Interfacing Networks of Workstations with the "Fastest LAN in the World" A Work in Progress Report

Gordon Brebner, Steve Cusack, Tim Hopkins and Rob Pooley
Department of Computer Science
University of Edinburgh
Edinburgh EH9 3JZ
rjp@dcs.ed.ac.uk

February 27, 1998

Abstract

ECOLE is a project concerned with the implementation of highly efficient parallel computation on a network of generic workstations connected by a very high speed optical LAN. It is founded on research conducted at British Telecom Laboratories, which has resulted in SynchroLan — a multi-gigabit LAN billed as the fastest LAN in the world. A holistic approach to system design is taken, to identify all possible hardware and software improvements that might allow a workstation to harness fully the raw capacity of a multi-gigabit network. Two important features of the work are the inclusion of configurable hardware to introduce overall flexibility without compromising efficiency by more than a small amount, and the use of standard PCs as nodes in the network, making the work potentially applicable at low cost to existing organisations. The work benefits from the use of advanced techniques for modelling and simulation of experimental architectures. This paper explains how the ECOLE hardware is being designed.

1 Introduction

This paper is intended as a progress report, not a final statement. The project presented, ECOLE (Edinburgh Configurable Optical LAN Environment), is ongoing and futher results , including a hard design for the initial interface card will be available by the time of HOT I.

[0]This work is funded by the Engineering and Physical Sciences Research Council of the UK under grant GR/L82403 "Locally Distributed Shared Memory Computing using Optically Switched Networking"

The emergence of low cost and high performance workstations and interconnection networks points to the use of network of workstation (NOW) architectures as a compelling alternative to expensive parallel supercomputers. Ideally, NOW architectures would employ totally off-the-shelf workstations and networking. However, the deliberate generality built into these components leads to greatly reduced performance when attempting to compete with supercomputer speeds. To improve the situation, a holistic approach to systems is desirable, reviewing the efficiency of all system components: workstation and networking hardware and software, and also interaction using communication protocols. It is desirable that the resulting systems do not completely sacrifice generality for efficiency, to allow support for a variety of applications with different needs.

The project seeks to apply in tandem several currently active areas of advanced systems research at the University of Edinburgh and British Telecom, to provide novel solutions to the problem of implementing efficient, but programmable, NOW architectures based on off-the-shelf workstations interconnected by a very high speed optical local area network. This paper introduces the different research areas, and indicates how their results contribute to the overall aim; an initial outline of the project focussing on the rôle of active protocols, was presented at CANPC '98 in Las Vegas [6]; future papers will report on experimentation. The three most central research threads are:

- SynchroLan: an optical LAN demonstrated by British Telecom Laboratories in 1996/97, and rated as the world's fastest;
- Configware: novel solutions to the problem of integrating soft circuitry based on Field-Programmable Gate Arrays (FPGAs) into total systems, particularly with System Level Integration (SLI), also being investigated at Edinburgh University; and
- Active protocols: a flexible and verifiable approach to protocol construction and use, being developed at the Edinburgh University.

and the experimental application of these is much enhanced by research into:

- system modelling using Stochastic Process Algebras; and
- system simulation of high-bandwidth networks and hierarchical architectures.

To complete the picture, the combined research can draw upon the skills of researchers involved in past development of pioneering and parsimonious operating systems for both mainframe computers and workstation computers. There is also ready access to the state-of-the-art supercomputers (e.g., a Cray T3D) of the Edinburgh Parallel Computing Centre (EPCC), allowing comparative benchmarking, as well as providing conventional brute-force computational power for simulation studies.

A technical impetus for the ECOLE project comes just from the raw communications technology. SynchroLan is already able to deliver data rates of 2.5

Gbit/second to workstations, and the technology allows upgrading to at least 10 Gbit/second. The problem is that nothing like such rates can currently be delivered to the applications running on workstations. As many other researchers in this area have noted, a main bottleneck lies in the network interface, since current interfaces are not specifically designed for frequent, high bandwidth communication. Apart from this bottleneck, the communications protocols used over the network are another main source of inefficiency. Given good solutions to these two main problems, further gains can be had from holistic improvement of system architectures.

