The Challenges and Practices of Network Stability in Alibaba's Large-Scale Computing Clusters

OCP GLOBAL SUMMIT

OCT 15–17, 2024
SAN JOSE, CA

Networking

NETWORKING

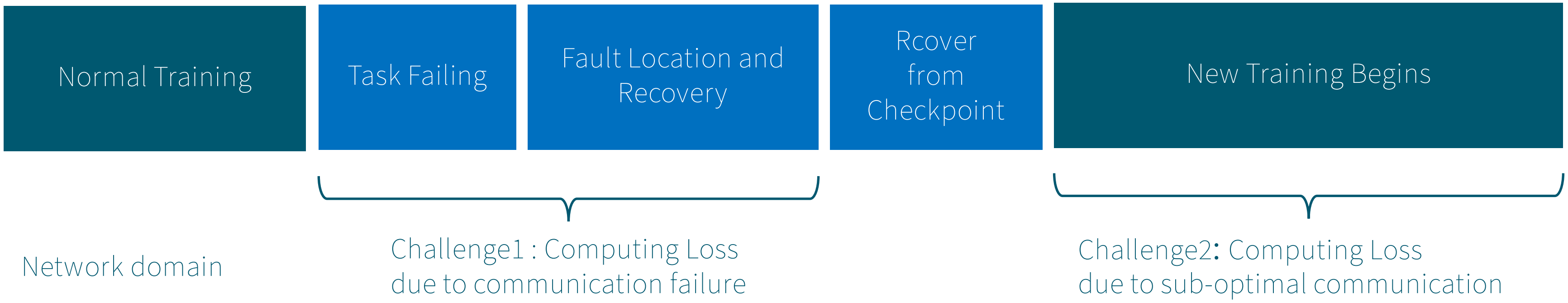
The Challenges and Practices of Network Stability in Alibaba's Large-Scale Computing Clusters

[Xuemei Shi, Alibaba]

[Surendra.anubolu, Broadcom]

OPEN PLATINUM™

Two Challenges we cover here



Normal Training

Task Failing

Fault Location and Recovery

Rcover from Checkpoint

New Training Begins

Network domain

Challenge1 : Computing Loss due to communication failure

Challenge2: Computing Loss due to sub-optimal communication

- Challenge1: How we detect and recover from communication failure ASAP
- Challenge2: How to observe the real network communication behavior

Larger scale, More components

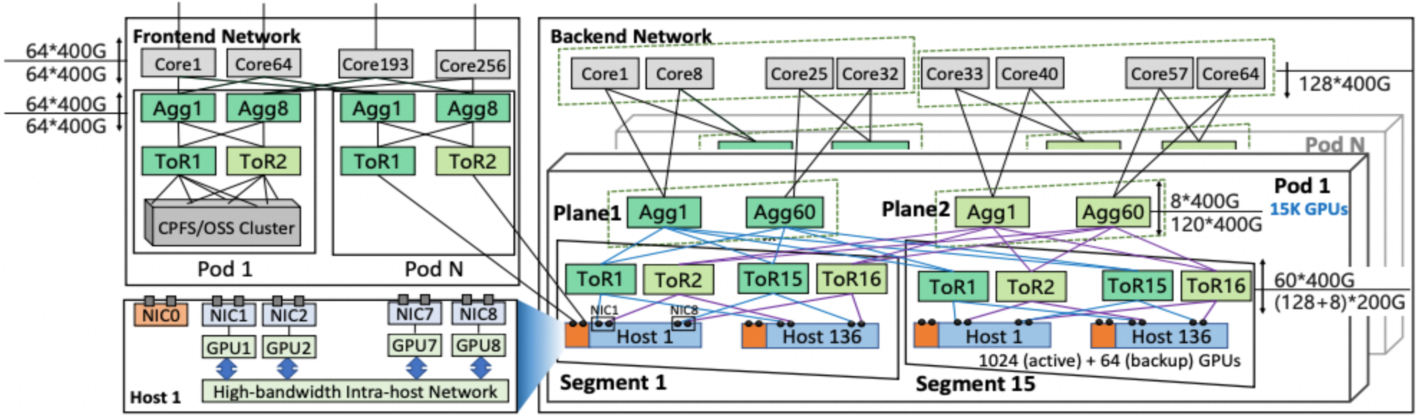
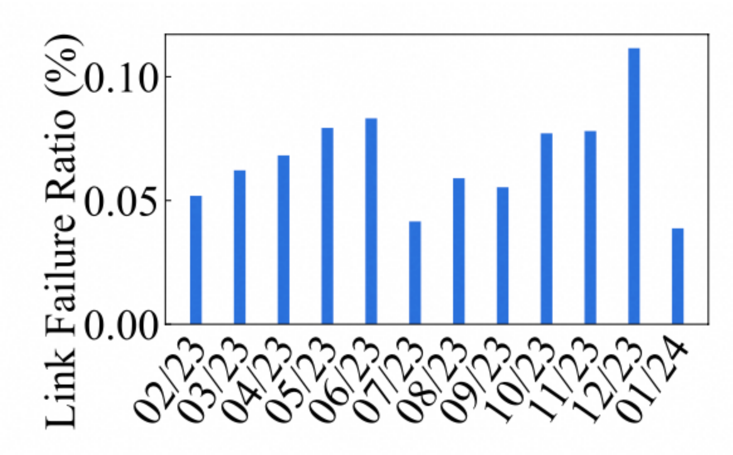


