Requirements of Low Power Photonic Networks for Distributed Shared Memory Computers

P.M.Watts, N. Barrow-Williams, S.W.Moore
Computer Laboratory, University of Cambridge, JJ Thompson Avenue, Cambridge, CB3 0FD, United Kingdom
philip.watts@cl.cam.ac.uk

Abstract: Using the PARSEC benchmark suite running on a 32-core Distributed Shared Memory computer system, photonic component and interconnection network characteristics required for reduced overall power consumption are determined.
OCIS codes: (200.4650) Optical Interconnect; (200.6715) Switching

1. Introduction

Power consumption of processor chips has become critical. With the rapid development of silicon photonics [1], polymer waveguides in standard PCBs [2], and 3D integration [3] technologies in recent years, photonic networks highly integrated within a chip multiprocessor (CMP) have been envisaged [4]. Recent advances point toward 2 - 4 DRAM layers on top of a multi-core substrate, reducing the power required to talk to DRAM, but only scalable to relatively small memory systems (≈1GB). Such modules could then be used to produce larger systems, with each module adding computation, memory and communication capabilities (Fig. 1). Under these conditions, it makes sense to use a distributed shared memory (DSM) architecture [5] in which each core has local memory shared through a global address space. Communication takes place directly between the local memories of each core with message sizes of the order of 8 – 32 B (i.e. one cache line). This is an extremely challenging application for photonic interconnect in which, due to the lack of practical optical memory, end-to-end paths must be setup for small packets. This paper investigates the requirements of photonic switching for reducing overall power consumption in multichip DSM machines based on an analysis of memory traces from a simulated 32-core DSM machine running the PARSEC benchmark suite. While future DSM machines of this type may consist of thousands of cores, the analysis of 32 cores (at the limits that can be modeled using full-system software simulation) gives an insight into the traffic characteristics of this class of computers.

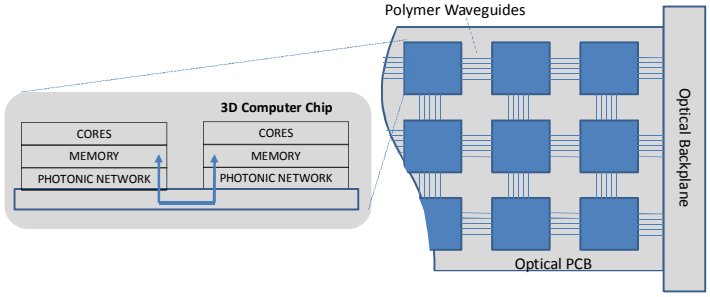

Fig. 1. Network of 3D integrated chip multiprocessors with distributed shared memory communications

We consider a centralized photonic switch thus minimizing the number of optical switching elements and hence power consumption compared with distributed switching. Setting up circuits on a per message basis has been shown to reduce energy efficiency where message sizes are low (10s to 1000s of bytes) [6] as in the DSM case. Instead we consider two approaches (1) A dual network consisting of a photonic circuit switch for large data flows combined with an electronic packet switched network [7]; (2) time division multiplexed (TDM) access to a switch fabric using short fixed time slots [8-10].

2. Methodology

In this work, we assess the benefits of photonic networks for DSM systems by analysis of the communication patterns generated by running the PARSEC benchmark suite [11] on a simulated 32 core x86 system running linux. PARSEC contains 12 algorithms covering financial, engineering and scientific applications designed for parallel processing on future multicore systems. Full details of the simulation parameters and cycle accurate memory trace generation were described in [12]. All communication between cores is memory to memory using 32B packets. A zero-latency infinite bandwidth crossbar interconnect was used to assess the interconnect requirements without simulating specific network schemes.

