

Study Item Proposal: Network for AI Computing

Lily Lyu (Huawei)

Jan 2024

Background

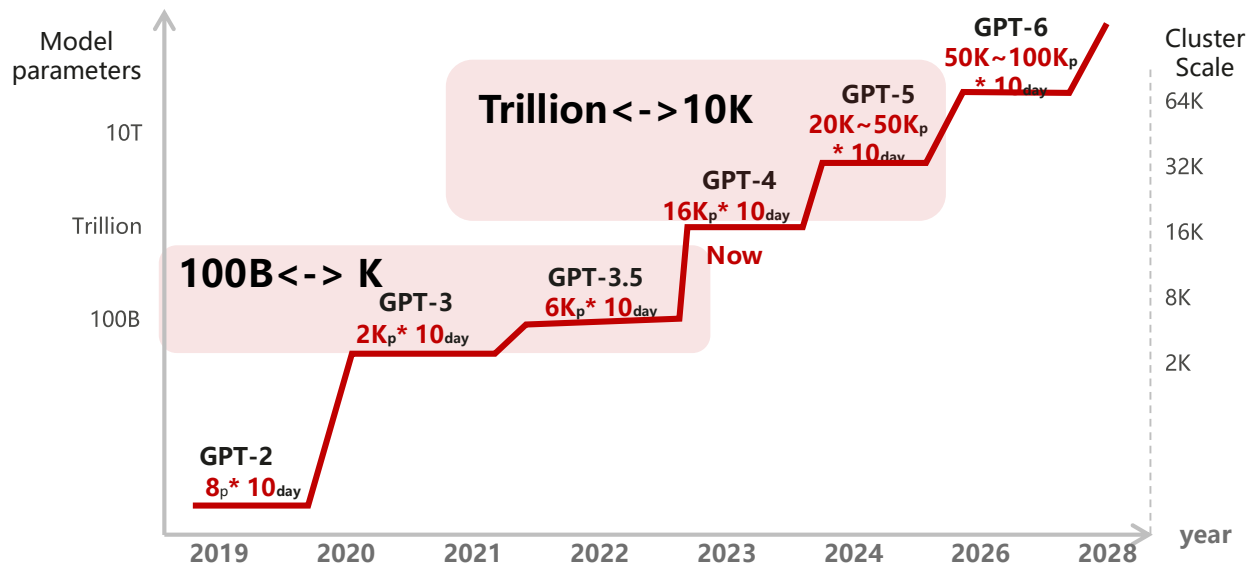
AI large model – new surge of AI computing

- AI large models show emergent abilities, attracting industry's attention.

Emergent abilities that are not present in smaller-scale models but are present in large-scale models, which are qualitative changes resulted by quantitative changes (training compute, number of model parameters and training dataset size)

--- Google&Stanford, 2022

- AI large models evolve very fast, requiring large scale network.



Network development

Industry activities:

UltraEthernet Consortium

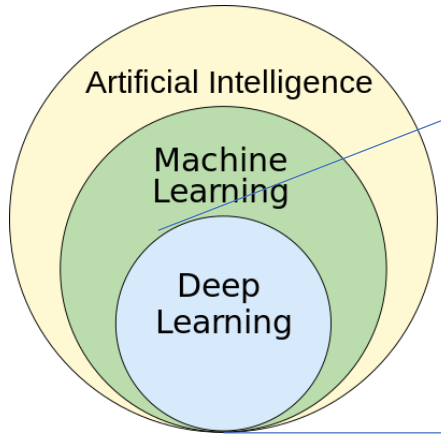
- UEC <https://ultraethernet.org/>
- IETF AI DC(datacenter) side meetings
<https://github.com/Yingzhen-ietf/AIDC-IETF117>
<https://github.com/Yingzhen-ietf/AIDC-IETF118>

Nendica contributions:

- Requirements for AI Fabric
- Congestion Signaling (CSIG)
- Network for AI datacenters
- Load balancing challenges in AI fabric

There's a lot of interest in network improvement in order to support AI large model.

