POSTER: BigBus: A Scalable Optical Interconnect

Eldhose Peter
Indian Institute of Technology
Hauz Khas, New Delhi 110016
Email: eldhose@cse.iitd.ac.in

Janibul Bashir
Indian Institute of Technology
Hauz Khas, New Delhi 110016
Email: janibbashir@cse.iitd.ac.in

Smruti R. Sarangi
Indian Institute of Technology
Hauz Khas, New Delhi 110016
Email: srsarangi@cse.iitd.ac.in

***Abstract*—This paper presents *BigBus*, a novel on-chip photonic network for a 1024 node system. The crux of the idea is to segment the entire system into smaller clusters of nodes, and adopt a hybrid strategy for each segment that includes conventional laser modulation, as well as a novel technique for sharing power across nodes dynamically. We represent energy internally as tokens, where one token will allow a node to send a message to any other node in its cluster. We allow optical stations to arbitrate for tokens and at a global level, we predict the number of token equivalents of power that the off-chip laser needs to generate.**

I. Introduction

A mere agglomeration of low power cores, and cache banks does not yield a high performing system, unless it is supplemented with a high performance interconnect. We propose one such interconnect in this paper called *BigBus*, which is a high performance optical interconnect. Our reasons for choosing photonics based technology are because of its inherent advantages in terms of latency, bandwidth, and possibly power efficiency if designed well (duly justified in Section III). We need such interconnects for such large systems of cores and caches.

By analyzing the behavior of some Parsec benchmarks, we observe that unless some spatial locality is created, we will have too many messages being sent all over the chip. After creating spatial locality, we note the variance in the optical power requirements of different transmitters, and try to intelligently size sub-networks to take the variance in traffic into account. We build on these insights in Section II, and show our results in Section III.

II. BigBus Architecture

In this poster, we propose a system with 768 cores and 256 cache banks (see Figure 1(a)). Each core in *BigBus* is a dual-issue in-order RISC processor similar to the cores used by the authors of ATAC [2].

We create square-shaped blocks of 4 cores or 4 cache banks, and refer to them as a *cluster*. Intra-cluster communication is electrical, and inter-cluster communication is optical. Each cluster has an optical station. We then proceed to build larger clusters: 16 clusters (arranged as a 4×4 square) form a *P-Cluster*, and each *O-Cluster* contains 4 *P-Clusters* (2×2). We thus have 4 *O-Clusters* in our system.

1) Power Network: We have four off-chip 1550 nm laser sources – connected to the chip at 4 separate points – for each *O-Cluster*. We use a tree based network to distribute power to the 4 constituent *P-Clusters*. At each *P-Cluster*, we use a comb splitter which is used to split a monochromatic optical signal (at 1550 nm) into 64 equally spaced wavelengths (between 1450-1650 nm). We use DWDM (dense wavelength division multiplexing) technology to transmit all these separate wavelengths on the same waveguide. Subsequently, we use a set of 16 cascading tunable optical splitters to split the DWDM signal into 17 parts. Now, each part of the optical signal is assigned to one waveguide: 16 backbone waveguides, and 1 token waveguide. All of these waveguides run parallel to each other. A station can source power from a backbone waveguide only if it has its corresponding token. The number of tokens that need to transmitted in each *P-Cluster* needs to be determined correctly, in order to decrease the optical power consumption. We divide time into fixed size durations called *epochs* and devise a prediction mechanism to predict the number of tokens that we require in the subsequent epoch. We then modulate the off-chip laser to produce just enough power in the next epoch (based on predicted activity).

Activity prediction is done in two stages. In the first stage, each station decides whether to increase or decrease tokens based on a function that has two inputs: wait time ($\mathscr{T}$) and the number of pending events ($\mathscr{N}$) and sends this information to the laser controller. In the second stage, the laser controller decides the number of tokens that should be created.

$$\mathscr{F}(\mathscr{T},\mathscr{N}) = \begin{cases} 3 & \mathscr{N} \geq T_p (T_p = 8) \\ 2 & \mathscr{T} \geq T_w \| T_p/2 \geq \mathscr{N} (T_w = epoch\_size/2) \\ 1 & T_w/2 \leq \mathscr{T} < T_w \| \mathscr{N} < T_p/2 \\ 0 & \mathscr{T} < T_w/2 \end{cases}$$

2) Data Network: The cache banks and cores in an *O-Cluster* are connected together by a serpentine structured optical link (called *O-Link*)(see Figure 1(a)). Additionally, we define a separate optical link to connect all the cache banks at the center of the chip called *CB-Link*. The set of all the cache banks connected by this link is called the *CB-Cluster* (see Figure 1(b)). Lastly, we define an optical link called the *top level link* that connects all the O-Links together. It is attached to each O-Link via a hub (containing a message queue).

In each *O-Link*, we have 64 parallel waveguides divided into four equally sized groups ($G_1 \ldots G_4$) (one group per *P-Cluster*). Stations in *P-Cluster i* can only transmit messages on any waveguide in group G_i, but they can receive messages on

