CONFIDENTIAL

Custom Memory Solutions for AI Applications

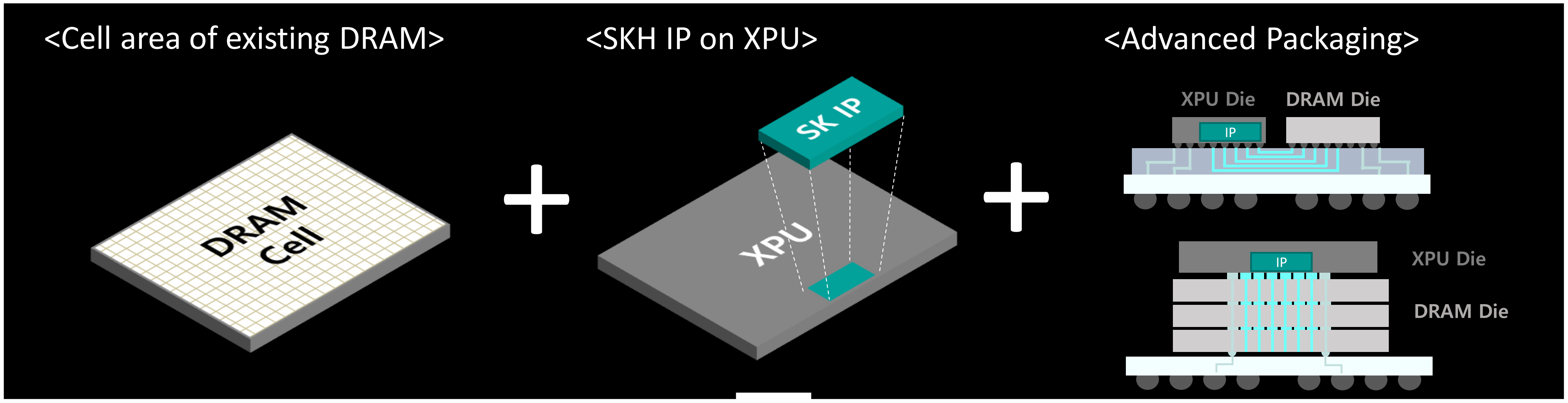
JAN 2024
US Engineering Center
Seongju Lee

SK hynix

Specifications and designs are subject to change without notice.
All data were deemed correct at time of creation. SK hynix is not liable for errors or omissions.

