

Where Does CHERI's Performance Cost Come From?

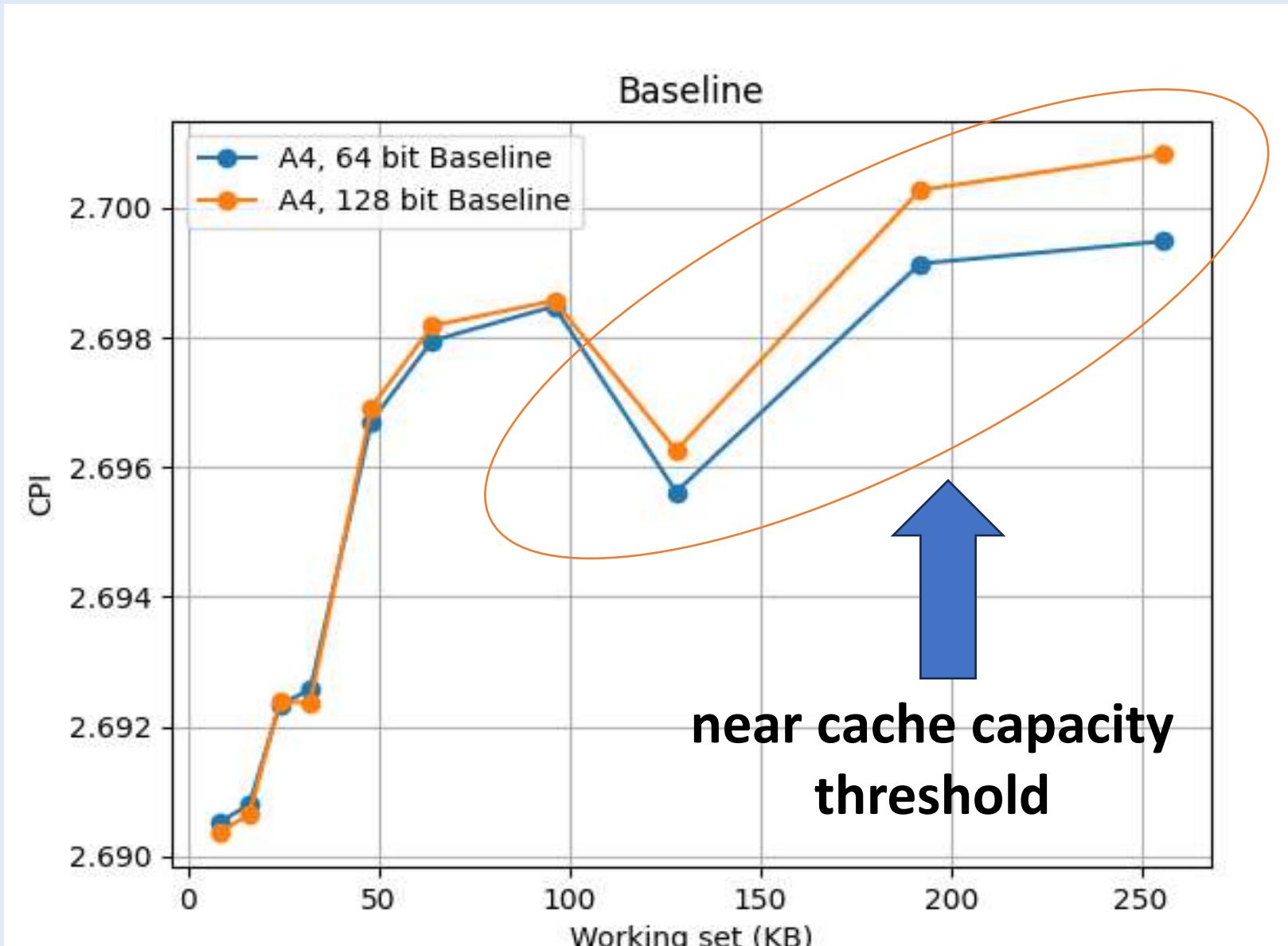
A Case Study on Capability Width and Memory Hierarchy Amplification

Jici Li, University of Edinburgh

Motivation & Setup:

- CHERI enhances memory safety via wide capabilities (128-bit)
- Wider capabilities inflate memory traffic
- Performance overhead is often treated as uniform
- **Question:** *Where is this overhead actually amplified?*

