香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

IEEE/ACM ICCAD 2025 INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN
44th Edition

AccelStack: A Cost-Driven Analysis of 3D-Stacked LLM Accelerators

Chen Bai[1] Xin Fan[1] Zhenhua Zhu[1,2] Wei Zhang[1] Yuan Xie[1]

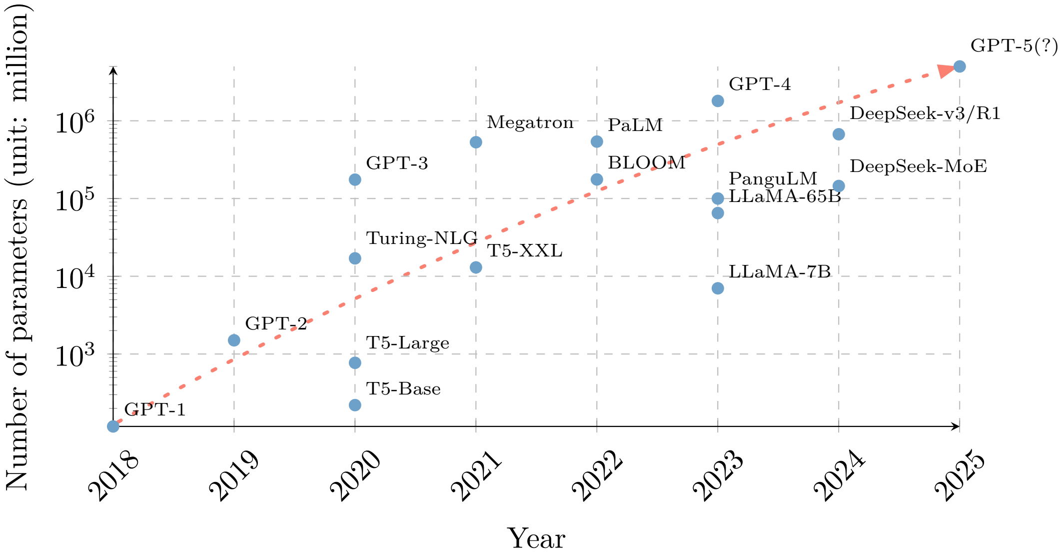
[1] The Hong Kong University of Science and Technology
[2] Tsinghua University

Oct. 28, 2025

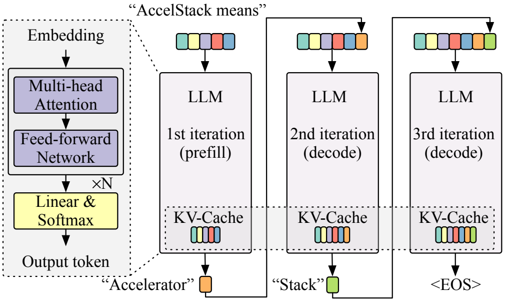
Outline

1. Introduction
2. AccelStack
3. Experiments
4. Conclusion

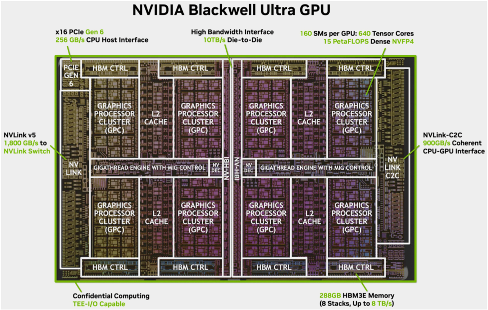
Introduction



The scaling of large langauge model (LLM).



Overview of LLM inference.



NVIDIA B200 die shot.

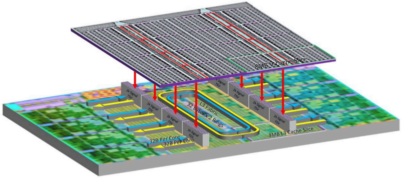
Innovations in LLM accelerators:

- Compute → 3D cube unit (Huawei HiSilicon), systolic arrays (Google), tensor cores (NVIDIA).
- Memory → High-bandwidth memory.
- Interconnect → Advanced 2.5D packaging.

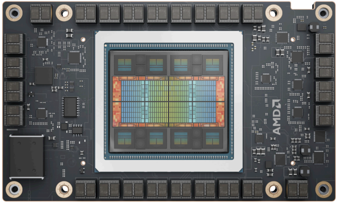
Introduction

Hybrid Bonding

It can stack chips with bumpless connections, *e.g.*, Cu-Cu bond[1].



AMD 3D V-Cache.



