Neu-NoC: A High-efficient Interconnection Network for Accelerated Neuromorphic Systems

Xiaoxiao Liu, Wei Wen, Xuehai Qian, Hai Li, **Yiran Chen**

University of Pittsburgh,

University of southern California,

Duke University

Outline

- Challenges in Computing System Design
- NoC for neuromorphic computing system

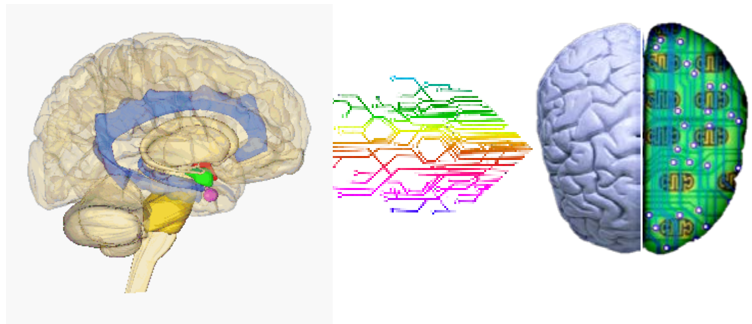
Neuromorphic Systems

IBM
TrueNorth

SRAM synapse
Digital spike
1M neurons/chip
256M synapse/chip

J. Hsu, *IEEE Spectrum*, 2014

A Brain Inspired Solution