Figure 7: HPN overview. A solid parallelogram represents a segment (containing 1024 active GPUs and 64 backup GPUs). Two dotted parallelograms represent dual-plane. A cube contains an entire Pod (containing 15K GPUs).

- 15K computing node, GPUs, within one Pod
- 10x communication node, network components, composing switches, modules and links.

SIGCOMM’24 Alibaba HPN: A Data Center Network for Large Language Model Training

Higher number of failures



- In each month, 0.057% of links fail.
- In each month, 0.051% of switches encounter critical errors and crashes.
- Lots of link issues happen every day due to optics failures.

Training is Sensitive to Communication Faults

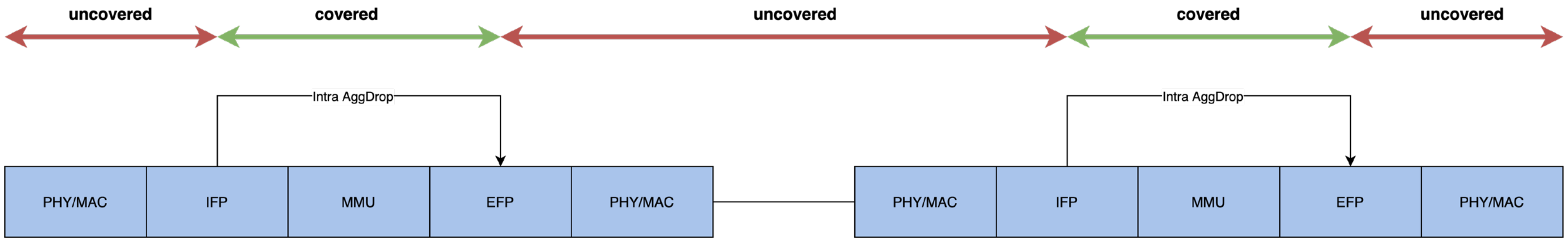
Fault Type	Fault Frequency	Fault Impact
GPU ECC	High	Task Failure
GPU Card malfunction	High	Task Failure
GPU Slow	High	Performance Decline/Fluctuation
Nvlink Fault	High	Task Failure
Pcie Fault	High	Performance Decline
Link Disruption	High	Performance Decline
Link Flapping	High	Performance Fluctuation
NIC Configuration Error	Medium	Task Failure/Performance Decline
NIC Flow Table Unloading Anomaly	Low	Performance Decline/Fluctuation
Congestion Control Anomaly	Low	Performance Decline/Fluctuation
Switch Interconnectivity Fault	Low	Task Failure
NIC FW Bug	Low	Task Failure
Storage Anomaly	Medium	Task Failure/Performance Decline
NCCL Environment Variable	High	Task Failure/Performance Decline
User Code Issue	High	ALL
Unknown Reason Hang	High	Task Failure

- A single link failure could cause 5%+ performance degradation, i.e. 50 cards in 1K scale
- Any component could lead to failure in a large scale production network

Solutions we have had in AI/ML network

- Ping style solutions
    - Not the live training traffic
- Piggyback style solutions
    - Not all of the live training traffic
- Counter style solutions
    - For all of live training traffic
    - But not cover all of network components, all of drop events

