



Intel Memory Summit

DRAM stacking for High Bandwidth Memory and CIM Tiles

Klaus Goebel, GEMS FTE EPAE PM GAR

July 2024



Novel DRAM stacking approach for High Bandwidth Memory and CIM Tiles

Currently existing solution: (Prior art)

HBM (High Bandwidth Memory) uses a stacked DRAM architecture. It involves placing multiple DRAM dies on top of a base logic die. This is currently done by stacking KGD (known good dies) and facilitating the D2D (Die-to-Die) interconnect by TSV's and uBump solder connections. Future solutions (HBM4 onwards) are expected to replace uBump solder connections by hybrid bond technology. The existing solution is expensive due to a significant number of costly process steps required to enable the stacking of multiple DRAM dies, such as formation of TSV's, TCFC (thermos compression flip chip) bonding and yield losses associated with a complex packaging process. The use of D2D hybrid bonding technology is expected in future (HBM5 or later) to enable stacking an even higher number of DRAMs. But this comes with significant challenges and will likely increase packaging cost even further.

Novel DRAM stacking approach for High Bandwidth Memory: (s. Fig 1)

The proposed multi-die-stacking is achieved by a two-step stacking approach: First WoW (Wafer-on-Wafer) hybrid bonding to combine two memory dies to a “DRAM sandwich”. In a second process the “DRAM sandwiches” are stacked by applying uBumps followed by TCFC (thermos compression flip chip) bonding or reflow process.

There are several advantages of the proposed HBM/DRAM stacking approach:

- Use of mature technologies: WoW hybrid bonding and TCFC (thermos compression flip chip) bonding are proven process technologies, which help to improve cost and time-to-market.
- The symmetric F2F stacking of DRAM wafers is expected to provide very good planarity and reduced warpage in following process steps.
- Lower z-height of a stacked DRAM module or (alternatively) a greater number of stacked dies at the same z-height.
- Improved thermal characteristics as hybrid bonding interfaces provide an excellent thermal path.

DRAM stacking approach for CIM (Compute-in-Memory) Tiles: (s. Fig.2)

Instead of bonding two DRAM wafers to a “DRAM sandwich”, it is also possible to use WoW hybrid bonding to create CIM (Compute-in-Memory) tiles, which form a “DRAM+Logic sandwich” tile.

The use of CIM tiles can be particularly advantageous in AI, Graphics, Simulations, and other related applications which require fast and very energy-efficient access to huge amounts of data. It is therefore possible to create a CIM tile which includes e.g., a complete GPU + DRAM-memory. To allow SW developers to make efficient use of CIM capabilities, it is important to standardize the interface and functionality of the CIM tile, and to consider this for the development of respective compiler architectures.

There are several advantages of the proposed DRAM+Logic (CIM) stacking approach:

- Shortest possible data path between compute- and memory resulting in a significantly improved power efficiency.
- Since most operations (e.g., MAC/Matrix/Vector operations, etc.) can run locally within the CIM tile, the requirements for bandwidth of the external interface can be reduced.
- Logic functionality which is typically implemented in the DRAM die (e.g., ECC, Self-Refresh, etc.) can be moved to the logic die and can be implemented more efficiently.

Example Solution (Data Center GPU): (s. Fig.3) and **Proposed Memory Architecture & Hierarchy** (s. Fig.4)