Important to Know How AI Works



Deep Learning
 ≈ Looking for a Function

- Speech Recognition

$$f(\text{audio waveform}) = \text{"How are you"}$$

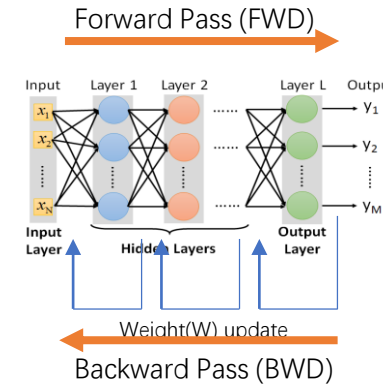
- Image Recognition

$$f(\text{cat image}) = \text{"Cat"}$$

- Dialogue System

$$f(\text{"Hi" (what the user said)}) = \text{"Hello" (system response)}$$

DNN-based Architecture for deep learning
 (DNN: Deep Neural Network)

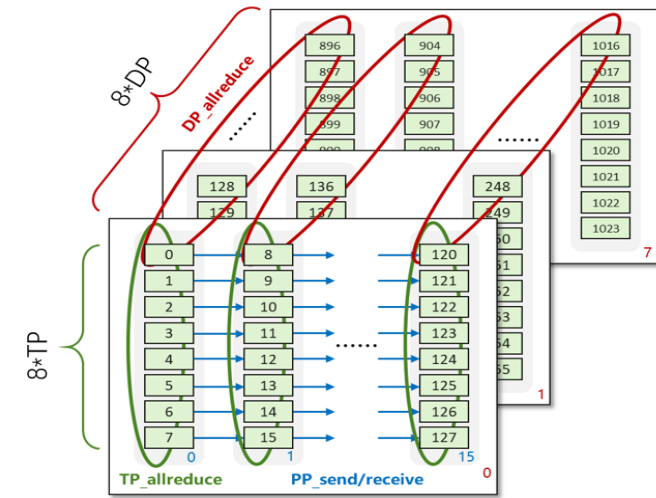


- ✓ Samples
- ✓ Parameters
- ✓ Gradients
- ✓

From Nendica contribution: "Network for AI datacenters"

Keys to AI Training:

- **Compute** (FLOPS, floating point operations per second) decides how fast to train a model.
 - Days trained * Number of GPUs * single GPU FLOPS ≈ (peta)FLOPS-day of model
- **Memory size** determines if the model can be trained.
 - Memory must be big enough to store model parameters and intermediate values generated during FWD and BWD.
 - Large model cannot fit into a single GPU memory, model parallelism has to be used.
- **Parallelism** enables model training.
 - Model parallelism and data parallelism



Example:

GPT3_175B, 1024 GPUs

16*PP

DP(data parallelism) = 8, TP(tensor parallelism)=8,

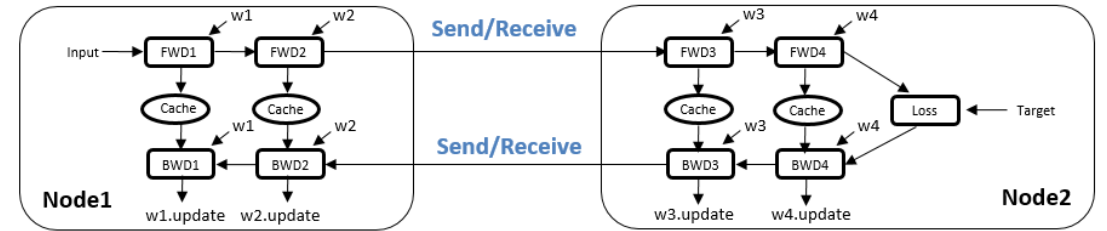
PP(pipeline parallelism)=16

Important to Understand Communication in AI (1/3)

Overlap communication and computation as much as possible to optimize training.