For the circuit switching case, the memory traces were divided into time intervals and the circuit configuration which maximizes the traffic over the circuit switch was determined. This approach gives an upper bound on the

benefits of introducing circuit switching without assuming any particular scheduling scheme. Time intervals from 300 clock cycles (120 ns at 2.5 GHz) up to the full algorithm run time were used. For each algorithm in the PARSEC benchmark, the total traffic carried on the circuit switch, total switch on-time and total switching activity were recorded. A lower bound on the total energy of the network can be given by:

$$E_{network} = N_{circuit} \cdot E_{end} + E_{switch\_total} + (N - N_{circuit}) E_{pkt} \quad (1)$$

where E_{end} and E_{pkt} are the photonic end-point and electronic packet network energy per packet, E_{switch_total} is the total switch energy, N is the total number of packets and $N_{circuit}$ is the total number of packets routed onto the circuit switch. For the TDM case, it is assumed that all traffic is carried over the photonic network ($N = N_{circuit}$). For switch technologies which operate using current injection, $E_{switch} = t_{switch} \cdot P_{switch}$ where t_{switch} is the total length of all circuits and P_{switch} is the on-state power per path. For switches which present a capacitive load, $E_{switch} = 2 \cdot N_{switch} \cdot E_{switch}$ where N_{switch} is the total number of switching operations and E_{switch} is the energy of a single switching operation. The photonic network is compared with an electronic 2D packet switched mesh network with total energy, $E_{network} = N \cdot E_{pkt}$. The energy per packet is given by:

$$E_{pkt} = E_{router} \cdot [H_{\min} + 1] + E_{link} \cdot H_{\min} \quad (2)$$

where E_{router} and E_{link} are the energies per packet of the routers and electronic links respectively and H_{min} is the average minimum hop count for the network. A 6.25 Gb/s chip-to-chip electronic transceiver specifically designed for low power in 90 nm CMOS consumed 2.1 pJ/bit [13] while a 5-port router design optimized for low power and speculative single cycle routing in 90 nm CMOS consumed 0.46 pJ/bit [14]. We can therefore estimate E_{router} = 118 pJ/pkt and E_{link} = 537 pJ/pkt giving E_{pkt} = 2.9 nJ/pkt for the 32 node network. We conservatively assume that the photonic end-point power per packet is equal to that of the electronic transceiver (537 pJ/pkt), as (while acknowledging that future integrated devices will reduce power), many functions of a transceiver (e.g. SERDES, clock recovery) are required in both cases.

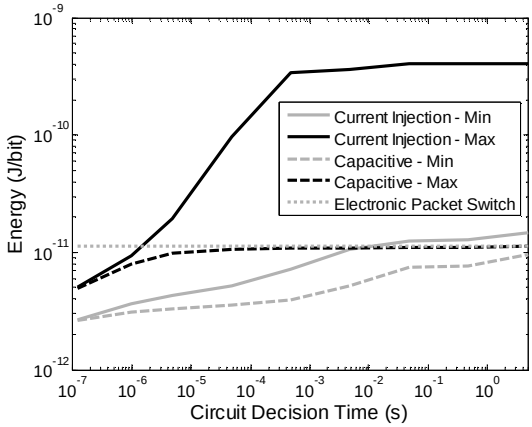

Fig. 2. Energy per bit for combined circuit switching and electronic packet switching