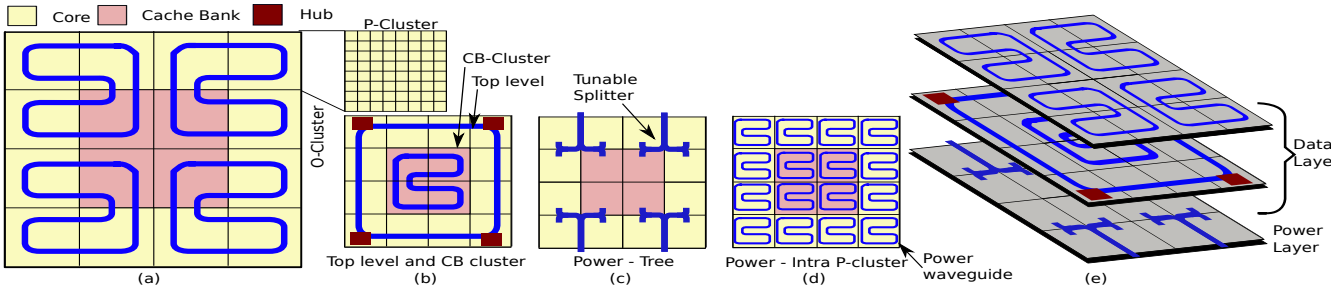

Fig. 1. Architecture

all the waveguides (even those assigned to other *P-Clusters*). To access the shared data waveguide we use the result of the arbitration for the power waveguide. If we get access to the i^{th} backbone waveguide in the bundle of power waveguides, we infer a permission to access the i^{th} data waveguide as well. In addition, if a station has multiple messages waiting in its queue, then it tries to grab tokens (1 token corresponding to each waveguide) for multiple waveguides in order to send multiple messages in parallel.

III. Experimental Results

We compare *BigBus* with three other state of the art photonics based multicore architectures namely Probe [6], ColdBus [3] and ATAC [2]. We use benchmarks from the Parsec [1] and the Splash2 [5] benchmark suites for simulations. We simulate all the designs using Tejas [4], a cycle accurate simulator. For a fair comparison, we have made some modifications to these architectures. We use the segmented power distribution(*P-Cluster based*) scheme for Probe, because a single power waveguide will not scale for a 1000 node system. While simulating Probe, we use its activity prediction and laser modulation schemes. We call this configuration *mProbe*. To simulate ATAC, we use their power network, which does not use any laser modulation techniques and we call this configuration *mATAC*.

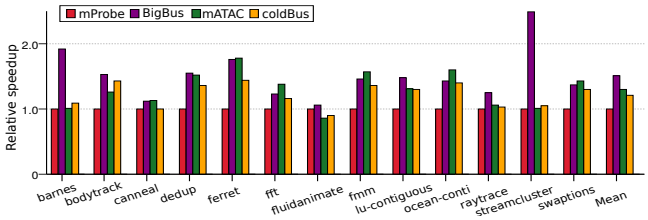

Fig. 2. Performance comparison

Compared to *mProbe*, ColdBus, and *BigBus* perform (in terms of simulated execution time) 18%, and 34% better respectively as shown in Figure 2. *BigBus* is the best configuration and is 14% faster than *mATAC*. The performance improvement of *BigBus* is due to the ability to send multiple messages at a time, and the ability to handle mispredictions due to the use of the shared network.

BigBus consumes 48% lesser laser energy as compared to *mProbe*. Because of the lack of laser modulation, *mATAC* is the most energy consuming configuration. It consumes 10 times more energy as compared to *BigBus*. Compared to ColdBus, *BigBus* consumes 12% lesser laser energy. In terms of energy-delay^2 product (ED^2), *BigBus* is the best configuration with a reduction of 61% as compared to *mProbe* (see Figure 3). ColdBus is the second best configuration (41% reduction).

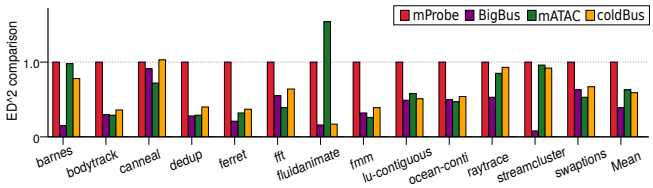

Fig. 3. ED^2 comparison

It is impossible to exactly ascertain the accuracy of prediction because optical power is not a binary value, rather it can take 32 values. Instead, we can use the "increase in queue occupancy due to unavailability of laser power" as an indirect measure. This will quantify the contention in queues due to lack of power. The average value of this parameter is less than 0.01 per hundred requests. This shows that the accuracy of our prediction mechanism is very high.

IV. Conclusion

In this paper, we proposed a novel optical network, *BigBus*, for the 1000 core era. We opted for a novel hybrid design that uses both laser modulation and power sharing across stations. The former approach is very effective in reducing static optical power, and the latter approach is effective in making the best utilization of the power that is available. To take both of these design decisions into account we proposed to use tokens for distributing both power and access to data waveguides. By using the currency of tokens, we could also simplify our design, and propose a predictor that predicted the number of tokens that we need to generate per epoch.

References

[1] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The PARSEC benchmark suite: characterization and architectural implications," in *PACT*, 2008.

[2] G. Kurian, J. E. Miller, J. Psota, J. Eastep, J. Liu, J. Michel, L. C. Kimerling, and A. Agarwal, "Atac: a 1000-core cache-coherent processor with on-chip optical network," in *PACT*, 2010.

[3] E. Peter, A. Thomas, A. Dhawan, and S. R. Sarangi, "Coldbus: A near-optimal power efficient optical bus," in *HiPC*, 2015.

[4] S. R. Sarangi, K. Rajshekar, K. Prathmesh, G. Seep, and P. Eldhose, "Tejas: A java based versatile micro-architectural simulator,," in *PATMOS*, 2015.

[5] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta, "The splash-2 programs: characterization and methodological considerations," *SIGARCH Comput. Archit. News*, vol. 23, pp. 24–36, May 1995.

[6] L. Zhou and A. K. Kodi, "Probe: Prediction-based optical bandwidth scaling for energy-efficient nocs," in *NOCS*, 2013.