Neural Network Assisted Reconfigurable Optical Network-On-Chip (ONoC)

Ryan Wong[1], Lei Zhang[2]
[1]University of Rochester, [2]University of Maryland Eastern Shore

Abstract

Optical interconnects provide higher bandwidth than metal based electric wiring. The fully optical network, Ouroboros, allows for any node or group of nodes to be connected, creating an independent subnetwork within the full network. This project attempts to optimize the configuration of the ONoC by analyzing the traffic patterns and predict which configurations will lead to low execution times.

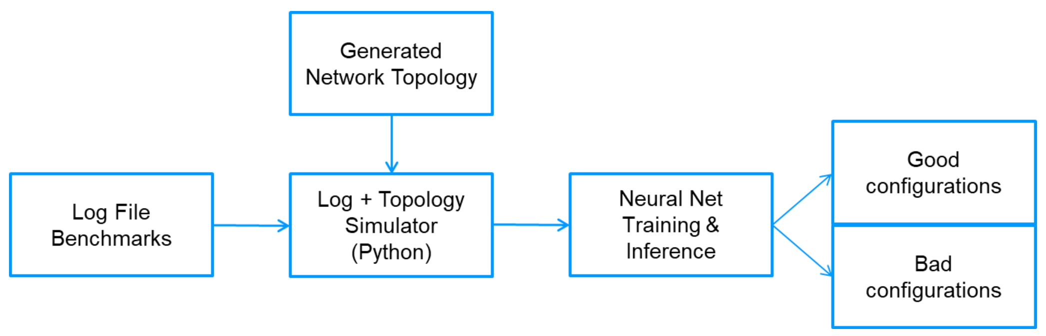

Figure 1

Method

An overview of the procedure is shown in Fig. 1. A custom built simulator is used for network simulation which given a configuration of nodes and a log file, outputs the expected execution time. Each test configuration is randomly generated. The log files are traces of the freqmine benchmark, taken from the PARSEC suite, and simulated on a 4x4 electrical NoC. Data is labeled according to its execution time compared to the in-order ring (0 to 15), where 'good' has a lower latency.

The neural network architecture is composed of two hidden layers of 15 neurons each, with a constant batch size of 100. Training and testing was performed using 60,000 and 600 configurations, respectively

For Dynamic Reconfiguration, the log files are partitioned based on transmission density, and volume factor. Determining the 'best' configurations to test is done by taking the 10 fastest and 10 slowest configurations simulated over the entire log file.

Ouroboros Network

This project uses an Ouroboros network organized in a ring configuration, with potential reconfigurations shown in Fig. 2. Similar to visible light, optical signals are blocked by the presence of another object, in this case other processors' optical devices on the bus. Thus, one of the optimizations is to have as maximize the number of transmissions on the network, while minimizing the number of blocking transmissions.

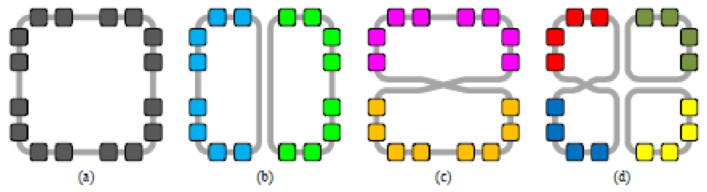

Figure 2

Results

Using the process outlined in Fig. 1, and a freqmine sub log, we observe a classification accuracy of roughly 66% for a 3-level classification. When going to a 5-level classification our accuracy decreases to roughly 40%, likely due to increased possible outputs. Figure 3 demonstrates the relative accuracy for both ReLU and Sigmoid activation functions. When using full logs instead of partial logs, these numbers vary only by 2 to 3 percent.

