



**ISCA Tutorial** 

June 19th, 2022

# **Enabling HW/SW Co-Design of Distributed Deep Learning Training Platforms**

ASTRA-sim Tutorial



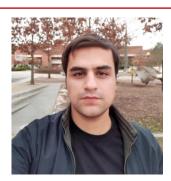
#### **Tushar Krishna**

Associate Professor, School of ECE Georgia Institute of Technology <a href="mailto:tushar@ece.gatech.edu">tushar@ece.gatech.edu</a>

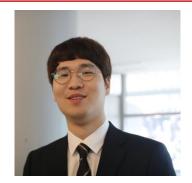
## Welcome



Tushar Krishna
Associate Professor, School of ECE
Georgia Institute of Technology
tushar@ece.gatech.edu



Saeed Rashidi
PhD Student, School of ECE
Georgia Institute of Technology
saeed.rashidi@ece.gatech.edu



Will Won

Ph.D. Student, School of CS
Georgia Institute of Technology
william.won@gatech.edu



Taekyung Heo
Postdoctoral Fellow, School of ECE
Georgia Institute of Technology
tkheo@casys.kaist.ac.kr



Srinivas Sridharan Research Scientist, Meta



**Presenters** 

# Agenda

Time (EDT)	Topic	Presenter
8:30 – 9:30	Introduction to Distributed Deep Learning Training Platforms	Tushar Krishna
9:30 – 10:30	ASTRA-sim	Saeed Rashidi
10:30 - 11:00	Coffee Break	
11:00 – 11:50	Demo and Exercises	William Won and Taekyung Heo
11:50 – 12:00	Extensions and Future Development	Taekyung Heo

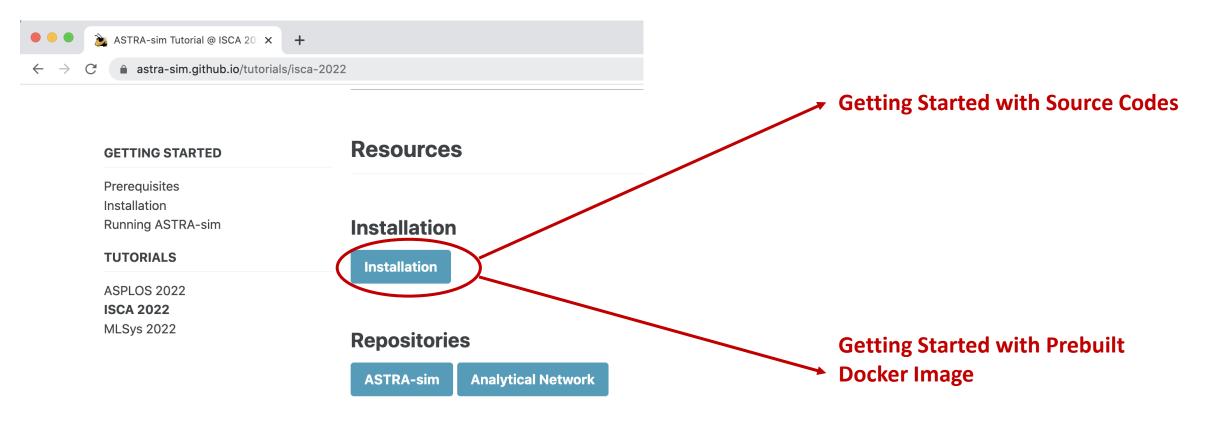
#### **Tutorial Website**

includes agenda, slides, ASTRA-sim installation instructions (via source + docker image) <a href="https://astra-sim.github.io/tutorials/isca-2022">https://astra-sim.github.io/tutorials/isca-2022</a>

**Attention:** Tutorial is being recorded

### **ASTRA-sim Installation**

- Please go ahead and install ASTRA-sim!
- Instructions here:



# Agenda

Time (EDT)	Topic	Presenter
8:30 – 9:30	Introduction to Distributed Deep Learning Training Platforms	Tushar Krishna
9:30 – 10:30	ASTRA-sim	Saeed Rashidi
10:30 - 11:00	Coffee Break	
11:00 – 11:50	Demo and Exercises	William Won and Taekyung Heo
11:50 – 12:00	Extensions and Future Development	Tushar Krishna and Saeed Rashidi

#### **Tutorial Website**

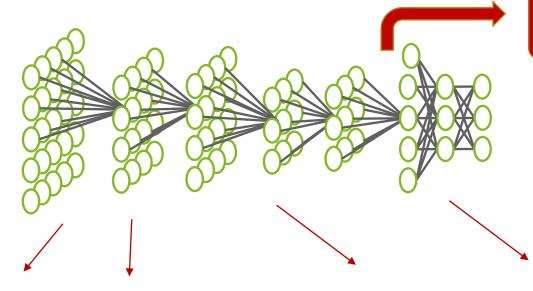
includes agenda, slides, ASTRA-sim installation instructions (via source + docker image) <a href="https://astra-sim.github.io/tutorials/isca-2022">https://astra-sim.github.io/tutorials/isca-2022</a>

**Attention:** Tutorial is being recorded

# The engine driving the AI Revolution







#### **Training**

Training a deep neural network (DNN) involves feeding it a training dataset to update its weights to model the underlying function representing the dataset



**Object Detection** 



**Speech Recognition** 



**Understanding** 



**Recommender Systems** 

# "Training" in the context of ML

- Machine Learning algorithms learn to make decisions or predictions based on data.
- We can categorize current ML algorithms based on the following three characteristics
  - Feedback from data
    - Supervised learning
    - Unsupervised learning
    - Semi-supervised learning
    - Reinforcement learning
  - Purpose / Task
    - Anomaly Detection
    - Classification
    - Clustering
    - Dimensionality Reduction
    - Representation Learning
    - Regression
  - Method (for hyperparameter optimization)
    - SGD
    - EA
    - Rule-based
    - Topic Models
    - ..

We focus on Supervised Learning with SGD --> most popular for DNNs

Source: A Survey on Distributed Machine Learning <a href="https://dl.acm.org/doi/abs/10.1145/3377454">https://dl.acm.org/doi/abs/10.1145/3377454</a>

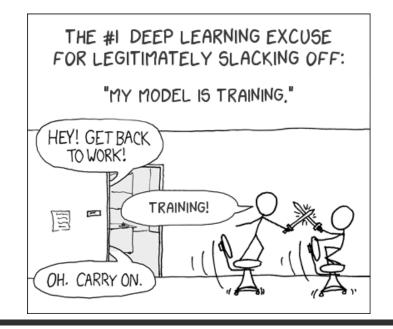
# DL Training: The Phases

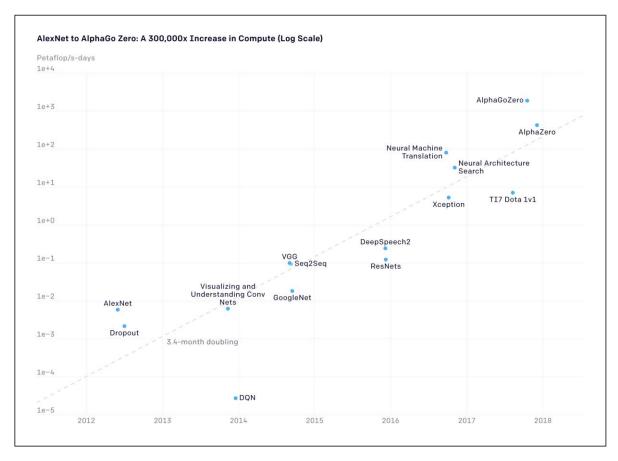
- Each training algorithm consists of 3 computation phases:
  - 1. Forward pass (inference):
    - The process of finding output activations using inputs and weights.
  - 2. Weight gradient computation:
    - The process of finding the gradient of weights (with respect to the loss function) using output gradients and inputs.
  - 3. Input gradient computation:
    - The process of finding the gradient of inputs (with respect to the loss function) using output gradients and weights.
- Operations 2 & 3 together are called backpropagation.
- The **training loop** dictates the order in which we issue the basic operations and (possibly) their related communication tasks.

# Deep Learning Training Challenge

#### Training time is increasing

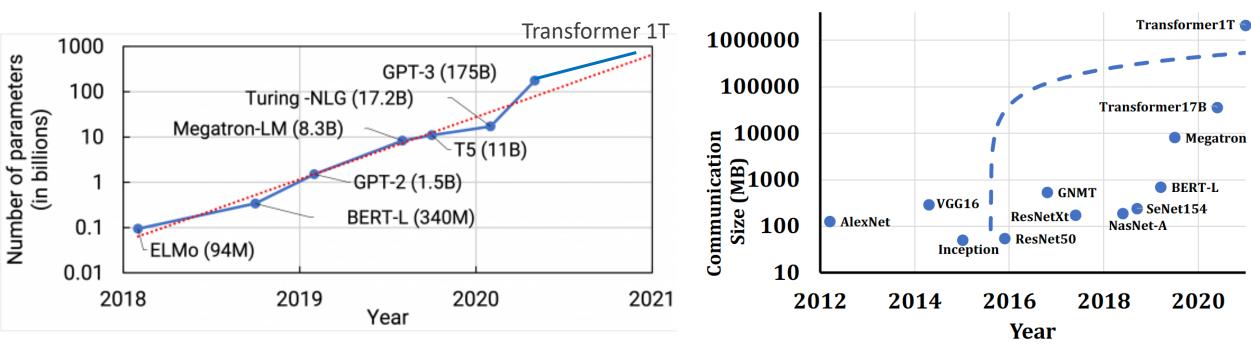
- DNNs are becoming larger
  - Turing NLG: 17.2 B Parameters
  - Megatron LM: 8.3B Parameters
- Training samples are becoming larger
- Moore's Law has ended





Source: <a href="https://openai.com/blog/ai-and-compute/">https://openai.com/blog/ai-and-compute/</a>

# Key Challenge: Large Models → Large Comms



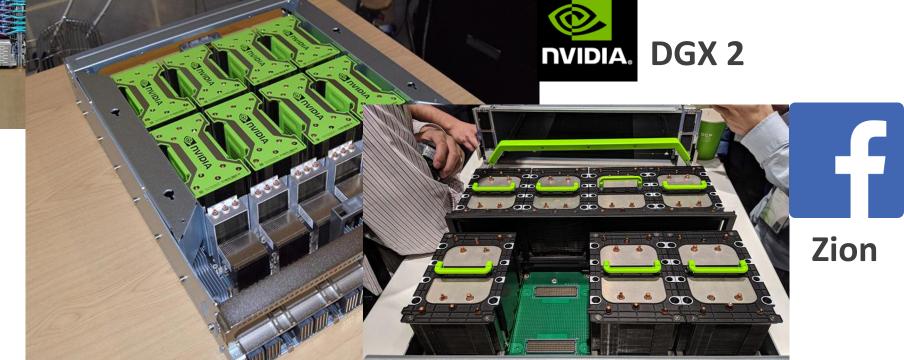
#### **Challenges:**

- Multiple NPUs are required to fit large-scale models
- e.g., 16 NPUs for GPT-3 (175B params)
   128 NPUs for Transformer-1T (1T params) (using ZeRO stage 2)

# **Enter: DL Training Platforms**



- ✓ Build customized chips for training
- ✓ Scale the training across more compute nodes

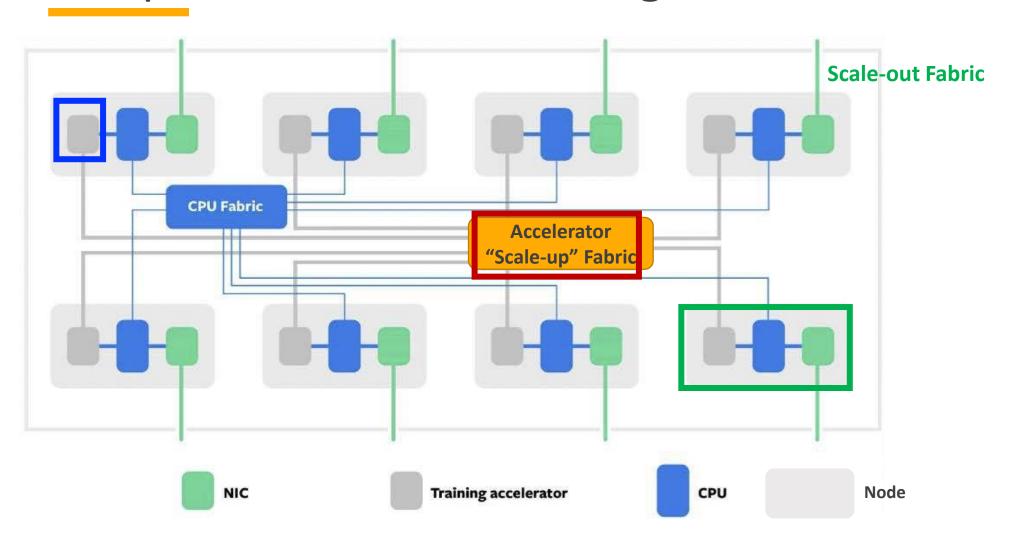


# And many more ...

- Cerebras CS2
- Tesla Dojo
- NVIDIA DGX + Mellanox SHARP switches
- Intel Habana
- IBM Blueconnect

•

# Components of a DL Training Platform



Modified version of source figure from: "Zion: Facebook Next- Generation Large Memory Training Platform", Misha Smelyanskiy, Hot Chips 31"

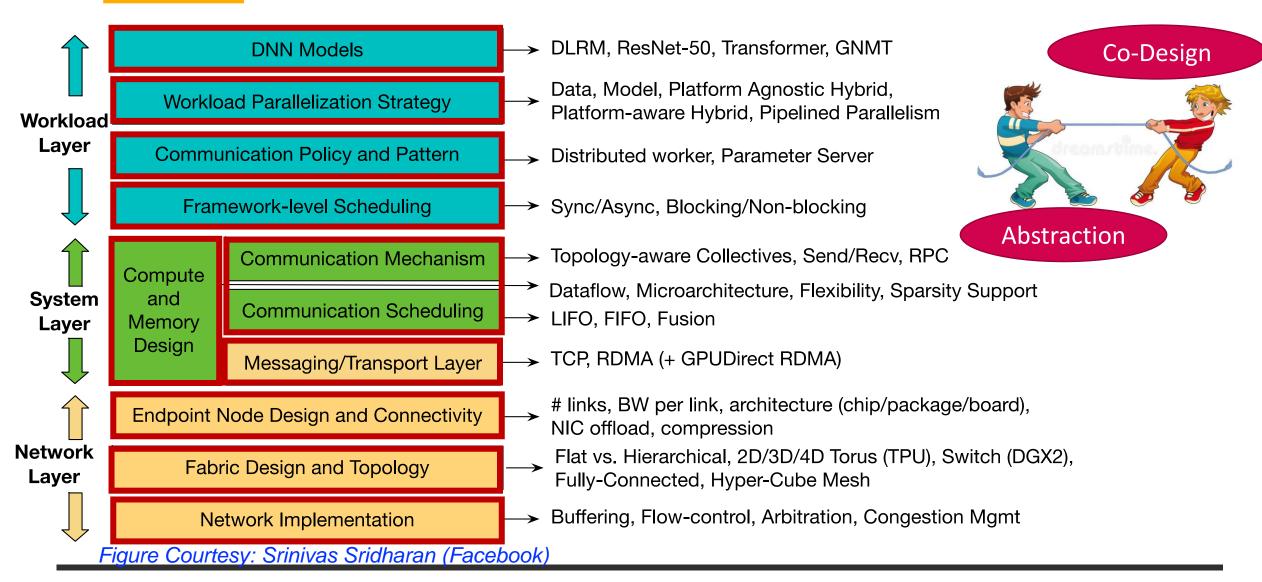
# Systems challenges with Distributed Training

- Communication!
  - Inevitable in any distributed algorithm

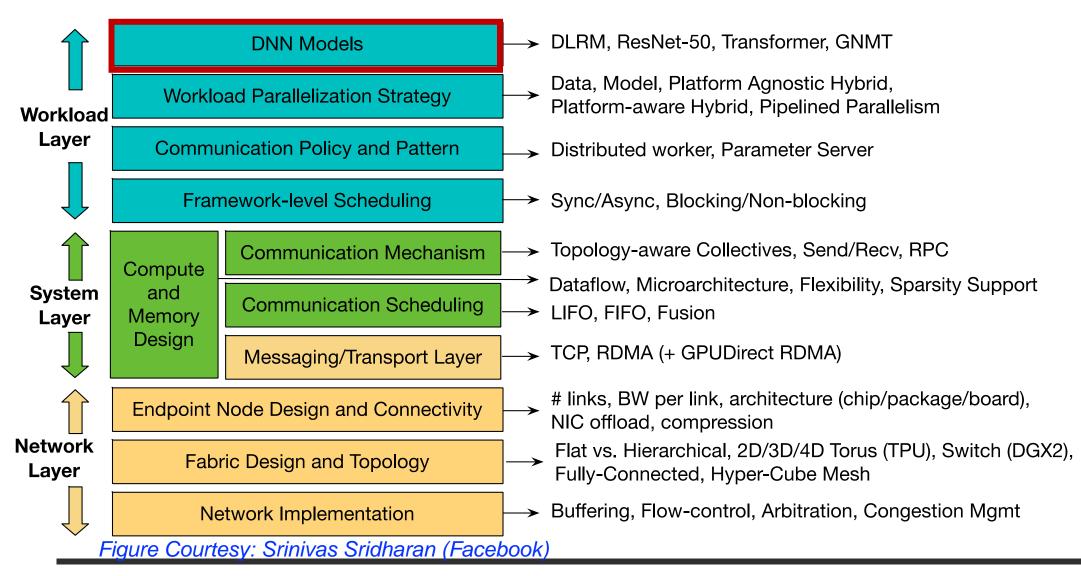
- What does communication depend on?
  - synchronization scheme: synchronous vs. asynchronous.
  - parallelism approach: data-parallel, model-parallel, hybrid-parallel., ZeRO ...

- Is it a problem?
  - Depends ... can we hide it behind compute?
  - How do we determine this?

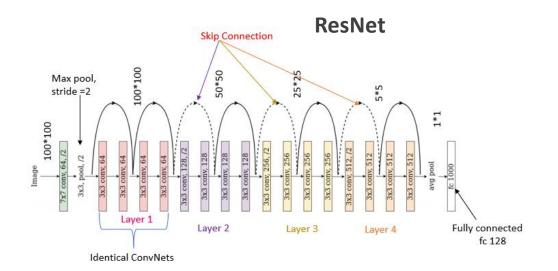
# Understanding DL Training design-space

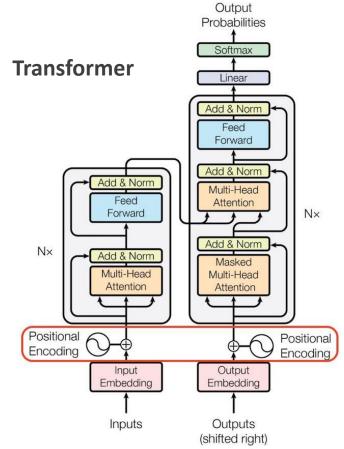


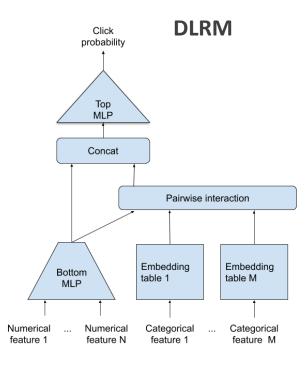
# Distributed Training Stack



## **DNN Models**



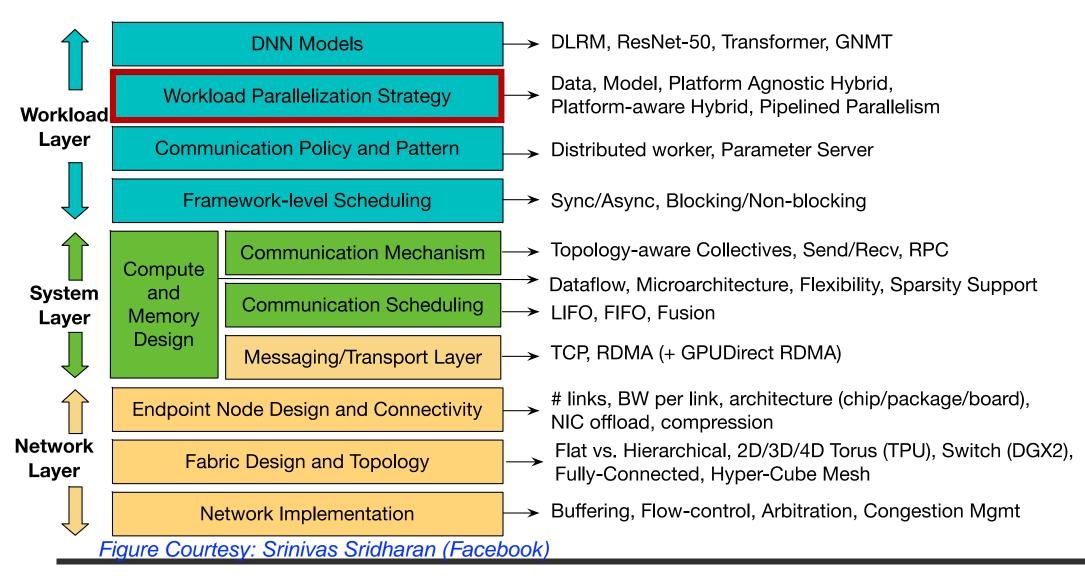




Layer Types: CONV2D, Attention, Fully-Connected, ...

Parameter sizes: Millions to Trillions

# Distributed Training Stack



## Parallelization Strategies

- The way compute tasks are distributed across different compute nodes. Multiple ways to split the tasks:
  - Split the minibatch (Data-Parallel)
  - Split the model (Model-Parallel)
  - Split the DNN layers: (Pipeline-Parallel)
  - •
- This also defines the communication pattern across different nodes.

## Parallelism: Data-Parallel Training

- Distribute Data across multiple nodes and replicate model (network) along all nodes.
- No communication during the forward pass.

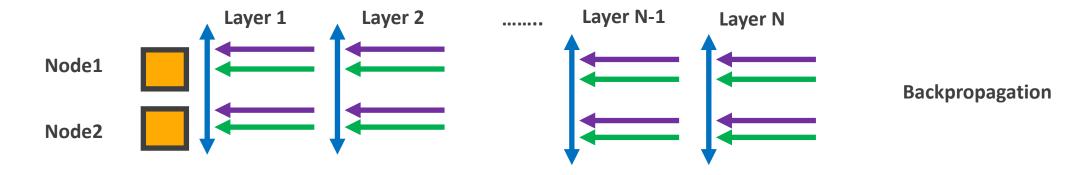


Flow-per-layer: 1.Compute output -> 2. go to the next layer



## Parallelism: Data-Parallel Training

- Distribute Data across multiple nodes and replicate model (network) along all nodes.
- Communicate weight gradients during the backpropagation pass.
  - Blocking wait during forward pass for collective of previous backpropagation for that layer.

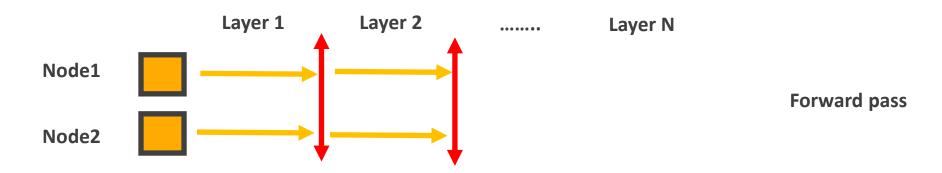


Flow-per-layer: 1.Compute weight gradient-> 2.issue weight gradient comm -> 3.compute input gradient -> 4. go to previous layer

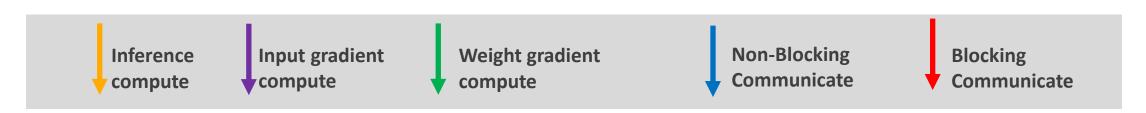


# Parallelism: Model-Parallel Training

- Distribute Model across all nodes and replicate data along all nodes.
- Communicate outputs during the forward pass.

