Reliability in AI Networks Gap Analysis, Problem statement, and Requirements

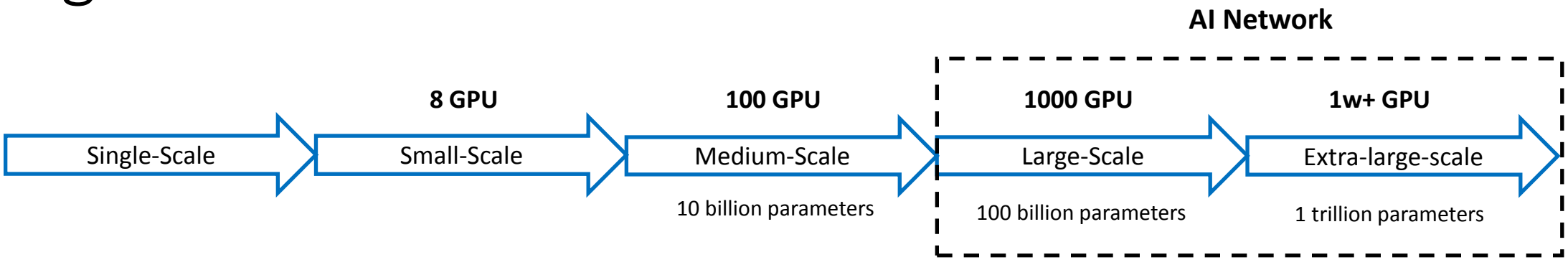
draft-cheng-rtgwg-ai-network-reliability-problem-00

Weiqiang Cheng (China Mobile)

Changwang Lin (New H3C Technologies)

Wenxuan Wang (China Mobile)

Background



The feature of AI model training

- **Large-scale:** most AI large models typically consist of large-scale networks with over **1,000 GPU** cards.
- **High bandwidth:** Distributed training is commonly used, and the larger the network scale, the higher the volume of communication data.
- **Long training duration:** AI model training typically takes days or even months. **Any network interruptions or failures** can lead to training interruptions, requiring the process to be restarted and wasting time and resources.

Some statistics about hardware failures

- **GPU Card Failure:** for a model trained at a scale of **1,000 cards**, the probability of encountering a failure **within a month is 60%.** If the AI network scale reaches 8,000 GPU cards, the probability of experiencing a card failure during a one-month training is 99%.
- **Optical Module Failure:** for an AI training network utilizing nearly 100,000 optical modules, on average, one optical module failure occurs every 4 days.

Existing Mechanism for Route Convergence

Detection mechanism	Technology	Convergence Time	Influencing factors
Fast failure detection 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	

- The convergence time for **local failure** handling includes local detection time and local fast switching time, usually taking tens of milliseconds.
- The convergence time for **global failure** handling includes local detection time, fault notification time, and global fast switching time, typically taking a few seconds.

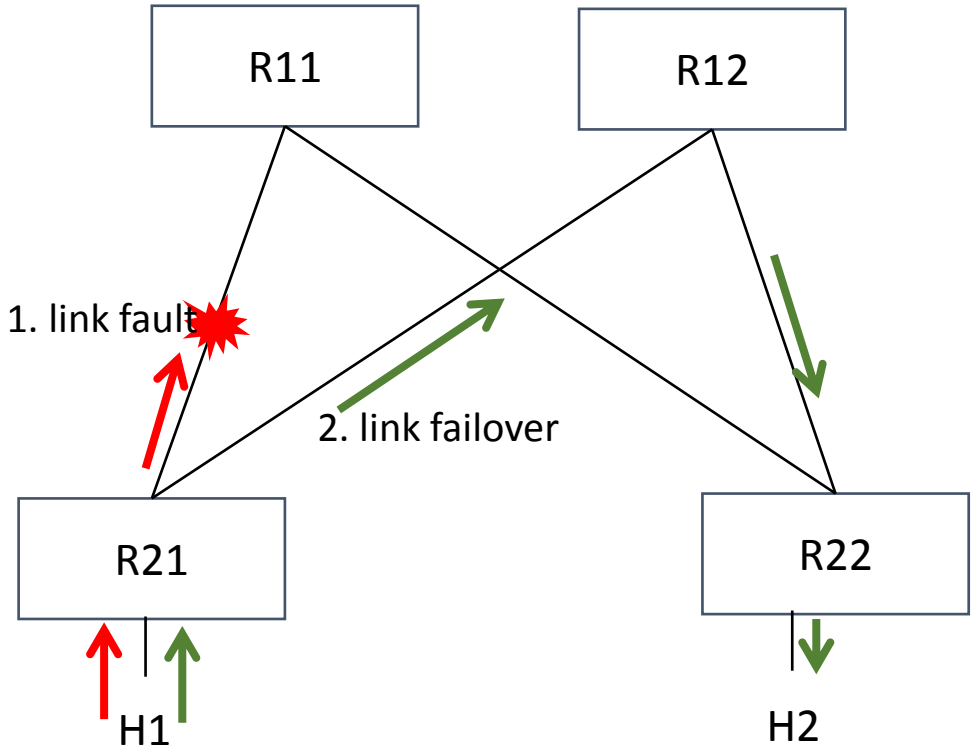
Gap Analysis 1: Local Fault in Spine-Leaf

