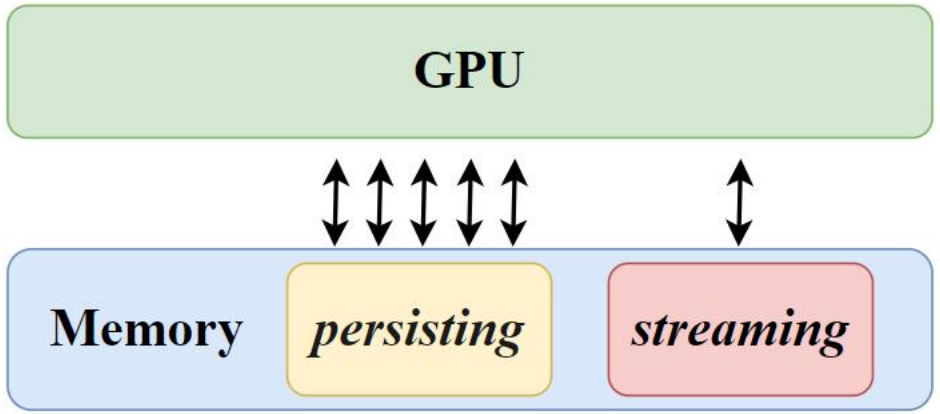
AUTOSCRATCH: ML-OPTIMIZED CACHE MANAGEMENT FOR INFERENCE-ORIENTED GPUS

Yaosheng Fu [1] **Evgeny Bolotin** [1] **Aamer Jaleel** [1] **Gal Dalal** [1] **Shie Mannor** [1] **Jacob Subag** [1] **Noam Korem** [1]
Michael Behar [1] **David Nellans** [1]

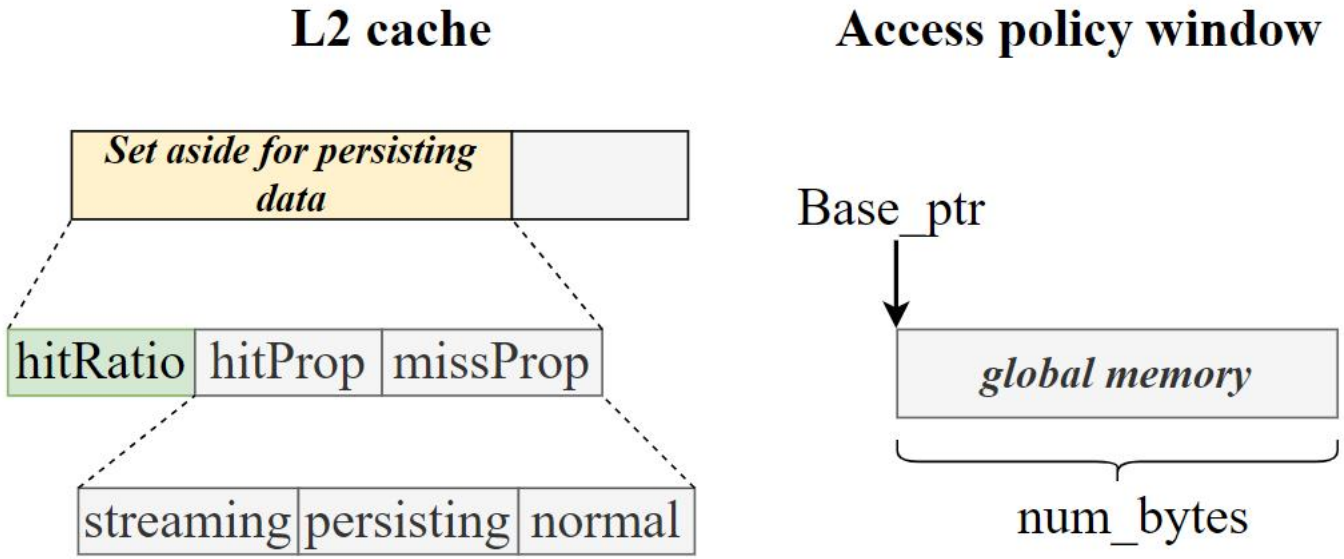
Proceedings of the 6th MLSys Conference

[1]NVIDIA.

