Resource Efficiency for AI

영남대학교
차세대 컴퓨터 시스템 연구실
석사과정생 이원호

Decoupled SSD: Reducing data movement on NAND-based Flash SSD

IEEE Computer Architecture Letters, 2021

차세대 컴퓨터 시스템 연구실

이원호

제안 배경

- 현대 NAND Flash memory 기반 SSD는 대량의 I/O 요청을 위해 멀티 채널, 멀티 메모리 칩 등을 포함한 여러 병렬화를 통해 높은 대역폭을 제공하도록 설계함

- SSD 시스템은 I/O 요청 뿐 아니라 플래시 메모리 관리 프로세스들 (e.g garbage collection)도 사용
  - 두 처리 과정은 시스템 자원을 공유함 → Conflict의 발생 가능성 → 전체 성능 저하

제안 사항 (Decoupled SSD)

- Front-end
  - Cores
  - DRAM
  - System bus
- Back-end
  - Flash memory-chip

- Error correction logic
  + flash controller
  + interconnect router
- I/O 경로와 garbage collection 경로를 구분지을 수 있음

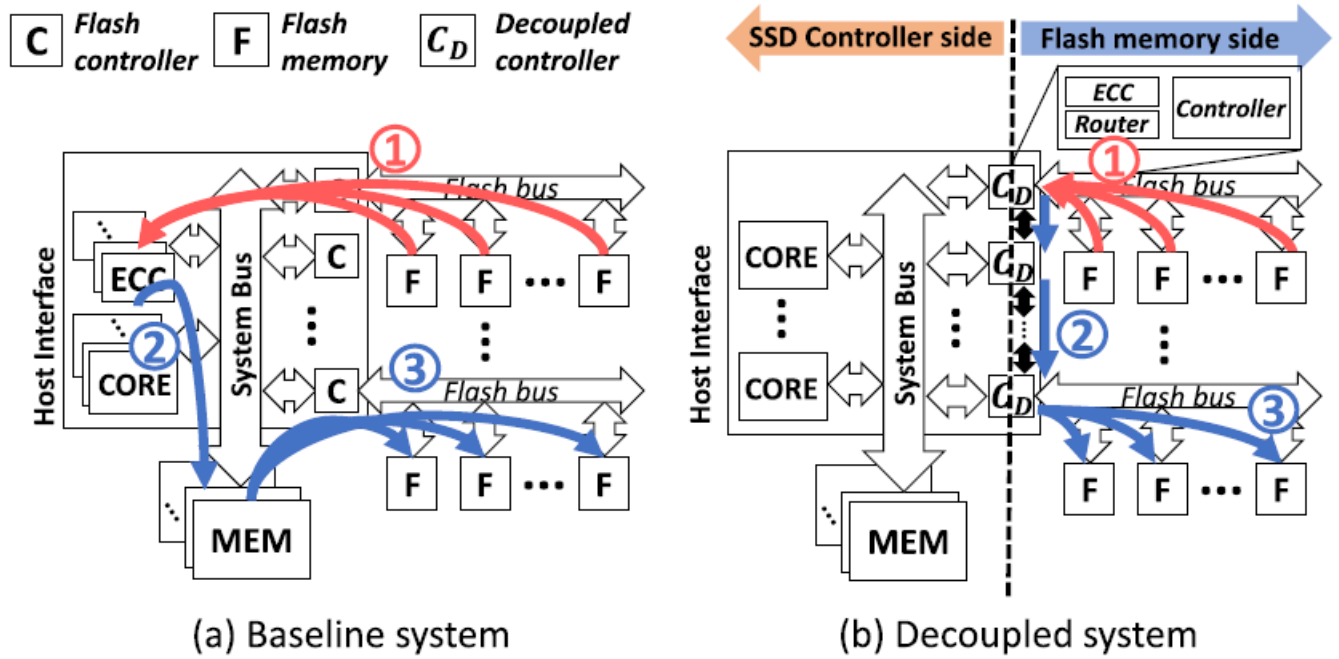


Fig. 1. (a) Execution path of garbage collection (GC) in current SSD system with heavy accesses of system resources, and (b) Execution path of GC without accessing system-bus and DRAM.

제안 사항 (Global copy-back)

- I/O 경로와 garbage collection 경로를 구분해 전역 copy-back 명령어의 사용이 가능해짐
- 여러 flash bus channel을 활용할 수 있게 해줌
- 병렬성 확장을 통한 성능 향상

