0418-0424

Target SLO의 변화에 따라서 각 모델들의 request rate에 따른 99th percentile latency 측정

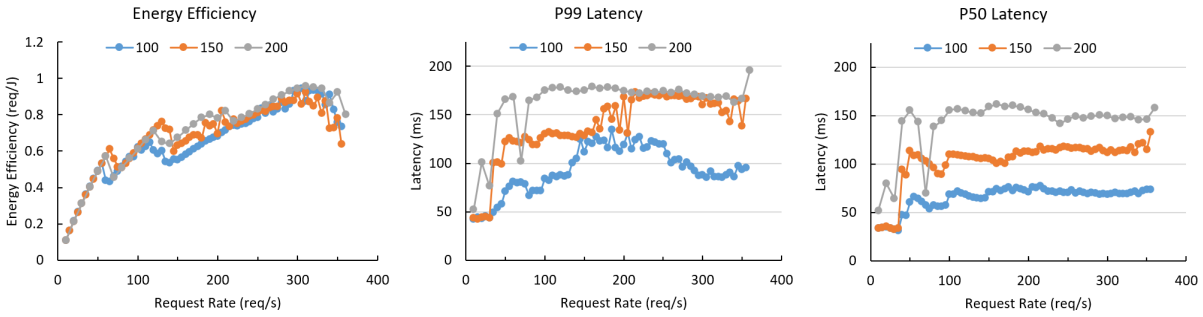
- Models: inception_v3, ResNet50, ResNet152, VGG19, BERT QA
- GPUs: Tesla V100, Titan RTX, RTX 3090
- Target SLO: 100ms, 150ms, (50ms의 경우는 일단 실험에서 제외)

실험 방법

- 모든 모델에 대해서 GPU의 모든 클럭에 대해서 inference time을 측정한 데이터를 바탕으로 알고리즘의 time2clock()을 위한 큐 길이별 요구되는 GPU클럭을 모두 계산. (~/time2clock/scale.py에서 확인가능)
- 빠르게 실험을 진행하기 위해서 5 rps 간격으로 100초동안 측정.
- 오류가 나는 경우는 일단 무시하고 넘어감.

실험 결과(TitanRTX-Inception 표시)

- 100, 150, 200ms SLO에 따른 latency, energy efficiency 비교. 200ms는 논문에 사용했던 데이터



- **Energy efficiency**
  - SLO가 낮아질수록 낮은 에너지 효율성을 가지는 구간이 늘어남
  - 특히 100 ms의 경우에는 120~140 rps구간에서 2→ 4개로 GPU 수가 빠르게 증가하는 것을 확인
  - 따라서 100~200 사이의 경우 100<150<200 순서로 에너지 효율성이 좋음

- **Latency**
  - 일단 낮은 rps에서는 목표로하는 latency를 잘 맞춰줌
  - 하지만 특정 구간(예를 들어, GPU를 4개 사용하는 경우 등)에서 tail latency가 조절이 안되는 결과를 보임.
  - (참고)P50의 경우 100ms부터 순서대로 평균 65.51ms, 103.6ms, 141.7ms

  Latency가 만족되지 않기때문에 실험을 다시 해야함

다음 주 계획

- 왜 중간 rate에서 latency violation이 발생하는지 확인 (다른 모델, GPU에서도 발생)
- 추가로 실험 진행. 높은 rate에 대해서는 기존보다 길게 측정 윈도우 설정
- Approach 파트 수정