➢ Local Fault Detect

Currently, link failure can be detected within **a few milliseconds**.

➢ Local Fast Failover

In response to a link failure, perform fast switchover using techniques such as ECMP and FRR to switch the link to a backup link. The switchover time for link failover can be achieved **within milliseconds**.



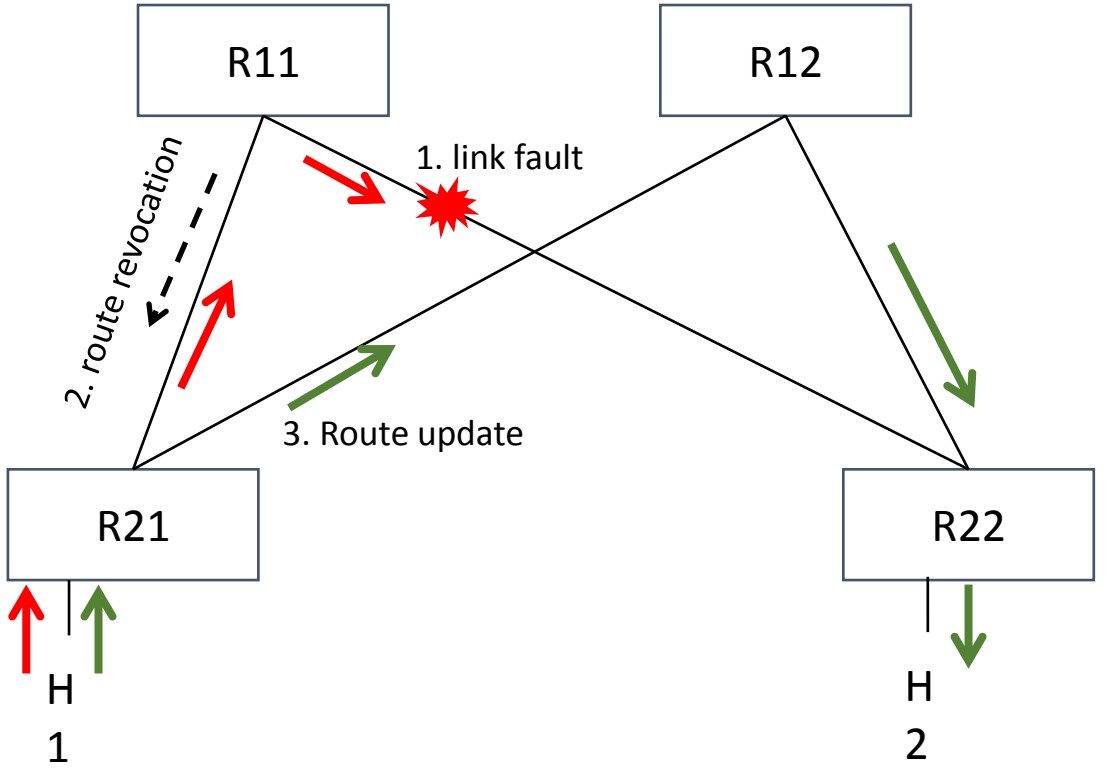
Gap Analysis 2: Remote Fault in Spine-Leaf

➢ Fault Notification

R11 detected a link failure, updates its routing, and notifies R21 to withdraw the route to R22. This process takes **a few milliseconds**.