Specialties of Our Solutions 1 : Competitive Price, Fast Time-to-Market

CONFIDENTIAL



<Cell area of existing DRAM>

<SKH IP on XPU>

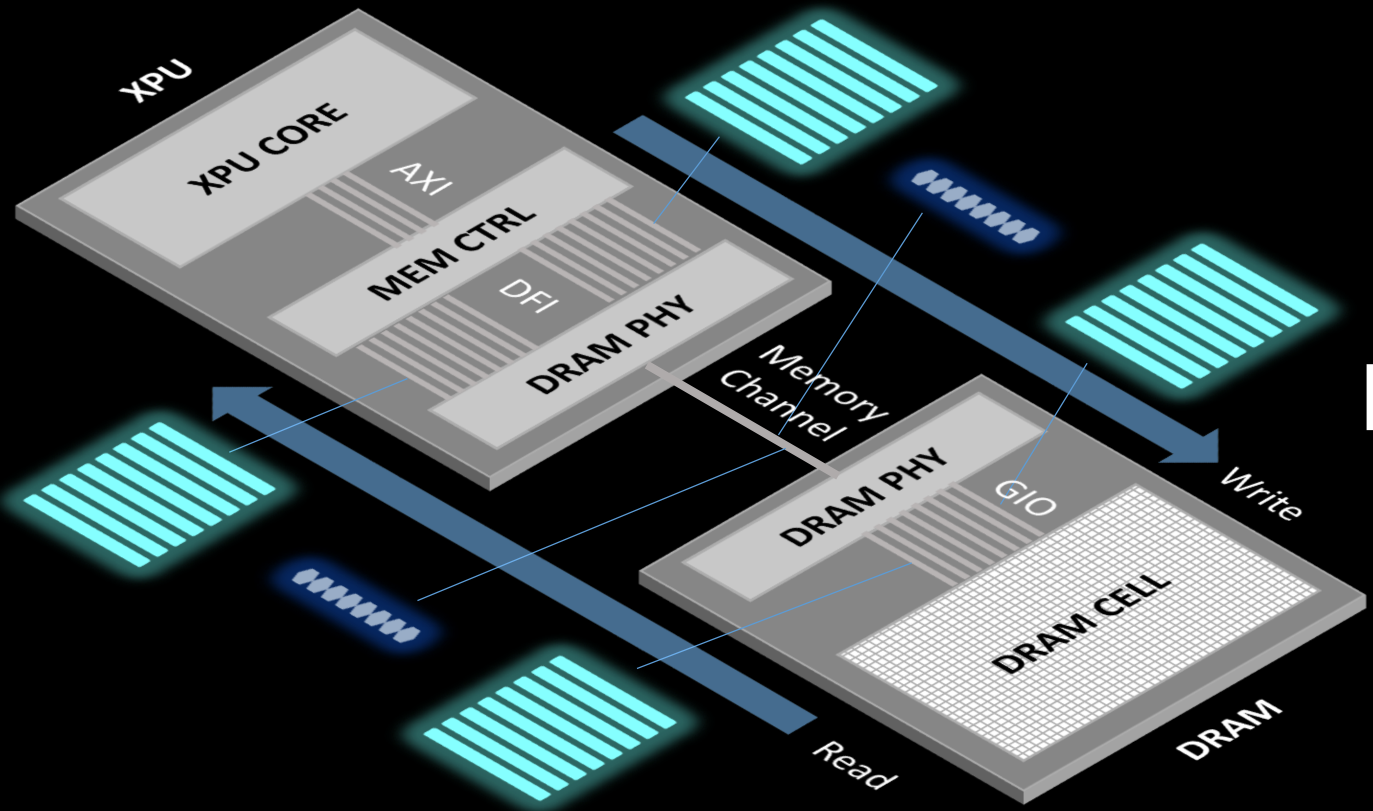
<Advanced Packaging>

Various Custom Memories

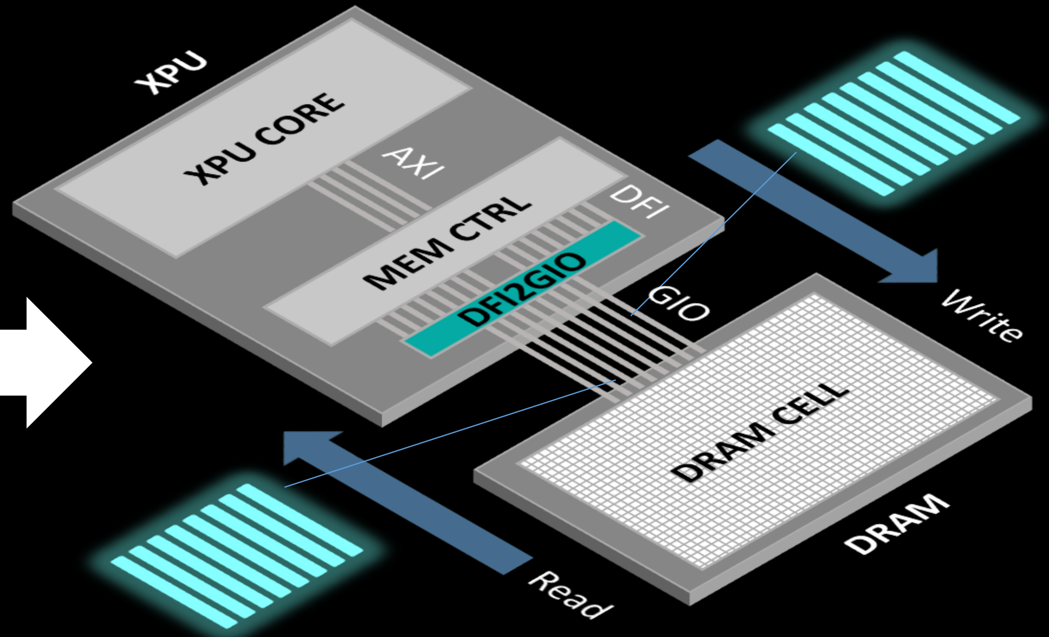
- Using cell area of existing DRAM without design modifications
- SKH IP integrated into XPU to operate the cell area
- DRAM & XPU are integrated in a package using advanced packaging technologies
- No modification on DRAM leads competitive price and fast time-to-market even though custom features

Specialties of Our Solutions 2 : Power & Latency Reduction

<Conventional Memory Subsystem>



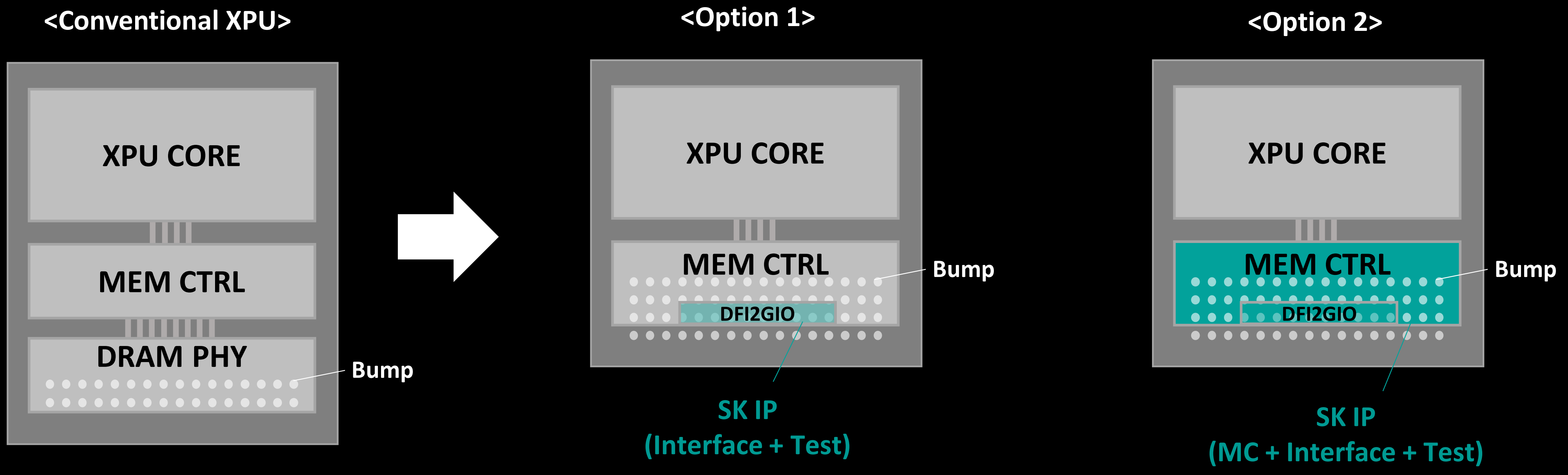
<Our Solutions>



- Tightly-coupled architecture of XPU and memory without complex high-speed PHYs on both sides
- PHY-less architecture reduces over 30% of power※ and 10~20ns latency

※ Memory related power (Memory Controller + Media)

Specialties of Our Solutions 3 : XPU Size Reduction

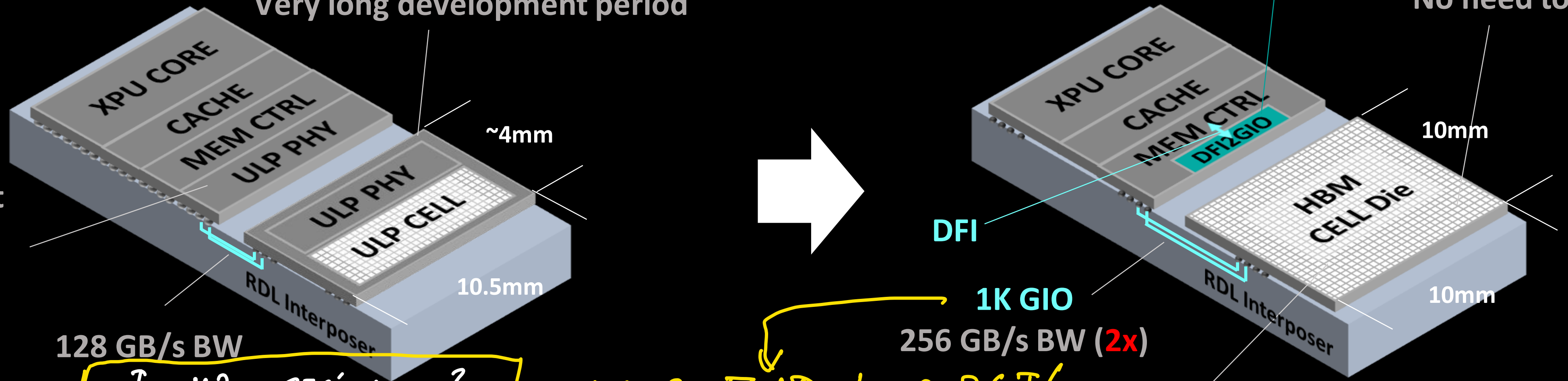


- High-speed complex PHY is located across the bump area

- In PHY-less architectures, the area between the bumps is almost empty except for simple CMOS TX/RX
- Memory controller w/ DFI2GIO can be located the empty area; overall XPU size can be reduced
- Two options are available;
  - ✓ **option 1** - SKH to provide DFI2GIO IP
  - ✓ **option 2** - SKH to provide memory controller IP including DFI2GIO

Custom Memory Solution 1. LPHBM for Edge Devices – XR, On-Device AI

CONFIDENTIAL



<Conventional ULP>

2Gb Capacity
Very long development period

XPU Must Develop dedicated ULP PHY

128 GB/s BW

<LPHBM>

SK IP
(Interface + Test)

