AI Computing Network

Lily Lyu (Huawei)
Jieyu Li (China Mobile)
Jose Duato (RSAC)
Jesus Escudero-Sahuquillo (UCLM)

March Plenary 2024

Recap

There are 2 proposals of study items on ‘AI’.

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)

Load balancing challenges in AI fabric

Weiqiang Cheng (China Mobile)
Dong Ye (Intel)
Yadong Liu (Tencent)
Jieyu Li (China Mobile)
Ruixue Wang (China Mobile)

Next Action

- **Propose a Study/Work Item：** Packet-Spray-Based AI Fabric
- **Scope：** Packet-Spray-Based Load Balance 、Packet-Spray-Based Congestion Control、Reorder、Ethernet QoS、Telemetry、high-precision OAM、Protection etc.

There are another 3 contributions talking about AI datacenter requirements. (1-23-0031-01-Icne(Weiqiang Cheng, Ruixue Wang)**，1-24-0001-00-Icne**(Jose Duato), **1-24-0009-02-Icne**(Jesus Escudero-Sahuquillo, Jose Duato)**)**

After discussion, we agree to make this joint proposal on AI computing network study item.

Motivation

To support AI large models, large scale and high performance networking is required.

Ethernet networking as the rich eco-system technology has opportunities to support AI clusters. However, it needs to be evolved in order to meet the requirements of AI computing network.

How does IEEE802 networking fit for AI cluster?

Start from study item, and aim to initialize a work item in order to output NENDICA report "AI Computing Network (AICN)":

- Analyzing network challenges for AI clusters
- Pointing out AI computing network technology trends
- Identifying IEEE802 standard gaps and opportunities

Contents of AICN Report

- Background/Use cases
- Requirements & Challenges of AICN
- Potential technologies
- Standard considerations

Background/Use Cases

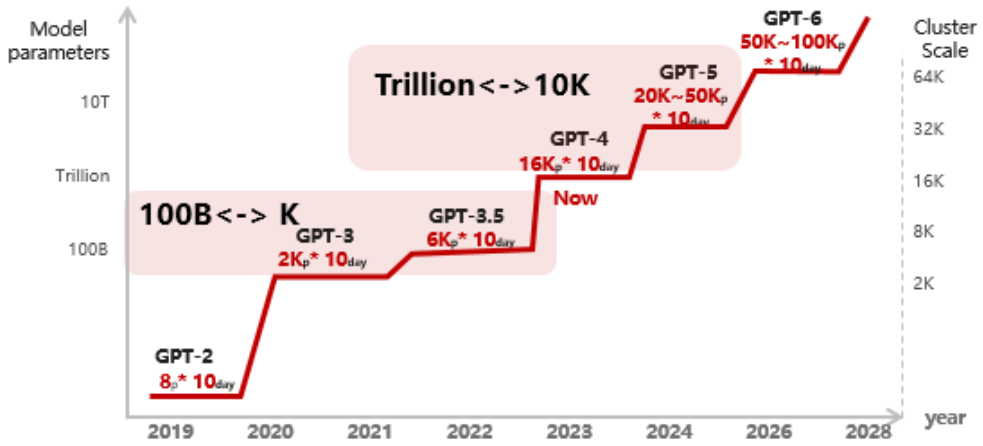
AI large model – new surge of AI computing

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

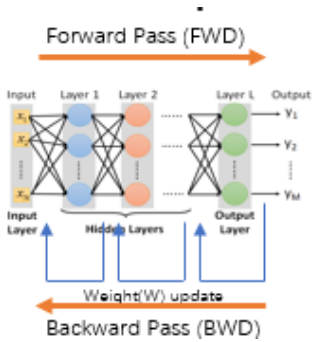
Emergent abilities that are not present in smaller-scale models but are present in large-scale models, which are qualitative changes resulted by quantitative changes (training compute, number of model parameters and training dataset size)

--- Google&Standford, 2022

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



"The last decade has witnessed a very rapid expansion of many DNN-based AI solutions."



- ✓ Samples
- ✓ Parameters
- ✓ Gradients
- ✓ .....

"The release of ChatGPT in Nov 2022 has garnered unprecedented attention, and triggered the recent boom of large language models (LLMs)"

From Nendica contribution: "Network for AI datacenters"

Requirements & Challenges of AICN

Total compute = single GPU compute * Scale * Efficiency * Availability

- **Scale:** "Expected size is on the order of 200K+ servers"

Network challenge：

Large scale CLOS/FATTREE topology has cost and performance issues.

