ICCAD 2026 | IEEE/ACM International Conference on Computer-Aided Design

Tri-Stack-PIM: A Heterogeneous Multi-Die 3D-Integrated Processing-in-Memory Architecture for Efficient LLM Inference

Shuao Jia[1], Zichao Ling[1], Chen Bai[2], Kang Zhao[1], Jianwang Zhai[1]

[1]Beijing University of Posts and Telecommunications

[2]Fudan University

{pdxqjiashuao, zhaijw}@bupt.edu.cn

Nov 10, 2026

北京郵電大學 Beijing University of Posts and Telecommunications

FUDAN UNIVERSITY 復旦大學

Outline

01 Introduction

02 Tri-Stack-PIM architecture

03 Evaluation

04 Conclusion

Introduction

One KV fetch feeds G query heads; batching adds B independent requests

The KV cache grows with context length; every token re-reads it.
GQA and MQA increase the MACs per fetched byte through KV reuse, while batching exposes additional parallel work.

Attention	Query / KV heads	Group ratio G	KV reuse
MHA: OPT-6.7B	32 / 32	1	Separate KV per head
GQA: LLaMA3-8B	32 / 8	4	KV shared within a group
MQA: PaLM-8B	16 / 1	16	One KV head for all queries

More MACs per fetched byte must come from the logic die

Prior PIM designs starve the compute die

Design	Advanced-node compute	In-stack softmax	Compute / I/O decoupling	Dynamic scheduling
AttAcc [9]	No	Yes	No	No
H2-LLM [5]	Yes	No	No	No
Duplex [11]	Yes	Yes	No	Yes
Tri-Stack-PIM	**Yes**	**Yes**	**Yes**	**Yes**

- **AttAcc** — compute sits in the DRAM die, far too little for $G \times B$
- **H2-LLM** — softmax runs on the host; the round trip cancels the gain

HB I/O and per-bank keep-out zones compete with compute logic: a fragmented floorplan, and less area left for MACs

[5] Cong Li, Yihan Yin, Xintong Wu, Jingchen Zhu, Zhutianya Gao, Dimin Niu, Qiang Wu, Xin Si, Yuan Xie, Chen Zhang, et al. 2025. H^2-LLM: Hardware-dataflow co-exploration for heterogeneous hybrid-bonding-based low-batch LLM inference. In *Proc. ISCA*. 194–210.

[9] Jaehyun Park, Jaewan Choi, Kwanhee Kyung, Michael Jaemin Kim, Yongsuk Kwon, Nam Sung Kim, and Jung Ho Ahn. 2024. AttAcc! Unleashing the power of PIM for batched transformer-based generative model inference. In *Proc. ASPLOS*, Vol. 2. 103–119.

[11] Sungmin Yun, Kwanhee Kyung, Juhwan Cho, Jaewan Choi, Jongmin Kim, Byeongho Kim, Sukhan Lee, Kyomin Sohn, and Jung Ho Ahn. 2024. Duplex: A device for large language models with mixture of experts, grouped query attention, and continuous batching. In *Proc. MICRO*. 1429–1443.

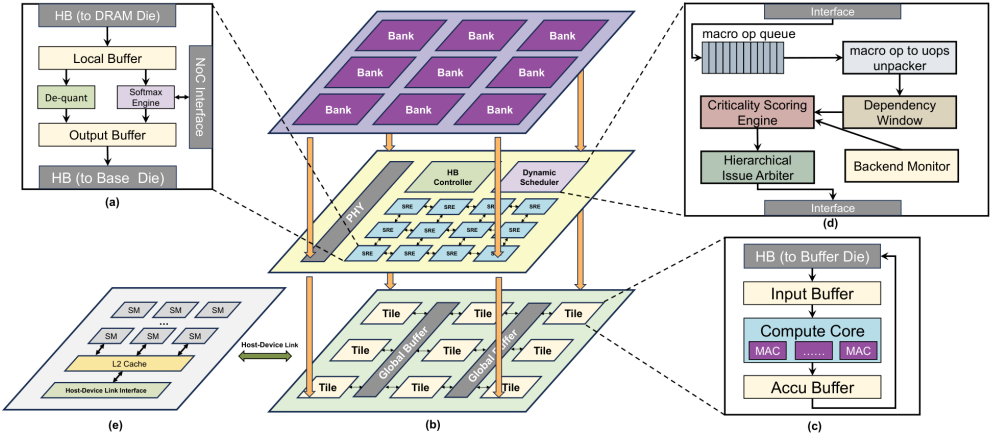
Motivation for a buffer die

Layer	Primary role
DRAM	Weights and KV cache
Buffer die	I/O, control and SREs
Bottom die	SIMD MACs and SRAM