24Gb Capacity (12x)
No need to new develop

DFI

1K GIO
256 GB/s BW (2x)

No Base Die,
No TSVs

Is ULP = SEC's LLC ?
x512 LPDDR @ 2GT/s

May/3/2022
2Gb
6841×3193
Xµm Yµm

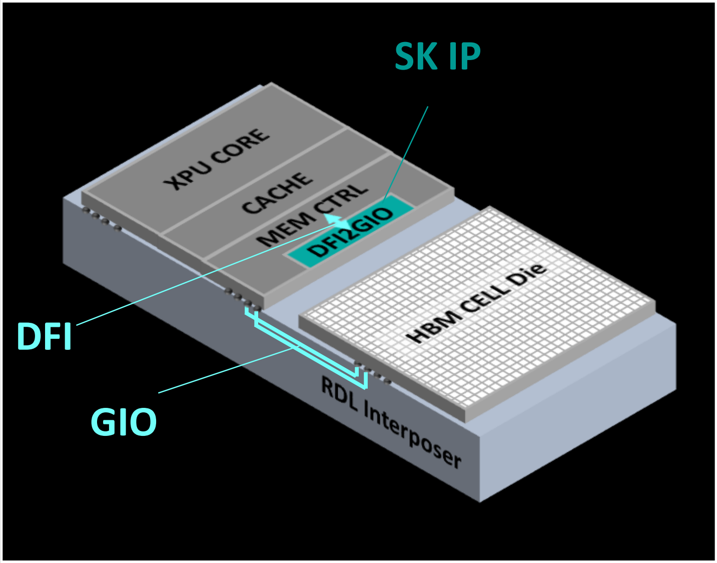
Full Duplex @ 2GT/s
UCIe is FD.@ 4GT/s

- SK developed several ULPs, low power/high bandwidth memories, required large resources and development of dedicated custom ULP PHY
- Our solution, LPHBM, **HBM cell die w/o base die** is used as a media and **Interface IP(DFI2GIO**) is placed on XPU to operate it
- Power/latency can be significantly improved by eliminating the PHYs from both of XPU & DRAM
- Fast time to market by developing Interface IP only, rather than new DRAM media