BACKGROUND Device Memory L2 Access Management Introduced by Ampere

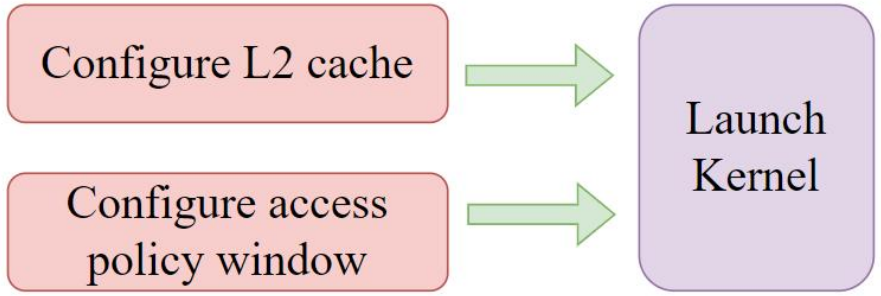


data access pattern



L2 access management

Basic workflow of L2 access management



https://docs.nvidia.cn/cuda/cuda-c-programming-guide/index.html#device-memory-l2-access-management

INTRODUCTION

Performance bottleneck

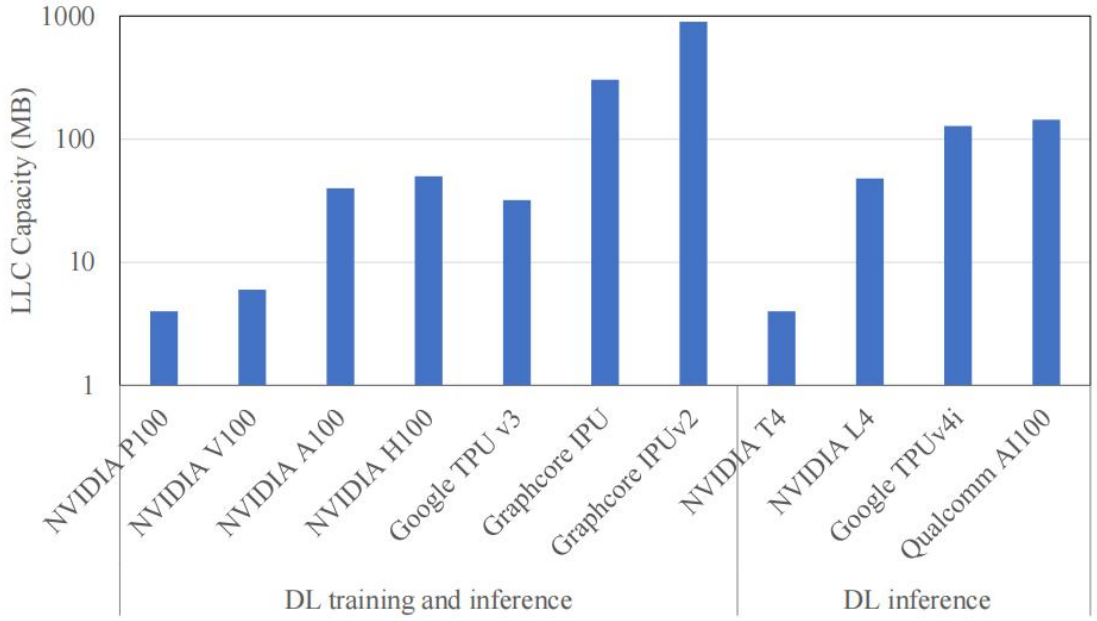
	FP16 throughput improvement	DRAM bandwidth improvement
V100	6x	0.25x
A100	2.6x	0.72x

