O-BAR: A Scalable, Optical Neuromorphic Communication Protocol

Mitchell A. Nahmias, Alexander N. Tait, Bhavin J. Shastri, and Paul R. Prucnal
Lightwave Communications Laboratory, Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA
mnahmias@princeton.edu

***Abstract*—We propose an address event representation (AER) broadcast optical networking scheme to alleviate communication bottlenecks in electronic neuromorphic processors. This protocol is collision, router and switch-free, potentially supporting large (10;000) neural fan-in and fan-out.**

I. Introduction

Spiking neural networks (SNNs) represent an alternative to the von Neumann paradigm of computation. They take advantage of distributed, sparse, and robust encoding schemes to perform computations with higher power efficiency and performance, and have been utilized both for biological network simulators and low-power data processors. Unlike interconnects in von Neumann processors, which are based on point-to-point communication between nodes, neural networks require many-to-one fan-in and significant multicasting, which can be a bottleneck in electronics.

Large-scale neuromorphic systems—such as IBM's TrueNorth architecture [1], SpiNNaker [2], and Brains on Silicon [3]—use packet switching, time division multiplexing (TDM), or crossbar arrays with complex routing protocols to side-step the electronic wiring bottleneck. Crossbar arrays can realize dense all-to-all interconnects, but are difficult to scale hierarchically to larger systems. In contrast, packet switching allows dense virtual interconnects, but at the cost of a communication and energy overhead.

Optical communication systems could provide an alternative, complementing the processing density of electronics with power-efficient, high bandwidth communication. Unlike electrical wires, optical signals consume less energy per bit, have a much larger bandwidth-distance product and a greater bandwidth-density per wire. The advantages are especially salient in a neuromorphic system, in which the primary system bottleneck is in communication rather than in processing. We propose O-BAR as an alternative to electronic packet switching: an Optical Broadcast Address event Representation network. The protocol is simple—free of collisions, routers, and switches—compatible with emerging trends in photonic integration, can support high-density interconnects, and can be scaled hierarchically to form large-scale neuromorphic systems.

II. Working Principles

Utilization of wavelength division multiplexing (WDM) is crucial in any parallelized interconnect fabric. Unfortunately, the simplest approach—associating a single wavelength channel $_i$ with each neuron—fails to utilize the full bandwidth of optics unless one operates at a faster time scale (as in [4]). Instead, we utilize a multiplexing scheme called address event represention (AER), in which a packet is release into the network containing only the identity of the neuron which has fired. This packet-based method allows multiple neurons to share a given wavelength channel to fully utilize available bandwidth.

The architecture separates communication and processing into optical and electronic hardware, respectively, and has many topological similarities to the ORnOC architecture [5]. Communication centers around a WDM ring network as shown in Figure 1, which implements all-to-all WDM broadcasting. Each node contains a cluster of neurons and synapses in electronics for processing and programming, using AER protocols at the sender to multiplex asynchronous spikes and broadcast them to every other node. Each node receives simultaneous signals from all other nodes in the loop, which are demultiplexed and converted to electronic signals that interface with the neural cluster.

A. Broadcast Node

Suppose there are N nodes, labeled from $i = 0;1;2:::N$. Each broadcast node contains densely packed neurons and synapses and asymmetrically receives all wavelength channels C while outputting on only a subset of channels $k = C=N$. Clusters of n neurons (where $n > k$) share the channels k for outputs. Since neurons will release pulses asynchronously based on their input activity, the E/O Packet Multiplexer (as shown in Figure 1) must perform arbitrage and buffering to prevent multiple access collision. The arbitrage can be done locally at the sender, keeping the network collision-free. Meanwhile, signals from other nodes are demultiplexed and converted to electrical signals which are compatible with dense, locally-packed synapses to provide inputs to the neural cluster.

B. Broadcast Loop

The broadcast loop is the medium through which all the AER packets are distributed to each node simultaneously, utilizing the extensive bandwidth of optics to perform high-throughput broadcasting. As Figure 2 shows, packets from a given cluster of neurons can coupled into the loop along a

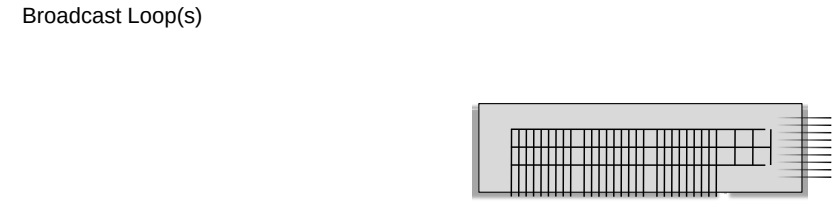

Fig. 1. A schematic of the proposed architecture with four clusters of nine neurons each. Each cluster *i* functions locally. When a neuron fires, its identity is sent to the E/O packet multiplexer, which sends a unique orthogonal code (packet) along a given wavelength $_i$. Each $_i$ channel is random access (i.e. neuron events are statistical multiplexed), but arbitrage is local to the sender so there are no network packet collisions. Balanced light-path splitting and filtering via drop-and-continue distributes the signal evenly to the other three clusters. The packets are demultiplexed and converted into electrical signals at the O/E packet demultiplexer, which modulating a densely clustered array of synaptic elements, driving the neural cluster. This miniature loop network alone supports 1296 synaptic connections. Multi-loop organization is possible through the sharing of channels or broadcast nodes across loops.