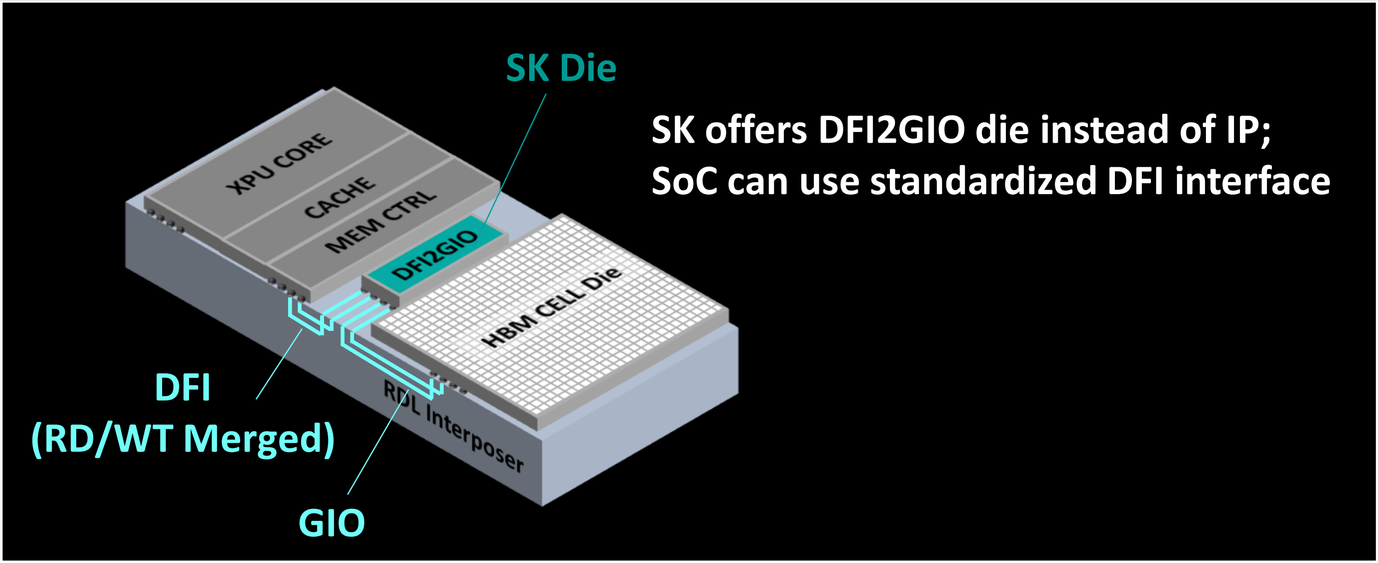
SK hynix

Custom Memory Solution 1. LPHBM for Edge Devices – XR, On-Device AI

CONFIDENTIAL

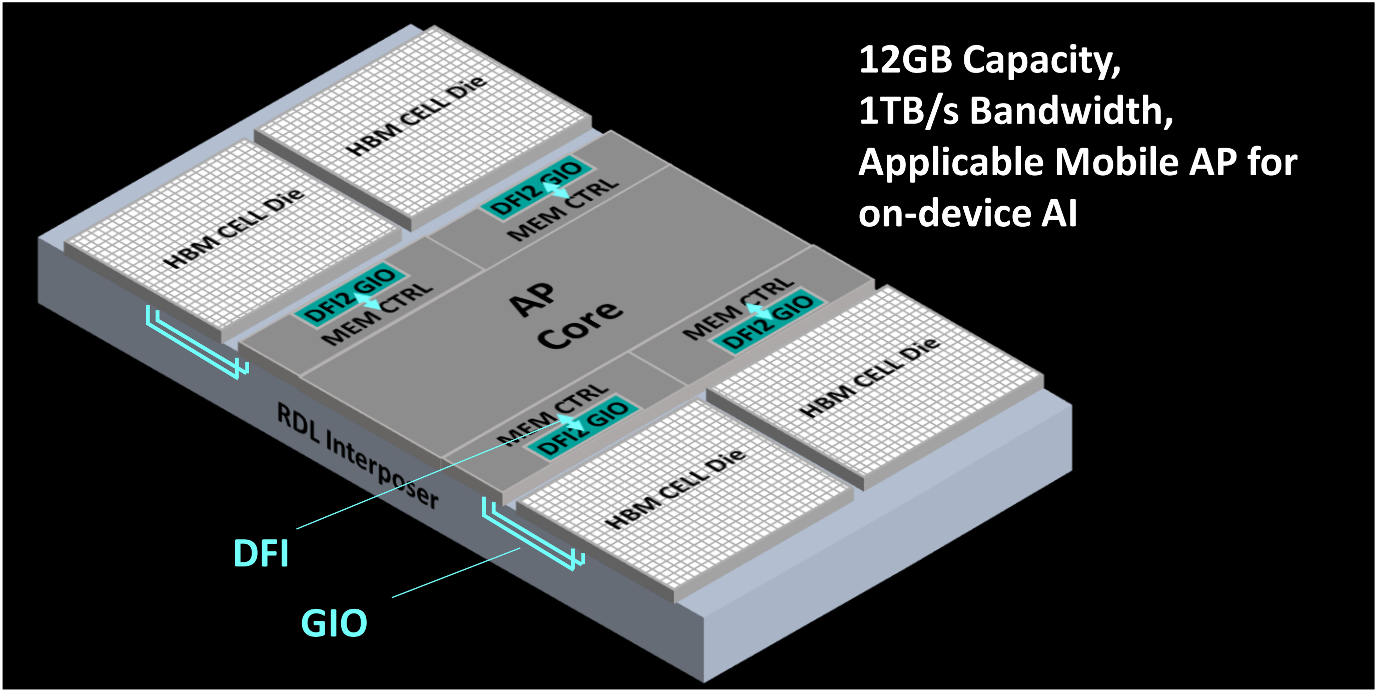


Standardized SoC Interface



SK offers DFI2GIO die instead of IP; SoC can use standardized DFI interface

Capacity & Bandwidth Expansion

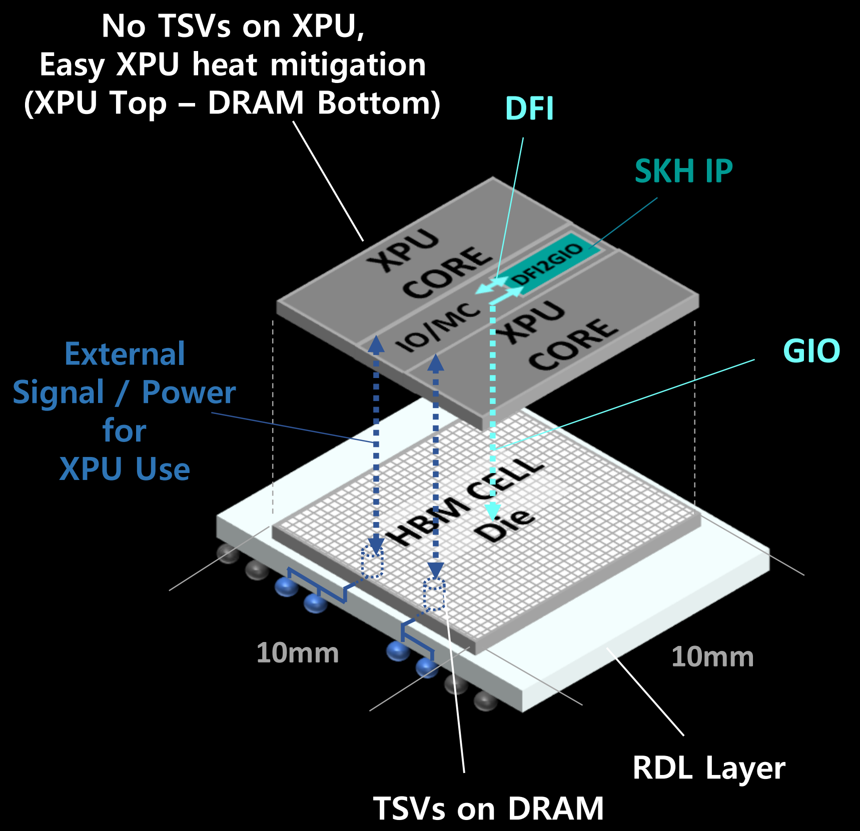


12GB Capacity,
1TB/s Bandwidth,
Applicable Mobile AP for on-device AI

- LPHBM can be implemented in a variety of ways depending on the needs

SK hynix

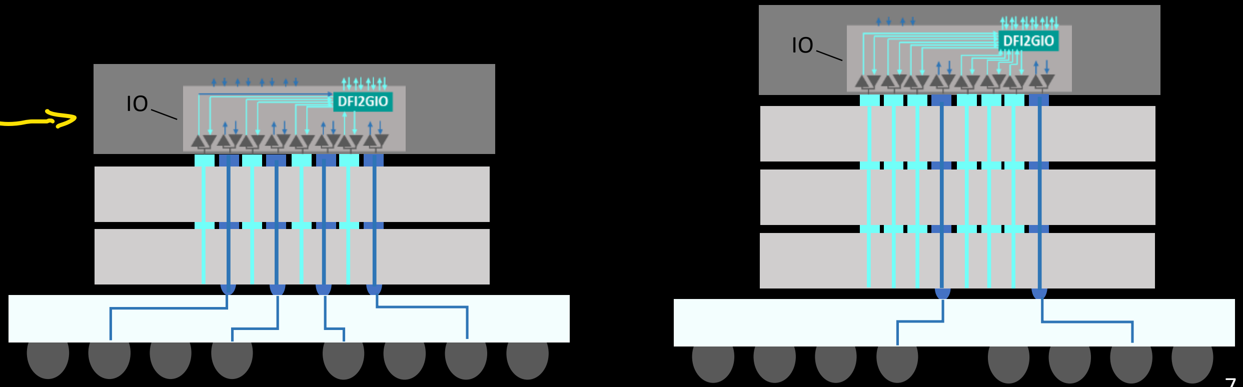
Custom Memory Solution 2. 3D LPHBM for Edge Devices – XR, On-Device AI



- 3D-structure can further increase bandwidth and decrease power
- Interface IP to operate HBM cell die integrated onto XPU
- TSV is not required on the XPU
- XPU communicates with the outside through the TSVs on DRAM
- Capacity and bandwidth can be expanded by stacking the DRAMs

	1 Stack	2 Stack		3 Stack	
	4 Channels	4 Channels	8 Channels	4 Channels	12 Channels
Capacity	3GB	6GB	6GB	9GB	9GB
Bandwidth	256GB/s	256GB/s	512GB/s	256GB/s	768GB/s

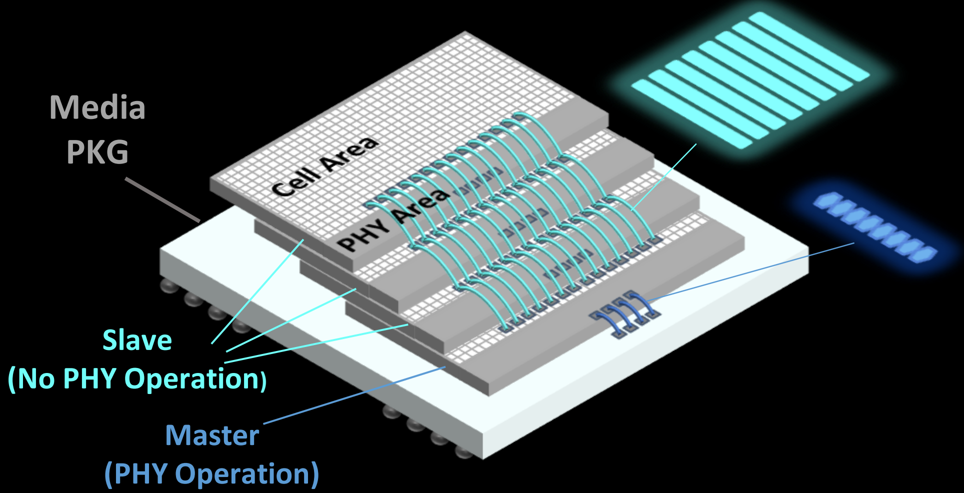
PI/SI ?
Reserve 4th TSVs for Top Die ?