Power consumption for HBM :
40 watts per 1TB/s bandwidth

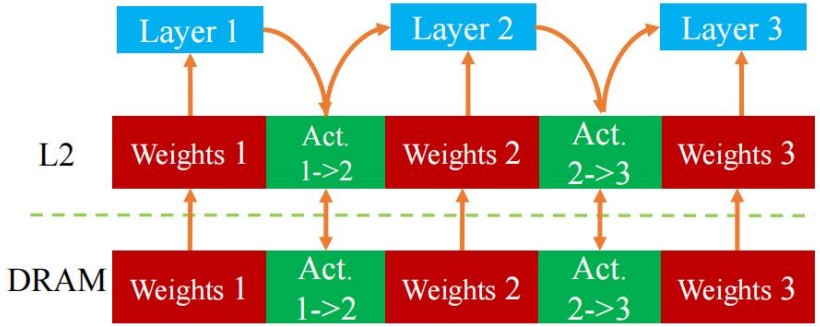
Reduce DRAM traffic <= Increase SRAM caches

Hardware managed VS Software managed

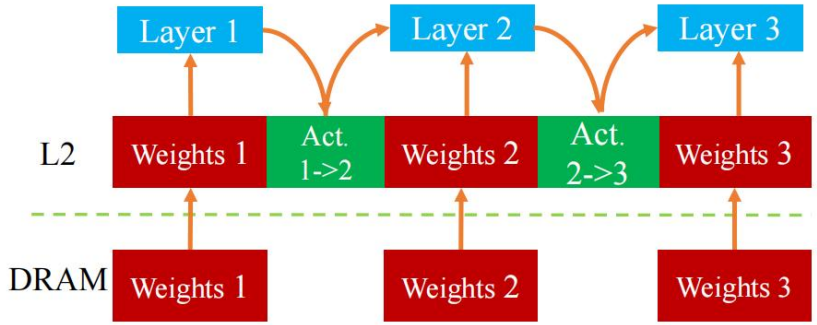
Programmer intuition VS Automatically discovers and optimizes



MOTIVATION



(a) Typically both activations and weights end up shuffling between L2 cache and DRAM in hardware-managed caches resulting in cache interference.



(b) Ideal reuse of L2 cache capacity results in cache resident activations, with only weights being fetched from DRAM, which are not shared between layers.

Figure 2. The memory access pattern for common DL inference workloads illustrating differing reuse patterns for weights and activations.