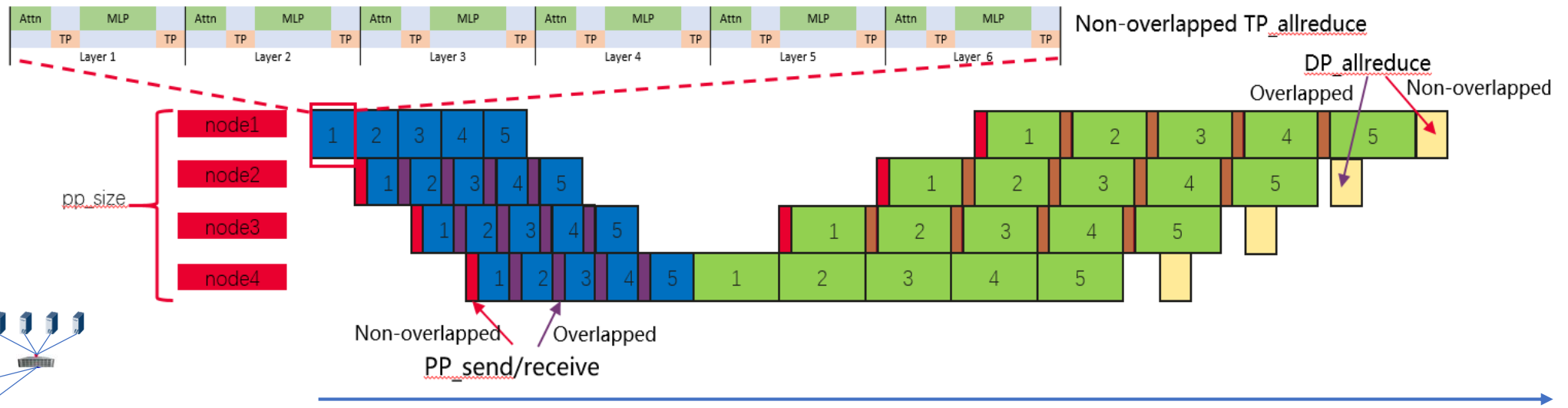
- TP Communication is hard to be overlapped with computation.
- PP Communication can be overlapped with computation.
- DP Communication can be overlapped with computation.

TP/PP/DP may have overlap.

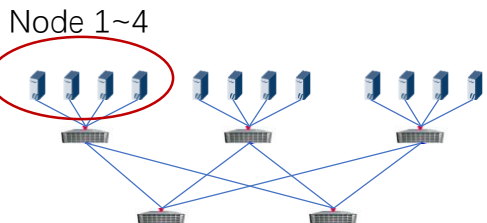


From Nendica contribution: "Network for AI datacenters"

Example:

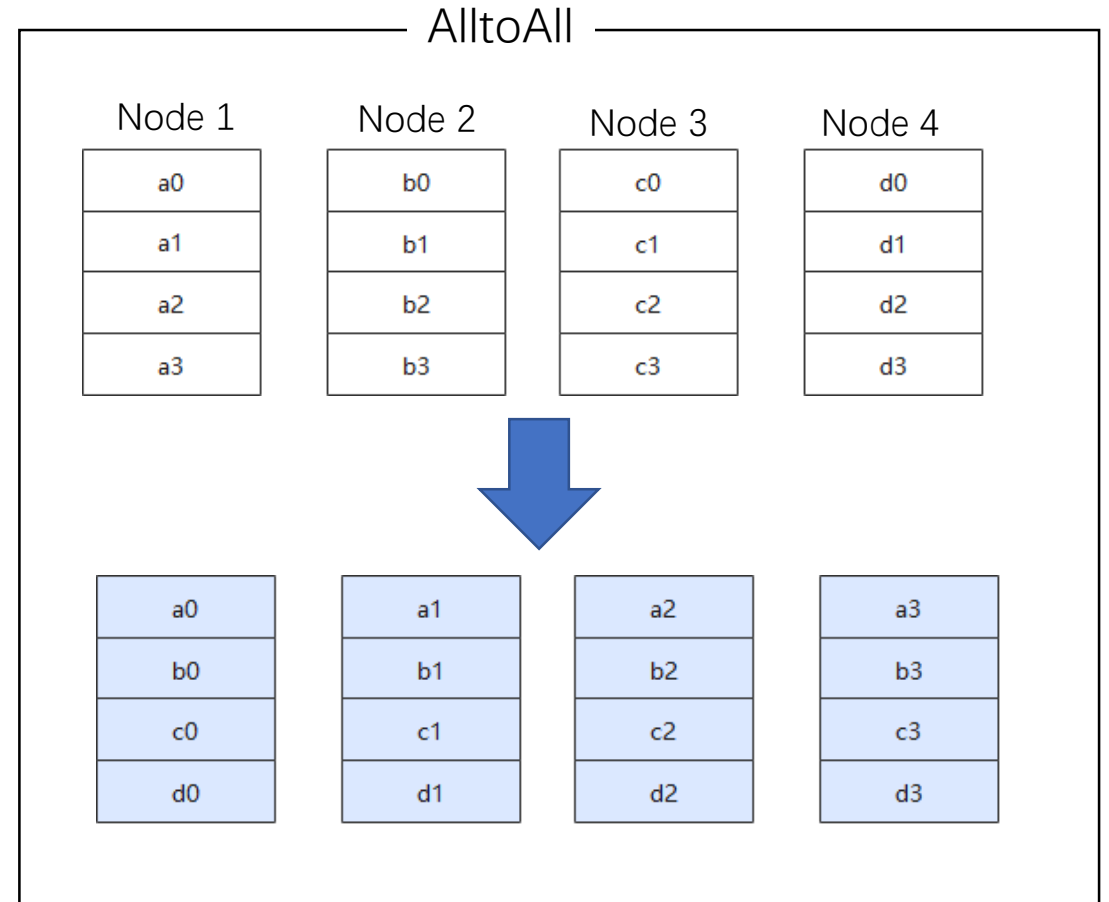
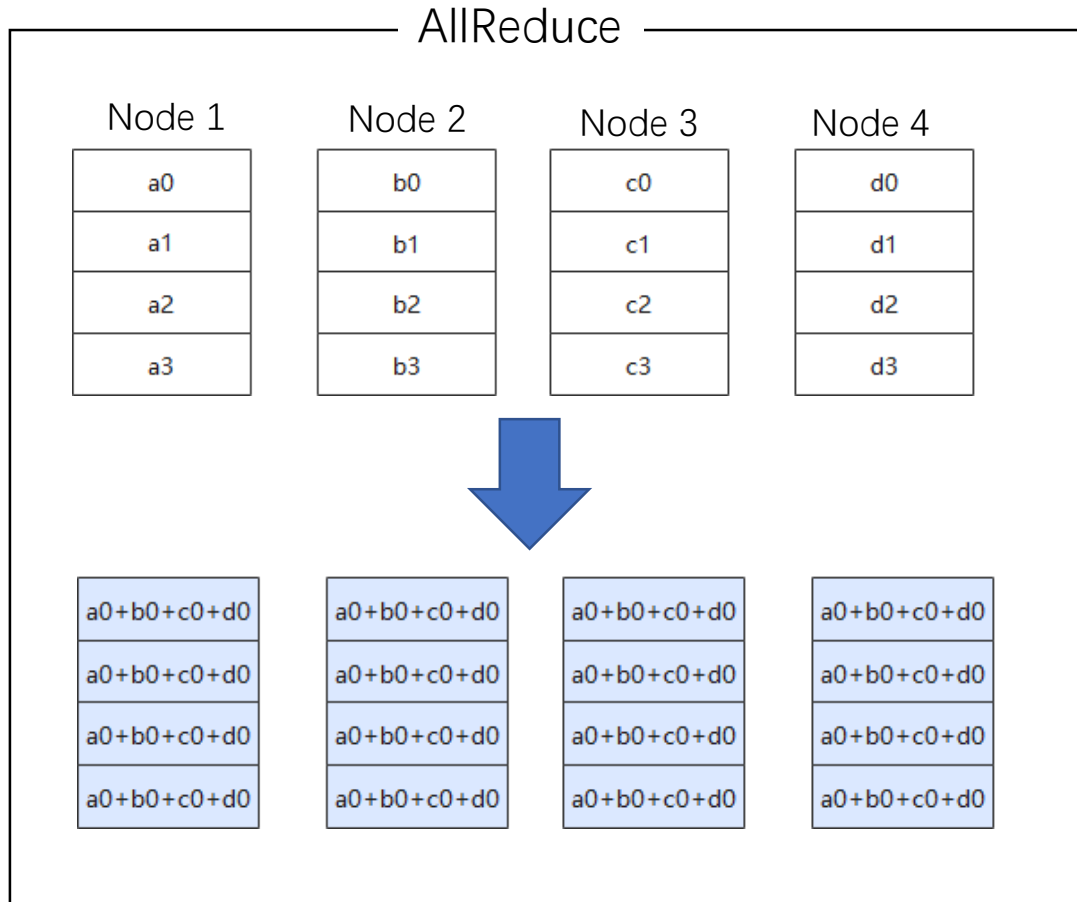