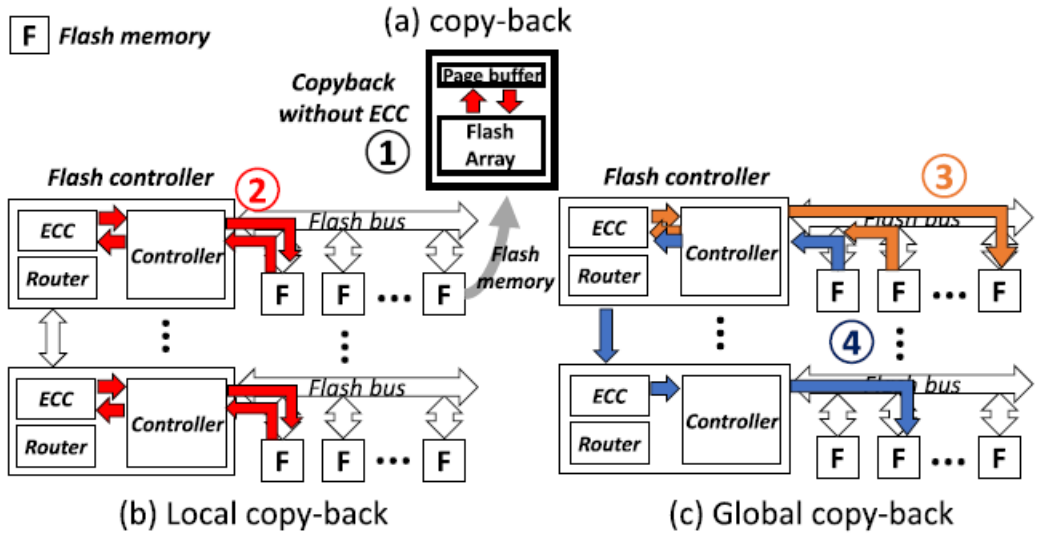


Fig. 3. (a) Local copy-back of the data within same chip either non-error corrected using copy-back command or error corrected using data transferring, (b) Global copy transfers the data from a flash chip to any flash chip either in the same channel or different channel.

HOME: A holistic GPU memory management framework for deep learning

IEEE Transactions on Computers, 2022

차세대 컴퓨터 시스템 연구실

이원호

제안 배경

- 최근 DNN 모델 크기의 증가로 인해 GPU의 메모리 부족 문제가 빈번히 발생
- DNN 모델 학습 과정에서 순전파 과정에서 각 레이어의 출력 (특징 맵)이 역전파 과정에서 가중치 갱신을 위해 **재사용**
- 이를 위해 효율적인 텐서 배치가 필요함

제안 배경

- GPU 메모리 부족 문제 해결 방안 (텐서 배치 정책)

- Data Compression
  - 특징 맵을 압축하는 방식
  - Memory footprint 감소 / 모델 성능 저하 발생
- Data Swapping
  - CPU 메모리를 GPU의 외부 메모리로 간주하고 사용
- Data Recomputation
  - 특징 맵이 필요할 때 재계산

- 기존의 작업들은 모델의 일부만 보고 텐서 배치 정책 중 선택

제안 사항 (HOME)

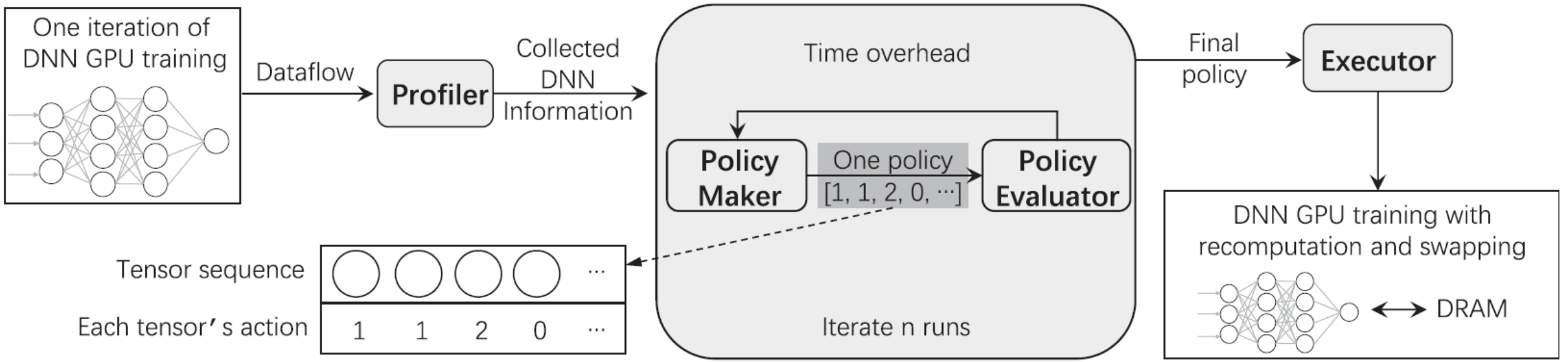


Fig. 3. The system overview of the HOME framework.

제안 사항 (HOME)

- 모델의 전체 사항을 보고 요구 메모리를 파악함
- 파악한 정보를 기반으로 각 텐서에 대해 swapping, recomputation, retaining 중 하나를 선택

