Expanding the Boundaries of the AI Revolution:

An In-depth Study of High Bandwidth Memory

Nayoung Lee & Sung Lee | March 2018

Table of Contents

1 THE MEMORY CHALLENGES of DEEP LEARNING

2 WHY HBM?

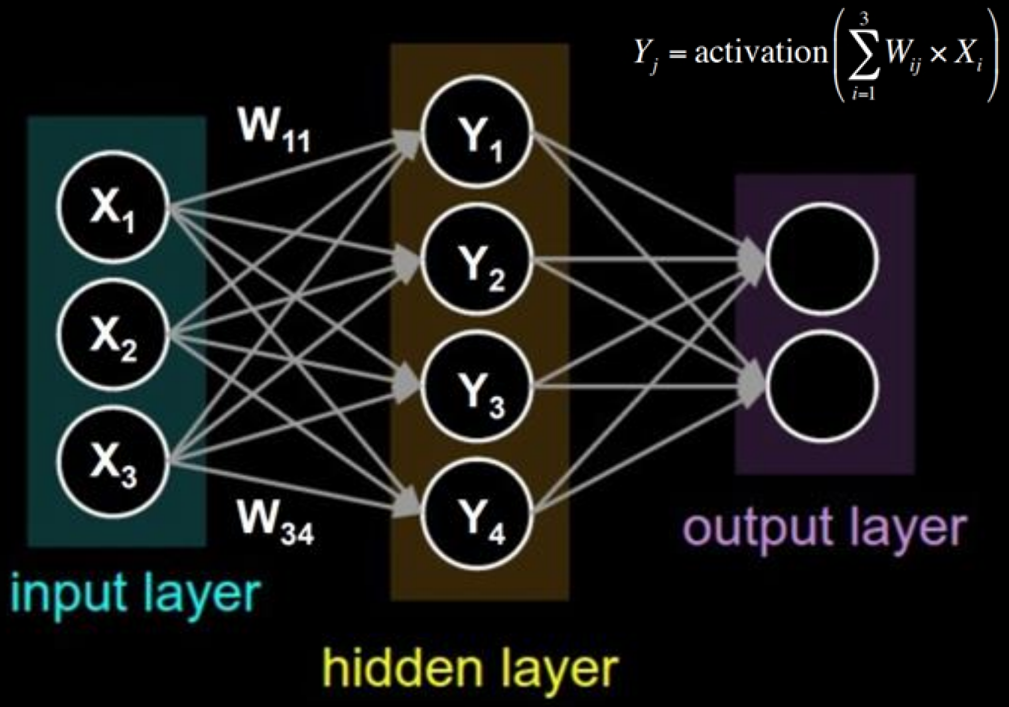
3 HIGH BANDWIDTH MEMORY DEEP-DIVE

1

THE MEMORY CHALLENGES of DEEP LEARNING

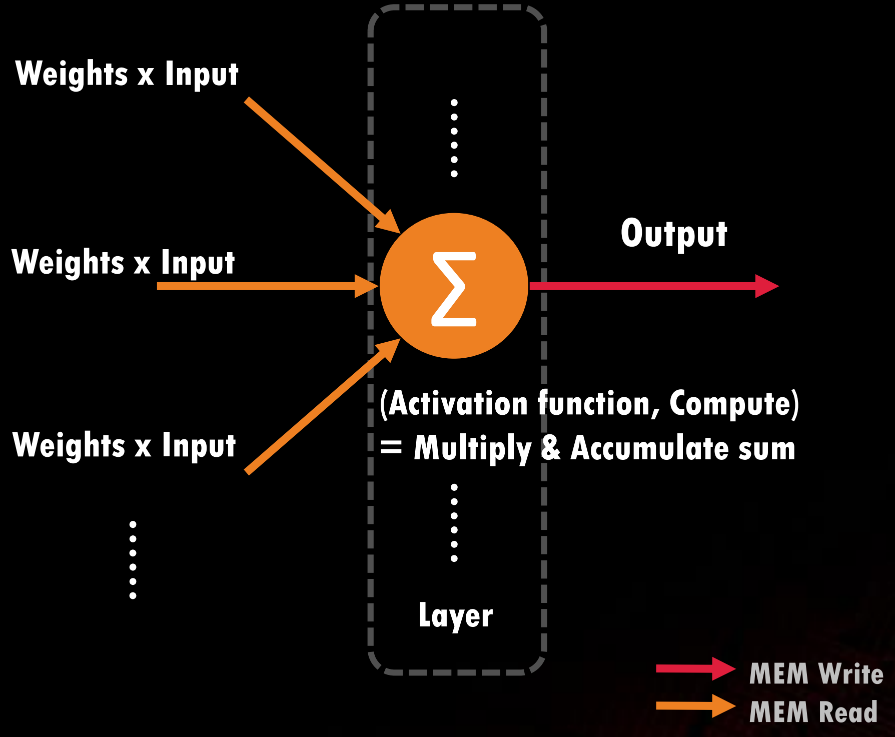
Deep Neural Network Fundamental Concepts