2 Configware and the network interface

Research at British Telecom laboratories has resulted in a single card that contains all of the components needed for the physical optoelectronic interface. The aim of the ECOLE project is to ensure that any extra hardware components needed to support an efficient, but flexible, NOW architecture can also be fitted on this single card, to be consistent with providing a low-cost solution. The card must interface to the system memory bus of the workstation, to provide the fast interaction with the main processor and the main memory that is needed to support application programs. The overall architecture is illustrated in Figure 1.

For ECOLE, low-cost workstations built from PC components will be used, with direct access to the system memory bus being made possible via the Accelerated Graphics Port (AGP) available with the latest Intel 440LX chipset for Pentium II PCs. This allows a point-to-point channel between a graphics controller and the system RAM, which can be exploited by a communications controller in ECOLE. The first generation of AGP interface (AGP-1x) offers a data rate of 266 MB/second, with a rate in excess of 1 GB/second expected by 1999 from the APG-4x. These rates are acceptable for the ECOLE prototyping work, the project being deemed a success if they can be fully harnessed.

In the short term, there are two approaches to ensuring that the extra hardware can be accommodated on the same card as the optoelectronic interface. The first is to design a custom ASIC once a solution has been selected from the possible design space; for example, this is the line pursued by the Utah Avalanche project [10]. The second is to include one or more Field Programmable Gate Array (FPGA) chips that can be configured with required circuitry. While the second option is more flexible, and allows prototyping, the performance is likely to be slower and the available number of gates will be rather smaller. For ECOLE, a longer-term perspective is being taken, strongly influenced by very active work on System Level Integration (SLI) at Edinburgh University.

With SLI (alternatively known as 'System On Silicon'), a number of major system components are located on a single chip. This is made possible by recent advances in deep submicron fabrication technologies. The complexity of SLI circuitry poses many challenges for the designer, and therefore much research is being carried out into new design methodologies and tools. One initiative is the