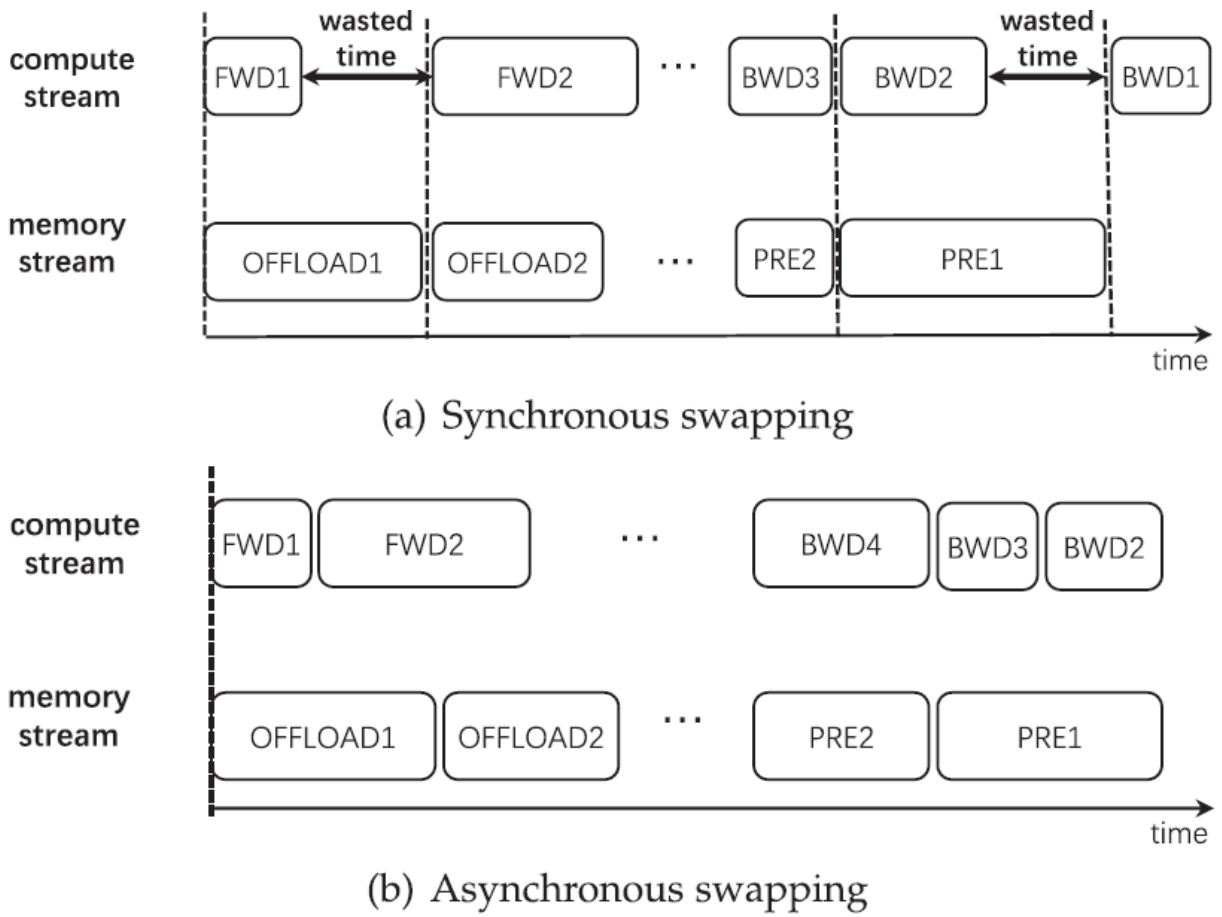


Fig. 1. Examples of layer-wise synchronous and asynchronous data swapping. FWD1 represents the forward propagation of the first layer in the DNN model, BWD1 represents the backward propagation of the first layer in the DNN model.

제안 사항 (HOME)

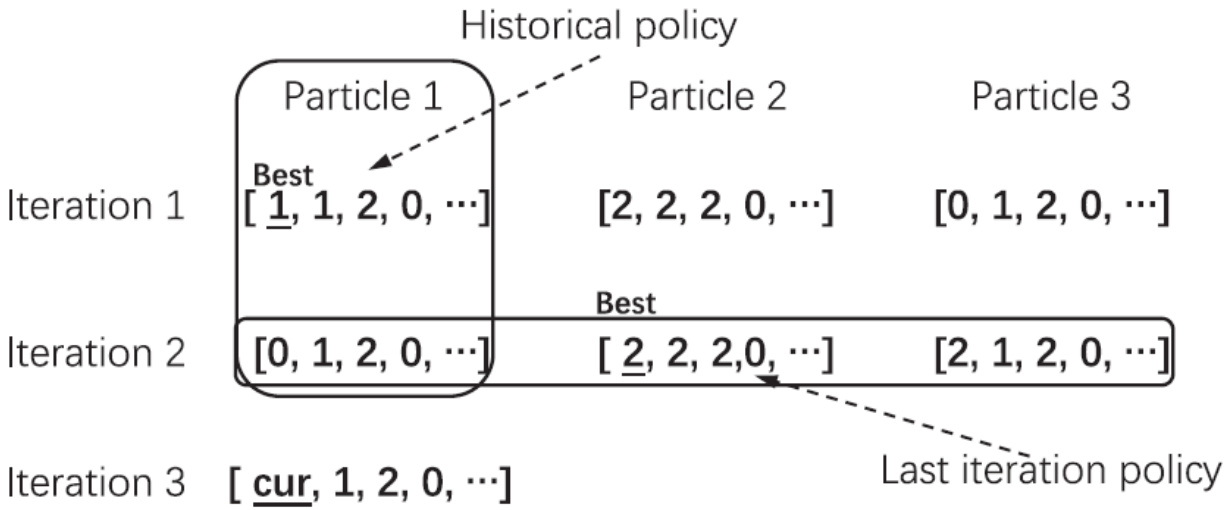


Fig. 5. An example of tensor placement policy update. The historical policy set: the best historical policies for all particles (each particle has its corresponding policy). The last iteration policy: the best placement policy among all particle policies in the last iteration (only one).