Attn: attention
MPL: multilayer perceptron



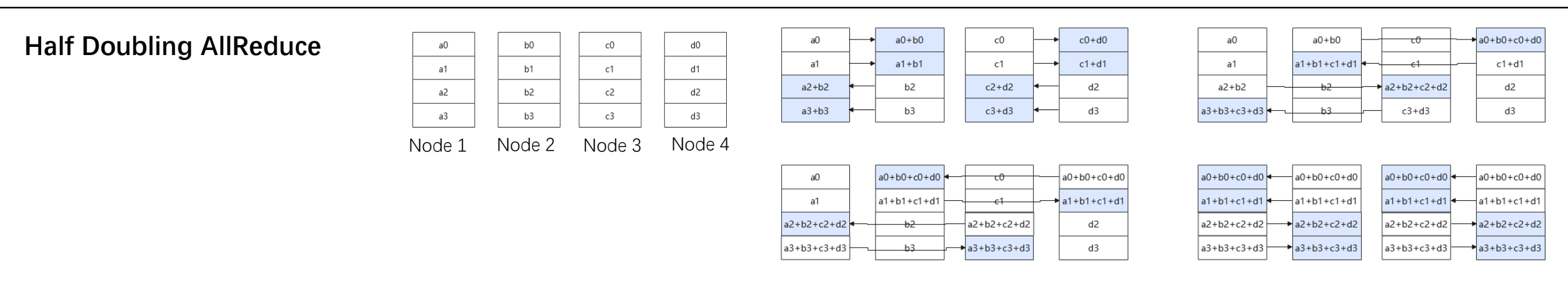
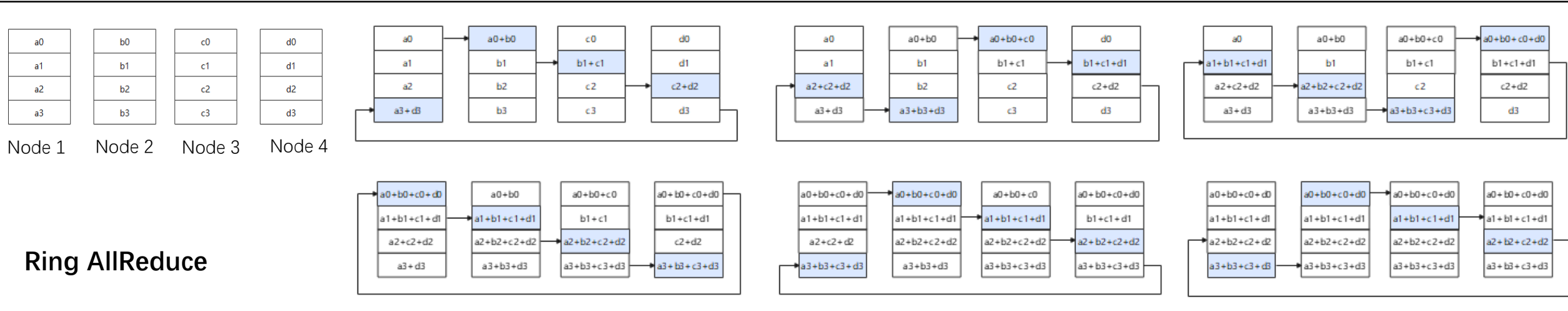
Important to Understand Communication in AI (2/3)

- AllReduce and AlltoAll are typical collective communication operations in AI training.



Important to Understand Communication in AI (3/3)

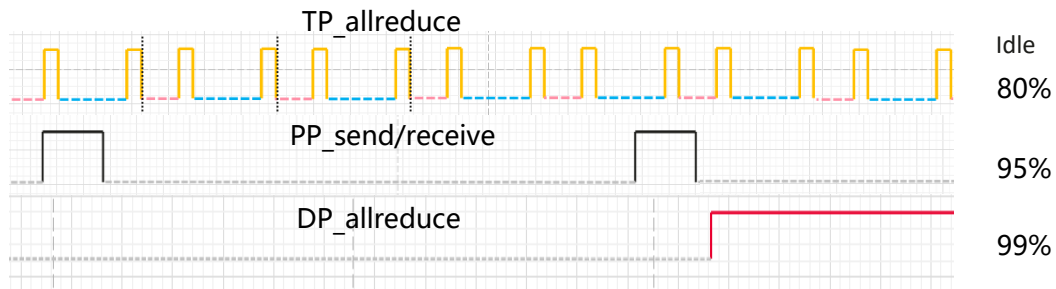
- Collective communication can have different implementations.
 - Needs comprehensive considerations (e.g. network topology, message size) to design proper implementation.



Need to Notice New Traffic Pattern

Sparse communication but requiring large bandwidth

- The distribution of traffic is regular in both space and time dimensions.
 - The flow of traffic is regular.
 - Communication pair is predictable.
 - Maximum number of connections on a GPU is $TP-1+DP-1+1$ (TP/DP/PP)
 - TP/DP/PP logical planes show periodic bursts of traffic.
 - The burst frequency : $TP > PP > DP$
 - Link is idle in most of time.