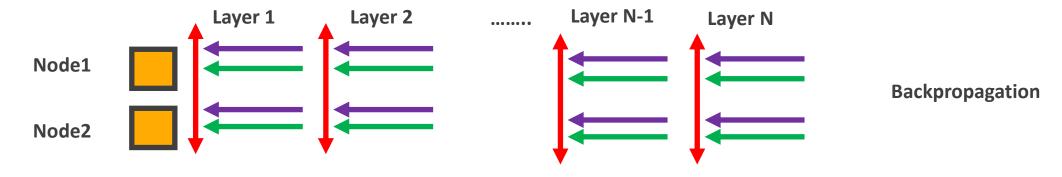


Flow-per-layer: 1.Compute output -> 2. issue output gradient comm -> 3.wait for gradient to be finished -> 4. go to the next layer



# Parallelism: Model-Parallel Training

- Distribute Model across all nodes and replicate data along all nodes
- Communicate input gradients during the backpropagation pass.

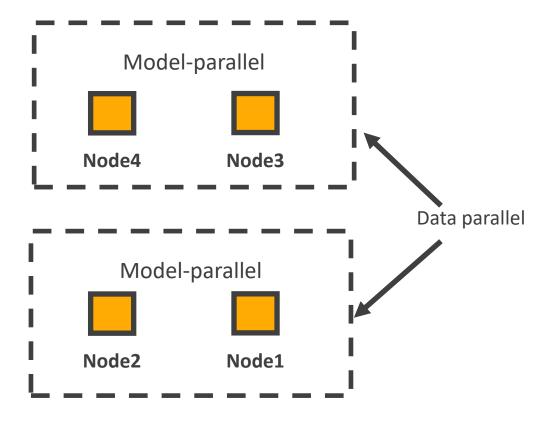


Flow-per-layer: 1.Compute input gradient-> 2.issue input gradient comm -> 3.compute weight gradient -> 4. wait for input gradient -> 5. go to previous layer



## Parallelism: Hybrid Parallel

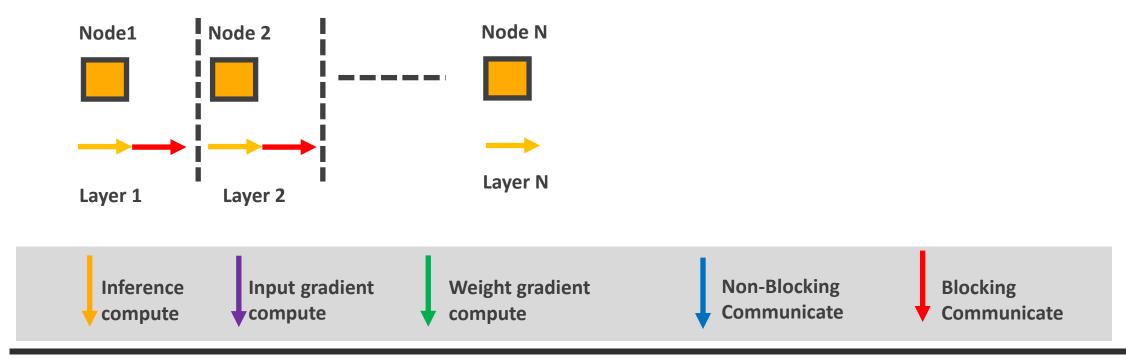
• Partition nodes into groups. Parallelism within a group is modelparallel, across the groups is data-parallel, or vice versa.



Parallelism	Activations during the forward pass	Weight gradients	Input gradients
Data		<b>✓</b>	
Model	<b>✓</b>		<b>✓</b>
Hybrid	partially	partially	partially

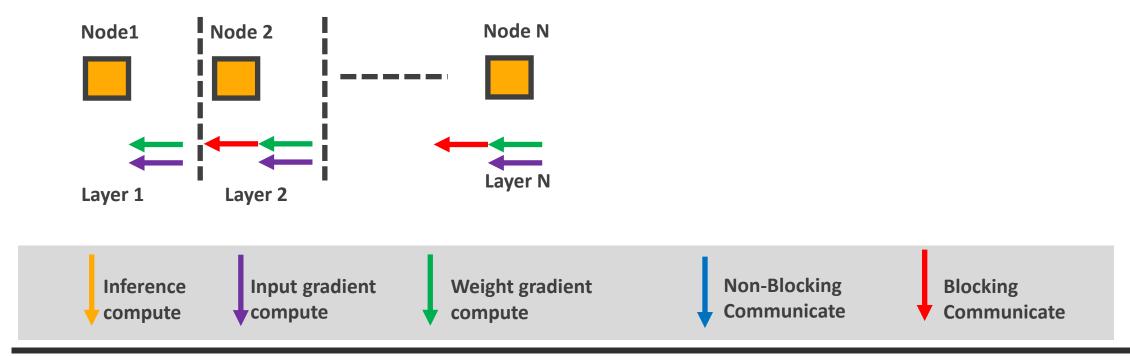
# Parallelism: Pipelined Parallel

- Distribute DNN layers across all nodes.
- Decompose minibatch into microbatches and propagate them to the pipeline in-order.
- Communicate outputs during the forward pass.



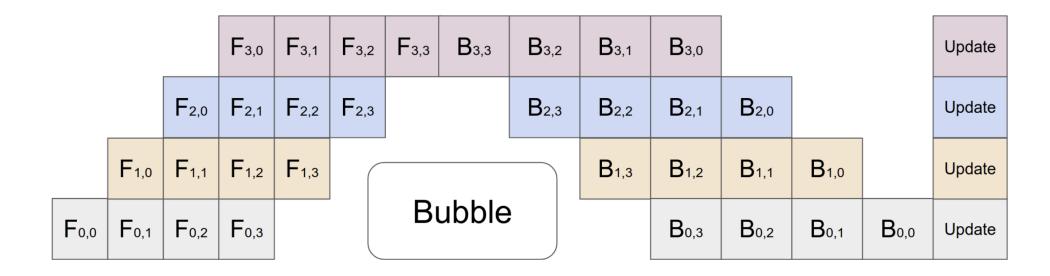
# Parallelism: Pipelined Parallel

- Distribute DNN layers across all nodes.
- Decompose minibatch into microbatches and propagate them to the pipeline in-order.
- Communicate input gradients during the backpropagation.



# Parallelism: Pipelined Parallel

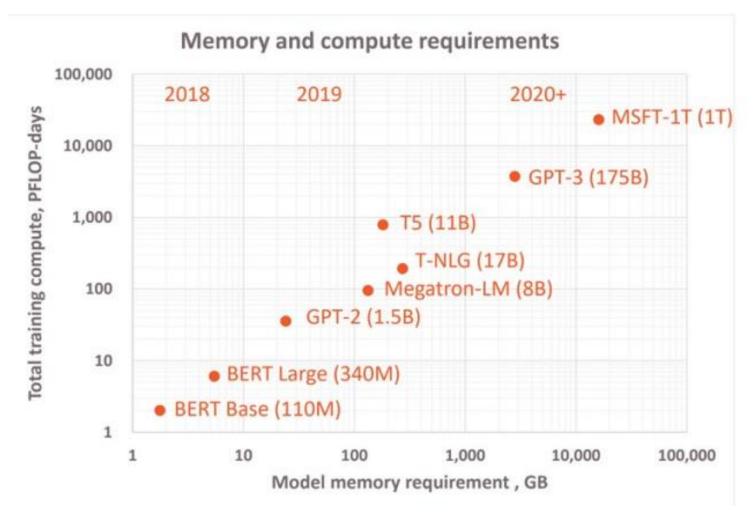
• How a minibatch is broken into micro-batches and pipeline is filled.



 $F_{m,n}$ : forward-pass corresponding to micro-batch #n at device #m.

B  $_{\rm m,n}$ : back-propagation corresponding to micro-batch #n at device #m.

# Need for more sophisticated schemes ...



1000x larger models 1000x more compute In just 2 years

**Today**, GPT-3 with 175 billion params trained on 1024 GPUs for 4 months.

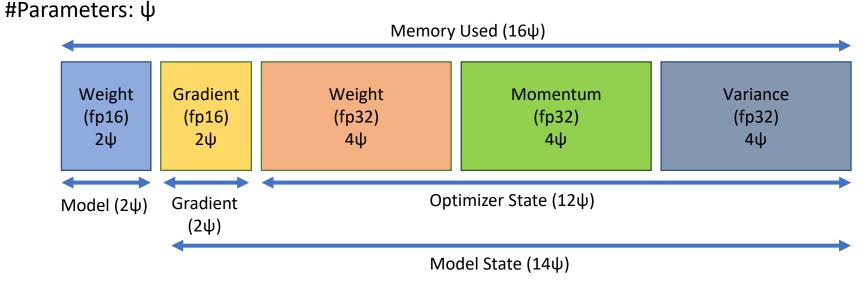
Tomorrow, multi-Trillion parameter models and beyond.

Source: Cerebras (Hot Chips 2021)

# Example 1: Microsoft ZeRO

#### Motivation

- Data Parallelism (DP): Cannot fit large models
- Model Parallelism (MP): Computations too fine-grained, Large communication overhead, Layer-dependent design
- Large Memory Overhead for Model + Optimizer state
  - 8x overhead over model state!



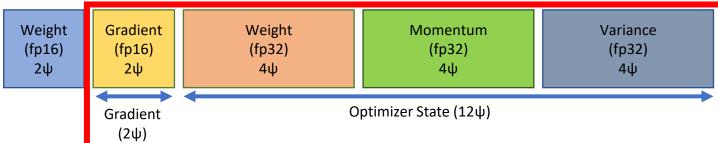
https://www.microsoft.com/en-us/research/blog/zero-deepspeed-new-system-optimizations-enable-training-models-with-over-100-billion-parameters/

# Example 1: Microsoft ZeRO

ZeRO: Zero Redundancy Optimizer

Reduce redundant Model State

- Partition Optimizer state
- Partition Gradient state
- Memory vs Communication



l .. . . . .

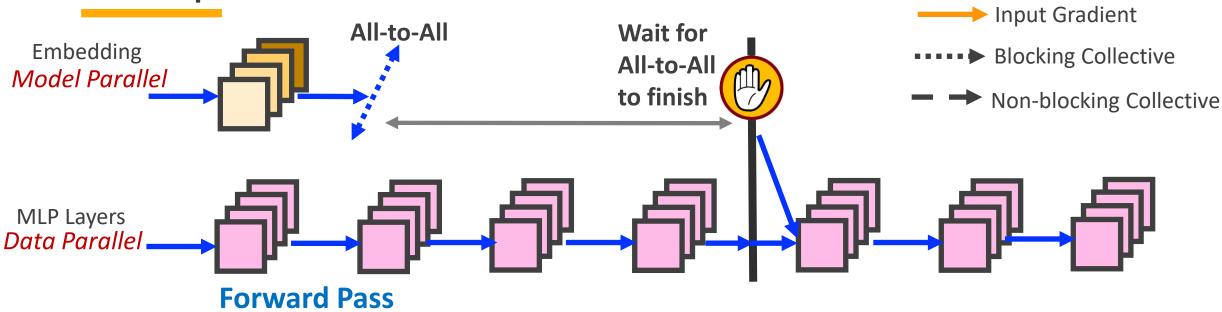
	gpu <sub>0</sub>		gpu <sub>i</sub>		gpu <sub>N-1</sub>	Memory Consumed	K=12 Ψ=7.5B N <sub>d</sub> =64
Baseline		•••		•••		$(2+2+K)*\Psi$	120GB
P <sub>os</sub>			T	•••		$2\mathbf{\Psi} + 2\mathbf{\Psi} + \frac{K * \mathbf{\Psi}}{N_d}$	31.4GB
P <sub>os+g</sub>				•••		$2\Psi + \frac{(2+K)*\Psi}{N_d}$	16.6GB
P <sub>os+g+p</sub>				•••		$\frac{(2+2+K)*\Psi}{N_d}$	1.9GB
	Parameters	5 = 0	Gradients		Optimizer Stat	es	

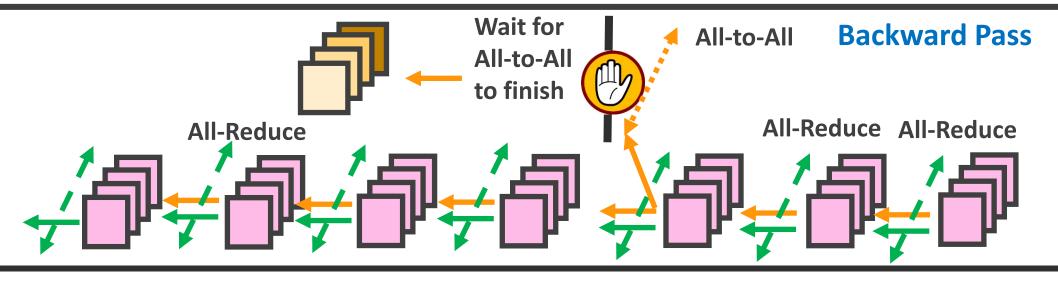
https://www.microsoft.com/en-us/research/blog/zero-deepspeed-new-system-optimizations-enable-training-models-with-over-100-billion-parameters/

Output Activation

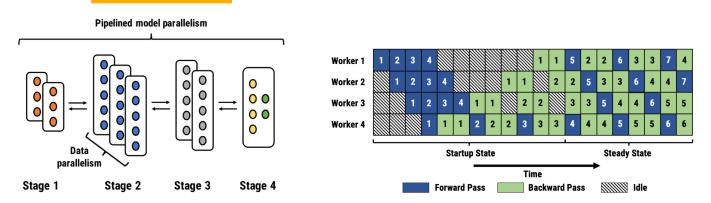
Weight Gradient

# Example 2: Facebook DLRM

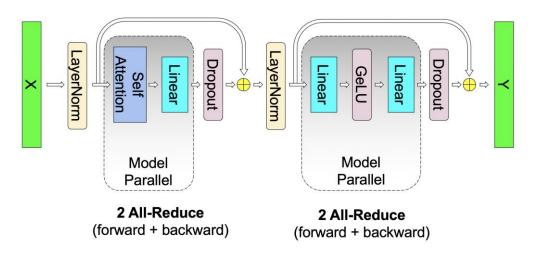




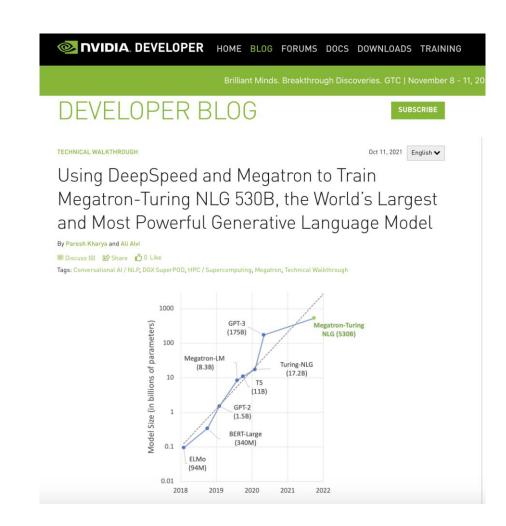
## More recent examples



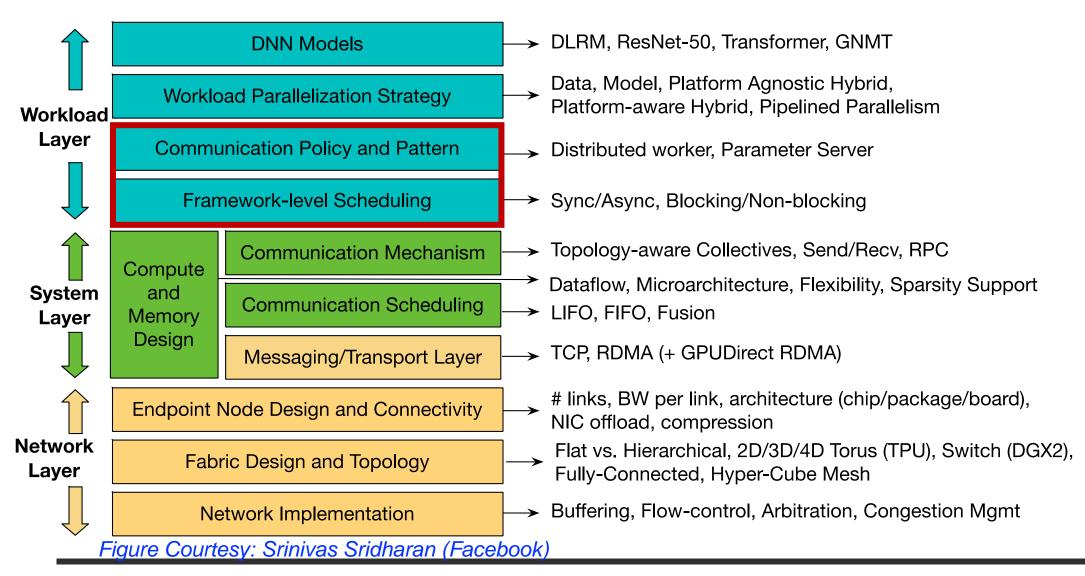
#### **PipeDream (Microsoft)**



#### MegatronLM (NVIDIA)



## Distributed Training Stack

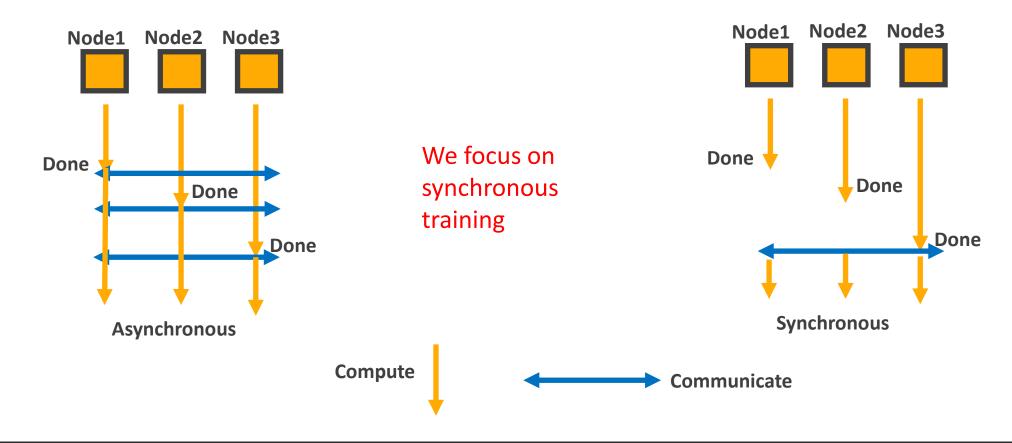


# Model Parameter Update Mechanisms

		Synchronization		
		Asynchronous	Synchronous	
Communication Handling	Parameter-server	Centralized or Distributed	Centralized or Decentralized	
	Collective-based	N/A	Distributed	

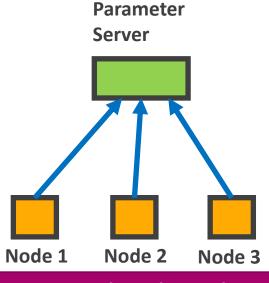
# Synchronization: Sync. vs. Async. Training

- Defines when nodes should exchange data
  - Affects convergence time

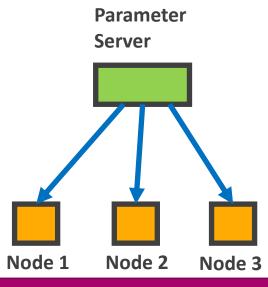


## **Communication Handling**

#### Parameter Server



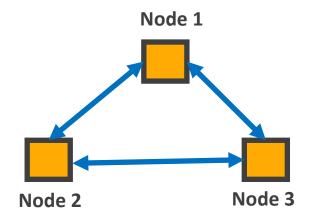
Step 1: Each node sends its model gradients to the parameter server to be reduced with other gradients and update the model

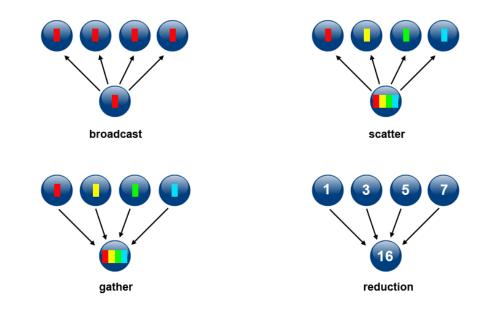


Step 2: The parameter server sends the updated model to the compute nodes to begin the new iteration.

### **Communication Handling**

• Collective-based: Compute Nodes directly talk to each other to globally reduce their gradients and update the model through *All-Reduce* communication pattern.





"Collective Communication" (from MPI)

More details later

#### **Exchanging Output Activations or Input Gradients:**

- It may be required depending on the **parallelization strategy** (discussed next)
- Handled either via collective based patterns or direct Node-to-Node sends/recvs (no parameter server is used).

### When are collectives needed?

	<b>Model Updates</b>	Input Gradient Exchange	Output Activation Exchange
Param-server	N	Data-parallel: <b>N</b> Model-parallel: <b>Usually</b> * Pipeline-Parallel: <b>N</b>	Data-parallel: <b>N</b> Model-parallel: <b>Usually</b> * Pipeline-Parallel: <b>N</b>
Collective-based	Y (All-Reduce)	Data-parallel: <b>N</b> Model-parallel: <b>Usually</b> * Pipeline-Parallel: <b>N</b>	Data-parallel: <b>N</b> Model-parallel: <b>Usually</b> * Pipeline-Parallel: <b>N</b>

<sup>\*</sup> All-reduce, All-gather, Reduce-scatter, All-to-All

## Different Kinds of Collective Algorithms

#### Reduce-Scatter:

- Used during input-output exchange due to model-parallelism
- Implementation Algorithms: Ring-Based, Direct-based, etc.

#### All-Gather:

- Used during input-output exchange due to model-parallelism
- Implementation Algorithms: Ring-Based, Direct-based, etc.

#### All-Reduce (Reduce-Scatter + All-Gather):

- Used during input-output exchange due to model-parallelism, or during model-parameter update.
- Implementation Algorithms: Ring-Based, Direct-based, Tree-based, Halving-doubling, etc..

#### • All-To-All:

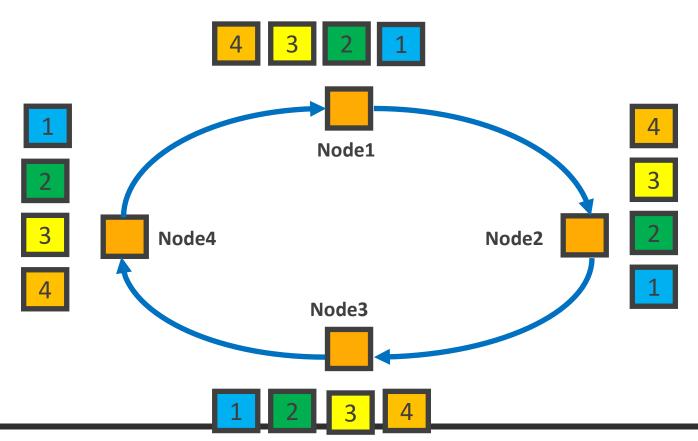
- Used during input-output exchange due to model-parallelism (e.g., distributed embedding layer on DLRM DNN.).
- Implementation Algorithms: Direct-based, Ring-Based, etc..