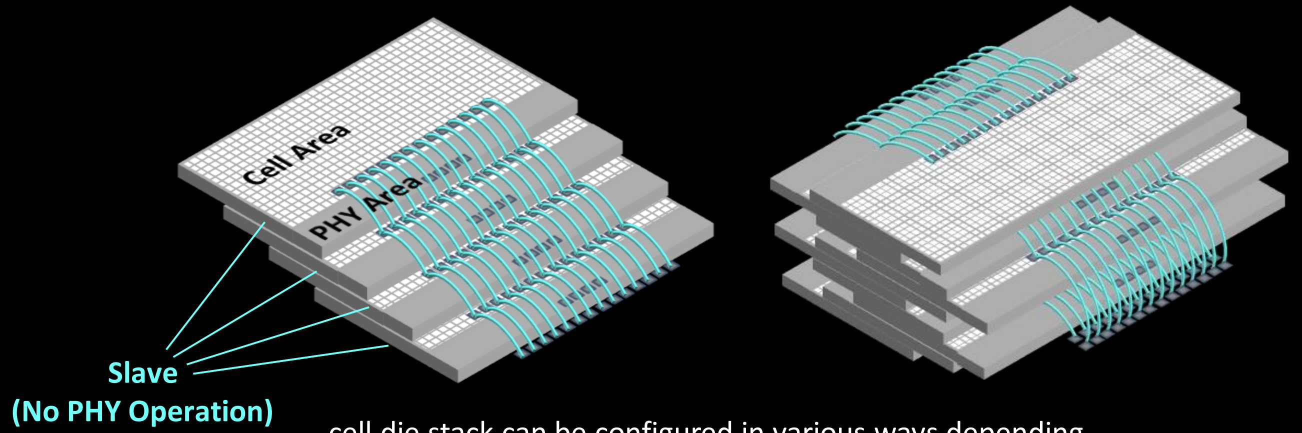
Appendix : PHY-less High Capacity Cell Die Stack

CONFIDENTIAL

<Cost-effective 3DS DDR5※ > ※ under developing



<Various Cell Die Stacks>

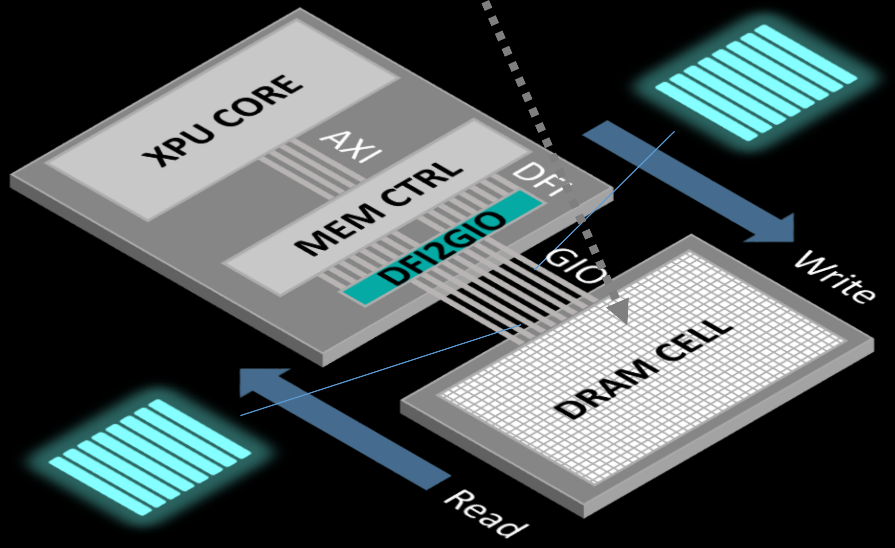


cell die stack can be configured in various ways depending on the desired capacity and bandwidth

- SK is developing cost-effective 3DS media using wire-bonding; the media has both of cell area and PHY area

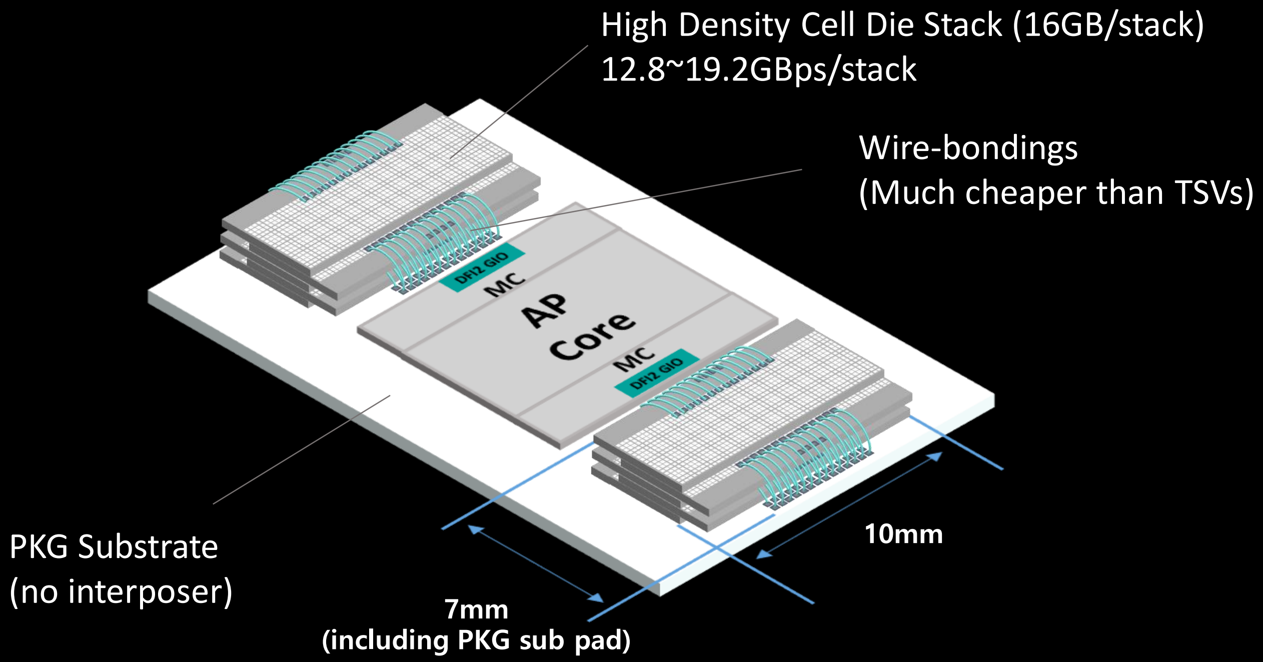
- The media can be used in our solutions; although the media has PHY, PHY-less operation is possible through cell die stack