AMD MI355X.

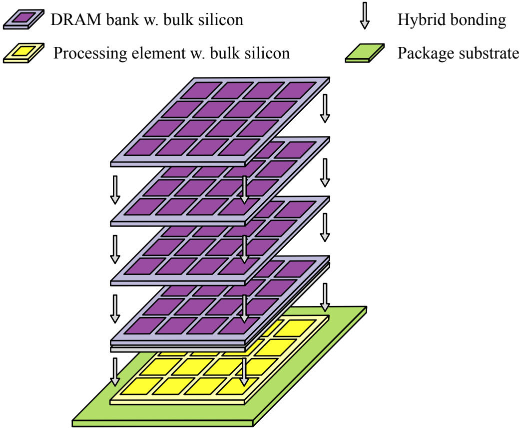
Compared to previous technology (microbump and 2.5D advanced packaging)[2]:

- $5 \sim 50\times$ improvement in pitches $\rightarrow$ Higher GB/s/mm^2.
- $5 \sim 20\times$ improvement in die-to-die distance $\rightarrow$ Smaller gate-to-gate delay.

[1] DB Ingerly et al. (2019). “Foveros: 3D Integration and the Use of Face-to-Face Chip Stacking for Logic Devices”. In: *IEEE International Electron Devices Meeting (IEDM)*. IEEE, pp. 19–6.

[2] John Wuu et al. (2022). “3D V-Cache: the Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU”. In: *IEEE International Solid-State Circuits Conference (ISSCC)*. vol. 65.

Introduction: 3D-Stacked LLM Accelerator



A possible architecture landscape in the future.

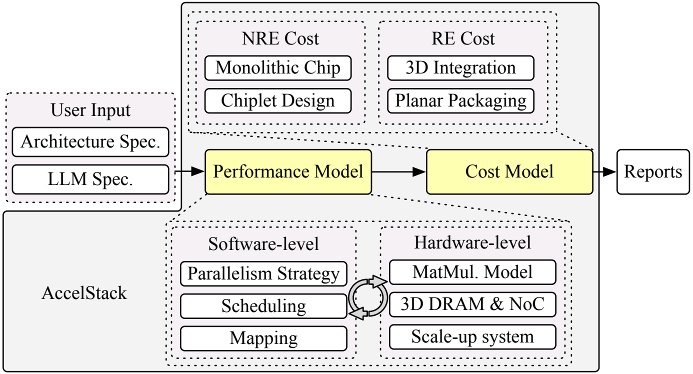
Problems:

- Performance: superiority over traditional LLM accelerators?
- Manufacturing cost: economic feasibility?

Can we decide on an investment in high-volume manufacturing (HVM) of the 3D-stacked LLM accelerators?

AccelStack

AccelStack: Cost-Driven Analysis of A 3D-Stacked Design

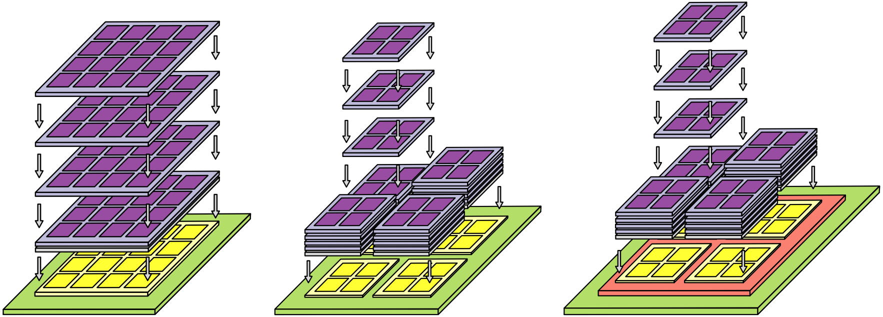


Overview of AccelStack.

In the following, we discuss:

- Focused 3D-stacked architecture assumptions.
- Performance model.
- Cost model.

Chipletized Design



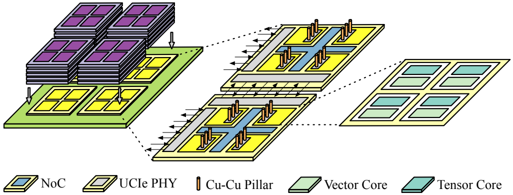
(a) Monolithic design (b) 3D chiplet design (c) 3.5D chiplet design

Chipletized versions of the architecture landscape.

