Carat: Unlocking Value-Level Parallelism for Multiplier-Free GEMMs

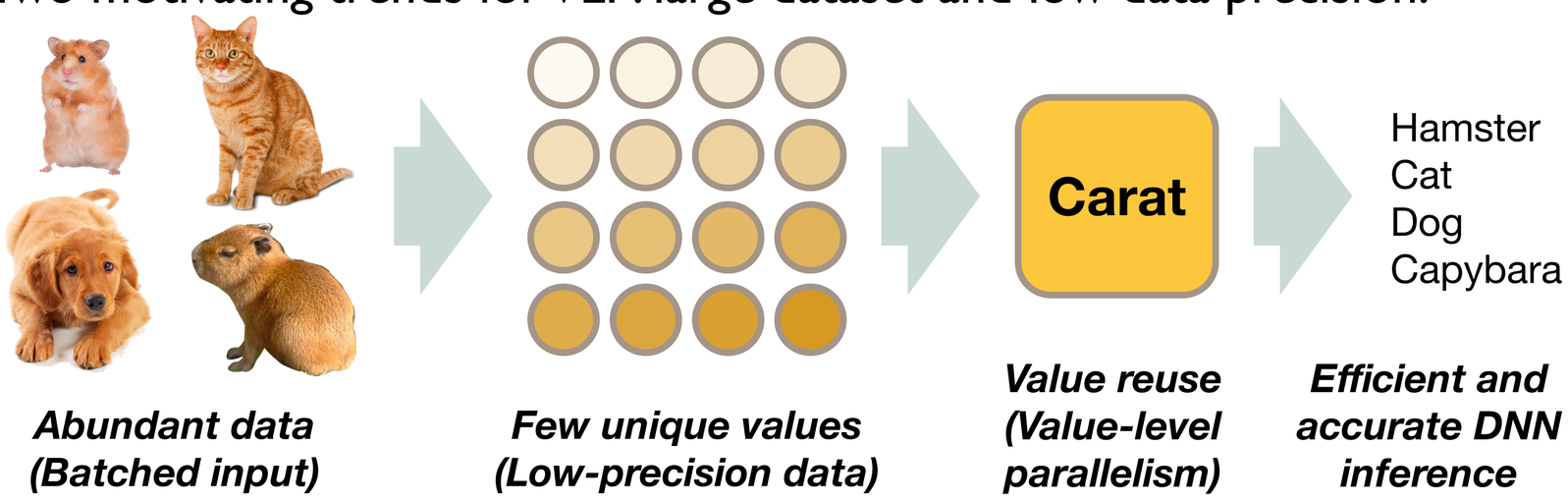
Zhewen Pan, Joshua San Miguel, Di Wu*

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

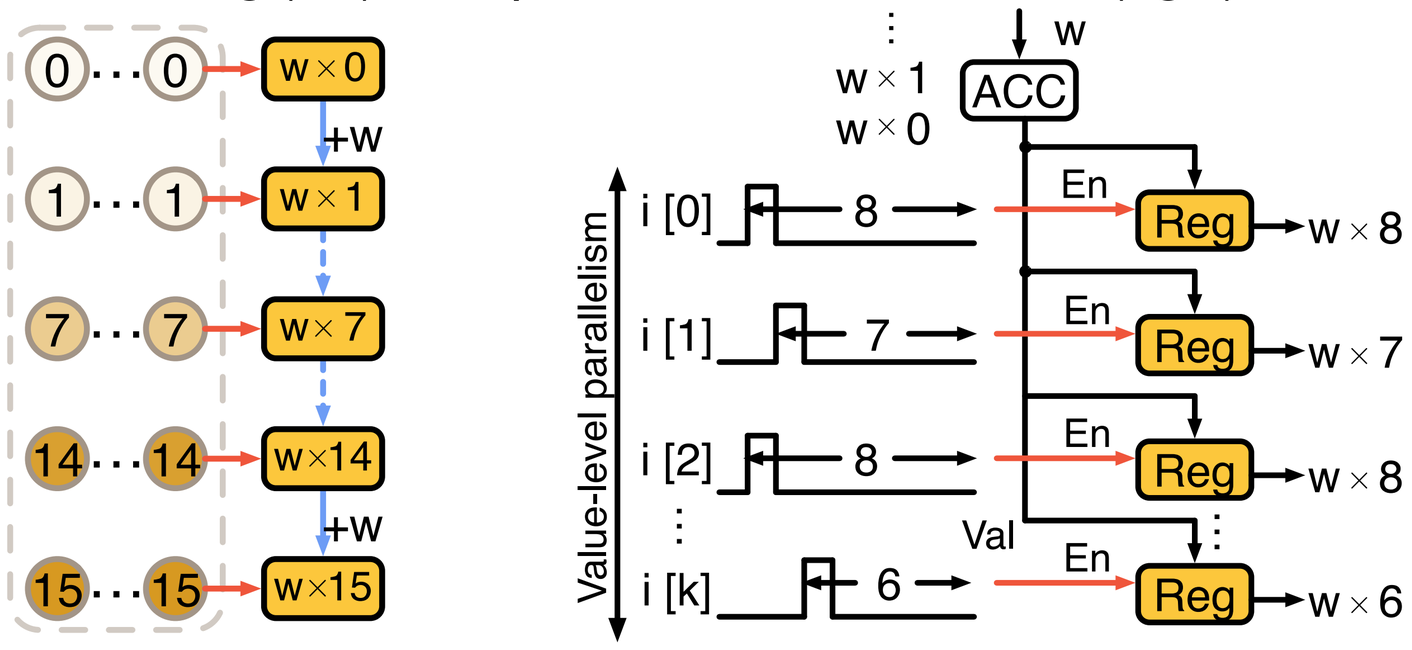
UCF
UNIVERSITY OF CENTRAL FLORIDA *

Value-Level Parallelism

Two motivating trends for VLP: large dataset and low data precision.