Node	Node	Node	Node	Node	Node	Node	Node
0	1	2	3_	0	1	2	3
$X_0^{(0)}$	$X_0^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$	$\sum_{j} X_0^{(j)}$	)		
$X_{1}^{(0)}$	$X_{1}^{(1)}$	$X_1^{(2)}$	$X_1^{(3)}$ _	-	$\sum_{j} X_{1}^{(j)}$	)	
$X_{2}^{(0)}$	$X_{2}^{(1)}$	$X_{2}^{(2)}$	$X_{2}^{(3)}$			$\sum_{j} X_2^{(j)}$	
$X_3^{(0)}$		$X_3^{(2)}$	$X_3^{(3)}$ Re	educe catter			$\sum_{j} X_3^{(j)}$
		Node	Node		Node	Node,	Node
0	1_	2	3_	_ 0_	$_{1}$	2	3
X0				X0	X0	X0	<i>X</i> 0
	<i>X</i> 1			<b>→</b> X1	<i>X</i> 1	<i>X</i> 1	X1
		<i>X</i> 2		<i>X</i> 2	<i>X</i> 2	<i>X</i> 2	X2
			<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	X3
	ı	I	l All-ga	other	<b> </b>	ı	

Node	Node	Node	Node	Node	Node	Node	Node
_0_	1	2	3	0	1	2	3_
$X_0^{(0)}$	$X_0^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$	$\sum_{j} X_0^{(j)}$			
$X_{1}^{(0)}$	$X_{1}^{(1)}$	$X_{1}^{(2)}$	$X_1^{(3)}$ .	$\sum_{j} X_{1}^{(j)}$			
$Y^{(0)}$	$X_{2}^{(1)}$	$X^{(2)}$	$X_{2}^{(3)}$				$\sum_{j} X_2^{(j)}$
$X_{3}^{(0)}$	$X_{3}^{(1)}$	$X_{3}^{(2)}$	$X_{3}^{(3)}$	$\sum_j X_3^{(j)}$	$\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{(j)}$
3			' All-1	reduce		٠	'

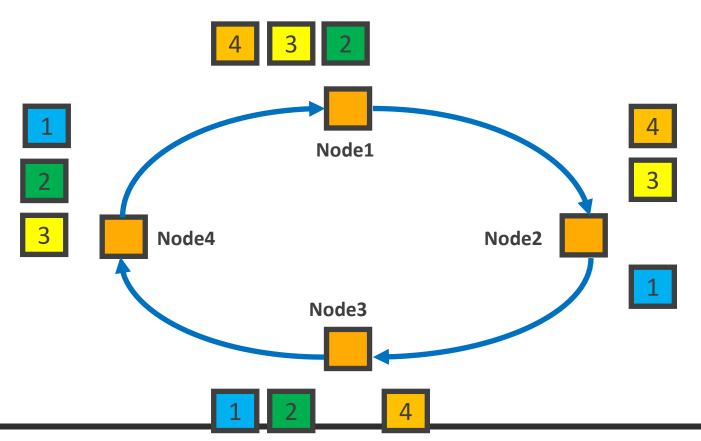
Noue	noue	Noue	noue	Noue	noue	Noue	Noue
0	1	2	3	0	1	2	3
$X_0^{(0)}$	$X_0^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$	$X_0^{(0)}$	$X_1^{(0)}$	$X_2^{(0)}$	$X_3^{(0)}$
1		$X_1^{(2)}$		$X_0^{(1)}$	$X_1^{(1)}$	$X_{2}^{(1)}$	$X_3^{(1)}$
		$X_2^{(2)}$	$X_2^{(3)}$	$X_0^{(2)}$	$X_1^{(2)}$	$X_2^{(2)}$	$X_3^{(2)}$
$X_3^{(0)}$	$X_3^{(1)}$	$X_3^{(2)}$	$X_{3}^{(3)}$	$X_0^{(3)}$	$X_1^{(3)}$	$X_2^{(3)}$	$X_3^{(3)}$
			AII-	to-an			

- A ring with N nodes partitions data to N messages
- Collective Communication Flow:



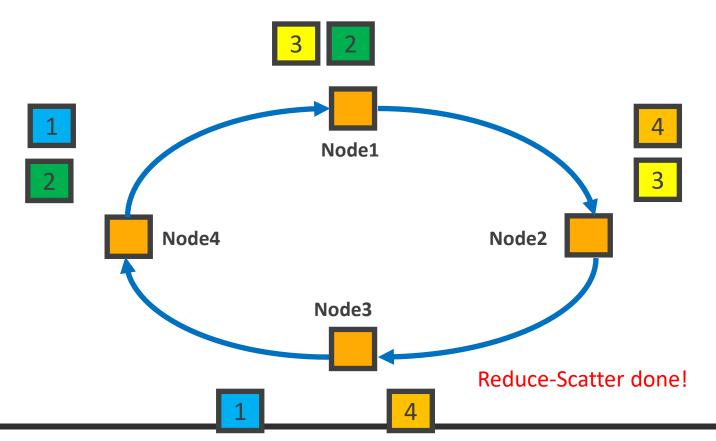
Node	Node	Node	Node	Node	Node	Node	Node
0	1_	2	3	_0	1	2	3
$X_0^{(0)} X_1^{(0)}$	$X_0^{(1)} \ X_{\cdot}^{(1)}$	$X_0^{(2)} \ X_{\cdot}^{(2)}$	$X_0^{(3)}$ $X_1^{(3)}$	$\sum_{j} X_0^{(j)}$	$\sum_{i} X_{1}^{(j)}$		
$X_{2}^{(0)} X_{3}^{(0)}$	$X_{2}^{(1)}$	$X_{2}^{(2)}$	$X_2^{(3)}$	educe		$\sum_{j} X_2^{(j)}$	$\sum_{i} X_{3}^{(}$
-	$A_3$	$A_3$	1 3 -se	catter	 	 	<b>~</b>
	Node	Node			Node	- 1	
_0	1	2	3	_0_	1	2	3
X0				X0	X0	X0	X0
	<i>X</i> 1		→	<b>→</b> X1	<i>X</i> 1	<i>X</i> 1	X1
		<i>X</i> 2		<i>X</i> 2	<i>X</i> 2	<i>X</i> 2	X2
			<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	X3
	ı	ı	All-ga	ather I	ı	I	
Node	Node	Node	Node	Node	,Node	Node	Node
0	1	2	3	0	1	2	3
$Y^{(0)}$	$X_{0}^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$	$\sum_{i} X_0^{(j)}$	$\sum_{i} X_0^{(j)}$	$\sum_{i} X_0^{(j)}$	$\sum_{i} X_0^{(j)}$
$V^{(0)}$	$Y^{(1)}$	$X^{(2)}$	$X_1^{(3)}$ -	$\sum_{i=1}^{n} X_{i}^{(j)}$	$\sum_{j} X_{1}^{(j)}$	$\sum_{i} X_{1}^{(j)}$	$\sum_{i}^{j} X_{i}^{(j)}$
$egin{array}{c} oldsymbol{\Lambda}_1 \\ oldsymbol{V}^{(0)} \end{array}$	$Y^{(1)}$	$X_1 X^{(2)}$	$\mathbf{v}_{(3)}$	$\sum_{j} X_{2}^{(j)}$	$\sum_{i} X_{2}^{(j)}$	$\sum_{i}^{j} X_{2}^{(j)}$	$\sum_{i=1}^{n} X_{i}^{(i)}$
$X_2^{(0)}$	$\mathbf{v}^{(1)}$	$X_2$ $Y^{(2)}$	$V^{(3)}$	$\sum_{j}^{j} X_{2}^{(j)}$	$\sum_{i=1}^{j} X_{i}^{(j)}$	$\sum_{i}^{j} X_{i}^{(j)}$	$\sum_{j=1}^{j} X_{2}^{j}$
$X_3^{(0)}$	$\Lambda_3$	$ X_3^{(2)} $	$\Lambda_{2}$	educe	$\angle_{j}^{-13}$	$\angle_{j}^{-13}$	<b>_</b> j - 3

- A ring with N nodes partitions data to N messages
- Collective Communication Flow:



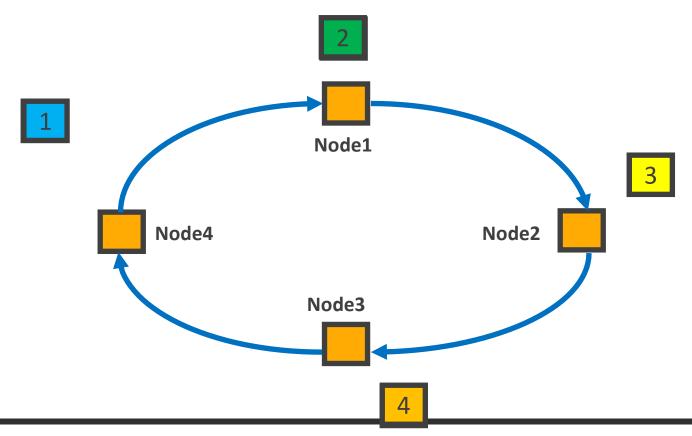
Node	Node	Node	Node	Node	Node	Node	Node
0_	1	2	3_	0	1	2	3
$X_0^{(0)} X_1^{(0)}$	$X_0^{(1)} \ X_1^{(1)} \ Y^{(1)}$	$X_0^{(2)} \ X_1^{(2)} \ Y^{(2)}$	$X_0^{(3)}$ $X_1^{(3)}$ $X_2^{(3)}$	$\sum_{j} X_0^{(j)}$	$\sum_j X_1^{(j)}$	$\sum X_{s}^{(j)}$	
$X_2^{(0)} X_3^{(0)}$	$X_{2}^{(1)}$ $X_{3}^{(1)}$	$X_{2}^{(2)} X_{3}^{(2)}$	$X_{2}^{(3)}$ Re	educe catter		<b></b> ij 2	$\sum_{j} X_3^{(}$
Node	Node	Node			Node,	Node,	Node
0	1	2	3_	_0_	$_{1}$	2	3
X0				X0	X0	X0	<i>X</i> 0
	<i>X</i> 1		_ →	- X1	<i>X</i> 1	<i>X</i> 1	X1
		<i>X</i> 2		<i>X</i> 2	<i>X</i> 2	<i>X</i> 2	<i>X</i> 2
			<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	X3
			All-ga	ther	ı	ı	
Node	Node	Node	Node	Node	Node	Node	Node
0_	_1_	2	3	0_	$\downarrow 1$	2	3
$X_0^{(0)}$	$X_0^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$	$\sum_j X_0^{(j)}$	$\sum\nolimits_{j} X_{0}^{(j)}$	$\sum_{j} X_0^{(j)}$	$\sum_j X_0^{(j)}$
$X_1^{(0)}$	$X_{1}^{(1)}$	$X_1^{(2)}$	$X_1^{(3)}$ -	$\sum_{j} X_1^{(j)}$	$\sum_{j} X_{1}^{(j)}$	$\sum_{j} X_1^{(j)}$	$\sum_{j} X_{1}^{(j)}$
$X_{2}^{(0)}$	$X_{2}^{(1)}$	$X_{2}^{(2)}$	$X_2^{(3)}$	$\sum_{j} X_2^{(j)}$	$\sum_{j} X_2^{(j)}$	$\sum_{j} X_{2}^{(j)}$	$\sum_{j} X_2^{(j)}$
$X_3^{(0)}$	$X_3^{(1)}$	$X_3^2$	$Y^{(3)}$	$\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{0}$
´-			All-re	eauce			

- A ring with N nodes partitions data to N messages
- Collective Communication Flow:



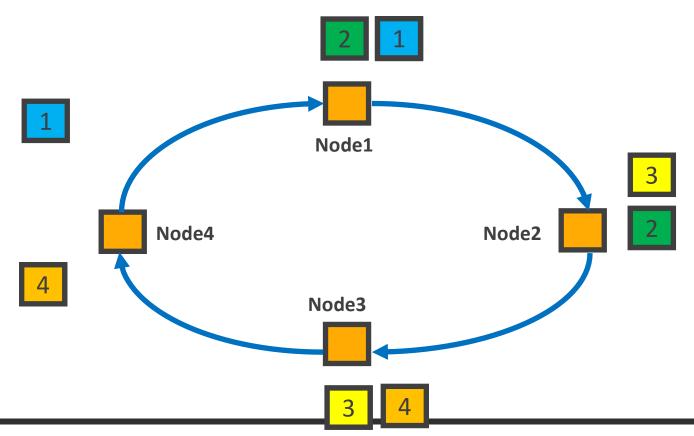
Node	Node	Node	Node	Node	Node	Node	Node
_0_	1	2	3	0_	1	2	3
$X_0^{(0)} \\ X_1^{(0)} \\ X_2^{(0)}$	$X_0^{(1)} \ X_1^{(1)} \ X^{(1)}$	$X_0^{(2)} \ X_1^{(2)} \ Y^{(2)}$	×1 0	$\sum_{j} X_0^{(j)}$	$\sum_j X_1^{(j)}$	$\sum_j X_2^{(j)}$	
$V^{(0)}$	$\mathbf{Y}^{(1)}$	$Y^{(2)}$	$X^{(3)}$ Re	duce			$\sum_{i} X_3^{(i)}$
Node	A3  Node	Node	Node	atter	ı Node,	l Nodo	l <sup>—</sup> ⁄ Node
0	1	2	3	0	1	2	Noue 3
			3	X0	<i>X</i> 0	X0	X0
X0	<i>X</i> 1		_	- X1	$X_1$	X1	X1
	$\Lambda$ I	<i>X</i> 2		X2	X2	X2	X2
		112	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3
			All-ga	ther	I	I	
Node.	Node.	Node	Node	Node	.Node	. Node	e Node
0	1	2	3	0	1	2	3
$ \begin{array}{c} X_0^{(0)} \\ X_1^{(0)} \\ X_2^{(0)} \\ Y_2^{(0)} \end{array} $	$X_0^{(1)} \ X_1^{(1)} \ X_2^{(1)} \ X_2^{(1)}$	$X_0^{(2)} \\ X_1^{(2)} \\ X_2^{(2)} \\ Y_2^{(2)}$	$ \begin{array}{c} X_0^{(3)} \\ X_1^{(3)} \\ X_2^{(3)} \\ Y_2^{(3)} \end{array} $	$\sum_{j} X_{0}^{(j)}$ $\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{2}^{(j)}$	$\sum_{j} X_{0}^{(j)} \sum_{j} X_{1}^{(j)} \sum_{j} X_{1}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)}$	$\sum_{j} X_{0}^{(j)}$ $\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{2}^{(j)}$	$\sum_{j} X_{0}^{(j)} \sum_{j} X_{1}^{(j)} \sum_{j} X_{1}^{(j)} \sum_{j} X_{2}^{(j)}$
A 3 1			All-re	duce			

- A ring with N nodes partitions data to N messages
- Collective Communication Flow:



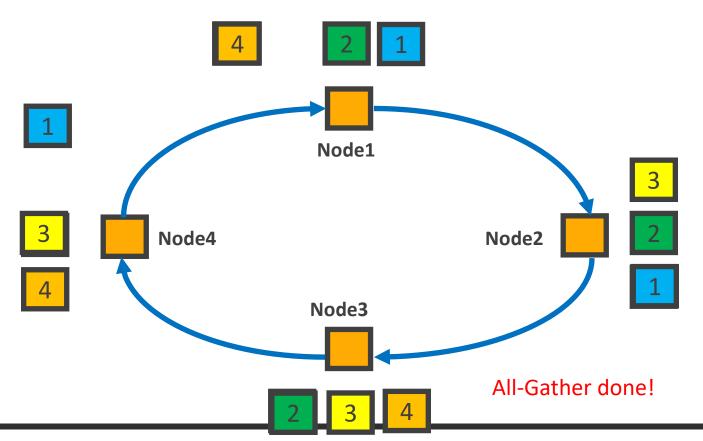
Node	Node	Node	Node	Node	Node	Node	Node
_0_	1	2	3	_0	1	2	3
$X_0^{(0)} X_1^{(0)}$		$X_0^{(2)} \\ Y^{(2)}$	U	$\sum_{j} X_0^{(j)}$	$\sum X^{(j)}$		
$X_1^{(0)}$	$X_1^{(1)}$ $X_2^{(1)}$	$X_1^{(2)}$ $X_2^{(2)}$	$X_{2}^{(3)}$			$\sum_j X_2^{(j)}$	\(\nabla_{\nu()}\)
$X_{3}^{(0)}$	$X_{3}^{(1)}$	$ X_3^{(2)} $	$X_3^{(3)}$ Re	duce atter			$\sum_{j} A_{3}^{c}$
Node	Node	Node	Node	Node,	Node,	Node,	Node
_0	1	2	3	_0_	1	2	3_
X0				<i>X</i> 0	X0	X0	<i>X</i> 0
	<i>X</i> 1		→	- X1	<i>X</i> 1	<i>X</i> 1	X1
		<i>X</i> 2		<i>X</i> 2	<i>X</i> 2	<i>X</i> 2	X2
			<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	X3
			All-ga	ther	ı	ı	
Node	Node	Node	Node	Node	Node	Node	Node
_0_	_1_	2	3_	0_	1	2	3
$X_0^{(0)}$	$X_0^{(1)}$ $Y^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$	$\sum_{j} X_0^{(j)}$	$\sum_{j} X_0^{(j)}$	$\sum_{j} X_0^{(j)}$ $\sum_{\mathbf{V}^{(j)}}$	$\sum_{j} X_0^{(j)}$
$X_1^{(0)} \ X_2^{(0)}$	$X_{1}^{(1)} X_{2}^{(1)}$	$X_1^{(2)}$ $X_2^{(2)}$	$X_1^{(3)} \rightarrow X_2^{(3)}$	$\sum_{j} X_{1}^{(j)}$	$\sum_{j} X_1^{(j)}$	$\sum_{j} X_{1}^{(j)}$	$\sum_{j} X_1^{(j)}$
$X_3^{(0)}$	$X_3^{(1)}$	$X_3^{(2)}$	$X_{3}^{(3)}$ All-re	$\sum_{j} X_3^{(j)}$ educe	$\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{(j)}$

- A ring with N nodes partitions data to N messages
- Collective Communication Flow:



Node	Node	Node	Node	Node	Node	Node
1_	2	3_	_0	1_	2	3
$X_1^{(1)}$	$X_0^{(2)} \ X_1^{(2)}$	$ A_1 $	$\sum_{j} X_0^{(j)}$	$\sum_{j} X_{1}^{(j)}$	\(\nabla_{\psi}(t)\)	
$X_{2}^{(1)} X_{3}^{(1)}$	$X_2^{(2)} \ X_3^{(2)}$	$X_{2}^{(3)}$ R			$\sum_{j} X_{2}^{j}$	$\sum_{j} X_3^{(j)}$
Node	Node	Node	Node	Node,	Node,	Node
1_	2	3_	_ 0	1	2	3_
	]		X0	X0	X0	X0
<i>X</i> 1		-	► X1	<i>X</i> 1	<i>X</i> 1	X1
	<i>X</i> 2		<i>X</i> 2	<i>X</i> 2	<i>X</i> 2	X2
		<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	X3
ı	I	l All-g	ather	ı	ı	
Node	Node	Node	Node	Node	Node	Node
_1_	2	3	0	1	2	3
$X_0^{(1)} \ X_1^{(1)} \ X_2^{(1)} \ X_3^{(1)}$	$X_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)}$	$X_0^{(3)}$ $X_1^{(3)}$ $X_2^{(3)}$ $X_2^{(3)}$	$\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$	$\sum_{j} X_{0}^{(j)} \sum_{j} X_{1}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)}$	$\sum_{j}^{-} X_{0}^{(j)}$ $\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$	$\sum_{j} X_0^{(j)}$ $\sum_{j} X_1^{(j)}$ $\sum_{j} X_2^{(j)}$ $\sum_{j} X_3^{(j)}$
	$\begin{array}{c} \mathbf{I} \\ X_0^{(1)} \\ X_1^{(1)} \\ X_2^{(1)} \\ X_3^{(1)} \\ \textbf{Node} \\ \mathbf{I} \\ X_1 \\ \end{array}$	$\begin{array}{c cccc} 1 & 2 \\ X_0^{(1)} & X_0^{(2)} \\ X_1^{(1)} & X_1^{(2)} \\ X_2^{(1)} & X_2^{(2)} \\ X_3^{(1)} & X_3^{(2)} \\ & & \mathbf{Node} \\ 1 & 2 \\ & & & & \\ \mathbf{Node} & \mathbf{Node} \\ 1 & & & & \\ 1 & & & & \\ & & & & \\ \mathbf{Node} & & & & \\ & & & & \\ & & & & & \\ & & & \\ & & $	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{ c c c c c }\hline 1 & 2 & 3 & 0 & 1\\ \hline X_0^{(1)} & X_0^{(2)} & X_0^{(3)} & \sum_j X_0^{(j)} \\ X_1^{(1)} & X_1^{(2)} & X_1^{(3)} & & \sum_j X_0^{(j)} \\ X_2^{(1)} & X_2^{(2)} & X_2^{(3)} & & & \sum_j X_1^{(j)} \\ X_2^{(1)} & X_3^{(2)} & X_3^{(3)} & \text{Reduce} \\ \hline X_3^{(3)} & \text{-scatter} \\ \hline Node Node Node Node Node Node Node Node $	$\begin{array}{ c c c c c } \hline 1 & 2 & 3 & 0 & 1 & 2 \\ X_{0}^{(1)} & X_{0}^{(2)} & X_{0}^{(3)} & \sum_{j} X_{0}^{(j)} & \\ X_{1}^{(1)} & X_{1}^{(2)} & X_{1}^{(3)} & & \sum_{j} X_{1}^{(j)} & \\ X_{2}^{(1)} & X_{2}^{(2)} & X_{3}^{(3)} & \text{Reduce} \\ X_{3}^{(1)} & X_{3}^{(2)} & X_{3}^{(3)} & \text{Reduce} \\ X_{3}^{(1)} & X_{3}^{(2)} & X_{3}^{(3)} & \text{Reduce} \\ \mathbf{Node} & \mathbf{Node} & \mathbf{Node} & \mathbf{Node} & \mathbf{Node} & \mathbf{Node} \\ 1 & 2 & 3 & 0 & 1 & 2 \\ & & & & & & & & & & & & & & & & \\ X_{1} & & & & & & & & & & & & & & & \\ X_{2} & & & & & & & & & & & & & & & & \\ X_{3} & & & & & & & & & & & & & & & & & & &$

- A ring with N nodes partitions data to N messages
- Collective Communication Flow:



Node	Node	Node	Node	Node	Node	Node	Node
_0_	1	2	3_	0	1	2	3
$X_0^{(0)} X_1^{(0)}$	$X_0^{(1)} \\ X^{(1)}$	$X_0^{(2)} \\ Y^{(2)}$	$X_0^{(3)}$ $X_0^{(3)}$	$\sum_{j} X_0^{(j)}$	$\sum_i X_1^{(j)}$		
$X_1 \ X_2^{(0)} \ \mathbf{v}^{(0)}$	$X_{2}^{(1)}$	$X_{2}^{(2)}$	$X_2^{(3)}$	educe		$\sum_{j} X_2^{(j)}$	$\sum_{i} X_3^{(i)}$
$A_3$	$A_3$	$A_3$		atter		 	<i>^</i>
Node	Node	Node			Node	- 1	
_0	$\stackrel{1}{\vdash}$	1 2	3_	_0_	$\frac{1}{4}$	2	3
X0				X0	X0	<i>X</i> 0	X0
	<i>X</i> 1		→	- X1	<i>X</i> 1	<i>X</i> 1	X1
		<i>X</i> 2		<i>X</i> 2	<i>X</i> 2	<i>X</i> 2	X2
			<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3
	'	•	All-ga	ther I	'	'	
Node	Node	Node	Node	Node	Node	Node	Node
0_	1	2	3	0	1	2	3
$X_0^{(0)} X_1^{(0)}$	$X_0^{(1)} X_1^{(1)}$	$X_0^{(2)} \ X_1^{(2)}$	$X_0^{(3)} \\ X_1^{(3)} \rightarrow$	$\sum_{j} X_0^{(j)}$ $\sum_{j} X_1^{(j)}$	$\sum_{j} X_0^{(j)}$ $\sum_{j} X_1^{(j)}$	$\sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} $	$\sum_{j} X_0^{(j)}$ $\sum_{j} X_1^{(j)}$
$X_{2}^{(0)} X_{3}^{(0)}$	$X_{3}^{(1)}$ $X_{3}^{(1)}$	$X_2^{(2)} X_3^{(2)}$	$X_2^{(3)} = X_2^{(3)}$	$\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ educe	$\sum_{j} X_2^{(j)}$ $\sum_{j} X_3^{(j)}$	$\sum_{j} X_2^{(j)}$ $\sum_{j} X_3^{(j)}$	$\sum_{j} X_2^{(j)}$ $\sum_{j} X_3^{(j)}$
<b>-</b>	<b>-</b>			-			

