0516-0522

Single GPU에 대한 SLO test

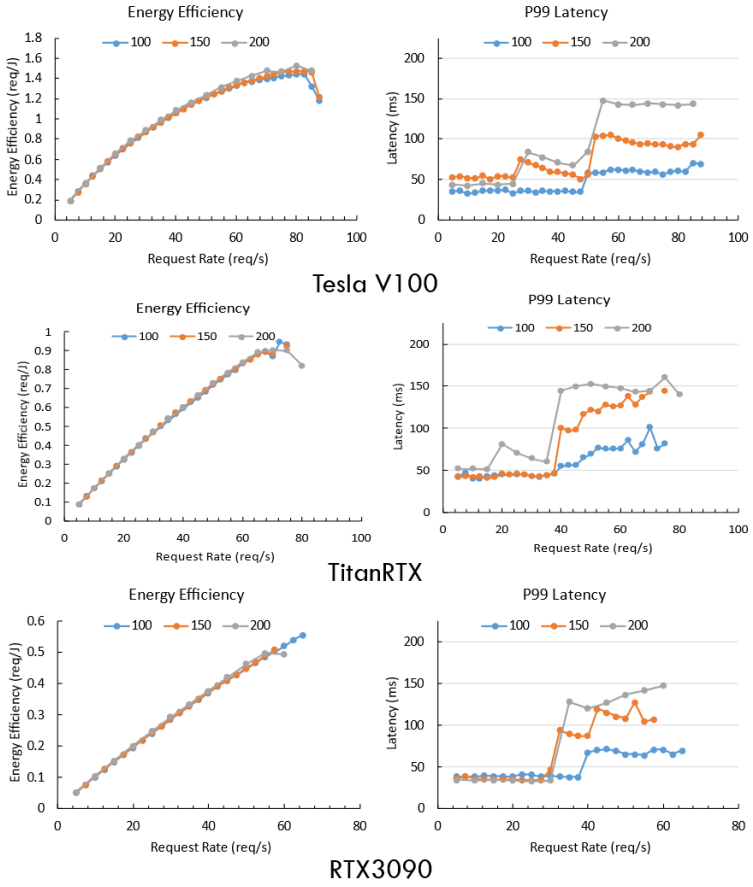


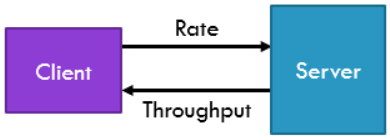
Figure 1. Inception Model

- Multi-GPU에서는 고려할 요소들이 많기 때문에 single-GPU에서 SLO 변화에 대한 효과를 확인해 보았음.

- Figure 1.은 Inception 모델에 대해서 3 종류의 GPU를 사용한 inference 에너지 효율성과 latency 결과이다. 200ms는 기존 논문 evaluation할 때 사용한 time2clock 데이터를 사용하고, 100ms, 150ms이 새로 측정된 것이다. SLO의 변화에 대해서 에너지 효율성에서는 큰 차이가 발생하지 않는다. 오히려 SLO에 여유가 생길수록 최대 throughput이 낮은 경향을 보인다. 에너지 효율성 그래프와 같이 보았을 때 마지막 포인트에서 효율성이 급격하게 낮아지는 것으로 보아 latency를 맞추기 위해서 매우 높은 클럭을 사용하고 있다. GPU 클럭이 제대로 설정되고 있지 않음을 알 수 있다.

- 200ms의 경우에는 지금 200ms의 90% 수준인 180ms 부근에서 latency가 형성되는 것이 아니라 150ms부근에서 형성되고 있다. 클럭 설정이 잘못되고 있음을 알 수 있다.

SLO가 미치는 영향



- 오른쪽 그림과 같이 client는 지정된 request rate로(동일한 request interval) 랜덤 데이터로 구성된 request를 server로 전달한다. Server는 이것을 받아서 처리하고, 우리는 서버에서 처리되는 request의 양을 통해서 throughput을 측정하고 있다.
- 실험에서는 rate를 고정하고 입력된 rate만큼의 throughput을 낼 수 있어야 하며, 이 때 latency를 위반하지 않아야 한다. 그렇기 때문에 request가 처리되는 시간은 request rate에 의해 결정되며 이것은 SLO에는 독립적이다.
- Inference time이 같으면 GPU의 클럭 또한 같다. 따라서 최대 throughput이 같아야 하며 에너지 효율성이 최대가 되는 지점도 동일해야 한다.
- Single-GPU 실험에서 에너지 효율성이 조금씩 차이가 발생하는 것은 SLO에 따라서 설정할 수 있는 GPU 클럭 후보군이 서로 다르기 때문에 GPU가 소비하는 에너지에서 미세한 차이를 보이기 때문이다.

계획

- SLO의 효과에 대한 내용을 evaluation 첫 번째 section에 추가
- Evaluation을 여러 번 수행한 결과로 업데이트