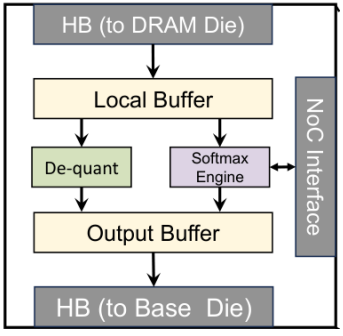
- **Bank I/O aggregation** — Per-bank bond pads are collected on the buffer die and re-routed once, instead of fanning out across the compute die.
- **Regular floorplan** — Scattered DRAM I/O no longer chops the compute region into fragments, so placement and routing close on a regular grid.
- **Mature-node control** — PHY, DMA and the SREs leave the advanced node: the area and the routing they would occupy go to MAC arrays and local SRAM.

Cost — one extra routing layer: TSV keep-out, plus fan-out wires that can exceed 3 mm.

Tri-Stack-PIM Architecture



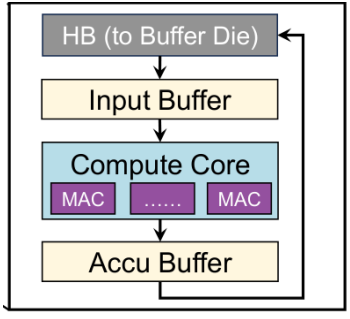
- **Top die:** DRAM storage (weights and KV cache)
- **Buffer die:** interface, control and SREs
- **Bottom die:** advanced-node compute tiles with local SRAM



Mature-node buffer die

- **Cross-layer re-routing** — Per-bank I/O is collected here and re-routed down to the compute die
- **Interface and control** — Host-facing PHY, HB controller, DMA and per-channel scheduling
- **Softmax and Reduction Engine (SRE)** — One SRE per bank, with local SRAM and a channel-local mesh NoC
- **In-stack nonlinear computation** — Cross-bank score reduction and streaming softmax avoid host round trips

Bottom die: data-triggered SIMD compute



Advanced-node compute

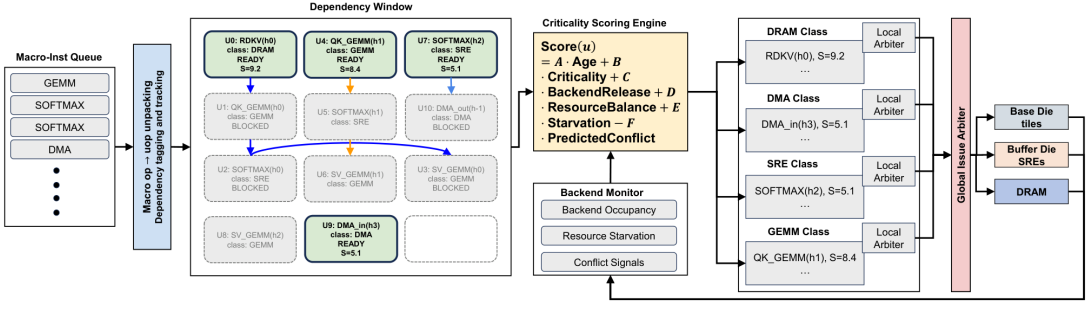
- **Local SRAM** — Stages incoming KV bursts and signals backpressure when full
- **Pipelined MAC array** — The same SIMD lanes execute QK^T and the weighted value sum
- **Per-burst compute cycles** — $\lceil (W_{HB}/b_{quant}) \times r/N_{lanes} \rceil$, where r is the query-head reuse count

Macro-instruction	Function
LAUNCH_GEMM	Tiled matrix–vector computation with query-head repeats
SRE_SOFTMAX	Softmax accumulation and partial-sum reduction
DMA_H2D	Host-to-device tile transfer
MODE_SWITCH	Global barrier between execution phases

