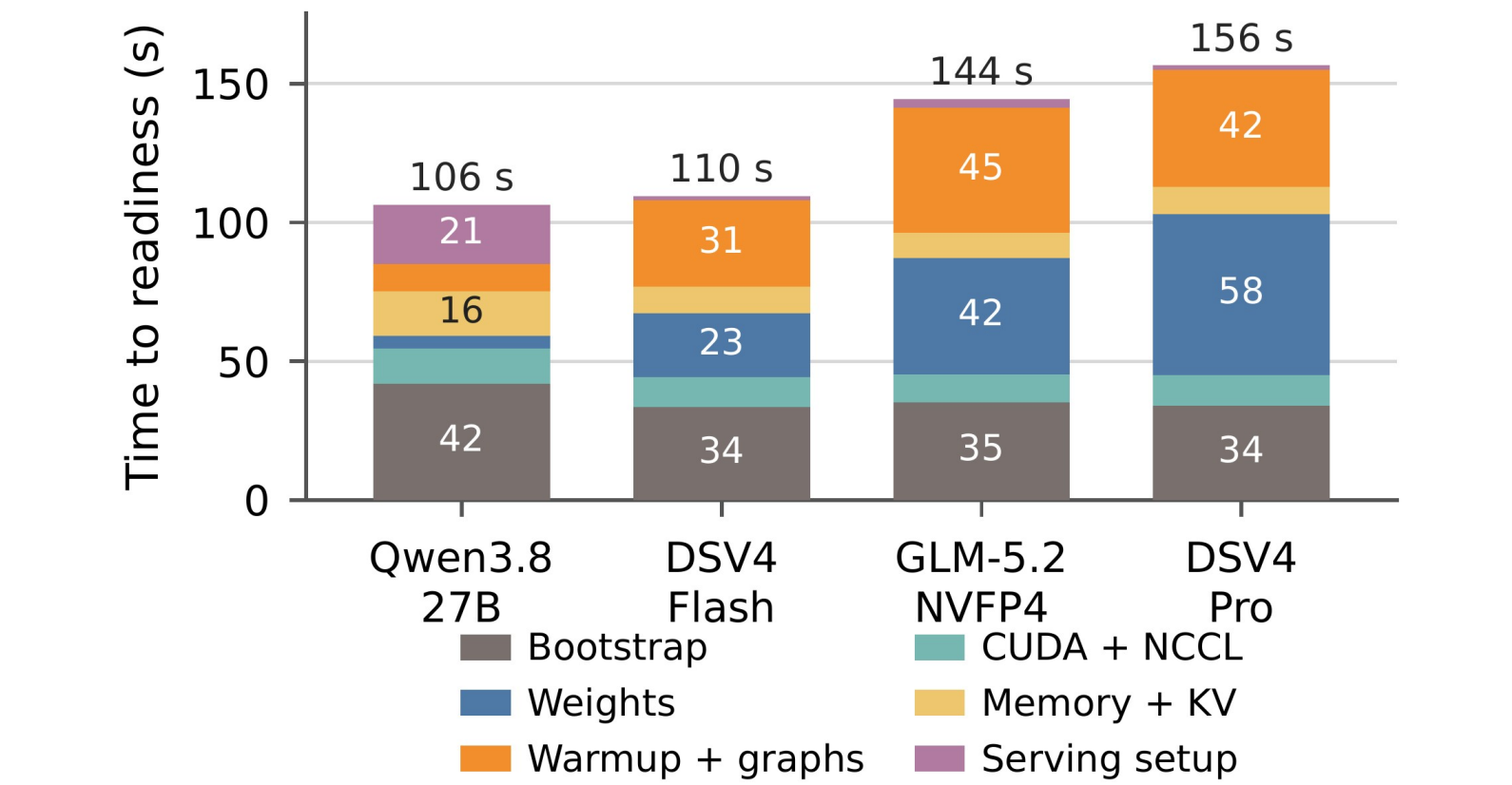


Schwinn Saereesitthipitak, Dan Feigin, Vikram Sharma Mailthody, Maksim Khadkevich NVIDIA

Recovery takes minutes



Restart time for vLLM 0.27.1 on 8×B200 GPUs (TP/TEP8). Compilation artifacts are pre-cached and weights pre-pulled into page cache.