- Single GPU requires large bandwidth for traffic communication

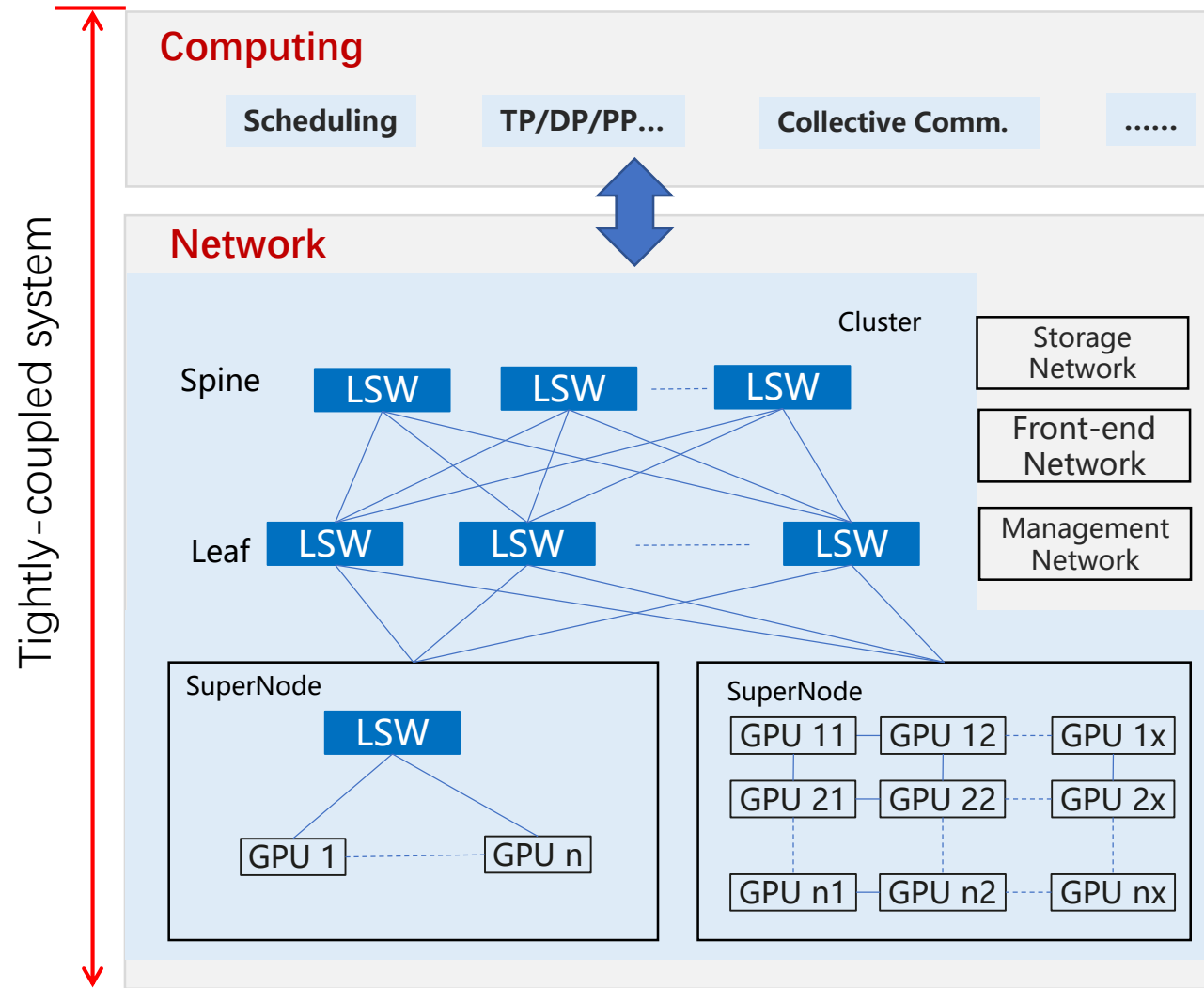
Parallel Mode	Communication (1 GPU 1 time)
TP	100s GB level
PP	100s MB level
DP	GB level

- E.g. Meta uses 200Gbps per GPU(A100) in its LLM models.