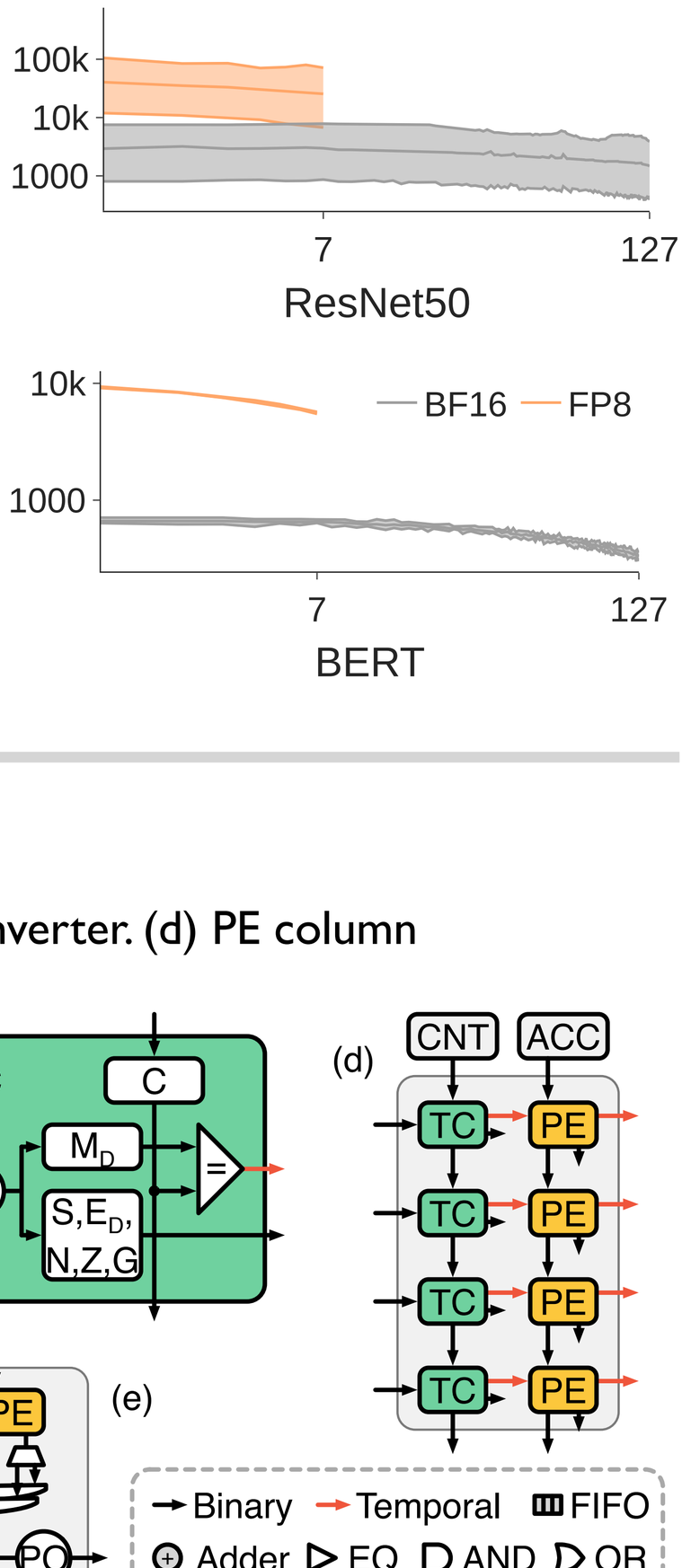


VLP concept of accumulation-based value reuse and subscription via temporal coding (left). Multiplier-free