Broadcast-and-weight Interconnects for Integrated Distributed Processing Systems

Alexander N. Tait, Mitchell A. Nahmias, Bhavin J. Shastri, and Paul R. Prucnal
Princeton University, Princeton, NJ, 08544 USA
atait@princeton.edu

***Abstract*—Silicon photonic platform development has revolved around point-to-point links for multi-core computing systems. We examine an opportunity for this technology to extend to unconventional architectures that rely heavily on interconnect performance. Broadcast-and-weight is a new approach for joining neuron-inspired processing and optical interconnect physics.**

I. Introduction

A modern generation of hardware-based neuron-inspired systems attempt to exploit the efficiency and robustness of sparse codes called spikes. Spiking systems could address important problems in recognition, optimization, and analysis of high-dimensional environments. Spiking models are central to several recent projects in microelectronic neuromorphic hardware, such as IBM's cognitive computing architecture [1], which address these domains where von Neumann machines perform poorly.

Utilization of analog physical dynamics represents a key step towards attaining the efficiency and functionality exhibited by biophysical information processors [2]. In most neural network models, each neuron multiplies signals from many other neurons by independent weights, sums them, performs some nonlinear dynamical process, and transmits copies of a single output signal. Long-range connections and high fan-in are thus critical capabilities that are fundamentally difficult for electronic interconnects (e.g. cross bars) to handle with both performance and scale. This fact has forced all modern spike processors to adopt various forms of address-event routing, which sacrifice bandwidth and efficiency by representing spikes as digital codes instead of physical pulses.

We present an on-chip interconnect protocol called broadcast-and-weight, which leverages recent advances in photonic integrated circuit technology to address interconnect challenges faced by distributed processing. In the past, optical neural networks have encountered barriers in reliability, scalability, and cost because they are difficult to integrate. WDM effectively channelizes available bandwidth without spatial or holographic multiplexing and avoids coherent interference effects during fan-in. High-bandwidth optical channels are compatible with recently proposed laser neuron devices [3], which could access a picosecond computational domain that impacts application areas where both complexity and speed are paramount (e.g. adaptive control, real-time embedded system analysis, and cognitive RF processing).

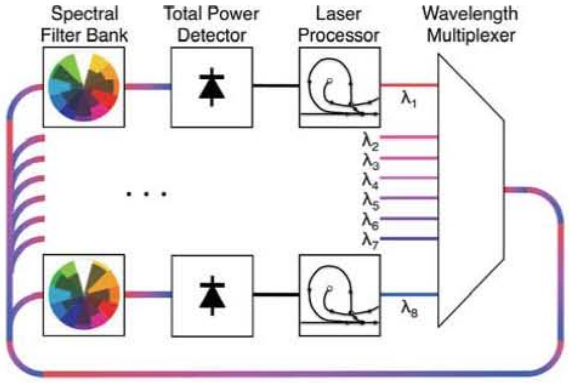

Fig. 1. An optical broadcast-and-weight network showing parallels with the neural pathway model of weighted addition front-end and large fan-out. An array of source lasers are wavelength multiplexed (WDM) in a single waveguide (multicolor). Independent weighting functions are realized by spectral filters (represented by gray color wheel masks) at the input of each unit. Instead of demultiplexing, the total optical power of each spectrally weighted signal is detected by a single photodetector, yielding the sum of the input channels. In this protocol, photodetectors act simultaneously as transducers and additive analog computational elements, solving both challenges of large, parallel fan-in and efficient many λ-conversion.

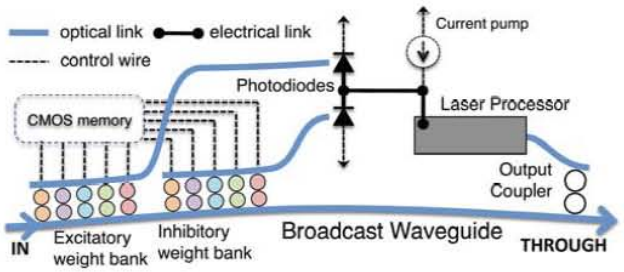

Fig. 2. Implementation of a processing-network node (PNN) in standard silicon photonic devices. A spectral filter bank is implemented by a bank of tunable microring resonator filters, whose drop weight is configured by CMOS drivers. Balanced photodetectors sum all weighted inputs by converting to the electronic domain. Photocurrent subtraction allows inputs to have a positive (excitatory) or negative (inhibitory) modulation effect, push-pull capabilities considered essential to any neuron-inspired model. A short wire modulates current injection into a hybrid evanescent excitable laser neuron [3], which performs both threshold detection and pulse generation. The output of the laser is coupled back into the broadcast waveguide and broadcast to other PNNs.

978-1-4799-2468-4/14/$31.00 ©2014 IEEE

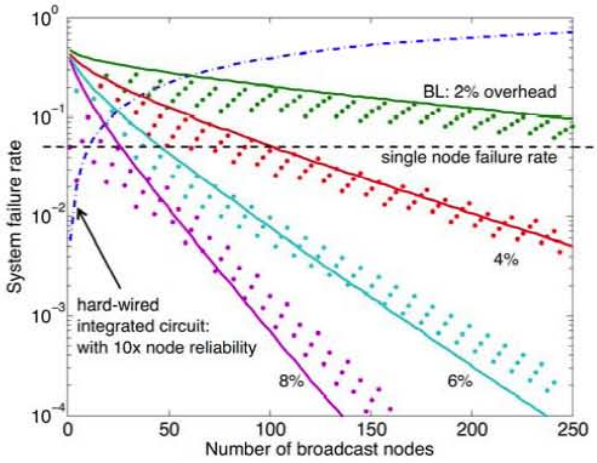

Fig. 3. System failure rate in networks of unreliable nodes. In constrained circuit-routed networks (blue dash-dot line), system failure approaches certainty with more nodes. Broadcast-and-weight systems with hardware overhead (solid lines) invert this trend, experiencing reliability that scales with complexity. The systemic reliability of a BL can even be better, sometimes by orders of magnitude, than that of a single node (black dotted line). Colors represent different overhead percentages from 2% to 8%. Circular markers are the exact BL behavior; solid curves are approximate error function models.

