AMOS - Analysis of an Optically Interconnected Beowulf Cluster

G. A. Russell, K. J. Symington, J. F. Snowdon

Department of Physics, Heriot-Watt University, Edinburgh, UK

I. Goulay, P. Dew

School of Computing, University of Leeds, Leeds, 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.

Introduction

Modern commodity processors can provide impressive performance (>2 Gflops) for little cost (<£50), therefore, clustering large numbers of such processors to form a parallel processing machine has become increasingly popular [1].

The main drawback with such Beowulf cluster systems is the large overhead on interprocessor communication due to the high latency and low bandwidth of the interconnection network used. An optical interconnection network can provide lower latency by being more connected and thus have less routing costs, collision penalties etc. It is less clear that the bandwidth can be used as the IO bus bottleneck remains.

If network level functionality is integrated with the optoelectronics to form a Smart Pixel Array (SPA) [2] low level functions such as load balancing can be performed without using the PC IO bus.

Hardware Model

The model of the hardware assumes an optical highway (OH) as described in [3]. This is a series of 4f image relays utilising polarisation to route channels into and out off the OH. The SPAs at each node were assumed to be optical field programmable gate arrays (OFPGAs), arrays of optical components flip-chip bonded to reconfigurable silicon.

With these assumptions, models of the optical power budget, aberration limits and semiconductor densities have been formulated. As an example Equation 1 below shows the power limited bandwidth of the OH, B'.

$$B' = \frac{\xi_{ed} \cdot C_w \cdot P_{VCSEL}^{\,2} \cdot \xi^{2 \cdot h_{\max}}}{8 \cdot NEP^2}$$

Equation 1

ξ_{ed} is the encoding efficiency, C_W the channel width, P_{VCSEL} the VCSEL optical power, ξ the efficiency of one optical stage, h_{max} the maximum number of stages in one hop and NEP the noise equivalent power of the detector. Values for these can be taken from available systems or modelled further. Similar latency models were considered.

The model for the PC assumes that the processor is stalled for at least the amount of time data is being transferred to and from main memory. Equation 2 shows the number of operations (N_{op}) available to a processor able to carry out N_{max} operations if B_{act}

bandwidth is requested. B_{per} and B_{mem} are the bandwidths of other peripheral devices and memory bus and p_{ov} and N_{ov} quantify the overhead.

$$N_{op} = N_{\max} - N_{\max}\left(\frac{(B_{per} + B_{act})p_{ov}}{B_{mem}}\right) - N_{ov}$$

Equation 2

Algorithmic Model

The algorithm chosen for this modelling was a random stealing load-balancing problem. In such an algorithm a divide and conquer type problem is distributed between the processors. As the queue of jobs (stack) on a processor reaches zero the processor randomly chooses a processor and takes excess jobs from it. In on our architecture the stack is held in the SPA layer so messages can be passed between stacks at high speed without stalling the processor.

The algorithm has been simulated using a discrete event analysis with the hardware models discussed above.

Results

Analysis and modelling has shown that the performance of the random stealing load-balancing algorithm above as well as sorting by random sampling and reducing-sum operations [4] can be improved. The hardware model suggests that highly connected networks are possible at high bandwidths and that it is the power budget, not aberration limits or semiconductor densities, that dominates.

Further Work

Two new projects have or are about to start continuing on from this work, POCA and HOLMS. Both projects move the work out of modelling and into real systems. POCA (Programmable Optoelectronic Computing Architecture) is an EPSRC/OSI funded project to investigate OFPGA technologies between Heriot-Watt and Edinburgh Universities. This is due to start in October. HOLMS (High-speed OptoeLectronic Memory Systems) is an European Union funded project to integrate a high bandwidth, low latency optical link into the processor-memory bus of a computer system. Heriot-Watt University and ETH in Zurich led this project.

Conclusion

This research has shown that an optical interconnect can be used in a commodity cluster system. However, the get the maximum benefit an extra layer of functionality is required at the SPA level.

References

[1] The Beowulf Project website –www.beowulf.org

[2] A.C. Walker, T.-Y. Yang, J. Gourlay, J.A.B. Dines, M.G. Forbes, S.M. Prince, D.A. Baillie. D.T. Neilson, R. Williams, L.C. Wilkinson, G.R. Smith, M.P.Y.Desmulliez, G.S. Buller, M.R. Taghizadeh, A. Waddie, I. Underwood, C.R. Stanley, F. Pottier, B. Vogele, and W. Sibbett. "Optoelectronic Systems Based On InGaAs-Ccomplementary-Metal-Oxide-Semiconductor Smart-Pixel Arrays and Free-Space Optical Interconnects." *Applied Optics*, Vol. 37, No. 14, 10 May 1998.

[3] B. Layet, J.F. Snowdon, "Comparison Of Two Approaches For Implementing Free-Space Optical Interconnection Networks," *Optics Communications,* Vol. 189, pp. 39-46, March 2001.

[4] G. Russell , J. Snowdon, T. Lim, I. Gourlay, P. Dew, "Modelling Of Optical Interconnects For Parallel Processing," Conference Publication, *Third Conference on Postgraduate Research in Electronics, Photonics, Communications and Software (PREP 2001),* EPSRC, April 2001.