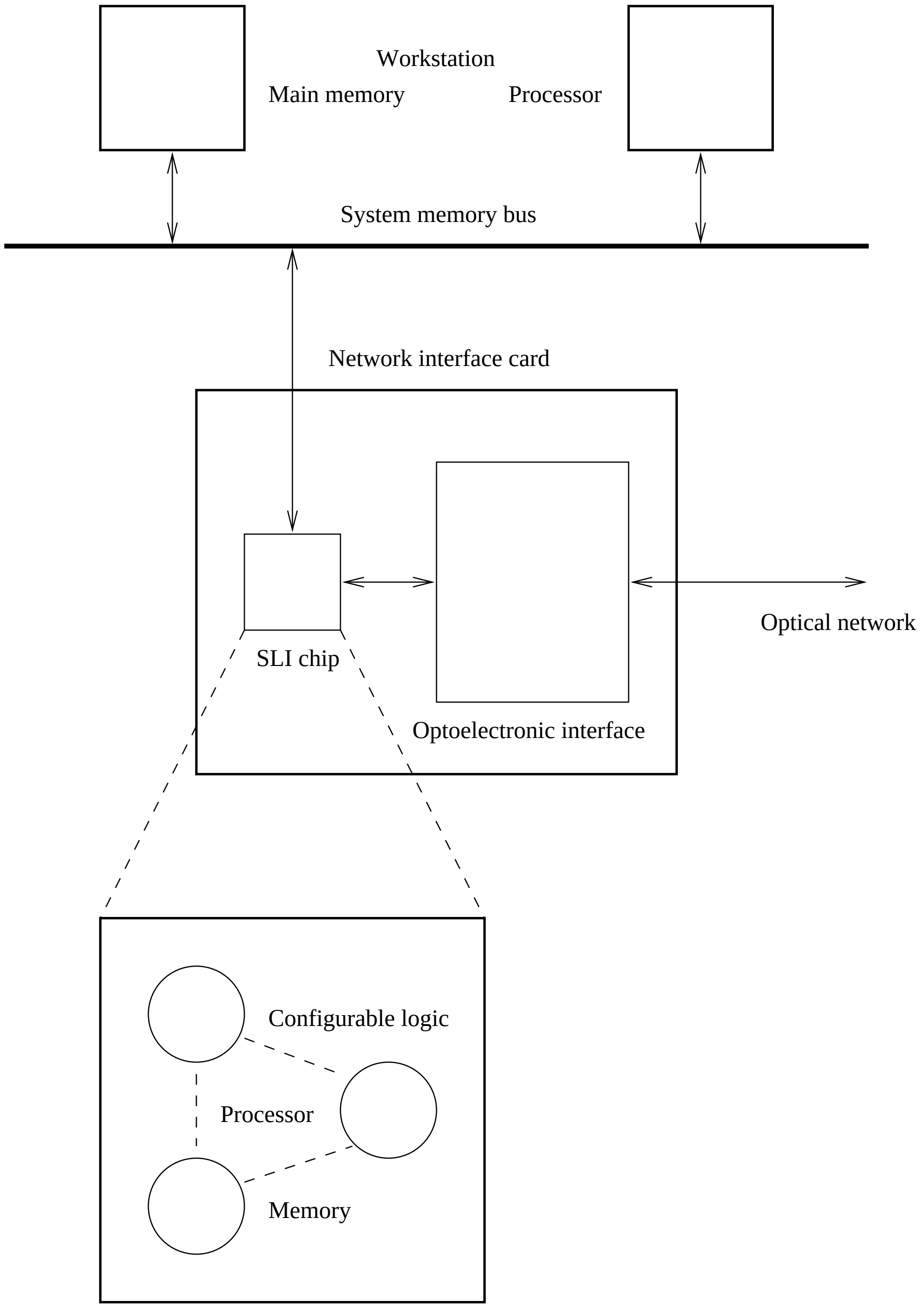

Figure 1: Overall network interface architecture

Virtual Socket Interface (VSI) [23], being developed within a consortium called the VSI Alliance; this is concerned with providing a standard way to integrate 'virtual components' from different sources, on a single chip. There are three major types of virtual component:

- processor cores (microprocessors, DSPs);
- memory devices; and
- configurable logic arrays.

and these can be integrated with more specialised types of component, such as analogue-digital convertors.

Apart from the obvious generic speed benefits of having processors, memories, etc., interfaced at chip level, the advent of SLI is particularly promising for the use of configware — configurable hardware — in this project. At present, configware is severely constrained by the pin interfaces to FPGA chips, or by the bus interfaces to boards containing FPGA chips. This penalises not only the speed-ups obtained when using configware as a configurable co-processor but also the time taken to reconfigure configware to change its co-processor function. SLI gives scope for rather more intimate, and fast, configware-processor-memory interactions.

For the ECOLE project, the presence of configware in the network interface is a central feature, in order to add flexible hardware support for communication protocols. This can be useful both for different network environments, for example, different latencies or reliabilities, or for different application requirements. The latter might be variations in model, for example, shared memory, message passing or bit pipe, or in quality requirements.

Previous configware research at Edinburgh has focused on how a configurable logic array can best be managed by a processor, resulting in general notions of 'virtual circuitry' — the configware analogue of virtual memory [3, 4]. The fruits of this work gain particularly from SLI, given an integrated processor running a lightweight virtual circuitry operating system. In addition the circuitry may implement accelerator components for communication protocols [11]. The general technology developed can be directly applied to particular protocols devised within the ECOLE project.

In summary, configware is not just included as a useful means of prototyping or implementing glue logic — the typical current uses of FPGA technology — although both of these are certainly of use here. It is included to provide support at hardware speeds for dynamically tailoring protocols to their operating environment. The use of SLI in the longer term should deliver satisfactory performance, as well as the means for enabling a single-card network interface.

3 Interfacing to SynchroLan.

SynchroLan supports up to 16 workstations connected as if they were connected by a crossbar network. Each workstation establishes an exclusive point-to-point connection to another and can have up to 2.5Gb/s of guaranteed bandwidth (irrespective of communications activity in other node pairs). Current implementation uses 16 pulses on the optical network to provide one logical bit for the communicating workstations, hence a bandwidth of 155Mb/s. Technically it is feasible to increase this per-channel bandwidth to around 600Mb/s without radically altering the construction of the current SynchoLan prototype network. Achieving the 2.5Gb/s maximum per-channel bandwidth is probably not an option with the current prototype.