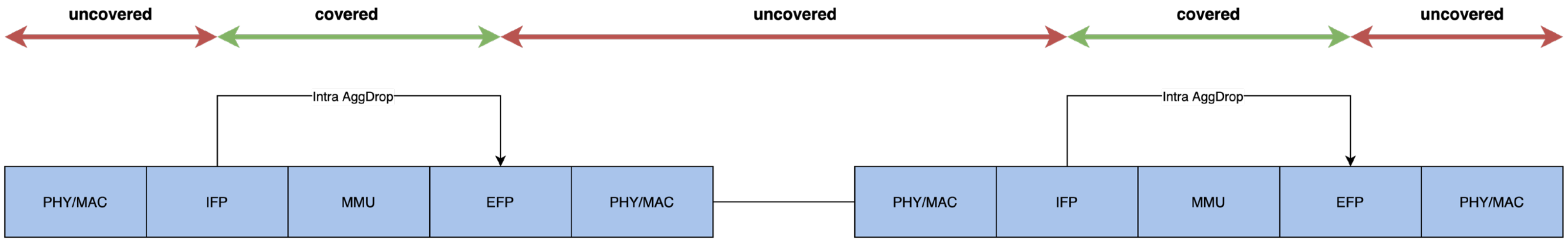
Counter style Intra Aggregate Drop



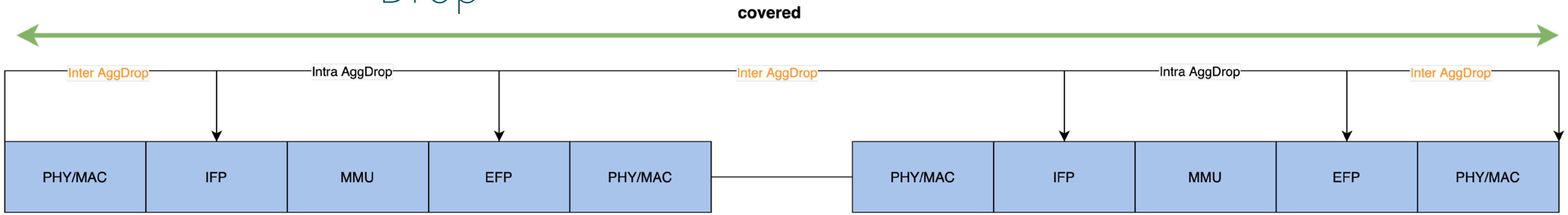
Theory of operation:

1. IFP(a.k.a. ACL) marks the metadata bit of the packets between 0 and 1 every second.
2. IFP counts the total number of packets marked as 0 and 1 separately, and EFP does the same.
3. During the period when the value is set to 1, obtain the packet loss in previous 0 cycle by calculating the difference in the number of packets marked as 0 between EFP and IFP.
4. The difference is the number of dropped packets marked as 0.
5. The same drop calculation for the packets marked as 1 during marking to 0.

Yet Another Counter style Communication Fault Detection

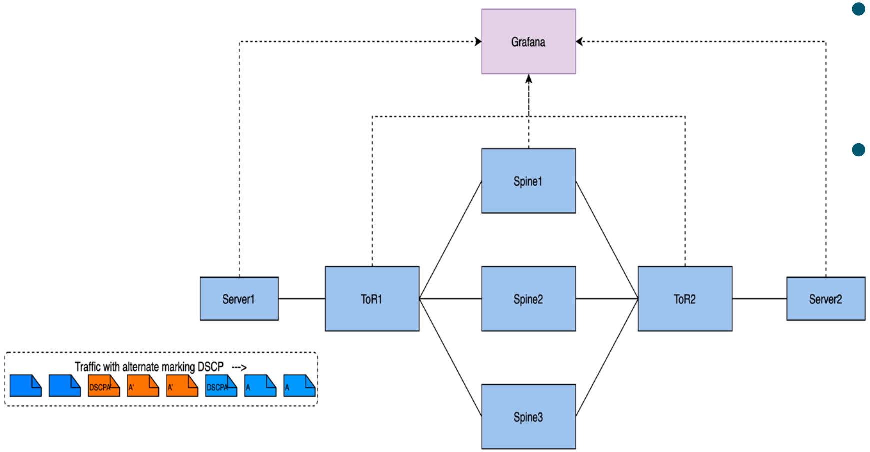


Per hop solution – Intra Aggregate Drop



Our Best Practice: End to End Solution – A.M.D

Alternative Marking DSCP(A.M.D)



