). The number of channels available is half the area of the lens aperture $(\pi\phi^2)/4$ divided by the spot area (s^2).

Current system scalability, or the maximum number of nodes, is limited by the number of transmitters N_{tx} that can be fabricated on a single device. The number of operations free for computation on the processing node is dependent on the volume of communications between the SPA and main memory via the IO bus.

2.3 Algorithmic View

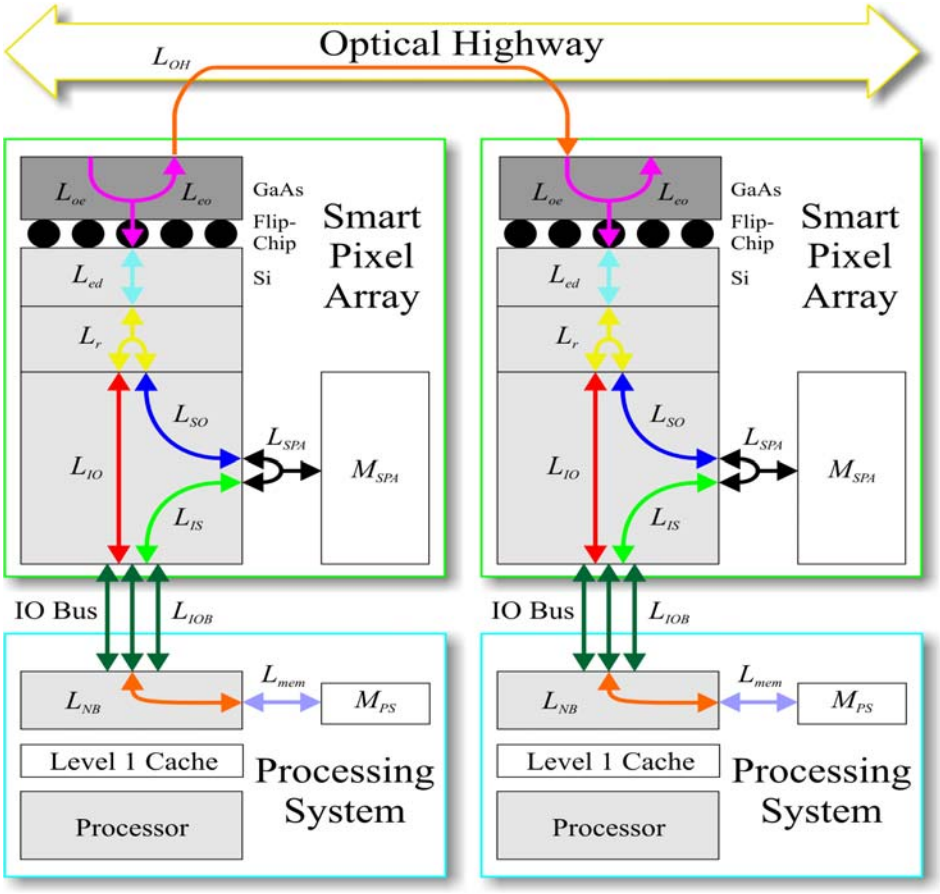

Figure 3: Node to Node Data Path Latency. The latency between selected source and destination nodes can be calculated by summing the above latencies for the appropriate path. Without describibg these latencies in detail, it may be noted that they parameterise real physical quantities for given system constraints and can be calculated and/or measured. These latency parameters (and their bandwidth couterparts) are generic and can be used with slight modification to describe any of the optoelectronic systems based on the various types of physical hardware we have studied. Similarly these parameters allow simulation at algorithm level. Thus the various levels of abstraction are linked through these parameters so that real performance data may be extracted at the top level based on experimentally validated physical hardware characteristics.

Algorithmic studies were based on the concept of using the computational functionality of the smart-pixel layer to manage communication between nodes, in such a way as to overcome the bandwidth mismatch between the layers.
In [5], an implementation of Bulk Synchronous Parallelism (BSP) is described, where each node(see figure 3) corresponds to a BSP processor. Data streaming between the PC and smart-pixel layer is used to tackle the bandwidth mismatch issue. Message combining for data to be communicated between any node-pair (i.e. BSP processor-pair) occurs in the smart-pixel layer. This results in an effective system bandwidth, B_{eff} associated with communication between BSP processors:

$$B_{eff} = \frac{1}{\frac{(2-s-r)}{B_{AVAIL}} + \frac{1}{B'}}$$

Here, for a given superstep, *r* is a lower bound on the fraction of data remaining in the smart-pixel layer associated with any PC when the local computation begins, while *s* is a lower bound on the fraction of data that resides in the smart-pixel layer for any PC, upon completion of the local computation. In general the BSP model lends itself well to the implementation of communication and load balancing tasks. In these, it is the the smart-pixel layer that is used to manage load balancing, relieving the PCs of this task. For heavily loaded PCs, some of the tasks are stored in the smart-pixel layer, allowing them to be rapidly communicated to other nodes, should they be required.

2.3 Discussion

In this brief summary we have outlined some of the models and methodology developed through the AMOS project. All of the parameter values and equations are solidly grounded in either previous demonstrator experiments or as part of the new experimental programme begun in AMOS to construct a fully reconfigurable optical highway. The algorithms used in performance measures make use of "standard" parallel processing techniques and models (such as BSP) and can thus be held to be generic. These models form the core of the next stage of development of optoelectronic computer hardware in which the next generation of demonstrators will be designed and built to tackle the currently burning issue of memory latency in commodity systems.

3. References

[1] Walker, A C, Tooley F A P, Taghizadeh M R, Desmulliez M P Y, Buller G S, Neilson D T, Prince S M, Baillie DA, Wherrett B S, Dines J A B, Wilkinson L C, Forbes M G, Snowdon J F, Stanley C R, Pottier F, Vass D G, Underwood I, Williams R, Sibbett W, Dunn M H, 'Development of an optoelectronic parallel data sorter based on CMOS/InGaAs smart pixel arrays' ,Technical Digest of International Conference on Optics in Computing (OC'97), Lake Tahoe, Nevada, OSA, 149-151 (1997).

[2] Gourlay,J.,Yang, T., Dines, J.A.B, Snowdon, J.F., Walker, A.C.
'Development of Free-Space Digital Optics in Computing',
Computer, **31**, 2, pp. 38- 44, (1998).

[3] J. A. B. Dines, J. F. Snowdon, M. P. Y. Desmulliez, D. B. Barsky, A. V. Shafarenko and C. R. Jesshope, "Optical Interconnectivity in a Scalable Data-Parallel System", *J. of Parallel and Distributed Computing*, vol. **41**, pp. 120-130, (1997)

[4] B. Layet, J.F.Snowdon, "Comparison of Two Approaches for Implementing Free-Space Optical Interconnection Networks", Optics Comms, **189**, pp. 39-46, (2001).

[5] D.B. Skillicorn, J.M.D. hill and W.F. McColl, 'Questions and Answers About BSP', Technical Report PRG-TR-15-96, Oxford University Computer Laboratory, (1996).