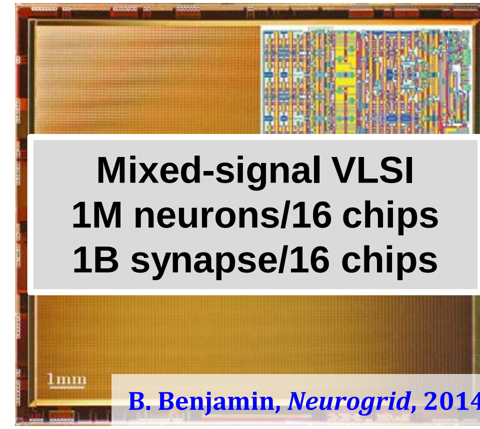


Integrated processing & storage

High-level parallelism

Large-scale memory

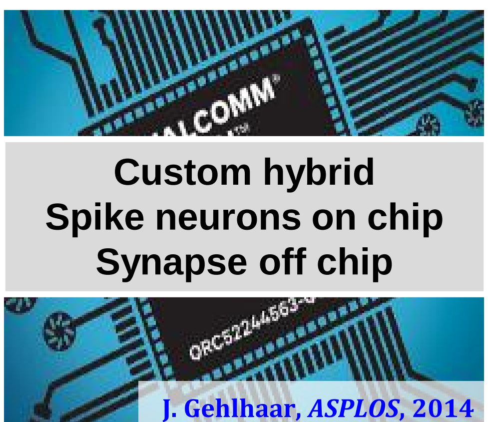
Stanford
Brain in Silicon



Mixed-signal VLSI
1M neurons/16 chips
1B synapse/16 chips

B. Benjamin, *Neurogrid*, 2014

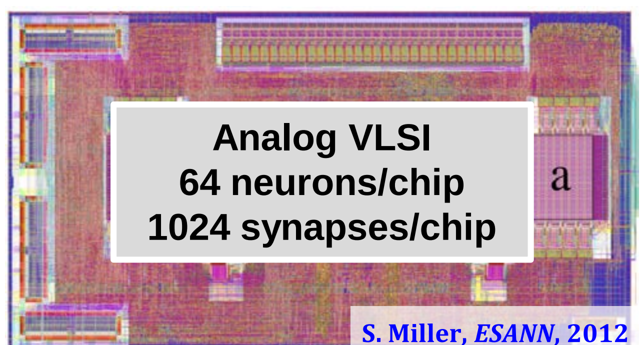
Qualcomm
Zeroth



Custom hybrid
Spike neurons on chip
Synapse off chip

J. Gehlhaar, *ASPLOS*, 2014

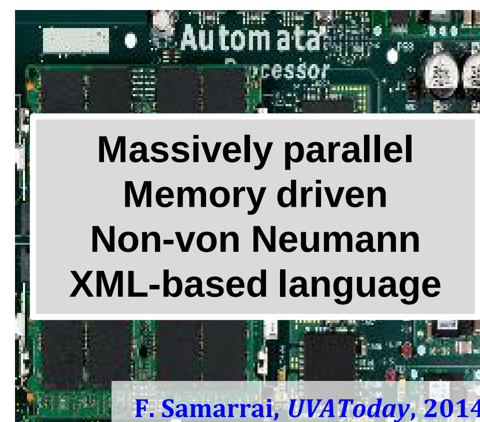
HBP
BrainScaleS



Analog VLSI
64 neurons/chip
1024 synapses/chip