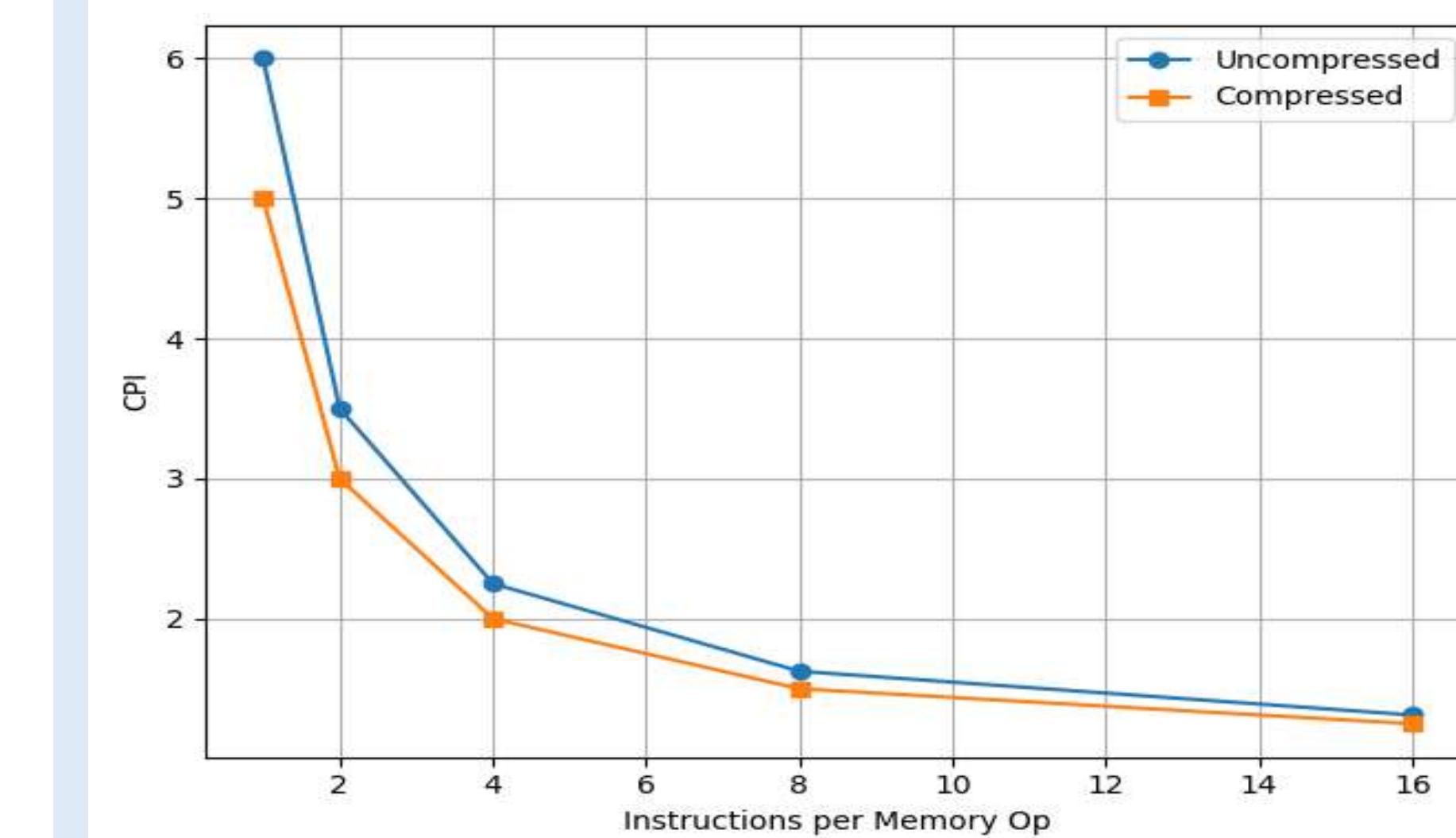
Key Observation:



Observation: CPI degradation peaks near cache capacity thresholds

Performance loss is non-linear.
Overhead concentrates when memory bandwidth and cache replacement are stressed.