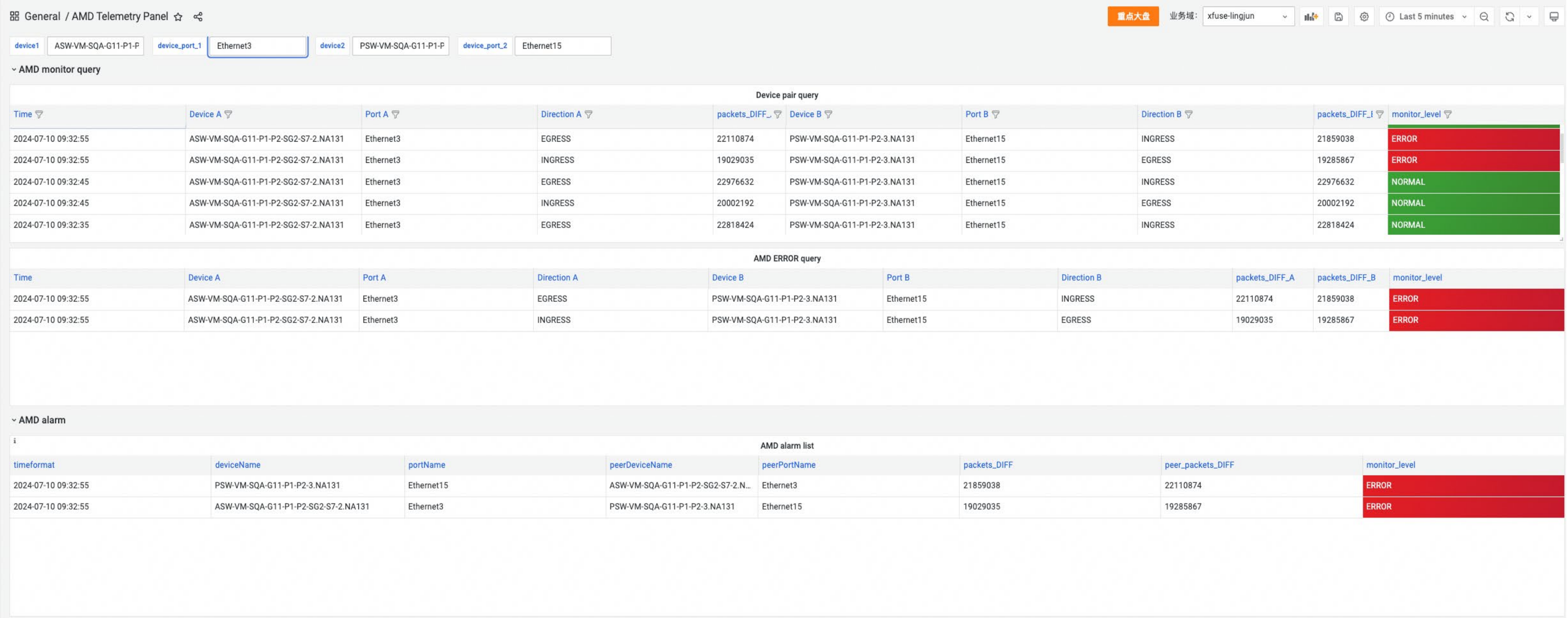
- Inspired by RFC8321(**Alternate-Marking Method for Passive and Hybrid Performance Monitoring)**
- Theory of operation:
  - Most is similar to intra Aggregate Drop, with exception:

1. Instead of internal metadata bit, mark DSCP for E2E.
2. Instead of the switch, the NIC performs this marking.
3. Instead of switch level, Count are refined to be based on port level. Packet loss on link can be calculated by exchanging counter between local port and its peer.

Alternative Marking DSCP(A.M.D)

1. Easy Deployment
   - Unlike reserve bits or extra-header, DSCP is conveniently deployable with wide support across components and platforms.
2. Noise-Free
   - Application or service based DSCP, focusing on packet-loss-sensitive services, where any packet loss is considered an anomaly
3. Real-time Fault Localization
   - Every device and link performs real-time calculations, enabling immediate detection of packet loss, and the discovered location is the fault location, allowing us to take action within seconds.
4. Comprehensive
   - Covers all network components, especially the last hop in the network, NIC to TOR.
   - Covers all communication of application since marking the real application packets.
5. Unified
   - Unified solution for both flow-based and packet-spray-based approaches in AI/ML networks.

