Machine Learning Intelligent Chip Design Final Project Report

311510205 黃煒恩

I. Neural Network Model

For the implemented DNN model, I build a 4-layer model for MNIST dataset. The DNN consists of 784, 256, 120, and 10 neurons in each layer. The ReLU function is used in the hidden layers. As for the output layer, the confidence level of each digit is calculated by the Softmax function.

Layer	# of neurons	activation function	# of Parameters
Input layer	784		
Hidden layer 1	256	ReLU	200960
Hidden layer 2	128	ReLU	32896
Output layer	10	Softmax	1290

II. NoC Architecture

The NoC architecture is a 2D mesh topology consisting of processing elements (PEs), network interfaces, routers, and channels. Each PE is connected to a router through the network interface. The PEs are capable of calculating MAC values and supporting different activation functions, including ReLU, softmax, sigmoid, and tanh. Additionally, the PEs can receive and send flits to the router.

To realize the required functionality, I built a modified version of NN-Noxim. This modified version includes the implementation of network interfaces, and also I added support for the softmax function.

III. Experimental Results

A. Average delay (cycles) under different group sizes and NoC size:

Group size	**4x4 NoC**	**8x8 NoC**	**12x12 NoC**
32		581.305	1317.19
64		540.419	705.097
128	293.588	295	552.353
256	265	145	145
512	78.5	77	77
1024	10	10	10

(Note: XY routing and dir_x mapping are adopted.)

When the group size is low, the impact of the NoC size on delay is more significant. The reason for the longer delay on a 12x12 NoC could be related to the mapping method used, which causes packets to require a longer transmission time before reaching their destination.

B. Performance under different group sizes and mapping methods:

Group size	Number of packet/Flit	Mapping Algorithm (unit: cycles)			
		Dir_x mapping	Dir_y mapping	Lyr_x mapping	Lyr_y mapping
32	236/7896	581.305	1095.77	558.025	1174.23
64	62/3900	540.419	737.968	509.387	712.065
128	17/1986	295	295	535.118	487.706
256	6/1180	145	145	265	265
512	4/1126	77	77	137	137
1024	3/1174	10	10	10	10

(Note: 8 x 8 mesh topology and XY routing are adopted.)

Based on the experimental results, we can observe that the larger group size leads to the lower latency. The larger group size means that each PE is capable of processing more neurons, and thus fewer data transactions are required. As a result, number of packets that need to be transferred on the NoC is reduced.

Furthermore, when dealing with smaller group sizes, the Lyr_x mapping exhibits shorter delays. This is because if the PEs, which need to transmit their computation results to each other, are more concentrated, the delay can be shorter. Conversely, the delay increases when the PEs are more distributed.

C. Average delay (cycles) under different mapping methods and routing algorithms:

	XY	westfirst	northlast	negativefirst	oddeven	Fully adaptive
Dir_x	540.419	618.565	624.387	595.935	743.935	624.387
Dir_y	737.968	698.113	747.355	698.226	756.79	698.113

(Note: 8 x 8 mesh topology and the group size is 64.)

Different routing methods also affect latency under different mapping schemes. The dir_X mapping exhibits the lowest latency under the XY routing algorithm, while the dir_Y mapping suffers from limitations in Y-direction routing due to the XY routing algorithm, resulting in poorer latency performance.

IV. Conclusion

From the experimental results, we can observe that the group size is the primary factor affecting latency, which is mainly limited by the processing capability of the PEs. Therefore, we can run the simulations to determine the optimal mapping method and

routing algorithm given a specified NoC architecture and group size.

V. Reference

[1] Chen, Kun-Chih, and Ting-Yi Wang. "NN-noxim: High-level cycle-accurate NoC-based neural networks simulator." 2018 11th International Workshop on Network on Chip Architectures (NoCArc). IEEE, 2018.