S. Miller, *ESANN*, 2012

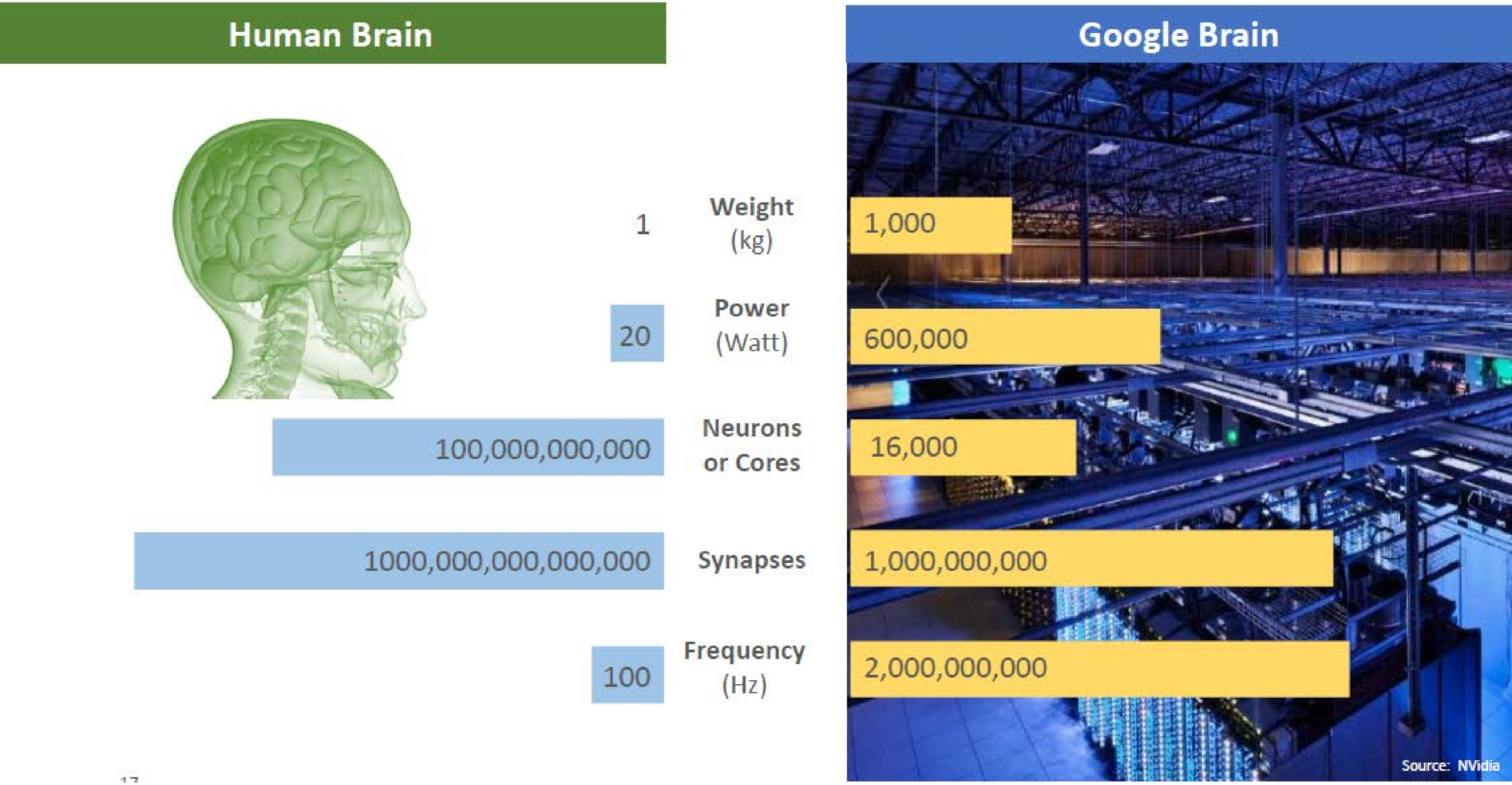
Micron
Automata



Massively parallel
Memory driven
Non-von Neumann
XML-based language

F. Samarrai, *UVAToday*, 2014

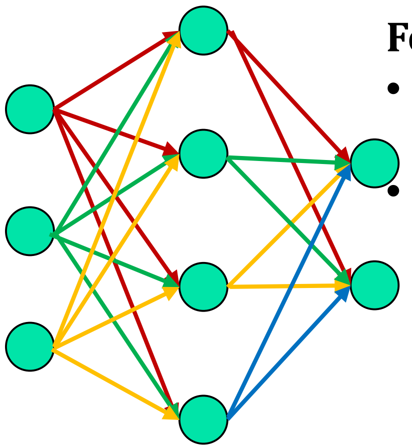
Neuromorphic System



The efficiency of Neuromorphic system is far behind human brain!

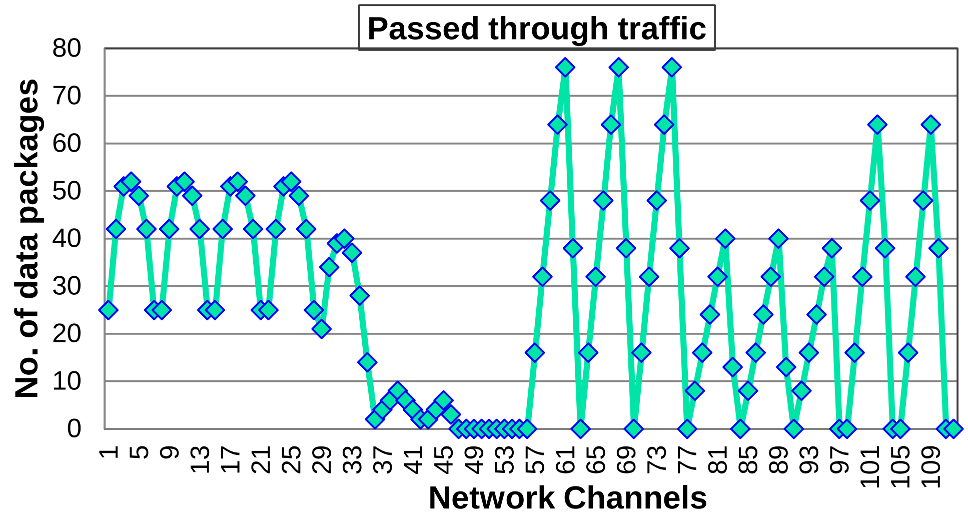
Communication in neuromorphic system

Neural Network

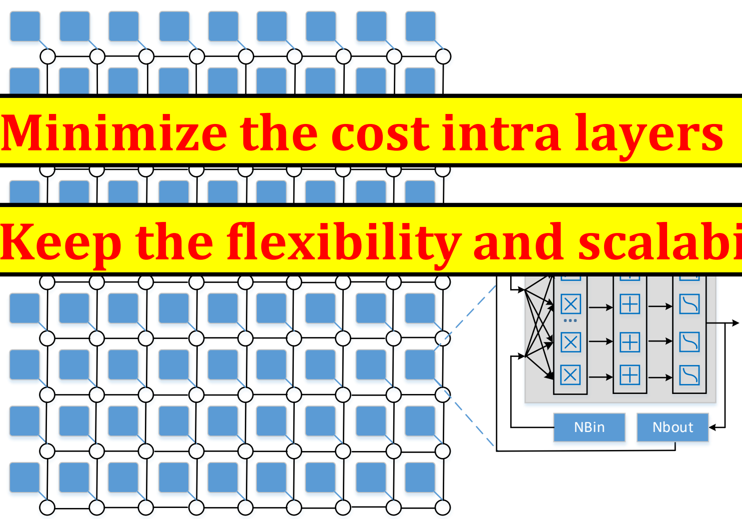


Features:
- **Data traffic only btw layers**
- **Same data sent to adjacent layers**