Compiler responsibility

Operator placement and macro dependencies encoded by tag / pre_tag

Hardware expands macros and schedules dependency-ready micro-operations



Readiness

Dependency count reaches zero

Issue also requires valid DRAM timing and available backend resources

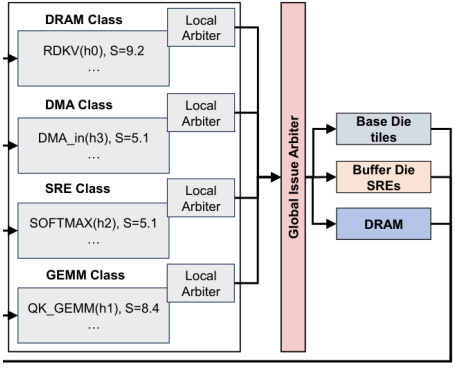
Priority among ready micro-operations

$$S(u) = \alpha T_{age} + \beta T_{critical} + \gamma T_{backend} + \delta T_{balance} + \varepsilon T_{starvation} - \zeta T_{conflict}$$

Score term	Priority effect
Age	Favors older entries with a saturation limit
Criticality	Favors statically assigned critical operation classes
Backend release	Prioritizes reads when the GEMM input FIFO is low
Resource balance	Discourages over-represented issue classes
Starvation	Raises priority after a waiting threshold
Predicted conflict	Penalizes DRAM row-buffer misses

Backend feedback helps keep the bottom-die MAC arrays supplied

Two-stage issue arbitration



- **Stage 1: local selection** — One ready champion from each class: DRAM, DMA, SRE and GEMM
- **Stage 2: global selection** — Choose the highest-scoring champion
- **Resource-check fallback** — Try another champion when the first candidate cannot issue

Domain	Explored parameters
Dataflow	KV head-to-bank mapping and sequence sharding
Bottom die	SIMD: 32–512 lanes; clock: 0.5–1.5 GHz Input SRAM: 32 KB–8 MB
Buffer die	Clock: 0.1–1.0 GHz; SRE SRAM: 32 KB–8 MB
Interconnect	HB width: 128–4096 bits

Multi-objective Bayesian optimization

Ax / BoTorch [2,8] minimizes latency, die cost and power jointly

Hardware capacity and workload mapping must match each other

[2] Maximilian Balandat et al. 2020. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization.In *Advances in Neural Information Processing Systems 33 (NeurIPS)*. arXiv:1910.06403.
[8] Miles Olson, Elizabeth Santorella, Louis C. Tiao, et al. 2025. Ax: A Platform for Adaptive Experimentation. In *Proc. 2nd AutoML Conference (JMLR W&CP)*, Vol. 293.

Evaluation

Workload	Configuration
Models	OPT-6.7B (MHA), LLaMA3-8B (GQA), PaLM-8B (MQA)
Batch sizes	1, 4, 8, 16, 32
Prompt / generation lengths	(128, 64), (512, 256), (1024, 64), (1024, 16)
Baselines	GPU-only, AttAcc [9] and H2-LLM [5]
Implementation models	Ramulator 2 [6]; ASAP7 [1] / NanGate 45 nm [7] synthesis FinCACTI [10] SRAM and HotSpot 7.0 [3,4] thermal analysis

Metric: average per-token attention latency over 32 decoder layers

[1] Lawrence T. Clark, Vinay Vashishtha, Lucian Shifren, Aditya Gujja, Saurabh Sinha, Brian Cline, Chandarasekaran Ramamurthy, and Greg Yeric. 2016. ASAP7: A 7-nm finFET predictive process design kit. *Microelectronics Journal* 53 (2016), 105–115.

[3] Jun-Han Han, Kevin Skadron, et al. 2021. HotSpot 7.0: Enhanced Compact Thermal Modeling for 3D-IC Cooling Solution Exploration. *ACM SIGDA University Demonstration at DAC*.

[4] Wei Huang et al. 2006. HotSpot: A Compact Thermal Modeling Methodology for Early-Stage VLSI Design. *IEEE Trans. VLSI Systems* 14, 5 (2006), 501–513.

[5] Cong Li, Yihan Yin, Xintong Wu, Jingchen Zhu, Zhutianya Gao, Dimin Niu, Qiang Wu, Xin Si, Yuan Xie, Chen Zhang, et al. 2025. H^2-LLM: Hardware-dataflow co-exploration for heterogeneous hybrid-bonding-based low-batch LLM inference. In *Proc. ISCA*. 194–210.