VLP architecture (right).



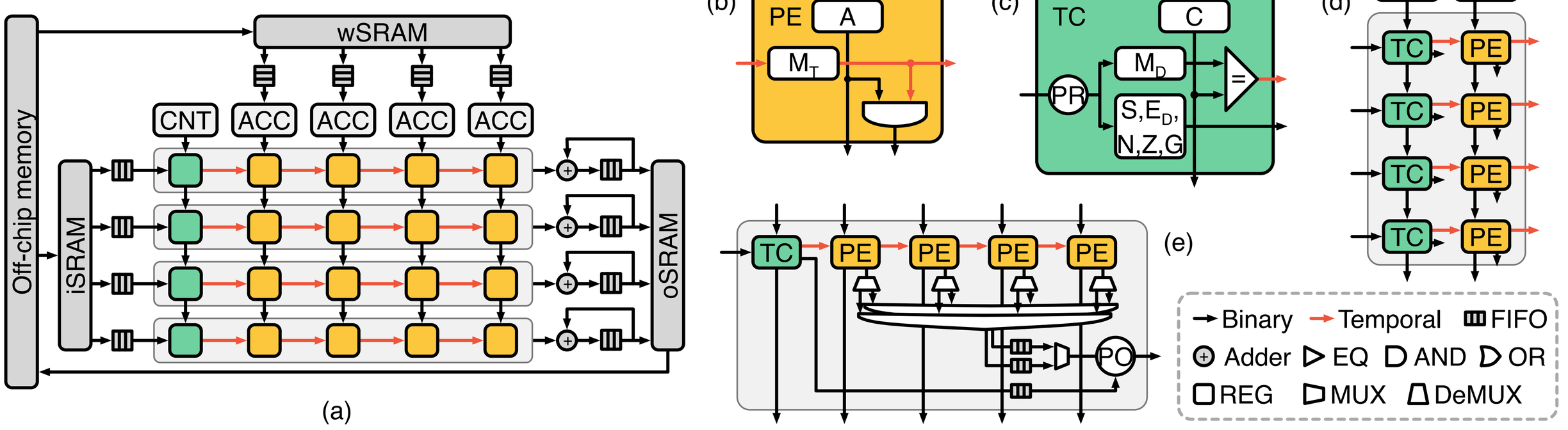
VLP opportunities

Number of inputs per unique mantissa values.