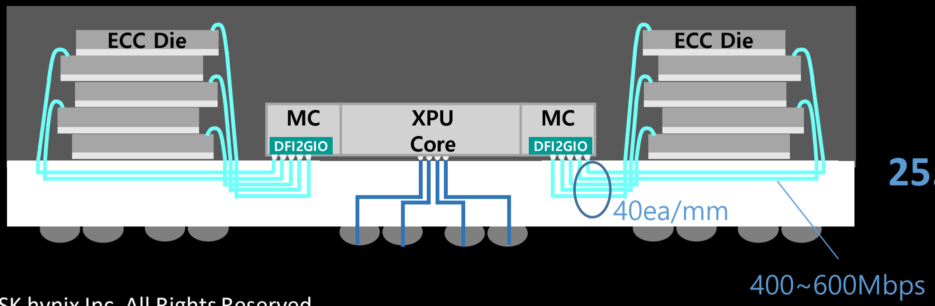
- The cell die stack is no more DDR; new concept of media w/ DDR capacity & 1/3 Lower power than LPDDR



Custom Memory Solution 3. PHY-less Post LPDDR

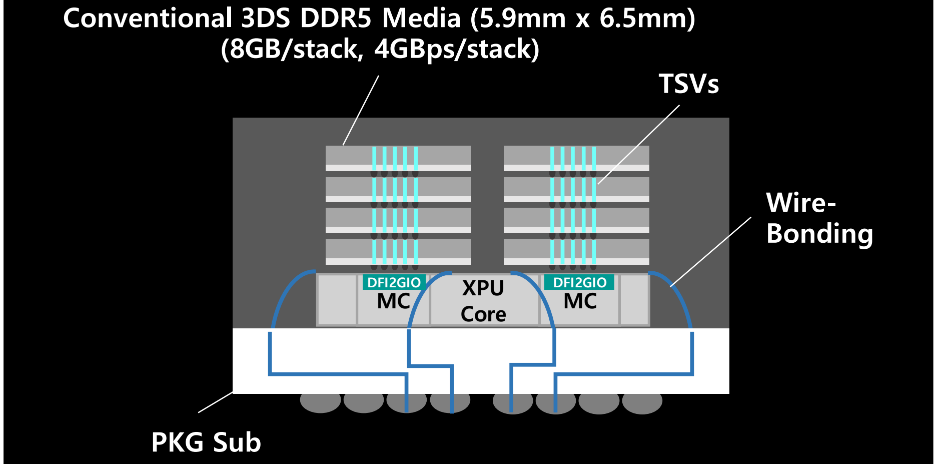
CONFIDENTIAL





32GB Density,
25.6~38.4GB/s Bandwidth,

<Alternative>



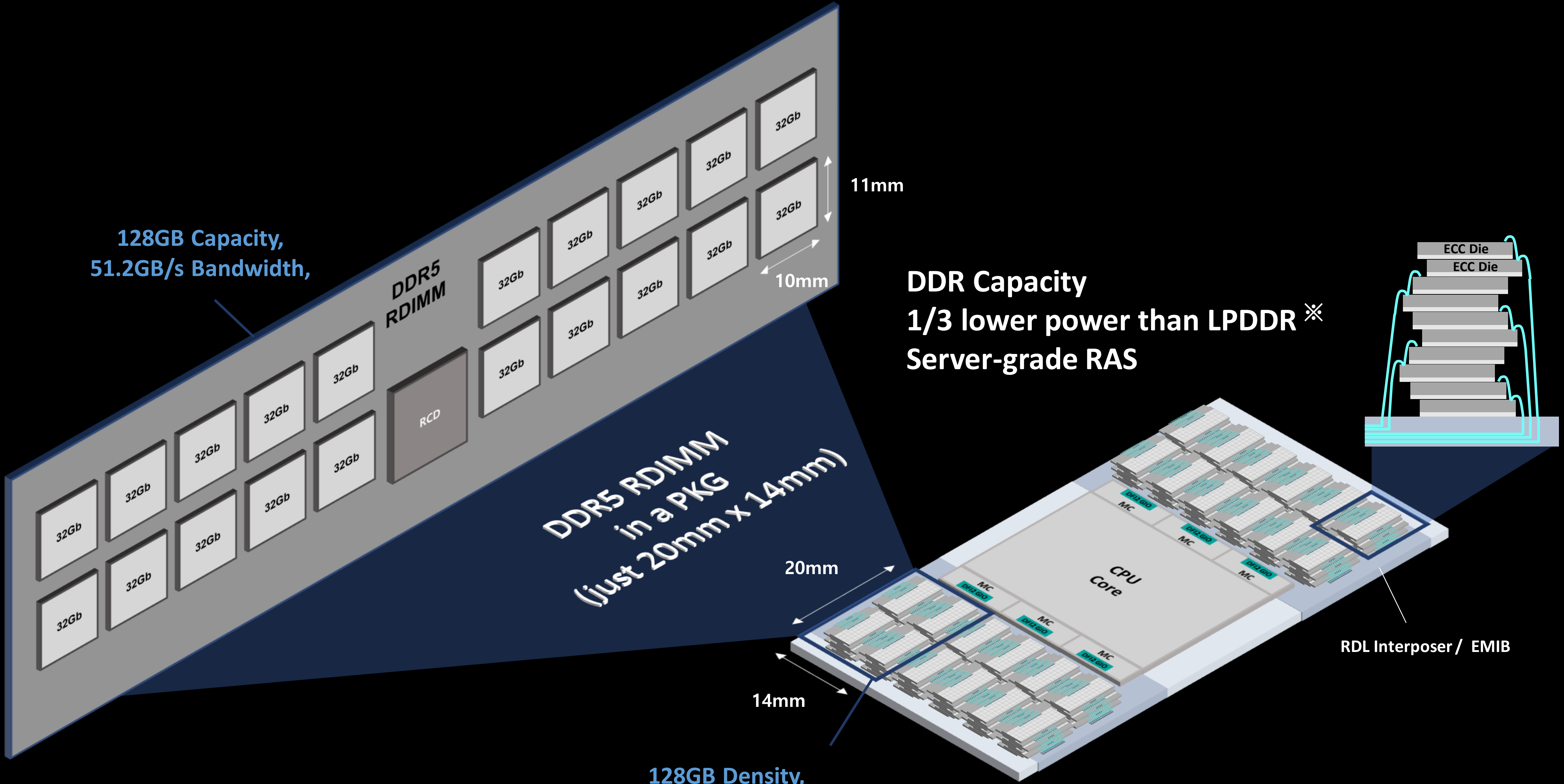
32GB Density (Stack x 4),
16GB/s Bandwidth,

- Conventional 3DS DDR5 media (TSV supported) can be used for small form-factor
- Bandwidth is lowered due to shared IO between the stack
- XPU heat & TSV cost must be considered

① Shoreline density
② Formfactor size

Custom Memory Solution 4. PHY-less Post DDR

CONFIDENTIAL



128GB Capacity,
51.2GB/s Bandwidth,

DDR5 RDIMM
in a PKG
(just 20mm x 14mm)

