ReST-KV Eviction Robustness and Throughput at Scale on NVIDIA A100 GPUs

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: Does the robustness of ReST-KV's eviction strategy maintain competitive throughput on NVIDIA A100 GPUs when scaling sequence lengths from 4096 to 32768 tokens compared to H2O and SnapKV methods. 11 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Benchmarking the Nvidia GPU Lineage: From Early K80 to Modern A100 with Asynchronous Memory Transfers. Research question: Does the robustness of ReST-KV's eviction strategy maintain competitive throughput on NVIDIA A100 GPUs when scaling sequence lengths from 4096 to 32768 tokens compared to H2O and SnapKV methods?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

4 papers retrieved. 11 claims extracted; 1 independently verified. Quality review score: 4.4/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The A100 delivers less performance increase than previous generations for the Rodinia benchmark suite.	✓	0.29
Two previous benchmarking studies on the A100 focused exclusively on sparse linear solvers.	×	0.05
Tsai et al. [25] and Anzt et al. [1] both showed a 1.8x improvement of A100 over V100 in sparse linear algebra computati	×	0.01
The K80 GPU (Kepler architecture) was released in Q4 2014 with 12 GB GDDR5 memory and a memory bandwidth of 240.6 GB/s.	×	0.03
The P100 GPU (Pascal architecture) was released in Q2 2016 with 16 GB HBM2 memory and a memory bandwidth of 732.2 GB/s.	×	0.02
The V100 GPU (Volta architecture) was released in Q3 2017 with 16 GB HBM2 memory and a memory bandwidth of 897.0 GB/s.	×	0.02
The A100 GPU (Ampere architecture) was released in Q3 2020 with 40 GB HBM2 memory and a memory bandwidth of 1555 GB/s.	×	0.05
The A100 GPU has 108 Streaming Multiprocessors (SMs) and a TDP of 250 Watts.	×	0.04
The Rodinia benchmark suite includes the 'Backprop' workload which uses the kernels bpnn_adjust_weights_cuda() and bpnn_	×	0.05
The Rodinia benchmark suite includes the 'CFD' solver which utilizes cudaMemcpy(D2D), cuda_compute_step_factor(), cuda_c	×	0.04
The paper presents a microbenchmark designed to gauge the performance implications of new asynchronous copy mechanisms.	×	0.07

References

- http://arxiv.org/abs/2605.09649v1
- http://arxiv.org/abs/2106.04979v2
- http://arxiv.org/abs/2605.08840v1