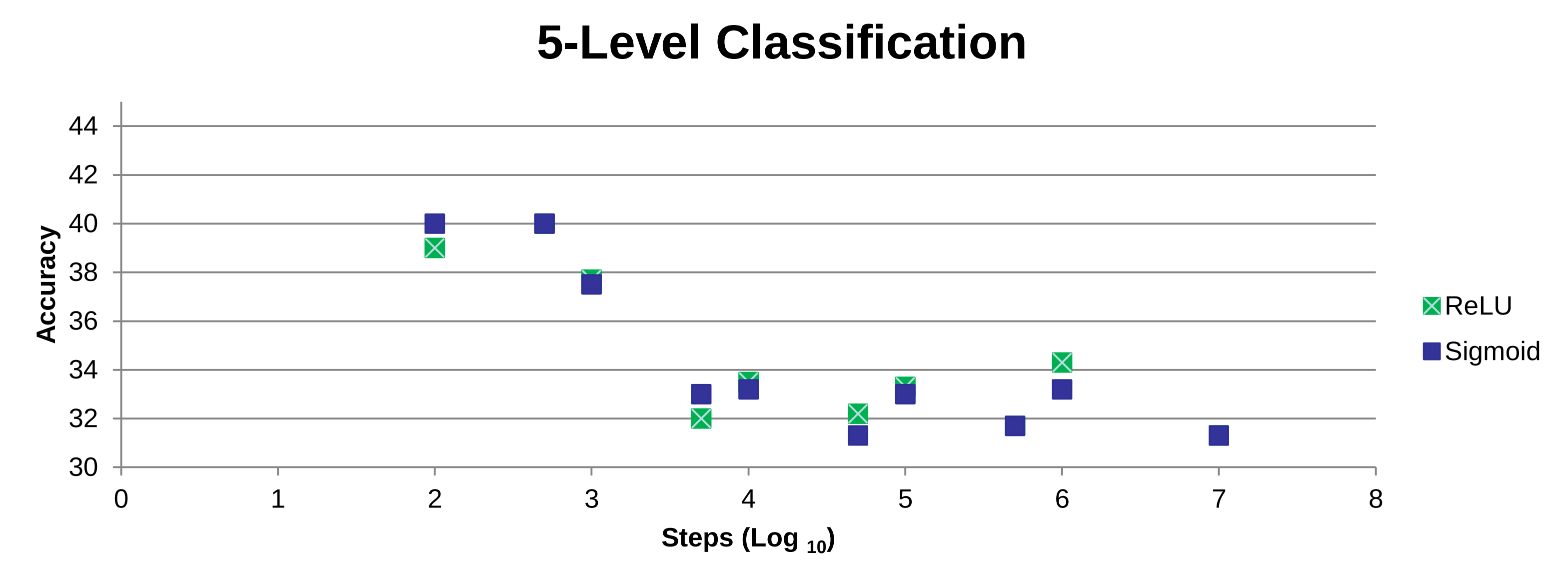

Figure 3

Volume Factor

Another important aspect of the log file is the volume factor, which influences a partition's dependence on configuration. It specifies the relative size of the transmission. We found that a low transmission density, in addition to a low volume factor is not sufficient to saturate the network. This leads to the 'Configuration Independent Zones'. In these zones there is only one transmission on the network, and thus any reconfiguration will not reduce latency. Examples of these zones can be seen in Fig. 4, which shows 65% of traffic occurs in the first 10% of the benchmark, and a maximum of 8% in each subsequent partition.

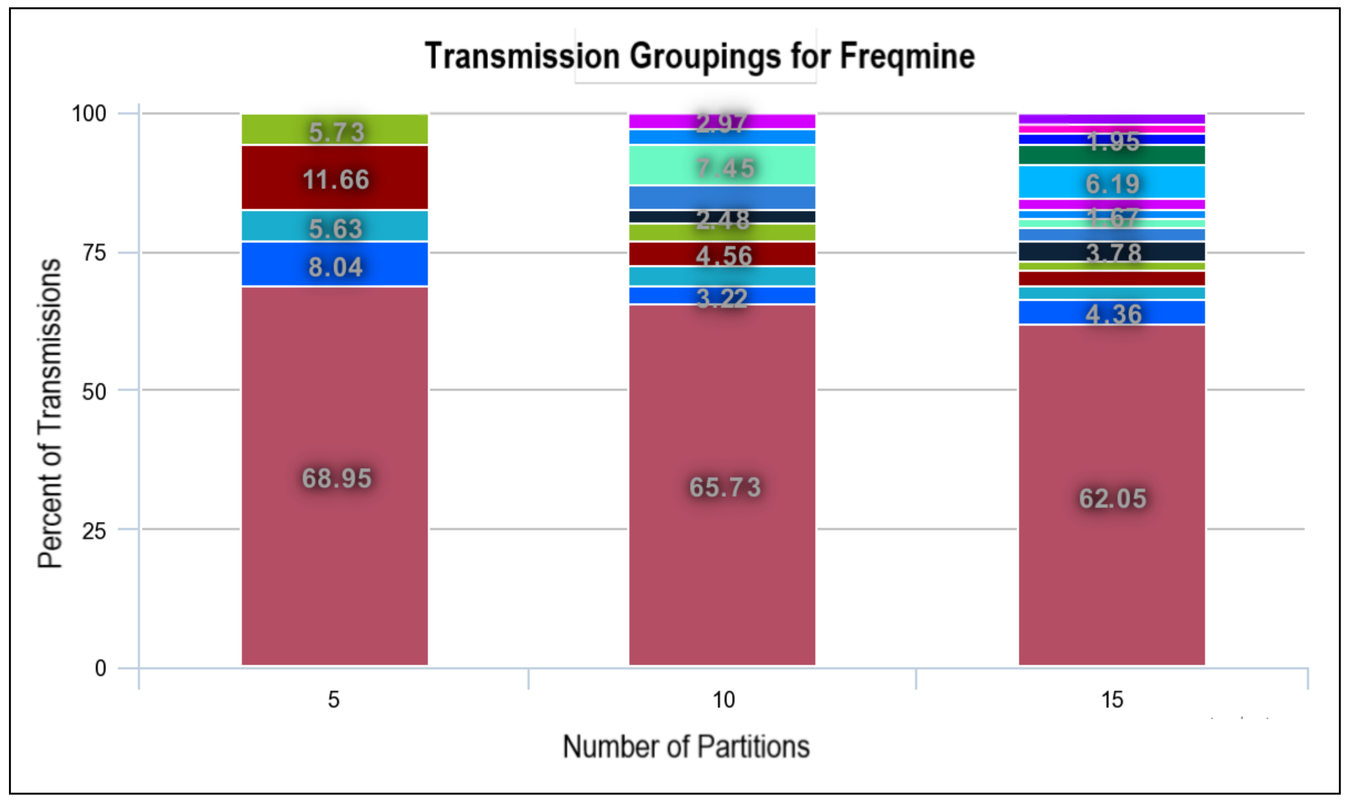

Figure 4

Conclusions

- Additional input data, in addition to the Ouroboros configuration, is required for precise network classification.
- Dynamic Reconfiguration is only effective after the Ouroboros network has been saturated.

Future Works

- Adjust the correctness of a classification by scaling a prediction based on how close it is to the ground truth. In the case of the 5-level classifier, knowing if configuration is faster than the in order ring is more important than knowing to what degree.
- Use of Dense Wavelength Division Multiplexing (DWDM), another way to increase network bandwidth.
- Exploration of multidimensional networks, starting with 2D torus.

Acknowledgments

Funding for this project is provided by the National Science Foundation CNS Award #1757017.