Deep Neural Network



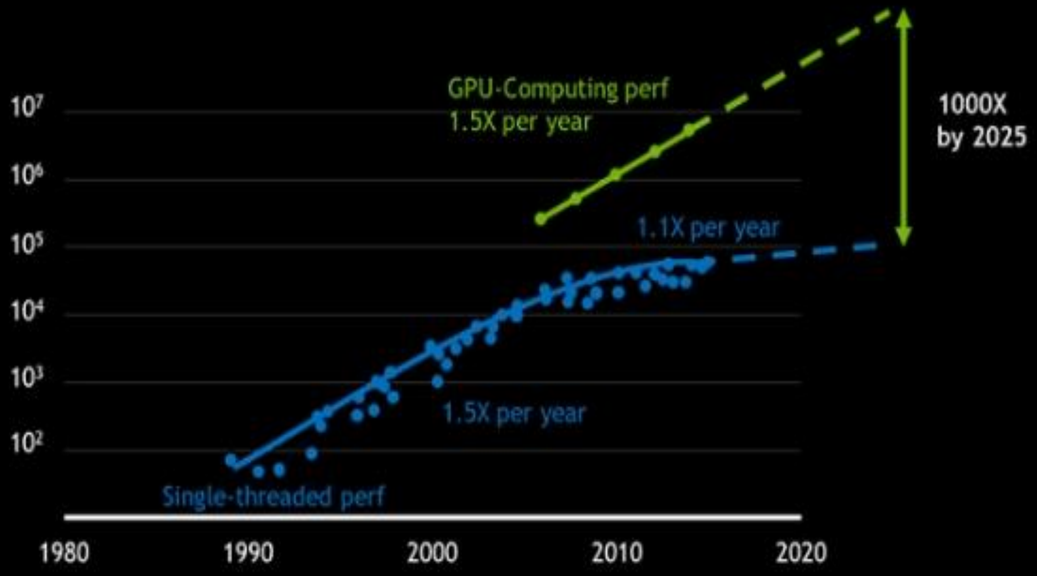
Source: Standford

Simple View

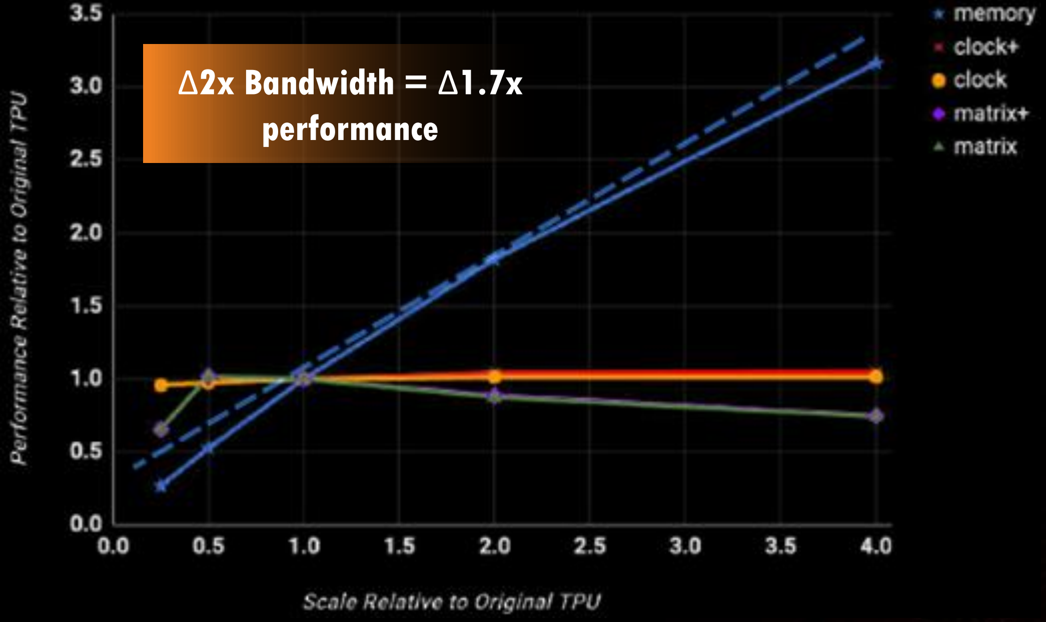


The Need for High Bandwidth Memory

GPU Computing

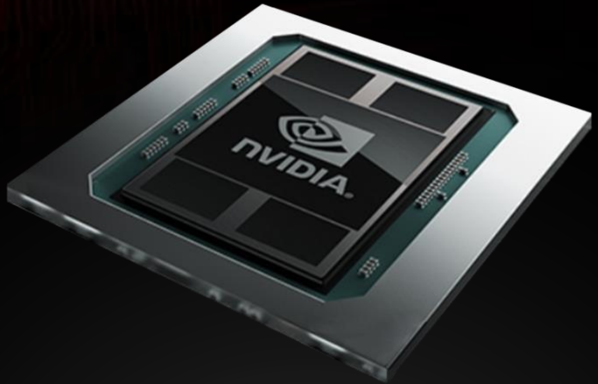


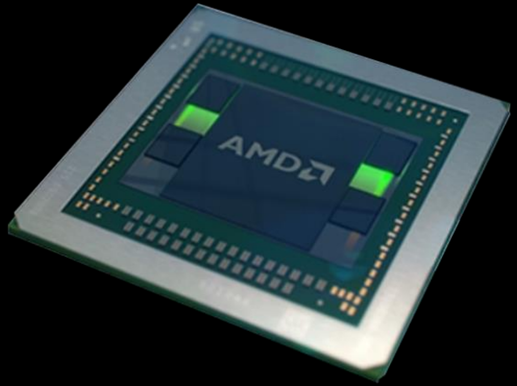
Performance bottleneck

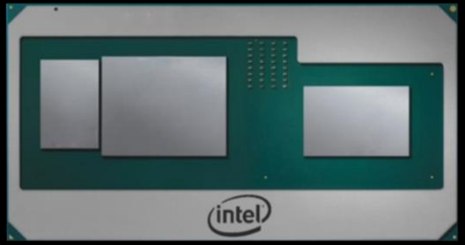

Δ2x Bandwidth = Δ1.7x performance

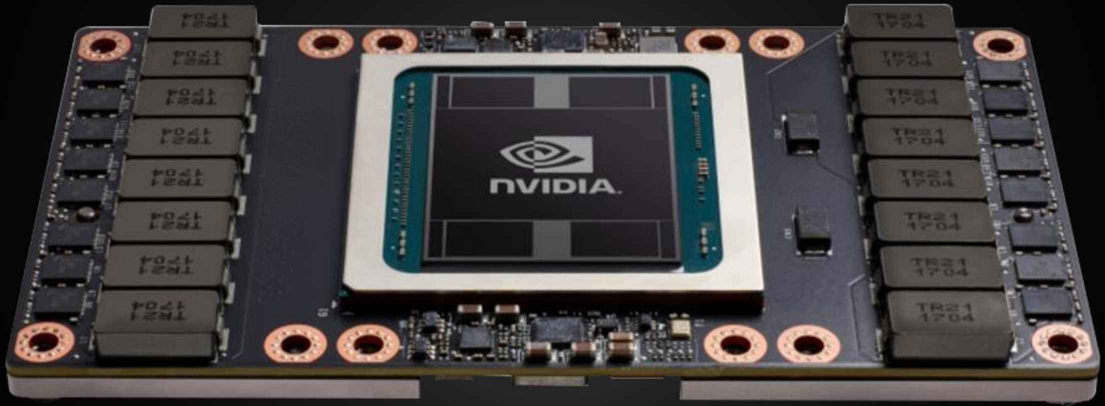
1) In-Datacenter Performance Analysis of a Tensor Processing Unit, Norm P. Jouppi et. al, (Google)

DRAM
XCVR
Intel® HyperFlex™
FPGA Architecture
XCVR
XCVR
XCVR
DRAM

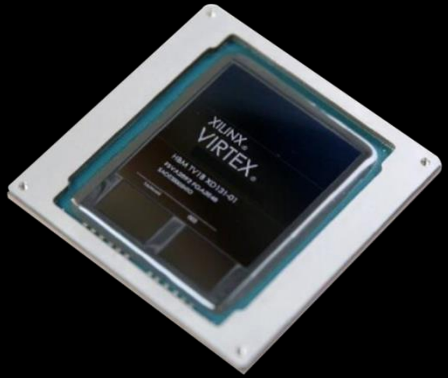

NVIDIA

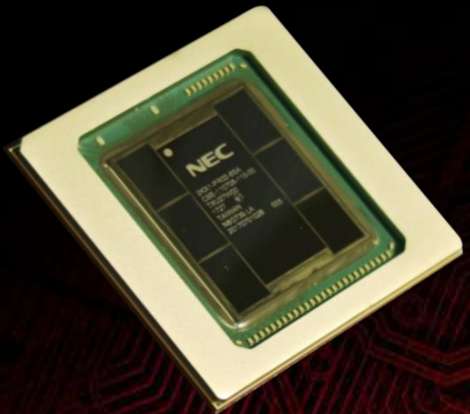

AMD


intel


NVIDIA

RADEON
VEGA


XILINX
VIRTEX


NEC

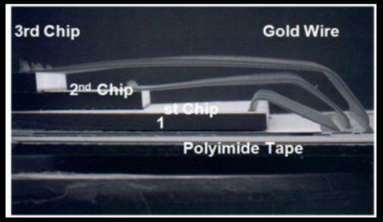
2

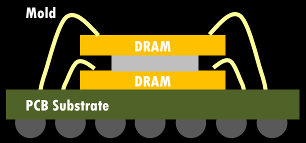
WHY HBM?

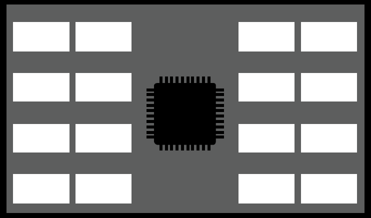
HBM, What's the difference?

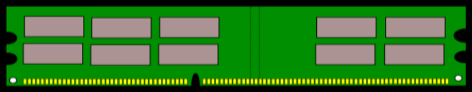
GDDR/DDR/LPDDR

➢ FBGA





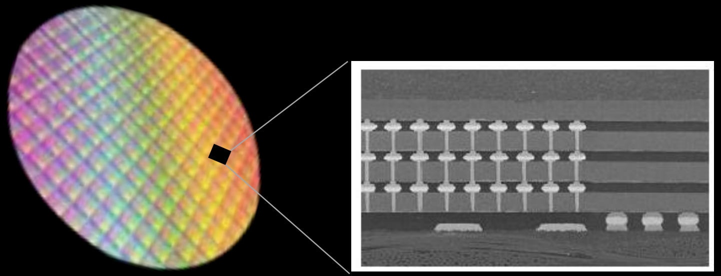




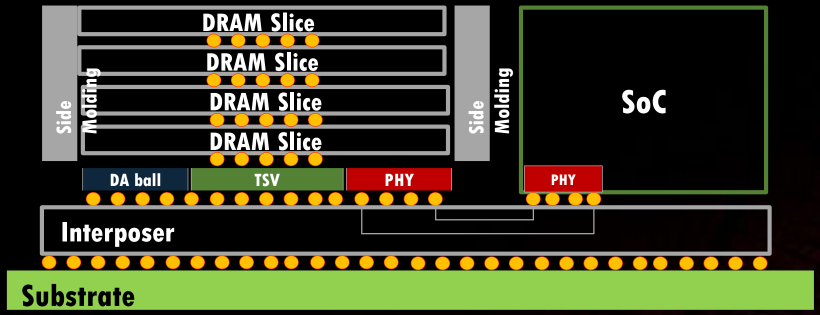
Soldered on PCB directly
Or
Use as DIMM Type

HBM

➢ KGSD



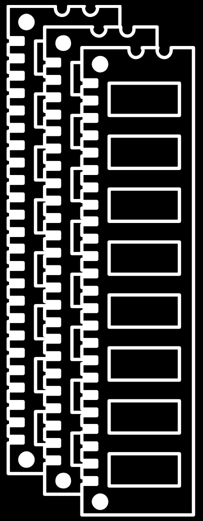
➢ HBM in 2.5D SiP



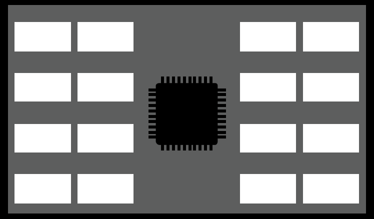
High Bandwidth Memory Delivers Small Form Factor