• *Unbalanced load*



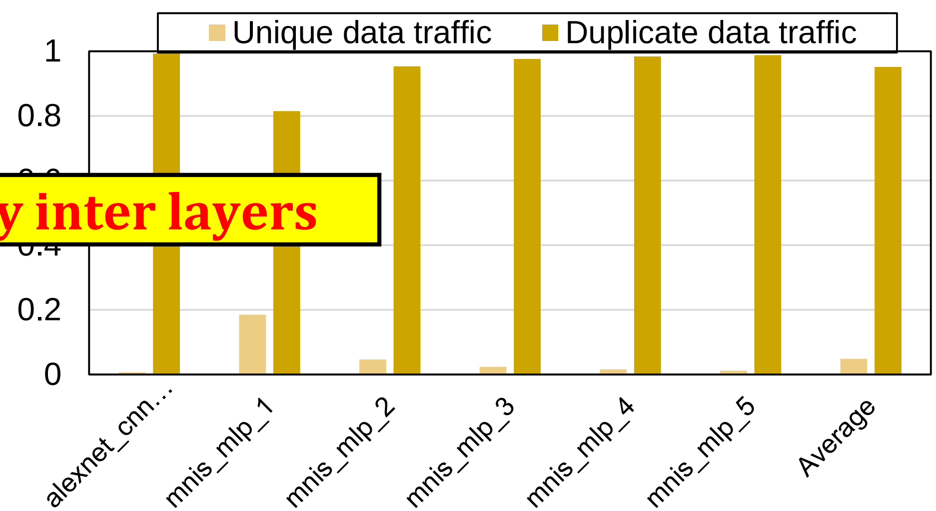
Conventional multicore system



Minimize the cost intra layers

Keep the flexibility and scalability inter layers

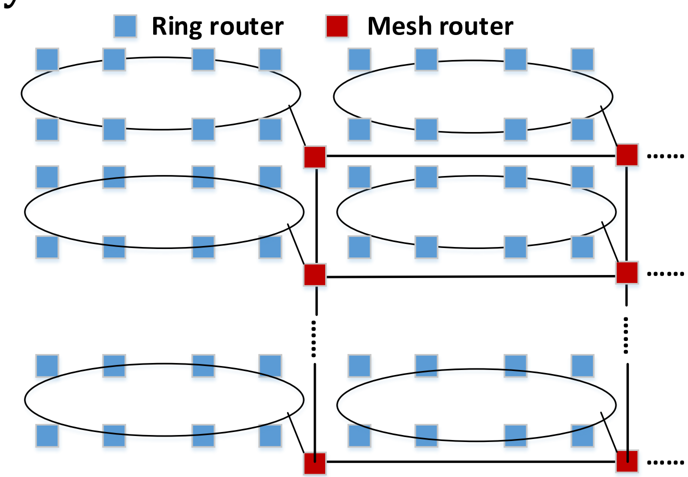
• *Redundant traffic*



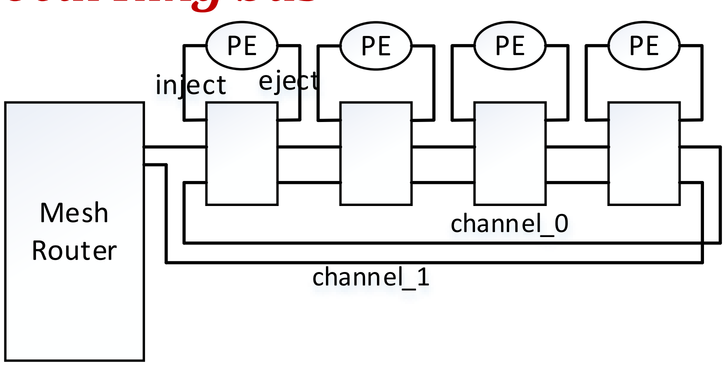