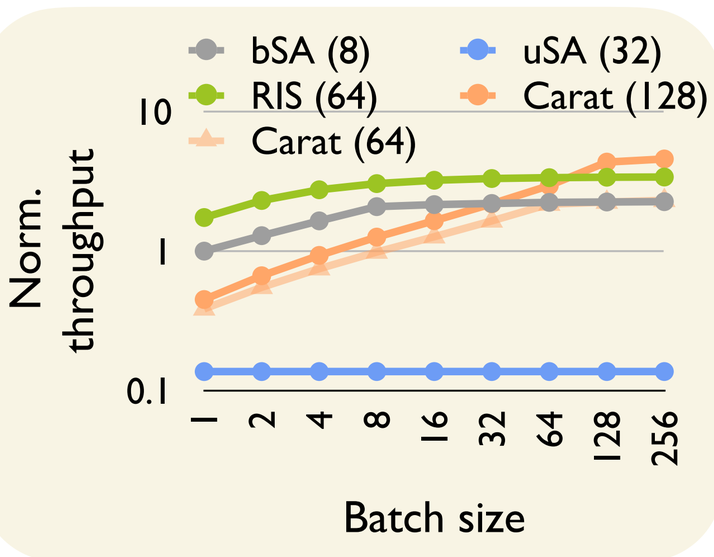


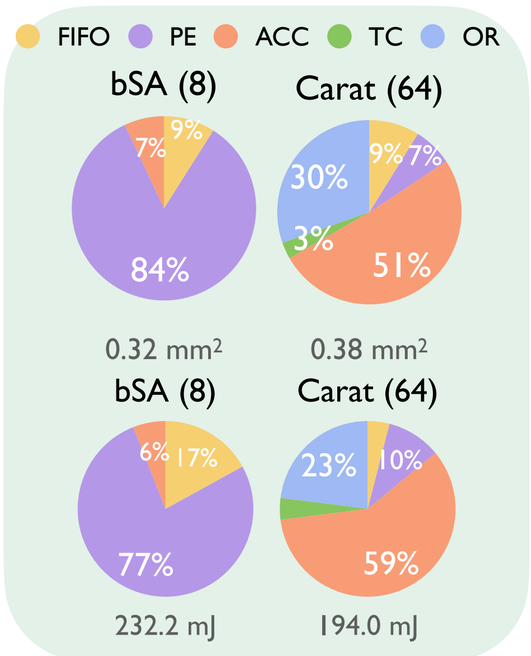
Architecture

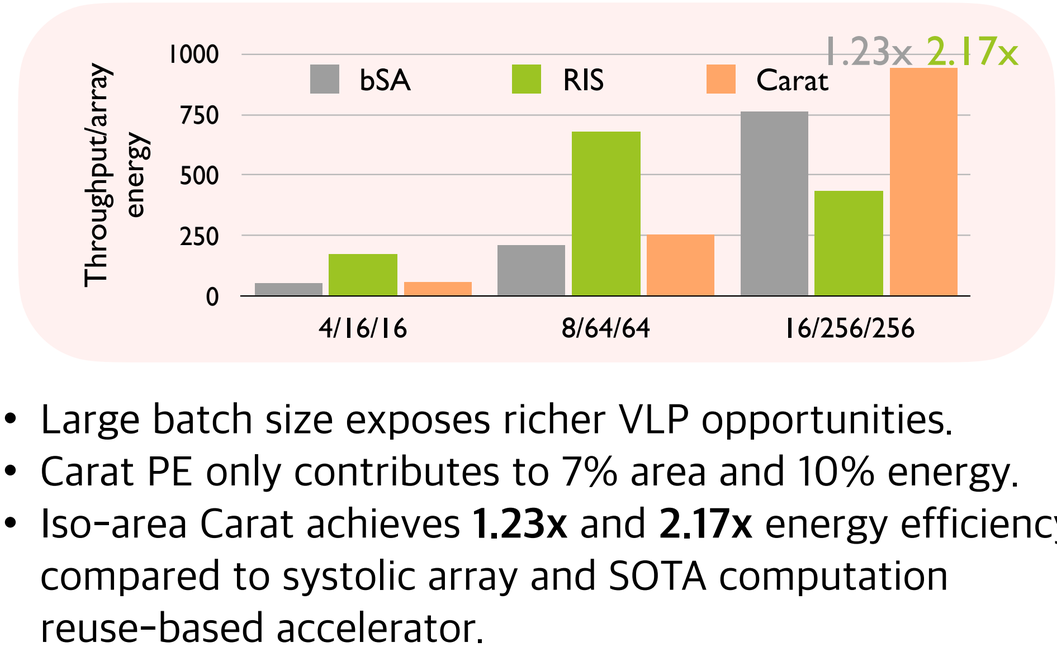
Carat architecture. (a) Overview. (b) accumulator-based PE. (c) Temporal converter. (d) PE column organization. (e) PE row organization.



Results







- Large batch size exposes richer VLP opportunities.
- Carat PE only contributes to 7% area and 10% energy.
- Iso-area Carat achieves **1.23x** and **2.17x** energy efficiency compared to systolic array and SOTA computation reuse-based accelerator.

ASPLOS'24