HBM provides highest bandwidth compare to other DRAM memories per unit area

To Achieve 1TB Bandwidth …



40ea of
DDR4-3200 Module



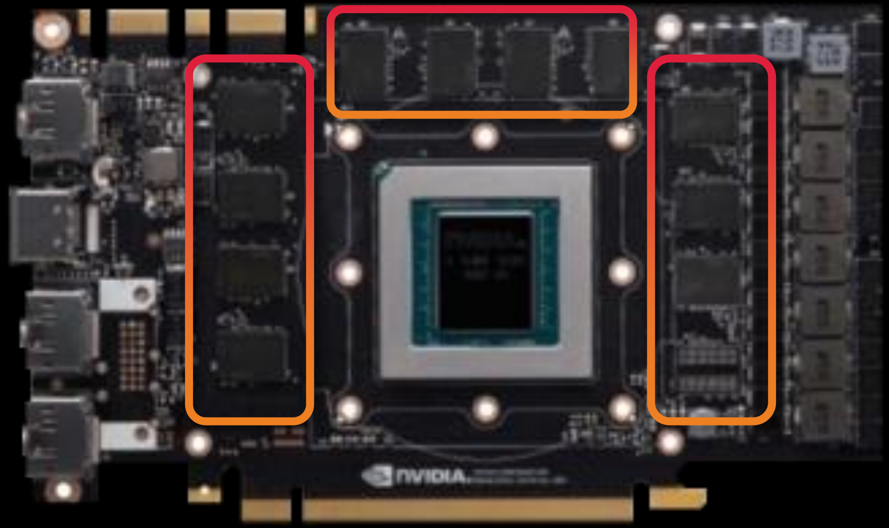
160ea of
DDR4-3200

Note: Advil is
a registered trademark

4ea HBM2 in
a single 50mm x 50mm Sip

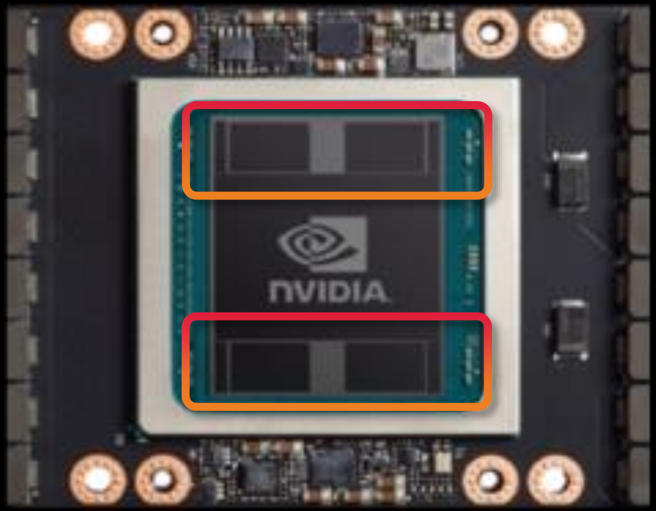
High Bandwidth Memory Delivers Small Form Factor

GDDR5(X)



Density	8Gb x 12 = 12GB
IO speed	8Gbps - 11Gbps
# of IO	384 bits
Bandwidth	384 – 528GB

HBM2



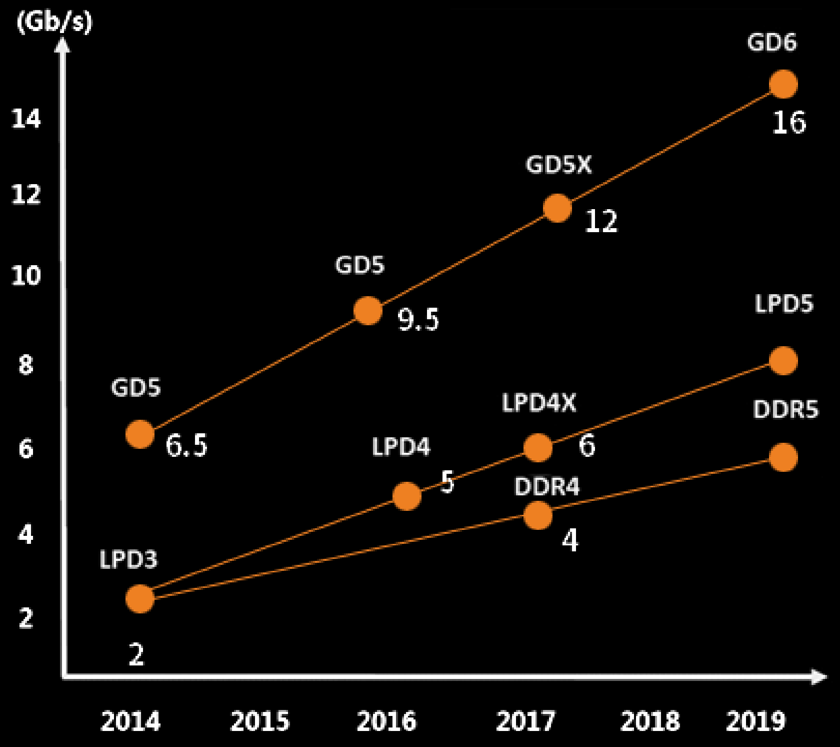
Density	8GB x 4 = 32GB
IO speed	2Gbps
# of IO	1024*4 = 4096
Bandwidth	1TB

SK hynix

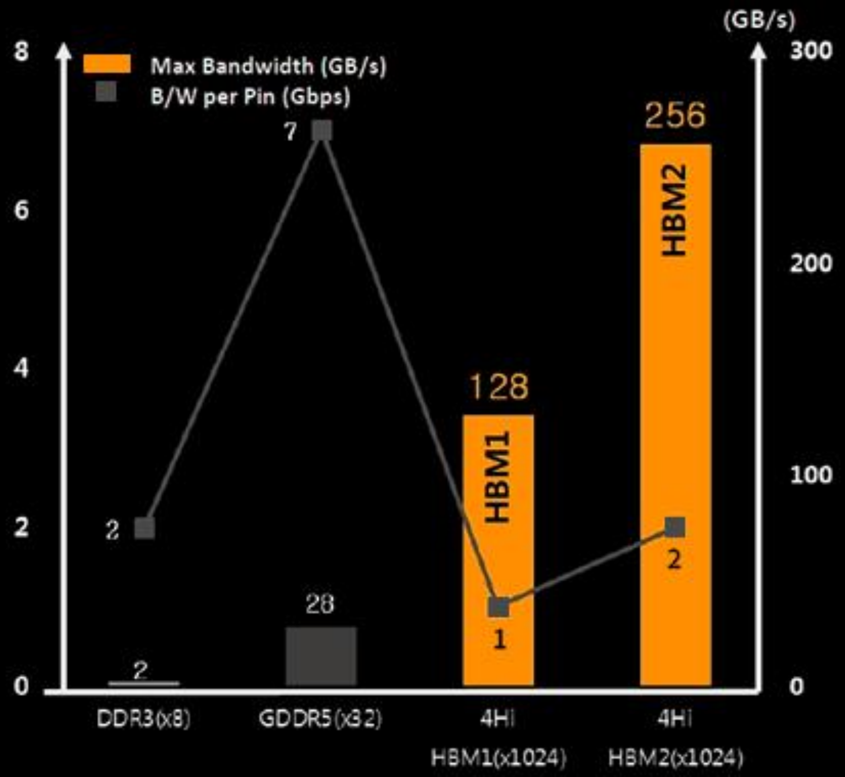
High Bandwidth Memory Delivers Unprecedented Bandwidth

HBM overcomes all DRAM bandwidth challenges

Bandwidth Challenges



High Bandwidth + High I/O

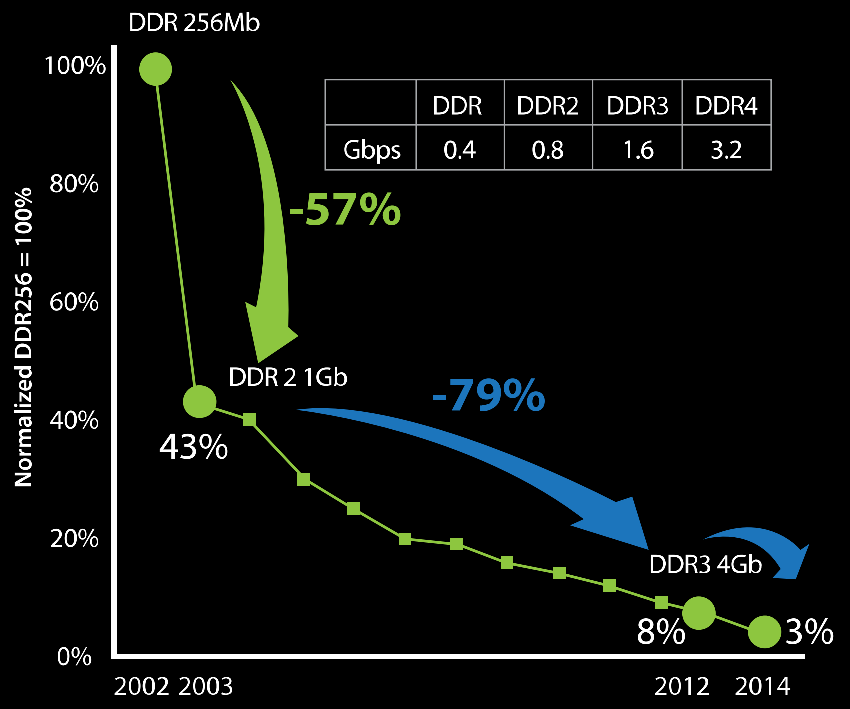


SK hynix

High Bandwidth Memory Delivers Power Efficiency

HBM low speed per pin & Cio reduces power consumption and increases power efficiency

