Lightweight Capability Compression for Bandwidth-Amplified CHERI Workloads

Anonymous Author(s)

Abstract

The Capability Hardware Enhanced RISC Instructions (CHERI) architecture enhances memory safety by extending pointers into 128-bit capabilities. Although this widened representation intuitively suggests a uniform performance penalty, our empirical study reveals that CHERI overhead is highly workload-dependent.

Across varying working-set sizes, 64-bit and 128-bit baselines exhibit nearly identical CPI within cache-resident regimes, indicating that widened capabilities do not inherently degrade pipeline performance. However, once the working set exceeds cache capacity, a pronounced divergence emerges, demonstrating that CHERI's cost is primarily driven by memory bandwidth amplification rather than per-instruction latency.

Motivated by this observation, we target memory traffic reduction instead of traditional pipeline optimizations. Prior compression techniques can reduce bandwidth demand but often introduce decompression latency and additional hardware complexity.

We propose a lightweight representation-level capability compression scheme based on mantissa truncation of bounds encoding. By trading minor bound precision for increased metadata density, our approach reduces memory traffic without modifying the processor pipeline.

Implemented on an iCE40 FPGA platform, our design achieves a 9.54% CPI improvement under memory-bound workloads while maintaining an identical hardware footprint of 321 total cells.

These results demonstrate that carefully designed representation-level approximation can reclaim CHERI performance at near-zero hardware cost by directly addressing bandwidth amplification.

ACM Reference Format:
Anonymous Author(s). 2026. Lightweight Capability Compression for Bandwidth-Amplified CHERI Workloads. In *Proceedings of EuroSys 2026 (EuroSys '26).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The Capability Hardware Enhanced RISC Instructions (CHERI) architecture strengthens memory safety by extending conventional pointers into 128-bit capabilities. This widened representation is often assumed to introduce a persistent performance penalty due to increased data width and memory traffic. However, our empirical study reveals a more nuanced behavior.

Figure 1 shows that the 128-bit baseline remains nearly indistinguishable from the 64-bit baseline when operating within cache-resident working sets (below approximately 120 KB). In this regime, the processor effectively hides the widened capability cost. A clear performance divergence only emerges after exceeding the cache capacity threshold, where memory traffic increases and bandwidth becomes the dominant bottleneck. This observation indicates that CHERI overhead is not an intrinsic per-cycle penalty, but rather a bandwidth-amplified effect triggered by cache evictions.

Motivated by this characterization, we target memory traffic reduction instead of pipeline optimization. We propose a lightweight architectural approximation based on mantissa truncation of capability bounds. By compressing capabilities at the write-back stage, our approach reduces cache-to-memory transfer volume without modifying pipeline depth or introducing decompression latency. The design preserves CHERI's bounds-checking semantics while trading minor bound precision for improved metadata density.

We evaluate the design using synthesis-driven resource analysis in the Yosys framework targeting the iCE40 FPGA architecture. As shown in Table 1, the proposed implementation maintains an identical hardware footprint of 321 total cells, indicating no increase in logic complexity. Performance evaluation (Figure 2) demonstrates a 9.54% CPI improvement under memory-bound workloads.

Together, these results suggest that CHERI's performance degradation is fundamentally bandwidth-driven, and that representation-level compression provides an effective mitigation strategy without additional hardware cost.

Contributions

The primary contributions of this work are:

- **Empirical Characterization of Bandwidth-Amplified Overhead.** We demonstrate that CHERI's performance penalty is workload-dependent rather than constant, and identify the cache-capacity threshold as the point where capability width inflation translates into measurable bandwidth pressure.
- **Logic-Neutral Representation-Level Compression.** We introduce an RTL-level mantissa truncation technique that reduces memory traffic without increasing cell count or modifying the processor pipeline.
- **Quantitative Performance Validation.** We show that the proposed approximation achieves a 9.54% CPI improvement in memory-bound scenarios while maintaining an identical hardware footprint on an iCE40 FPGA target.

2 Observation

3 Design and Implementation

Table 1: Synthesis Results (iCE40 FPGA)

Component	Baseline	Our Work
SUB$_{L}UT4$	134	134
SUB$_{D}FFESR$	2	2
SUB$_{L}DFFSR$	30	30
SUB$_{C}ARRY$	59	59
Total Cells	321	321

EuroSys '26, Edinburgh, UK
2026. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

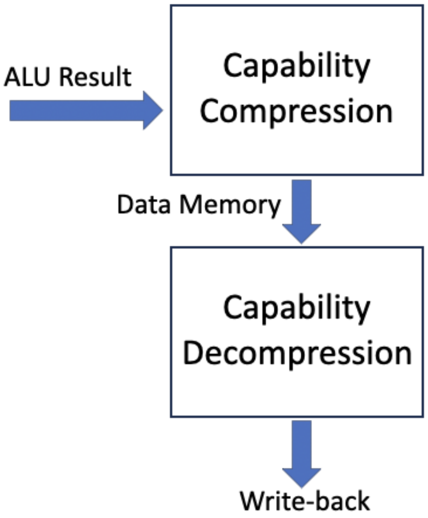

Figure 3: The RTL path of the design

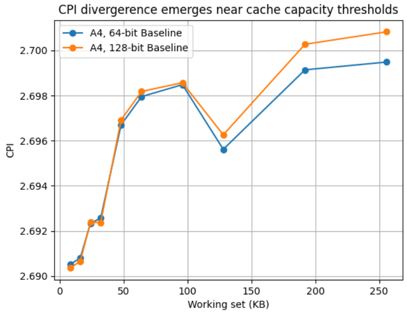

Figure 1: CPI comparison between 64-bit and 128-bit baselines across varying working-set sizes. Performance divergence emerges only beyond the cache-resident regime.

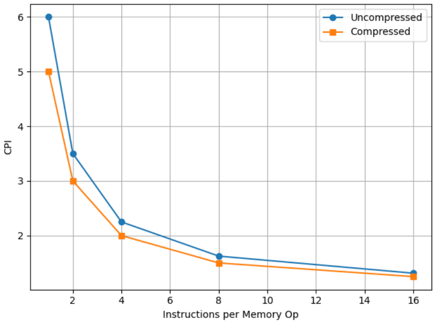

Figure 2: CPI improvement of compressed CHERI relative to uncompressed baseline under varying memory intensity.

4 Conclusion

This work presents an initial investigation into the performance implications of widened capability representations in CHERI systems. Through empirical characterization, we demonstrate that CHERI's performance overhead is not a uniform per-cycle penalty, but a workload-dependent phenomenon that emerges when cache capacity is exceeded and memory bandwidth becomes saturated.

Based on this observation, we propose a lightweight representation-level compression technique using mantissa truncation of capability bounds. By reducing memory transfer volume without modifying pipeline structure or increasing hardware cell count, our approach achieves a measurable CPI improvement in memory-bound scenarios while preserving CHERI's bounds-checking semantics.

These findings suggest that CHERI's performance challenges are fundamentally bandwidth-driven rather than pipeline-bound, and that representation-level approximation provides a practical and low-cost mitigation strategy. While this study focuses on a prototype FPGA implementation, future work includes broader workload evaluation, precision–performance trade-off analysis, and exploration of applicability in larger-scale systems.

Overall, this work demonstrates that carefully designed structural data optimization can reclaim a portion of CHERI-induced bandwidth amplification without increasing architectural complexity.