Node Node Node Node

X0

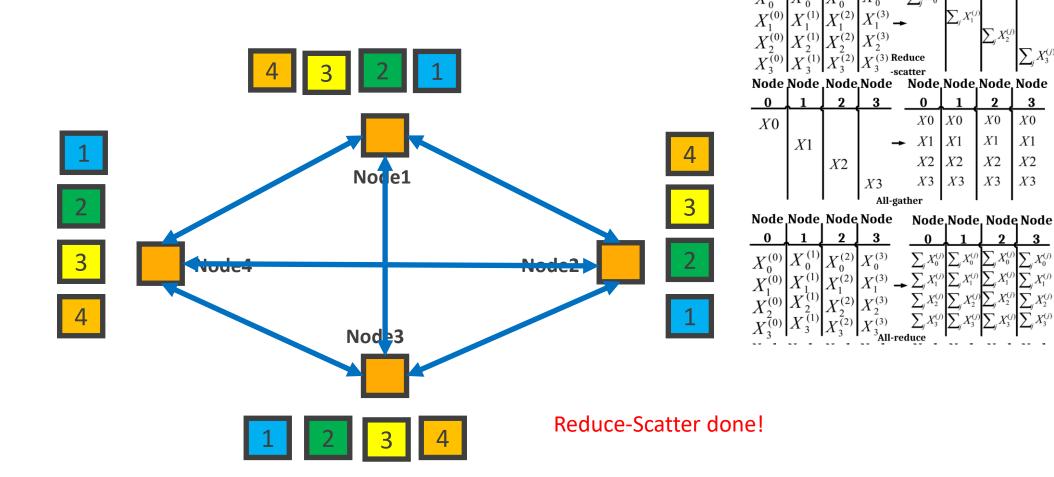
X1

X2

X3

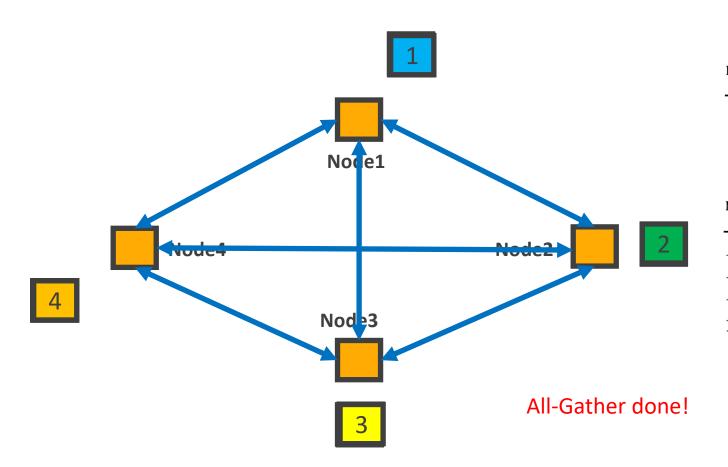
*X*3

# Example: Direct All-Reduce



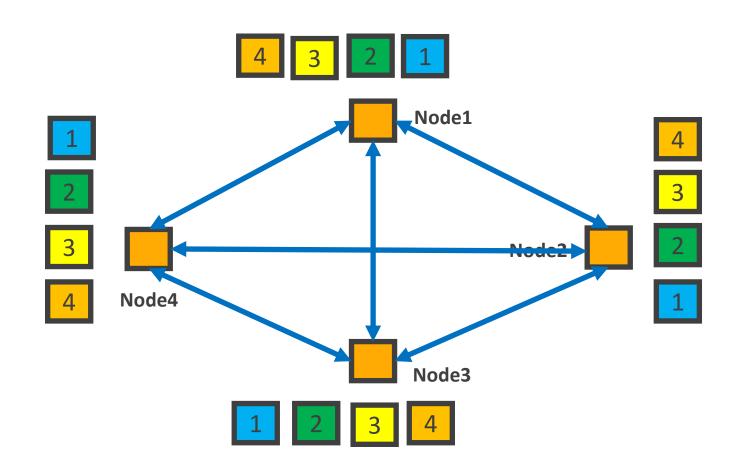
Node Node Node Node

# Example: Direct All-Reduce



Node	Node	Node	Node	Node	Node	Node	Node
0_	1	2	3	0	1	2	3
$X_0^{(0)}$	0	$A_0$	0	$\sum_{j} X_0^{(j)}$			
$X_1^{(0)}$	<b></b> 1	$\alpha_1$	<b>-</b> 1	<b>→</b>	$\sum_{j} X_{1}^{(j)}$	) 	
$X_{2}^{(0)}$	$X_{2}^{(1)}$	$X_{2}^{(2)}$	$X_{2}^{(3)}$			$\sum_{j} X_{2}^{(j)}$	<u></u>
$X_3^{(0)}$	$X_{3}^{(1)}$	$X_3^{(2)}$	$X_3^{(3)}$ R	teduce scatter			$\sum_{j} X_3^{(j)}$
Node	Node	Node			Node,	Node,	Node
_0	1_	2	3	_ 0	1	2	3
X0				<i>X</i> 0	X0	X0	<i>X</i> 0
	<i>X</i> 1		-	<b>→</b> X1	<i>X</i> 1	<i>X</i> 1	<i>X</i> 1
		<i>X</i> 2		<i>X</i> 2	<i>X</i> 2	<i>X</i> 2	X2
			<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	X3
	ı	ı	All-g	gather	I	ı	
Node	Node	Node	Node	Node	Node	Node	e Node
0	_1_	2	3	0	1	2	3
$X_0^{(0)}$	$X_0^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$	$\sum_{j} X_0^{(j)}$	$\sum_{j} X_0^{(j)}$	$\sum_{j} X_0^{(j)}$	$\sum_{j} X_0^{(j)}$
$X_{1}^{(0)}$	$X_{1}^{(1)}$	$X_1^{(2)}$	$X_1^{(3)}$ .	$\sum_{j} X_{1}^{(j)}$	$\sum_{j} X_1^{(j)}$	$\sum_{j} X_{1}^{(j)}$	$\sum\nolimits_{j} X_{1}^{(j)}$
$X_{2}^{(0)}$	$X_{2}^{(1)}$	$X_{2}^{(2)}$	$X_{2}^{(3)}$	$\sum_{j} X_2^{(j)}$	$\sum_{j} X_2^{(j)}$	$\sum_{j} X_{2}^{(j)}$	$\sum_{j} X_2^{(j)}$
$X_3^{(0)}$	$X_3^{(1)}$	$X_3^{(2)}$	$X_{3}^{(3)}$	$\sum_{j} X_3^{(j)}$ reduce	$\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{(j)}$
´-	<b>-</b>		AII-I	reduce			<b>-</b>

# Example: All-to-All



Node	Node	Node	Node	Node Node Node Node				
_0	_1	2	3_	_0_	_1_	2	3	
$X_0^{(0)}$	$X_0^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$			$X_2^{(0)}$		
		$X_1^{(2)}$		$X_0^{(1)}$	$X_1^{(1)}$	$X_{2}^{(1)}$	$X_3^{(1)}$	
$X_2^{(0)}$	$X_2^{(1)}$	$X_2^{(2)}$	$X_2^{(3)}$	$X_0^{(2)}$	$X_1^{(2)}$	$X_2^{(2)}$	$X_3^{(2)}$	
$X_3^{(0)}$	$X_3^{(1)}$	$X_3^{(2)}$	$X_3^{(3)}$ All-	$X_0^{(3)}$	$X_1^{(3)}$	$X_2^{(3)}$	$X_3^{(3)}$	

## Collectives on Sophisticated Training Platforms

Torus 3D
Similar to Google TPU

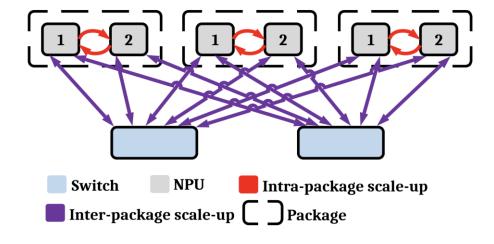
**NPU** Intra-package scale-up Inter-package scale-up

#### Hierarchical all-reduce:

- Reduce-scatter within package
- All-reduce across rows
- All-reduce across columns
- All-gather within package

All-To-All

Similar to NVIDIA DGX2



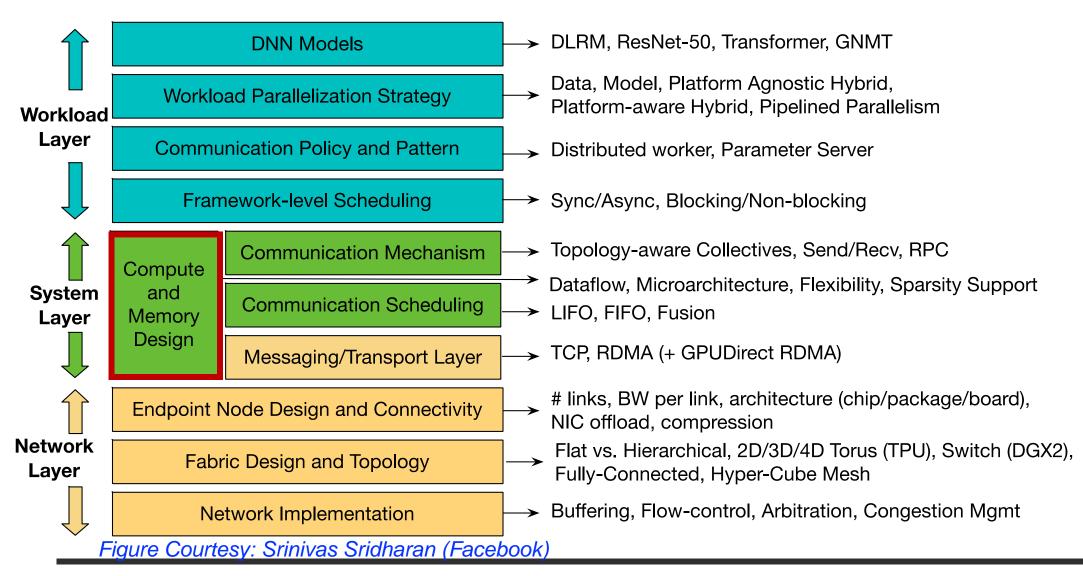
### Hierarchical all-reduce:

- Reduce-scatter within package
- All-reduce across switch
- All-gather within package

Heterogeneous Bandwidth

Multi-phase Collectives

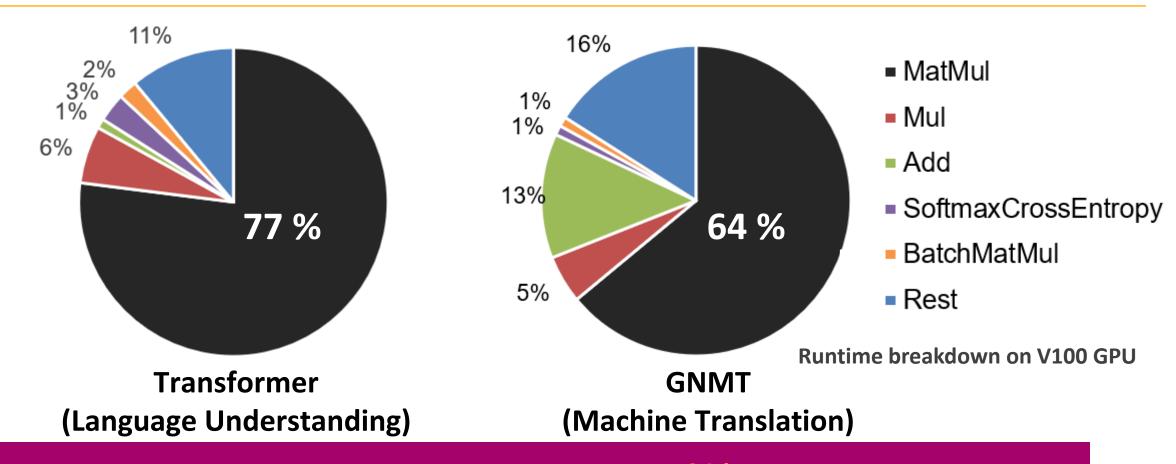
### Distributed Training Stack



## DL Training: The Compute

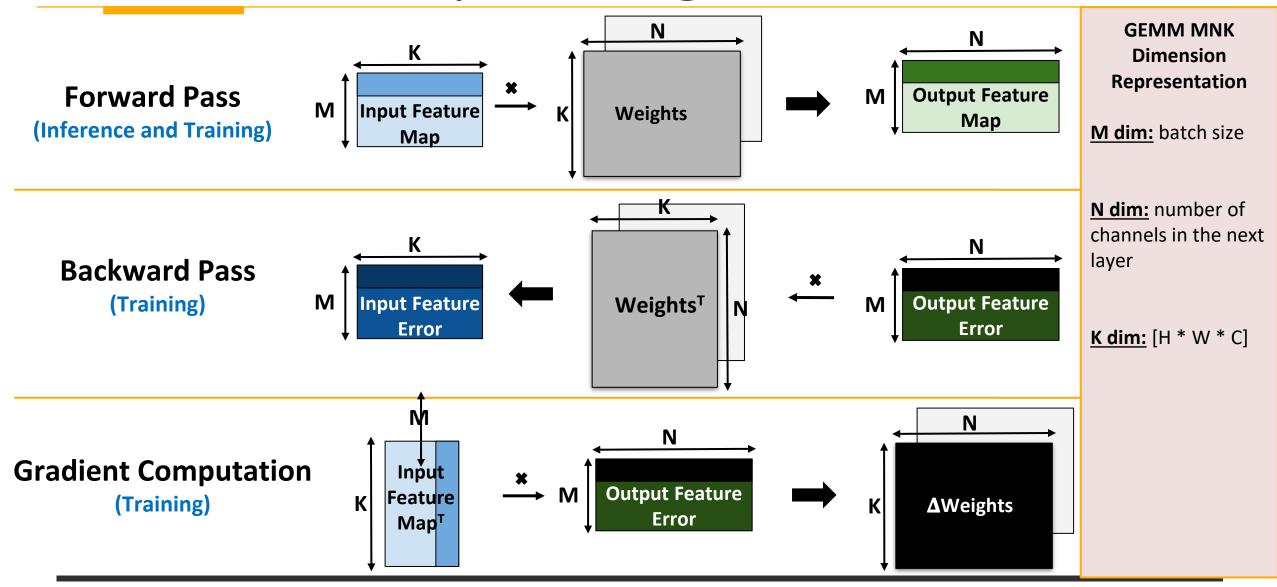
**GEMM MNK** N N **Dimension** Representation **Forward Pass Output Feature** M **Input Feature** Weights Map (Inference and Training) M dim: batch size Map N dim: number of channels in the next layer **Backward Pass Input Feature Output Feature** M (Training) **Weights**<sup>T</sup> **Error Error K dim:** [H \* W \* C] **Gradient Computation** Input **Output Feature Feature** (Training) **ΔWeights Error** Map<sup>T</sup>

# Key Compute Kernel during DL Training



Matrix multiplications (GEMMs) consume around **70%** of the total runtime when training modern deep learning workloads.

# **GEMMs** in Deep Learning

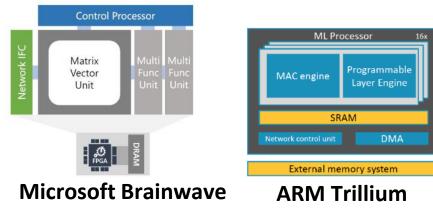


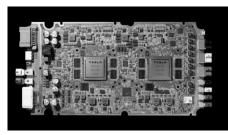
# Hardware for Accelerating GEMMs

#### **SIMT Architectures**



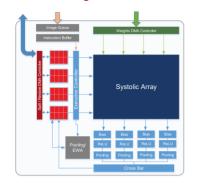
#### **SIMD Architectures**

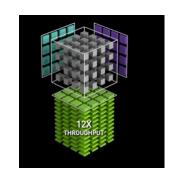




**Tesla FSDC** 

#### **Systolic Architectures**





Xilinx xDNN

**Nvidia Tensor Cores** 



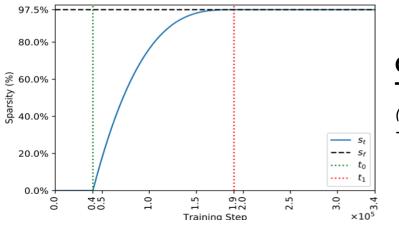
**Google TPU** 

Key Feature: Specialized support for GEMMs

### GEMMs in Modern DL

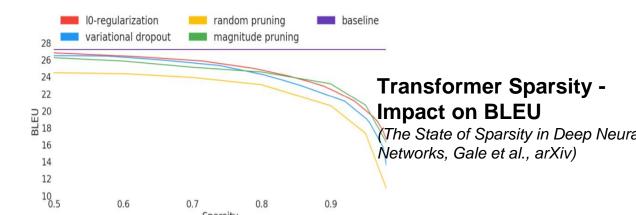
Workload	Application	<b>Example Dimensions</b>		
		M	N	K
GNMT		128	2048	4096
	Machine	320	3072	4096
	Translation	1632	36548	1024
		2048	4096	32
DeepBench	General	1024	16	500000
	Workload	35	8457	2560
Transformer	Language	31999	1024	84
	Understanding	84	1024	4096
NCF	Collaborative	2048	1	128
	Filtering	256	256	2048

GEMMs are irregular (non-square)!



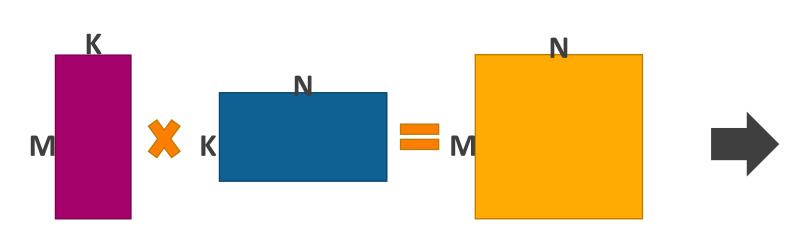
## **GNMT Pruning - Temporal Sparsity**

(https://www.intel.ai/compressing-gnmt-models)



GEMMs are Sparse! Weight sparsity ranges from **40%** to **90%**. Activation sparsity is approximately **30%** to **70%** from ReLU, dropout, etc.

## Mapping GEMMs on Accelerators



What determines utilization?

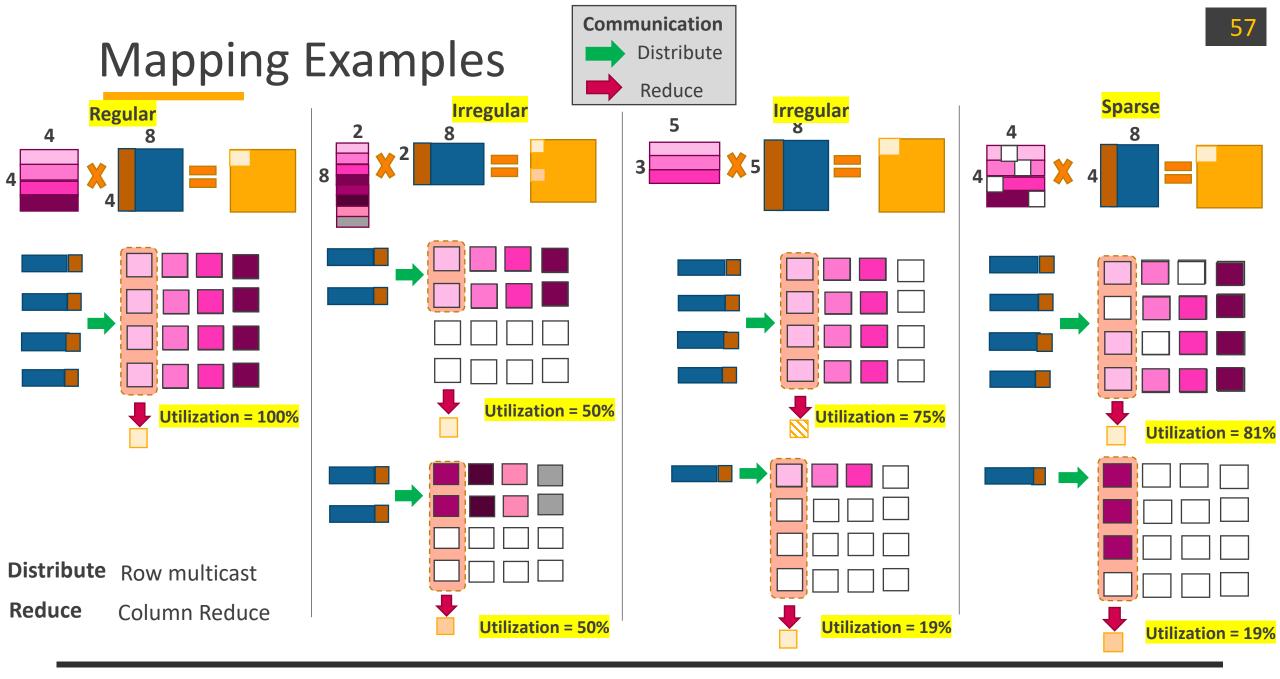
**Mapping Efficiency** 

What determines stalls?

Memory/Interconnect Bandwidth

**TPU (Systolic Array)**  $15 \cdot \cdot \cdot 31 \cdot \cdot \cdot 47 \cdot \cdot \cdot 63 \cdot \cdot \cdot 79 \cdot \cdot \cdot 95 \cdot \cdot 111 \cdot \cdot 127$ 

\*\* Assuming MK matrix is streaming and KN matrix is stationary. (aka weight stationary)



### Mechanisms to increase utilization

### Handling Irregular GEMMs

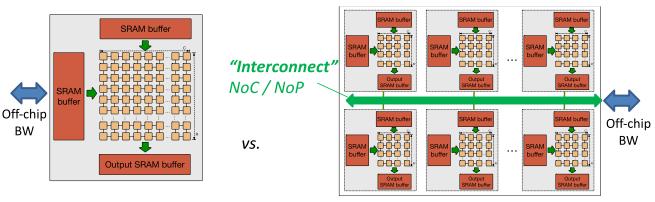
• One large array (e.g., Google TPU) versus several smaller arrays (e.g., NVIDIA

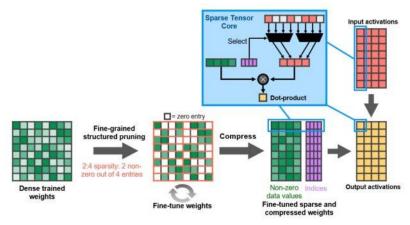
Tensor cores)

• Trade-off: reuse vs utilization

### Handling Sparse GEMMs

- Structured Sparsity Support
  - E.g., NVIDIA A100
- Unstructured Sparsity Support
  - Active research going on

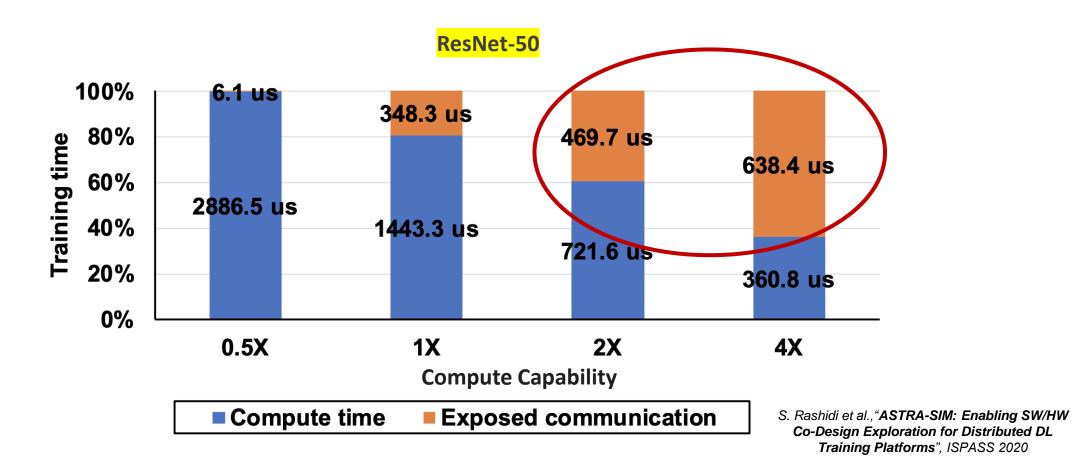




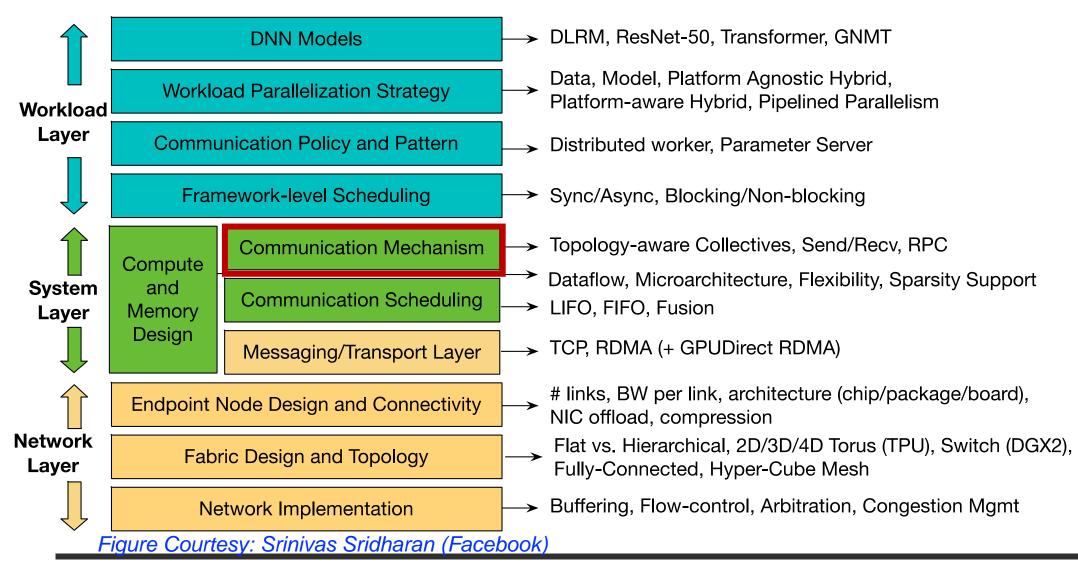
**NVIDIA A100 supports 4:2 structured sparsity** 

### Effect of Enhanced Compute Efficiency on Training

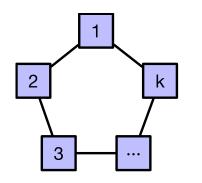
• A Torus 3D with total of 32 nodes (2X4X4) is used.

