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.

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][2][3], 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.

NA (Network availability) 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 general, NA is a fairly simple concept. But in the AI computing data center, it can become a complicated one. There are various components in an AICN, E.g. Power system, cooling system, cables, optical modules, intra-node switch, storage unit, accelerator and management system. Each component is an integration of essential hardware and software. Many factors contribute to the availability of an AICN. The academic community has already conducted some modeling analysis on the availability of a complicated network system [4][5]. From the perspective of the practical network deployment, link level errors and network level errors decrease the NA. To improve the overall cluster availability, some new technologies may be needed to achieve both lossless fabric and fast error recovery.Four technical requirements within the scope of IEEE802 are included in this report.

Error predictionIn 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](https://info.support.huawei.com/info-finder/encyclopedia/en/BFD.html) ([BFD](https://info.support.huawei.com/info-finder/encyclopedia/en/BFD.html)) 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 take sub-milliseconds, which would help the network to achieve faster link failure recovery or shorter congestion relief time.



Link Level Reliability (LLR)

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 [6] 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 BER. 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 most of the bit errors and then checks the CRC. If this check fails, the receiver initiates a simple link-layer retransmission protocol to request the data again. A light weighted FEC save dozens of nanoseconds to all the frames [7], the LLR only cost several microseconds to a tiny portion of all the frames. So a tradeoff between latency and link level availability 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. <https://en.wikipedia.org/wiki/RDMA_over_Converged_Ethernet#RoCE_v2>
2. https://community.fs.com/article/an-indepth-guide-to-roce-v2-network.html
3. Datacenter Ethernet and RDMA: Issues at Hyperscale， <https://arxiv.org/abs/2302.03337>
4. S. K. Chaturvedi, Network Reliability: Measures and Evaluation. Hoboken, NJ, USA: Wiley, 2016
5. Availability Model for Data Center Networks with Dynamic Migration and Multiple Traffic Flows, https://ieeexplore.ieee.org/document/10037235
6. What Is a Pod? What Is a Cluster? <https://blogs.nvidia.com/blog/what-is-a-cluster-pod>
7. https://www.ieee802.org/3/dj/public/adhoc/optics/0423\_OPTX/brown\_3dj\_optx\_01b\_230413.pdf