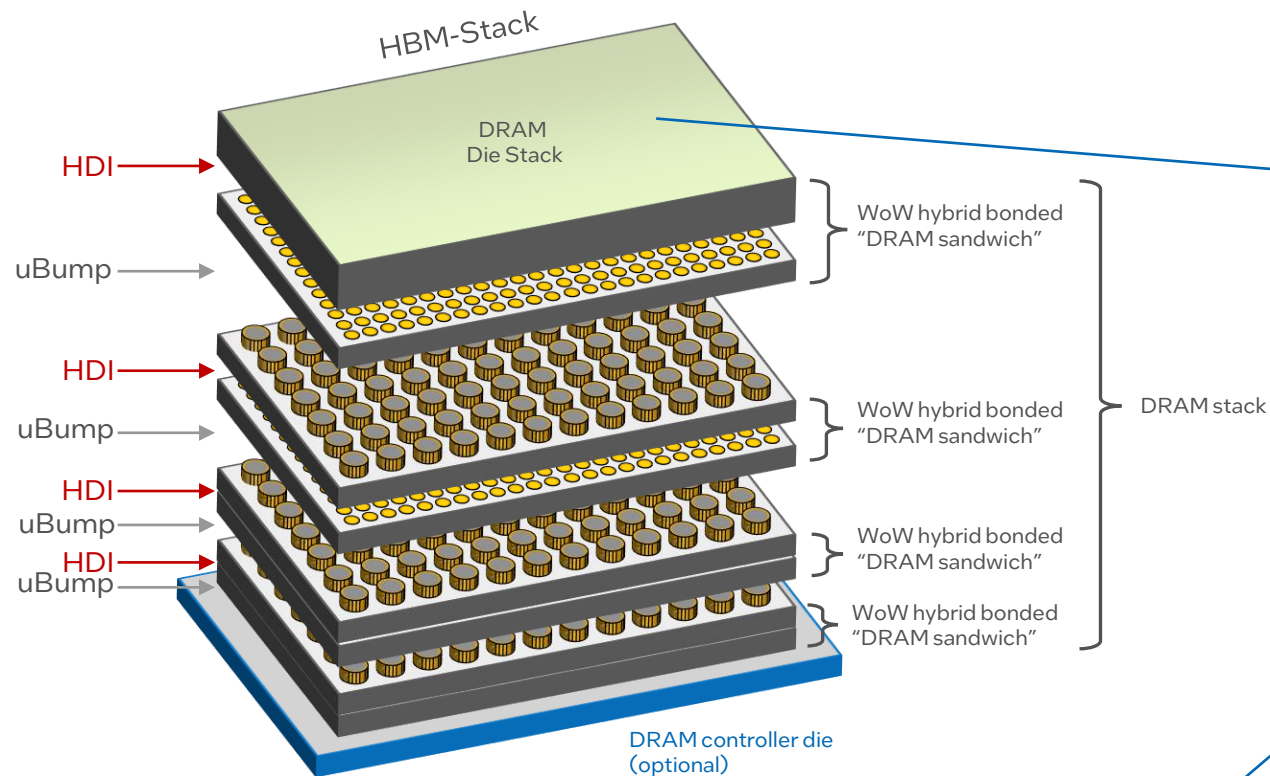


Fig. 1: HBM stack

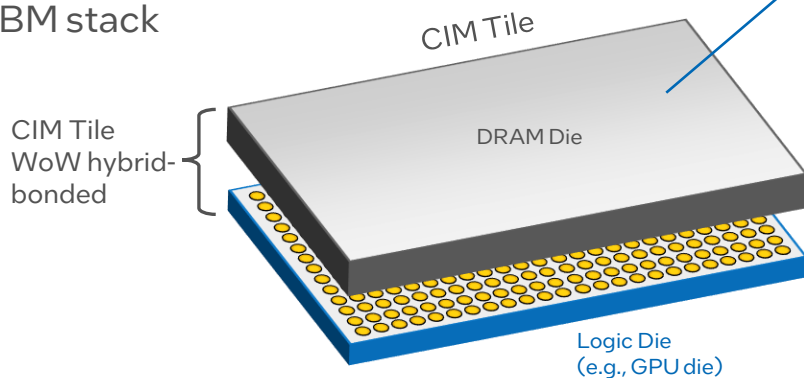


Fig. 2: CIM tile

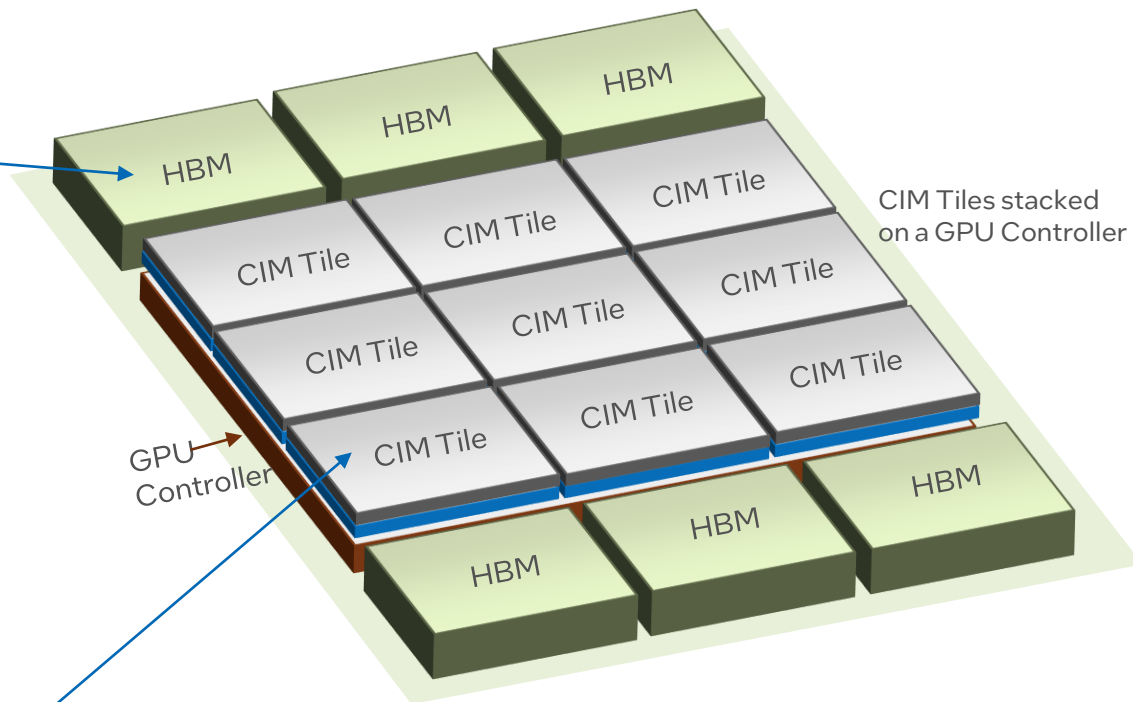


Fig. 3: Example Application (Data Center GPU)

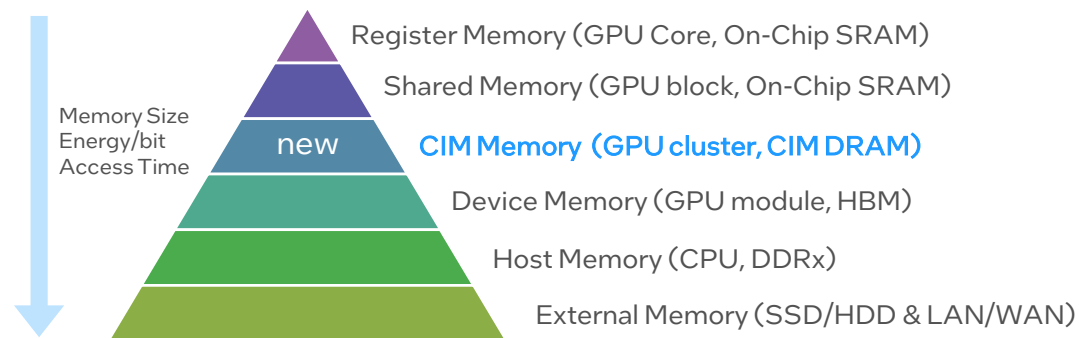


Fig. 4: Memory Architecture/Hierarchy

The Intel logo is centered on a dark blue background. It features the word "intel" in a white, lowercase, sans-serif font. A small, bright blue square is positioned above the letter "i". To the right of the word "intel" is a registered trademark symbol (®).

intel®