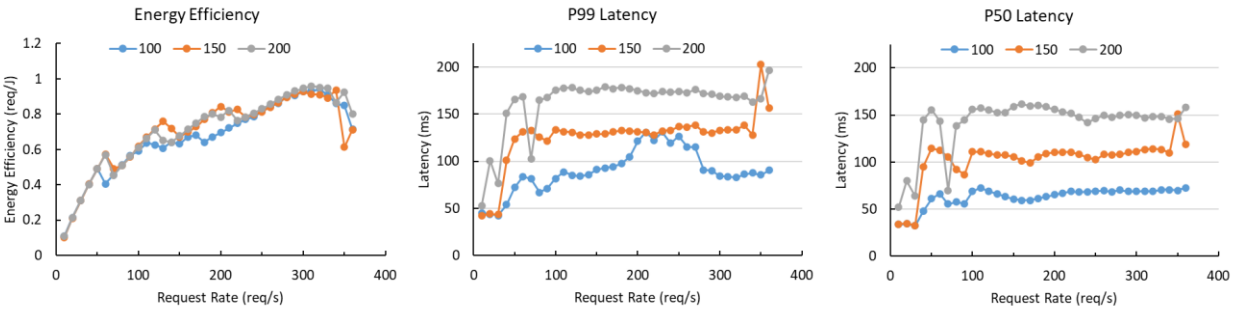
0425-0501

Triton dynamic scheduler – clock 변경 방식 수정

- 실험 결과에서 latency가 위반되는 이유가 스케줄링 과정에서 clock 조절 수행방법에 영향을 받는 것을 확인했음.
- 기존 코드에서는 변경할 GPU 개수를 파라미터로 주어지는 클럭 조절 함수를 async하게 실행했다. 루프를 돌면서 GPU마다 개별 nvml 쿼리를 날리는 것과 비교했을 때, 실행 속도는 async하게 동작하는 쪽이 최소 20% 더 느림.
- (추측) 오히려 새롭게 thread를 생성하는 오버헤드가 추가적으로 발생하고, 이것이 누적되면서 latency를 맞출 수 없는 상황이 발생할 것이다.

Single GPU, various SLOs



- 스케줄러 코드에서 클럭 조절 부분을 수정하여 다시 측정하였음. 위의 그림은 TitanRTX-Inception 모델 조합에서 100, 150, 200ms multi-GPU 환경 측정 결과.
- 정리 1: SLO에 따라서 최대 throughput이 차이가 없다. 조금 더 나아가서 GPU를 4개 사용하게 되는 request rate에서는 에너지 효율성이 최대가 되는 throughput이 비슷하거나 같은 것으로 보인다.
- 정리 2: 100ms의 경우에는 latency를 위반하는 경우가 생긴다. 이 부분은 GPU scale이 정상적으로 동작하지 않는 것으로 추측된다.

계획

- 정리 1, 2에 대한 이유 파악
- 논문 Approach 수정