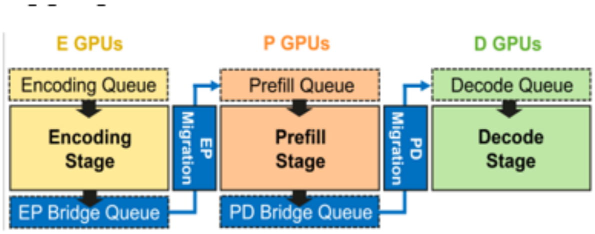
Open Challenges Proposal

Call Identifier: OCF-OC1-Challenge4

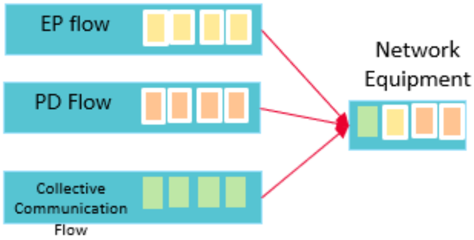
Challenge: Multimodal AI Inference Traffic Scheduling and Performance Optimization.

Background: The encoding phases in multi-modal inference consume significant resources. Video-to-token encoding traffic and prefilling traffic interfere with each other. Additionally, both the LLM model and the KV cache require high storage capacity. As a result, the processing rate of images or videos is limited due to resource competition.

Technical Challenge: Currently, the EPD separation mode is being utilized. Traffic transmitted between the Encode-Prefill and Prefill-Decode is introduced, which interferes with the inferred aggregate communication traffic, such as all2all.



From EPD Co-Partitioning to EPD Disaggregation



Different traffic between multiple tasks interferes with each other.

Main Objectives: Identify and isolate different types of inference traffic on the network, and perform differentiated scheduling increasing inference processing by 10%.

Deliverables: Publication

Duration: 6 Months

Available Funding: 60k€

Sponsor: Huawei – Ding Jiutao (dingjiutao@huawei.com)