USC University of Southern California

USC Viterbi School of Engineering

Ming Hsieh Department of Electrical Engineering

Bandwidth Optimizations for 3D Memory Processing

Shreyas G. Singapura, Rajgopal Kannan and Viktor K. Prasanna

Introduction & Motivation

- High performance processor: Frequency, # cores
- Memory Wall
  - Speed gap b/n memory and processors
  - Low bandwidth, high latency
- Solution: **3D Memories**

Challenges

- Large design space due to large number of parameters
- Row activation overhead for every access
- Low page hit rate degrades bandwidth
- Accessing different rows: activation energy

Modeling 3D Memory

- Timing parameters
  - different column
  - different rows
  - different banks
  - different layers

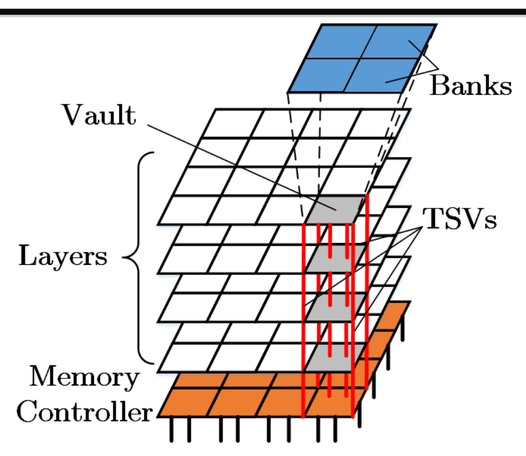


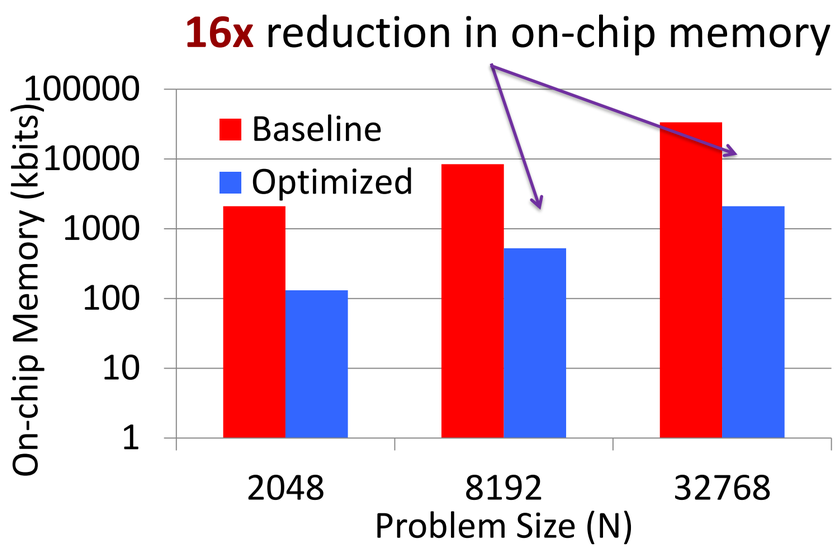
Fig: 3D Memory Architecture

Optimized Data Layout

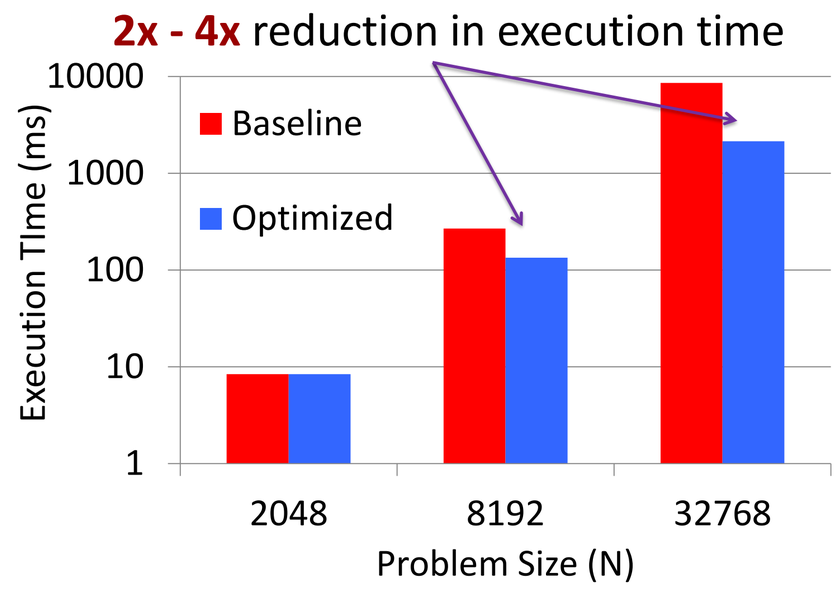
- Exploit parallelism at all levels:
  - Distribute elements across vaults
  - $Inter\text{–}layer\ pipelining$ (t_{layer})
- Exploit large number of banks:
  - Hide t_{col} and t_{row}