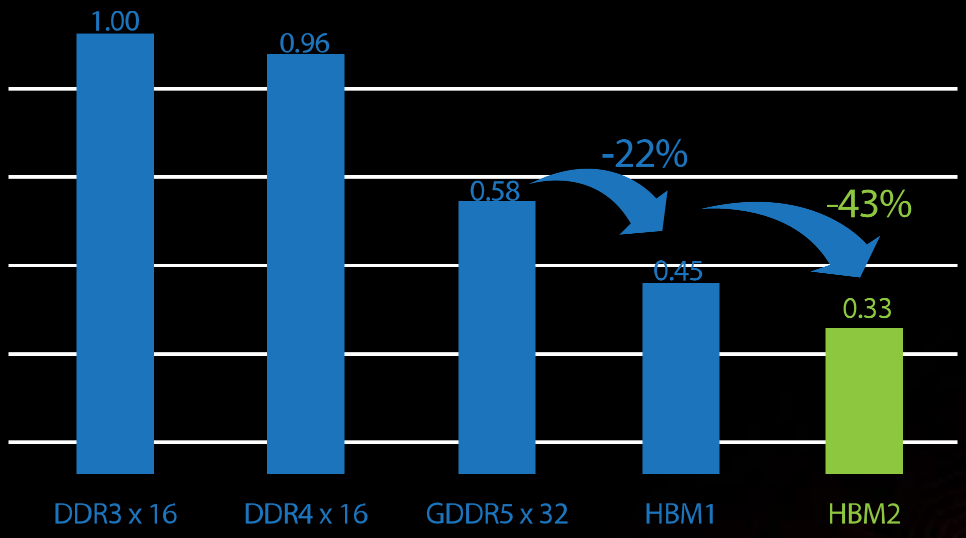
Power Efficiency



	DDR	DDR2	DDR3	DDR4
Gbps	0.4	0.8	1.6	3.2

Power Consumption

(mW/Gbps/pin)



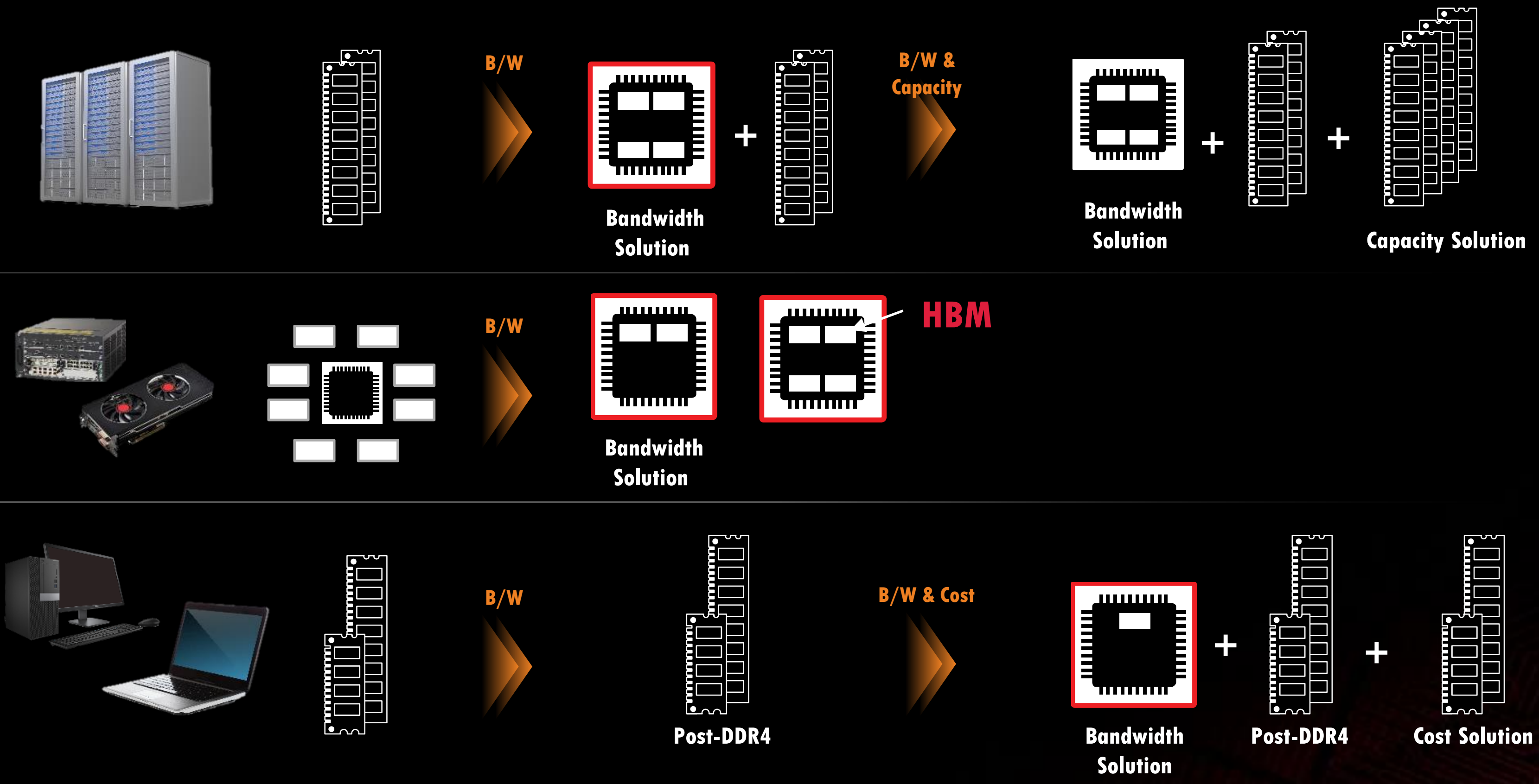
Next Generation System Architectures Leveraging HBM

HBM and 2.5D integration unlock new system architectures

HPC & Server
(B/W & Capacity)

Network & Graphics
(B/W)

Client-DT & NB
(B/W & Cost)



B/W

Bandwidth Solution +

B/W & Capacity

Bandwidth Solution + + Capacity Solution

B/W

Bandwidth Solution

HBM

B/W

Post-DDR4

B/W & Cost

Bandwidth Solution + Post-DDR4 + Cost Solution

3

HIGH BANDWIDTH MEMORY DEEP-DIVE

1) Innovative Design
2) Revolutionary Technological Features
3) Next Generation Line-up Considerations

Introduction

Did You Know?

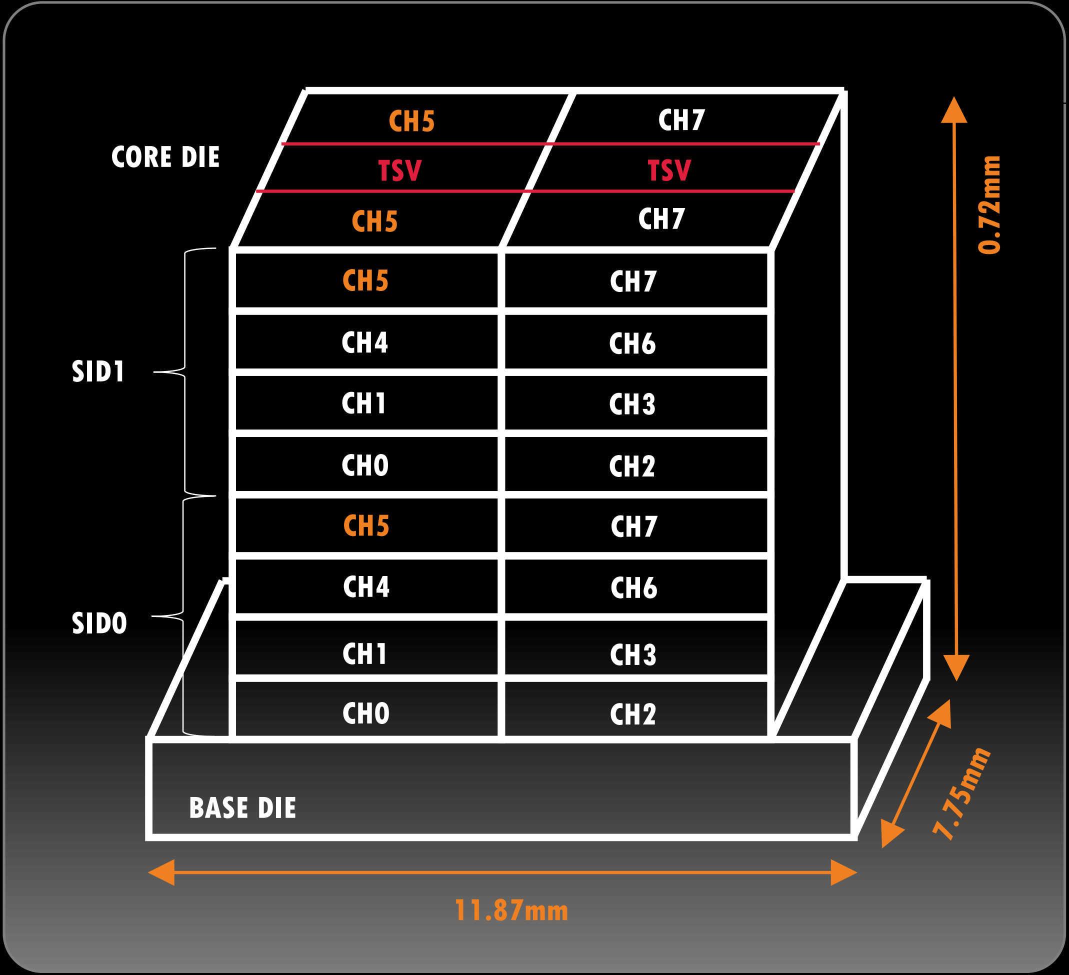
HBM standard adopted by the Joint Electron Device Engineering Council(JEDEC) in 2013, and the current 2nd generation HBM in 2016.