Neu-NoC design

Hierarchical Structure of Neu_NoC

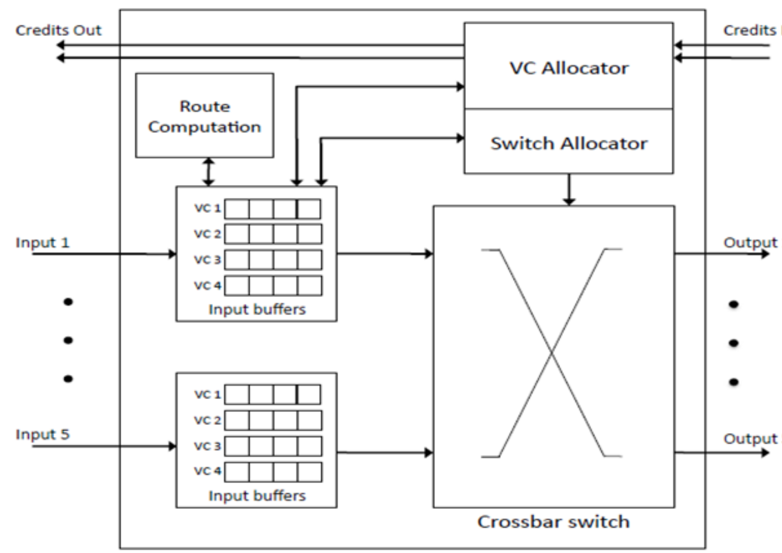
- **Hybrid Ring-Mesh NoC**
- **Simple in local and flexible in global**
- **broadcast to receive data from upper layer**
- **Reduce the number of destination in lower layer**



