Study Item Proposal: Network for AI Computing

Lily Lyu (Huawei)

Feb 2024

Background

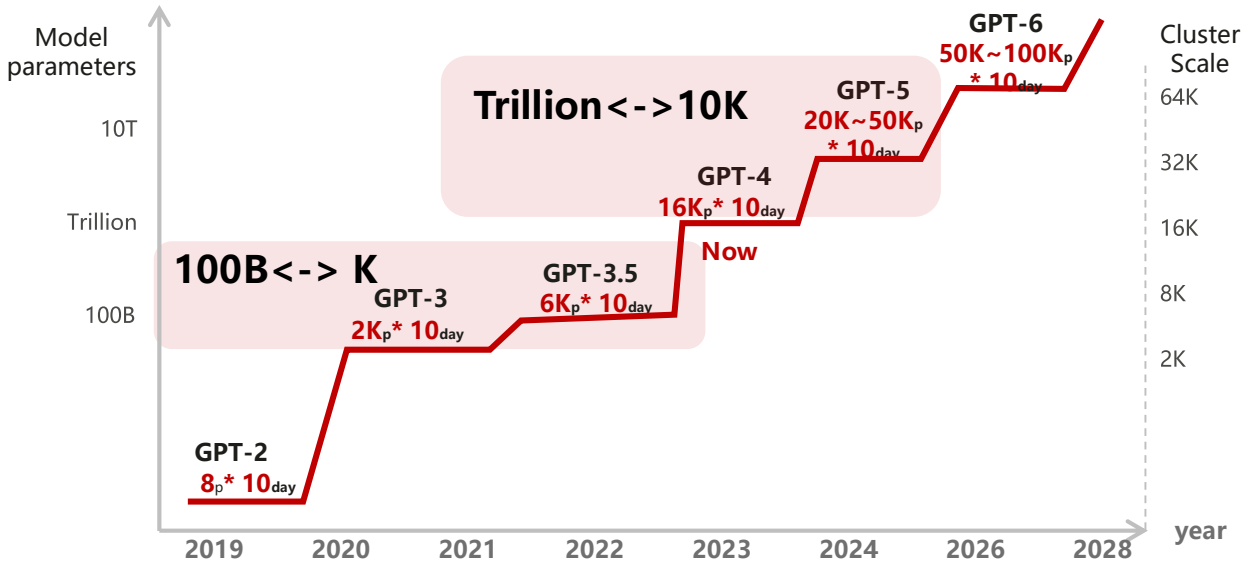
AI large model – new surge of AI computing

- 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&Standford, 2022

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



Network development

Industry activities:

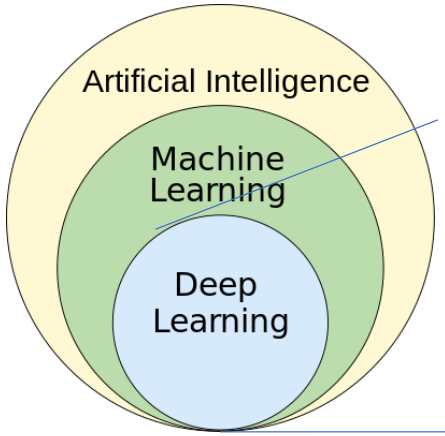
- UEC https://ultraethernet.org/
- IETF AI DC(datacenter) side meetings
  https://github.com/Yingzhen-ietf/AIDC-IETF117
  https://github.com/Yingzhen-ietf/AIDC-IETF118

Nendica contributions:

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

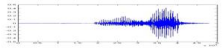
There's a lot of interest in network improvement in order to support AI large model.