MODEL NAME	RELEASE DATE	MODEL SIZE	DATASET SIZE	TRAINING ZETA (1E21) FLOPS	TRAINING HW (COMPUTE)	TRAINING HW (NETWORK)	GPU HOURS (# GPU X HOURS)
OPT	May 2022	175 B	300 B	430	1K A100	IB 200Gbps per GPU 25.6 TB/s bisection BW	800K
LLaMA	Feb 2023	65 B	1.4 T	600	2K A100	IB 200Gbps per GPU 51.2 TB/s bisection BW	1M
LLaMA2	July 2023	34 B	2 T	400	2K A100	RoCE 200Gbps per GPU 51.2 TB/s bisection BW	1M
LLaMA2	July 2023	70 B	2 T	800	2K A100	IB 200Gbps per GPU 51.2 TB/s bisection BW	1.7M

From <https://www.nextplatform.com/2023/09/26/meta-platforms-is-determined-to-make-ethernet-work-for-ai/>

Systematic View On AI Computing Network (1/2)



Systematic View On AI Computing Network (2/2)

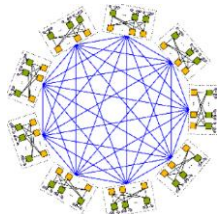
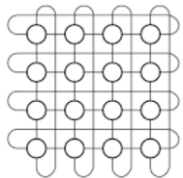
$$\text{Total compute} = \text{single GPU compute} * \text{Scale} * \text{Efficiency} * \text{Availability}$$

Challenge:

- Interconnection of large number of GPUs (K->10K->100K)

Consideration:

- Topology optimization for super-node and cluster network
 - Direct topology, e.g. torus, dragonfly+

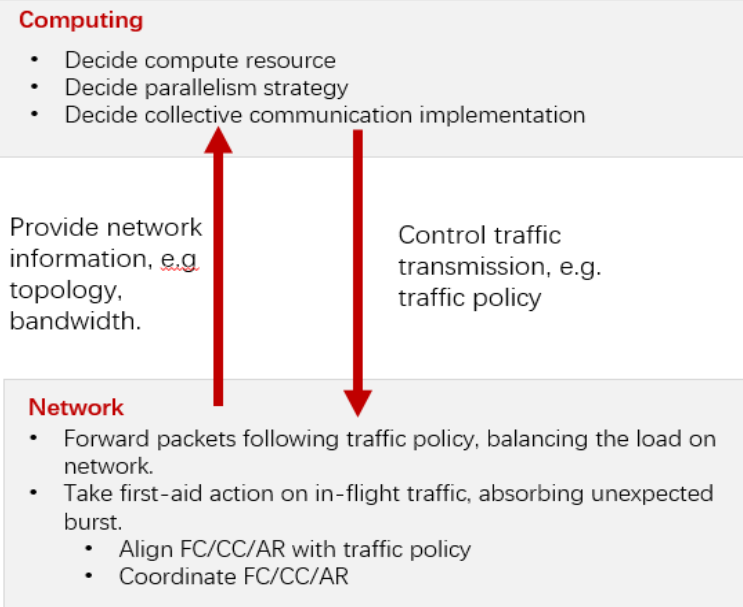


Challenge:

- Communication costs hinder linear expansion of computing power

Consideration:

- Coordination between computing and network.



Challenge:

- Components in large scale system frequently fail.

Consideration:

- Combination of hot swap, automatic path migration, and checkpointing
- Backtracking to the last checkpoint has a high penalty
- Avoid it whenever possible with APM plus load balancing, followed by retransmission of lost packets
- Combine with AR for immediate response after failure detection

Quote from Nendica contribution: "Network for AI datacenters"

Study Item Proposal

Study item: AI computing Network

Purpose:

- Understand the requirement of network for AI computing.
- Look for potential standardization opportunity in IEEE802.

Scope:

- Study main factors (parallelism, collective communication) in AI training which impact traffic.
- Analyze the major challenges for the network.
- Investigate future network technologies.
- Identify potential standard work.

Deliverables:

- Informal report documenting, including
 - AI computing network requirements and challenges
 - Potential technologies
 - Possible standardization needs
 - Work item proposal

Schedule:

- Start in Jan 2024
- Propose work item in March 2024

Leader:

Lily Lyu (Huawei)

Supporters:

José Duato (Royal Spanish Academy of Sciences)

Liang Guo (China Academy of Informational and Communication Technology)

Jesús Escudero (UCLM)

Thank You!