- Recovering a failed replica requires restarting the inference engine. This takes several minutes even when all caches are warm.
- Over 18 weeks, one Dynamo cluster saw 5,762 in-place restarts with a median recovery of 3.5 minutes and 2,967 GPU-hours lost.
- Most failures are device-preserving. The engine process fails while the GPU and its resident allocations remain intact.
- Recovery should therefore preserve initialized engine state rather than reconstruct it.

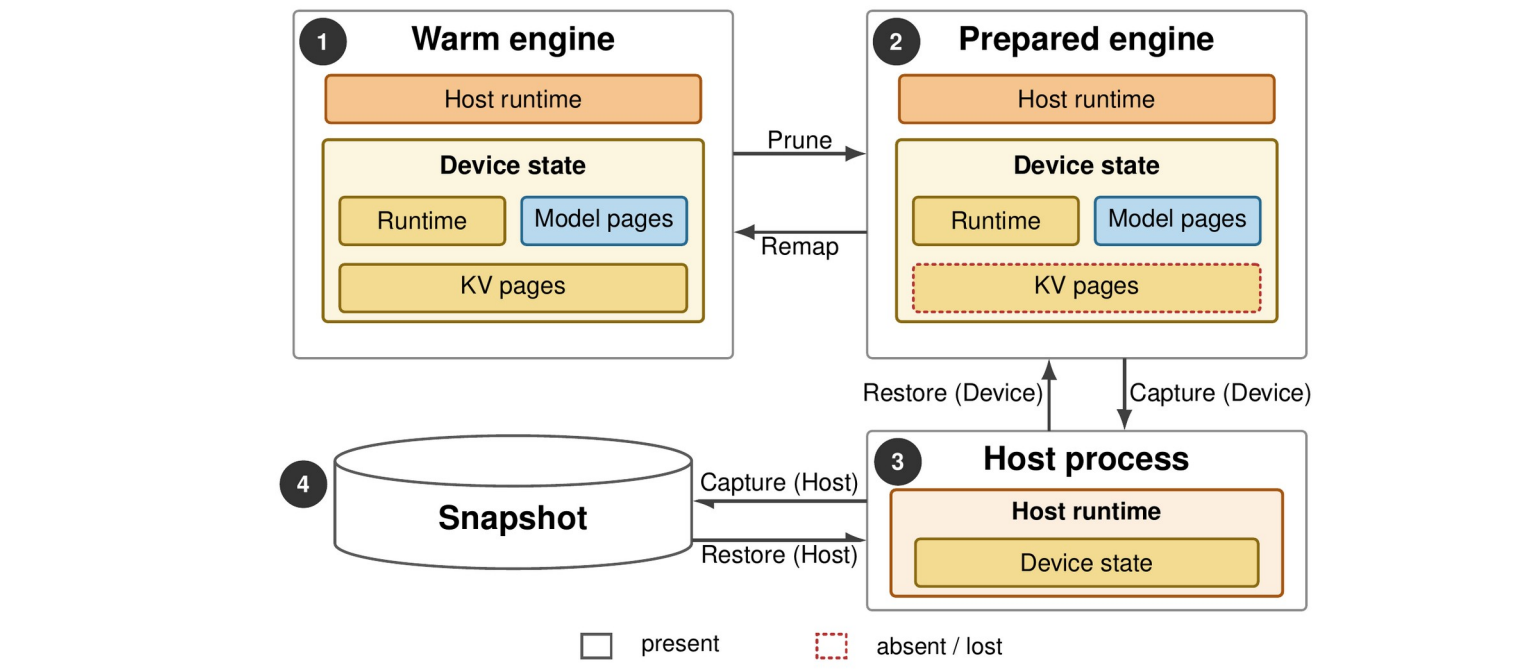
Engine state by lifetime

State	Contents	Preserved by
Model	Rank-local weight tensors	GMS, resident on the GPU
Runtime	CUDA context, CUDA graphs, communicators, execution buffers	Snapshot (durable) or Shadow Engine (resident)
Request	Scheduler metadata, KV mappings and contents	Rebuilt by replaying prefill