Algorithm 3. Particles Update

Require: m: the number of particles in one iteration; n: the numbers of layers; $policy[m][n]$: a list of actions given to each tensor;

1: $P_1[n] = getLastIterBest()$ ▷ get the best policy in the last interation
2: **for** $i = 1, 2, ..., m$ **do** ▷ m particles
3: $P_2[n] = getHistoricalBest(i)$ ▷ get the best historical policy of particle i
4: **for** $j = 1, 2, ..., n$ **do** ▷ n layers (tensors)
5: $policy[i][j] = P_1[j]$ ▷ set the action for tensor j
6: $O_1 \leftarrow$ evaluate($particle_i$) ▷ call Algorithm 2
7: $policy[i][j] = P_2[j]$
8: $O_2 \leftarrow$ evaluate($particle_i$)
9: **if** $O_1 < O_2$ **then**
10: $policy[i][j] = P_1[j]$ ▷ update policy
11: **else**
12: $policy[i][j] = P_2[j]$
13: **end if**
14: **end for**
15: **end for**

Zico: Efficient GPU Memory Sharing for Concurrent DNN Training

USENIX Annual Technical Conference , 2021

차세대 컴퓨터 시스템 연구실

이원호

제안 배경

- 여러 딥러닝 모델의 동시 학습을 위해 GPU 메모리 공유가 필수적

- DNN 모델 학습 과정에서 순전파 과정에서 각 레이어의 출력 (특징 맵)이 역전파 과정에서 가중치 갱신을 위해 **재사용**
  - 메모리 소비 패턴이 순환 구조를 가짐
  - 순전파 과정에서 점진적으로 증가했다가, 역전파 과정에서 점진적으로 감소함

- 생성된 중간 데이터 (특징 맵)들의 공유를 통한 전체 memory footprint 감소
  - 효율적인 공유 패턴의 필요성

제안 배경 (메모리 사용 패턴)

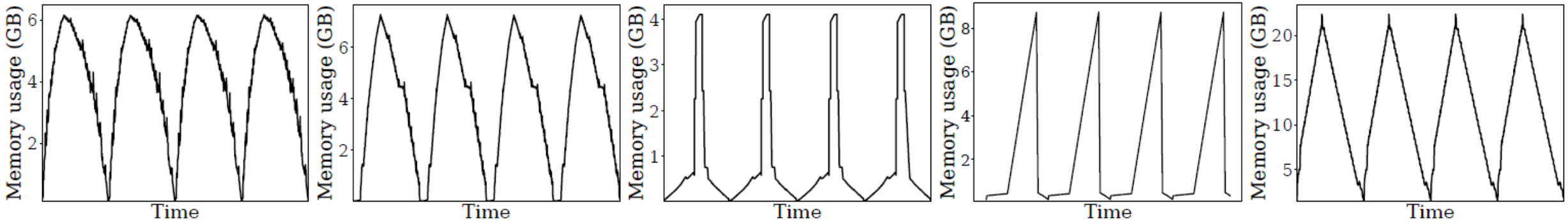
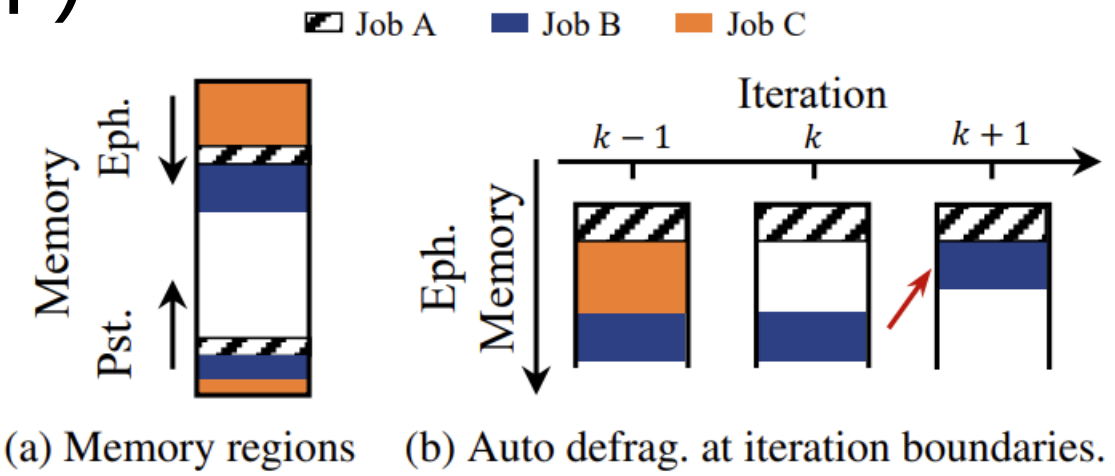


Figure 2: Memory usage patterns for different DNN models over time.

