184QPS/W 64Mb/mm²
3D Logic-to-DRAM Hybrid Bonding with Process-Near- Memory Engine for Recommendation System

Dimin Niu¹, Shuangchen Li¹, Yuhao Wang¹, Wei Han¹, Zhe Zhang², Yijin Guan², Tianchan Guan³, Fei Sun¹, Fei Xue¹, Lide Duan¹, Yuanwei Fang¹, Hongzhong Zheng¹, Xiping Jiang⁴, Song Wang⁴, Fengguo Zuo⁴, Yubing Wang⁴, Bing Yu⁴, Qiwei Ren⁴, Yuan Xie¹

¹Alibaba DAMO Academy, Sunnyvale, CA, ²Alibaba DAMO Academy, Beijing, China, ³Alibaba DAMO Academy, Shanghai, China, ⁴UniIC, Xi’an, China
Self Introduction

Education Background

- B.S and M.S degree in electronic engineering from Tsinghua University
- Ph.D. degree in computer engineering from Pennsylvania State University

Work Experience

- Computing Technology Lab, DAMO Academy, Alibaba since 2019
- Memory Solutions Lab, Samsung Semiconductor 2014 - 2019

Research Interests

- Computer Architecture, Computing in/with Memory, Non-volatile memory
Outline

- Motivation
- System and Chip Architecture
  - 3D Logic-to-DRAM Hybrid Bonding
  - PNM Engine for Recommendation System
- Measurement Results
- Conclusion
Outline

■ Motivation

■ System and Chip Architecture
  ● 3D Logic-to-DRAM Hybrid Bonding
  ● PNM Engine for Recommendation System

■ Measurement Results

■ Conclusion
Memory Wall in AI Era

Scaling Out Solution
- More hardware
- More computation time

Solution Needed
- Limited by physical limit
- New chip architecture or new memory technology

750x / 2 years  3.1x / 2 years  1.4x / 2 years
Al model computation requirement  Hardware computation capability  Memory system capability
Memory-Bound Applications

Natural Language Processing

Recommendation Systems

Graph Neural Network

Multi-Task Online Inference
State-of-the-art PNM/CIM Solutions

Traditional

2D CIM

Memory & Computation

2D PNM

Memory & Computation

2.5D PNM

Memory

Computation

Interposer

3D HB-PNM

Memory

Computation

3D TSV-PNM

Memory

Memory

Memory

Computation

PNM: Process Near Memory
CIM: Compute In Memory
HB: Hybrid Bonding
TSV: Through-silicon Via
Outline

- Motivation
- System and Chip Architecture
  - 3D Logic-to-DRAM Hybrid Bonding
  - PNM Engine for Recommendation System
- Measurement Results
- Conclusion
3D Logic-to-DRAM Hybrid Bonding

- Logic-to-DRAM face-to-face Hybrid wafer Bonding
- 25nm DRAM technology with 36 x 1Gbits array
- 1Gbits DRAM core with 8 banks and on-chip ECC
- Each bank with 128 bits I/O, and implemented with HB

![Diagram showing 3D Logic-to-DRAM Hybrid Bonding](image-url)
Hybrid-bonding Interconnection

- Cu-Cu direct fusion with low bonding temperature (< 350°C)
- Up to 110,000/mm² integration density
- Small pitch size of 3μm
- Align marker with high precision of 0.35μm
Outline

- Motivation
- System and Chip Architecture
  - 3D Logic-to-DRAM Hybrid Bonding
  - PNM Engine for Recommendation System
- Measurement Results
- Conclusion
Typical Recommendation System

- A two-step Recommendation System
  - Feature Generation
    - Classification, object detection and feature extraction
    - Computation-bound
    - Typically executed on GPU
  - Matching & Ranking
    - Coarse-grained matching and fine-grained ranking
    - Memory-bound
    - Typically executed on CPU and commercial DRAM as external memory
    - Consumes most latency (89.87%) and energy (82.97%)
    - Requires high-bandwidth, large-capacity and energy-efficient memory

*Item feature can be extracted from different methods. Here is a typical case for image queries.
Ranking & Matching

- **Coarse-grained Matching**
  - Coarse-grained features with 1bit x 512 dimensions
  - Matching: L2 distance calculation
  - Top-1000 items selected from 40K items

- **Fine-grained Ranking**
  - Fine-grained features with 8bits x 1024 dimensions
  - Similarity prediction: three-layer MLP (2048-256-64-1)
  - Top-100 ranking results selected from 1K items
Overall Architecture

- **Memory**
  - 4 x 1Gb blocks with 4096 bits I/O
  - 38.4GB/s on-chip bandwidth per block

- **Compute**
  - Match Engine: Coarse-grained Matching
  - Neural Engine: Fine-grained Ranking

- **Dual-mode Interface**
Match Engine Architecture (1)

- Address Generator
  - Multi-mode to support different access patterns
  - Configurable via registers
  - Build-in performance evaluation mode & performance counter
Match Engine Architecture (2)

- **Distance Calculator**
  - Compare the similarity between input feature and query
  - Compute the Hamming distance of two 512-bit features
  - Filtered by rot of max-heap
Match Engine Architecture (3)

- **Top-k Engine**
  - Maintain a max-heap hardware block
  - Receives input every two cycles
  - Alternately heapifies nodes in odd layers and even layers
  - Stores the top-1000 matching results
Neural Engine Architecture (1)

- Vector Process Unit
  - Activations
    - LUT based design
    - Supports GeLU & Exp
  - Transpose
    - Transpose 16x16 matrix with ping-pong array
    - Implemented with 2D register file array
    - Supports row-based writes and column-based reads
Neural Engine Architecture (2)

