Availability challenges and requirements of AICN

Hesong Li, Yuehua Wei, Yuanbin Zhang

ZTE Corporation

In AICN report draft 1-24-0028-00-ICne-aicn-report-draft.pdf, “Availability” is listed as one topic in the chapter “Requirements and Challenges of AI computing Networks”. This document describes availability issues and proposes text to AICN report.

Network availability (NA)is a measure of how well a computer network can respond to the connectivity and performance demands placed on it. It is also known as network uptime. Network availability is calculated by dividing the uptime by the total time in any period. The goal is 100% availability, although another commonly referenced goal is known as “five nines,” or 99.999% availability. That’s the equivalent of only about 5 minutes of downtime in a year.

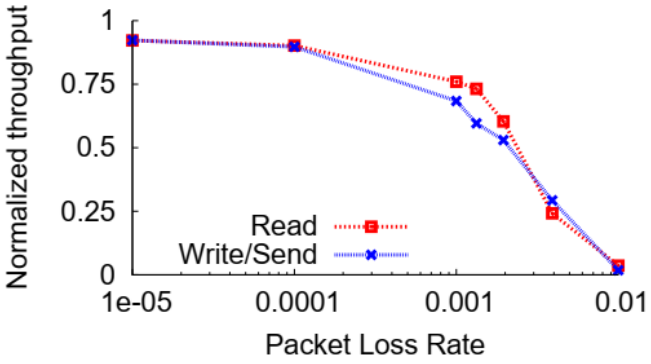
Typical KPIs of NA are MTBF (Mean Time Between Failure) and MTTR (Mean Time To Repair):

Availability = MTBF/(MTBF + MTTR)

In a datacenter network (DC), both link level errors and network level events matter the network high availability (NHA). Link level errors degrade network performance, and network level events affect operation KPI such as bandwidth utilization, latency, etc. if not properly dealt with, will cause service interruption.

Redundancy and failover is essential for HA in a DC network, but the actual performance is use case specific. CRC, statistical counters and error detection tools can achieve sub-50ms link level error recovery; Network management station (NMS) is typically used to detect network level events and do service recovery in minutes to hours.

There’s obvious differences between AICN and cloud computing DCN. Training of AI models leads to rapidly increasing number of parameters, entries of embedding tables, and words of context buffers. As a result, large cluster is a requirement. And the AI computing is bandwidth hungry with dominant remote direct memory access (RDMA) traffic. In the AICN, training jobs run for long periods of time with elephant flows and synchronized and bursty traffic. Tail latency impacts the AI training's job completion time significantly. According to an experiment of RDMA over RoCEv2 with go-back-n loss recovery model[1], the throughput degrades significantly once the packet loss rate exceeds 0.1%. When packet loss rate reaches 1%, the throughput is 0.



To support the emerging AI workloads, AICN needs enhancements to provide high utilization, reliable transport, and predictable low tail latency to minimize job completion time for these workloads. Availability however, is the prerequisite of high performance. To improve the overall cluster availability, some new technologies may be needed to achieve both lossless fabric and fast error recovery.

Error prediction

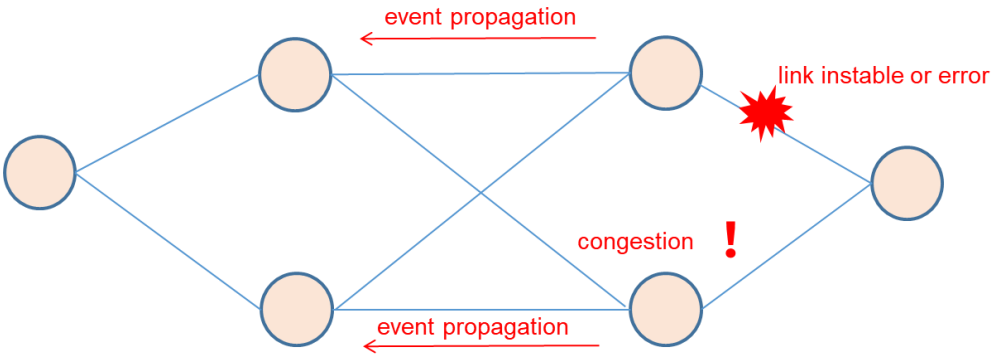
In a large scale AICN, many GPUs work simultaneously. The computing resources are connected by a large number of NICs, switches, optical modules. Link failure and node failure leads to interruption of the training jobs. The failure detection and recovery efficiency is critical to computing resource utilization.

Error prediction based on statistical counters and real-time states of the network may help to achieve sub-ms detection and recovery. The efficiency is better than 50ms recovery based on heart-beat detection mechanism.

Data Plane Fast Recovery

Traditional route convergence technologies rely on dynamic routing protocols or Bidirectional Forwarding Detection (BFD) to detect faults and recalculate the path. The route convergence takes hundreds of milliseconds. In a large-scale AICN it may take seconds, which will result in computing jobs interruption.

Fault detection and event propagation via data plane brings significantly fast link failure or congestion events propagation.



Link Level Reliability

AICN is bandwidth hungry and latency sensitive, 400G and beyond Ethernet rate need Pulse-Amplitude Modulation 4-level (PAM4) instead of Non-Return to Zero (NRZ) because PAM4 effectively doubles the bit rate compared to NRZ for a given baud rate, enhancing efficiency in high-speed optical transmission, and significantly reduces signal loss in the transmission channel for PAM4 signaling. But PAM4 signaling becomes more susceptible to noise, resulting in a higher bit error rate (BER). Suppose post-FEC BER is 1e-12, a typical 256 GPU AI POD will be estimated to suffer 2700 error frames per second. PAM4 may implement advanced Forward-Error Correction (FEC) to enable linked systems to achieve the desired Bit Error Rate. But the more complex FEC mechanisms may increase the latency significantly.

An alternative approach has been adopted by Peripheral Component Interconnect Express (PCIe) and InfiniBand. The idea is that the receiver first uses a light weighted FEC to correct some bit errors and then checks the CRC. If this check fails, it initiates a simple link-layer retransmission protocol to request the data again. So a tradeoff between latency and buffer consumption may be achieved.

Reduced Lane Mode

Optical module is the primary source of AICN failures. Standard high speed Ethernet interface is multi-laned, currently a single-lane failure will cause the entire physical interface going down, which results in frequent training job interruptions. Reduced lane mode may come in handy. Faults of a single lane can be discovered in time and effectively isolated, service may be removed from the failed physical lanes, and the port remains operable in properly scaled reduced rate, which is of great significance for improving the availability of AI cluster.

References

[1] Datacenter Ethernet and RDMA: Issues at Hyperscale https://arxiv.org/abs/2302.03337