Submission #12

Towards On-the-Fly Snapshot Memory Compression for Low-Latency Elastic Inference Serving Systems

Radostin Stoyanov, Viktória Spišaková, Adrian Reber, Andrei Vagin and Rodrigo Bruno

Introduction

AI inference platforms increasingly rely on large-scale, multi-tenant GPU clusters with scale-to-zero policies [2] that lead to significant **cold-start overheads from model initialization** [1]. Snapshot-based approaches mitigate initialization cost but shift the bottleneck to snapshot size and I/O, while existing compression methods introduce additional bandwidth inefficiencies [3].

We present CRIU-LZ4, **a restore-optimized, on-the-fly compression system integrated within the CPU–GPU checkpointing pipeline** [4]. By eliminating intermediate storage and post-processing, **CRIU-LZ4 reduces restore latency by up to 3× and snapshot size by up to 6×.**

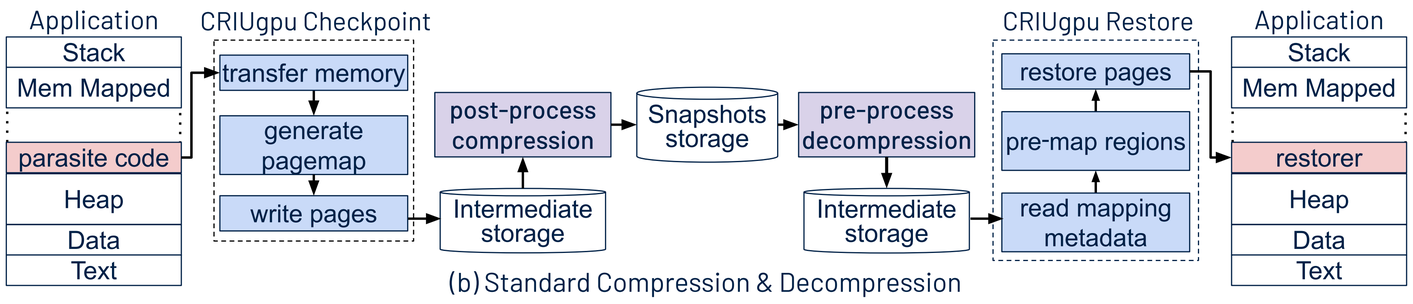
https://criu.org

On-the-Fly Memory Compression with CRIU-LZ4

Container snapshots enable fast startup of inference serving engines by transparently capturing the execution state of fully initialized model-serving instances and avoiding repeated initialization. This makes snapshot restore the critical path, where large memory images are read from storage and transferred to GPU memory.
Our analysis shows that **disk I/O is the primary bottleneck for restore latency (Figure 1)**.

To overcome this limitation, CRIU-LZ4 integrates compression and decompression directly into the memory path, operating at page granularity as data is streamed between CPU and GPU, as shown in Figures 2 and 3. This design **reduces the amount of data transferred during checkpoint and restore** while avoiding intermediate storage, preserving page-level access and maintaining compatibility with incremental snapshots and deduplication.

(a) On-the-fly Memory Pages Compression



(b) Standard Compression & Decompression

Figure 2: Overview of (a) CRIU-LZ4 and (b) standard file-based compression mechanisms for container checkpointing systems with an intermediate local storage.

Figure 1: Throughput of NVMe disk I/O vs. CUDA C/R APIs on an NVIDIA RTX 5090 (PCIe 5.0 x16).

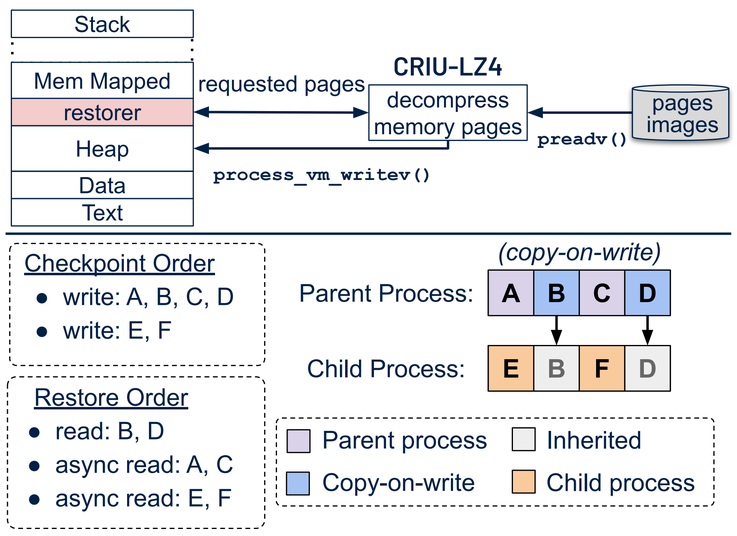


Figure 3: (Top) Restoring compressed memory for a PIE-compiled context; (Bottom) sequence of saving and restoring process tree memory.