Current BT hardware configuration (per workstation) consists of a standard 155Mb/s ATM NIC (network interface card), a 10Mb/s Ethernet (to provide out-of-band signalling capabilities) and a standard D/A converter card (to provide read channel selection). As signalling occurs out-of-band on relatively slow Ethernet, the current testbed is designed to show communication of large datasets requiring infrequent node pair changes.

The first choice for designing a bespoke SynchoLan network interface card (NIC) is that of choosing an appropriate channel bandwidth in the region of 155 - 600 Mb/s. This choice heavily influences the hardware required to construct such a card and the interface options for a standard PC. Other options will include the variety of signalling used (in or out of band) and how to provide the analogue signal to drive the SynchoLan data channel read select.

- Clearly, the higher the channel bandwidth selected the faster the circuitry on the NIC will have to operate. Especially with configurable logic, this could be a serious design limit.

- AGP provides a fast interface to the host computer's memory system and can achieve rates of 528MB/s in the current x2 mode. Even though a SynchoLan channel bandwidth of 600Mb/s is only 75MB/s, AGP is a dedicated bus to memory which doesn't have to contend with any other devices and hence would not provide a bottleneck. If the NIC was connected to the PCI bus of the PC the data to and from SynchroLan will have to contend with other PCI devices sharing the bus.

- Channel read select is performed by applying an analogue voltage to a microwave phase shifter within SynchoLan. This device has a nominal switching time of around 100ns in the current testbed and provides the upper bound on switch performance. As the current signalling is performed out-of-band over standard Ethernet then delays in using the OS and this hardware solution will dominate the time required to switch channels. Little knowledge exists of the latency incurred by having to frequently switch read channels in order to receive small amounts of data.

- In-band signalling would be possible by making hosts "listen" to a dedicated control channel and then perform appropriate read channel selection based on this. Problems include:

    1. the fact that each station is statically configured to write to a unique

channel and such a signalling system could "waste" a node (and there are only three nodes available in the current SynchoLan

2. There is no feedback mechanism readily available for acknowledging messages on the in-band system.

4 Proposed (possible) solutions

AGP is a very new interface standard and thus development tools and technologies are sparse. Designing an AGP interface will be a big job in itself and would require either fast programmable logic devices or the production of a custom ASIC for the job (the AGP bus runs at 66MHz). However it is possible to obtain a synthesizable AGP master core (as a piece of virtual chip intellectual property) which could possibly be realised in a fast programmable device or used in a custom ASIC design. (See http://www.vchips.com/agp/agp.html) Such developments are encouraging in that it will be likely that AGP prototyping board systems with ASICs to perform the interface functions will be available in the future. Bearing this in mind, it is feasible to start the design and testing of a suitable NIC based on the PCI bus now and then migrate the development to AGP once the AGP development kits become available.

For a project with a relatively short time to produce results, moving the NIC design to use the standard PCI bus is a good option. Companies such as AMCC (see http://www.amcc.com/) produce PCI prototyping boards which provide ASICs to perform the PCI interface operations. Such boards effectively provide a simpler data interface to the designer. However, as noted above, PCI boards have to share the PCI bus with any other devices on the system. This has the potential to be a bottleneck especially at higher SynchroLan channel bandwidths.

Data to and from the SynchroLan interface is bit-serial and thus serial⇔parallel conversion will have to be performed on the NIC. As 155Mb/s and 622Mb/s rates are getting more popular in items such as ATM NICs, obtaining suitable chipsets to perform the serial/parallel conversion should not be a major problem.

Some means of out-of-band signalling and a means of providing an analogue signal for the channel read select will have to be provided and could be realised with 100Mb/s Ethernet cards and standard PC DAC cards. Clearly it would be more "elegant" to have this functionality on the SynchoLan NIC but that will be a goal for later stages of the project.

