29.1 184QPS/W 64Mb/mm² 3D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System

Dimin Niu[1], Shuangchen Li[1], Yuhao Wang[1], Wei Han[1], Zhe Zhang[2], Yijin Guan[2], Tianchan Guan[3], Fei Sun[1], Fei Xue[1], Lide Duan[1], Yuanwei Fang[1], Hongzhong Zheng[1], Xiping Jiang[4], Song Wang[4], Fengguo Zuo[4], Yubing Wang[4], Bing Yu[4], Qiwei Ren[4], Yuan Xie[1]

[1]Alibaba DAMO Academy, Sunnyvale, CA; [2]Alibaba DAMO Academy, Beijing, China
[3]Alibaba DAMO Academy, Shanghai, China; [4]UniIC, Xian, China

The era of AI computing brings significant challenges to traditional computer systems. As shown in Fig. 29.1.1, while the AI model computation requirement increases 750× every two years, we only observe a very slow-paced improvement of memory system capability in terms of both capacity and bandwidth. There are many memory-bound applications, such as natural language processing, recommendation systems, graph analytics, graph neural networks, as well as multi-task online inference, that become dominating AI applications in modern cloud datacenters. Current primary memory technologies that power AI systems and applications include on-chip memory (SRAM), 2.5D integrated memory (HBM [1]), and off-chip memory (DDR, LPDDR, or GDDR SDRAM). Although on-chip memory enjoys low energy access compared to off-chip memory, limited on-chip memory capacity prevents the efficient adoption of large AI models due to intensive and costly off-chip memory access. In addition, the energy consumption of data movement of off-chip memory solutions (HBM and DRAM) is several orders of magnitude larger than that of on-chip memory, bringing the well-known "memory wall [2]" problem to AI systems. Process-near-memory (PNM) and computing-in-memory (CIM) have become promising candidates to tackle the "memory wall" problem in recent years.

As summarized in Fig. 29.1.1, implementation and integration methods result in different types of PNM [3, 4] and CIM [5] systems. This work focuses on one practical architecture, 3D HB-PNM with mature chip fabrication and bonding process [7], to provide a high-density and high-energy-efficiency PNM solution built with logic-to-DRAM hybrid bonding (HB) technology for AI applications. In addition to the significant on-chip memory capacity advantage, the hybrid wafer-to-wafer bonding technology is capable of delivering a high bandwidth integrated solution with >200× better energy efficiency compared to off-chip memory solutions such as HBM or DDR SDRAM and at least 2× better energy efficiency compared to state-of-the-art (SOTA) PNM solutions. Applied to a recommendation system, our PNM macro with 64Mb/mm² on-chip memory density together with 2.4GB/s/mm² bandwidth density is capable to deliver high performance of 184.11QPS/W (queries per second per watt).

As shown in Fig. 29.1.2, the entire chip consists of a DRAM die with 25nm technology stacking on top of a logic die with 55nm technology, connected by uniformly distributed copper metal pads in the bond interface with a pitch size of 3μm. The bonding pads serve as both electrical wires and supporting materials between two dies. The DRAM die and the logic die have the same dimension of 25.24×23.86mm². The DRAM is composed of 6×6 equally-sized blocks, and each block is 4×4mm² with 1Gb capacity. The memory controllers and PHYs of each DRAM block are located at the corresponding position of the logic die, which naturally partitions the logic die into 6×6 blocks as well. Each logic block has direct access to its counterpart DRAM block while being able to access all other memory blocks via an on-chip bus. With a block-aware design and placement strategy, the compute engine on the logic die can span multiple logic blocks and access multiple DRAM blocks, enabling a flexible architecture design. The 4Gb DRAM blocks serve as the memory for two computation engines, a match engine (ME) and a neural engine (NE).

Figure 29.1.3 shows the execution flow of an industrial recommendation system for item retrieval matching and the overall architecture, including two compute engines and 4Gb DRAM. Typically, a recommendation system has two steps: the first step is feature extraction, which is usually compute-bound and requires a GPU to perform classification, object detection, and feature extraction. The second step includes coarse-grained matching and fine-grained ranking, which is memory-bound and conventionally performed on CPUs. To tackle the memory-bound part, a PNM architecture is proposed. An ME with a distance calculation module and a max-heap top-k module fulfills the matching phase of the recommendation system. The top-1000 items selected from 40K items are then delivered to the NE for DNN operations. The NE provides 600GOPS (@300MHz) computation capability for similarity prediction. The similarity prediction is accomplished by a three-layer MLP (2048-256-64-1), which is trained and quantized (INT8) on an internally used dataset. After the fine-grained re-ranking in the ME, the top-100 item indices are selected for recommendation. The two compute engines connect to the DRAM by a dual-mode interface module, which can switch between lockstep (of 8 banks) mode for full bandwidth and single channel (any 1 of the 8 banks) mode for flexibility. More flexible recommendation algorithms are available with arbitrary combinations of ME and NE. They are orchestrated and configured by a controller which takes commands from an external test platform.