Alternative Marking DSCP(A.M.D)



General / AMD Telemetry Panel

业务域: xfuse-lingjun

device1: ASW-VM-SQA-G11-P1-P

device_port_1: Ethernet3

device2: PSW-VM-SQA-G11-P1-P

device_port_2: Ethernet15

AMD monitor query

Device pair query

Time	Device A	Port A	Direction A	packets_DIFF_	Device B	Port B	Direction B	packets_DIFF_I	monitor_level
2024-07-10 09:32:55	ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131	Ethernet3	EGRESS	22110874	PSW-VM-SQA-G11-P1-P2-3.NA131	Ethernet15	INGRESS	21859038	ERROR
2024-07-10 09:32:55	ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131	Ethernet3	INGRESS	19029035	PSW-VM-SQA-G11-P1-P2-3.NA131	Ethernet15	EGRESS	19285867	ERROR
2024-07-10 09:32:45	ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131	Ethernet3	EGRESS	22976632	PSW-VM-SQA-G11-P1-P2-3.NA131	Ethernet15	INGRESS	22976632	NORMAL
2024-07-10 09:32:45	ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131	Ethernet3	INGRESS	20002192	PSW-VM-SQA-G11-P1-P2-3.NA131	Ethernet15	EGRESS	20002192	NORMAL
2024-07-10 09:32:35	ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131	Ethernet3	EGRESS	22818424	PSW-VM-SQA-G11-P1-P2-3.NA131	Ethernet15	INGRESS	22818424	NORMAL

AMD ERROR query

Time	Device A	Port A	Direction A	Device B	Port B	Direction B	packets_DIFF_A	packets_DIFF_B	monitor_level
2024-07-10 09:32:55	ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131	Ethernet3	EGRESS	PSW-VM-SQA-G11-P1-P2-3.NA131	Ethernet15	INGRESS	22110874	21859038	ERROR
2024-07-10 09:32:55	ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131	Ethernet3	INGRESS	PSW-VM-SQA-G11-P1-P2-3.NA131	Ethernet15	EGRESS	19029035	19285867	ERROR