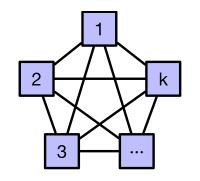


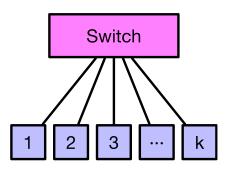
### Distributed Training Stack



## Topology-aware Collectives





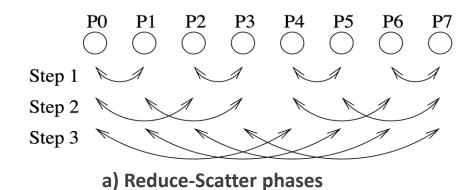


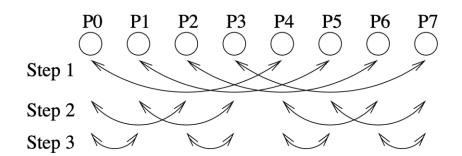
(a) Ring(k)

(b) FullyConnected(k)

(b) Switch(k)

Topology Building Block	Topology-aware Collective Algorithm		
Ring	Ring		
FullyConnected	Direct		
Switch	HalvingDoubling		

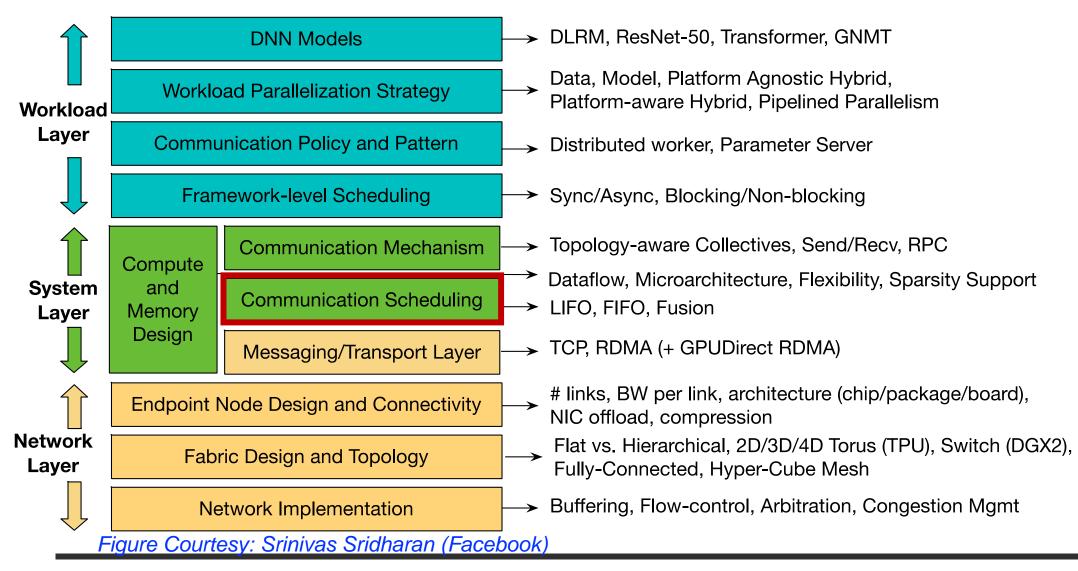




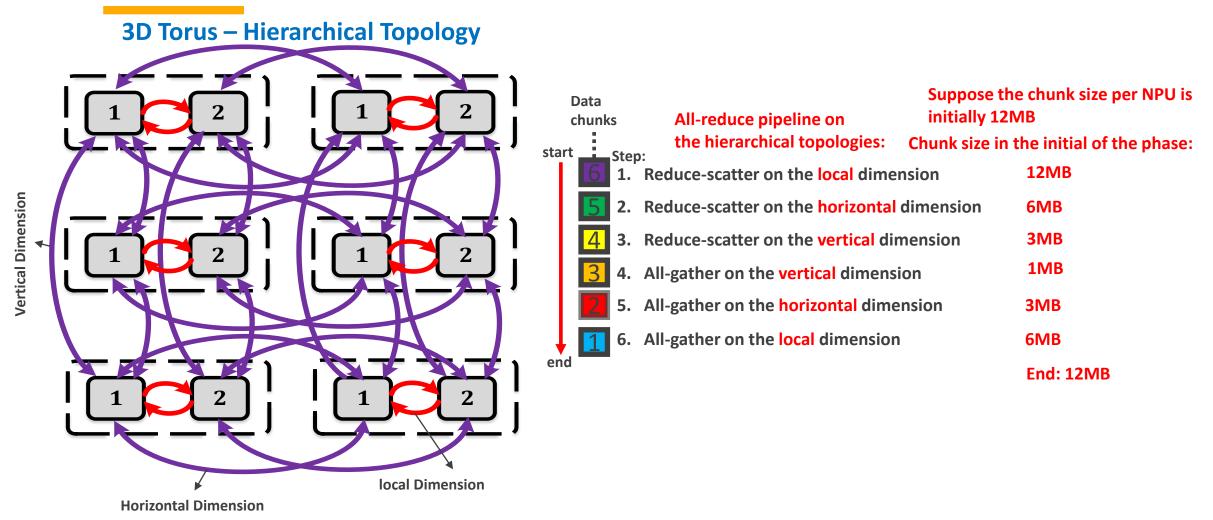
b) All-gather phases

#### HalvingDoubling All-Reduce

### Distributed Training Stack

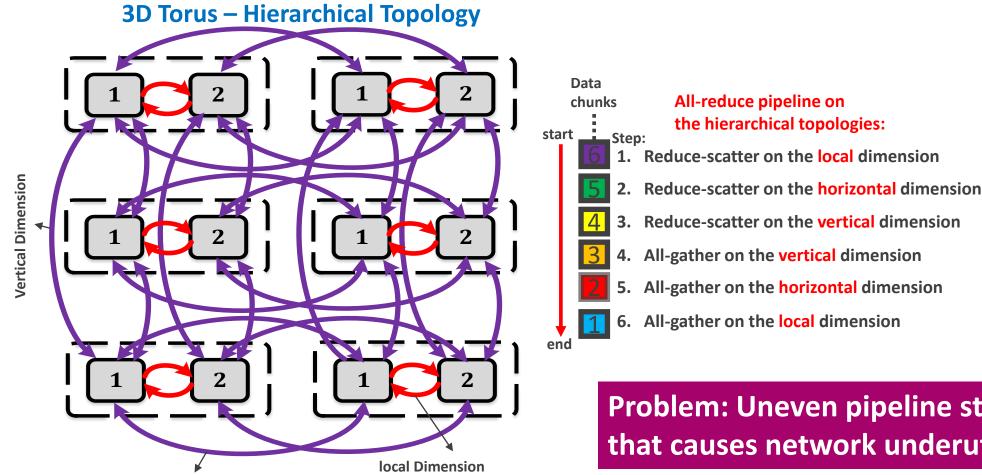


# Baseline All-Reduce on the Hierarchical Topologies



S. Rashidi et al., "Themis: A Network Bandwidth-Aware Collective Scheduling Policy for Distributed Training of DL Models". ISCA 2022.

# Baseline All-Reduce on the Hierarchical Topologies



**Pipeline Stage latency:** Idle time Idle time Idle time Idle time

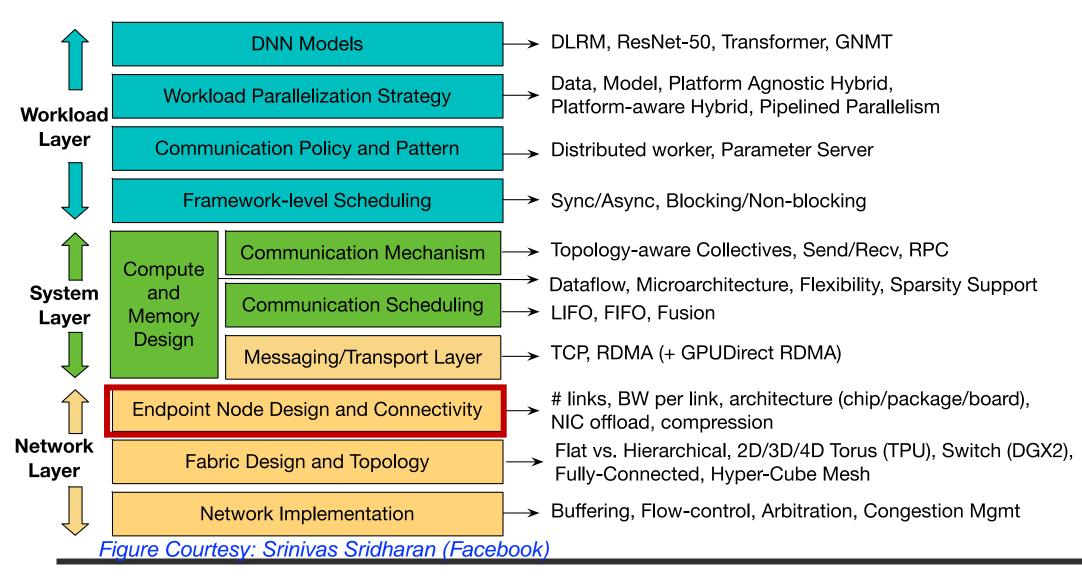
**Problem: Uneven pipeline stage latencies** that causes network underutilization

For solution to this problem, check out our talk on Themis on Tuesday, June 21, Session 7B, 2:30 – 2:50 PM

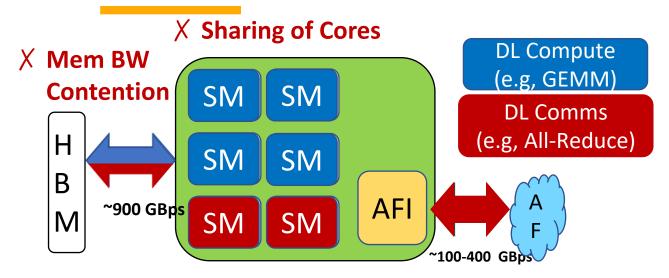
S. Rashidi et al.. " Themis: A Network Bandwidth-Aware Collective Scheduling Policy for Distributed Training of DL Models". ISCA 2022.

**Horizontal Dimension** 

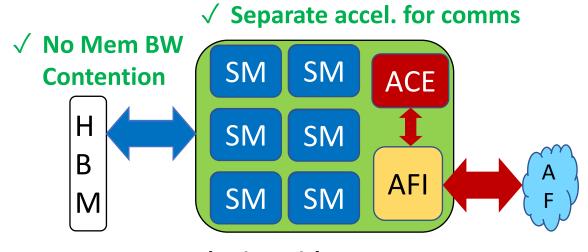
### Distributed Training Stack



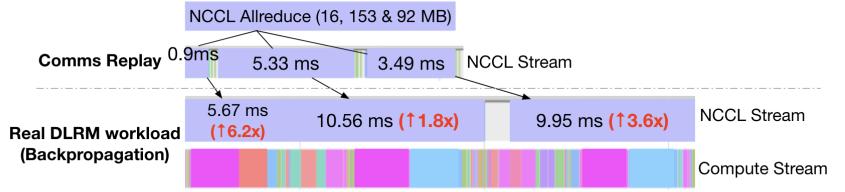
### Resource Contention at End-point



**Endpoint in Baseline** 



**Endpoint with Accelerator Collectives Engine (ACE)** 

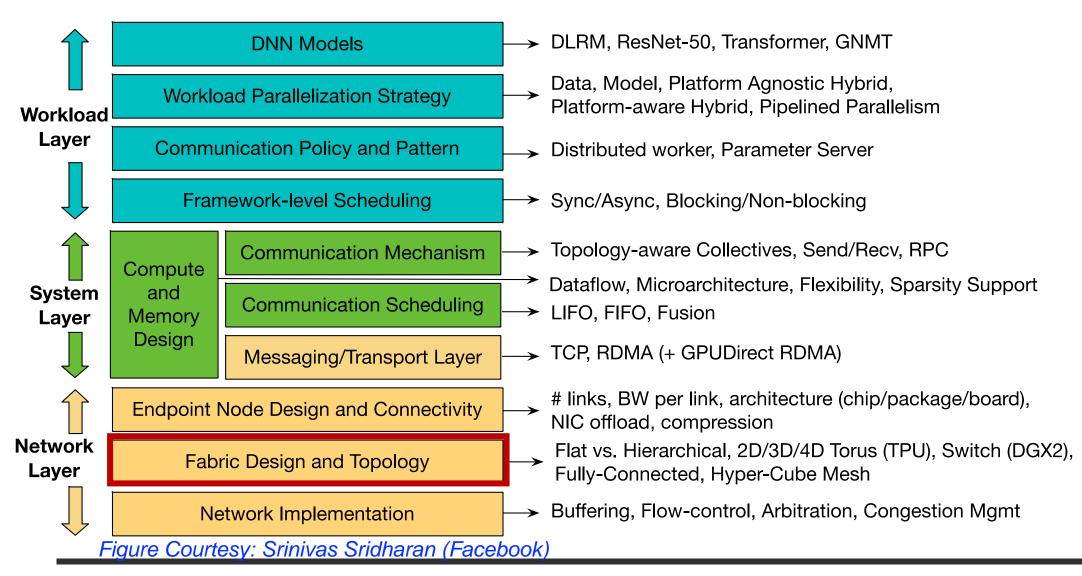


Alternate approach: offload to switch (e.g., ISCA 2020)

(b) Impact of compute-comms overlap on a real-world production-class DLRM workload

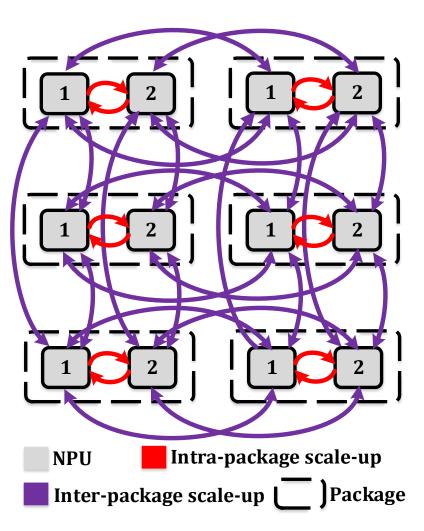
S. Rashidi et al., "Enabling Compute-Communication Overlap in Distributed Deep Learning Training Platforms". ISCA 2021

### Distributed Training Stack



# Target System

#### Torus 3D



X \* Y\* Z dimension

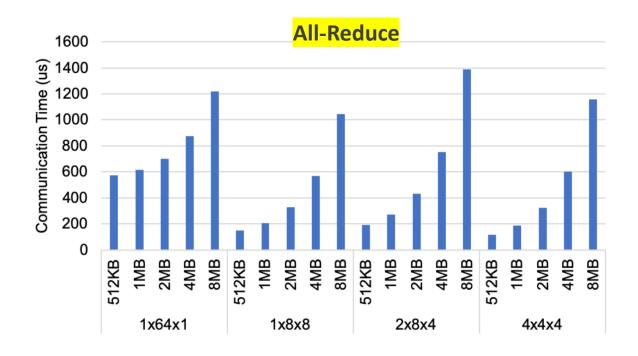
X= cores within a package

Y= packages in horizonal dimension

Z= packages in vertical dimension

### Impact of 1D/2D/3D Torus

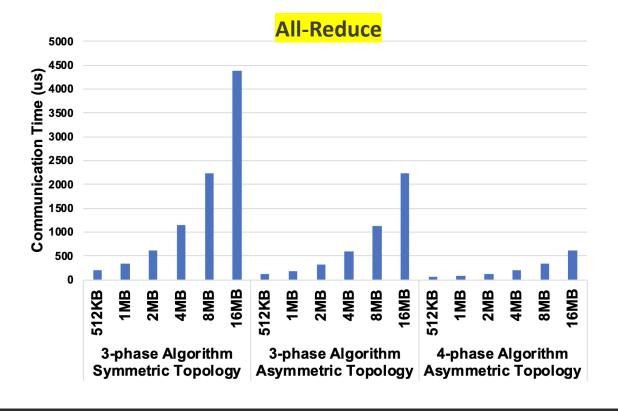
- Adding a dimension decreases the number of steps per collective.
  - For example, going from 1X64X1 to 1X8X8.
- Adding a dimension might increase amount of data each node sends out (depends on the algorithm).
  - For example, going from 1X8X8 to 2X8X4.
- Hence, choosing a topology is a tradeoff between the above effects.



S. Rashidi et al., "ASTRA-SIM: Enabling SW/HW
Co-Design Exploration for Distributed DL
Training Platforms", ISPASS 2020

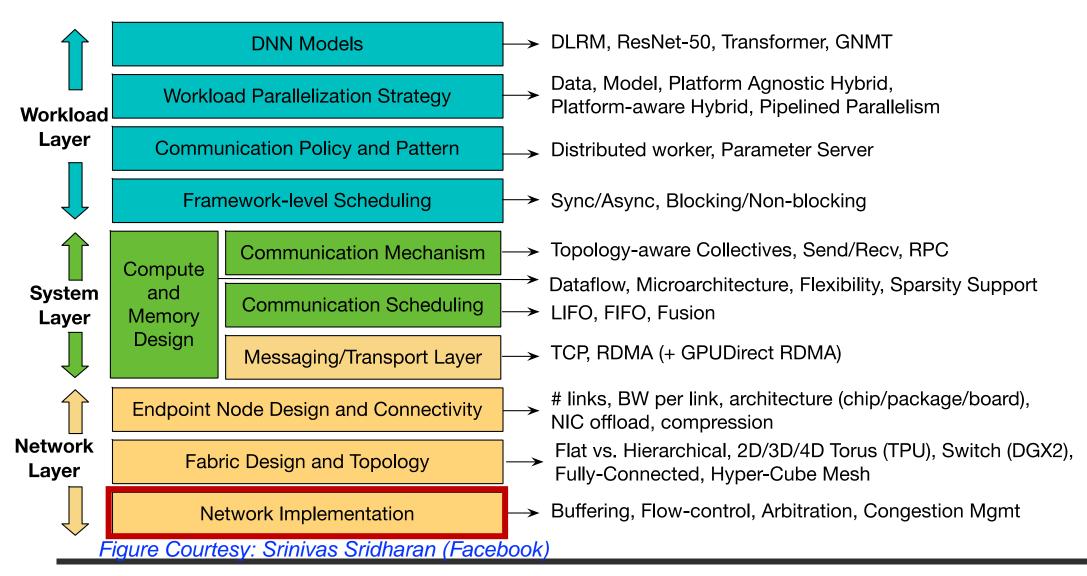
## Impact of Asymmetric Hierarchical Topology

- Having higher intra-package BW improves the performance.
- We can further improve performance by changing the algorithm to leverage this asymmetric BW.

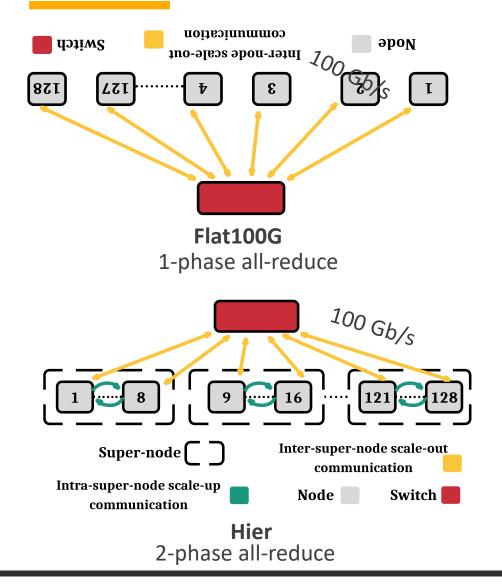


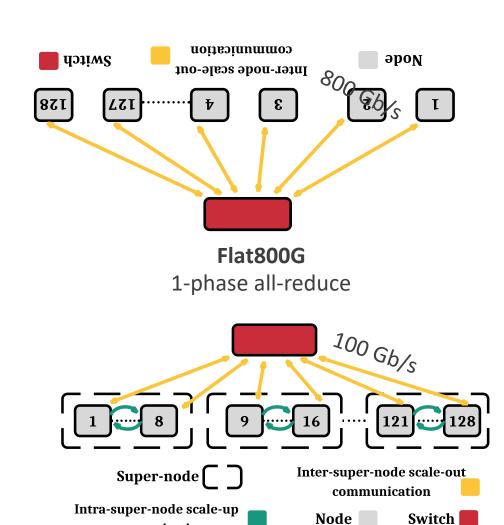
S. Rashidi et al., "ASTRA-SIM: Enabling SW/HW
Co-Design Exploration for Distributed DL
Training Platforms", ISPASS 2020