- Cost issue: require many switches and more hops to connect tens of thousands of nodes
- Performance issue: increase chance of congestion and long tail latencies

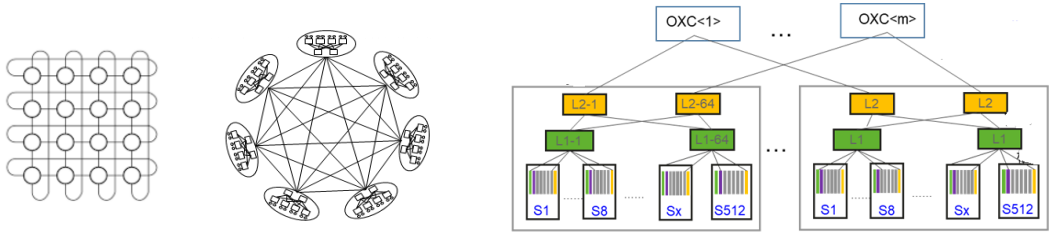
	Scale (K=64)	Scale (K=128)
2 layer CLOS	2048	8192
3 layer CLOS/Fattree	65536	524288

K: radix of switch

Potential Technologies & IEEE802 Considerations

New technologies

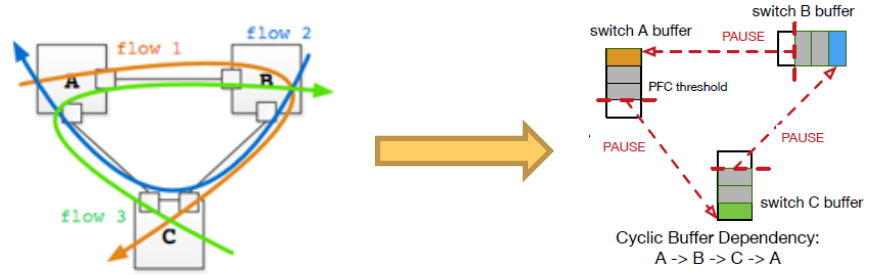
- Direct topology, e.g. torus/dragonfly(+)
- Changeable topology, e.g. OXC(optical cross connect)



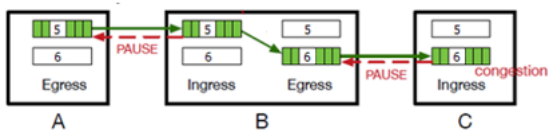
IEEE802 standard considerations

- PFC deadlock prevention

How does PFC deadlock happen?



Principle of preventing PFC deadlock

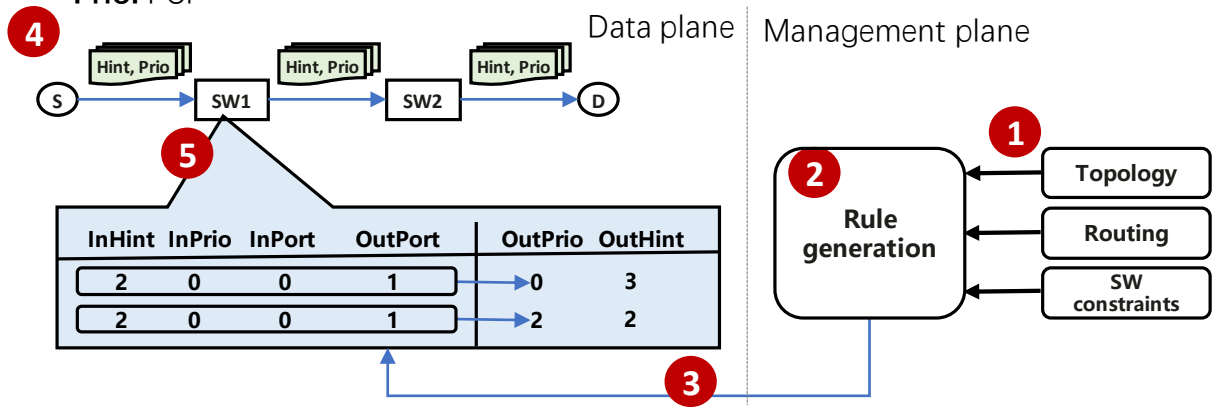


Break the loop by switching priority queue

1) Based on topology and routing, construct a dependency graph and find out all the loops in it. Each port of switches is a point in the dependency graph.
2) Divide the dependency graph into multiple DAGs (Directed Acyclic Graph), and make sure there's no loops within and between DAGs.
3) Set up the rules of priority queue switching.

