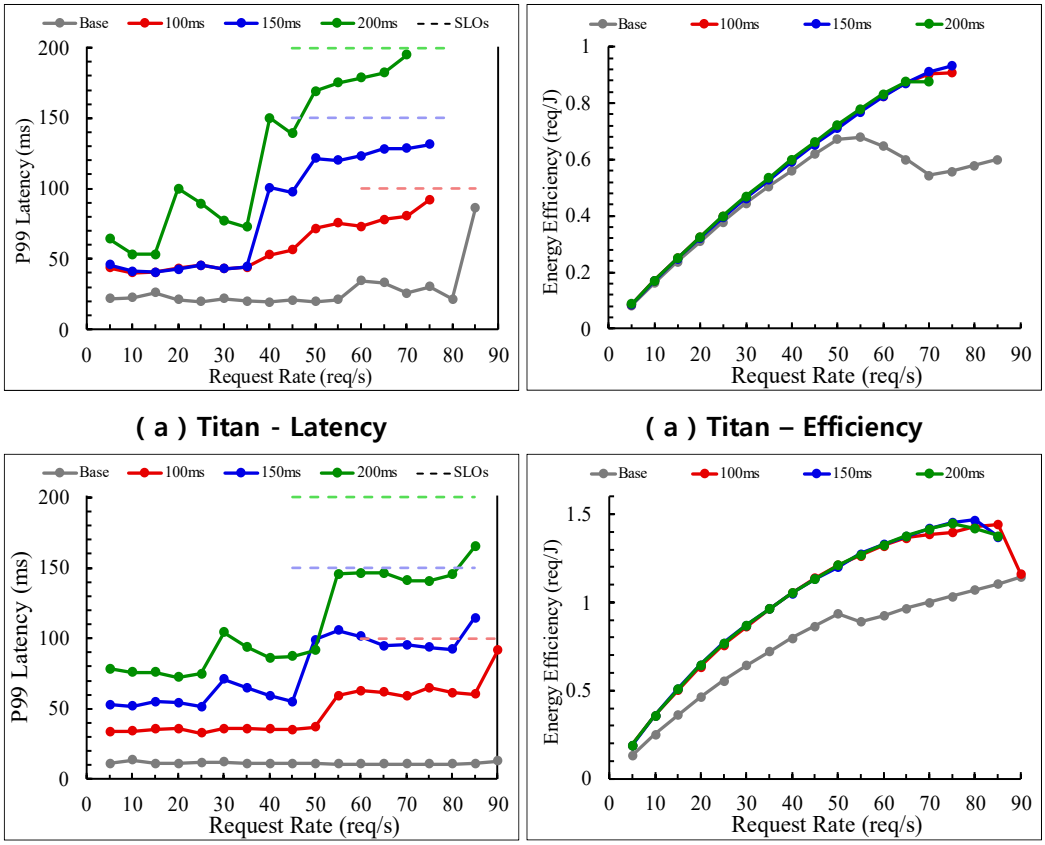
0606-0612

Single GPU에 대한 SLO test (계속)



(a) Titan - Latency **(a) Titan – Efficiency**

(c) V100 - Latency **(d) V100 - Efficiency**

Figure 1. Inception model with single GPU

- Figure 1.은 inception 모델에 대해서 두 GPU의 latency와 energy efficiency를 나타내고 있다. Baseline을 포함하여 3개 SLO target에 대해서 실험을 진행했다.
- Issue 1. Baseline보다 최대 throughput은 조금 낮게 측정되었다. 이것은 클럭을 변경하는 오버헤드로 인해서 발생할 수 있다.
- Issue 2. V100의 경우, 50 req/s 보다 높은 경우에 SLO target에 비해 빠르게 처리된다. SLO target의 90%를 만족하도록 GPU clock을 조절하고 있지만 200ms (초록색)을 보면 150ms 즘에서 99th percentile latency가 형성되고 있다. 이것은 시간을 clock으로 바꾸는 함수의 raw data가 맞지 않을 가능성이 있다.
- Issue 3. TitanRTX의 경우, latency는 target SLO에서 형성되고 있지만 정상적으로 처리되고 있지 않다. SLO가 커질수록 max throughput이 낮아지는 결과를 보인다.

문제점 분석 및 해결 방안

- 두 GPU에서 서로 다른 경향을 보이고 있지만 공통적인 문제점은 필요한 클럭이 필요한 시점에 제공되지 않는다는 것이다. V100의 경우에는 과하게 높은 클럭으로 설정되고 있어 target SLO보다 빠르게 처리되고, TitanRTX의 경우에는 낮은 클럭으로 설정되어 throughput이 낮아지는 결과를 보이고 있기 때문이다.
- 이를 해결하기 위해서는 각 GPU에서 inference time을 다시 측정하는 것이다. 처음 이 프로젝트를 시작할 때와 동일한 환경을 유지한다고 볼 수 없기 때문에 clock마다 inference time을 다시 측정해서 기존 데이터와 비교할 필요가 있다.

계획

- Inference time을 다시 측정하고 기존 데이터와 비교. Inference server에 사용할 time2clock 테이블 재구축
- 새로 얻은 데이터로 SLO마다 다시 실험
- SLO의 효과에 대한 내용을 evaluation 첫 번째 section에 추가
- Evaluation을 여러 번 수행한 결과로 업데이트