Figure 29.1.4 shows the detailed design of ME. ME is responsible for performing the coarse-grained matching operation on 512-dimension binary feature vectors in our recommendation model. The address generator (AddrGen) generates the address of the input query stored in DRAM according to predefined configurations. The feature vectors are fetched from DRAM and delivered to the distance calculator and ranking engine to perform Hamming distance calculation and top-k sorting, respectively. The top-k engine is implemented by a max-heap hardware block, and the top-1000 shortest distance results from the distance calculator. The max-heap block maintains a max-heap data structure with 1000 nodes, in which each node contains an <address, distance> pair. This block receives input every two cycles and alternately heapifies nodes in the odd layers and the even layers. Since the maximum throughput of max-heap block is only half of the distance calculator's, the input of max-heap block is filtered by comparison logic: only if the latest calculated distance is smaller than the root node of the max-heap, it replaces the root node and heapifies the max-heap in a top-to-bottom manner.

Figure 29.1.5 demonstrates the other key module in our design, NE. NE consists of interface sub-modules, a vector processing unit (VPU), general matrix multiplication (GEMM), and control sub-modules. The interface bridge is in charge of the data transmission between memory and the VPU/GEMM and can work in either lock-step mode or single-bank mode on demand. VPU is responsible for vector operations, specifically activation functions and transpose operations. With proper control signals, the activation unit performs activation functions with 8 LUTs, and the transpose unit performs transpose operations on 2 register arrays. GEMM computes matrix multiplications with a fully-pipelined systolic array (32×32 INT8), and the partial sums are then accumulated by the accumulator (INT32). According to the configuration from registers, AddrGen generates the read/write commands and corresponding addresses for memory access. The central control module of NE is NE Ctrl, which is an FSM with one idle state and five working states. Each working state is for one instruction, and all the working states are independent. In addition to the sub-modules above, NE is also equipped with two utility sub-modules: PERF. for recording the execution cycles of NE, and DEBUG for enabling lock-step debug mode by clock gating.

Figure 29.1.6 illustrates the FPGA-based evaluation platform, comparison with prior PNM designs, and a performance evaluation of the HB chip and a conventional CPU-DRAM system on the recommendation system. The test board is capable to mount up to 4 HB chips, which can be tested simultaneously. The FPGA is responsible for writing/reading data and generating the configuration to/from the chip register. Thanks to the hybrid bonding, our HB-PNM chip outperforms prior PNM chips significantly, in terms of bandwidth-to-capacity ratio (307) and energy cost (0.88pJ/b). During the test, query embedding is uploaded to the chip via QSPI, and our PNM engine generates top-100 results out of the pre-loaded features in memory. The CPU for comparison is Intel Xeon Gold 5220@2.20GHz. With an industrial recommendation model, the theoretical throughput of our chip is 512QPS, and the measured performance is 401QPS. The performance gap arises from initialization overhead and memory subsystem overhead (random access, refresh, etc.). Compared to the CPU-DRAM system, our chip achieves 9.78× speedup. Note that the throughput and memory capacity can be further improved by scaling up the number of hybrid bonding blocks or using more advanced process technologies to serve more complicated recommendation models. In terms of energy efficiency, which is significant in memory-bound applications, our work achieves 184.11QPS/W, which outperforms the CPU-DRAM system by 317.43×. In terms of area efficiency, the high-density hybrid bonding improves QPS/mm² by 660×.

Figure 29.1.7 shows the die photos of DRAM die, NE and ME. The detailed specifications of the logic die and DRAM die are also listed in Fig. 29.1.7. The DRAM die is fabricated using 25nm technology within a 602.22mm² die area. The area of the DRAM blocks corresponding to NE and ME are both 32mm². The DRAM die operates with 1.1V supply at 150MHz, and the power consumption is 300mW/1Gb. The die-to-die bandwidth is 1.38TB/s in total, and the bandwidth corresponding to NE and ME is 153.60GB/s. To meet the dimension requirements of hybrid bonding technology, the logic die occupies exactly the same area as the DRAM die. The logic die is fabricated with 55nm technology, and it integrates NE, ME and MCs. The compute engines with INT8 precision occupy area: NE's area is 5.90mm², and ME's area is 7.02mm². Both NE and ME are equipped with 16 memory controllers. The logic die runs at 300MHz with 1.2V supply, and its power consumption is 977.70mW.