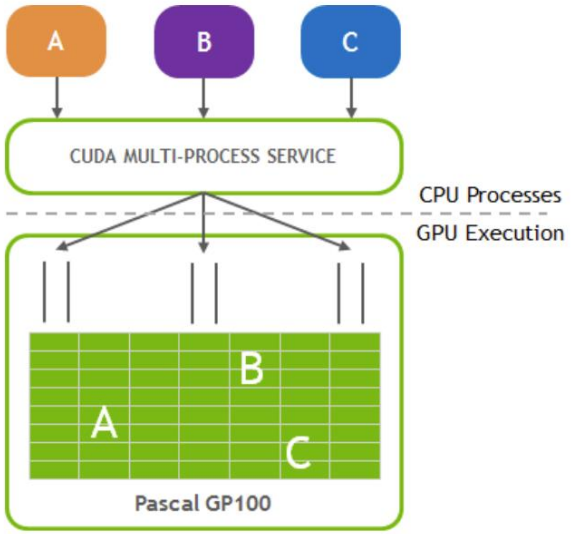
제안 배경 (GPU 공유)

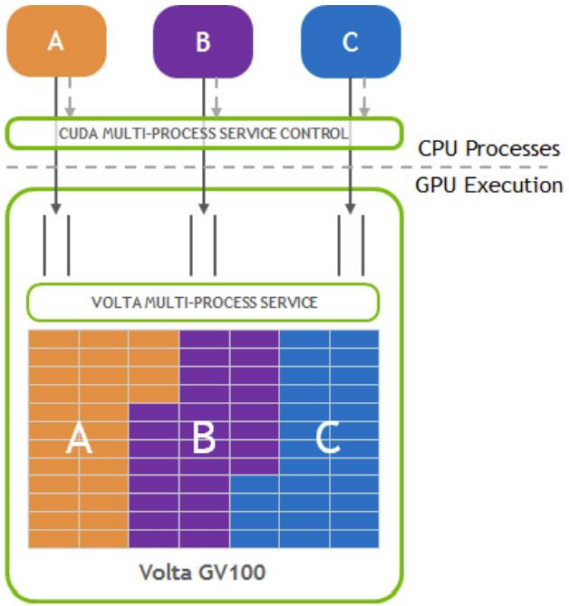
- Temporal Sharing
  - GPU 자원 공유
  - 단일 GPU에서 한 학습에 대해 세분화함
  - 데이터 병렬화 정도가 낮은 경우 낮은 GPU 사용량을 일으킴

- Spatial Sharing
  - GPU 컴퓨팅 유닛 공유
  - Temporal sharing보다 더 높은 처리량을 제공함
  - 동시 작업들의 작업 셋 크기에 따라 적용에 제약이 있음



(a) Memory regions (b) Auto defrag. at iteration boundaries.





제안 사항(Zico)

- 동시 학습을 위한 메모리 공유를 진행하는 DNN 플랫폼

1. 학습 작업의 연산 과정을 모니터링
   - 모니터링 정보를 기준으로 DNN 커널을 위한 메모리 할당과 해제 (GPU 관점 메모리)
2. 런타임 통합, 작업 스케줄러 실행
3. 전체 GPU 공간을 elastic memory pool로 구성

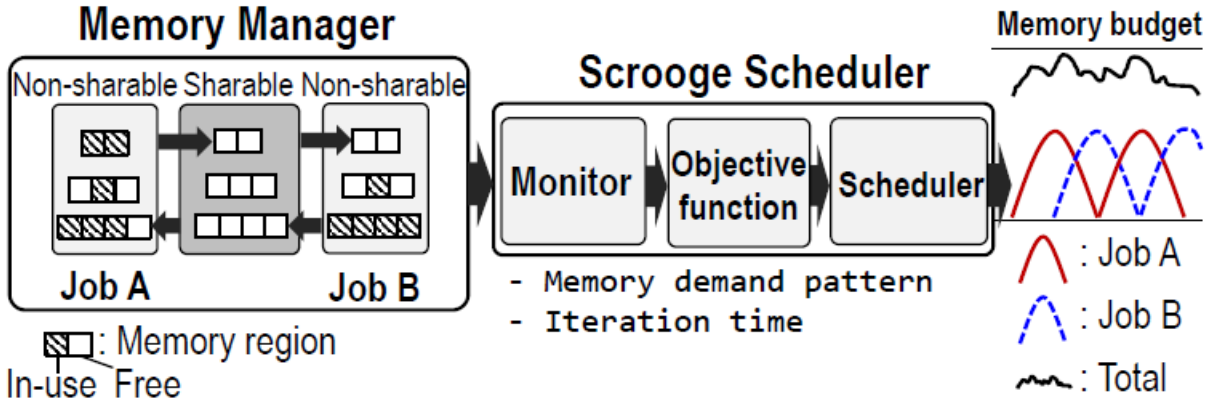


Figure 4: System architecture in Zico.

