USC University of Southern California

USC Viterbi School of Engineering

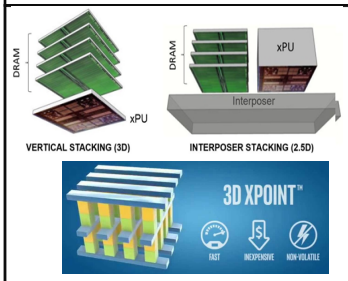
Ming Hsieh Department of Electrical Engineering

Exploiting 3D Memory for Energy-Efficient Memory-Driven Computing

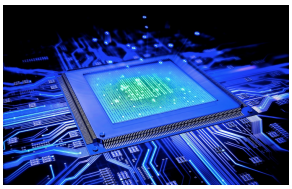
PI:Viktor K Prasanna

Co-PI: Charalampos Chelmis

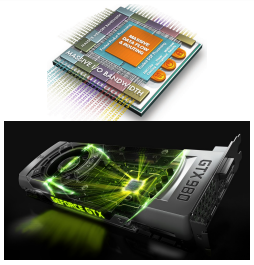
Motivation



Memory: 3D Memory and 3D XPoint



Processor

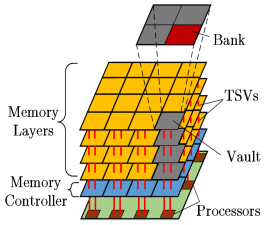


Accelerators: FPGAs and GPUs

Challenges

- Traditional challenges still exist
- Row activation overhead
- Low page hit rate

- Optimizations specific to 3D Memory
- Structure and organization
- Access pattern of the application
- Data layout in the memory



Modeling 3D Memory

- Timing Parameters of 3D memory are:

- t_{vault}: accesses to different vaults
- t_{layer}: accesses to different layers in a vault
- t_{bank}: accesses to different banks in a layer in a vault
- t_{row}: accesses to different rows in a bank
- t_{col}: accesses to different columns in a row

Signal Processing

- 2D FFT: Row phase (sequential) and Column phase (strided)

- Trade row activations for on-chip memory

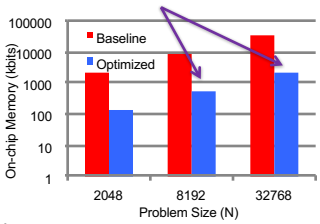
- Exploit parallelism at all levels: Inter-Layer Pipelining and Parallel Vault access

- Exploit large number of banks: Hide t_{col} , t_{row} , t_{bank}

- Optimized data layout for 2D FFT on 3D memory
  - Streaming data sequential and strided accesses
  - $\boldsymbol{O(N)}$ on-chip memory for $N \times N$ problem size
  - $\sqrt{\mathbf{c}}$ × reduction in on-chip memory against state-of-the-art for c columns in a bank row
  - **2 × to 4 ×** improvement in execution time on limited on-chip memory architectures

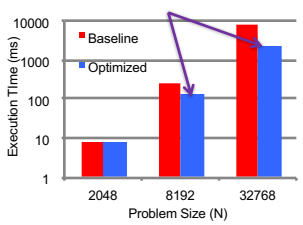
- On-chip memory reduced to $\boldsymbol{O(N)}$ from $\boldsymbol{O(\sqrt{c}N)}$[1]

16x reduction in on-chip memory



- Execution time reduced by a factor of (t_{bank} / t_{layer}) to (t_{col} / t_{layer})[1]

2x - 4x reduction in execution time

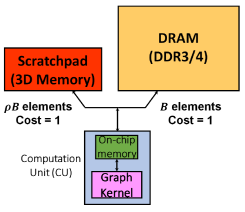


[1] "On-chip Memory Efficient Data Layout for 2D FFT on 3D Memory Integrated FPGA", Shreyas G. Singapura, Rajgopal Kannan and Viktor K. Prasanna, HPEC 2016 (Best Paper Finalist)

Graph Analytics

- 2 Level Main memory:
  - High bandwidth, small capacity scratchpad
  - Low bandwidth, large capacity DRAM

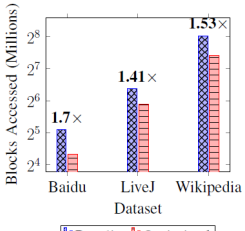
- Trade DRAM accesses for Scratchpad accesses
- Reuse data in scratchpad



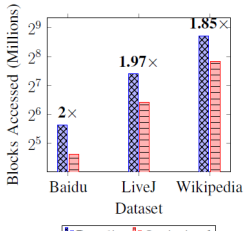
- Optimized data placement for graph processing
  - 2× to 3× reduction in algorithm iterations
  - 1.7× to 2.7× reduction in DRAM accesses
  - 1.4× to 2× improvement in total memory accesses

TABLE VI: Iterations for BFS

Dataset	DRAM		Scratchpad
	Baseline (d_1)	Optimized (d_2)	Optimized ($d_{S_1}^{max}$)
Baidu	21	7	12
LiveJ	15	5	10
Wikipedia	48	11	41



(a) BFS



(b) SSSP

{singapur, rajgopak, prasanna}@usc.edu

http://ganges.usc.edu/