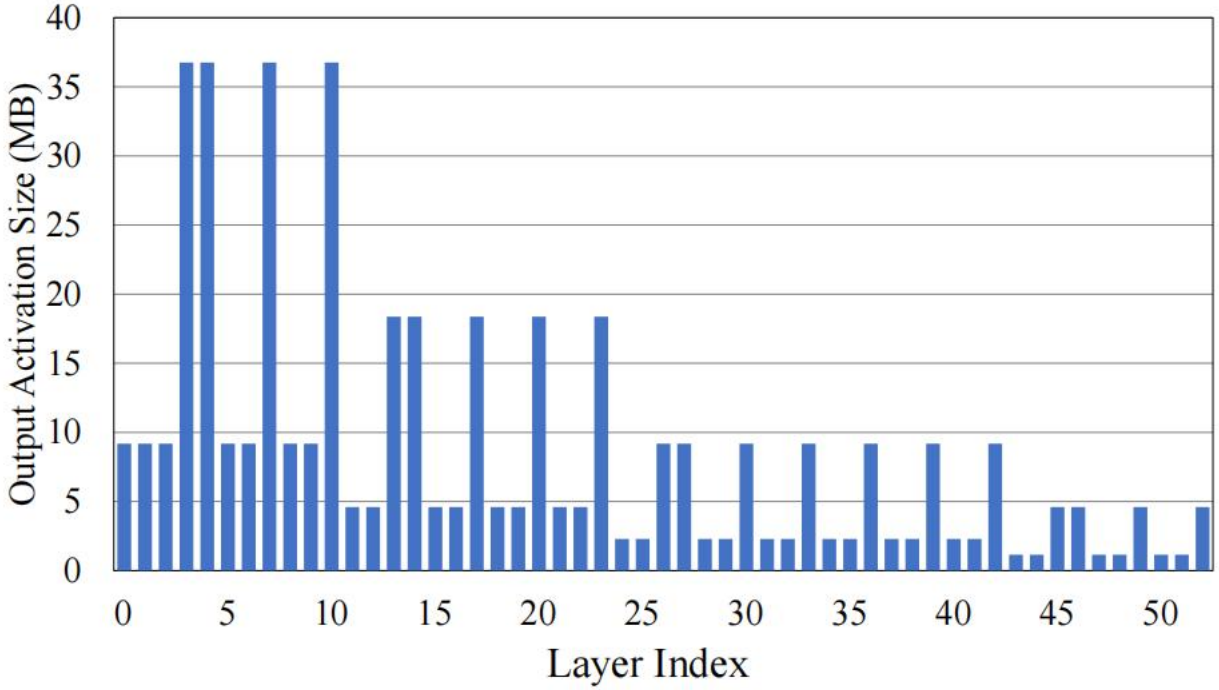


Figure 3. The per-layer activation sizes for int8 datatype in *resnet50* inference, with a batch size of 48. The largest per-layer activation size is less than 37MB.

AUTOSCRATCH FRAMEWORK

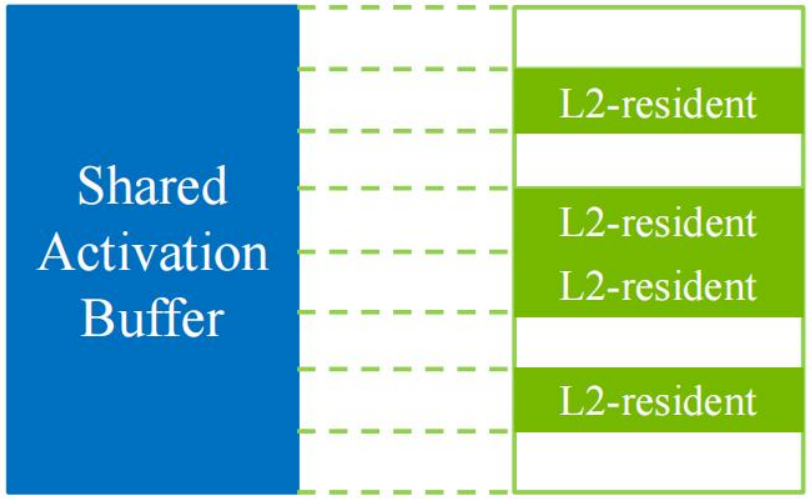


Figure 4. L2 residency selections within the shared activation buffer.

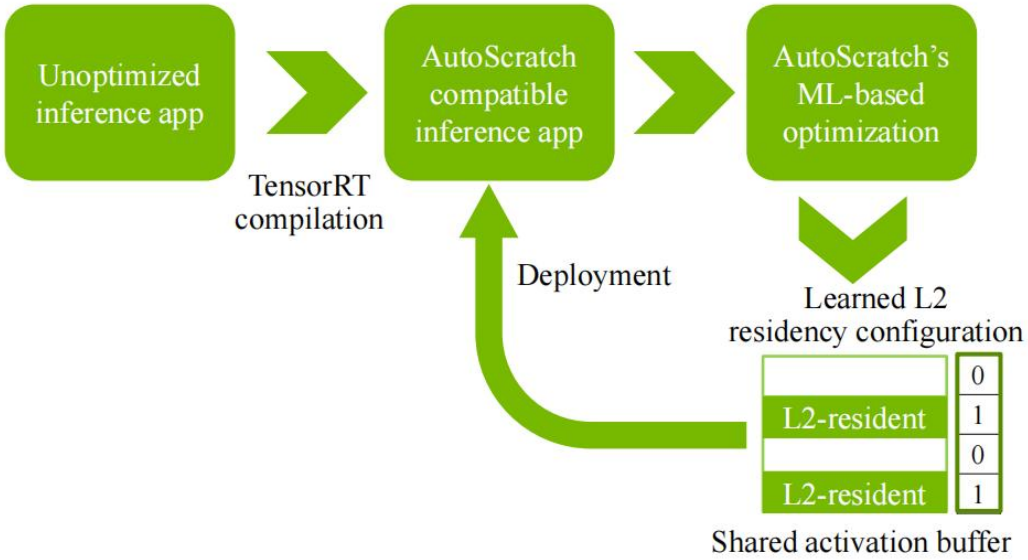


Figure 5. The high-level architecture and optimization flow of AutoScratch.