References:

[1] D. Lee et al., "A 128Gb 8-High 512GB/s HBM2E DRAM with a Pseudo Quarter Bank Structure, Power Dispersion and an Instruction-Based At-Speed PMBIST," *ISSCC*, pp. 334-336, 2020.

[2] M. Horowitz, "Computing's Energy Problem (and what we can do about it)," *ISSCC*, pp. 10-14, 2014.

[3] Y. C. Kwon et al., "A 20nm 6GB Function-In-Memory DRAM, Based on HBM2 with a 1.2TFLOPS Programmable Computing Unit Using Bank-Level Parallelism, for Machine Learning Applications," *ISSCC*, pp. 350-351, 2021.

[4] F. Devaux et al., "The True Processing in Memory Accelerator," *IEEE Hot Chips Symp.*, pp. 1-24, 2019.

978-1-6654-2800-2/22/$31.00

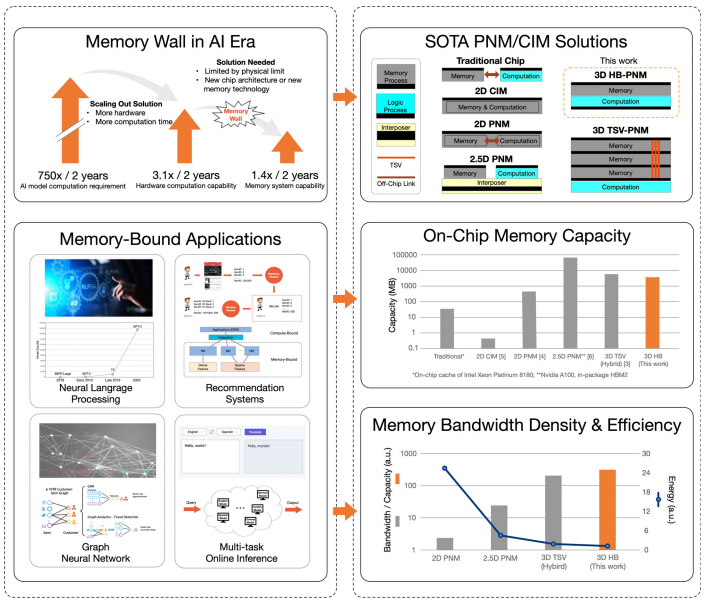
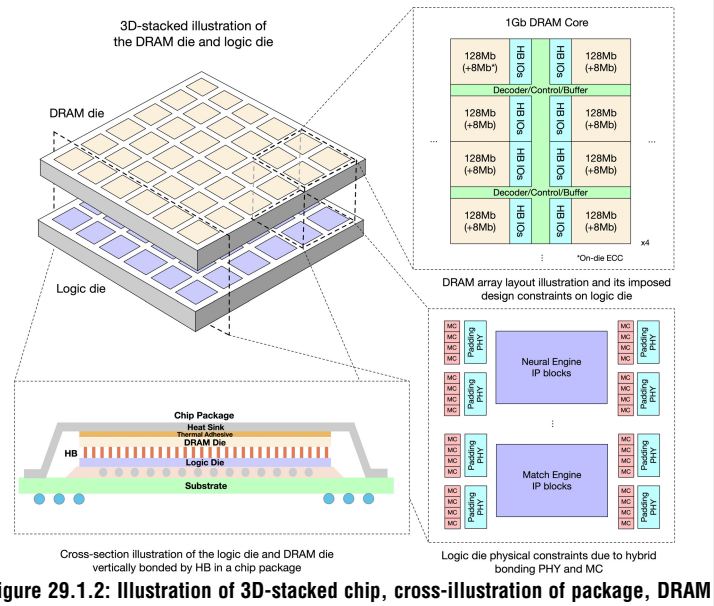


Figure 29.1.1: Motivations and comparison of state-of-the-art PNM/CIM architectures.



Cross-section illustration of the logic die and DRAM die vertically bonded by HB in a chip package

Logic die physical constraints due to hybrid bonding PHY and MC

Figure 29.1.2: Illustration of 3D-stacked chip, cross-illustration of package, DRAM array layout and design blocks on logic die.