5 A first design

We are currently developing a strategy to allow us to develop the system incrementally. Our objective is to prove the concept before we move to a fully featured design.

- Base the design on the PCI bus and use a commercial PCI bus based development board.
- Use the current 155Mb/s SynchroLan channel bandwidth to simplify design issues in the first prototype. This should make demands more bearable on programmable logic speeds and serial/parallel conversion chipsets.
- Use 100Mb/s Ethernet card to perform the out-of-band signalling required for channel read select. Providing a network dedicated to this function will help mimimise the latencies incurred.
- Use a standard PC DAC board to provide the analogue signal for the read channel selection.
- For the board itself, the grey area is in how much programmable logic to provide and what it should do. Clearly a design will need a datapath leading from the PCI bus to the serial/parallel converters required to drive SynchoLan. To add programmable functionality, buffers should be provided along the datapath and programmable logic which can access the contents of these buffers as well and interface the logic providing the PCI bus gateway. Such a deployment of programmable logic should provide the flexibility to perform functions such as memory address remapping which might be required by the shared memory model implemented in the software for the system.

6 Related work, and the future ECOLE work programme

The ECOLE project is one of many that are looking the problem of achieving fast parallel computation using a network of workstations. The adoption of a holistic approach is a central feature, something that is relatively uncommon in other research, one notable exception being the Utah Avalanche project [10]. A major physical difference from Avalanche and other work is that a next-generation general purpose LAN forms the interconnection mechanism, rather than an interconnect such as Myrinet [2]. This is to ensure that workstations can still be used in their more conventional independent manner, as well as just in a NOW architecture.

Another physical aspect of ECOLE is that normal workstations are used, in contrast to more specialised architectures of other work, such as the MIT Alewife [1] or the Stanford FLASH [13] architectures. Indeed, the choice of a PC-based architecture is targeted at the most commonly available workstation — and one that can deliver much-improved performance these days, rather than the HP multiprocessor workstations or Sun SPARCstations used in other work, for example Avalanche, HP Lab's Hamlyn [8] or Illinois Fast Messages [18]. The aim is to add just one extra card, responsible both for physical network interfacing and for tailored hardware/software functions inside the network interface.

A final innovative physical aspect is the inclusion of configurable logic as an integral part of the network interface design, to provide scope for dynamic configurability of hardware. This has not been a feature of other NOW research, due to the relative inflexibility of much current FPGA hardware and to the speed penalties compared with custom ASICs. ECOLE seeks to benefit from emergent improvement in configurable logic technologies that will greatly reduce current performance penalties.

Reduced-weight communication protocols are also important to the holistic approach of the ECOLE project. These have also featured in other NOW research, for example, the DD protocol of the Avalanche project [10] or the PRP protocol of the PARMA project [17]. Gains over more conventional protcols, such as TCP/IP, come from introducing application specificity and from simplied physical channel assumptions, such as error freedom or perfect sequencing. For ECOLE, the framework of active protocols offers a way of experimenting with new stripped-down protocols to find those that fit the application/physical environment efficiently. Moreover, the configurable interface gives a means to avoid the need of selecting a single protocol for all uses.

The work plan of the ECOLE project involves extensive use of modelling and simulation of design alternatives, before any prototyping using PC-based workstations and the SynchoLan optical LAN. The first stage of the holistic approach involves an examination of high-level architectural design alternatives, together with alternative communication protocols. The second stage involves examining major system components. Two particular aspects of this are an investigation of the uses of reconfigurable logic as part of an overall architecture, and an investigation of improved software architectures to support applications. The third stage involves prototyping of chosen designs, by which time it is intended that SLI chip facilities will be available to allow a complete system implementation.

Acknowledgement

The authors thank British Telecom Laboratories, particularly David Cotter, for inspiring this work, and for access to experimental optical network prototypes.

References

[1] Agarwal A., R. Bianchini, D. Chaiken and K. Johnson, "The MIT Alewife Machine: Architecture and Performance", Proc. 22nd Annual International Symposium on Computer Architecture, June 1995, pp.2–13.

[2] Boden N. et al, "Myrinet — A Gigabit-per-second Local Area Network", IEEE MICRO Vol 15 No 10, February 1995, pp.29–36.