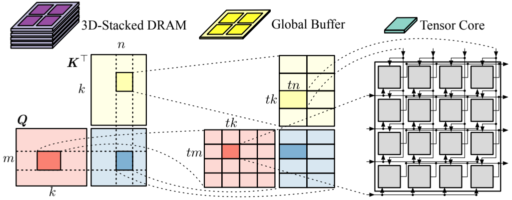
- Chiplet-to-chiplet interconnections are modeled according to the UCIe specification[3].
- We model CoWoS, EMIB, and MCM packaging techniques.

[3] *Universal Chiplet Interconnect Express 2.0 Specification* (2024). https://www.uciexpress.org/specifications.

Performance Model



(a) The microarchitecture of the 3D chiplet design.



(b) Two-level loop tiling for GEMM operations.

The tensor core is a $p \times p$ array.
P denotes the number of outputs per cycle.
α is the utilization.
B is the link bandwidth.

- Computing latency:

$$t_{\text{comp}} = \frac{m + tm - 1}{tm} \cdot \frac{n + tn - 1}{tn} \cdot \frac{k + tk - 1}{tk} \cdot \frac{2 \cdot (tm + p - 1) \cdot (tn + p - 1) \cdot tk}{p^2 \cdot \alpha \cdot P}. \quad (1)$$

- Communication latency:

$$t_{\text{comm}} = \frac{V}{B} + l + o, \quad V = \left\lceil \frac{n}{\texttt{payload}} \right\rceil \cdot \texttt{flit}. \quad (2)$$

The scale-up system is modeled with a Clos-based topology[4].

[4] Pengfei Zuo et al. (2025). "Serving Large Language Models on Huawei CloudMatrix384". In: *arXiv preprint arXiv:2506.12708*.

Cost Model (Non-recurring engineering cost, NRE)

$$C = C_{\text{NRE}} + C_{\text{RE}}, \tag{3}$$

We use a first-order approximation to evaluate C_{NRE}.
Monolithic design:

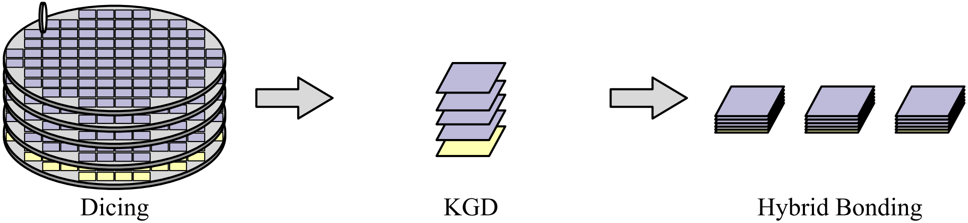
$$C_{\text{NRE}} = \left(\sum_{m_i \in \text{chip}} \alpha_{m_i} A_{m_i}\right) + \beta_{\text{chip}} A_{\text{chip}} + C_{\text{fixed}}. \tag{4}$$

Chiplet design:

$$C_{\text{NRE}} = \sum_{m_i \in M} \alpha_{m_i} A_{m_i} + \left(\sum_{c \in \text{chip}} \beta_c A_c + C_{\text{fixed}_c}\right) + C_{\text{fixed}}. \tag{5}$$

α_{m_i} is a cost coefficient. A is the silicon area. β accounts for physical design and verification efforts. C_{fixed} refers to the requisite fixed costs.

Cost Model (Recurring engineering cost, RE)



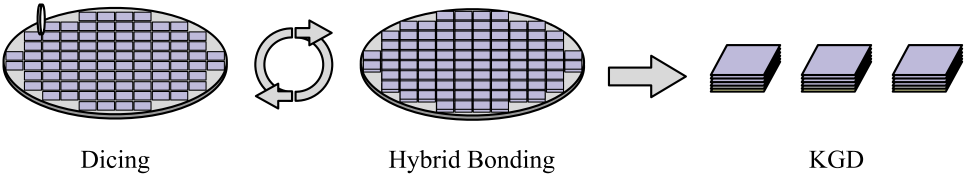
Die-on-die manufacturing flow (DoD).

$$C_{3D} = \frac{\sum_{i=1}^{n} C_{\text{die}_i} + (n-1)C_{\text{DoD}}}{\prod_{i=1}^{n-1} Y_{\text{DoD}}}, \tag{6}$$

where

$$C_{\text{die}_i} = \left(\frac{C_{\text{wafer}}}{N_{\text{die}}} + C_{\text{test}} + C_{\text{misc.}}\right)/Y_{\text{die}}. \tag{7}$$

