Open Challenges Proposal

Call Identifier: OCF-OC1-Challenge3

Challenge: Enhance AI clusters resilience and reactivity to data-plane failures by leveraging on sparse in-network resources (e.g. memory and compute limited switches).

Background: Access link faults, i.e., failures between the GPU NIC and its ToR switch, are among the most damaging in large-scale LLM (large-Language Model) training and inference clusters. Unlike interconnect faults, they halt jobs instantly, forcing expensive checkpoint recovery and hours of lost progress. They also undermine GPU fault-recovery mechanisms, as checkpoint reloads can also fail when the network is unstable due to packet loss. Studies show that access link faults account for more than half of all network failures, making them the most critical type of faults to address for reliable LLM training and inference. While end-host solutions exist to mitigate their impact, they are typically slower and less accurate than switch-centered approaches, since switches provide a direct real-time view of links health and ports state. In such a context, in-network data-plane devices failure detection capabilities can be coupled with end-host collective communication pattern adaptation in order to enhance reactivity and robustness.

Technical Challenge: The main technical challenge of this project comes from the strict hardware limitations of datacenter switches and the complexity of detecting subtle network failures. Switches are designed for fast forwarding with very limited memory and compute, which makes storing and processing state information at packet or flow level non-trivial. Furthermore, while detecting obvious hard failures (e.g., complete link-down event) is straightforward, identifying intermittent or gray failures (e.g., when a link remains up but performance degrades due to packet loss) is far more difficult when relying only on the limited resources of datacenter switches.

Main Objectives: Leverage in-network capabilities, in particular switch-level fault detection, together with end-host collaboration, to design and implement a fault-tolerant solution that aims to maintain collective communication performance as close as possible to fault-free situation, under network faults.

Deliverables: Publication

Duration: 6 Months

Available Funding: 60k€

Sponsor: Huawei – John Stephan (john.stephan@huawei.com)