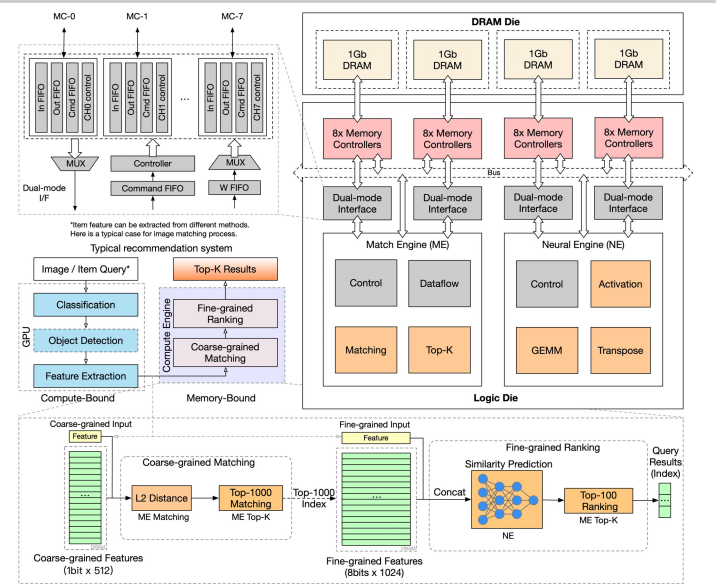


Figure 29.1.3: Overall architecture of PNM logic. Detailed flow of typical recommendation system.

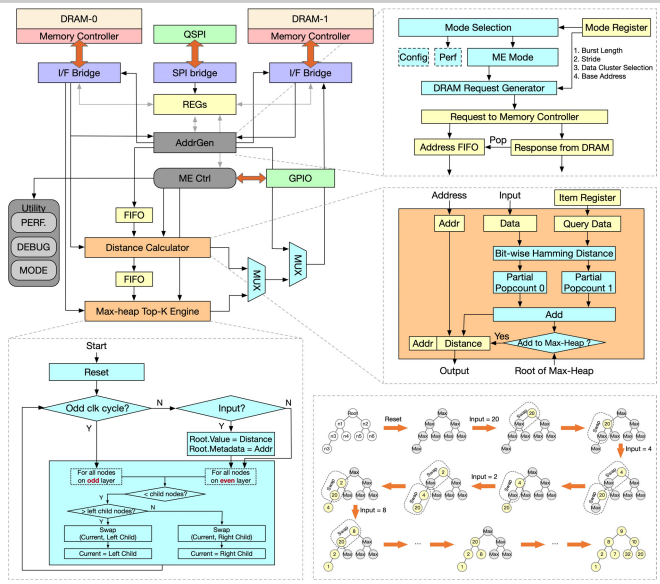


Figure 29.1.4: Detailed design of Match Engine (ME), showing internal data-path micro-architecture of AddGen, distance calculator, and top-K engine.

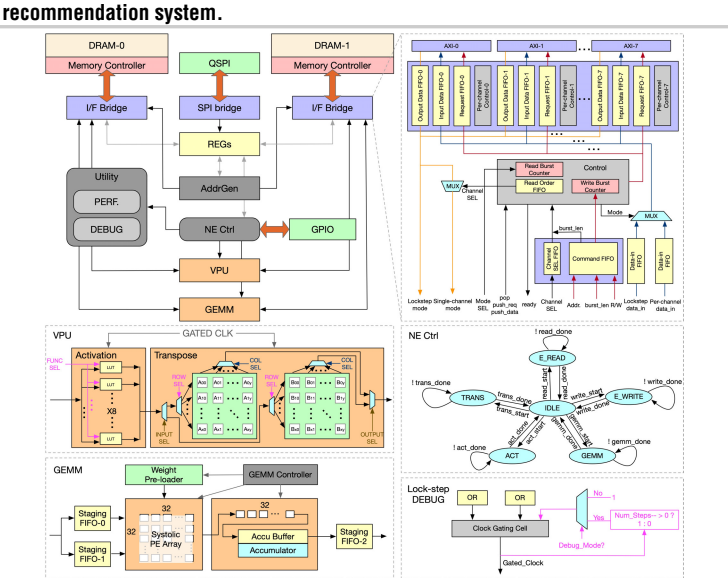
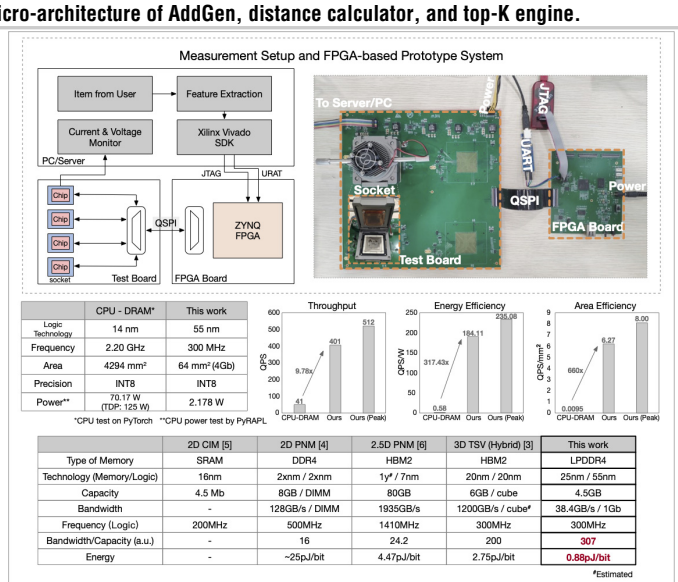