5 steps to implement PFC deadlock prevention

Hint: new added info in L2 header, implying DAG and potential default routing change(e.g. due to port fault)
Prio: PCP

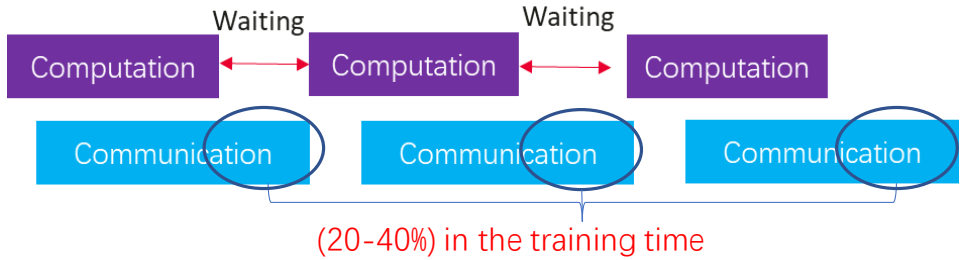


Requirements & Challenges of AICN

Total compute = single GPU compute * Scale * Efficiency * Availability

- **Efficiency:** increase GPU utilization

- Due to memory and interconnect bottlenecks, currently, GPU utilization when training large AI models is around 30-40%.
- Communication consumes a non-negligible proportion (20-40%) in the training time, and the situation gets worse when AI model size increases (more GPUs).



Reducing communication time and overhead reduces GPU waiting time, thereby increasing computation efficiency. (1-23-0031-01-ICne)

Network challenge:

Big pressure on bandwidth.

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

- Network congestion
  - ECMP is invalid for AI traffic load balancing.
  - Uncoordinated congestion management schemes deteriorate performance.
- Decoupled computation and network
  - Un-optimized traffic injection
  - Application-agnostic network QoS

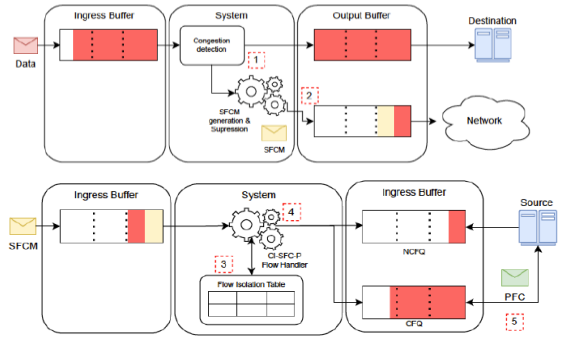
Potential technologies & IEEE802 Considerations

New technologies

- Congestion management coordination (1-24-0009-02-ICne)
  - LB/AR/CC cooperation

Solution: Multi-path routing combined with CC that distinguishes between in-network and incast congestion

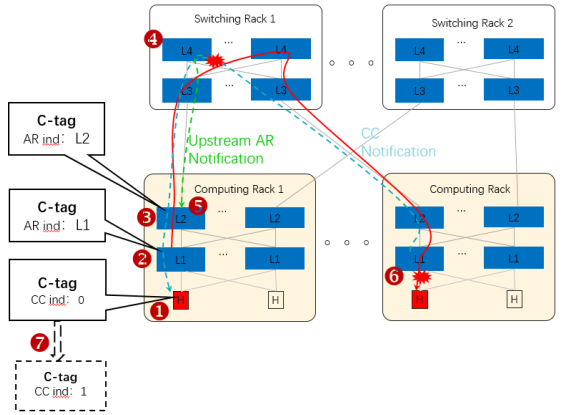
- Combination of SFC, CI, and DCQCN



- Computation and network coordination
  - Network status based collective communication
  - Collaborative configurations of FC, CC and Transmission selection

IEEE802 standard considerations

- In-cast and in-network congestion differentiation mechanism



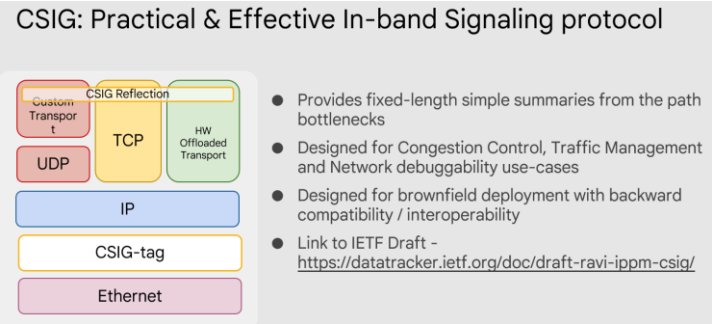
  - AR/CC notifications
  - Packets distinctions on switches
    - switches use adaptive routing to alleviate in-network congestion or deterministic routing for in-cast congestion

