Predicting the Performance of Reconfigurable Interconnects in Distributed Shared-Memory Systems

W. Heirman*,[1],[2], J. Dambre*,
I. Artundo†, C. Debaes†, H. Thienpont†,
D. Stroobandt*, J. Van Campenhout*

* *ELIS, Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium*
† *TONA, Free University of Brussels, Pleinlaan 2, B-1050 Brussel, Belgium*

ABSTRACT

Reconfigurable optical interconnect technologies will allow the fabrication of run-time adaptable networks for connecting processors and memory modules in shared-memory multiprocessor machines. Since switching is typically slow compared to the memory access time, reconfiguration exploits low frequency dynamics in the network traffic patterns. These are however not easily captured in tools employing statistical traffic genereration, which is commonly used for fast design space exploration. Here, we present a technique that can predict network performance based on actual traffic patterns, but without the need to perform slow full-system simulations for every parameter set of interest. This again allows for a quick comparison of different network implementations with good relative accuracy, narrowing down the design space for more detailed examination.

KEYWORDS: prediction, reconfiguration, interconnect, shared-memory

[1] E-mail: wim.heirman@elis.UGent.be

[2] This paper presents research results of the PHOTON Inter-university Attraction Poles Program (IAP-Phase V).

1 Introduction

The electrical interconnection networks connecting the different processors and memory modules in a modern large-scale multiprocessor machine, are running into several physical limitations [Mill97]. In shared-memory machines, where the network is part of the memory hierarchy, the ability to overlap memory access times with useful computation is severely limited by inter-instruction dependencies. Hence, a network with high latencies causes a significant performance bottleneck.

It has been shown that optical interconnection technologies can alleviate this bottleneck [Benn05]. Mostly unhindered by crosstalk, attenuation and increased capacitive bus load, these technologies will soon provide a cheaper, faster and smaller alternative to electrical interconnections, on distances from a few centimeters upward. Massively parallel inter-chip optical interconnects [Scha05] are already making the transition from lab-settings to commercial products.

Optics may provide another advantage: the optical pathway can be influenced by components like steerable mirrors, liquid crystals or diffractive elements. In combination with tunable lasers or photodetectors these components will enable a runtime reconfigurable interconnection network [Artu06] that supports a much higher bandwidth than what is achievable through electrical reconfiguration technology. From a viewpoint higher in the system hierarchy, this would allow us to redistribute bandwidth or alter the network topology such that node-pairs that communicate intensely have a direct connection. Since there is no longer interference from other traffic streams, this results in higher available bandwidth and lower packet latency (which would otherwise be increased by congestion and switching delays).

However, the switching time for most of these components is such that reconfiguration will necessarily take place on a time scale that is significantly above that of the individual memory accesses. The performance of such a network therefore strongly depends on the temporal behavior of the interprocess data transfer patterns. These temporal aspects of the workload therefore need to be incorporated at an early stage in the design flow for these networks. Typically these early stages consist of large-scale design space explorations, requiring a fast estimation method for the performance of thousands of network implementations. Fast methods, such as those employing statistical traffic generation, are usually good enough for modeling static traffic patterns. This can suffice to evaluate non-reconfigurable networks. The temporal behavior of the network traffic, on which our reconfigurable networks depend, is however not sufficiently modeled in existing traffic generators, which precludes their use for our purposes. Execution driven simulation on the other hand perfectly models all temporal traffic behavior, but is very detailed and usually too slow (up to several days for a single simulation) to be practical for large-scale explorations.

Our method, which is an extension of the methode previously described in [Heir06], builds upon another well known technique, called *trace-driven simulation*. Here the network traffic is recorded while doing a single execution-driven simulation. This recorded traffic flow can be played back on a large number of different network architectures. Since the reconfiguration of the network manifests itself only as a periodical modification of the network topology, it is possible to determine the distance a packet will travel in the new network as compared to the old network. In this approach, it is not necessary to do a slow, cycle-by-cycle simulation of the flow of packets. In contrast, it allows us to predict new packet latencies and the resulting network performance, with an acceptable degree of accuracy. As such, our method allows for a rapid exploration of the design space of an interconnection network.

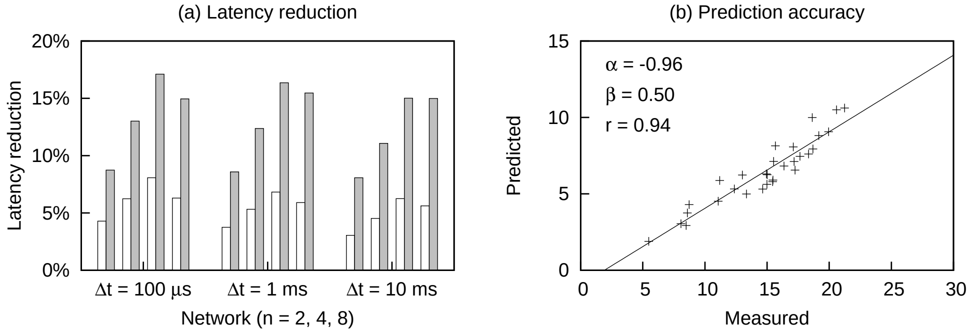

Figure 1: Left: latency reduction after adding elinks, estimated (white) and measured (gray) values, for 3 reconfiguration intervals, and 3 networks (with 2, 4 and 8 elinks). Right: estimated versus measured latency reduction for a variety of network implementations. α, β and r represent a linear regression of the form $P = \alpha + \beta \times M$ and its correlation coefficient.