Important to Know How AI Works



Deep Learning ≈ Looking for a Function

- Speech Recognition

$f($  $)=$ "How are you"

- Image Recognition

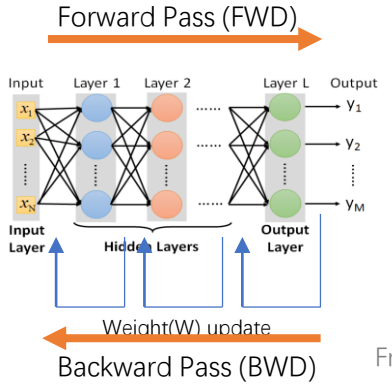
$f($  $)=$ "Cat"

- Dialogue System

$f($ "Hi" $)=$ "Hello"

(what the user said) (system response)

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

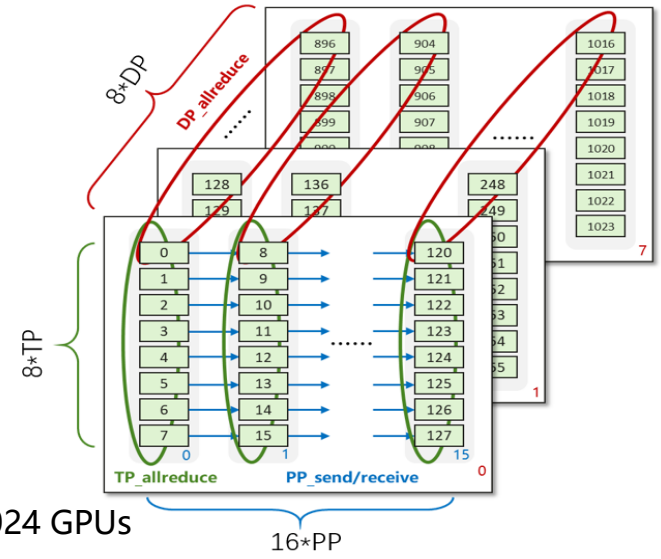


- ✓ Samples
- ✓ Parameters
- ✓ Gradients
- ✓ ……

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

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

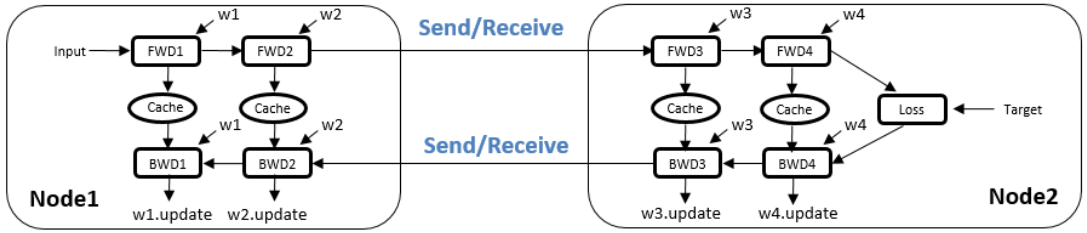
PP(pipeline parallelism)=16

Important to Understand Communication in AI (1/3)

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

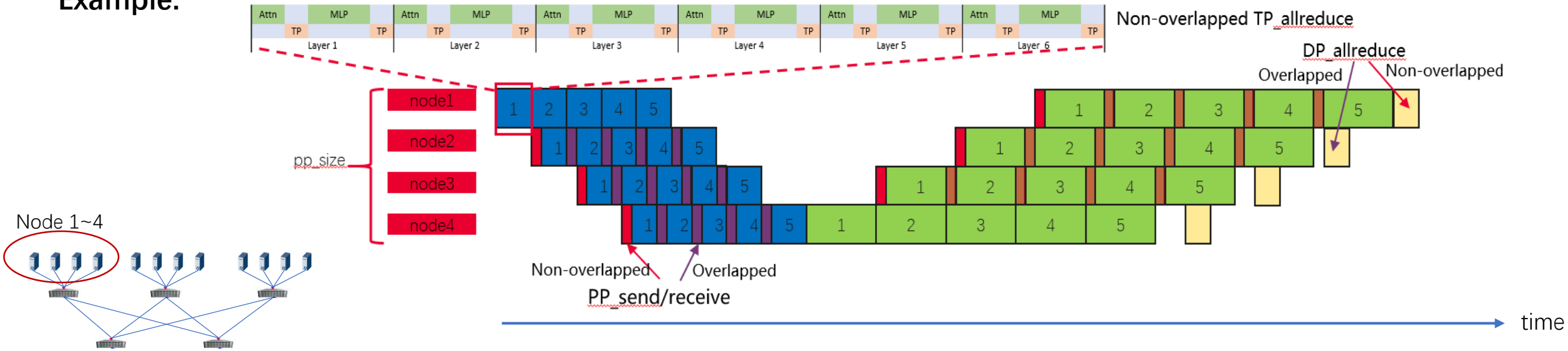
- TP Communication is hard to be overlapped with computation.
- PP Communication can be overlapped with computation.
- DP Communication can be overlapped with computation.