### Distributed Training Stack



## **Target Systems**





**HierOpt** 

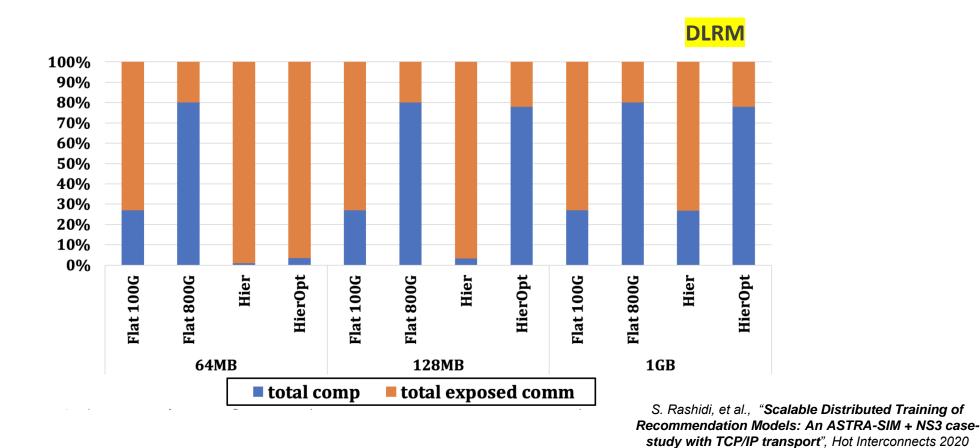
3-phase all-reduce

communication

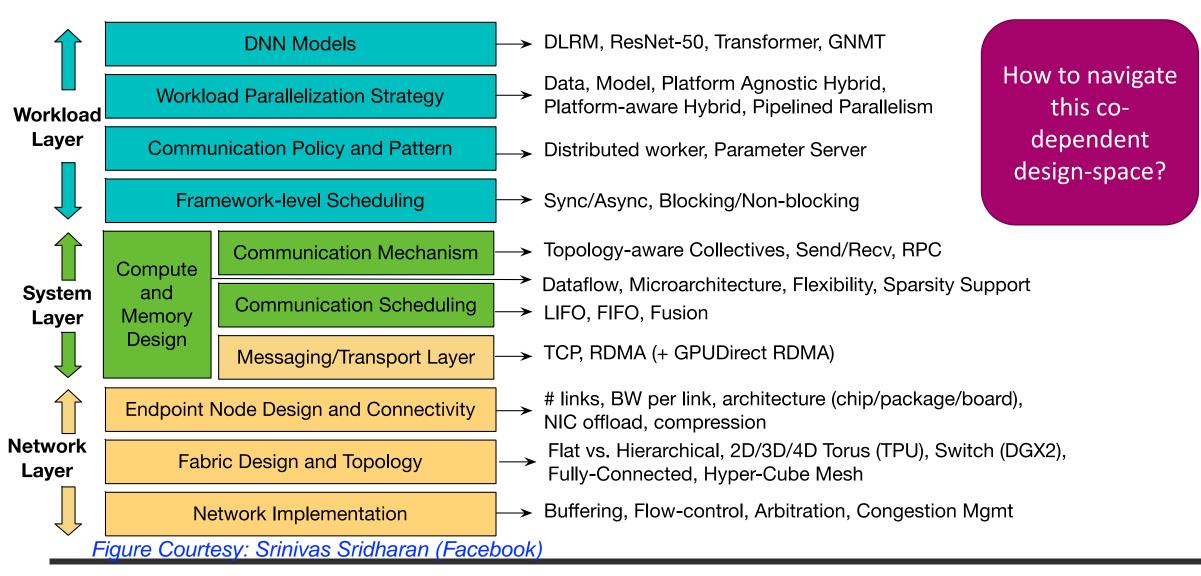
### Effect of Size of Switch Buffer

#### Observations:

• Flat vs. Hierarch different Sensitivity to global switch size



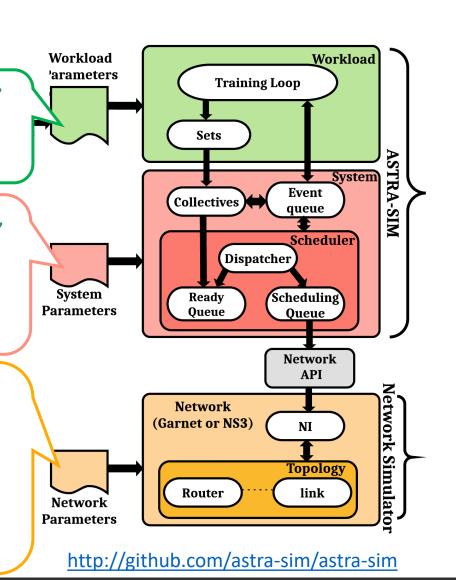
### **Distributed Training Stack**



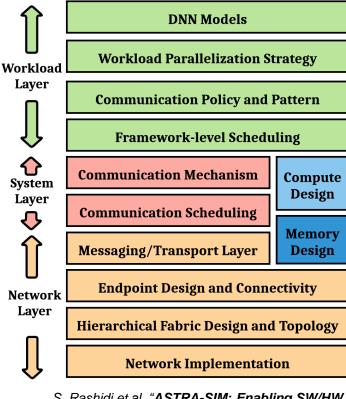
### Introducing ASTRA-sim

75 STRA\* SIM\*

- ✓ Released
- In progress
- ✓ Supports Data-Parallel, Model-Parallel, Hybrid-Parallel training loops
- ✓ Extensible to more training loops
  - Graph-based input from PyTorch
- ✓ Ring based, Tree-based, AlltoAll based, and multi-phase collectives
- √ Variety of scheduling policies
- ✓ Compute times fed via offline system measurements or compute simulator
- ✓ Various topologies, flow-control, link bandwidth, congestion control
- ✓ Plug-and-play options
  - ✓ Analytical (roofline)
  - Analytical with congestion
  - ✓ Garnet (credit-based)
  - ➤ NS3 (TCP, RDMA)



#### **DL Training Co-Design Stack**



S. Rashidi et al., "ASTRA-SIM: Enabling SW/HW
Co-Design Exploration for Distributed DL
Training Platforms", ISPASS 2020

S. Rashidi, et al., "Scalable Distributed Training of Recommendation Models: An ASTRA-SIM + NS3 casestudy with TCP/IP transport", Hot Interconnects 2020

### What Does ASTRA-sim Report?

### ASTRA-sim Reports:

- 1. End-to-end training time.
- 2. Total communication time for each communication operation.
- 3. The amount of **exposed communication** for each communication operation.
- 4. Total Exposed communication and total computation.
- More detailed stats such as average message latency per each hierarchical collective phase.

### Network Backend Specific Reports (Depends on the network backend type):

- 1. Network BW utilization
- Communication protocol stats, such as packet drops, # of retransmissions, etc.
- 3. Network switch buffer usage
- 4. ...

### Summary and Takeaways

- Large Model distributed training is an ongoing open-research area
- Many emerging supercomputing systems being designed specifically for this problem!
  - Cerebras CS2
  - Tesla Dojo
  - NVIDIA DGX + Mellanox SHARP switches
  - Intel Habana
  - IBM Blueconnect
  - •
- Co-design of algorithm and system offers high opportunities for speedup and efficiency

### Agenda

Time (EDT)	Topic	Presenter
8:30 – 9:30	Introduction to Distributed Deep Learning Training Platforms	Tushar Krishna
9:30 – 10:30	ASTRA-sim	Saeed Rashidi
10:30 - 11:00	Coffee Break	
11:00 – 11:50	Demo and Exercises	William Won and Taekyung Heo
11:50 – 12:00	Extensions and Future Development	Taekyung Heo

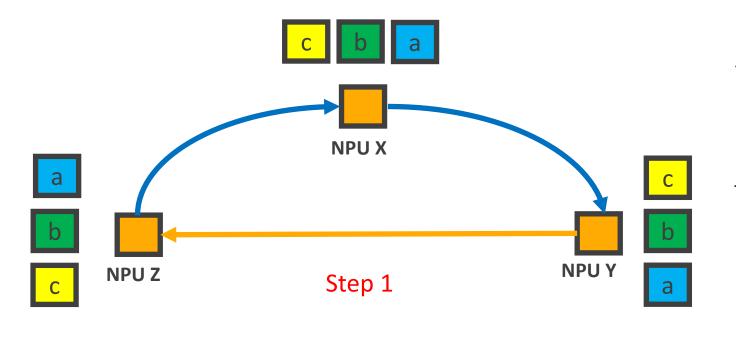
#### **Tutorial Website**

includes agenda, slides, ASTRA-sim installation instructions (via source + docker image) <a href="https://astra-sim.github.io/tutorials/isca-2022">https://astra-sim.github.io/tutorials/isca-2022</a>

**Attention:** Tutorial is being recorded

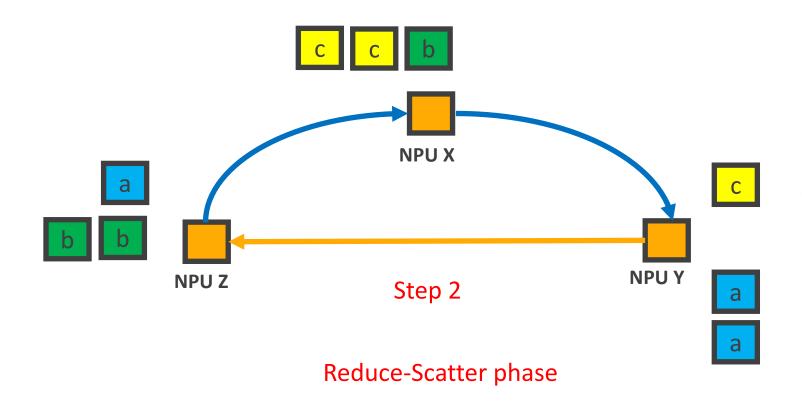
# Backup

- A ring with N nodes partitions data to N messages
- Collective Communication Flow:



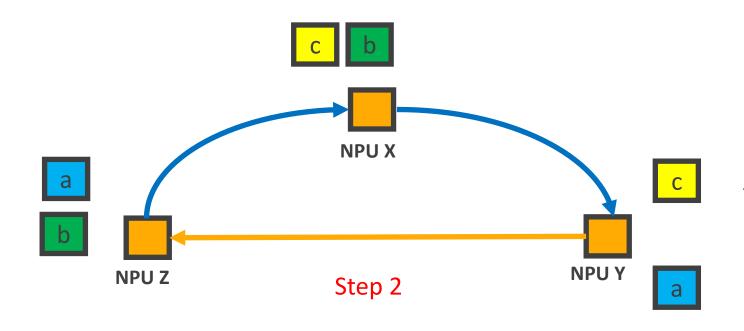
Node	Node	Node	Node	Node	Node	Node	Node
_0_	1_	2	3_	0	1	2	3
$X_0^{(0)}$	$X_0^{(1)}$	U		$\sum_{j} X_0^{(j)}$			
$X_1^{(0)}$	$X_1^{(1)}$	$X_1^{(2)}$	$X_1^{(3)} \rightarrow$		$\sum_{j} X_1^{(j)}$		
$X_{2}^{(0)}$	$X_{2}^{(1)}$	$X_{2}^{(2)}$				$\sum_{j} X_2^{(j)}$	<b>\</b>
$X_3^{(0)}$	$X_3^{(1)}$	$X_3^{(2)}$	$X_3^{(3)}$ Re	duce atter			$\sum_{j} X_3^{(j)}$
Node	Node	Node	Node	Node,	Node,	Node,	Node
_0_	1	2	3	_0	1	2	3_
X0		)		X0	X0	X0	X0
	<i>X</i> 1		-	- X1	<i>X</i> 1	<i>X</i> 1	X1
		<i>X</i> 2		<i>X</i> 2	<i>X</i> 2	<i>X</i> 2	X2
			<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3
	ı		All-ga	ther I	'	- 1	
Node	Node	Node	Node	Node	Node	Node	Node
_0_	_1_	2	3	0	1	2	3
$X_0^{(0)}$	$X_0^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$	$\sum_{j} X_0^{(j)}$	$\sum_{j} X_0^{(j)}$	$\sum_{j} X_0^{(j)}$	$\sum_{j} X_0^{(j)}$
$X_1^{(0)}$	$X_{1}^{(1)}$	$X_1^{(2)}$	$X_1^{(3)} \rightarrow$	$\sum_{j} X_1^{(j)}$	$\sum_{j} X_1^{(j)}$	$\sum_{j} X_1^{(j)}$	$\sum_{j} X_1^{(j)}$
$X^{(0)}$	$X_{2}^{(1)}$	$X_{2}^{(2)}$	$X_{2}^{(3)}$	$\sum_{j} X_2^{(j)}$	$\sum_{j} X_2^{(j)}$	$\sum_{j} X_2^{(j)}$	$\sum_{j} X_2^{(j)}$
$X_3^{(0)}$	$X_3^{(1)}$	$X_3^{(2)}$	$X_{2}^{(3)}$	$\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{(j)}$
	Node	Node	Ап-ге	<sup>duce</sup> Node	.Node	. Node	Node
•	_						3
	L 1 .	12.	1 3	0	ΙI	] 2	ו נ
$X_0^{(0)}$	$X_0^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$		$X_1^{(0)}$	$\leftarrow$	←
$X_0^{(0)}$ $X_1^{(0)}$			$X_0^{(3)}$ $X_1^{(3)}$	$X_0^{(0)}$ $X_0^{(1)}$		$X_{2}^{(0)}$	$X_3^{(0)}$
	$X_1^{(1)}$	$X_1^{(2)}$	$X_1^{(3)} - X_2^{(3)}$	$X_0^{(0)}$ $X_0^{(1)}$	$X_1^{(1)} X_1^{(2)}$	$X_2^{(0)} X_2^{(1)}$	$X_3^{(0)} X_3^{(1)}$

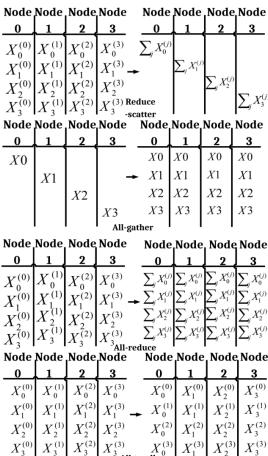
- A ring with N nodes partitions data to N messages
- Collective Communication Flow:



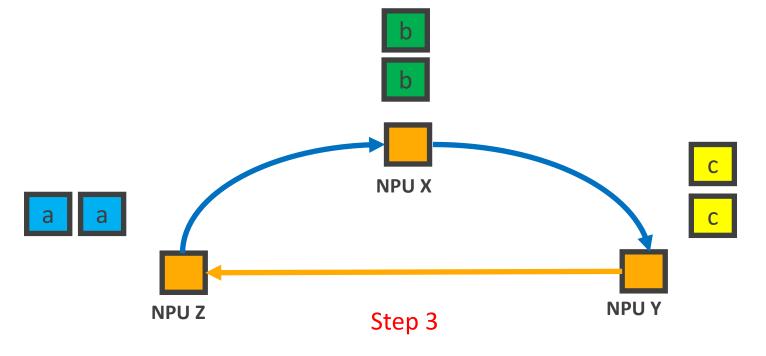
Node	Node	Node	Node	Node	Node	Node	Node
_0_	1	2	3	_0	1_	2	3_
$ \begin{array}{c} X_0^{(0)} \\ X_1^{(0)} \\ X_2^{(0)} \\ X_3^{(0)} \end{array} $	$X_1^{(1)} \\ X_2^{(1)}$	$X_1^{(2)} \\ X_2^{(2)}$	$X_1^{(3)} \rightarrow$	$\sum_{j} X_0^{(j)}$	$\sum_j X_1^{(j)}$	$\sum_j X_2^{(j)}$	$\sum_i X_3^{(j)}$
				utter	 N= 4 - 1	 N= 4 -	— <sup>,</sup>
	Noae	Node		Node	- 1	- 1	
_0_	1	2	3	0	1	2	3_
X0				X0	X0	<i>X</i> 0	X0
	<i>X</i> 1		-	· X1	<i>X</i> 1	<i>X</i> 1	X1
		<i>X</i> 2		<i>X</i> 2	<i>X</i> 2	<i>X</i> 2	X2
			<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	X3
	l		All-ga	<sub>ther</sub>	ı	ı	
Node	Node	Node	Node	Node	Node	Node	Node
_0_	_1_	2	3	0	1	2	3
	$X_{1}^{(1)}$	$X_0^{(2)} \ X_1^{(2)}$	$X_0^{(3)} \\ X_1^{(3)} \rightarrow$	$\sum_{j} X_0^{(j)}$ $\sum_{i} X_1^{(j)}$		$\sum_{j} X_0^{(j)}$ $\sum_{i} X_1^{(j)}$	$\sum_{j} X_0^{(j)}$ $\sum_{i} X_1^{(j)}$
$X_{2}^{(0)}$	$X_{2}^{(1)}$	$X_{2}^{(2)}$	$X_2^{(3)}$	$\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{2}^{(j)}$		$\sum_{j}^{j} X_{2}^{(j)}$ $\sum_{j} X_{2}^{(j)}$	$\sum_{j}^{J} X_{2}^{(j)}$ $\sum_{j} X_{2}^{(j)}$
$X_3^{(0)}$	$X_3^{(1)}$	$X_2^{(2)} X_3^{(2)}$	$X_{3}^{(3)}$ All-re	$\sum_{j} X_3^{(j)}$ duce	$\sum_{j} X_2^{(j)}$ $\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{(j)}$	$\sum_{j} X_2^{(j)} $ $\sum_{j} X_3^{(j)}$
$X_3^{(0)}$	$X_3^{(1)}$	$X_2^{(2)}$ $X_3^{(2)}$ Node	$X_{3}^{(3)}$ All-re	$\sum_{j} X_3^{(j)}$ duce	$\sum_{j} X_2^{(j)}$ $\sum_{j} X_3^{(j)}$	$\sum_{j} X_3^{(j)}$	$\sum_{j}^{J} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ Node
$X_3^{(0)}$	$X_3^{(1)}$ Node	$X_{2}^{(2)} \ X_{3}^{(2)}$ Node 2	$X_{3}^{(3)}$ Node	$\sum_{j} X_{3}^{(j)}$ duce Node	$\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ Node	$\sum_{j} X_{3}^{(j)}$ Node	3
$X_{3}^{(0)}$ Node $X_{0}^{(0)}$	$X_3^{(1)}$ Node $X_0^{(1)}$	$X_{2}^{(2)} \ X_{3}^{(2)}$ Node $X_{3}^{(2)}$	$X_{3}^{(3)}$ All-re Node $X_{3}^{(3)}$	$\sum_{j} X_{3}^{(j)}$ duce Node	$\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ Node	$\sum_{j} X_{3}^{(j)}$ Node	3
$X_{3}^{(0)}$ Node $X_{0}^{(0)}$	$X_3^{(1)}$ Node $X_0^{(1)}$	$X_{2}^{(2)} \ X_{3}^{(2)}$ Node $X_{3}^{(2)}$	$X_{3}^{(3)}$ Node	$\sum_{j} X_{3}^{(j)}$ duce Node	$\sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)}$ <b>Node</b> $\frac{1}{X_{1}^{(0)}}$	$\begin{array}{c c} \sum_{j} X_3^{(j)} \\ \hline \mathbf{Node} \\ \hline 2 \\ X_2^{(0)} \end{array}$	$X_3^{(0)}$
$X_3^{(0)}$ Node	$X_{3}^{(1)}$ Node $X_{0}^{(1)}$ $X_{1}^{(1)}$	$X_{2}^{(2)} \ X_{3}^{(2)}$ Node $X_{0}^{(2)} \ X_{1}^{(2)}$	$X_{3}^{(3)}$ All-re Node 3 $X_{0}^{(3)}$ $X_{1}^{(3)}$ $X_{2}^{(3)}$	$\frac{\sum_{j} X_{3}^{(j)}}{\text{duce}}$ <b>Node</b> $\frac{0}{X_{0}^{(0)}}$	$\sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)}$ Node $\frac{1}{X_{1}^{(0)}}$ $X_{1}^{(1)}$ $X_{1}^{(2)}$	$\begin{array}{c c} \sum_{j} X_{3}^{(j)} \\ \hline \textbf{Node} \\ \hline \textbf{2} \\ X_{2}^{(0)} \\ X_{2}^{(1)} \end{array}$	$X_3^{(0)}$ $X_3^{(1)}$

- A ring with N nodes partitions data to N messages
- Collective Communication Flow:





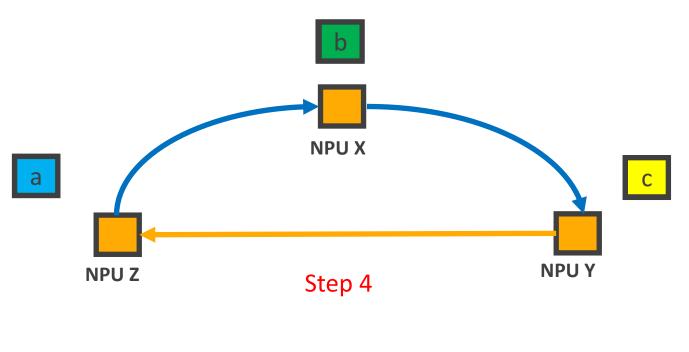
- A ring with N nodes partitions data to N messages
- Collective Communication Flow:



Reduce-Scatter phase done!