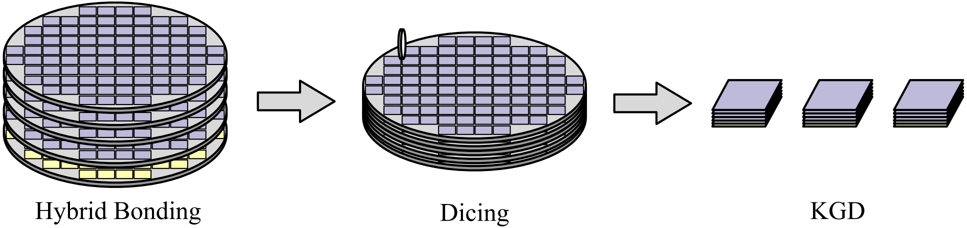
Cost Model (Recurring engineering cost, RE)



Die-on-wafer manufacturing flow (DoW).

$$
\begin{aligned}
C_{3D} &= (\frac{C_{\text{logic wafer}} + C_{\text{DoW}}}{N_{\text{logic}}} + C_{\text{DRAM}_{i+1}} + C_{\text{test}} + C_{\text{misc.}})/(Y_{\text{logic}} \cdot Y_{\text{DoW}}), \\
C_{\text{DRAM}_{i+1}} &= (\frac{C_{\text{DRAM wafer}} + C_{\text{DRAM}_i} + C_{\text{DoW}}}{N_{\text{DRAM}}} + C_{\text{test}} + C_{\text{misc.}})/Y_{\text{DoW}}.
\end{aligned} \tag{8}
$$

Cost Model (Recurring engineering cost, RE)



Wafer-on-wafer manufacturing flow (WoW).

$$
\begin{aligned}
C_{3D} &= \left(\frac{\sum_{i=1}^{n} C_{\text{wafer}_i} + (n-1)\cdot C_{\text{WoW}}}{\min\{N_{\text{logic}}, N_{\text{DRAM}}\}} + C'\right)/Y', \\
Y' &= Y_{\text{logic}} \cdot Y_{\text{DRAM}} \cdot \prod_{i=1}^{n-1} Y_{\text{WoW}}.
\end{aligned} \tag{9}
$$

Cost Model (Recurring engineering cost, RE)

CoWoS, EMIB, and MCM packaging:

$$\begin{aligned} C_{\text{package}} = & C_{\text{raw package}} + \\ & C_{\text{interposer}} \cdot \left(\frac{1}{Y_1 \cdot Y_2^n \cdot Y_3} - 1\right) + \\ & C_{\text{substrate}} \cdot \left(\frac{1}{Y_3} - 1\right) + \\ & C_{\text{3D}} \cdot \left(\frac{1}{Y_2^n \cdot Y_3} - 1\right). \end{aligned} \tag{10}$$

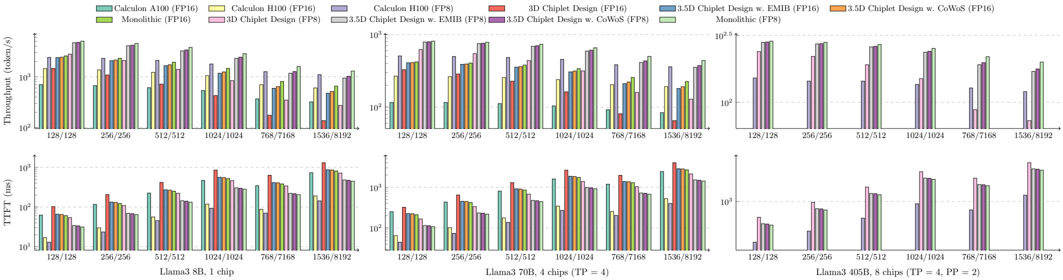
Y_1 is the yield of the interposer or bridge structure.
Y_2 is the bonding yield of chips.
Y_3 is the bonding yield of the interposer or bridge structure.

Experiments

Architecture configurations

Metric	Value
Silicon area of the monolithic design	32×25 mm^2
# of DRAM stacks	4
Total DRAM capacity	64 GB
Aggregate DRAM bandwidth	9.6 TB/s
# of processing elements (PEs)	16
Computing power	786 TFLOPS@FP8, 393 TFLOPS@FP16
NoC bisection bandwidth of the monolithic design	1.5 TB/s
NoC bisection bandwidth of 3D chiplet design w. CoWoS	1.1 TB/s
NoC bisection bandwidth of 3D chiplet design w. EMIB	1.0 TB/s
NoC bisection bandwidth of 3D chiplet design (MCM)	255.0 GB/s
Price ratio (DoD:DoW:WoW)	4 : 2 : 1
Hybrid bonding yield	0.95