Multi-resource interleaving for deep learning training

Proceedings of the ACM SIGCOMM 2022 Conference , 2022

차세대 컴퓨터 시스템 연구실

이원호

제안 배경

- 기존 딥러닝 스케줄러들은 딥러닝 학습에서 GPU가 병목 현상의 주된 요인이라는 가정
  - 실제 ResNet과 같은 초기 딥러닝 모델들은 GPU가 병목 현상을 일으킴

- 딥러닝 모델의 크기와 종류가 다양해짐에 따라 GPU가 병목현상의 주된 요인이 아닌 경우가 발생
  - 순전파와 역전파 과정 (GPU)
  - 강화학습 시뮬레이션 (CPU)
  - 분산 학습 (Network I/O)
  - IoT 기기를 위한 소형 모델 학습 (Storage I/O)
- **다른 자원들(Storage I/O, CPU, Network I/O)도 스케줄링의 고려사항**

제안 사항(Muri)

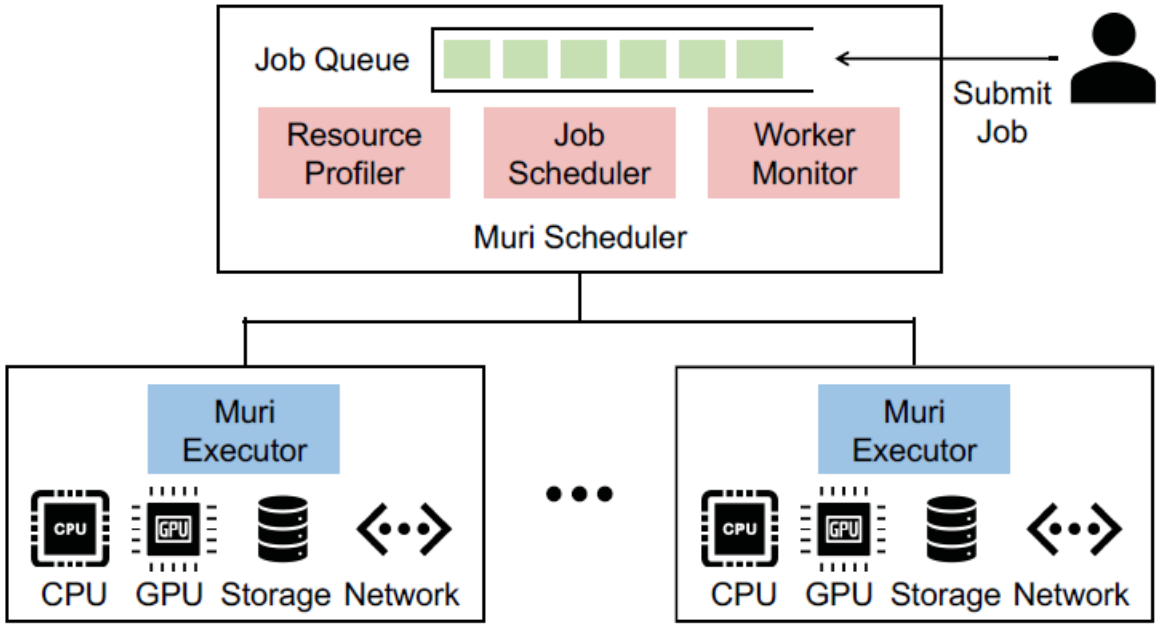


Figure 3: Muri architecture.