[3] Brebner G., *Computers in Communication*, London:McGraw-Hill 1997.

[4] Brebner G., "A virtual hardware operating system for the Xilinx XC6200", Proc. 6th International Workshop on Field Programmable Logic and Applications, Darmstadt, September 1996, Springer LNCS 1142, pp.327–336.

[5] Brebner G., "The Swappable Logic Unit: a Paradigm for Virtual Hardware", Proc. 5th Annual IEEE Symposium on Custom Computing Machines, Napa, April 1997, IEEE Computer Society Press.

[6] Brebner G. and Pooley R. "ECOLE: A Configurable Environment for a Local Optical Network of Workstations" in Proceedings of the Workshop on Communication and Architecture for Network-Based Parallel Computing – CANPC '98, Las Vegas, February 1998, published as Springer LNCS 1362

[7] Brebner G. and S. Haeck, "Active Protocols", submitted to SIGCOMM'98.

[8] Buzzard G. et al, "An Implementation of the Hamlyn Sender-managed Interface Architecture", Proc. 2nd Symposium on Operating System Design and Implementation, October 1996.

[9] Cotter D., "Ultrafast Computer Networking" (Invited Paper), Optics in Computing '97, Incline Village, NV USA, 18–21 March 1997.

[10] Davis A., M. Swanson and M. Parker, "Efficient Communication Mechanisms for Cluster Based Parallel Computing", Proc. 1st Internation Workshop on Communication and Architectural Support for Network-Based Parallel Computing, San Antonio, February 1997, Springer LNCS 1199, pp.1–15.

[11] Donlin A., "A Dynamically Self-Modifying Processor Architecture and its Application to Active Networking", Internal report, Department of Computer Science, University of Edinburgh, September 1997.

[12] Gilmore S. and J. Hillston, "The PEPA Workbench: A Tool to Support a Process Algebra-based Approach to Performance Modelling", Proc. 7th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation, Vienna, May 1994, Springer LNCS 794.

[13] Heinrich M. et al, "The Performance Impact of Flexibility in the Stanford FLASH Multiprocessor", Proc. 6th Symposium on Architectural Support for Programming Languages and Operating Systems, October 1994, pp.274–285.

[14] Hillston J., A. Opdahl and R. Pooley, "A case study using the IMSE experimentation tool", Proc. 3rd Conference on Advanced Information Systems Engineering, Trondheim, Springer LNCS 498, pp.284–306.

[15] Ibbett R., P. Heywood and F. Howell, "HASE: A Flexible Toolset for Computer Architects", Computer Journal, Vol 38 No 10, 1995.

[16] Lucek J., D. Cotter, K. Smith and P. Gunning, "Ultrafast Photonic Data Networks" (Invited Paper), IEEE LEOS'96 Annual Meeting, Boston, 18–21 November 1996.

[17] Marenzoni et al, "An Operating system Support to Low-overhead Communications in NOW Clusters", Proc. 1st Internation Workshop on Communication and Architectural Support for Network-Based Parallel Computing, San Antonio, February 1997, Springer LNCS 1199, pp.130–143.

[18] Paikin S., M. Lauria and A. Chien, "High Performance Messaging on Workstations: Illinois Fast Messages (FM) for Myrinet", Proc. Supercomputing '95, San Diego, December 1995.

[19] Pooley R., *Formalising the Description of Process Based Simulation Models*, Ph.D. Dissertation, University of Edinburgh, 1995.

[20] Pooley R., "Integrating Behavioural and Simulation Modelling", Proc. 8th International Conference on Techniques and Tools for Computer Performance Modelling, Heidelberg, September 1995.

[21] Pooley R., "Exploiting Functional Properties to Reduce Complexity in Simulating ATM", Proc. Telecommunication, Distribution and Parallelism (TDP '96), La Londe les Maures, France, June 1996, pp 23–34.

[22] Tennenhouse D. and D. Wetherall, "Active Networks", Proc. 15th Symposium on Operating System Principles, December 1995.

[23] VSI Alliance Architecture Document, VSI Alliance 1997.