- GEMM
  - 32 by 32 systolic PE array (INT8)
  - Partial sum accumulated by the accumulator (INT32)
  - 600GOPS (@300MHz)
Interface Bridge & Debug

- Support both single-channel mode and lockstep-mode
- Read/write counter to support burst requests
- Support cycle-wise debug with clock gating
Outline

- Motivation
- System and Chip Architecture
  - 3D Logic-to-DRAM Hybrid Bonding
  - PNM Engine for Recommendation System
- Measurement Results
- Conclusion
Die Photo and Summary

### DRAM Die

<table>
<thead>
<tr>
<th></th>
<th>25nm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology</strong></td>
<td>602.22 mm²</td>
</tr>
<tr>
<td><strong>Area</strong></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>602.22 mm²</td>
</tr>
<tr>
<td>Neural Engine</td>
<td>32 mm²</td>
</tr>
<tr>
<td>Match Engine</td>
<td>32 mm²</td>
</tr>
<tr>
<td><strong>Voltage</strong></td>
<td>1.1 V</td>
</tr>
<tr>
<td><strong>Frequency (max)</strong></td>
<td>150 MHz</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>300 mW per 1Gb</td>
</tr>
<tr>
<td><strong>Bandwidth</strong></td>
<td>153.60 GB/s / 1.38 TB/s</td>
</tr>
</tbody>
</table>

### Logic Die

<table>
<thead>
<tr>
<th></th>
<th>55nm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology</strong></td>
<td>602.22 mm²</td>
</tr>
<tr>
<td><strong>Area</strong></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>602.22 mm²</td>
</tr>
<tr>
<td>Neural Engine</td>
<td>5.90 mm²</td>
</tr>
<tr>
<td>Match Engine</td>
<td>7.02 mm²</td>
</tr>
<tr>
<td><strong># of MC</strong></td>
<td>16 per IP</td>
</tr>
<tr>
<td><strong>Voltage</strong></td>
<td>1.2 V</td>
</tr>
<tr>
<td><strong>Frequency</strong></td>
<td>300 MHz</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>977.70 mW</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>INT8</td>
</tr>
</tbody>
</table>
Evaluation Platform

- Test board capable to mount up to 4 HB
- FPGA board responsible to write/read data and generate configuration to the chip register
Performance

<table>
<thead>
<tr>
<th></th>
<th>CPU - DRAM*</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic Technology</td>
<td>14 nm</td>
<td>55 nm</td>
</tr>
<tr>
<td>Frequency</td>
<td>2.20 GHz</td>
<td>300 MHz</td>
</tr>
<tr>
<td>Area</td>
<td>4294 mm²</td>
<td>64 mm² (4Gb)</td>
</tr>
<tr>
<td>Precision</td>
<td>INT8</td>
<td>INT8</td>
</tr>
<tr>
<td>Power**</td>
<td>70.17 W (TDP: 125 W)</td>
<td>2.178 W</td>
</tr>
</tbody>
</table>

* CPU: Intel Xeon Gold 5220@2.20GHz, tested on Pytorch
** CPU power measured by PyRAPL

- Measurement vs. Peak: ~20% initialization and memory subsystem overhead
# Comparison

<table>
<thead>
<tr>
<th></th>
<th>2D CIM *</th>
<th>2D PNM **</th>
<th>2.5D PNM ***</th>
<th>3D TSV (Hybrid) ****</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of Memory</td>
<td>SRAM</td>
<td>DDR4</td>
<td>HBM2</td>
<td>HBM2</td>
<td>LPDDR4</td>
</tr>
<tr>
<td>Technology (Memory/Logic)</td>
<td>16nm</td>
<td>2xnm / 2xnm</td>
<td>1y# / 7nm</td>
<td>20nm / 20nm</td>
<td>25nm / 55nm</td>
</tr>
<tr>
<td>Capacity</td>
<td>4.5 Mb</td>
<td>8GB / DIMM</td>
<td>80GB</td>
<td>6GB / cube</td>
<td>4.5GB</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>-</td>
<td>128GB/s / DIMM</td>
<td>1935GB/s</td>
<td>1200GB/s / cube#</td>
<td>38.4GB/s / 1Gb</td>
</tr>
<tr>
<td>Frequency (Logic)</td>
<td>200MHz</td>
<td>500MHz</td>
<td>1410MHz</td>
<td>300MHz</td>
<td>300MHz</td>
</tr>
<tr>
<td>Bandwidth/Capacity (a.u.)</td>
<td>-</td>
<td>16</td>
<td>24.2</td>
<td>200</td>
<td>307</td>
</tr>
</tbody>
</table>

- High off-chip bandwidth
- High bandwidth per capacity
- Low energy per bit

** F. Devaux et al, Hotchip 2019
*** J. Choquette et al, Hotchip 2020
**** Y. C. Kwon et al, ISSCC 2021

#Estimated
Outline

- Motivation
- System and Chip Architecture
  - 3D Logic-to-DRAM Hybrid Bonding
  - PNM Engine for Recommendation System
- Measurement Results
- Comparison
- Conclusion
Conclusion

- Memory-bound application can significantly benefit from process-near-memory and computing-in-memory
- A 3D Logic-to-DRAM Hybrid Bonding Chip with Process-Near-Memory Engine for Recommendation System is demonstrated featuring with:
  - High-bandwidth and energy-efficient memory with hybrid bonding
  - High-throughput streaming processing units for matching and ranking
  - 2.4GB/s/mm² bandwidth density and 0.88pJ/bit energy consumption
  - ~10x performance improvement and over 300x energy-efficiency improvement over conventional CPU+DRAM system