Divert, not Throttle: Colocating Batched Jobs with Online Services in Datacenters

Online Services running in datacenters have pre-defined Service Level Objectives (SLOs) that the Datacenter Provider must strive to meet, such as a threshold tail latency.

However, these services commonly run well below the load at which their SLOs get violated.

The excess resources are utilized by Datacenter Providers to run non-latency-critical batched applications.

However, the online services' load patterns are generally bursty.

Colocating batched jobs with latency-critical online services eats up the slack in the latter's SLOs that would have otherwise accommodated the load spikes.

Therefore, during periods of bursty load, the Datacenters generally throttle the batched applications.

This causes significant degradation in the batched jobs' throughput.

We observe that the throughput of the colocated batched applications is bound by the available memory bandwidth and not the memory latency. This observation contradicts prior studies.

We propose a design that utilizes expanded memory using the CXL protocol. Our design replicates a portion of the batched jobs' dataset on both the low-latency local DRAM and the CXL-attached memory. During periods when online services face bursty load, our design diverts the memory accesses of batched jobs towards the CXL memory. This minimizes their interference with the colocated online services ensuring that their SLOs are not violated, while experiencing a much lower degradation in throughput compared to the approach where they are throttled.

Our work is still in progress. We are currently conducting preliminary experiments to motivate our approach.