➢ Global Failover

R21 receives the route withdrawal, recalculates its routing, and updates the routing to point to the correct path. This process **takes a few seconds**.



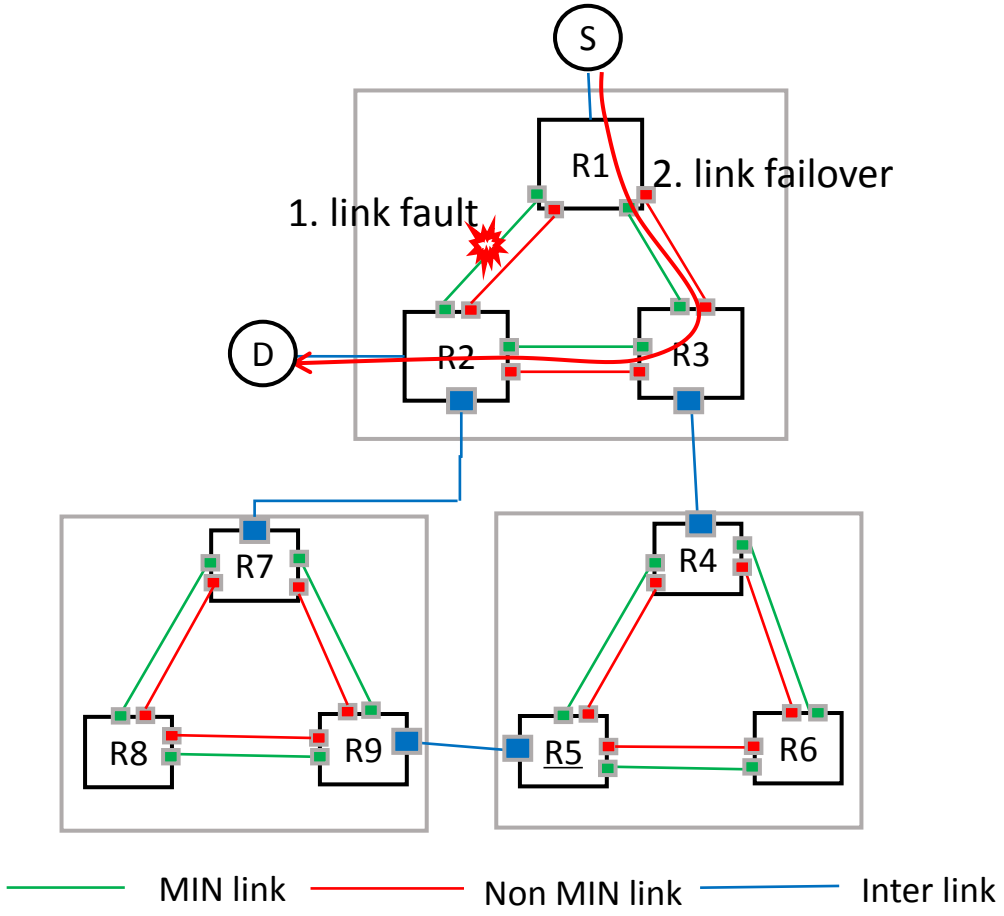
Gap Analysis 3 : Intra-Group Fault in DragonFly

➢ Intra-Group Fault Detect

Router R1 detects a failure in the Intra-Link through fast link detection technology. which can detect link failures within **a few milliseconds**.

➢ Intra-Group Failover

R1 responds to the link failure by switching the link from the Min Link(R1->R2) within the group to the Non-Min Link(R1->R3) . The switchover time for link failover can be achieved **within milliseconds.**