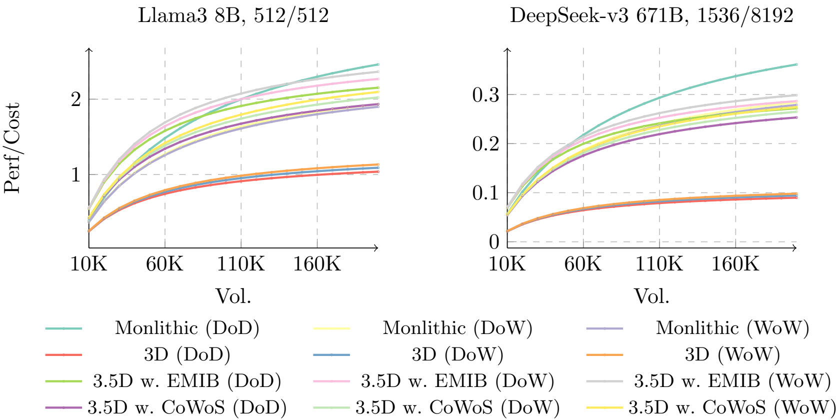
Compare w. A100/H100 Simulation Results



Compare w. A100 & H100 simulation results.

- The monolithic design vs. A100 (FP16) and H100 (FP8): 7.17× and 2.09× higher token/s.
- Slow-thinking (7 ∼ 8K output tokens), 3D-stacked design vs. H100 (FP8): 1.36× greater throughput.

Performance-per-cost Comparison

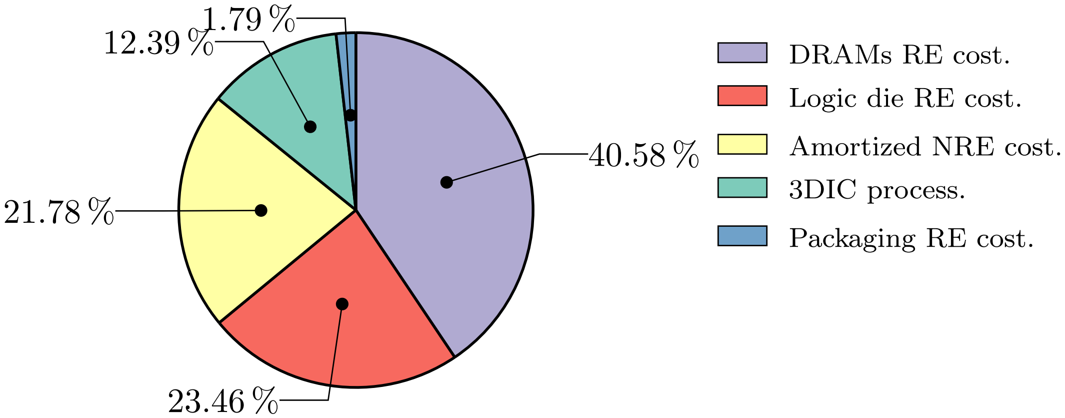


Performance-per-cost comparison.

For Llama3:

- The 3.5D chiplet design with EMIB using the WoW is the winner for shipment volumes below 140K (approximately 3500 logic wafers).
- Monolithic design with DoD is the winner for shipments more than 140K.

Detailed Cost Analysis



Cost breakdown for monolithic design with WoW under 200K shipment volumes.

- Much of the cost contributes to the 3D-stacked DRAM.

Conclusion

Conclusion

- A performance model for 3D-stacked LLM accelerators.
- A cost model supports DoD, DoW, and WoW hybrid bonding manufacturing flows.
- 3D-stacked accelerators can achieve up to 7.17× and 2.09× faster inference than A100 (FP16) and H100 (FP8) simulation results, with chiplet-based designs reducing recurring engineering costs by 38.09% versus monolithic implementations.
- Open-source address: `https://github.com/baichen318/accel-stack`!

THANK YOU!

DB Ingerly et al. (2019). "Foveros: 3D Integration and the Use of Face-to-Face Chip Stacking for Logic Devices". In: *IEEE International Electron Devices Meeting (IEDM)*. IEEE, pp. 19–6.

Universal Chiplet Interconnect Express 2.0 Specification (2024). https://www.uciexpress.org/specifications.

John Wuu et al. (2022). "3D V-Cache: the Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU". In: *IEEE International Solid-State Circuits Conference (ISSCC)*. Vol. 65. IEEE, pp. 428–429.

Pengfei Zuo et al. (2025). "Serving Large Language Models on Huawei CloudMatrix384". In: *arXiv preprint arXiv:2506.12708*.