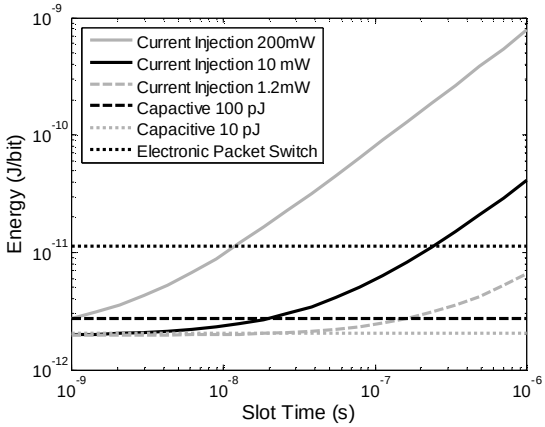

Fig. 3. Energy per bit in the TDM case

3. Results

Fig. 2 shows the minimum and maximum average energy per bit across all 12 algorithms in the PARSEC benchmark for both current injection and capacitive switches as the time on which circuit decisions are made is varied. The current injection case assumes an on-state power of 1.2 mW, consistant with a Clos switch using ring resonators sized to allow transmission of 16 wavelengths with 250 GHz spacing [15]. For the capactive case, a switching energy of 100 pJ is assumed, consistant with electro-optic Mach-Zehnder switches [16]. However, lower energy capacitive switches do not substantially change the results as the overall energy is dominated by the photonic end point and packet switch energy. Higher power switch technologies such as SOAs substantially increase the overall energy in the current injection case. It can be observed that circuit decisions must be made on microsecond timescales or less to obtain significant and consistant energy advantage over packet switching. No significant energy advantage can be obtained in the static case. These results are not significantly changed by using 2 or 3 nodes per core. It is notable that for the shortest circuit decisions times, the mean number of packets per circuit interval is small (varying over 1.0 - 3.2 between algorithms). The maximum peak circuit bandwidth observed is a challenging but achievable 97.3 Gb/s (for the x264 algorithm). The presence of the circuit switch allows the peak electronic bandwidth requirement to be reduced by 23 – 74 % depending on algorithm considered.

Fig. 3. shows the average energy per bit for TDM as the slot time is varied. If we assume a minimum slot time for the 32B packet of 6 ns or 15 clock cycles (processor clock rate of 2.5 GHz, optical bit rate of 10λ x 10 Gb/s, 50% coding overhead and 1 ns switching time), the effective bandwidth per port is 43.7 Gb/s. Ideally current injection switches require on-state power of around 10 mW on this timescale. However, SOA devices (200mW per path [17]) can also have energy advantage over packet switches. In the capacitive case, switching energies of around 10 pJ ensure that the overall energy is dominated by the end point. Low latency in the TDM case results from inter-arrival times (IAT) being substantially less than the slot time. As shown in Fig. 4, all algorithms have less than 5% of packets with IAT below 15 clock cycles with the exception of x264 and streamcluster. However, it is notable in x264 that a very high percentage of adjacent packets are to the same destination. An analysis of the number of packets in each 300 cycle interval (Table 1) shows that while peak bandwidth requirements per port can be very high (e.g. 112 packets per interval for Streamcluster = 238 Gb/s), 90 % of intervals have no more than 8 packets with x264 again providing the greatest challenge.

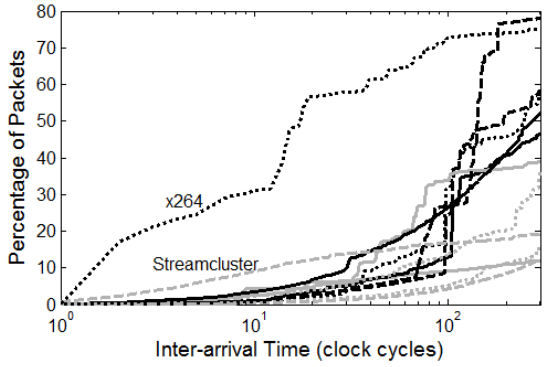

Fig. 4: PARSEC benchmark cumulative distributions of inter-arrival times

Table 1. Number of Packets in each 300 cycle intervals (per port)

	Max.	*Cum. Dist. 90%*	*Cum. Dist. 99%*
Blackscholes	*57*	*1*	*3*
Bodytrack	*18*	*3*	*4*
Canneal	*9*	*1*	*2*
Dedup	*19*	*3*	*4*
Facesim	*54*	*2*	*3*
Ferret	*20*	*3*	*5*
Fluidanimate	*16*	*3*	*3*
Freqmine	*39*	*2*	*6*
Streamcluster	*112*	*1*	*3*
Swaptions	*14*	*3*	*5*
Vips	*19*	*1*	*3*
X264	*57*	*8*	*16*