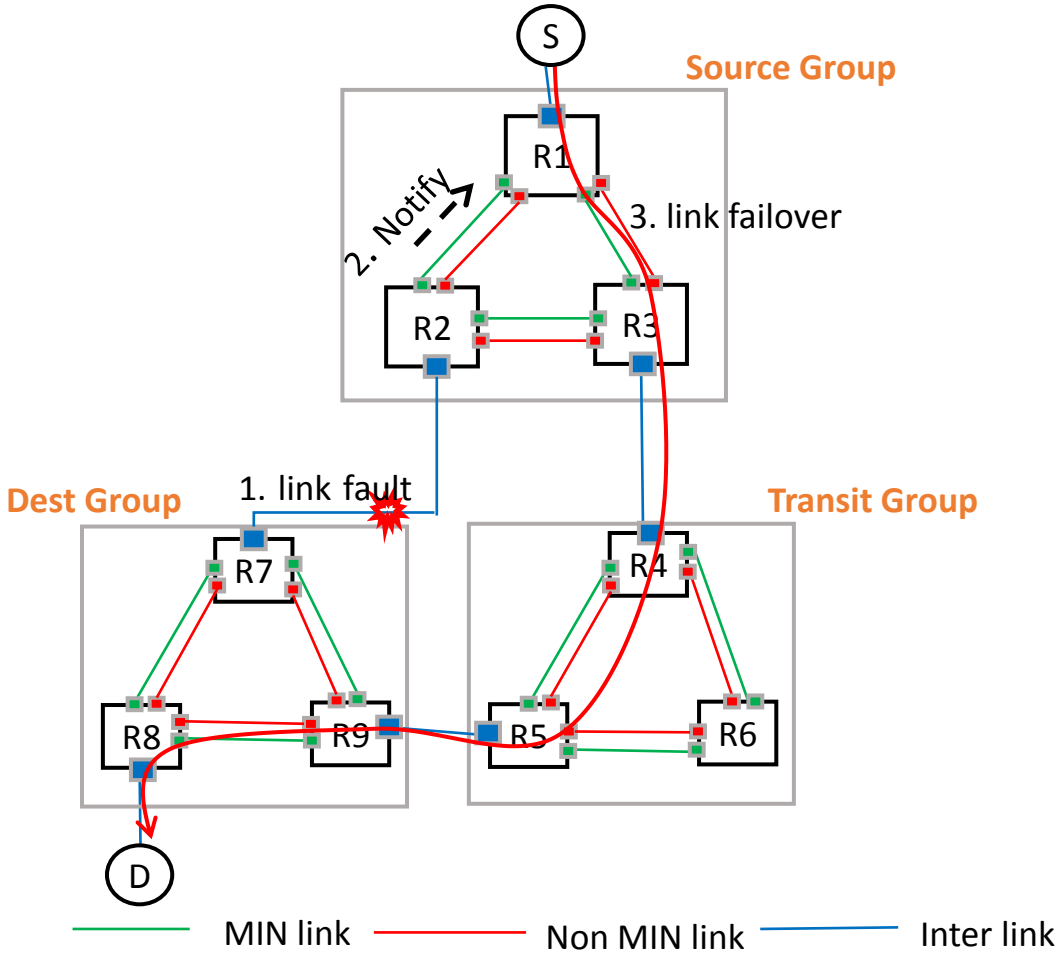
Gap Analysis 4 : Inter-Group Fault in DragonFly

➢ Fault Notification

Router R2 detects a failure and notifies R1 of the link failure event through a fast notification message. The notification process takes **a few milliseconds.**

➢ Inter-Group Failover

R1 responds to the Inter-Link failure by switching the link to the path passing through the Transit Group(R1->R3->R4->R5->R9->R8). The switchover time for link failover can be achieved **within seconds.**



Problem Statement

- Local Fault Detection

The local fault detection time is too long and does not meet the requirements.

- Local Fast Failover

The local fault switchover is too long and does not meet the requirements.

- Fault Notification

There is a lack of notification mechanism for remote faults such as remote links in spine-leaf topology or inter-group links in dragonfly topology. The IGP Link-State flooding and BGP route updates are both too slow.

- Global Fast Failover

Currently, we only have a fast switching mechanism for local failures, but we lack a mechanism for fast switching when responding to remote faults.

Requirements for Reliability in AI Networks

- Fast Fault Detection Mechanism

This mechanism must be capable of detecting faults at sub-millisecond level.

- Local Fast Failover

This mechanism must be able to perform local ECMP or FRR switching, with fault recovery time in the sub-millisecond range.

- Fast Fail Notification

This mechanism must provide sub-millisecond fault notifications.

- Fast Global Failover

This mechanism must provide global fast fault switching, with recovery time in the sub-millisecond range.

- Topology Awareness

This mechanism must be capable of detecting changes in the spine-leaf or dragonfly topology and performing rapid switching accordingly.

Any questions or comments are welcomed.