DDR Capacity
1/3 lower power than LPDDR ※
Server-grade RAS

128GB Density,
51.2GB/s Bandwidth

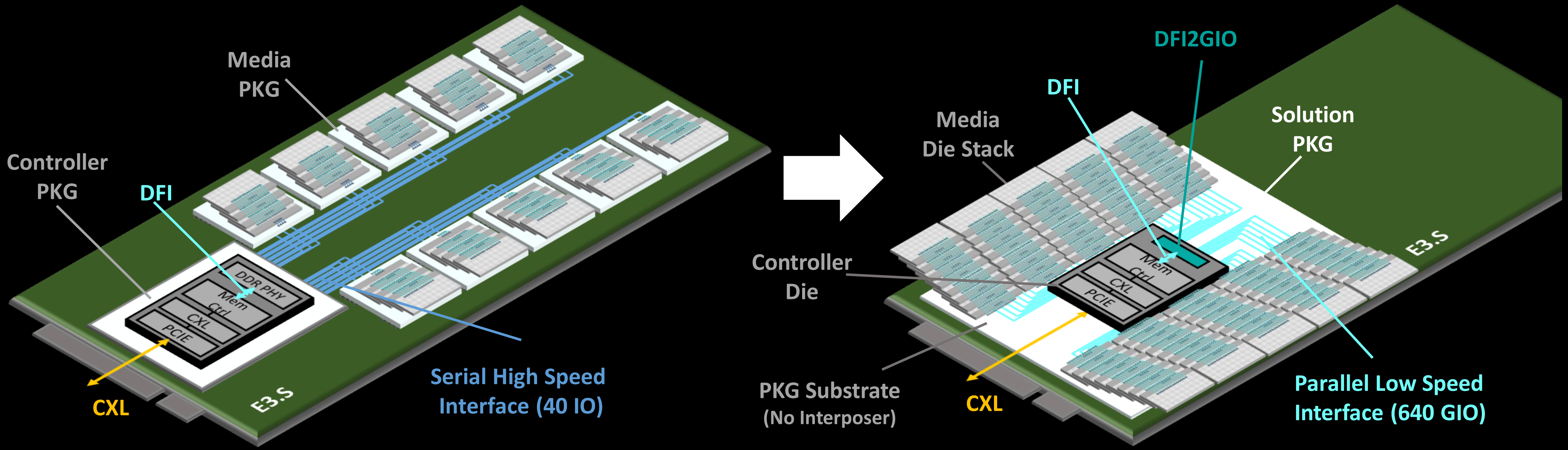
※ Power related to memory (Memory Controller + Media)

Custom Memory Solution 5. PHY-less CXL Device

CONFIDENTIAL

<SK hynix CXL Device >

<Next Gen. CXL Device>

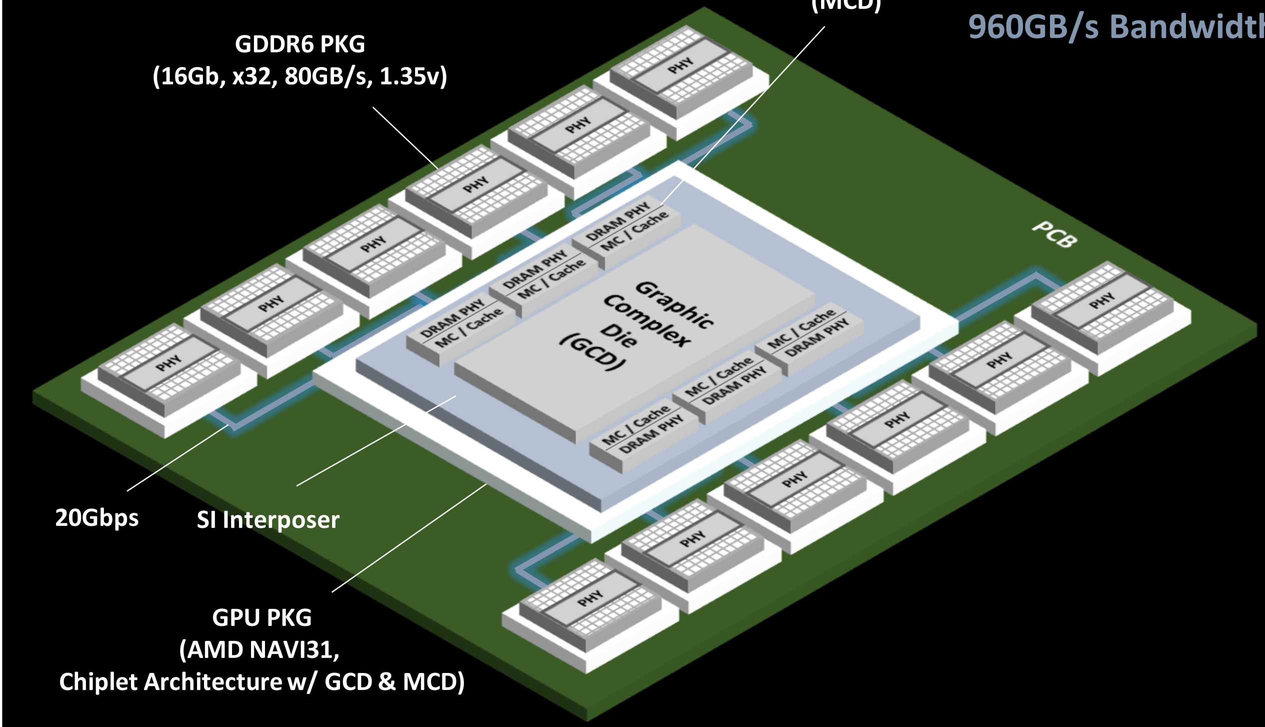


- The PHY-less structure using the cell die stack can also be applied to a CXL device
- Long latency in conventional CXL device can be reduced by 20ns
- No high speed module-level validation is required

Custom Memory Solution 6. LPHBM for Gaming GPU

CONFIDENTIAL

<GPU system w/ GDDR6>



Memory Cache Die (MCD)

24GB Density
960GB/s Bandwidth

GDDR6 PKG
(16Gb, x32, 80GB/s, 1.35v)

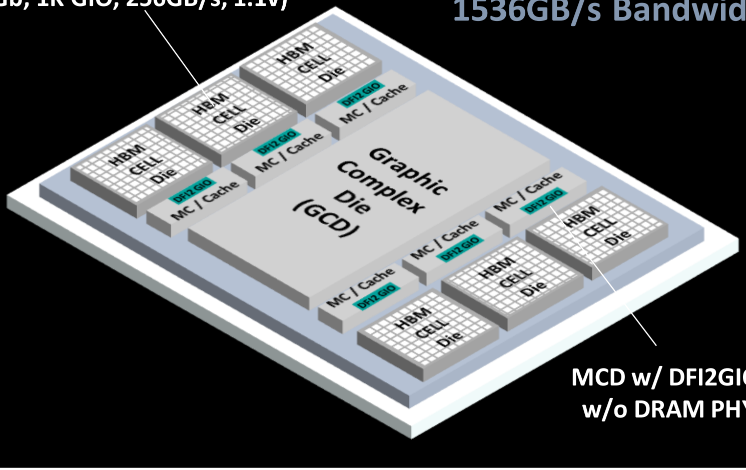
20Gbps

SI Interposer

GPU PKG
(AMD NAVI31,
Chiplet Architecture w/ GCD & MCD)