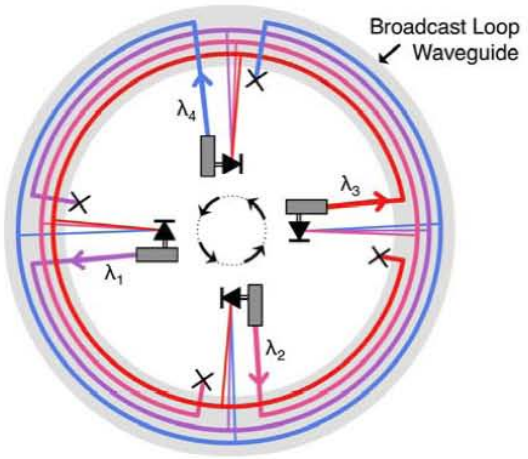

Fig. 4. Conceptual diagram of a broadcast loop. The loop waveguide carries WDM channels from all participating PNNs. Each PNN detects a linear subset of the present channels. The PNN laser then outputs its signal, a function of those inputs, on its unique wavelength channel. Once a signal transverses the BL, it is terminated by its originating unit to avoid further interference. Filter banks and inhibitory pathways not shown.

II. Interconnect Architecture

1) Broadcast-and-weight: Broadcast-and-select is a fiber WDM protocol that obtains collision-free, circuit-routed, and densely parallel interconnection. The active connection is selected, not by altering the intervening medium, but by tuning a filter at the receiver [4]. We propose a similar protocol called "broadcast-and-weight," which allows multiple inputs to be selected simultaneously and with intermediate strengths between 0% and 100%. A group of nodes shares a common medium in which the output of every node is assigned a unique transmission wavelength for broadcast (Figure 1).

2) Processing-network node: Participants in the network called processing-network nodes (PNN) perform both roles of processing (weighting, addition, nonlinear dynamics) and networking (routing, λ-fan-in, WDM carrier generation) in a compact set of standard devices. A silicon photonic PNN implementation is depicted in Figure 2. This circuit technique could also generalize to future PIC platforms. λ-fan-in in a photodetector strips WDM signals of any trace of their origin, a side-effect that corrupts digital signals, typically neccessitating demultiplexing and dedicated detection. In the neurocomputing context, however, this channel destruction is precisely correspondant with the summation function, so dumultiplexing is not required. A λ-fan-in front-end was found to yield input commutativity resulting in combinatorial robustness to device failures, so the systemic reliability of a PNN network can *decrease* with network size (Figure 3). A "receiver-less" front-end, it is not subject to well-known optical-electronic-optical (O/E/O) conversion overhead, whose assumed cost, energy, and complexity are due to the digital electronic receiver (amplifier, sampler, quantizer) in most communication links, not to the physical conversion itself.

3) Broadcast loop: A physical medium is efficiently implemented by a ring waveguide (Figure 4). Each PNN drops a fraction of total power, allowing most to continue. Drop-and-continue is a physical solution to optical multicasting that can enhance virtual interconnect density for a given network traffic [5]. In this context, the broadcast loop (BL) is fully multiplexed and capable of supporting N^2 interconnects in just 1 link with N WDM channels, where an electronic interconnect would require, at best, $N(N-1)/2$ links.

Waveguide rings with WDM channelization have been proposed as an implementation of broadcast-and-*select* for efficient multicast in multi-core networks on-chip [6]. However, demultiplexing and dedicated detection can negate area and energy savings and create a buffering bottleneck. In contrast, physical λ-fan-in through total power detection does not require demultiplexing or an electronic receiver. By addressing interconnect challenges of efficient parallelism and fan-in, broadcast-and-*weight* could extend the promise of silicon photonics to high-performance neuromorphic architectures.

References

[1] D. S. Modha *et al.*, "Cognitive computing," *Commun. of the ACM*, vol. 54, no. 8, pp. 62–71, Aug. 2011. [Online]. Available: http://doi.acm.org/10.1145/1978542.1978559

[2] K. Boahen, "Neurogrid: emulating a million neurons in the cortex," in *engineering in medicine and biology society, IEEE int. conf. of*, 2006.

[3] M. A. Nahmias *et al.*, "An evanescent hybrid silicon laser neuron," in *IEEE Photonics Conf. (IPC)*, Sep. 2013.

[4] R. Ramaswami, "Multiwavelength lightwave networks for computer communication," *IEEE Comm. Mag.*, vol. 31, no. 2, pp. 78–88, 1993.

[5] X. Zhang *et al.*, "Constrained multicast routing in WDM networks with sparse light splitting," *J. Lightwave Tech.*, vol. 18, no. 12, pp. 1917–1927, 2000.

[6] S. Le Beux *et al.*, "Optical ring network-on-chip (ORNoC): Architecture and design methodology," in *Design, Automation Test in Europe Conference Exhibition (DATE)*, 2011, pp. 1–6.