AUTOSCRATCH FRAMEWORK

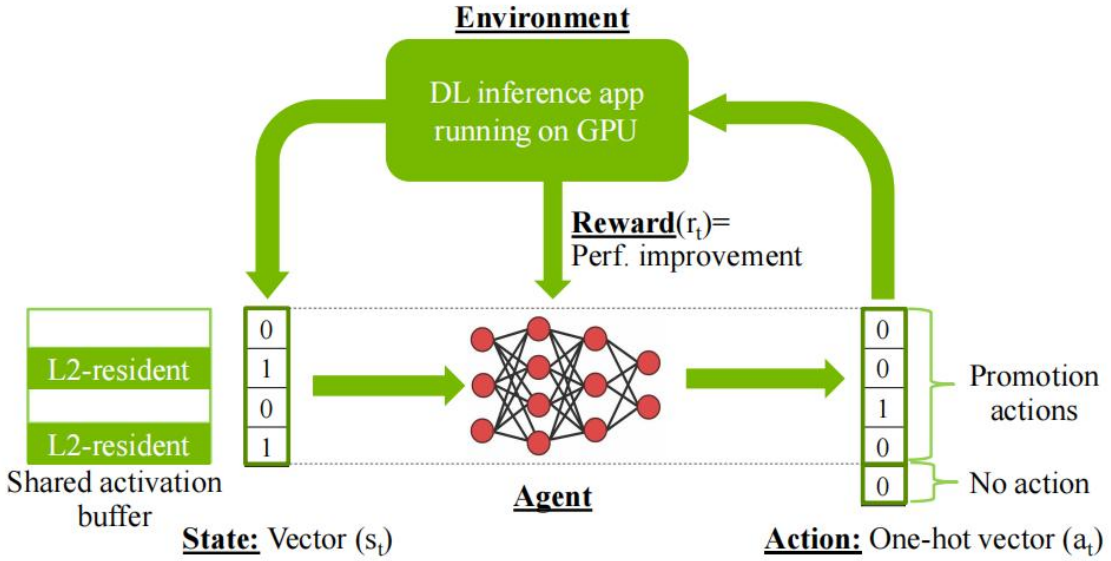


Figure 6. The AutoScratch-RL optimization framework for tuning GPU's L2 residency configuration.

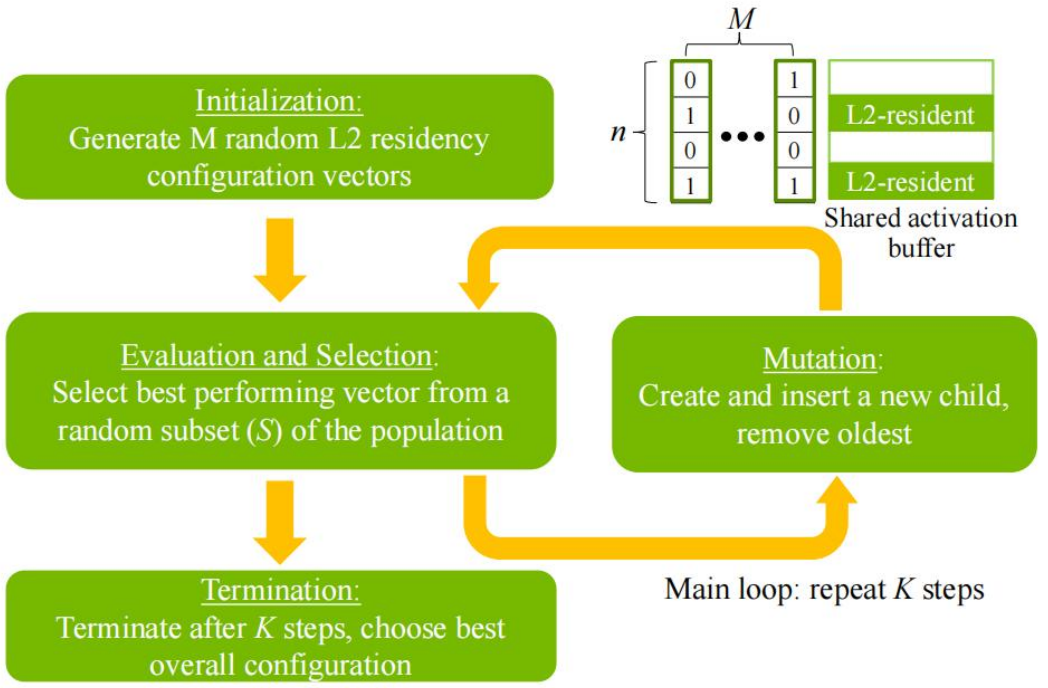
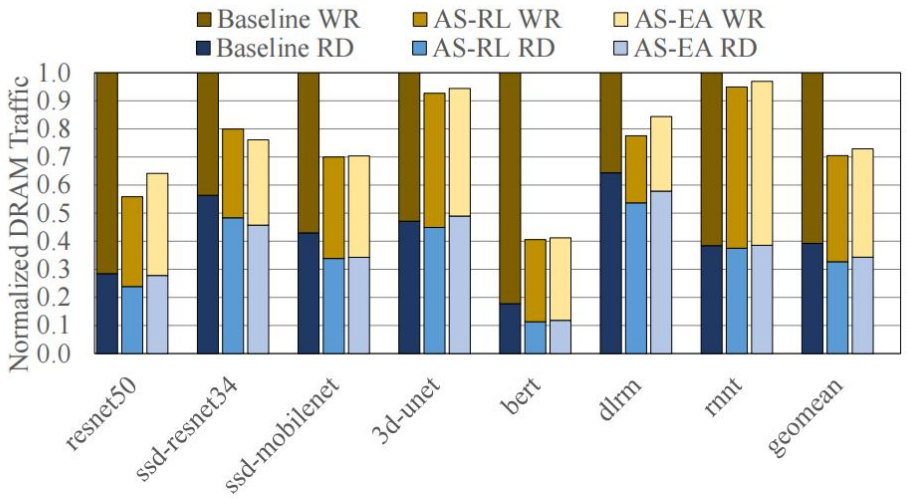