Recovery configurations

Configuration	Recovery action
Warm restart	Reinitialize runtime, load model
Snapshot	Restore Snapshot
Snapshot + GMS	Restore runtime, attach model read-only
GMS + Shadow Engine	Promote shadow, install KV, replenish shadow from Snapshot

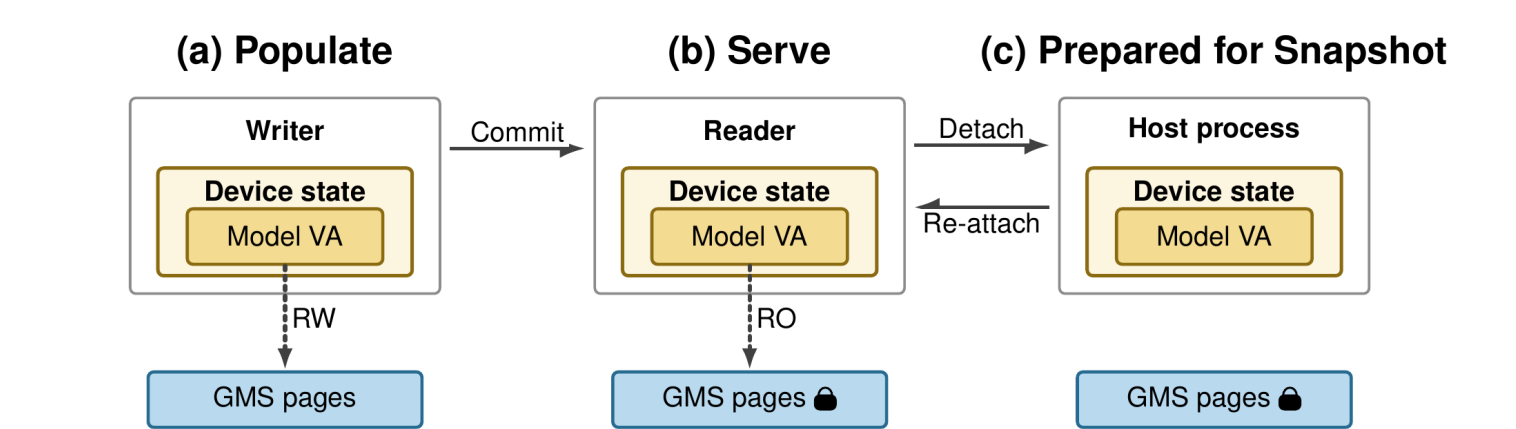
Snapshots



Capture releases empty KV pages, stages device state in the host runtime, and checkpoints the process. Restore reverses these steps.

- A Snapshot is a durable checkpoint of an initialized engine: the host process tree and the CUDA state, including virtual-address mappings.
- KV pages are released before capture while their virtual-address reservations are retained. This preserves the addresses embedded in CUDA graphs.
- Recovery latency is dominated by moving the serialized state, so the artifact excludes state that need not be checkpointed.
- cuda-checkpoint does not capture peer and multicast mappings between GPUs. An LD_PRELOAD shim records them, removes them before capture, and rebuilds them at the original addresses after restore.