- Document how to better use the congestion management mechanisms in 802.1 standard

Potential technologies & IEEE802 Considerations

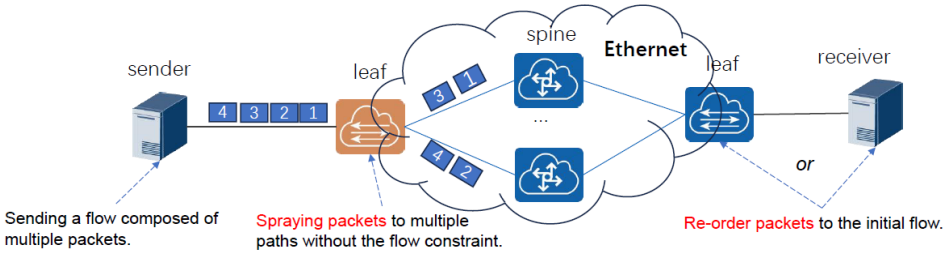
New technologies

- CSIG (1-23-0034-01-Icne)



Detailed in : https://datatracker.ietf.org/doc/draft-ravi-ippm-csig/

- Packet-based load balancing (1-24-0004-05-Icne)



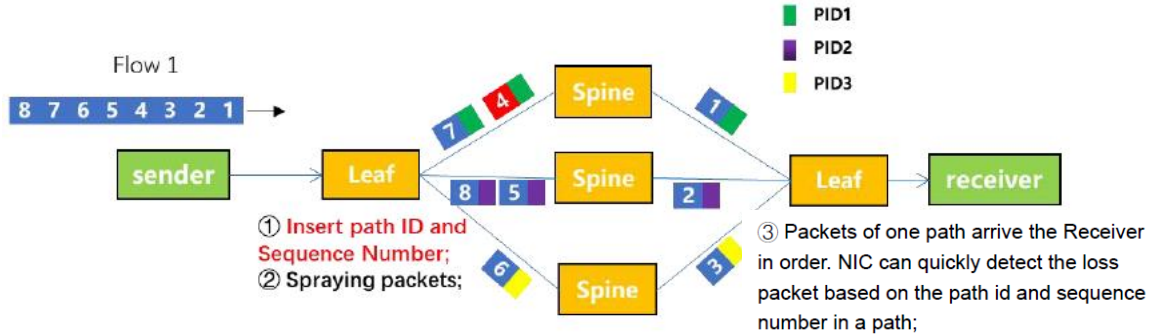
IEEE802 standard considerations

- L2 telemetry

ARPA	dstmac / srcmac / **csig-tag** / ethertype / payload
802.1q:	dstmac / srcmac / vlan-tag / **csig-tag** / ethertype / payload
802.1ad	dstmac / srcmac / vlan-tag / vlan-tag / **csig-tag** / ethertype / payload
802.1ad tunnel	dstmac / srcmac / vlan-tag / vlan-tag / vlan-tag / vlan-tag / **csig-tag** / ethertype / payload
802.1ae	dstmac / srcmac / security-tag / vlan-tag / **csig-tag** / ethertype / payload

Insert telemetry field in L2 header, carrying signal type, signal value and location of bottleneck point.

- Packet loss detection mechanism.



Insert path information of packets forwarding to help receiver detecting the loss more quickly then booting fast retransmission.

Requirements & Challenges of AICN

Total compute = single GPU compute * Scale * Efficiency * Availability

- **Availability**: failure recovery within sub-millisecond.

  The features of AI training place higher demands on network availability:

  - Large-scale: 1000 GPU cards, the probability of a failure within a month is 60%; Further, AI network involves nearly 100,000 optical modules, one optical module failure occurs every 4 days on average.
  - High bandwidth and long training duration: Any network interruptions or failures can lead to training interruptions, requiring the process to backtrack to the last checkpoint and wasting a large amount of time and resources.