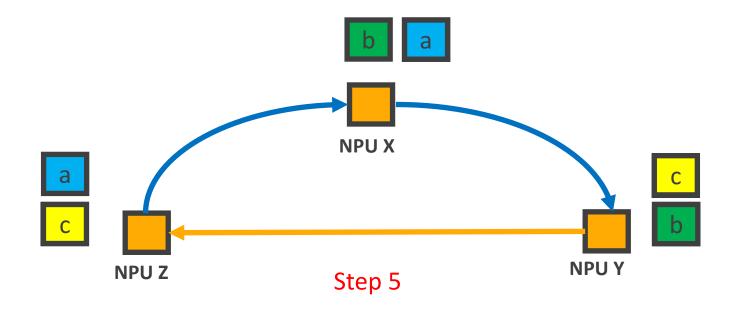
Node	Node	Node	Node	Node	Node	Node	Node
0	$\begin{bmatrix} 1 \end{bmatrix}$	2	3	0	$\begin{bmatrix} 1 \end{bmatrix}$	2	3
$X_0^{(0)}$	$X_{0}^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$	$\sum X_0^{(j)}$			
$X_1^{(0)}$			$X_1^{(3)} \rightarrow$	<b>_</b> j 0	$\sum X_1^{(j)}$		
$X_{2}^{(0)}$	$X_1^{(1)}$			•		$\sum_{i} X_{2}^{(j)}$	
	$X_2^{(1)}$			luce		<b>_</b> j <sup>2</sup>	$\sum X_2^{(j)}$
$X_3^{(0)}$			$X_3^{(3)}$ Rec	ittei	l	l	$\Delta_{i}^{-3}$
	Node	Node	Node	Node	Node <sub> </sub>	- 1	
0	1	2	3 .	0	1	2	3_
X0				X0	X0	X0	X0
	<i>X</i> 1		→	<i>X</i> 1	<i>X</i> 1	<i>X</i> 1	X1
		<i>X</i> 2		<i>X</i> 2	<i>X</i> 2	<i>X</i> 2	X2
			<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	X3
	I	ı	l <sub>All-gat</sub>	her l	I	I	
Node	Node	Node	Node	Node	Node	Node	Node
0	_1_	2	3	0	1	2	3
	$X_0^{(1)}$	$\overline{}$			$\sum_{j} X_0^{(j)}$	$\sum_{j} X_0^{(j)}$	$X_0^{(j)}$
$X_0^{(0)}$		$X_0^{(2)}$	$X_0^{(3)}$	$\frac{0}{\sum_{j} X_{0}^{(j)}}$		$\sum_{j} X_0^{(j)}$	$egin{array}{c} {\bf 3} \ \sum_j X_0^{(j)} \ \sum_i X_1^{(j)} \end{array}$
$X_0^{(0)} X_1^{(0)}$	$X_0^{(1)}$	$X_0^{(2)} X_1^{(2)}$	$X_0^{(3)}$ $X_1^{(3)} \rightarrow$	0		$\sum_{j} X_0^{(j)}$	$ \begin{array}{c} 3 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \end{array} $
$X_0^{(0)} X_1^{(0)}$	$X_0^{(1)} X_1^{(1)}$	$X_0^{(2)} X_1^{(2)} X_2^{(2)}$	$X_0^{(3)} \\ X_1^{(3)} \to X_2^{(3)}$	$\frac{0}{\sum_{j} X_{0}^{(j)}}$		$\sum_{j} X_0^{(j)}$	$ \begin{array}{c} 3 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \end{array} $
$X_0^{(0)} \ X_1^{(0)} \ X_2^{(0)} \ X_3^{(0)}$	$X_0^{(1)} X_1^{(1)} X_1^{(1)} X_2^{(1)} X_3^{(1)}$	$X_0^{(2)} \\ X_1^{(2)} \\ X_2^{(2)} \\ X_3^{(2)}$	$X_0^{(3)}$ $X_1^{(3)} \rightarrow X_2^{(3)}$ $X_3^{(3)}$ All-rec	$\begin{array}{c} 0 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \end{array}$	$\sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)}$	$ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} $	$ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} $
$X_0^{(0)} \ X_1^{(0)} \ X_2^{(0)} \ X_3^{(0)}$	$X_0^{(1)} X_1^{(1)}$	$X_0^{(2)} \\ X_1^{(2)} \\ X_2^{(2)} \\ X_3^{(2)}$	$X_0^{(3)}$ $X_1^{(3)} \rightarrow X_2^{(3)}$ $X_3^{(3)}$ All-rec	$\begin{array}{c} 0 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \end{array}$	$\sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)}$	$ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} $	$\begin{array}{c} 3 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \end{array}$ $\begin{array}{c} \mathbf{Node} \end{array}$
$X_0^{(0)} \ X_1^{(0)} \ X_2^{(0)} \ X_3^{(0)}$	$X_0^{(1)} X_1^{(1)} X_1^{(1)} X_2^{(1)} X_3^{(1)}$	$X_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)}$ Node	$X_0^{(3)}$ $X_1^{(3)}$ $X_2^{(3)}$ $X_3^{(3)}$ All-reconde	$\begin{array}{c} 0 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \end{array}$	$\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ Node	$\sum_{j} X_{0}^{(j)} \sum_{j} X_{1}^{(j)} \sum_{j} X_{1}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)}$ Node	$\sum_{j} X_0^{(j)}$ $\sum_{j} X_1^{(j)}$ $\sum_{j} X_2^{(j)}$ $\sum_{j} X_3^{(j)}$ Phode 3
$X_0^{(0)} \ X_1^{(0)} \ X_2^{(0)} \ X_3^{(0)}$	$X_0^{(1)} X_1^{(1)} X_1^{(1)} X_2^{(1)} X_3^{(1)}$	$X_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)}$ Node	$X_0^{(3)}$ $X_1^{(3)}$ $X_2^{(3)}$ $X_3^{(3)}$ All-reconde	$\begin{array}{c} {\bf 0} \\ \sum_{j} X_0^{(j)} \\ \sum_{j} X_1^{(j)} \\ \sum_{j} X_2^{(j)} \\ \sum_{j} X_3^{(j)} \\ {\bf Node} \end{array}$	$\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ Node	$\sum_{j} X_{0}^{(j)} \sum_{j} X_{1}^{(j)} \sum_{j} X_{1}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)}$ Node	$\sum_{j} X_0^{(j)}$ $\sum_{j} X_1^{(j)}$ $\sum_{j} X_2^{(j)}$ $\sum_{j} X_3^{(j)}$ Phode 3
$X_0^{(0)} \ X_1^{(0)} \ X_1^{(0)} \ X_2^{(0)} \ X_3^{(0)}$ Node $0$	$X_0^{(1)} \ X_1^{(1)} \ X_2^{(1)} \ X_3^{(1)}$ Node	$X_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)}$ Node $\mathbf{z}$	$X_0^{(3)}$ $X_1^{(3)} \rightarrow X_2^{(3)}$ $X_2^{(3)}$ $X_3^{(3)}$ All-recondense $X_0^{(3)}$	$\begin{array}{c} {\color{red} {\bf 0}} \\ {\color{red} {\sum_j} X_0^{(j)}} \\ {\color{red} {\sum_j} X_1^{(j)}} \\ {\color{red} {\sum_j} X_2^{(j)}} \\ {\color{red} {\sum_j} X_3^{(j)}} \\ {\color{red} {\bf duce}} \\ {\color{red} {\bf Node}} \\ {\color{red} {\bf 0}} \\ {\color{red} {X_0^{(0)}}} \end{array}$	$\sum_{j} X_{1}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)} \sum_{j} X_{3}^{(j)}$ Node  1 $X_{1}^{(0)}$	$\sum_{j} X_{0}^{(j)} \sum_{j} X_{1}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)}$ Node 2 $X_{2}^{(0)}$	$\begin{array}{c} \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \end{array}$ $\begin{array}{c} \text{Node} \\ 3 \\ X_{3}^{(0)} \end{array}$
$X_0^{(0)} \ X_1^{(0)} \ X_1^{(0)} \ X_3^{(0)} \ X_3^{(0)} \ Node \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \$	$X_0^{(1)} \ X_1^{(1)} \ X_2^{(1)} \ X_3^{(1)}$ Node $1$ $X_0^{(1)} \ X_1^{(1)}$	$X_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)} \ $ Node $2$ $X_0^{(2)} \ X_1^{(2)}$	$X_0^{(3)}$ $X_1^{(3)}$ $X_2^{(3)}$ $X_2^{(3)}$ $X_3^{(3)}$ All-rec Node $X_0^{(3)}$ $X_0^{(3)}$	$\begin{array}{c} 0 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \text{duce} \\ \mathbf{Node} \\ 0 \\ X_{0}^{(0)} \\ \end{array}$	$\sum_{j} X_{1}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)}$ Node $\frac{1}{X_{1}^{(0)}}$	$\sum_{j} X_{0}^{(j)}$ $\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ Node $\frac{2}{2}$ $X_{2}^{(0)}$ $X_{2}^{(0)}$	$\sum_{J} X_0^{(J)} \sum_{J} X_2^{(J)} \sum_{J} X_2^{(J)} \sum_{J} X_3^{(J)}$ P. Node $\frac{3}{X_3^{(0)}}$ $X_3^{(1)}$ $X_2^{(2)}$
$X_0^{(0)} \ X_1^{(0)} \ X_1^{(0)} \ X_2^{(0)} \ X_3^{(0)}$ Node $0$	$X_0^{(1)} \ X_1^{(1)} \ X_2^{(1)} \ X_3^{(1)}$ Node $1$ $X_0^{(1)} \ X_1^{(1)}$	$X_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)}$ Node $\mathbf{z}$ $X_0^{(2)} \ X_1^{(2)}$	$X_0^{(3)}$ $X_1^{(3)}$ $X_2^{(3)}$ $X_2^{(3)}$ $X_3^{(3)}$ All-rec Node $X_0^{(3)}$ $X_0^{(3)}$	$\begin{array}{c} {\color{red} {\bf 0}} \\ {\color{red} {\sum_j} X_0^{(j)}} \\ {\color{red} {\sum_j} X_1^{(j)}} \\ {\color{red} {\sum_j} X_2^{(j)}} \\ {\color{red} {\sum_j} X_3^{(j)}} \\ {\color{red} {\bf duce}} \\ {\color{red} {\bf Node}} \\ {\color{red} {\bf 0}} \\ {\color{red} {X_0^{(0)}}} \end{array}$	$\sum_{j} X_{1}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)}$ Node $\frac{1}{X_{1}^{(0)}}$ $X_{1}^{(1)}$ $X_{1}^{(2)}$	$\sum_{j} X_{0}^{(j)} \sum_{j} X_{1}^{(j)} \sum_{j} X_{1}^{(j)} \sum_{j} X_{3}^{(j)} \sum_{j} X_{3}^{(j)}$ Node 2 $X_{2}^{(0)}$	$\sum_{J} X_0^{(J)} \sum_{J} X_2^{(J)} \sum_{J} X_2^{(J)} \sum_{J} X_3^{(J)}$ P. Node $\frac{3}{X_3^{(0)}}$ $X_3^{(1)}$ $X_2^{(2)}$

- A ring with N nodes partitions data to N messages
- Collective Communication Flow:



Node	Node	Node	Node	Node	Node	Node	Node
0	1	2	3	_0_	1	2	3
$X_0^{(0)}$	$X_{0}^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$	$\sum_{i} X_0^{(j)}$	]		
$X_1^{(0)}$	$X^{(1)}$	$X_1^{(2)}$	-	<b>—</b> ,	$\sum_{i} X_{1}^{(j)}$		İ
$X_{2}^{(0)}$	$\mathbf{Y}^{(1)}$	$X_{2}^{(2)}$				$\sum_{i} X_2^{(j)}$	
$X_{3}^{(0)}$			$X_3^{(3)}$ Rec	luce			$\sum_{i} X_3^{(j)}$
5			-80	ittei	l Nodel	l Nodo	l—, Node
	Noae	Noae	Node	Node	Node	- 1	
0	1	2	3	_0 ∤	<del></del> {	2	3_
X0				X0	X0	X0	X0
	<i>X</i> 1		-	<i>X</i> 1	<i>X</i> 1	<i>X</i> 1	X1
		<i>X</i> 2		<i>X</i> 2	<i>X</i> 2	<i>X</i> 2	X2
			<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	X3
			All-gat	her	ı	ı	
Node	Node	Node	Node	Node	Node	Node	Node
_			_			١ ۾	_
0	$\lfloor 1 \rfloor$	2	3	0		<u> </u>	<u> </u>
	<u> </u>	-			$\sum_{i} X_0^{(j)}$	$\sum_{i} X_0^{(j)}$	$\sum_{i} X_0^{(j)}$
$X_0^{(0)}$	$X_0^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$	$\sum_{j} X_0^{(j)}$		$\frac{2}{\sum_{j} X_0^{(j)}}$ $\sum_{j} X_1^{(j)}$	$\sum_{i} X_0^{(j)}$
$X_0^{(0)} X_1^{(0)}$	$X_0^{(1)}$	$X_0^{(2)} X_1^{(2)}$	$X_0^{(3)}$ $X_1^{(3)} \rightarrow$	$\sum_{j} X_0^{(j)}$ $\sum_{j} X_1^{(j)}$	$\sum_{j} X_1^{(j)}$	$\sum_{j} X_1^{(j)}$	$\frac{\sum_{j} X_0^{(j)}}{\sum_{j} X_1^{(j)}}$
$X_0^{(0)} \ X_1^{(0)} \ X_2^{(0)}$	$X_0^{(1)} X_1^{(1)}$	$X_0^{(2)}$	$X_0^{(3)} \\ X_1^{(3)} \to X_2^{(3)}$	$\sum_{j} X_0^{(j)}$	$\sum_{j} X_1^{(j)}$	$\sum_{j} X_{0}^{(j)} \ \sum_{j} X_{1}^{(j)} \ \sum_{j} X_{2}^{(j)} \ \sum_{j} X_{2}^{(j)}$	$\frac{\sum_{j} X_0^{(j)}}{\sum_{j} X_1^{(j)}}$
$X_0^{(0)} \\ X_1^{(0)} \\ X_2^{(0)} \\ X_3^{(0)}$	$X_0^{(1)} X_1^{(1)} X_2^{(1)} X_3^{(1)}$	$X_0^{(2)} \\ X_1^{(2)} \\ X_2^{(2)} \\ X_3^{(2)}$	$X_0^{(3)}$ $X_1^{(3)} \rightarrow X_2^{(3)}$ $X_3^{(3)}$ All-rec	$\sum_{j} X_{0}^{(j)}$ $\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ duce	$\sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)}$	$\frac{\sum_{j} X_{1}^{(j)}}{\sum_{j} X_{2}^{(j)}}$ $\frac{\sum_{j} X_{3}^{(j)}}{\sum_{j} X_{3}^{(j)}}$	$ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} $
$X_0^{(0)} \\ X_1^{(0)} \\ X_2^{(0)} \\ X_3^{(0)}$	$X_0^{(1)} X_1^{(1)} X_2^{(1)} X_3^{(1)}$	$X_0^{(2)} \\ X_1^{(2)} \\ X_2^{(2)} \\ Y_2^{(2)}$	$X_0^{(3)}$ $X_1^{(3)} \rightarrow X_2^{(3)}$ $X_3^{(3)}$ All-rec	$\sum_{j} X_{0}^{(j)}$ $\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ duce	$\sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)}$	$\frac{\sum_{j} X_{1}^{(j)}}{\sum_{j} X_{2}^{(j)}}$ $\frac{\sum_{j} X_{3}^{(j)}}{\sum_{j} X_{3}^{(j)}}$	$\frac{\sum_{j} X_0^{(j)}}{\sum_{j} X_1^{(j)}}$
$X_0^{(0)} \\ X_1^{(0)} \\ X_2^{(0)} \\ X_3^{(0)}$	$X_0^{(1)} X_1^{(1)} X_2^{(1)} X_3^{(1)}$	$X_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)}$ Node	$X_0^{(3)}$ $X_1^{(3)}$ $X_2^{(3)}$ $X_3^{(3)}$ All-reconde	$\sum_{j} X_{0}^{(j)}$ $\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ duce	$\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ Node	$\sum_{j} X_{1}^{(j)} $ $\sum_{j} X_{2}^{(j)} $ $\sum_{j} X_{3}^{(j)} $ $\sum_{j} Node$	$ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} $ Node
$X_0^{(0)} \\ X_1^{(0)} \\ X_2^{(0)} \\ X_3^{(0)}$	$X_0^{(1)} X_1^{(1)} X_2^{(1)} X_3^{(1)}$	$X_0^{(2)} \ X_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)}$ Node $\mathbf{z}$	$X_0^{(3)}$ $X_1^{(3)} \rightarrow X_2^{(3)}$ $X_3^{(3)}$ Node $X_0^{(3)}$	$\sum_{j} X_{0}^{(j)}$ $\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ duce	$\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ Node	$\sum_{j} X_{1}^{(j)} $ $\sum_{j} X_{2}^{(j)} $ $\sum_{j} X_{3}^{(j)} $ $\sum_{j} Node$	$ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} $ Node
$X_0^{(0)} \ X_1^{(0)} \ X_2^{(0)} \ X_3^{(0)}$ Node	$X_0^{(1)} \ X_1^{(1)} \ X_2^{(1)} \ X_3^{(1)}$ Node	$X_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)}$ Node $\mathbf{Z}$	$X_0^{(3)}$ $X_1^{(3)} \rightarrow X_2^{(3)}$ $X_3^{(3)}$ Node $X_0^{(3)}$	$\sum_{j} X_{0}^{(j)}$ $\sum_{j} X_{1}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{3}^{(j)}$ duce Node	$\sum_{j} X_{1}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)}$ Node  1 $X_{1}^{(0)}$	$\sum_{j} X_{1}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)} \sum_{j} X_{3}^{(j)}$ Node 2 $X_{2}^{(0)}$	$\sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \text{Node} \\ 3 \\ X_{3}^{(0)}$
$X_0^{(0)} \ X_1^{(0)} \ X_1^{(0)} \ X_3^{(0)}$ Node $X_0^{(0)} \ X_0^{(0)}$	$X_0^{(1)} \ X_1^{(1)} \ X_2^{(1)} \ X_3^{(1)}$ Node	$X_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)}$ Node $\mathbf{z}$ $X_0^{(2)} \ X_1^{(2)}$	$X_0^{(3)}$ $X_1^{(3)}$ $X_2^{(3)}$ $X_2^{(3)}$ $X_3^{(3)}$ Node $X_0^{(3)}$ $X_0^{(3)}$	$\frac{\sum_{j} X_{0}^{(j)}}{\sum_{j} X_{1}^{(j)}} \cdot \sum_{j} X_{2}^{(j)}$ $\sum_{j} X_{2}^{(j)}$ duce $\frac{0}{X_{0}^{(0)}}$	$\sum_{j} X_{1}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)} $ Node $\frac{1}{X_{1}^{(0)}}$ $X_{1}^{(1)}$	$\sum_{j} X_{1}^{(j)} \sum_{j} X_{2}^{(j)} \sum_{j} X_{3}^{(j)} \sum_{j} X_{3}^{(j)}$ Node 2 $X_{2}^{(0)}$	$\sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \text{Node} \\ 3 \\ X_{3}^{(0)} \\ X_{3}^{(1)}$

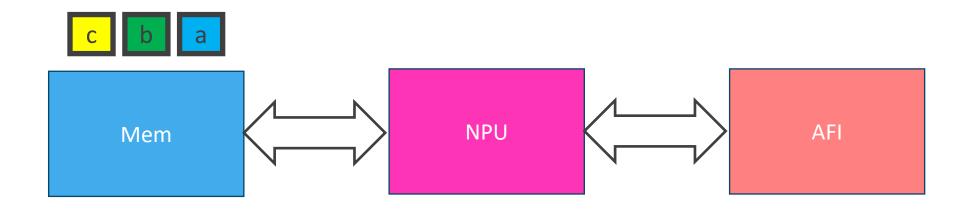
- A ring with N nodes partitions data to N messages
- Collective Communication Flow:



All-Reduce done

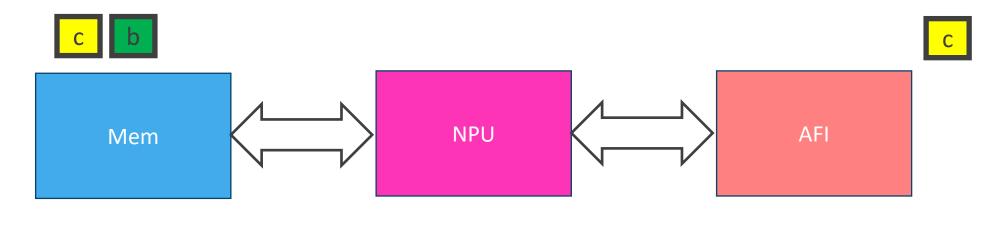
Node	Node	Node	Node	Node	Node	Node	Node
_0_	1	2	3	_0	1_	2	3
	$X_{2}^{(1)}$	$X_1^{(2)} \ X_2^{(2)}$	$X_1^{(3)} \rightarrow X_2^{(3)}$	$\sum_{j} X_0^{(j)}$	$\sum_j X_1^{(j)}$	$\sum_{j} X_2^{(j)}$	<b>\</b>
$X_{3}^{(0)}$	$X_3^{(1)}$	$X_3^{(2)}$	$X_3^{(3)}$ Rec	duce			$\sum_{j} X_{3}^{(j)}$
	-		Node	Node,	Node,	Node,	Node
_0_	1	2	3_	_0_	1	2	3
X0		]	]	X0	X0	X0	X0
	<i>X</i> 1		→	- X1	<i>X</i> 1	<i>X</i> 1	<i>X</i> 1
	11.1	<i>X</i> 2		<i>X</i> 2	<i>X</i> 2	<i>X</i> 2	<i>X</i> 2
			<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3	<i>X</i> 3
	l	l	All-gat	ther	I	ı	
Mode	Modo	Node	Nodo	Mode	Mode	Mode	Node
Noue,	Noue,	Noug	Noue	Noue	moue	, Nout	, Nouc
0	Noue 1	2	3	0	1	2	3
_0_	_1_	2	3	0_	1	2	3_
$\frac{0}{X_0^{(0)}}$	$X_0^{(1)}$	$X_0^{(2)}$	$X_0^{(3)}$	$\frac{0}{\sum_{j} X_{0}^{(j)}}$	$\sum_{j} X_0^{(j)}$	$\sum_{j} X_0^{(j)}$	$\frac{3}{\sum_{j} X_0^{(j)}}$
$X_0^{(0)} X_1^{(0)}$	$X_0^{(1)} X_i^{(1)}$	$X_0^{(2)}$ $X_1^{(2)}$	$\begin{array}{c} 3 \\ X_0^{(3)} \\ X_1^{(3)} \rightarrow \end{array}$	$\frac{0}{\sum_{j} X_{0}^{(j)}}$ $\sum_{j} X_{1}^{(j)}$	$\sum_{j} X_0^{(j)} \sum_{j} X_1^{(j)}$	$egin{array}{c} 2 \ \sum_{j} X_0^{(j)} \ \sum_{j} X_1^{(j)} \end{array}$	$\begin{array}{c} 3 \\ \sum_{j} X_0^{(j)} \\ \sum_{i} X_1^{(j)} \end{array}$
$ \begin{array}{c}                                     $	$\begin{array}{c} 1 \\ X_0^{(1)} \\ X_1^{(1)} \\ X_2^{(1)} \end{array}$	$X_0^{(2)}$ $X_1^{(2)}$ $X_2^{(2)}$	$ \begin{array}{c} 3 \\ X_0^{(3)} \\ X_1^{(3)} \\ X_2^{(3)} \end{array} $	$ \frac{0}{\sum_{j} X_{0}^{(j)}} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} $	$\sum_{j} X_{0}^{(j)} \sum_{j} X_{1}^{(j)} \sum_{j} X_{1}^{(j)}$	$egin{array}{c} {f 2} \ \sum_j X_0^{(j)} \ \sum_j X_1^{(j)} \ \sum_j X_2^{(j)} \ \end{array}$	$ \begin{array}{c} 3 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \end{array} $
$\begin{array}{c c} 0 \\ X_0^{(0)} \\ X_1^{(0)} \\ X_2^{(0)} \\ X_3^{(0)} \end{array}$	$\begin{array}{c} 1 \\ X_0^{(1)} \\ X_1^{(1)} \\ X_2^{(1)} \\ X_3^{(1)} \end{array}$	$\begin{array}{c} \mathbf{Z} \\ X_0^{(2)} \\ X_1^{(2)} \\ X_2^{(2)} \\ X_3^{(2)} \end{array}$	$\begin{array}{c} 3 \\ X_0^{(3)} \\ X_1^{(3)} \\ X_2^{(3)} \\ X_3^{(3)} \end{array} \rightarrow \begin{array}{c} 3 \\ X_2^{(3)} \\ X_3^{(3)} \end{array}$	$ \frac{0}{\sum_{j} X_{0}^{(j)}} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ $ duce	$\begin{array}{c} 1 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \end{array}$	$egin{array}{c} {f 2} \ \sum_j X_0^{(j)} \ \sum_j X_1^{(j)} \ \sum_j X_2^{(j)} \ \sum_j X_3^{(j)} \ \end{array}$	$\begin{array}{c c} 3 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \end{array}$
$\begin{array}{c c} 0 \\ X_0^{(0)} \\ X_1^{(0)} \\ X_2^{(0)} \\ X_3^{(0)} \end{array}$	$\begin{array}{c} 1 \\ X_0^{(1)} \\ X_1^{(1)} \\ X_2^{(1)} \\ X_3^{(1)} \end{array}$	$X_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)}$ Node	$\begin{array}{c} 3 \\ X_0^{(3)} \\ X_1^{(3)} \\ X_2^{(3)} \\ X_3^{(3)} \\ \text{Node} \end{array}$	$\begin{array}{c} {\bf 0} \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ {\bf duce} \\ {\bf Node} \end{array}$	$\begin{array}{c} 1 \\ \sum_{j} X_0^{(j)} \\ \sum_{j} X_1^{(j)} \\ \sum_{j} X_2^{(j)} \\ \sum_{j} X_3^{(j)} \end{array}$ $\begin{array}{c} \mathbf{Node} \end{array}$	$egin{array}{c} {f 2} \ \sum_j X_0^{(j)} \ \sum_j X_1^{(j)} \ \sum_j X_2^{(j)} \ \sum_j X_3^{(j)} \ \end{array}$	$egin{array}{c} {\bf 3} \\ {\sum_j X_0^{(j)}} \\ {\sum_j X_1^{(j)}} \\ {\sum_j X_2^{(j)}} \\ {\sum_j X_3^{(j)}} \\ {f Node} \\ {f $
$egin{array}{c} oldsymbol{0} \ X_0^{(0)} \ X_1^{(0)} \ X_2^{(0)} \ X_3^{(0)} \ \end{pmatrix}$ Node	$egin{array}{c} {f 1} \ X_0^{(1)} \ X_1^{(1)} \ X_2^{(1)} \ X_3^{(1)} \ \end{array}$ Node	$egin{array}{c} {f Z}_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)} \ \end{array}$ Node	$\begin{matrix} 3 \\ X_0^{(3)} \\ X_1^{(3)} \\ X_2^{(3)} \\ X_3^{(3)} \\ \mathbf{All-re} \\ \mathbf{Node} \\ 3 \end{matrix}$	$\begin{array}{c} {\bf 0} \\ \sum_j X_0^{(j)} \\ \sum_j X_1^{(j)} \\ \sum_j X_2^{(j)} \\ \sum_j X_3^{(j)} \\ {\bf duce} \\ {\bf Node} \\ {\bf 0} \end{array}$	$\begin{array}{c} 1 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \sum_{j} X_{3}^{(j)} \end{array}$ $\begin{array}{c} 1 \\ \mathbf{Node} \\ 1 \end{array}$	$egin{array}{c} {f 2} \ \sum_j X_0^{(j)} \ \sum_j X_1^{(j)} \ \sum_j X_2^{(j)} \ \sum_j X_3^{(j)} \ \sum_j X_3^{(j)} \ \end{array}$	$egin{array}{c} {\bf 3} \\ {\sum_j X_0^{(j)}} \\ {\sum_j X_1^{(j)}} \\ {\sum_j X_2^{(j)}} \\ {\sum_j X_3^{(j)}} \\ {f Node} \\ {f 3} \\ \end{array}$
$\begin{array}{c} {\bf 0} \\ X_0^{(0)} \\ X_1^{(0)} \\ X_2^{(0)} \\ X_3^{(0)} \\ {\bf Node} \\ {\bf 0} \\ X_0^{(0)} \end{array}$	$\begin{matrix} 1 \\ X_0^{(1)} \\ X_1^{(1)} \\ X_2^{(1)} \\ X_3^{(1)} \end{matrix}$ Node $\begin{matrix} 1 \\ X_0^{(1)} \end{matrix}$	$egin{array}{c} {f Z} & {f Z}_0^{(2)} & X_1^{(2)} & X_1^{(2)} & X_2^{(2)} & X_3^{(2)} & {f Node} & {f Z} & X_0^{(2)} & {f Z}_0^{(2)} & $	$\begin{array}{c} {\bf 3} \\ X_0^{(3)} \\ X_1^{(3)} \\ X_2^{(3)} \\ X_3^{(3)} \\ {\bf Node} \\ {\bf 3} \\ X_0^{(3)} \end{array}$	$\begin{array}{c} 0 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \mathbf{duce} \\ \mathbf{Node} \\ 0 \\ X_{0}^{(0)} \end{array}$	$\begin{array}{c} 1 \\ \sum_{j} X_0^{(j)} \\ \sum_{j} X_3^{(j)} \\ \sum_{j} X_3^{(j)} \\ \mathbf{Node} \\ 1 \\ X_1^{(0)} \end{array}$	$\begin{array}{c c} 2 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \end{array}$ $\begin{array}{c c} \mathbf{Node} \\ 2 \\ X_{2}^{(0)} \end{array}$	$\begin{array}{c c} 3 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \end{array}$ $\begin{array}{c c} \mathbf{Node} \\ 3 \\ X_{3}^{(0)} \end{array}$
$\begin{array}{c} {\color{red}0} \\ {\color{blue}X_0^{(0)}} \\ {\color{blue}X_1^{(0)}} \\ {\color{blue}X_1^{(0)}} \\ {\color{blue}X_3^{(0)}} \\ {\color{blue}\mathbf{Node}} \\ {\color{blue}0} \\ {\color{blue}X_1^{(0)}} \end{array}$	$X_0^{(1)} \ X_0^{(1)} \ X_2^{(1)} \ X_3^{(1)}$ Node $X_0^{(1)} \ X_0^{(1)}$	$egin{array}{c} {f 2} \ X_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)} \ \end{array}$ Node $f 2$ $egin{array}{c} X_0^{(2)} \ X_1^{(2)} \ \end{array}$	$\begin{array}{c} {\bf 3} \\ X_0^{(3)} \\ X_1^{(3)} \\ X_2^{(2)} \\ X_3^{(3)} \\ {\bf Node} \\ {\bf 3} \\ X_0^{(3)} \\ X_1^{(3)} \\ \end{array}$	$\begin{array}{c} {\bf 0} \\ \sum_{j} X_0^{(j)} \\ \sum_{j} X_1^{(j)} \\ \sum_{j} X_2^{(j)} \\ \sum_{j} X_3^{(j)} \\ {\bf duce} \\ {\bf Node} \\ {\bf 0} \end{array}$	$\begin{array}{c} 1 \\ \sum_{j} X_{0}^{(0)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ 1 \\ X_{1}^{(0)} \\ X_{1}^{(1)} \end{array}$	$\begin{array}{ c c } \textbf{2} \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \textbf{Node} \\ \textbf{2} \\ X_{2}^{(0)} \\ X_{2}^{(1)} \end{array}$	$\begin{array}{c c} 3 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \end{array}$ $\begin{array}{c c} \mathbf{Node} \\ 3 \\ X_{3}^{(0)} \\ \end{array}$
$egin{array}{c} oldsymbol{0} \ X_0^{(0)} \ X_1^{(0)} \ X_2^{(0)} \ X_3^{(0)} \ \end{pmatrix}$ Node	$X_0^{(1)} \ X_0^{(1)} \ X_2^{(1)} \ X_3^{(1)}$ Node $X_0^{(1)} \ X_0^{(1)}$	$egin{array}{c} {f 2} \ X_0^{(2)} \ X_1^{(2)} \ X_2^{(2)} \ X_3^{(2)} \ \end{array}$ Node $f 2$ $egin{array}{c} X_0^{(2)} \ X_1^{(2)} \ \end{array}$	$\begin{array}{c} {\bf 3} \\ X_0^{(3)} \\ X_1^{(3)} \\ X_2^{(3)} \\ X_3^{(3)} \\ {\bf Node} \\ {\bf 3} \\ X_0^{(3)} \end{array}$	$\begin{array}{c} 0 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \mathbf{duce} \\ \mathbf{Node} \\ 0 \\ X_{0}^{(0)} \end{array}$	$\begin{array}{c} 1 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ 1 \\ X_{1}^{(0)} \\ X_{1}^{(1)} \end{array}$	$\begin{array}{c c} 2 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{2}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ 2 \\ \mathbf{X}_{2}^{(0)} \\ X_{2}^{(1)} \\ X_{2}^{(2)} \end{array}$	$\begin{array}{c c} 3 \\ \sum_{j} X_{0}^{(j)} \\ \sum_{j} X_{1}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \sum_{j} X_{3}^{(j)} \\ \end{array}$ $\begin{array}{c c} \mathbf{Node} \\ 3 \\ X_{3}^{(0)} \end{array}$