TP/PP/DP may have overlap.



From Nendica contribution: "Network for AI datacenters"

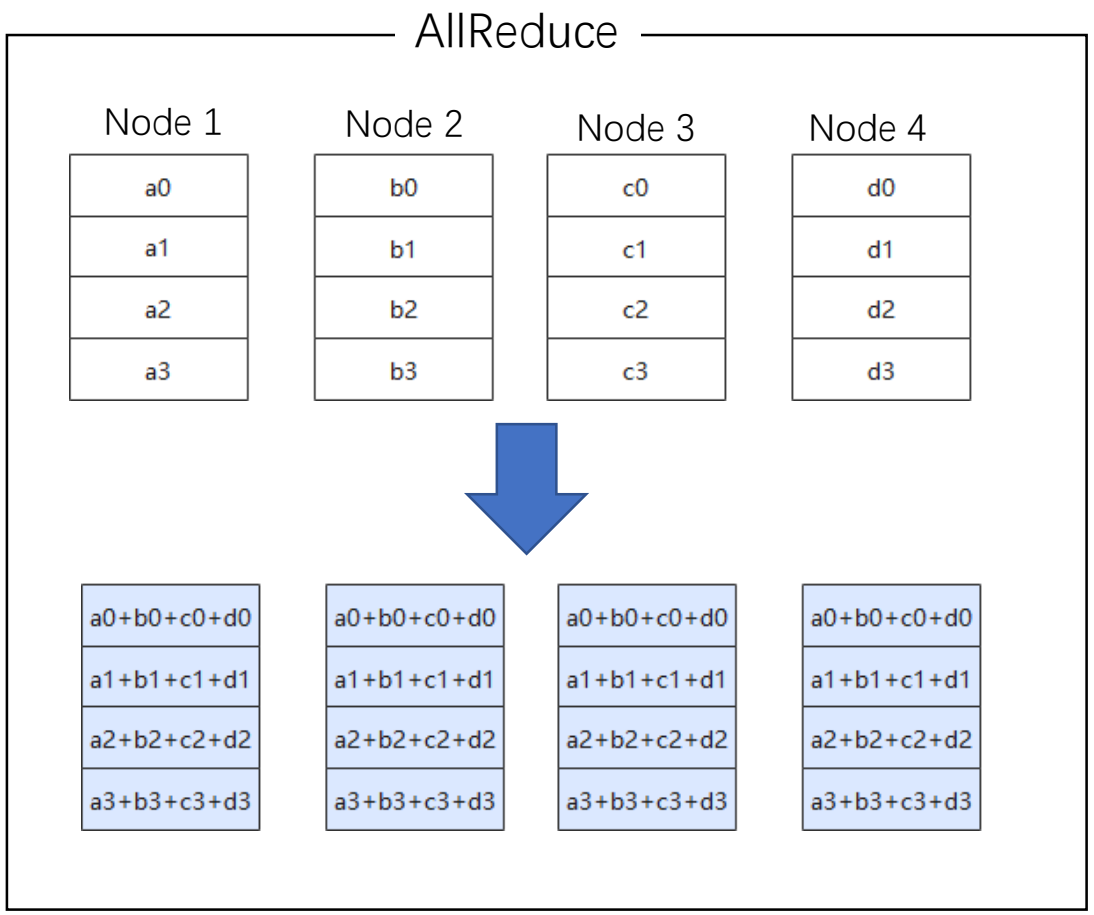
Example:

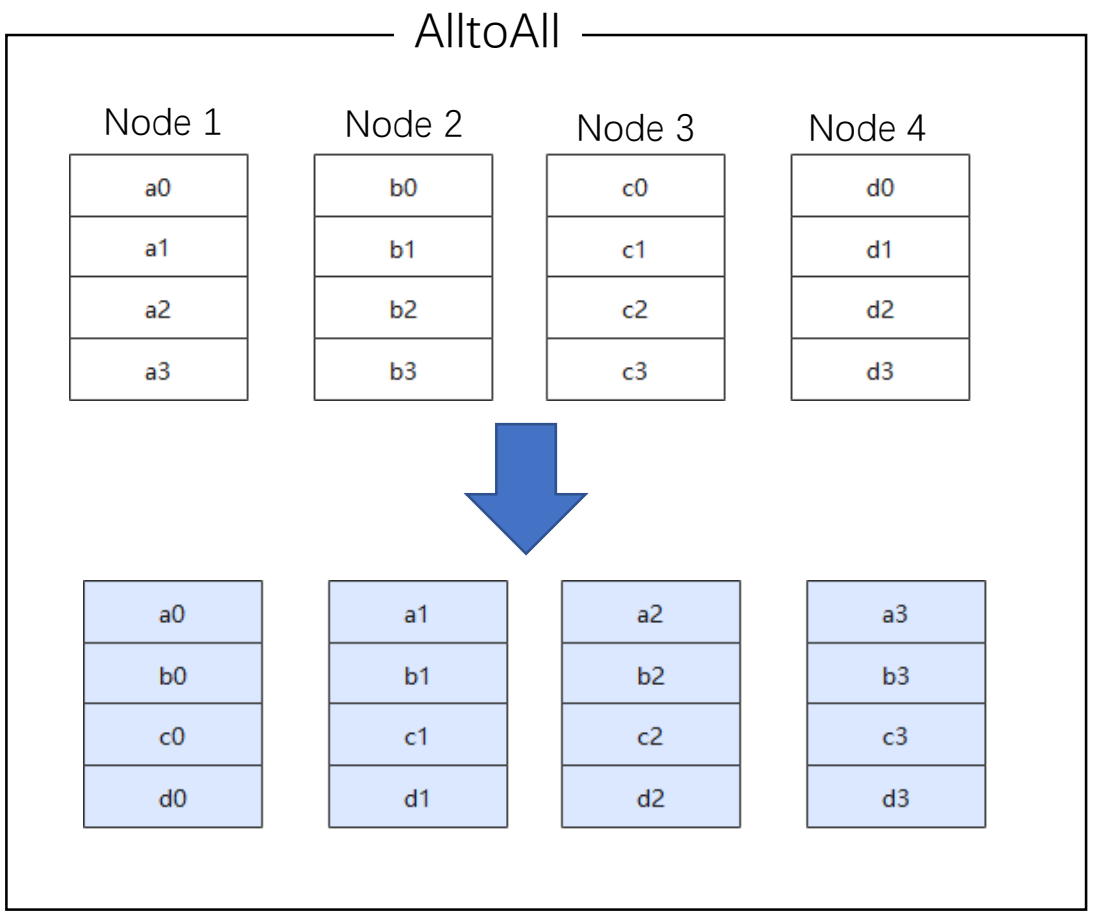


Attn: attention
MPL: multilayer perceptron

Important to Understand Communication in AI (2/3)