Figure 7. AutoScratch-EA with regularized evolutionary optimization for tuning GPU's L2 residency configuration.

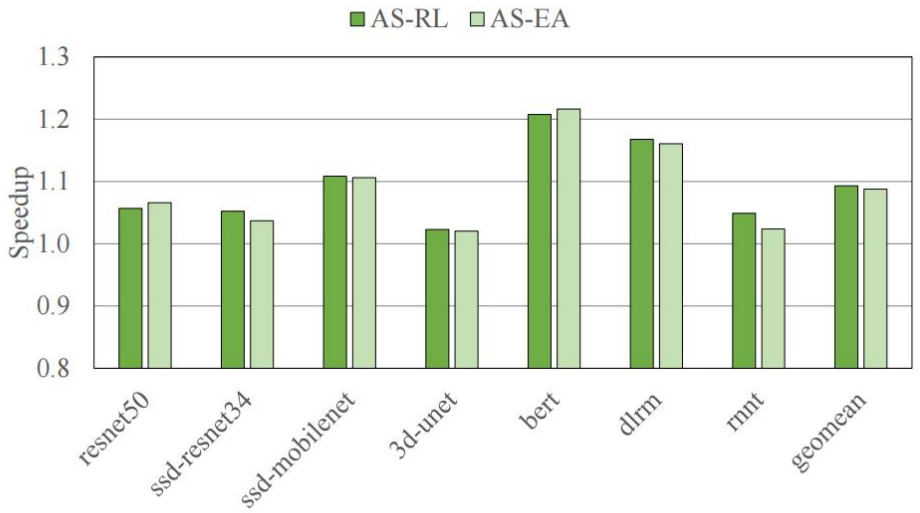
EVALUATION

Table 1. MLPerf inference benchmark settings in AutoScratch.

BENCHMARK	PRECISION	BATCH SIZE	ACTIVATION BUFFER SIZE (MB)
RESNET50	INT8	32	63
SSD-RESNET34	INT8	6	104
SSD-MOBILENET	INT8	64	140
3D-UNET	INT8	1	278
BERT	INT8	32	81
DLRM	INT8	51200	106
RNNT	FP16	2048	4175



(a) Offchip DRAM Traffic



(b) Performance Speedup

THANKS & QA