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

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

FFT

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

16x reduction in on-chip memory

Baseline Optimized

2048 8192 32768

Problem Size (N)

Execution Time (ms)

100000
10000
1000
100
10
1

Execution Time reduced factor of

Baseline Optimized

2048 8192 32768

Problem Size (N)

PARSEC 2.0 Benchmark
• Memory is organized as a set of $N$ blocks
• Pattern decided by user/algorithm $\rightarrow$ random

Normalized Access Time Comparison

Baseline Optimized Improvement

Swaptions Blackscholes Bodytrack Dedup Ferret

0 0.5 1 1.5 2 2.5

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

(singapur, rajgopak, prasanna)@usc.edu,
http://fpga.usc.edu/