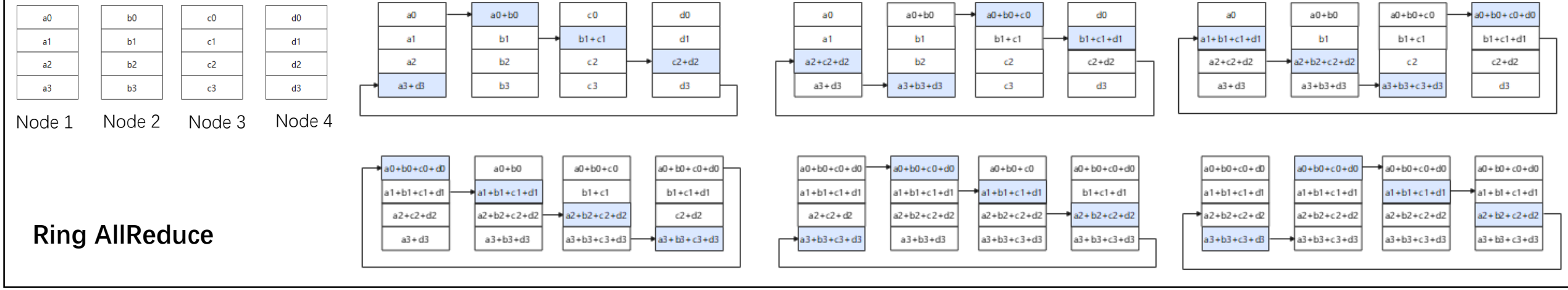
- **AllReduce and AlltoAll are typical collective communication operations in AI training.**

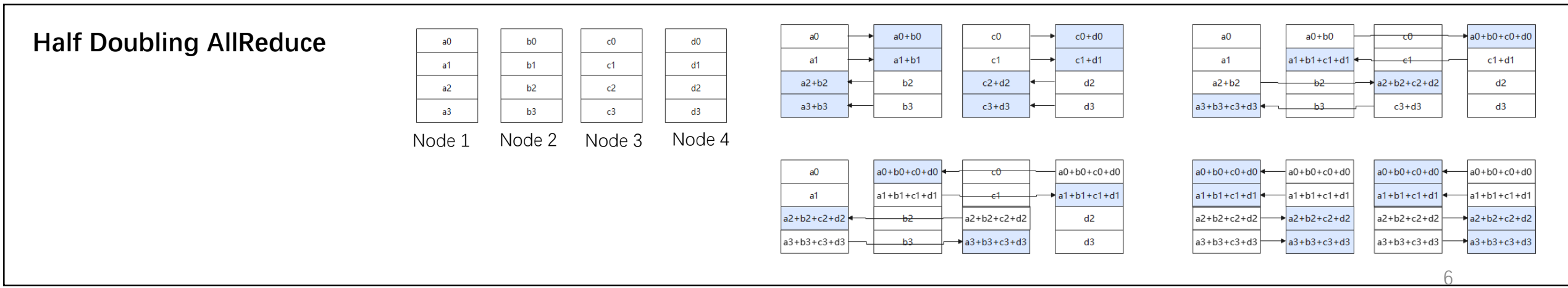




Important to Understand Communication in AI (3/3)