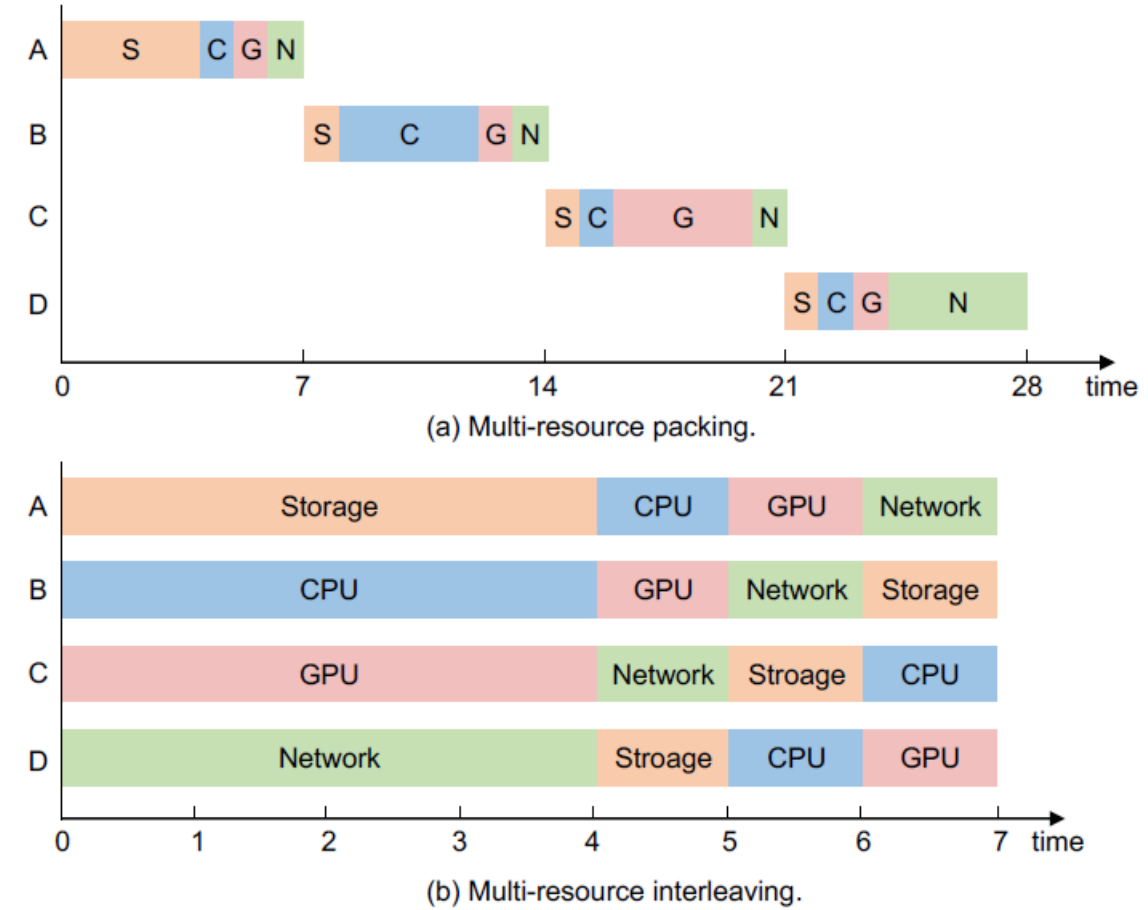


Figure 1: Benefits of multi-resource interleaving compared with multi-resource packing. (a) Multi-resource packing can not pack jobs when the peak usage of at least one resource type is high. (b) Multi-resource interleaving can interleave jobs over time to run multiple jobs concurrently.

제안 사항(Muri)

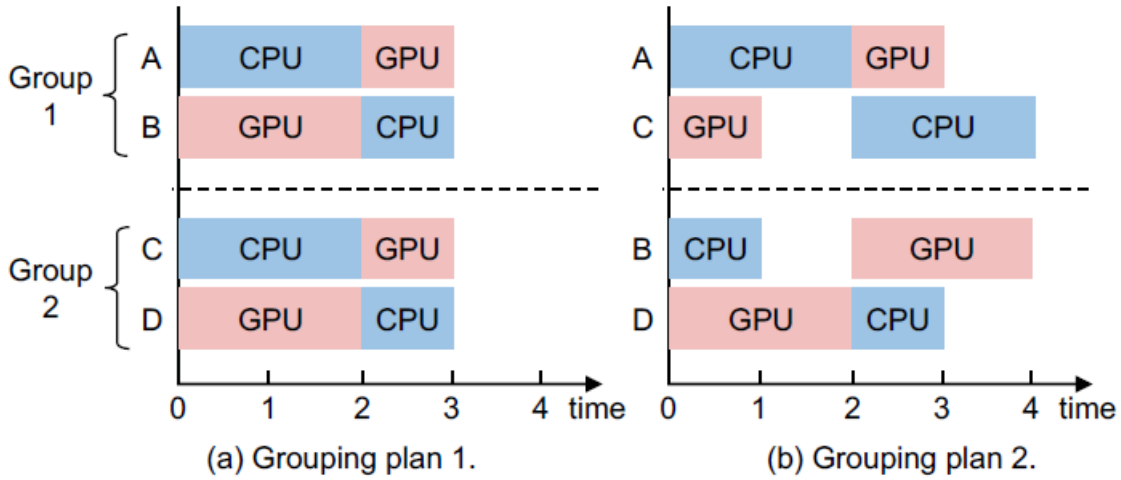


Figure 4: Impact of grouping plans on interleaving efficiency. Grouping plan 1 can perfectly overlap four jobs on two GPUs, while grouping plan 2 increases the per-iteration time.

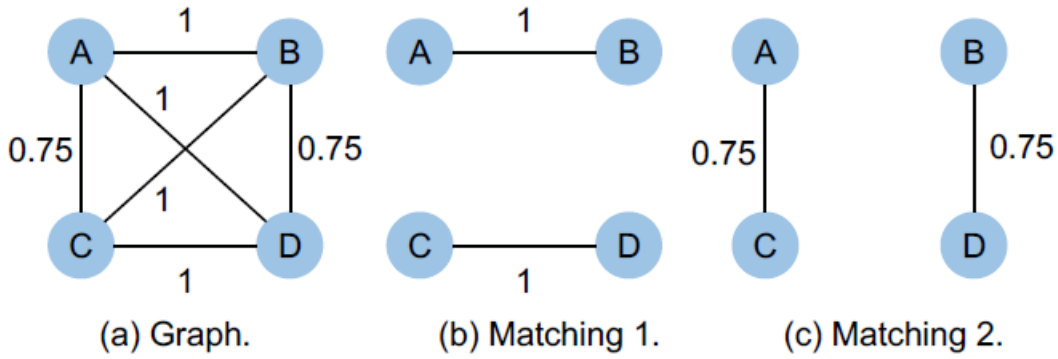


Figure 5: Grouping plans are computed with maximum weighted matching. The edge weights in (a) represent the interleaving efficiency for each pair of jobs. Grouping plan 1 and 2 in Figure 4 correspond to (b) and (c), respectively. Matching 1 in (b) has higher total weights than matching 2 in (c).