#### **Baseline system**



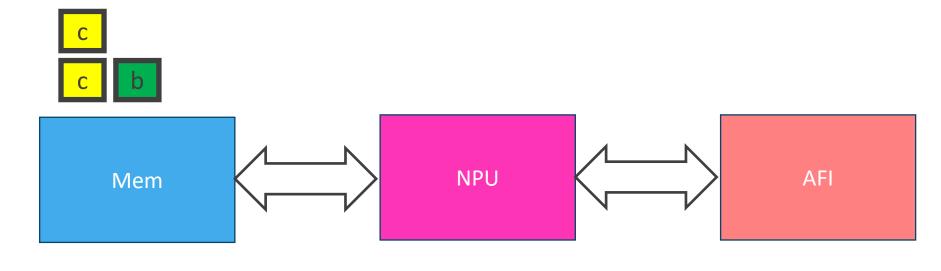
Step 1

#### **Baseline system**



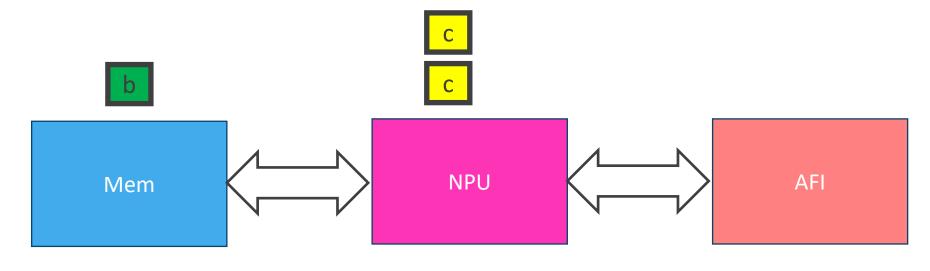
Step 2

#### **Baseline system**



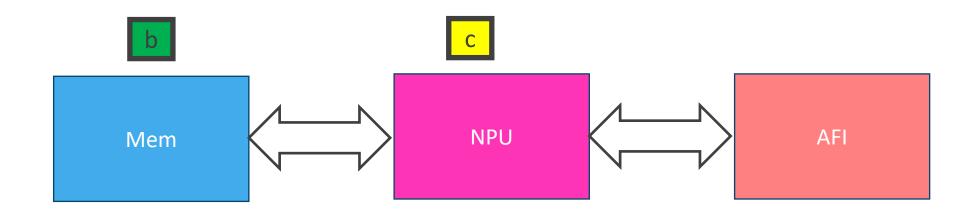
Step 2

#### **Baseline system**



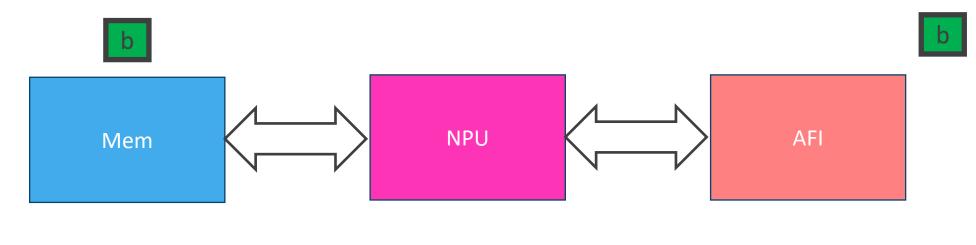
Step 2

#### **Baseline system**



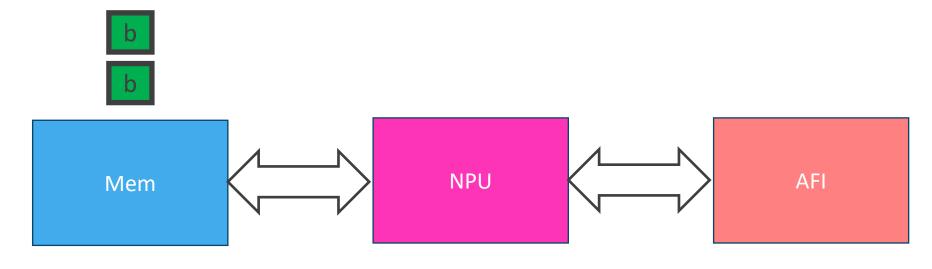
Step 2

#### **Baseline system**



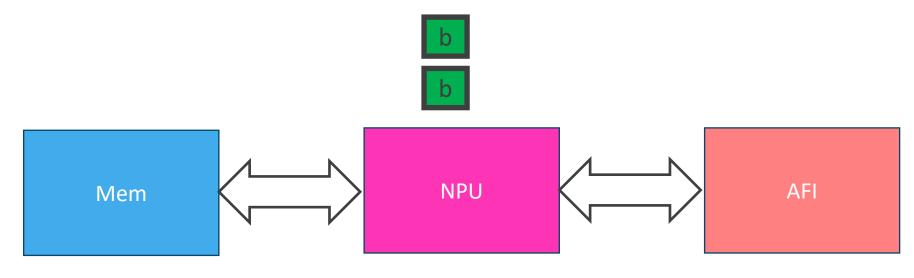
Step 3

#### **Baseline system**



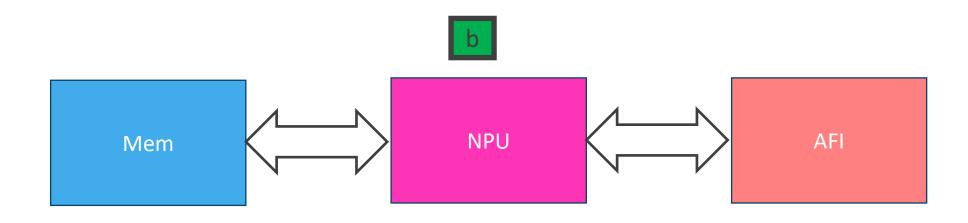
Step 3

#### **Baseline system**



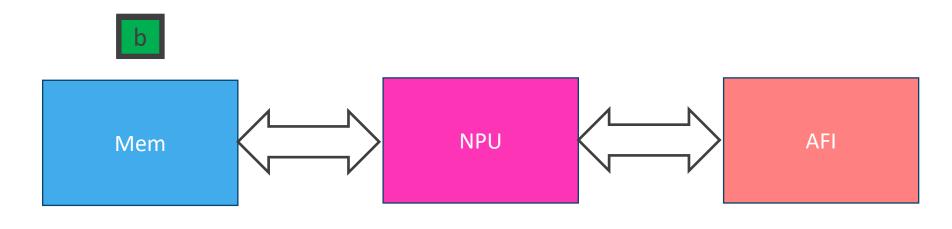
Step 3

#### **Baseline system**



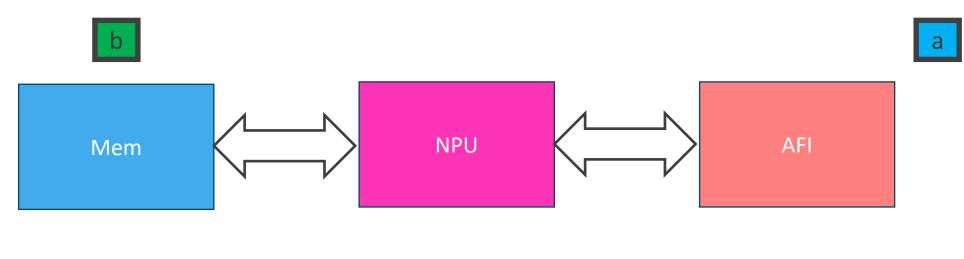
Step 3

#### **Baseline system**



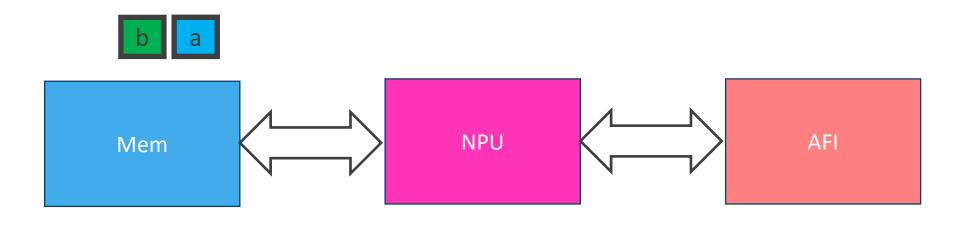
Step 4

#### **Baseline system**



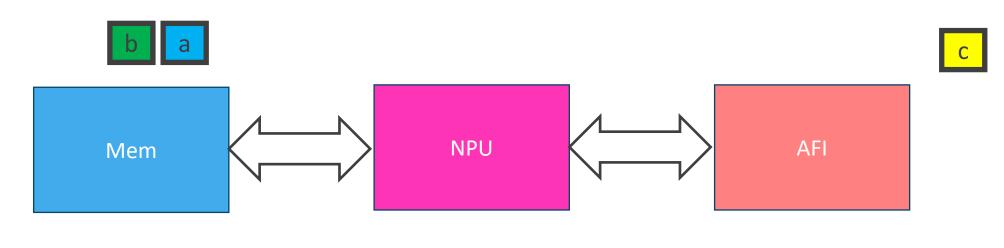
Step 5

#### **Baseline system**



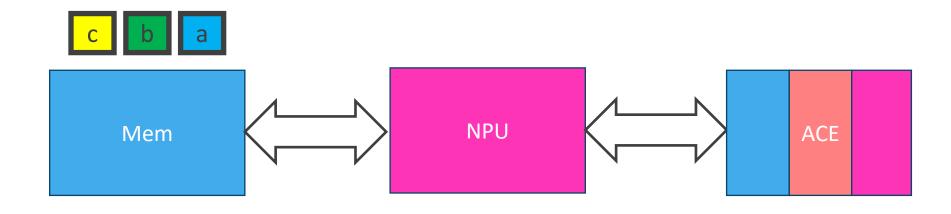
Step 5

#### **Baseline system**



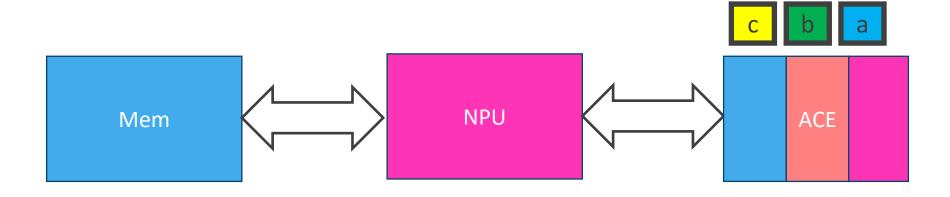
Step 5

#### **ACE**



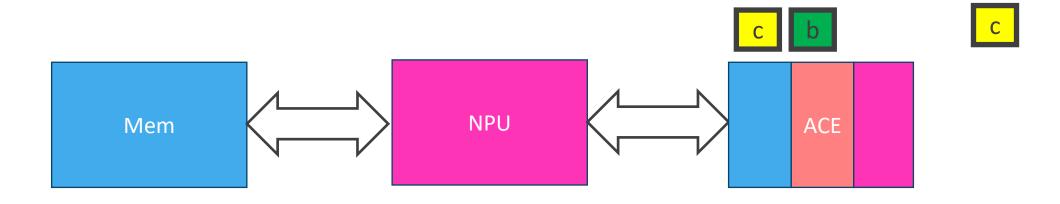
Step 1

**ACE** 



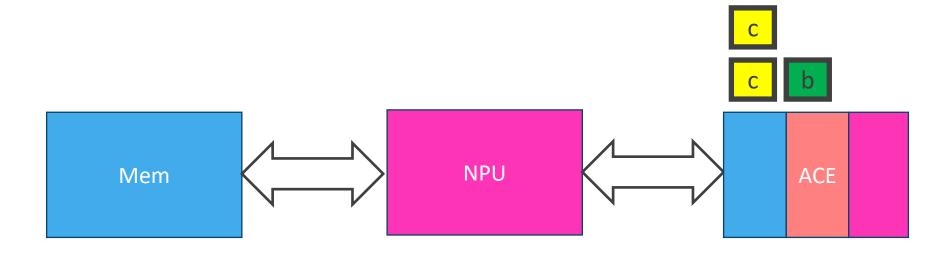
Step 1

**ACE** 



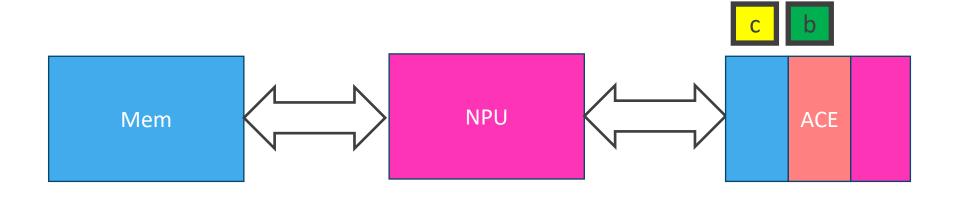
Step 2

#### **ACE**



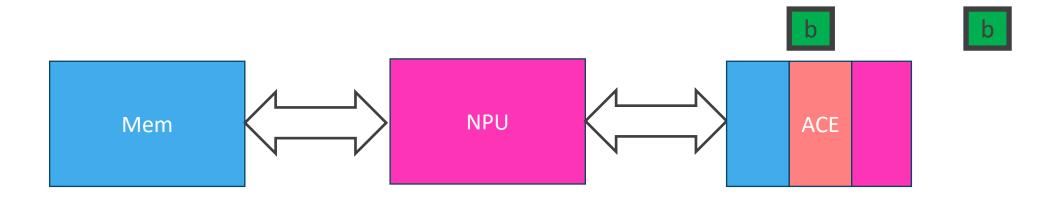
Step 2

**ACE** 



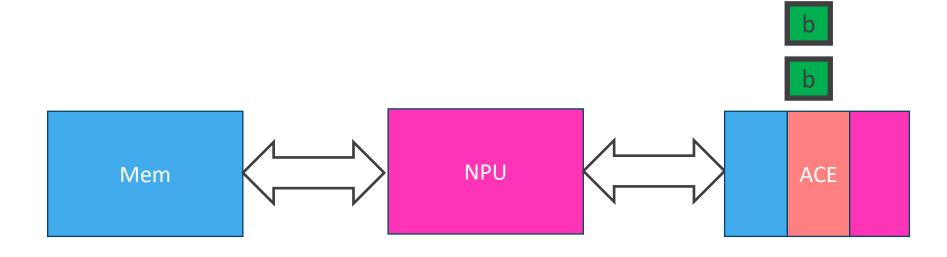
Step 2

**ACE** 



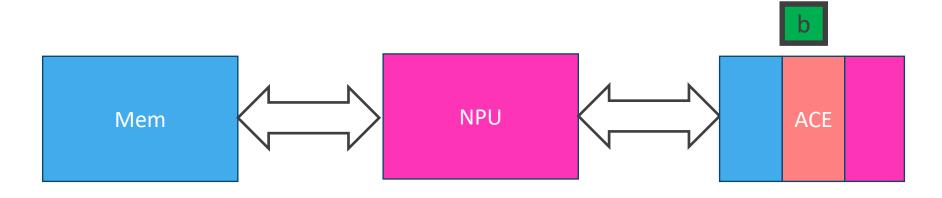
Step 3

#### **ACE**



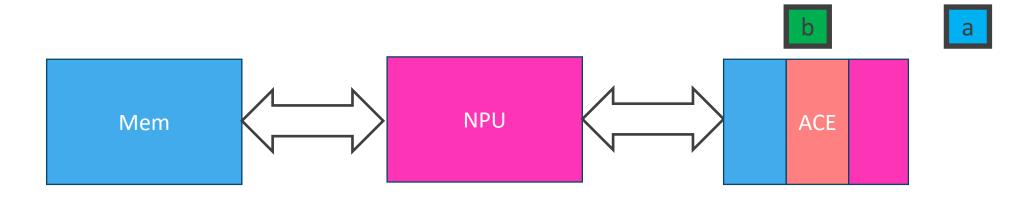
Step 3

**ACE** 



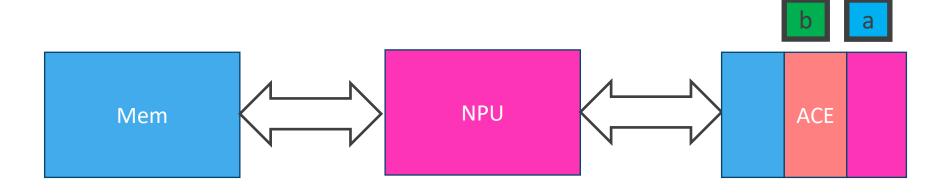
Step 4

**ACE** 



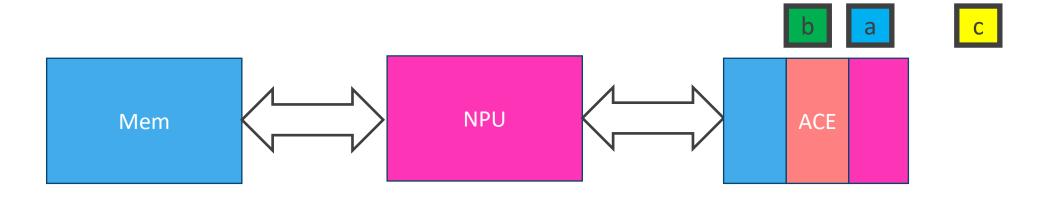
Step 5

**ACE** 



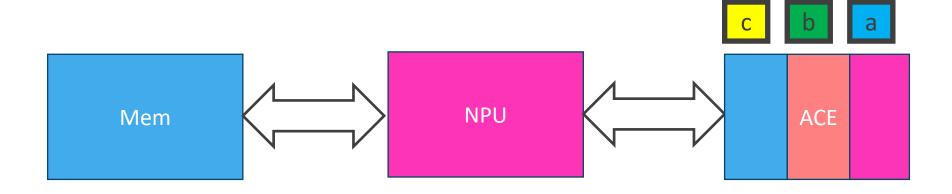
Step 5

**ACE** 



Step 5

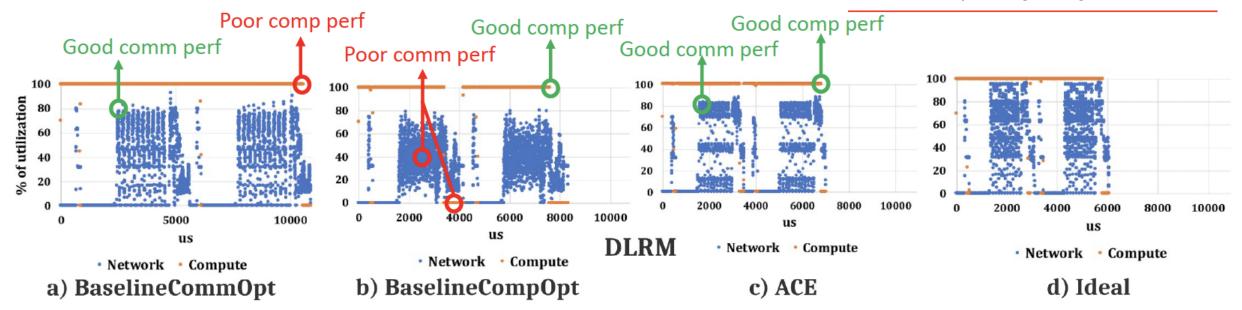
**ACE** 



Step 5

### Speed-Up Results

S. Rashidi et al., "Enabling Compute-Communication Overlap in Distributed Deep Learning Training Platforms". ISCA 2021



**End-to-end training iteration speed-ups** 

Workload	Speedup (avg.)	Speedup (max)
Resnet-50	1.41X	1.51X
GNMT	1.12X	1.17X
DLRM	1.13X	1.19X

#### **Takeaways:**

- Comp-comm overlap is required for high-perf training.
- · Comp-comm overlap is challenging in the baseline system.
- ACE reduces comp-comm resource contention by handling the communication, enabling efficient comp-comm overlap.
- Please check out our paper and full talk for more details!