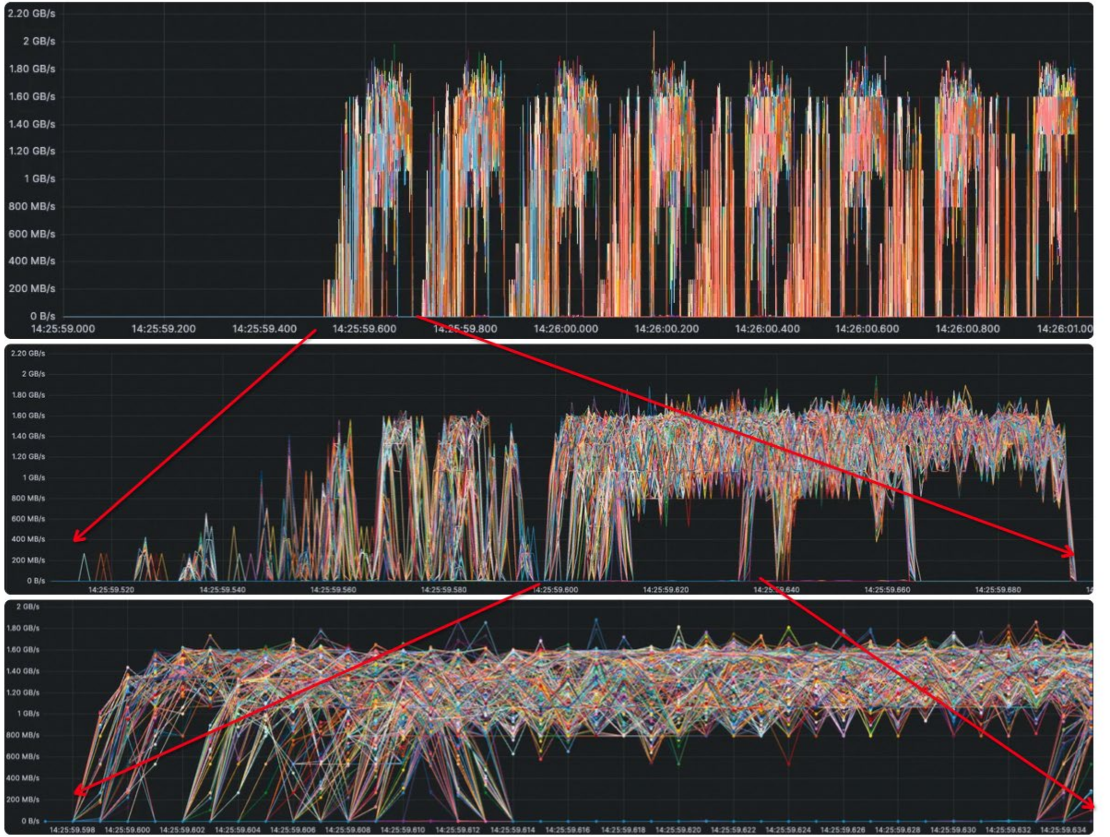
AMD alarm

AMD alarm list

timeformat	deviceName	portName	peerDeviceName	peerPortName	packets_DIFF	peer_packets_DIFF	monitor_level
2024-07-10 09:32:55	PSW-VM-SQA-G11-P1-P2-3.NA131	Ethernet15	ASW-VM-SQA-G11-P1-P2-SG2-S7-2.N...	Ethernet3	21859038	22110874	ERROR
2024-07-10 09:32:55	ASW-VM-SQA-G11-P1-P2-SG2-S7-2.NA131	Ethernet3	PSW-VM-SQA-G11-P1-P2-3.NA131	Ethernet15	19029035	19285867	ERROR

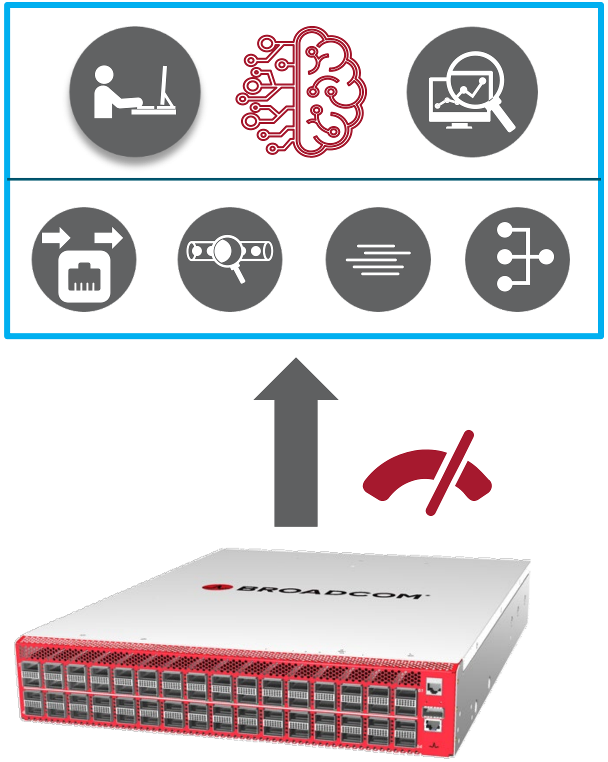
Granular Telemetry

- Chart showing actual utilization at high frequency
- Telemetry with counter polling is slow → 100s of mSec
- Large amount of stat data
    - Need efficient mechanism
- Need fast response for link events
- Granular telemetry required to collect data at high frequency



Streaming Telemetry for AI - Details

- Reports w/ port, queue & buffer stats are sent at ~100 microseconds frequency to host CPU
- Immediate data availability supports rapid decision-making and quick response / remediation
- Reports can be sent to a remote collector as well
- Scalability and efficiency: reduces overhead on the network device by minimizing polling and improve scalability
- Enabled efficient A.M.D. at Alibaba for a high performance AI fabric



Call to Action

- Problem to solve

  1. NIC supports more flexible DSCP setting like flow-based marking, counting, and being set periodically at sub-second level.
  2. Switch supports dedicated counters resource instead of the expensive TCAM for A.M.D counting.
  3. Switch supports dedicated implementation to faster recover from network fault detected by A.M.D.
  4. Switch supports streaming telemetry with granular data at high frequency
  5. The SAI definition for the resource listed above

Thank you!