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

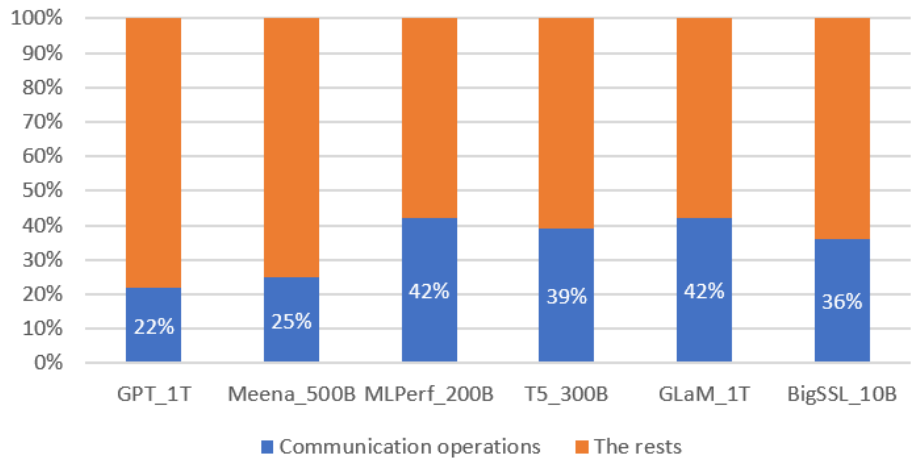




Analysis on Communication Time

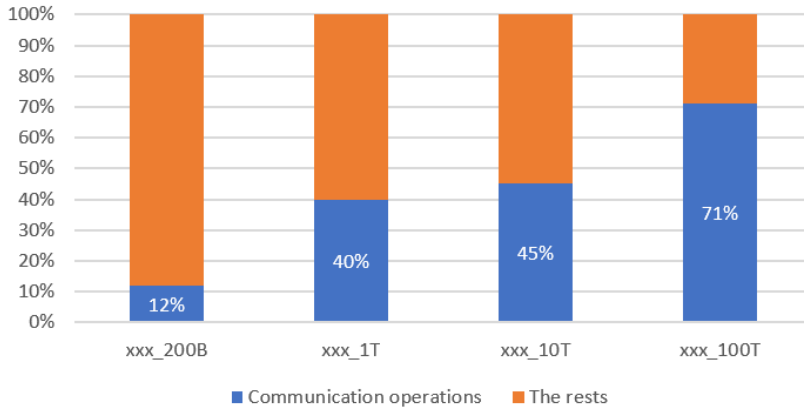
Communication consumes a non-negligible proportion in the training time, and the situation gets worse when AI model size increases (more GPUs).

Substantial percentage of the training time on communication in different models.



Data from ASPLOS '23 paper "Overlap Communication with Dependent Computation via Decomposition in Large Deep Learning Models"

Larger AI model, higher proportion of communication time

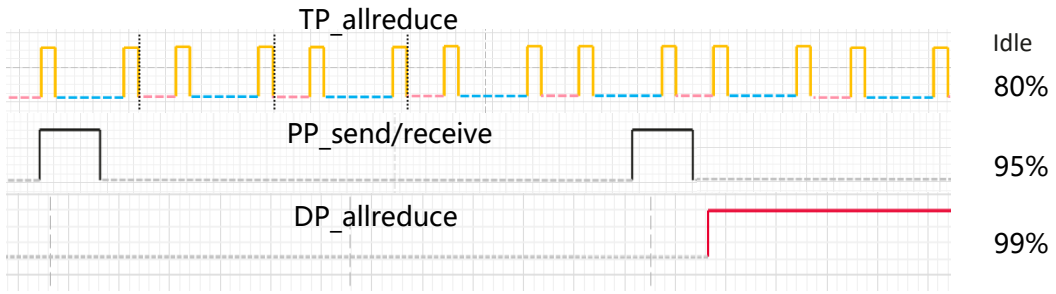


Calculated communication time percentage when training AI models of different model sizes using the same strategy

