0613-0619

Inference 시간 – GPU clock 변환

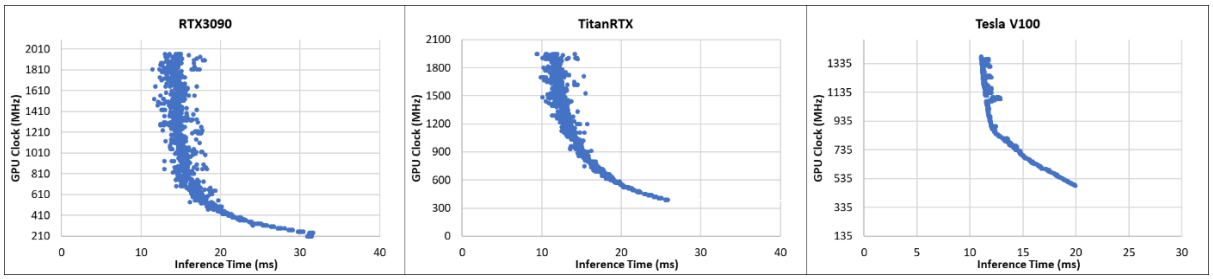

Figure 1. Inference time to GPU clock

- Figure 1은 GPU clock을 고정한 상태로 10번씩 최저 clock부터 최대 clock까지 측정한 inception모델의 inference time을 나타낸 그래프이다. 이 때 perf_analyzer에서 concurrency 옵션의 값을 2로 주어 측정했다. Inference time은 GPU로 data copy하는 시간과 GPU에서 inference가 수행되는 시간의 합이다.
- 주목해볼 점은 RTX3090, TitanRTX, Tesla V100 순서대로 편차가 줄어든다는 것이다. 왼쪽 GPU 부터 inference time의 표준편차가 2.39, 1.73, 0.49ms이다. 기존에는 같은 클럭에서 측정된 값들의 평균을 사용했다. 하지만 편차가 큰 3090 GPU를 사용하게 되면 예상했던 시간보다 더 오래 걸릴 수 있다. 따라서 GPU에 따라서 우리가 예상했던 것보다 더 느리게 처리되는 request가 존재하고 있다.
- time2clock 테이블 구성을 위해서 기존에 사용하면 다항식에서 반비례함수로 변경했다. y = 1/(mx+b)의 꼴로 x와 1/y에 대해서 linear regression하면 m과 b를 구할 수 있다.
- Figure 2.에서 TitanRTX-Inception 조합의 regression 결과를 볼 수 있다. 각 클럭에서 하위 20%, 상위 20%의 값을 제외하고 평균을 내어 사용했다.

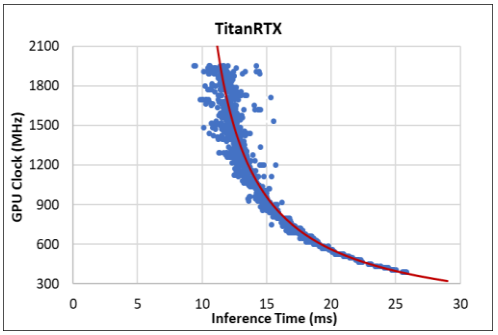

Figure 2. Regression 결과

Scheduler thread 수정

- GPU 클럭 조절은 enqueue 하기 전, dequeue 한 후에 이루어졌다. 이상적으로 모든 request가 예상하는 시간 내에 실행된다면 문제가 없다. 하지만 Figure 1.에서 살펴본 것처럼 같은 클럭에서도 inference time은 편차가 존재하기 때문에 enqueue과정에서 조절되는 클럭으로는

현재 queue에 있거나 미래에 들어올 request들의 latency가 점점 밀려 위반되는 상황이 발생할 수 있다.

- 우리는 dequeue하여 GPU로 dispatch하기 전에 해당 request의 남은 시간을 고려해서 클럭을 조절하는 과정을 추가했다. 각 request에는 queue에 들어갈 때 시간을 기록하고 있으므로 queue delay 시간을 구하고 현재 GPU 클럭으로 동작할 때 기대되는 inference time을 더한다. 이 때 target SLO와 비교하여 남은 시간이 없다면 클럭을 상승시킨다.

- GPU 클럭을 최대로 상승시키면 빠르게 latency가 밀리는 현상을 해결할 수 있다. 하지만 request rate가 낮은 경우, 낮은 클럭을 유지하다가 갑자기 큰 폭으로 클럭이 상승하여 에너지 효율성에서 손해를 본다. 현재 클럭으로 latency가 위반되는 상황을 살펴본 결과 대부분 request rate가 높을 때 발생하며, 클럭을 한 단계 증가시키는 것(queue 사이즈가 +1되었을 때 설정되는 클럭)으로 최대 클럭이 되었다. 따라서 위반될 것으로 감지되면 클럭을 한 단계 증가시키도록 수정했다.

- Figure 3.는 새롭게 측정한 time2clock 테이블을 기반으로 서로 다른 target SLO에 대해서 실험한 결과이다. V100의 경우에 이전 결과보다 더 target SLO에 가깝게 tail latency가 형성되고 있다. 하지만 inference time편차가 큰 다른 두 GPU 모델에 대해서는 정상적으로 동작하지 않는다. 특히 TitanRTX에서는 더 여유로운 SLO에서 throughput이 낮아지는 결과를 얻었다. 이것은 필요한 클럭보다 낮게 설정되고 있음을 의미한다.

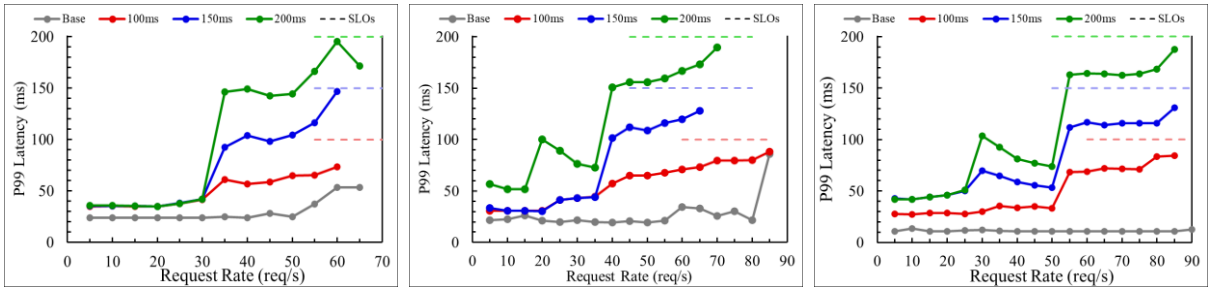

Figure 3. 서로 다른 target SLO에 따른 tail latency (왼쪽부터 RTX3090, TitanRTX, TeslaV100)

계획

- Latency가 지켜지지 않은 원인 분석. 클럭 조절이 필요한 만큼 적절한 시간에 수행되는지 위주로 확인

0613-0619