제안 사항(Muri)

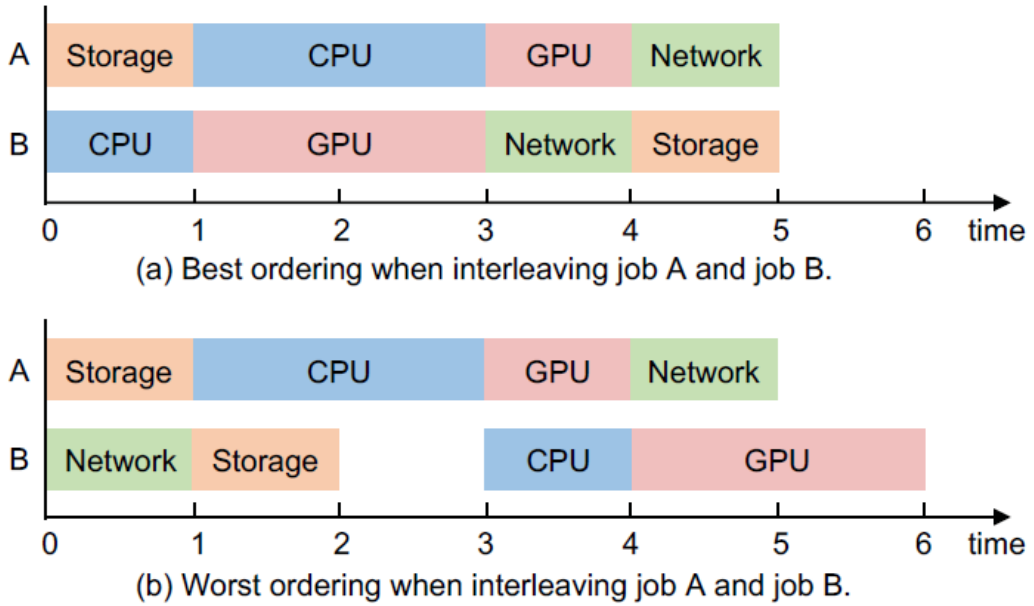


Figure 6: The ordering of jobs affects the interleaving efficiency under multiple resource types.

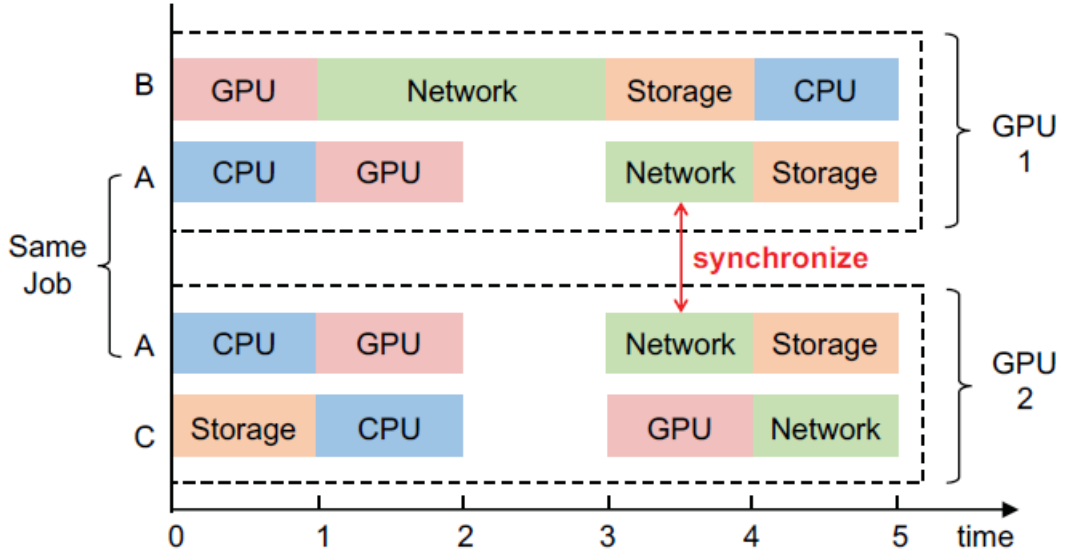


Figure 7: Interaction between inter-job interleaving and intra-job synchronization. The per-iteration time of C is increased.

Reference

- Decoupled SSD: Reducing data movement on NAND-based Flash SSD
  IEEE Computer Architecture Letters, 2021

- HOME: A holistic GPU memory management framework for deep learning
  IEEE Transactions on Computers, 2022

- Zico: Efficient GPU Memory Sharing for Concurrent DNN Training, USENIX Annual Technical Conference , 2021
  - Salus: Fine-Grained GPU Sharing Primitives for Deep Learning Applications. MLSys , 2020
  - NVIDIA Corp. Multi-process service.2020

- Multi-resource interleaving for deep learning training, Proceedings of the ACM SIGCOMM 2022 Conference , 2022