High bandwidth, high power efficiency and compact form factors have propelled HBM collaboration engagements covering all IT sectors.
e.g. Graphics, AI/Deep Learning, HPC, SVR, NTW Router/Switches etc.

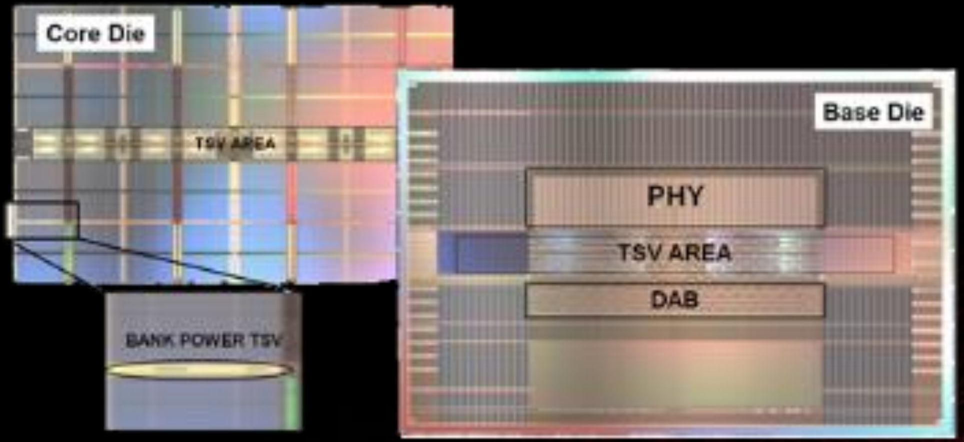
Total HBM (+HMC) market expected to increase from $922.7M in 2018 to $3,842.5M by 2023, resulting in CAGR 33%. (Source: RESEARCH AND MARKETS)

Innovative Design

HBM KGSD Architecture

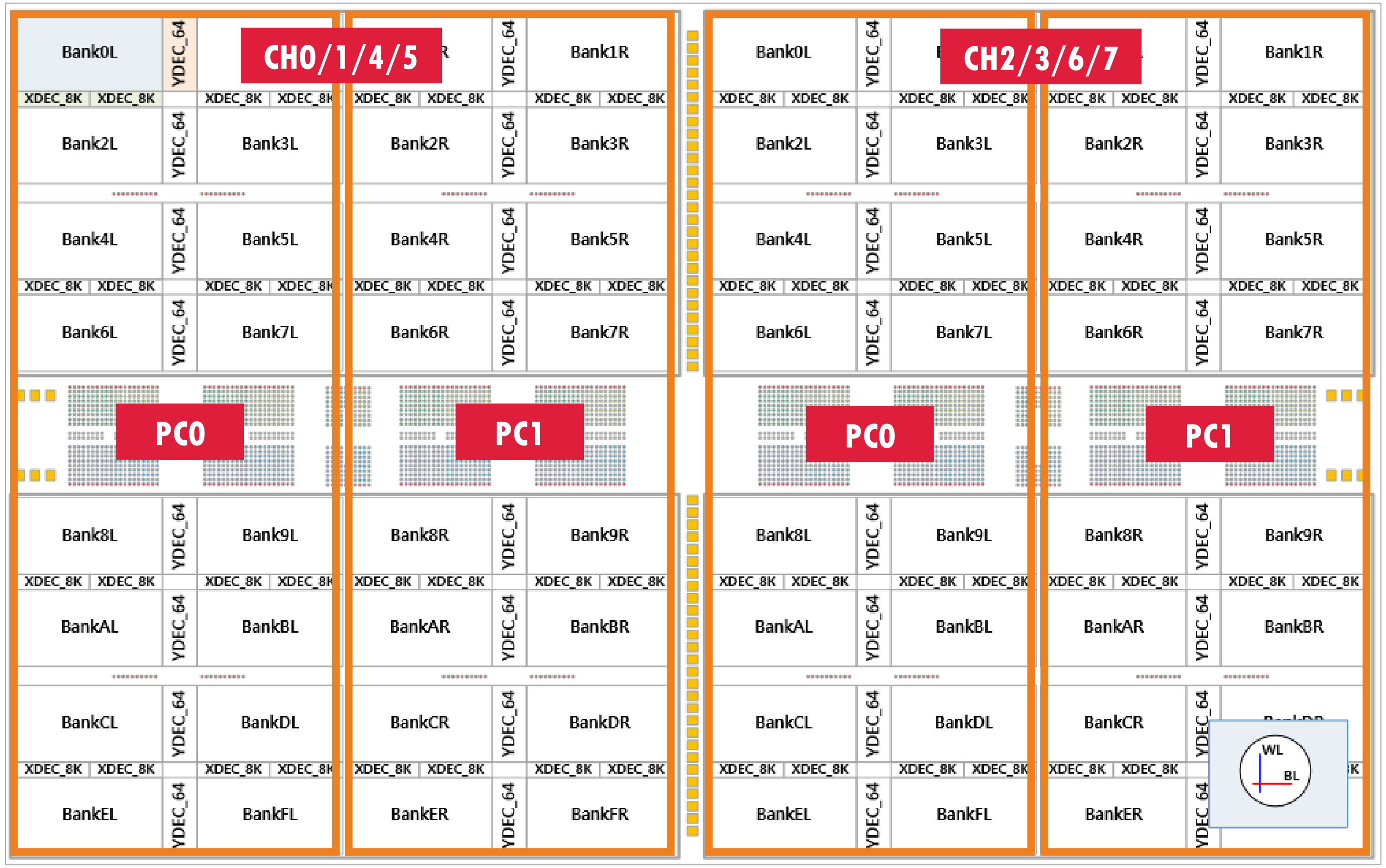


- 11.87x7.75x0.72mm PKG dimension
- 9Gb per cell array (Optional 1Gb ECC cell)
- 4/8GB density per mKGSD stack
- Max 2.4Gbps data transmission speed enabling 307GB/s B/W performance