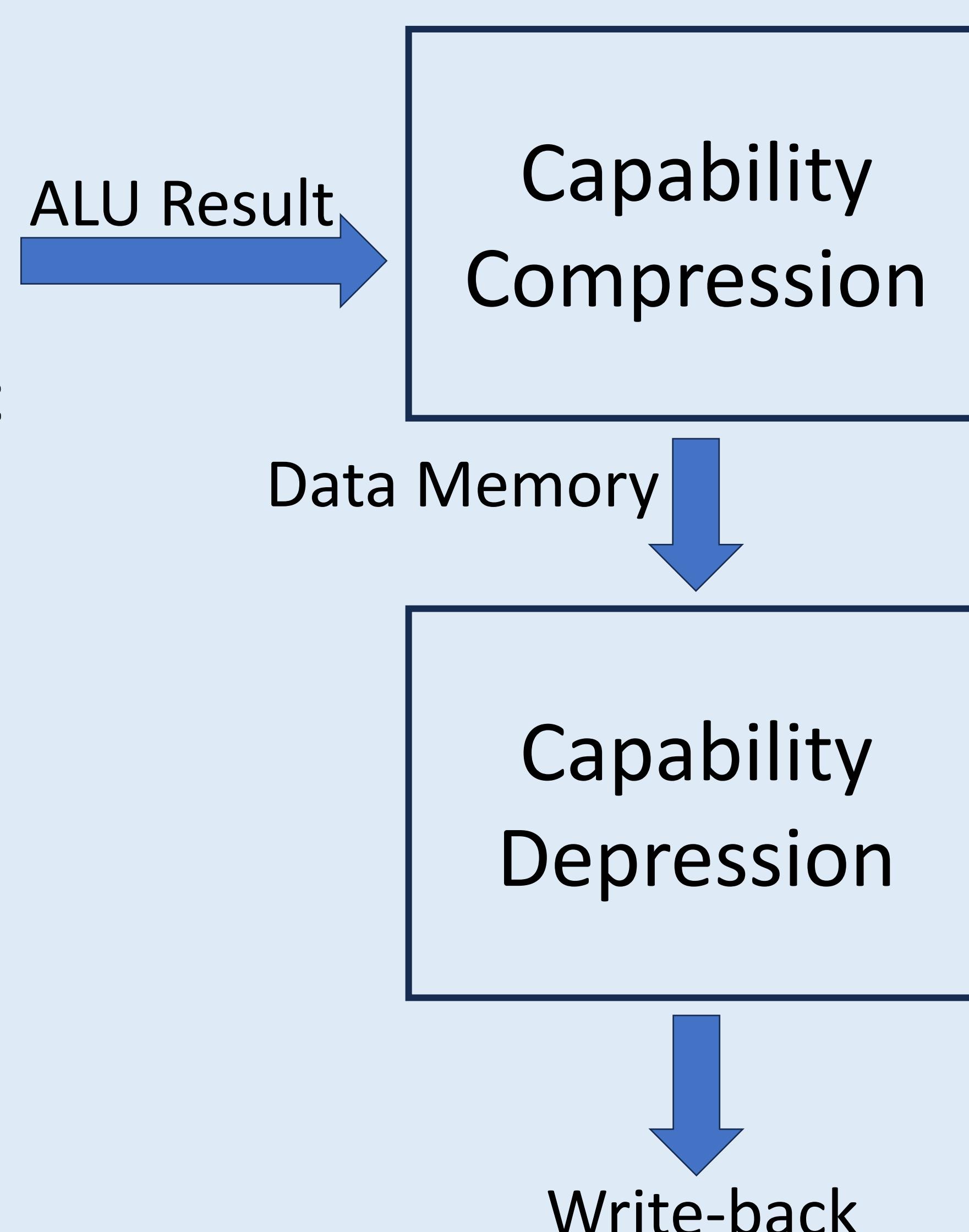
Results:



Architecture-level approximation mitigates bandwidth-amplified overhead in memory-bound regimes.

Approaches & Implementation:

- Capability width inflates memory traffic
- Approximation at representation level
- Mantissa truncation (design point)
- Cuts memory traffic **without modifying pipeline or checks**



Fixed latency,
RTL-level, no
pipeline
restructuring

Cost & Trade-off:

	Compressed	Uncompressed
SB_LUT4	134	134
SB_DFFESR	2	2
SB_CARRY	30	30
SB_CARRY	59	59
total_cells	321	321

Mantissa truncation introduces a bounded loss of spatial precision ($\leq 2^k$ bytes), while preserving capability upper bounds.

Limitation: Application-dependent trade-offs remain future work.