Evaluation Results

Restoring from snapshots reduces vLLM startup time from 75–124 s to 39–51 s (46–59% improvement). While uncompressed restore is faster (25–35 s), it requires significantly larger snapshots (36–40 GB vs. 5.4–20 GB).

CRIU-LZ4 achieves a balance between near-baseline latency and significantly reduced storage requirements.

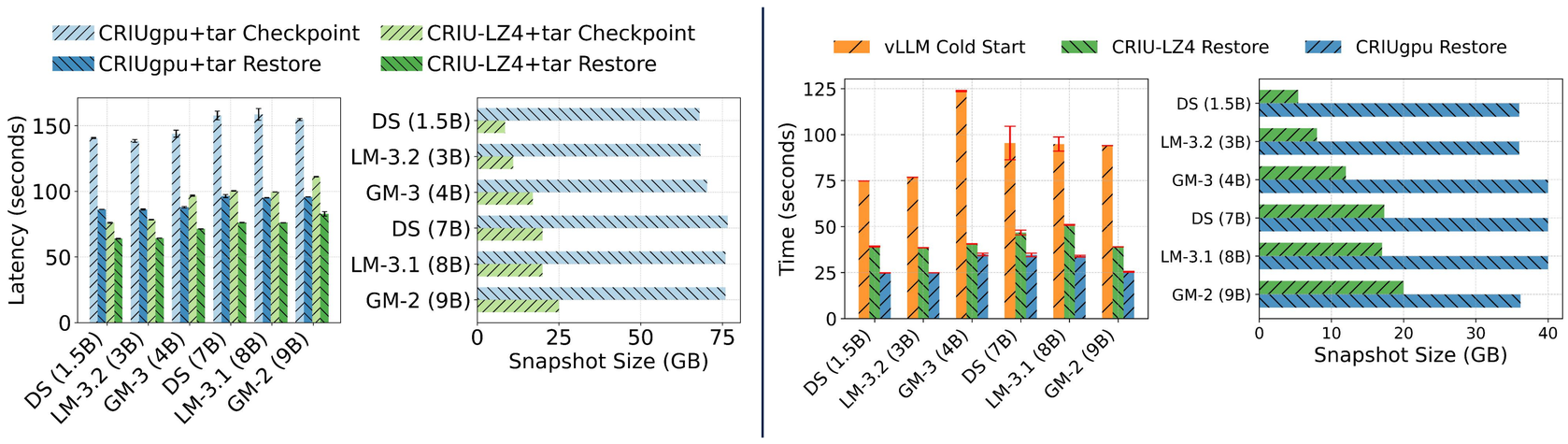


Figure 4: (Left) Checkpoint/restore latency and snapshot size; (Right) cold-start vs. snapshot-based restore performance for vLLM inference, both comparing CRIUgpu and CRIU-LZ4 on an NVIDIA H100 (80 GB).

Performance Analysis

Reducing the amount of snapshot data read from storage directly improves restore latency, as disk I/O dominates the critical path.

Despite the additional compression stage of CRIU-LZ4, restore times remain close to the uncompressed baseline, demonstrating that I/O savings outweigh the compression overhead.

Conclusion

CRIU-LZ4 reduces snapshot size by up to 6× and achieves 46–59% faster startup than cold-start initialization, while maintaining near-baseline restore latency. Our evaluation shows that it provides a practical approach to improving efficiency in inference serving systems.

References

[1] Kwon et al., Efficient Memory Management for Large Language Model Serving with PagedAttention. (SOSP 2023)
[2] Fu et al., ServerlessLLM: LowLatency Serverless Inference for Large Language Models. (USENIX OSDI 2024).
[3] R. Stoyanov et al., Engine-Agnostic Model HotSwapping for Cost-Effective LLM Inference. (CANOPIE-HPC 2025).
[4] R. Stoyanov et al., CRIUgpu: Transparent Checkpointing of GPU-Accelerated Workloads. (https://arxiv.org/abs/2502.16631) 2025

Red Hat · Masaryk University · IST TÉCNICO LISBOA · UNIVERSITY OF OXFORD

The 6th Workshop on Machine Learning and Systems (EuroMLSys) – April 27, 2026, Edinburgh, UK