Innovative Design

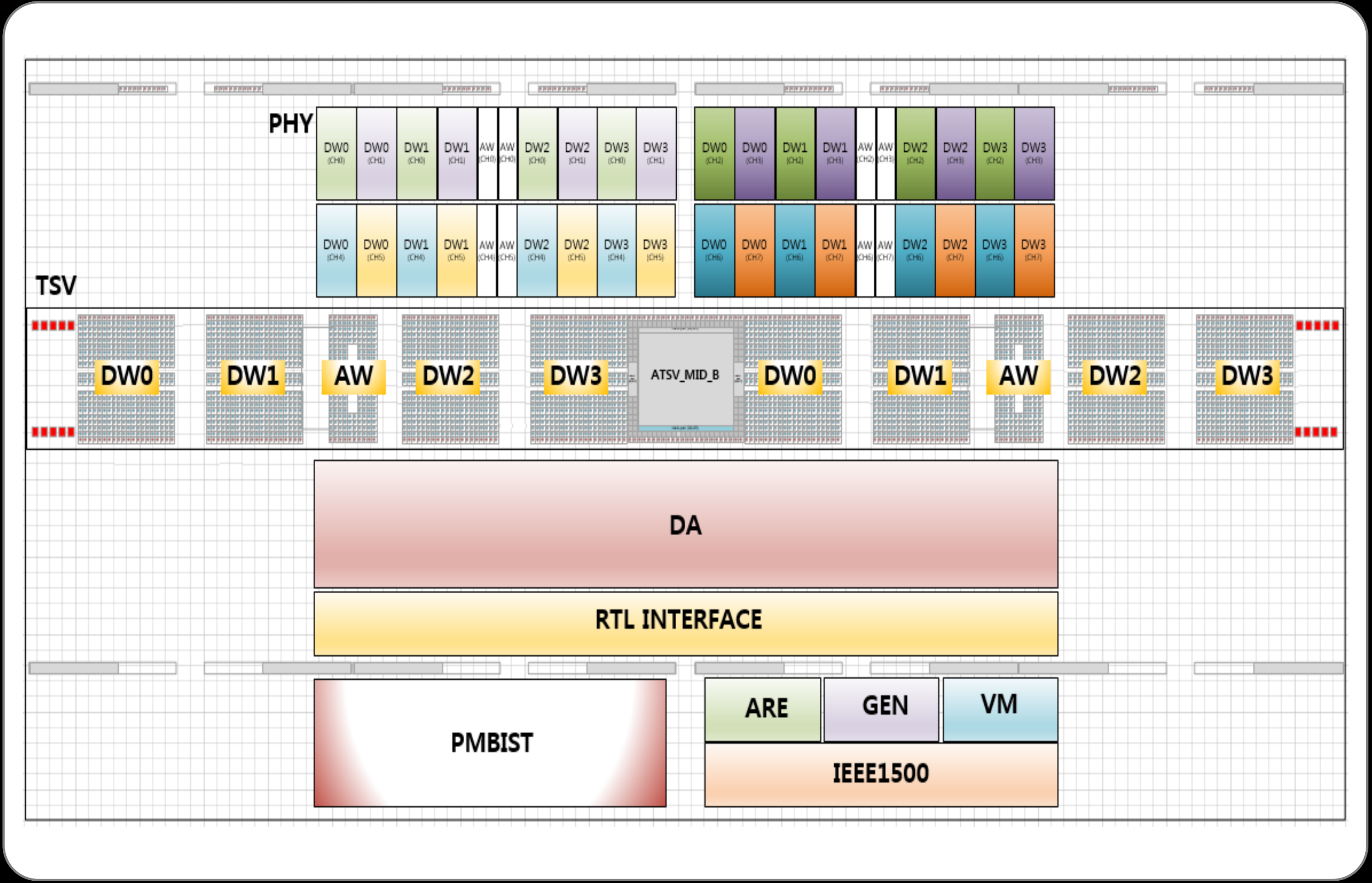
HBM Gen2 Core Die



- 10.63mm x 6.65mm
- Supports Pseudo CH mode
- 2 individual sub-CH of 64bits I/O, 16 banks
- Two seamless array access w/ Burst Length 4
- 256b Prefetch per PCH

Innovative Design

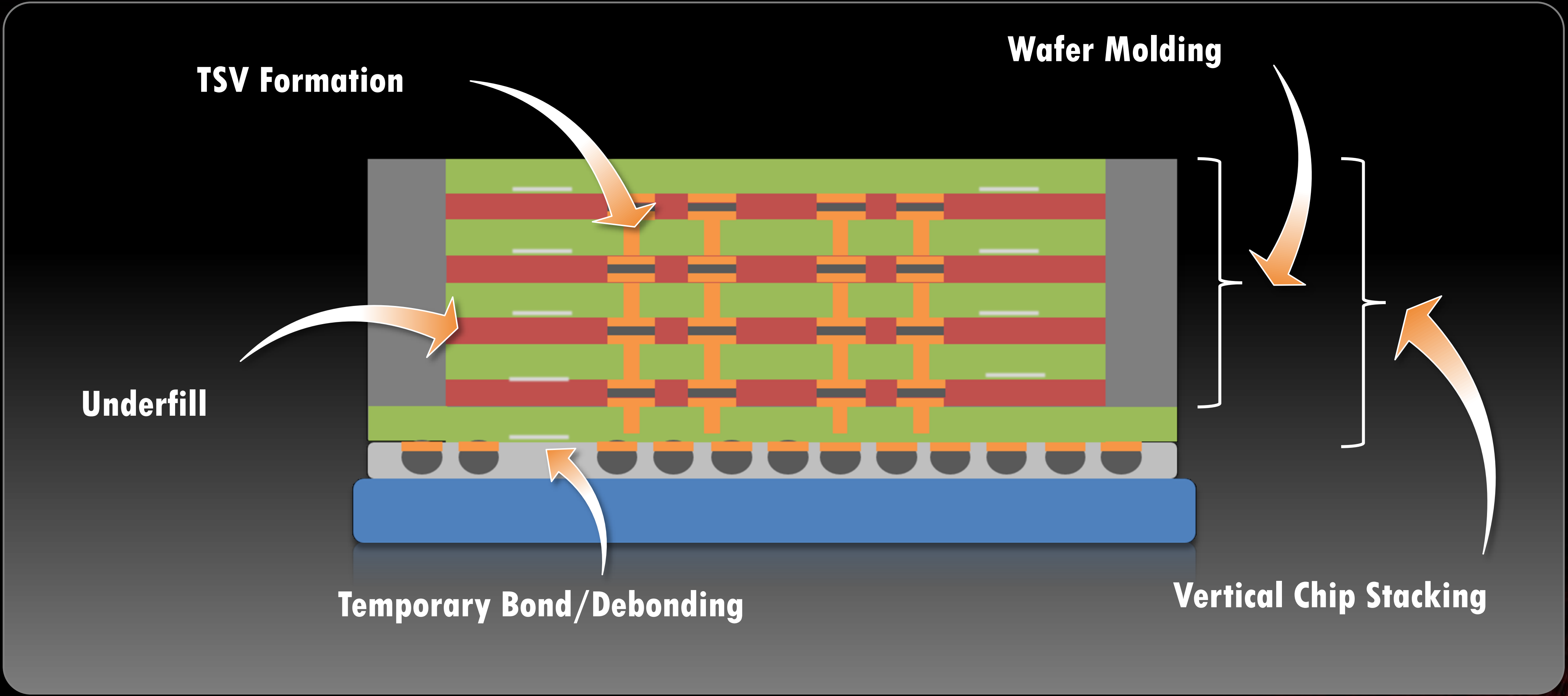
HBM Gen2 Base Die



- 11.87mm x 8.87mm
- Programmable Memory Built-In Self Test
- Direct Access
- IEEE1500
- PHY