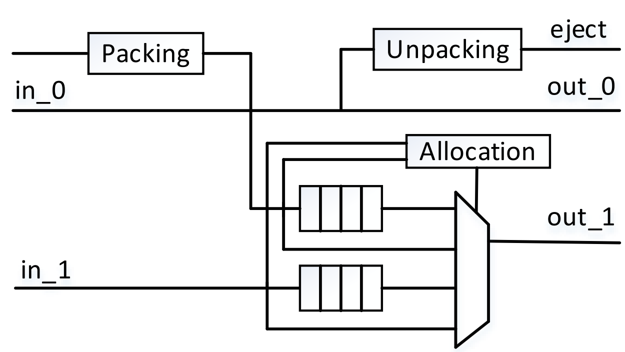
Local Ring bus



Router in Mesh NoC



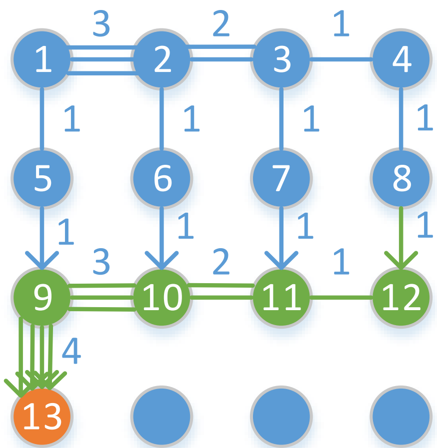
Router in Ring bus



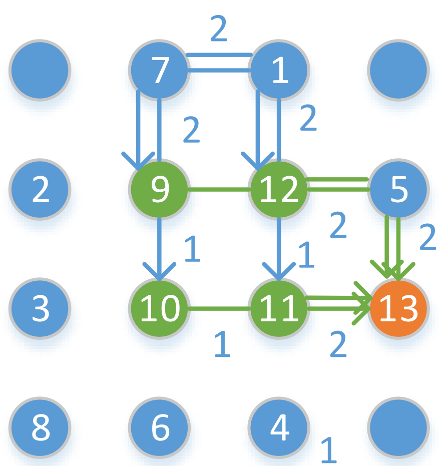
Implementation of Neu_NoC

Mapping influence in Neu_NoC

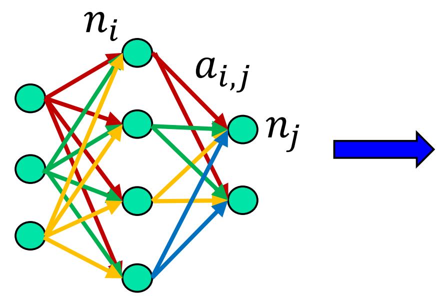
(a) Hop count=124, max load=16



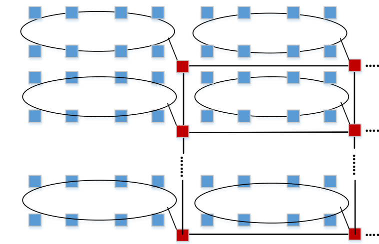
(b) Hop count=76, max load=4



Neural Network-aware mapping



Neural Network Communication Graph $NNCG(N, A)$



An architecture characterization graph $ARCG(U, L)$

Find a $map()$ from $NNCA(N, A)$ to $ARCG(U, L)$ has minimum of hop count and max load

Algorithm:

For $n_i \in NNCG(N, A)$ map n_i to $u_x \in ARCG(U, L)$

generate $map(n_i) \in U$

For each $a_{i,j}$ in A,

count $Hop\ Count = \sum(|row_i - row_j| + |col_i - col_j|)$

Find $map(n_i)$ has $Hop\ Count_{min}$

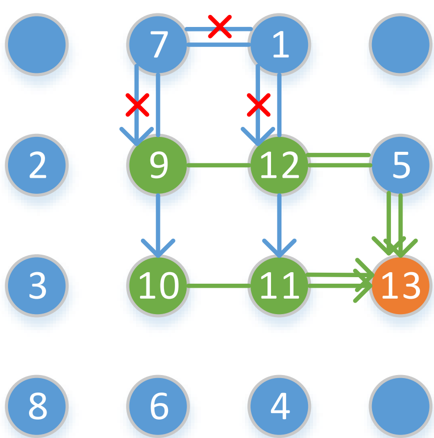
Count $load\ sum\ of\ l_i \in L$ for $map(n_i)$

Output $map(n_i)$ with smallest $load\ sum$

Implementation of Neu_NoC

Benchmark	Topology	Hops count		Max load	
		Random	NN	Random	NN
Mnist_mlp_1	784-300-100-10	12	4	3	1
Mnist_mlp_2	784-1000-500-10	124	76	16	4
Mnist_mlp-3	784-1500-1000-500-10	550	421	24	14
Mnist_mlp_4	784-2000-1500-1000-500-10	1774	1308	36	11
Mnist_mlp_5	784-2500-2000-1500-1000-500-10	3958	3074	27	16
Alexnet_cnn_cla	9216-4096-4096-1000	8218	7024	90	54

Reduce traffic load --- Multicast Transmission



Original	Multicast
1->7->9	
1->7->9>10	1->7->9>10
1->12	
1->12->11	1->12->11
9->12->5->13	9->12->5->13
12->5->13	12->5->13
10->11->13	10->11->13
11->13	11->13

Bit string encoding for multicasting

1->7->9>10:

Router ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Packet Header	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0