Snapshots with GMS

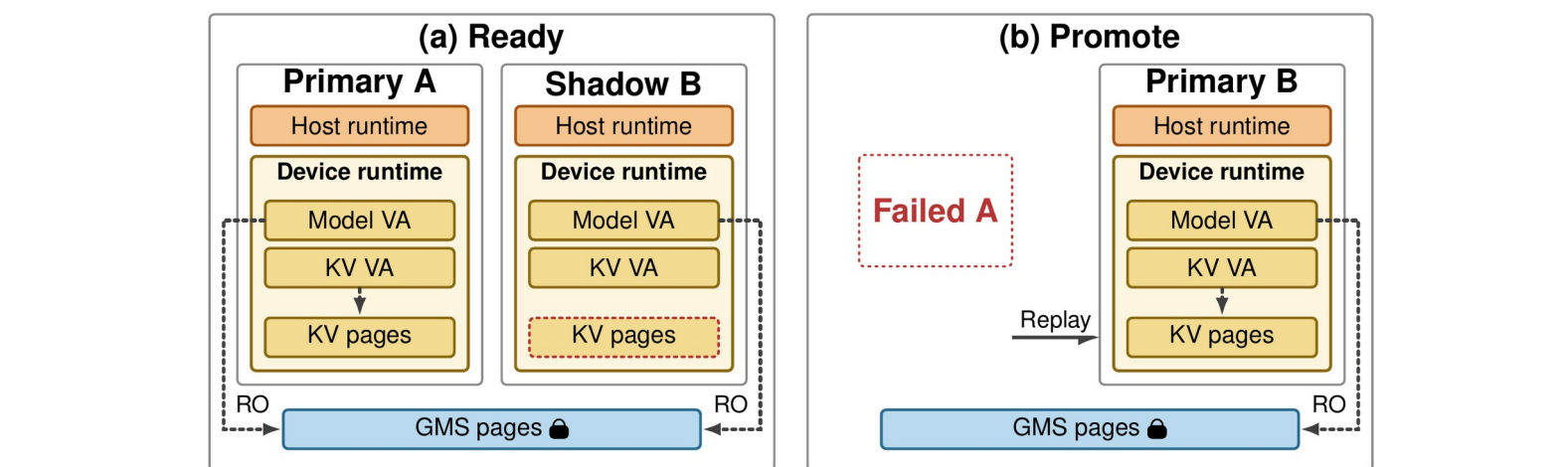


- The checkpoint contains the virtual addresses of the model tensors and none of their contents.
- After a device-preserving failure, the engine imports the surviving GMS allocations at the retained addresses.
- After a device-invalidating failure, model state is reconstructed in fresh GMS allocations in parallel with runtime restoration.

GMS: GPU Memory Service

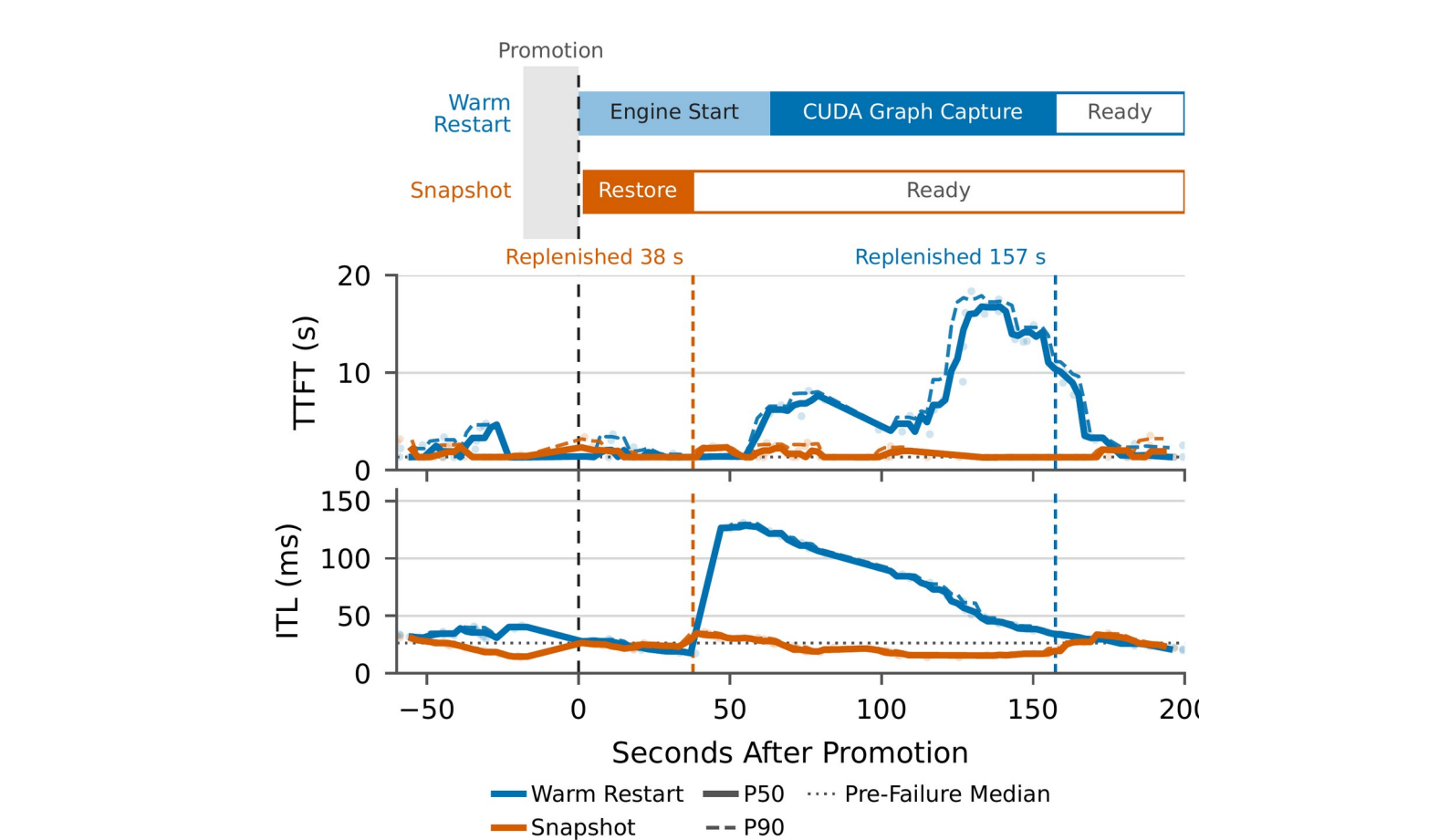
- GMS owns the physical model allocations in a separate process on each GPU and exports their handles to engines over Unix-domain sockets.
- A cold-starting engine populates the allocations under an exclusive read-write lease and commits them. Commit makes the model state immutable.
- GMS never maps the pages, so committed model state survives a device-preserving engine failure. Engines map it read-only, and an invalid write faults only the offending engine (XID 31).