FFT

On-chip memory reduced by a factor of $\sqrt{c}$ ×



Execution time reduced factor of $\left(\frac{t_{bank}}{t_{layer}}\right)$ - $\left(\frac{t_{col}}{t_{layer}}\right)$ ×



PARSEC 2.0 Benchmark

- Memory is organized as a set of N blocks
- Pattern decided by user/algorithm → random

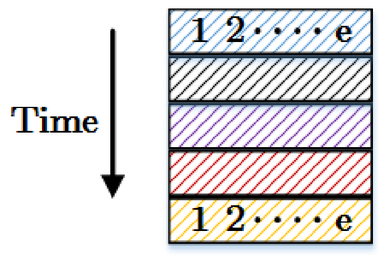


Fig: Random block access

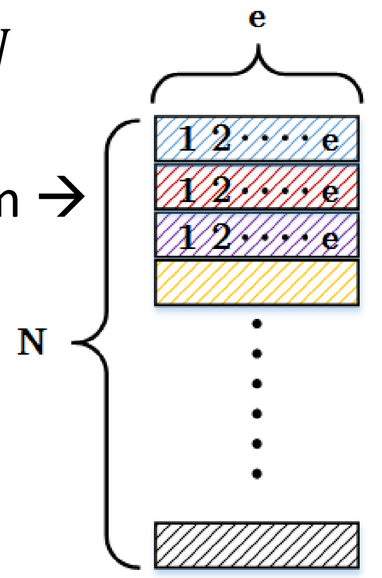
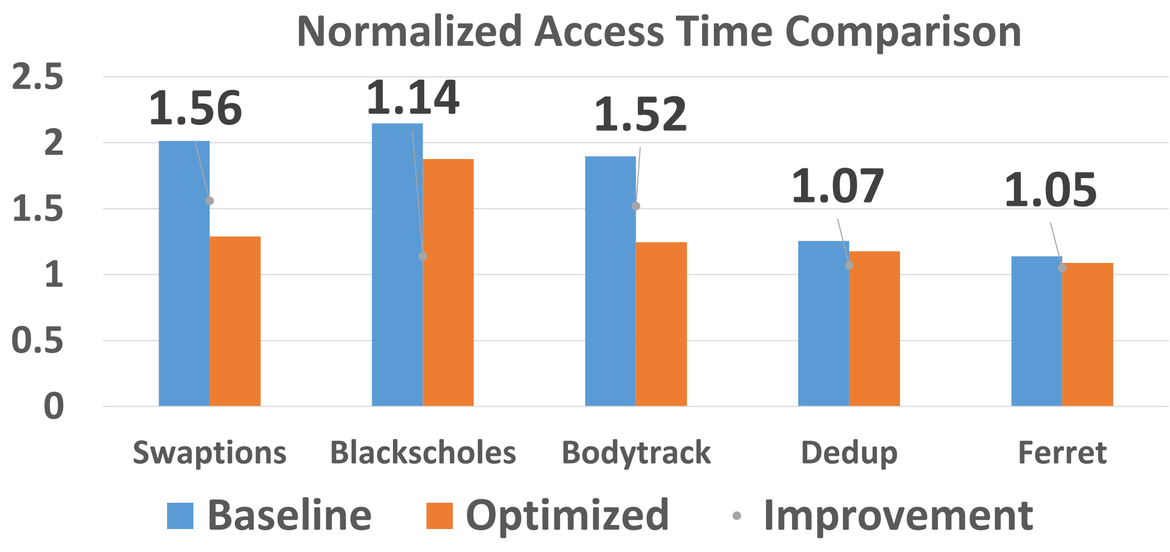


Fig: Memory Layout



{singapur, rajgopak, prasanna}@usc.edu,
http://fpga.usc.edu/

This work has been supported by AFRL under Grant number FA 8750-15-1-0185