The CSI Journal on Computing Science and Information Technology
Vol. 13, No. 2, 2016

A High-Performance and Low-Power Network-on-Chip Architecture for Neural Networks

Nasrin Akbari **Bita Dabiri** **Mehdi Modarressi**

School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran

ABSTRACT

Emerging neural network accelerators are often implemented as a many-core chip and rely on a network-on-chip to handle the huge amount of inter-neuron traffic. The well-known mesh and tree are the most popular topologies in prior many-core neural network implementations and research proposals. However, these conventional topologies suffer from high diameter, low bisection bandwidth, and poor collective communication support. In this paper, we present a customized version of the Dragonfly topology for Neural Networks. The capability of dragonfly to support multicast and broadcast traffic in a simple and efficient way, as well as its low diameter, is the major motivation behind proposing a customized version of this topology as the communication infrastructure of neural network accelerators. We also apply a conflict-free static scheduling for neurons to send their data to the network, thereby enable the network to use very simple circuit-switched routers to further improve power/performance profile. We compare Dragonfly with some state-of-the-art NoC topologies adopted in recent neural network hardware accelerators and show that it yields lower average message hop count and higher throughput.

Keywords: Network-on-Chip, Neural Network, Circuit-Switching, Low-Power, Dragonfly Topology.