Reduce traffic load

Feature map sparsity: reducing traffic

The sparsity analysis of MNIST feature maps

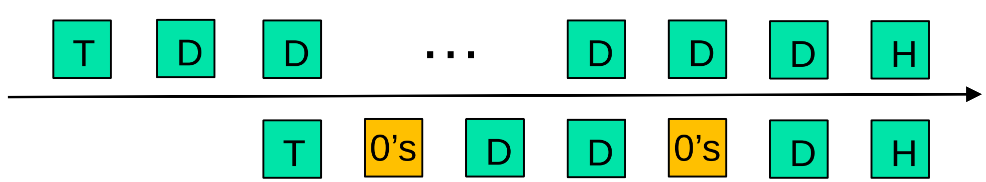
Benchmark	Accuracy	Accuracy drop 1%			Accuracy drop 2%		
		4 0's	8 0's	16 0's	4 0's	8 0's	16 0's
Mnist_mlp_1	98%	43.3%	272.%	4.9%	69.1%	56.3%	18.7%
Mnist_mlp_2	98.27%	52.7%	33.1%	8.2%	60.1%	39.7%	11.0%
Mnist_mlp-3	98.28%	49.7%	26.9%	4.4%	64.8%	44.7%	17.1%
Mnist_mlp_4	98.32%	42.2%	22.0%	4.1%	59.1%	42.6%	12.9%
Mnist_mlp_5	98.22%	43.6%	26.1%	6.7%	63.4%	44.1%	11.3%
Alexnet_cnn_cla	56.60%	58.3%	36.1%	9.7%	64.0%	42.5%	12.7%

Format of data package

H	dest	No	Unused
P	Data		
P	Data		
PP	No. of "0"		
P/PP	...		
T	Unused		

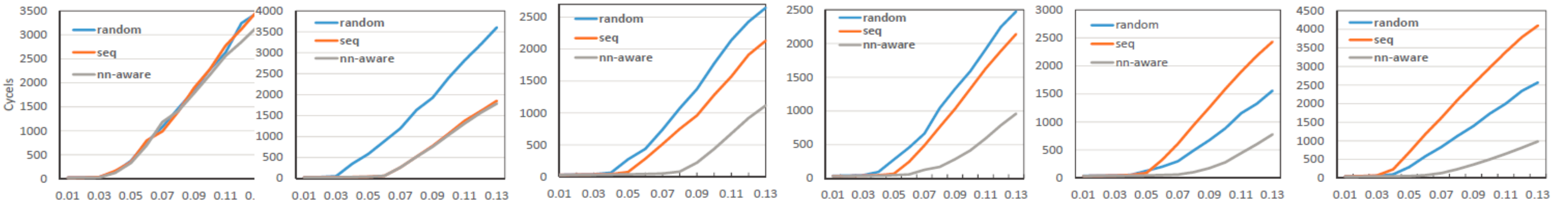
H: head (01), P: payload (10),
PP: packed (11), T: tail (00)

Data traffic reduction based on 0's number



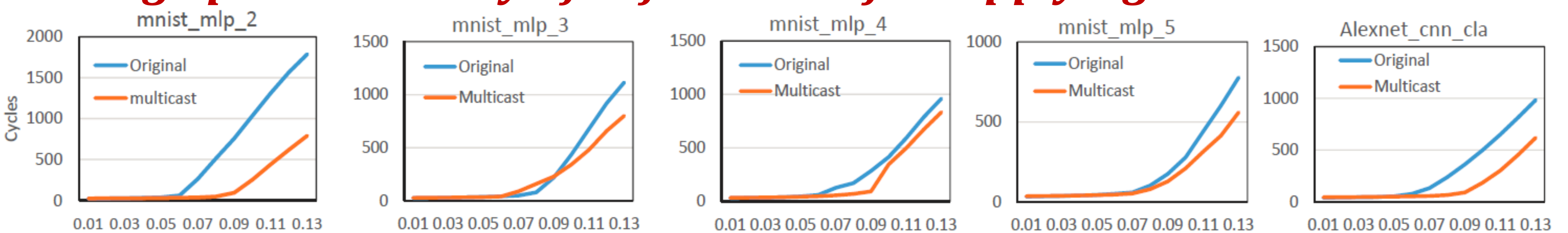
Evaluation of Neu_NoC

Average packet latency of different mappings



(a) mnist_mlp_1 (b) mnist_mlp_2 (c) mnist_mlp_3 (d) mnist_mlp_4 (e) mnist_mlp_5 (f) alexnet_cnn_cla

Average packet latency of before and after applying multicast



Normalized average packet latency and energy

(a) latency

(b) Energy