Figure 29.1.5: Detailed design of Neural Engine (NE), showing internal datapath, interface modules, micro architecture of VPU and GEMM, FSM of control modules and lock-step debug module.



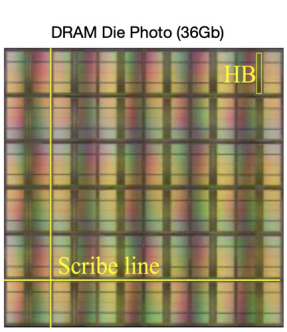
	CPU - DRAM*	This work
Logic Technology	14 nm	55 nm
Frequency	2.20 GHz	300 MHz
Area	4294 mm²	64 mm² (4Gb)
Precision	INT8	INT8
Power**	70.17 W (TDP: 125 W)	2.178 W

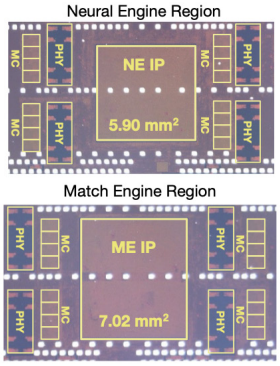
*CPU test on PyTorch **CPU power test by PyRAPL

	2D CIM [5]	2D PNM [4]	2.5D PNM [6]	3D TSV (Hybrid) [3]	This work
Type of Memory	SRAM	DDR4	HBM2	HBM2	LPDDR4
Technology (Memory/Logic)	16nm	2xnm / 2xnm	1y# / 7nm	20nm / 20nm	25nm / 55nm
Capacity	4.5 Mb	8GB / DIMM	80GB	6GB / cube	4.5GB
Bandwidth	-	128GB/s / DIMM	1935GB/s	1200GB/s / cube#	38.4GB/s / 1Gb
Frequency (Logic)	200MHz	500MHz	1410MHz	300MHz	300MHz
Bandwidth/Capacity (a.u.)	-	16	24.2	200	307
Energy	-	~25pJ/bit	4.47pJ/bit	2.75pJ/bit	0.88pJ/bit

#Estimated

Figure 29.1.6: Illustration of FPGA-based evaluation platform, comparison with prior near-memory processing designs, and end-to-end performance evaluation of our HB chip and CPU-DRAM system on recommendation application.





DRAM Die		
Technology	25nm	
Area	Total*	602.22 mm²
	Neural Engine	32 mm²
	Match Engine	32 mm²
Voltage	1.1 V	
Frequency (max)	150 MHz	
Power	300 mW per 1Gb	
Bandwidth**	153.60 GB/s / 1.38 TB/s	

Logic Die		
Technology	55nm	
Area	Total*	602.22 mm²
	Neural Engine	5.90 mm²
	Match Engine	7.02 mm²
# of MC	16 per IP	
Voltage	1.2 V	
Frequency	300 MHz	
Power	977.70 mW	
Precision	INT8	

* Total die size of test chip
** Memory bandwidth of NE and ME / Total bandwidth of DRAM die

Figure 29.1.7: Die micrographs of DRAM die, NE and ME. Detailed specifications of DRAM die and logic die.

Additional References:

[5] H. Jia et al., "A Programmable Neural-Network Inference Accelerator Based on Scalable In- Memory Computing," *ISSCC*, pp. 236-237, 2021.
[6] J. Choquette et al., "Nvidia A100 GPU: Performance & Innovation for GPU Computing," *IEEE Hot Chips Symp.*, pp. 1-43, 2020.
[7] B. Fujun et al., "A Stacked Embedded DRAM Array for LPDDR4/4X using Hybrid Bonding 3D Integration with 34GB/s/1Gb 0.88pJ/b Logic-to-Memory Interface," *IEDM*, pp. 123-126, 2020.

978-1-6654-2800-2/22/$31.00