Revolutionary Technical Features

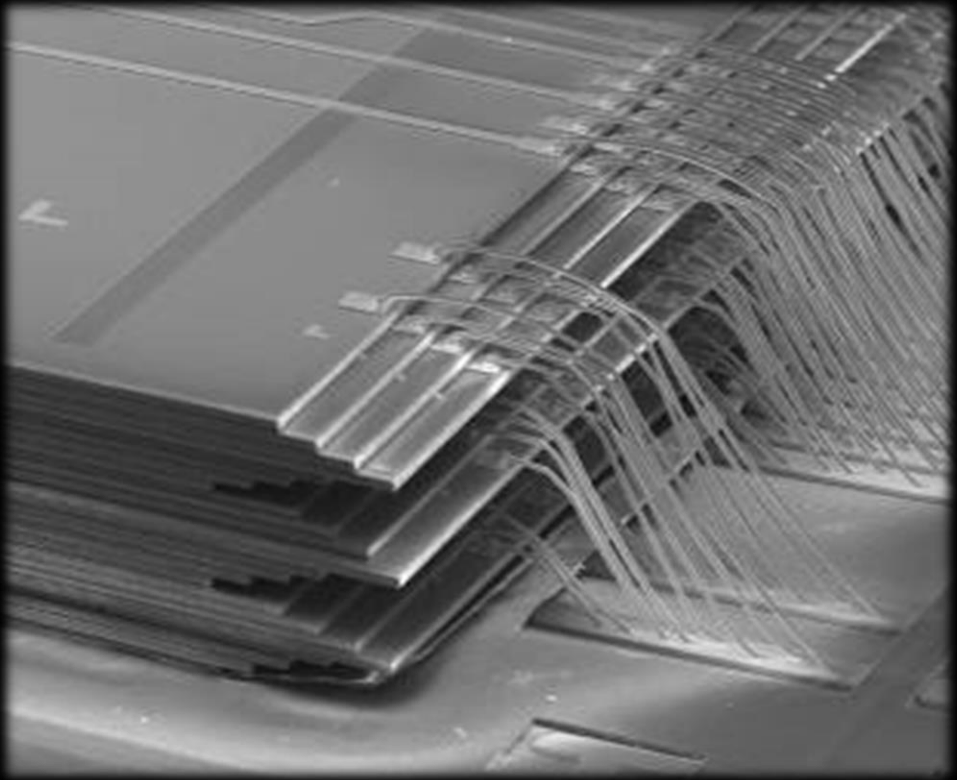
PKG Stacking & Interconnection



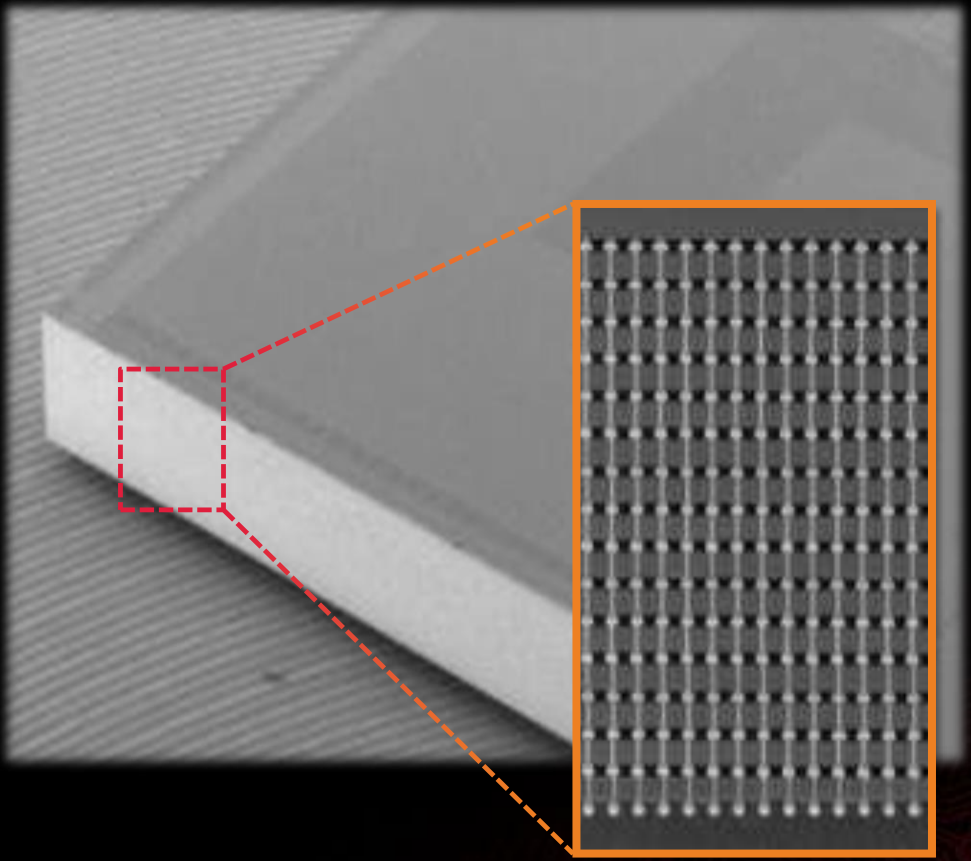
Revolutionary Technical Features

PKG Stacking & Interconnection

Wire Bonding



Through Silicon Via

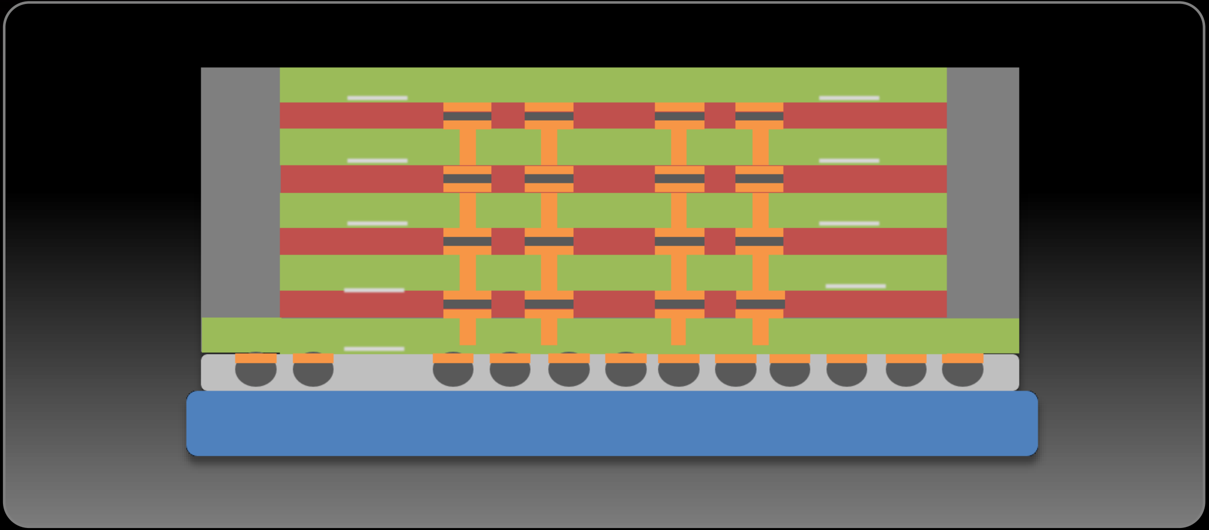


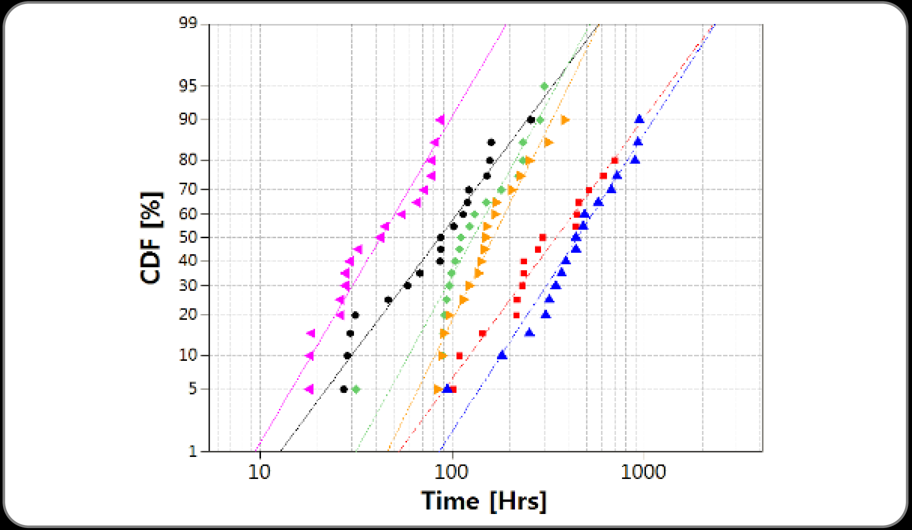
Revolutionary Technical Features

Wafer & KGSD PKG Level Reliability

Wafer-level Process Qualification	PKG-level Product Qualification
Time Dependent Dielectric Breakdown	EFR, HTOL, LTOL (Lifetime)
Hot Carrier Injection	TC, THB, HAST, uHAST, HTS w/ Preconditioning (Environmental)
Negative Bias Temp Instability	Electrostatic Discharge
Electro Migration	Latch-up
Stress Migration	Package Construction Analysis
TSV, uBump Electromigration	Electrical Characterization

Wafer & KGSD PKG Level Reliability



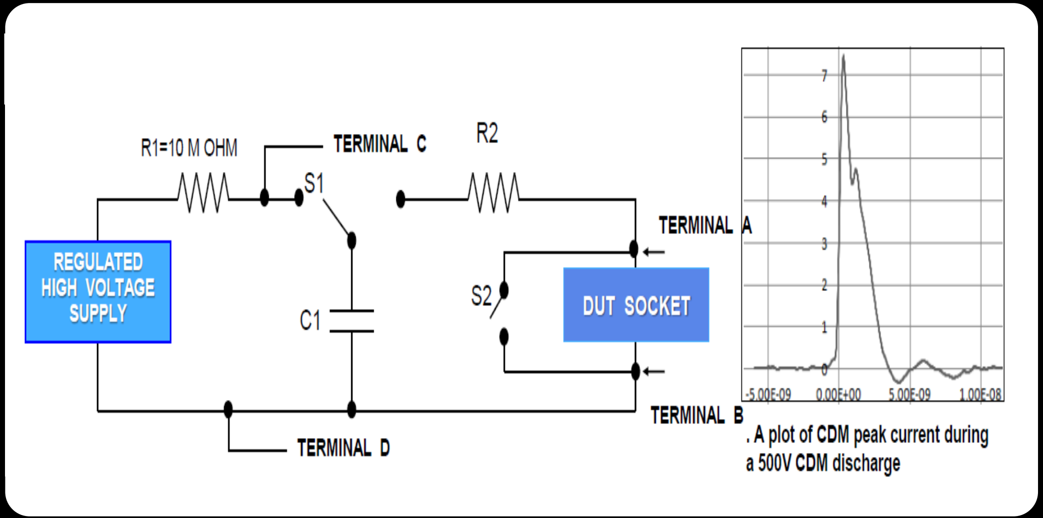


Type	Direction	T0.1% Lifetime	Criteria
Core Die	VDD	>> 10 years	• $\Delta R/R_0$ x 100> 20% • F(10yrs) < 0.1% @ use condition
	VSS		
Base Die	VDD		
	VSS		
TSV	VDD		
	VSS		

Revolutionary Technical Features

Wafer & KGSD PKG Level Reliability

Direct Access Bump



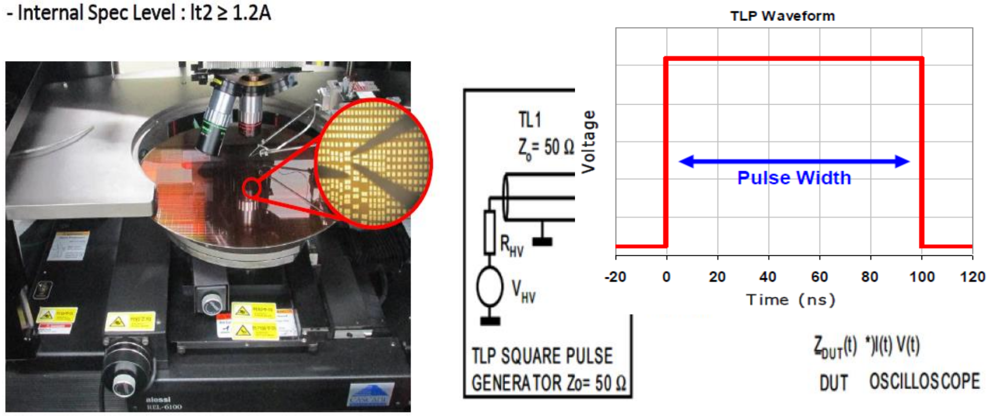
. A plot of CDM peak current during a 500V CDM discharge