Need to Notice New Traffic Pattern

Sparse communication but requiring large bandwidth

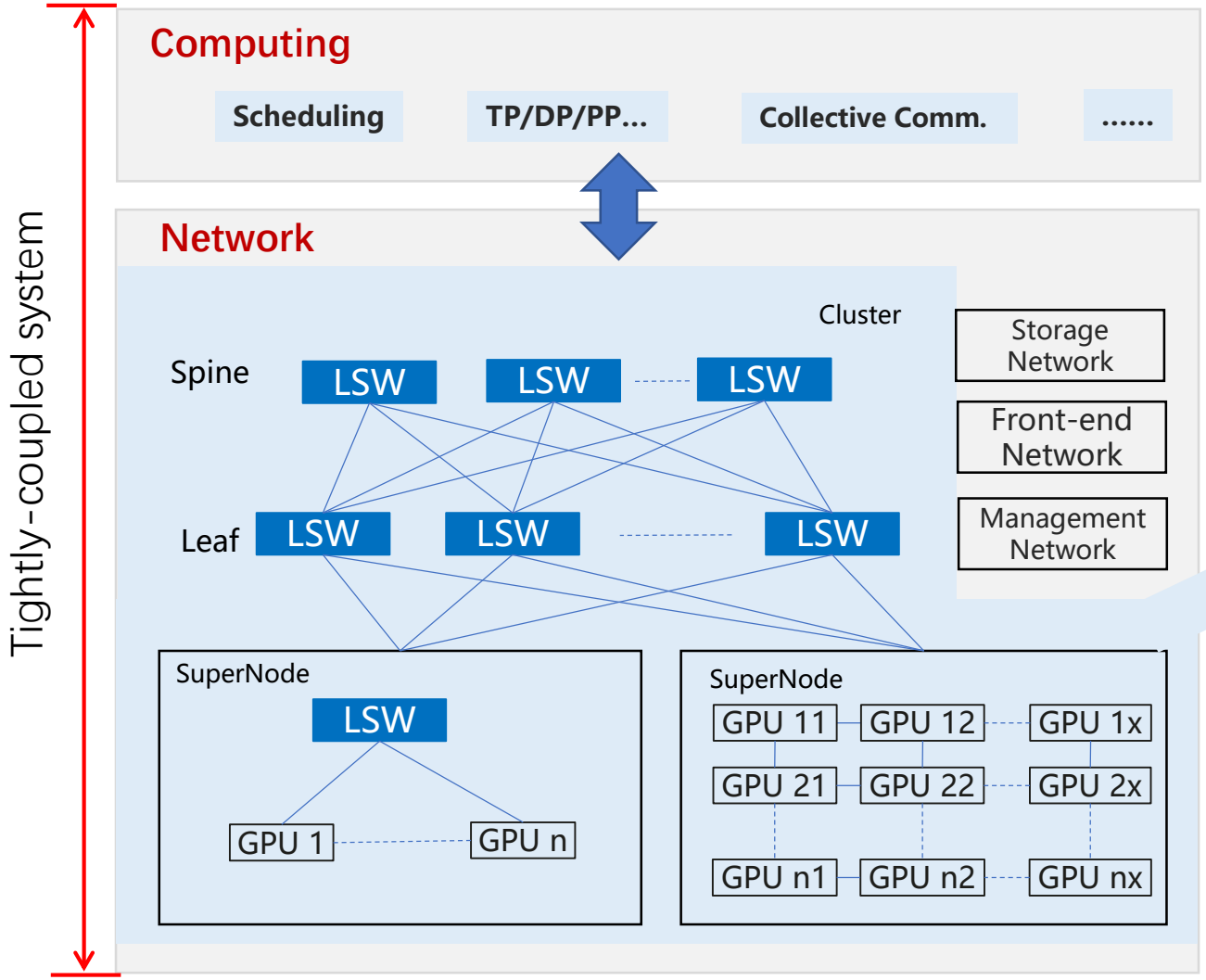
- The distribution of traffic is regular in both space and time dimensions.
  - The flow of traffic is regular.
    - Communication pair is predictable.
    - Maximum number of connections on a GPU is TP-1+DP-1+1 (TP/DP/PP)
  - TP/DP/PP logical planes show periodic bursts of traffic.
    - The burst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

Systematic View On AI Computing Network (1/2)