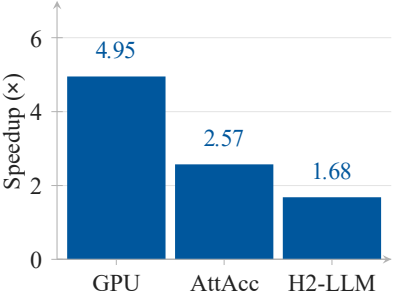
[6] Haocong Luo, Yahya Can Tugrul, F. Nisa Bostanci, Ataberk Olgun, A. Giray Yaglikci, and Onur Mutlu. 2023. Ramulator 2.0: A modern, modular, and extensible DRAM simulator. *IEEE Computer Architecture Letters* 23, 1 (2023), 112–116.

[7] Nangate Inc. 2011. Nangate Open Cell Library. Version 1.3, standard-cell library for the FreePDK45 technology.

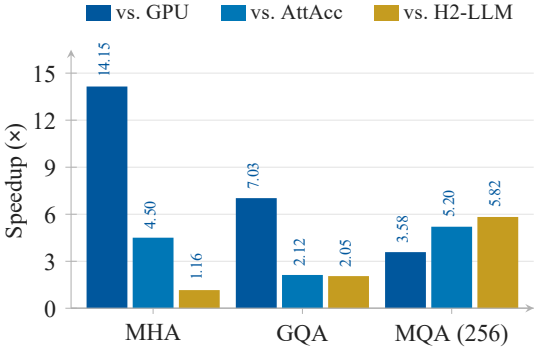
[9] Jaehyun Park, Jaewan Choi, Kwanhee Kyung, Michael Jaemin Kim, Yongsuk Kwon, Nam Sung Kim, and Jung Ho Ahn. 2024. AttAcc! Unleashing the power of PIM for batched transformer-based generative model inference. In *Proc. ASPLOS*, Vol. 2. 103–119.

[10] Alireza Shafaei, Yanzhi Wang, Xue Lin, and Massoud Pedram.2014. FinCACTI: Architectural analysis and modeling of caches with deeply-scaled FinFET devices. In *Proc. IEEE ISVLSI*. 290–295.

(a) Geometric mean, all workloads



(b) Per attention mechanism



Each bar uses the named design as its baseline. MHA and GQA use the default 64-lane configuration; the MQA results use 256 lanes.

In-order blocked cycles with another ready micro-operation (%)

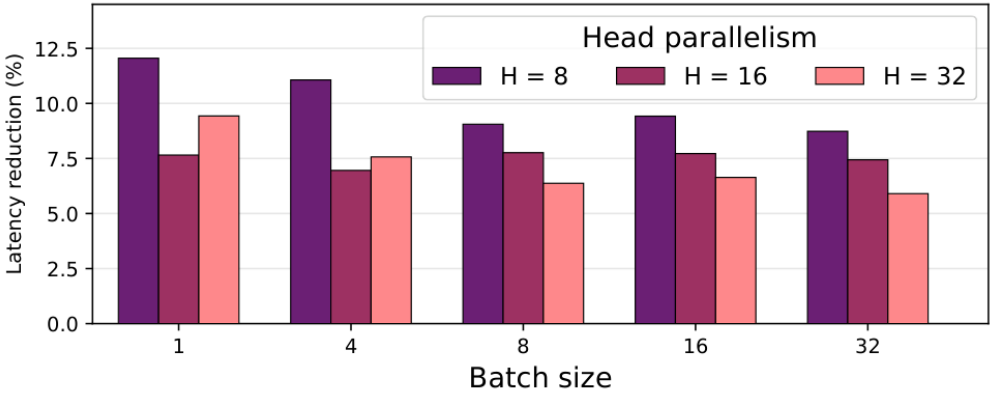
Batch / heads	1	8	16	32
1	3.8	**63.5**	**77.5**	**83.5**
4	44.8	**82.3**	**87.1**	**88.5**
8	61.1	**86.7**	**88.8**	**89.5**
16	76.2	**88.4**	**89.3**	**90.2**
32	81.2	**89.3**	**89.9**	**90.3**

Model S is the smallest of the three model-size panels of Fig. 5 (one panel per evaluated model); the M and L panels show the same trend.

Batch and head parallelism each expose independent work; both saturate by $B, H \geq 8$.