55% Power Reduction ※
15ns Latency Reduction
25% Lower Capacity

<GPU system w/ LPHBM>



HBM3E Core Die
(24Gb, 1K GIO, 256GB/s, 1.1v)

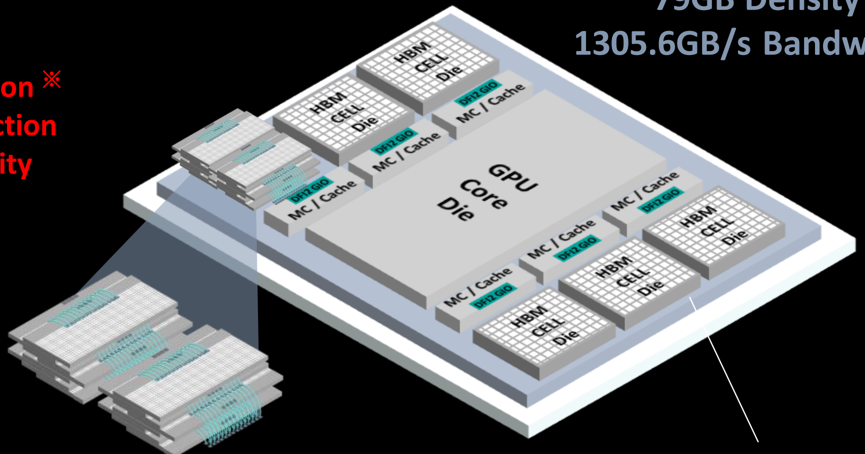
18GB Density
1536GB/s Bandwidth

MCD w/ DFI2GIO,
w/o DRAM PHY

55% Power Reduction ※
15ns Latency Reduction
230% Higher Capacity

<GPU system w/ 2-Tier Memory>

79GB Density
1305.6GB/s Bandwidth



Tier 2 Memory :
Cell Die Stack x2
(512Gb, 512 GIO, 25.6GB/s 1.1v)

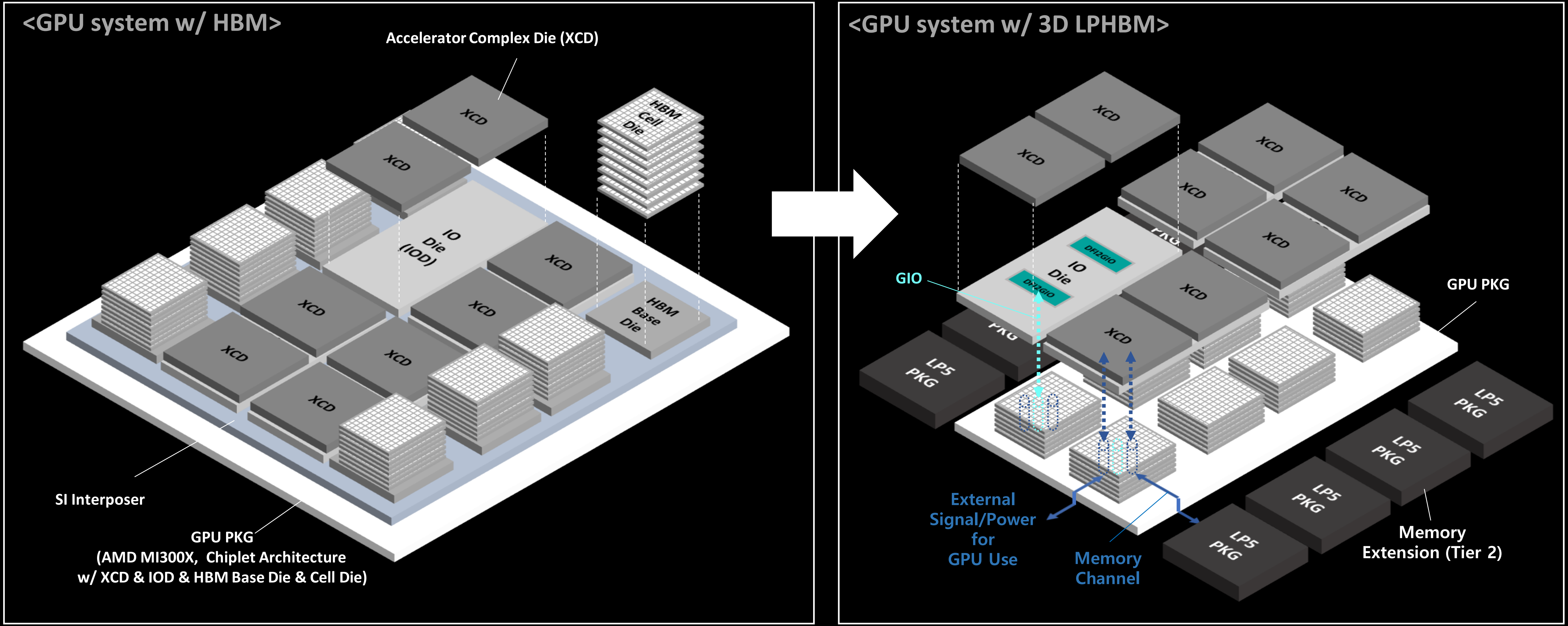
Tier 1 Memory :
HBM3E Core Die
(24Gb, 1K GIO, 256GB/s, 1.1v)

- LPHBM can be applied to GPU system instead of GDDR6
- Applying LPHBM on chiplet based GPU is relatively easy by modifying MCD only & reusing GCD
- 2-tier memory system replacing some media with cell die stack can expand the capacity largely

※ 55% reduction is memory related power (Memory Controller + Media)
※ 55% reduction by PHY-less architecture + lower voltage

Custom Memory Solution 7. 3D LPHBM for AI Server GPU

CONFIDENTIAL


<GPU system w/ HBM>
Accelerator Complex Die (XCD)
XCD
HBM Cell Die
IO Die (IOD)
HBM Base Die
SI Interposer
GPU PKG
(AMD MI300X, Chiplet Architecture
w/ XCD & IOD & HBM Base Die & Cell Die)
<GPU system w/ 3D LPHBM>
XCD
IO Die
DFI2GIO
GIO
GPU PKG
PKG
LP5 PKG
External Signal/Power for GPU Use
Memory Channel
Memory Extension (Tier 2)

- 3D LPHBM can be applied to AI server GPU system instead of HBM
- PHY-less 3D structure without base die and silicon interposer can significantly reduce power, latency, and cost
- Capacity can be expanded by further deploying LPDDR PKG by taking advantage of the smaller GPU PKG size