The uniqueness of AI computing network

- ✓ **Predictable traffic**
- ✓ **Large amount of traffic for each burst**

LSW: LAN switch

Systematic View On AI Computing Network (2/2)

Total compute = single GPU compute * Scale * Efficiency * 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
  - Combination of different topologies, e.g. clos+torus

Challenge:

- Communication costs hinder linear expansion of computing power

Consideration:

- Collaboration between computing and networking.
  - Computing: -- 'static' planning
    - Pre-plan/update traffic strategy based on network information
  - Networking: -- 'dynamic' adjustment
    - Follow traffic strategy, maximize network resource to handle in-flight traffic

Challenge:

- Components in large scale system frequently fail.

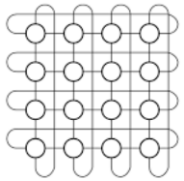
Consideration:

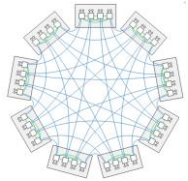
- Combination of hot swap, automatic path migration , and checkpointing
- Backtracking to the last checkpoint has a high penalty
- Avoid it whenever possible with APM plus load balancing, followed by retransmission of lost packets
- Combine with AR for immediate response after failure detection

Quote from Nendica contribution: "Network for AI datacenters"

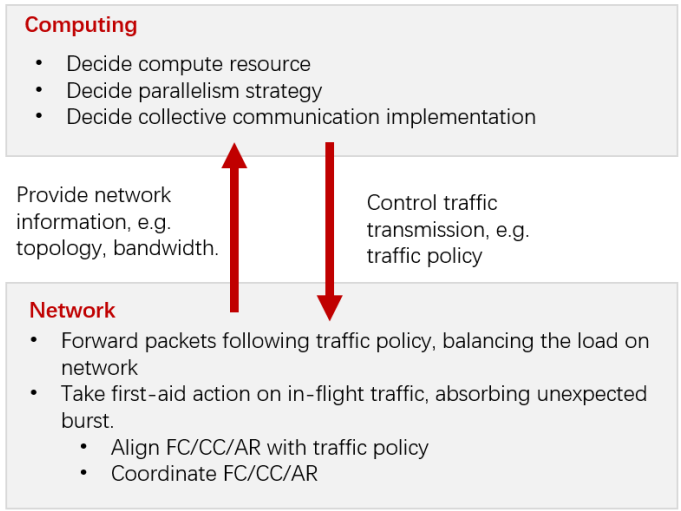
Potential Technologies and Standardization Considerations

Topology optimization





Computing and networking collaboration



Network reliability

Potential technologies
(Underline marked technologies may involve standard work in IEEE802)

- Routing protocol for direct topologies
- PFC deadlock prevention
- QoS optimization
  - Collaboratively configure FC, CC and Transmission selection
- CC/AR coordination
- Load balancing
  - Packet based load balancing
    - Load-aware packet spray
    - Path-aware packet re-ordering
- Data plane fast failure recovery
- Link layer retransmission

Basic capability to support the technologies

- Topology recognition (LLDP)
- 'Path associated signaling'
  - Hop by hop update signal, such as L2 telemetry
  - Fixed indication signal, such as path ID
- Fast feedback of link/port/queue status
  - Hop by hop notification
  - Remote notification

FC: flow control
CC: congestion control
AR: adaptive routing

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 Feb 2024
- Propose work item in July 2024

Leader:

Lily Lyu (Huawei)

Supporters:

José Duato (Royal Spanish Academy of Sciences)
Liang Guo (CAICT)
Jesús Escudero (UCLM)

Motion Discussion

There was discussion on the study item name “computing network” in interim meeting.

To initiate a Nendica study item on computing network

Proposed: Lily Lyu
Second: Nader Zein

Proposed new text for motion:

Option1：To initiate a Nendica study item on AI computing network

Option2：To initiate a Nendica study item on computing network for AI Large Model

Thank You!