About 90% recoverable issue slack at high concurrency

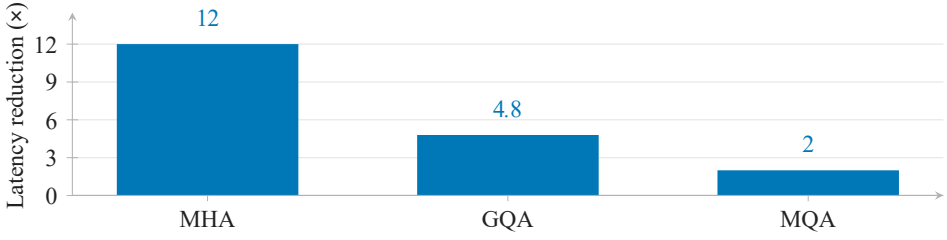
Scoring adds value beyond out-of-order issue



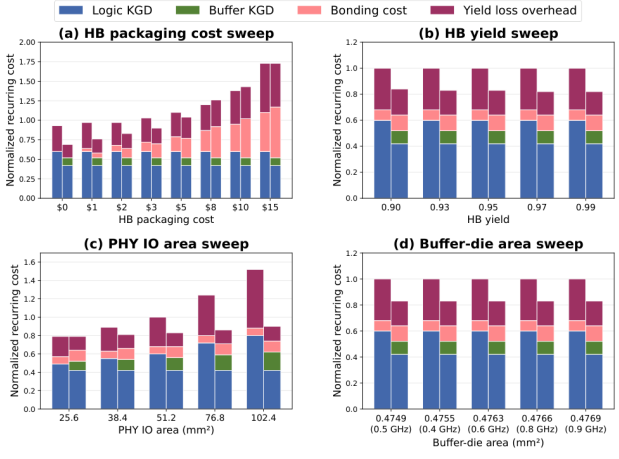
Matched hardware: SIMD 32, HB 2048 bits, bottom 500 MHz, buffer 1 GHz

Runtime scoring provides additional latency reduction over age-first issue

HB width: 128 to 4096 bits



- **SIMD width** — 32 to 256 lanes: 4.5× for MQA, but <1.1× for MHA / GQA
- **KV mapping** — 2.3×–9.6× latency variation at a fixed HB width
- **Joint exploration** — 33×–66× lower EDP than poor parameter choices

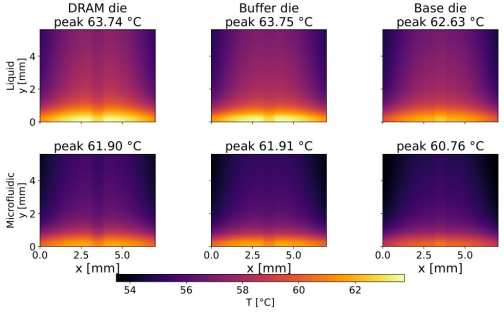


$$C_{RE} = C_{logic}/Y_{logic} + C_{buffer}/Y_{buffer} + C_{package}$$

The model excludes PHY redesign and qualification NRE.

- **Technology partitioning** — PHY wafer-cost contribution is 4.7× lower at the mature node.
- **Additional bonding step** — Cost benefit holds below about $5 per HB step in this model.

Thermal feasibility under modeled cooling



Cooling scheme	Peak stack temperature	DRAM limit
Advanced liquid cooling	63.75 °C	85 °C
Microfluidic cooling	61.91 °C	85 °C

HotSpot 7.0 [3,4] steady-state simulation at 11.97 W total stack power

[3] Jun-Han Han, Kevin Skadron, et al. 2021. HotSpot 7.0: Enhanced Compact Thermal Modeling for 3D-IC Cooling Solution Exploration. *ACM SIGDA University Demonstration at DAC*.
[4] Wei Huang et al. 2006. HotSpot: A Compact Thermal Modeling Methodology for Early-Stage VLSI Design. *IEEE Trans. VLSI Systems* 14, 5 (2006), 501–513.

Tri-Stack-PIM

- **Heterogeneous three-die partition** — DRAM storage, mature-node interface/control/SREs, and advanced-node SIMD compute
- **Dynamic micro-operation scheduling** — Dependency tracking and backend feedback reduce idle issue slots
- **Joint architecture–dataflow exploration** — Bandwidth, SIMD capacity and KV mapping determine the operating point
- **Attention speedup** — **4.95× vs. GPU-only, 2.57× vs. AttAcc, 1.68× vs. H2-LLM**

THANK YOU

Questions