Network Challenge：

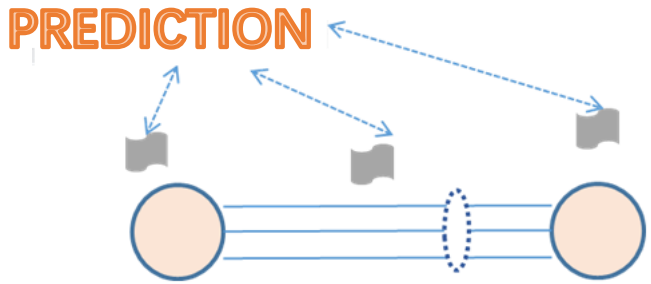
Detection mechanism	Technology	Convergence Time	Influencing factors
Fast Failure Detection	BFD	a few milliseconds	/
	CFM	milliseconds to seconds	/
Local Fast Failover	ECMP	a few milliseconds	convergence time primarily depends on the fault detection time.
	FRR	a few milliseconds	
Failure Notification	IGP LinkState propagation	milliseconds to seconds	The notification time depends on the network size and the number of routes.
	BGP route updates	milliseconds to seconds	
Global Fast Failover	BGP PIC	milliseconds to seconds	The convergence time depends on both the fault notification time and the network size.
	IGP route calculation convergence	several hundred milliseconds to a few seconds	

- Failure Detection
  - The current failure detection like BFD (Bidirectional Forwarding Detection), requires at least several tens of millisecond.
- Failure Notification
  - The current failure notification mechanism in DC mainly rely on the control plane protocol like IGP or BGP, and its convergence time is in the millisecond to seconds range.

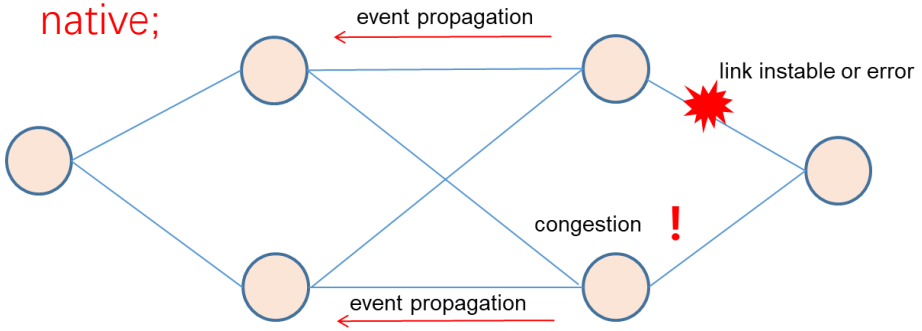
Potential technologies & IEEE802 Considerations

New technologies

- Fast link failure detection
  - Solution: Error prediction instead of heart-beats detection;
  - Data based prediction: sub-MS detection & recovery possible;



- Fast failure propagation
  - Solution: Data plane propagation instead of control plane protocol;
  - Signal generation and propagation are all hardware native;



IEEE802 standard considerations

- Modeling based on already available data of ethernet:
  - Define prediction algorithm based on available statistical counters & states from physical or link layer of ethernet.
- Failure event based on data plane propagation
  - Like PFC signal, but for failure/congestion event propagation;
  - May need encryption;
  - Need standard;

Summary

- Industry shows interest in AI computing network.
- AICN is deserved to be thoroughly studied and characterized, so to identify and recommend future standard activities.
- Propose the main content of AICN report, which includes background/use cases, requirements and challenges, potential technologies, as well as standard consideration.
- Encourage more contributions and ask for the group's opinion if AICN study item should be initialized.
    - Plan of potential study item is to draft initial version of AICN report, then propose a work item to complete the report.

Study Item Proposal

Study item: AI computing Network (AICN)

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:

- Initial draft of AICN report, including
  - Background/Use cases
  - AI computing network requirements and challenges
  - Potential technologies
  - Standardization considerations

Schedule:

- 3~4 months to draft initial version of AICN
- Propose work item afterwards depending on feedback of AICN report draft

Co-Leader: Lily Lyu (Huawei), Jieyu Li (China mobile)

Contributors/Supporters List

- Jieyu Lie (China Mobile)
- Jose Duato (RSAC)
- Jesus Escudero-Sahuquillo (UCLM)
- Liang Guo(CAICT)
- Lily Lyu (Huawei)
- Weiqiang Cheng(China Mobile)
- Ruixue Wang(China Mobile)
- Yuehua Wei(ZTE)
- Yadong Liu(Tencent)

Motion

To initiate a Nendica study item on AI computing network

Proposed:
Second：