Updates to 1-24-0031-00-ICne

Yuehua Wei (wei.yuehua@zte.com.cn)

IEEE 802 Plenary, Montreal, CA.

Nendica Meeting, 2024-07-18

Updates

- Addressed comments collected from the Nendica meeting held on 2024-05-30.
- Re-phrased several paragraphs of version 00.
- Added more references.
- Ver01： https://mentor.ieee.org/802.1/dcn/24/1-24-0031-**01**-ICne-availability-challenges-and-requirements-of-aicn.docx

How to apply typical NA KPI to large scale AICN?

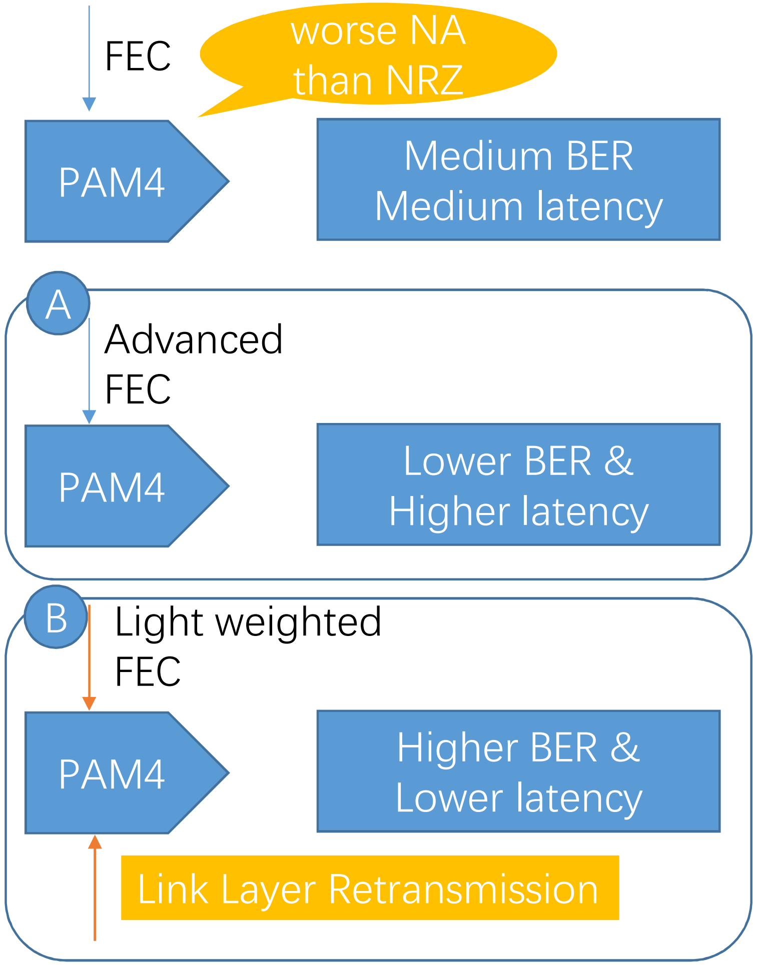
- Typical KPIs of Network availability (NA) are MTBF (Mean Time Between Failure) and MTTR (Mean Time To Repair):

Availability = MTBF/(MTBF + MTTR)

- It’s simple to apply this equation to a single service or an individual component.
  - It’s complex to apply it to a whole AICN
  - Various components
- The academia has already conducted some modeling analysis on the availability of a complicated network system, several papers are listed for reference.
- This contribution only list several technical factors that may contribute to NA within the scope of IEEE802.

The logic of introducing LLR to improve the NA

- AICN is bandwidth hungry and latency critical
  - 400G and beyond Ethernet rate need PAM4 instead of NRZ because PAM4 doubles the bit rate.
  - But PAM4 signaling becomes more susceptible to noise, resulting in a higher bit error rate (BER)
    - **A higher BER (than NRZ) results in a worse NA.**
    - Implement advanced FEC can achieve the desired BER. But increase the latency significantly.
    - Use a light weighted FEC to correct most of the bit errors and then checks the CRC
      - If this check fails, it initiates a simple link-layer retransmission protocol to request the data again.
  - A light weighted FEC saves dozens of nanoseconds to all the frames, the LLR only costs several microseconds to a tiny portion of all the frames.


FEC
worse NA than NRZ
PAM4
Medium BER Medium latency
A
Advanced FEC
PAM4
Lower BER & Higher latency
B
Light weighted FEC
PAM4
Higher BER & Lower latency
Link Layer Retransmission

The relation between congestion and NA

- Network congestion will result in packet drop, that could interrupt AI computing jobs. So from this perspective, network congestion may affect network availability.
- The requirement in this contribution is not talking about a particular congestion control mechanism. Only possible congestion state propagation via dataplane message.
    - More efficient than control protocol.

AI PoD in AICN

- Typical 256 GPU AI POD means a POD consists:
  - 32 computing nodes
    - Each node comprises 8 GPUs

Add more 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

Next step

- Merge Ver01 of this contribution to the AICN report

Thank You!