Method	Target
Human Body Model	≥ 2,000V
Charged Device Model	≥ 500V

PHY Bump

Standardization of the Transmission Line Pulse(TLP) Methodology for Electrostatic Discharge (ESD)

- Internal Spec Level : It2 ≥ 1.2A



VF-TLP(CDM like) : 1.25ns

Method	Target
VF-TLP (CDM-like)	It2 ≥ ~ 1.xA

* Very Fast Transmission Line Pulse

Wafer & KGSD PKG Level Reliability

KGSD HBM Test Flow

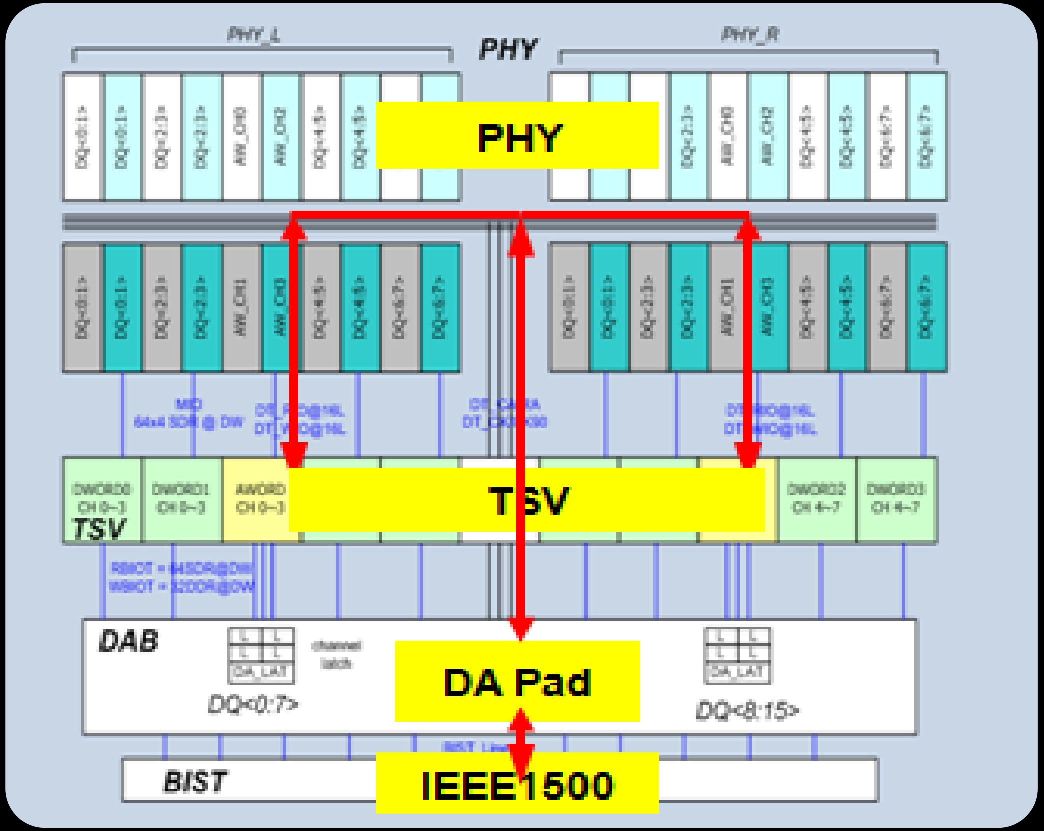
Core Die	Base Die
WFBI	Logic Test
Hot & Cold Test	
Repair	

KGSD
TSV Scan
Built-In Stress
Hot & Cold Test
Speed Test

Revolutionary Technical Features

Wafer & KGSD PKG Level Reliability

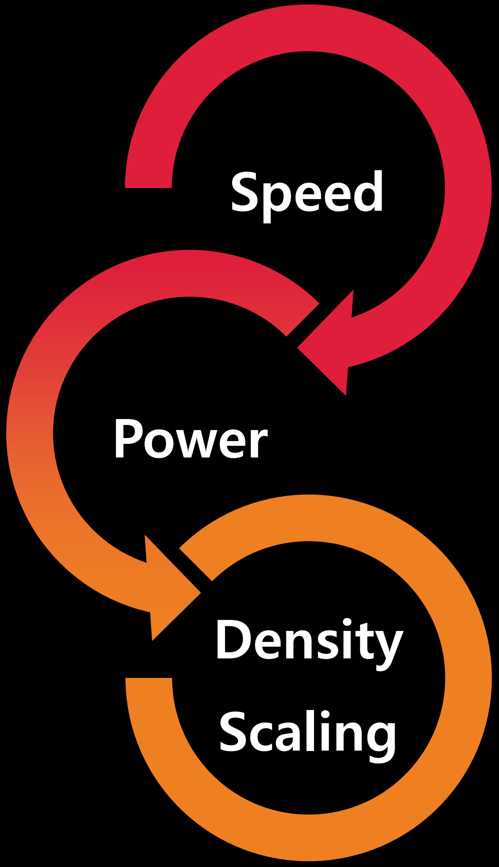
KGSD HBM Test Coverage



Area	Type	Comment
PHY	Function Test	RD/WT,CL,BL
	Margin Test	Speed, VDD, Setup/Hold Timing
TSV	Function Test	RD/WT,CL,BL,TSV interface
	OS Check	TSV Open/Short Check
Logic	Function Test	IEEE1500, Function, BIST, Repair
	Margin Test	VDD, Speed, Setup/Hold
Core	Function Test	RD/WT, Self Ref, Power Down
	Margin Test	Speed, VDD, Async, Refresh
	Repair	Cell Repair

Next Generation Line-up

Key Performance Considerations



- Transistor performance between DRAM process and Logic Process (2.8Gbps~3Gbps may be the realistic max speed on DRAM)
- TSV lines to be doubled to secure valid window

- Speed increasing makes worse power consumption
- All possible solution should be considered for power reduction

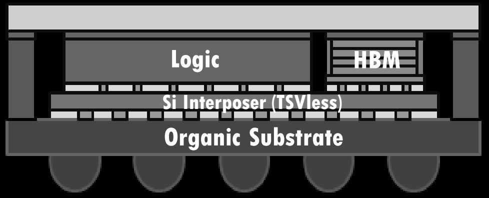
- Additional HBM cubes
- DRAM density and process are limited by SiP size
- Higher DRAM stack has to be considered to increase density

Next Generation Line-up

Key Performance Considerations

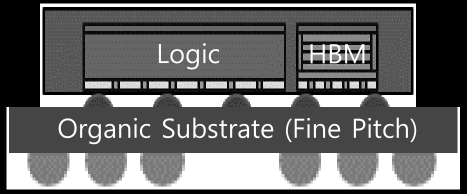
Cost Effective Solutions

TSVless Si-Interposer



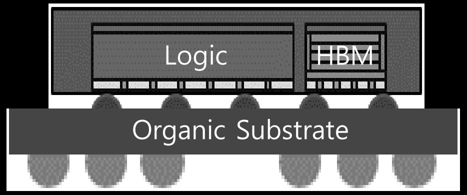
- Removing Si to expose BEoL layer (as RDL)

2.1D SiP



- Fine pitch organic substrate allows direct interconnection w/o interposer

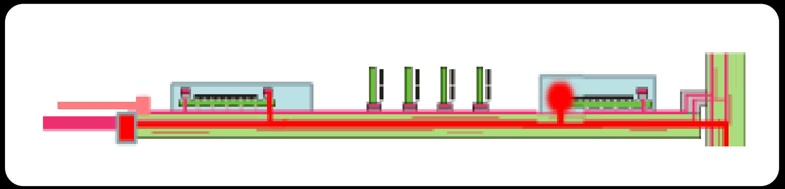
Fan Out SiP on Sub.



- Removing Si-interposer thanks to fine pitch RDL trace of Fan Out Package

High Speed Signal Transmission

Si Photonics in 2.5D SiP

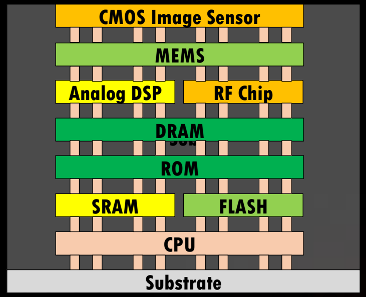


- Chip to chip optical signal transmission through embedded wave guide in Si-interposer

Source : CEA-Leti

Low Power and Small Form Factor

Hetero-generous 3D Stack



- More chips in a package with TSV stack

SK hynix

Thank you

Come visit us at **Booth #711** and learn more about SK hynix memory solutions

SK hynix