4. Implications for Future Photonic Networks

Our results show that with appropriate switching technology, both the circuit switch and TDM techniques can reduce power consumption compared with an electronic packet switched network. For circuit switching, circuits must be setup on microsecond timescales to route a large proportion of the traffic onto the circuit switch. For DSM traffic we do not see long-lived flows that made the technique attractive for supercomputers with millisecond timescale switches [7]. Obtaining the latency benefit of circuit switching on microsecond timescales requires an efficient scheduler. Given that on average only 1-3 packets are transmitted in a single time interval at the shortest timescales, it is unlikely that real time network monitoring could efficiently detect circuit requirements. It is an open question as to whether circuit requirements could be efficiently detected by the compiler or programmer. The results assume ideal circuit decisions and non-ideal scheduling will reduce the energy advantage. The TDM technique has attractive energy properties, even given our conservative assumption of equal power consumption for electronic and photonic transceivers. The disadvantage of TDM is the relatively high latency, demonstrating the importance of minimizing the slot time through increasing bit rates, reducing clock recovery and switching times and jitter control. The low IAT observed for algorithms with high proportions of same destination traffic, indicate a requirement for supporting short circuits (≈ 10s of ns) within the scheduling algorithm of a TDM network.

[1] N. Sherwood-Droz, *et al.*, "Optical 4x4 hitless silicon router for optical Networks-on-Chip (NoC)", *Opt. Exp.* **16**, 15915-15922 (2008).
[2] I. H. White and R. V. Penty, "Optical interconnects for backplane and chip-to-chip photonics", *Networks-on-Chip Symposium (2008)*.
[3] J. U. Knickerbocker, *et al.*, "Three-dimensional silicon integration," *IBM J.of Research and Developmen* **52**, 553-569 (2008).
[4] U. Vlasov, "Silicon photonics for next generation computing systems," Proc. *ECOC 2008*.
[5] J. L. Hennessy and D. A. Patterson, *Computer Architecture, A Quantitative Approach*, 4th ed.: Morgan Kaufmann, 2007.
[6] G. Hendry, *et al.*, "Analysis of Photonic Networks for a Chip Multiprocessor using scientific applications", *Networks on Chip Symp.* (2009).
[7] K. J. Barker, *et al.*, "On the Feasibility of Optical Circuit Switching for High Performance Computing", *Supercomputing Conference (2005)*.
[8] R. Luijten, *et al.*, "Viable opto-electronic HPC interconnect fabrics" in *Supercomputing Conference 2005*.
[9] M. Glick, *et al.*, "SWIFT: A testbed with optically switched data paths for computing applications", *Proc. ICTON (2005)*.
[10] A. Shacham and K. Bergman, "Building ultralow-latency interconnection networks using photonic integration", *IEEE Micro* 27, 6-20 (2007).
[11] C. Bienia, *et al.*, "The PARSEC Benchmark Suite", *Proc. Conf. on Parallel Architectures and Compilation Techniques* (2008).
[12] N. Barrow-Williams, *et al.*, "A Communication Characterisation of Splash-2 and Parsec," in *Proc. Symp. Workload Characterisation (*2009).
[13] J. Poulton, *et al.*, "A 14-mW 6.25-Gb/s transceiver in 90-nm CMOS," *IEEE J. of Solid-State Circuits* 42,2745-2757 (2007).
[14] A. Banerjee, *et al.*, "An Energy and Performance Exploration of Network-on-Chip Architectures" *IEEE Trans. VLSI* 17, 319-329 (2009).
[15] A. W. Poon, *et al*, "Cascaded Microresonator-Based Matrix Switch for Silicon On-Chip Interconnection", *Proc. IEEE* 97, 1216-1238 (2009).
[16] R. S. Tucker, "Green Optical Communications---Part II: Energy Limitations in Networks" to appear in *IEEE J. STQE (2011)*.
[17] H. Wang, *et al.*, "Demonstration of lossless monolithic 16x16 QW SOA switch", *Proc ECOC (2009)*.