2 A reconfigurable network architecture

Our network architecture starts from a base network with a fixed topology. In addition, we provide a second network that can realize a limited number of connections between arbitrary node pairs – these will be referred to as *extra links* or *elinks*. Reconfiguration takes place at specific intervals, the length of each interval being a (fixed) parameter of the network architecture. Traffic is observed by a reconfiguration entity during the course of an interval, and total traffic between each node pair is computed. At the end of each interval, new positions of the elinks are determined, such that node pairs that exchanged the most data in the previous interval will be 'closer together': the distance, defined as the number of hops a packet sent between the pair must traverse in the new network topology, is minimized. This way, a large percentage of the traffic has a short path and a correspondingly low uncongested latency, also congestion is lowered because heavy traffic is no longer spread out over a large number of links.

The physical implementation of this network can be done in a number of different ways. The light path can be influenced directly by using switchable mirrors, or indirectly through a combination of wavelength tunable lasers, selective receivers and broadcasting elements. Depending on the technology used, reconfiguration can take from tens of microseconds up to several milliseconds. More about reconfigurable optical networks can be found in [Artu06], which gives an overview of different technologies and their expected performance.

3 Predicting network performance

Our performance prediction starts by doing one full-system, execution-driven simulation of each benchmark. Here the memory accesses and network traffic are recorded. Next we perform a post-processing step on this data that is parameterized on the properties of the reconfigurable network under investigation, and where the performance of this candidate network is predicted. This second step can be done much faster than doing an execution-driven simulation, taking minutes of computation time as opposed to hours, while the data of the initial simulation can be reused for a large number of candidate networks.

In figure 1(a), we show the results of our prediction method. We have done the initial execution-driven simulation using a non-reconfigurable network, and used our model to estimate the improvement in memory access latency for a number of reconfigurable networks (white bars). Next, we ran a number of new execution-driven simulations, one for each reconfigurable network, measuring the actual performance (gray bars) to asses the accuracy of our prediction. This was done for networks with different numbers of elinks ($n = 2, 4, 8$) and reconfiguration intervals ($\Delta t = 100\ \mu$s, 1 ms, 10 ms). We used 5 different benchmark applications from the SPLASH-2 suite (Barnes, Cholesky, FFT, Ocean.cont and Radix) and averaged the results per network.

Since our prediction method makes some simplified assumptions about the network (for instance, congestion is not modeled), the *absolute* predictions shown in figure 1(a) are always about a factor of 2 too low. However, the *relative* prediction accuracy over different network parameters, which is the most important value when comparing different suggested network implementations, is much better. To show this, figure 1(b) plots the predicted versus the actual latency improvement for a number of different networks. A linear regression of the form $P = \alpha + \beta \times M$ is calculated (with P and M the predicted and measured latency reduction, respectively). The correlation coefficient r turns out to be high, so a strong, linear correlation exists between measurement and prediction. Therefore, although our method does not allow one to accurately predict the performance of one specific network, it can be used to very quickly compare different proposals for network parameters. This makes it a very useful tool for design-space explorations, where the optimum solution needs to be found from a large collection of candidate networks.

References

[Artu06] I. ARTUNDO, L. DESMET, W. HEIRMAN, C. DEBAES, J. DAMBRE, AND J. VAN CAMPENHOUT. Selective Optical Broadcast Component for Reconfigurable Multiprocessor Interconnects. *IEEE Journal on Selected Topics in Quantum Electronics: Special Issue on Optical Communication*, 12(4):828–837, July 2006.

[Benn05] A. BENNER, M. IGNATOWSKI, J. KASH, D. KUCHTA, AND M. RITTER. Exploitation of optical interconnects in future server architectures. *IBM Journal of Research and Development*, 49(4/5):755–776, 2005.

[Heir06] W. HEIRMAN, J. DAMBRE, I. ARTUNDO, C. DEBAES, H. THIENPONT, D. STROOBANDT, AND J. CAMPENHOUT. Predicting Reconfigurable Interconnect Performance in Distributed Shared-Memory Systems. *Integration, the VLSI Journal: Special Issue on System Level Interconnect Prediction*, 2006. To appear.

[Mill97] D. MILLER AND H. OZAKTAS. Limit to the bit-rate capacity of electrical interconnects from the aspect ratio of the system architecture. *Journal of Parallel and Distributed Computing*, 41(1):42–52, 1997.

[Scha05] L. SCHARES AND OTHERS. Terabus – A Waveguide-Based Parallel Optical Interconnect for Tb/s-Class On-Board Data Transfers in Computer Systems. In *Proceedings of the 31st European Conference on Optical Communication (ECOC 2005)*, volume 3, pages 369–372, Glasgow, Scotland, September 2005. The Institution of Electrical Engineers.