Shadow Engines



Primary and shadow share read-only GMS model state. Only the primary holds KV pages. After failure, the shadow installs fresh KV pages and serves replayed requests.

- A Shadow Engine keeps a second initialized runtime on the same GPUs. Its separate CUDA context survives a device-preserving failure of the primary.
- It needs neither a second copy of the model nor physical KV pages. Its footprint is 4.3–8.1 GiB per GPU while shared model state grows from 8 to 124 GiB.



A new shadow is replenished while the promoted engine serves GLM-5.2. A warm restart captures CUDA graphs on the same GPUs and raises the primary's TTFT and ITL. Restoring the shadow from a Snapshot does not.

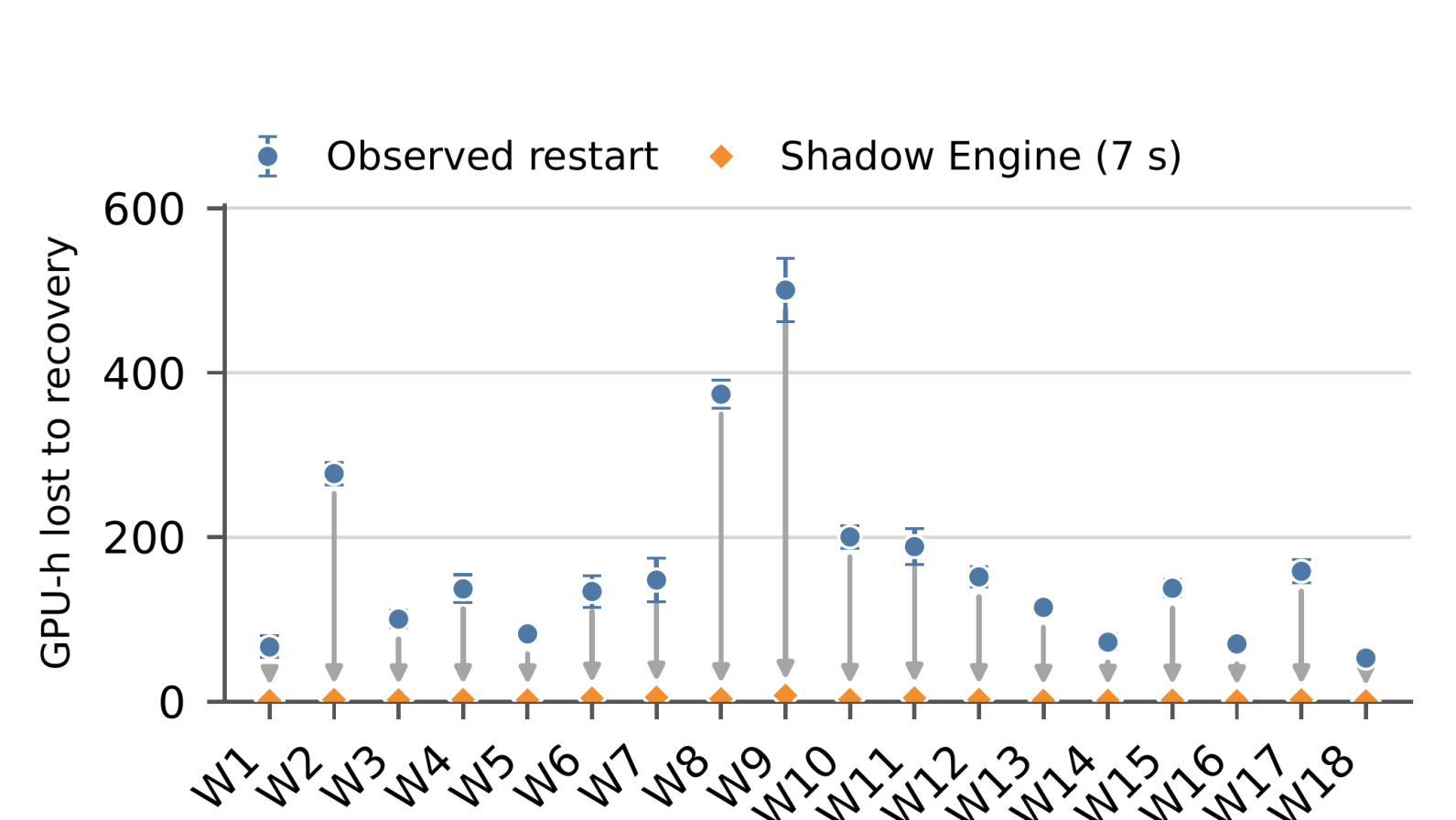
Recovery latency

Single GPU, seconds	Qwen3-0.6B	Qwen3-8B	gpt-oss-120b
Cold start	52.7	58.7	85.5
Snapshot	3.5	8.1	31.1
Snapshot + cold GMS, NFS	3.1	3.8	10.5
Snapshot + cold GMS, NVMe	2.9	3.1	4.1

Single-GPU vLLM, measured from the restore trigger. Cold start excludes base container startup. With weights striped across 8 local NVMe SSDs, gpt-oss-120b restores in 4.1 s, 21× faster than its cold start.

Multi-GPU, seconds	Qwen3.8-27B	DSV4 Flash	GLM-5.2	DSV4 Pro
Warm restart	106.4	109.5	144.4	156.5
GMS, runtime rebuilt	80.0	85.3	82.0	91.9
Snapshot + cold GMS	21.9	26.8	26.8	50.0
Snapshot + warm GMS	16.6	24.1	24.9	25.7
GMS + Shadow Engine	4.6	5.6	5.6	5.4

vLLM on 8×B200 (TP/TEP8). DSV4 is DeepSeek-V4. Failure is injected with SIGKILL and measured from container start until the engine is ready to serve.



Replaying the 18-week production trace with 7 s shadow promotions reclaims 2,339 GPU-hours, 79% of the GPU-hours lost to in-place recovery.



Blog post
tinyurl.com/nvsnsp



Snapshot
github.com/ai-dynamo/snapshot



Dynamo
github.com/ai-dynamo/dynamo