
# HBM: Memory Solution for Bandwidth-Hungry Processors

**Joonyoung Kim and Younsu Kim**

**SK hynix Inc.**

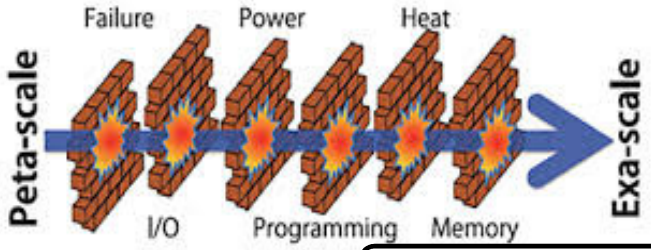
**August 2014**

# Memory requirement

## HBM: Memory Solution for Density & Bandwidth-Hungry Processors

High-End Graphics


40G/100G Ethernet




Various walls to

Exa-scale HPC

**< Exa-scale Roadmap >**

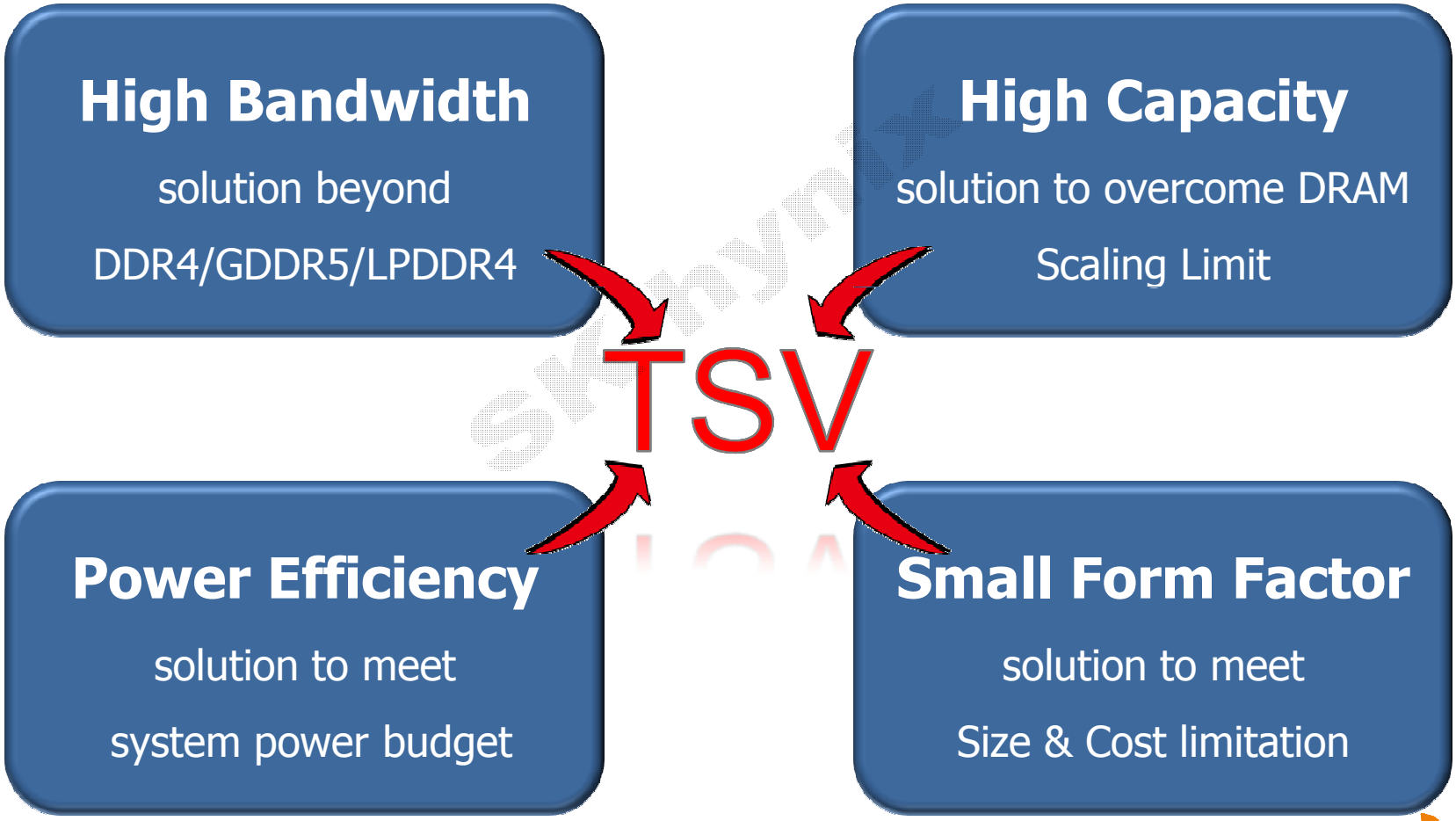
| Systems | 2009 | 2011 | 2015 | 2018 |
|---|---|---|---|---|
| System Peak Flops/s | 2 Peta | 20 Peta | 100-200 Peta | 1 Exa |
| System Memory | 0.3 PB | 1 PB | 5 PB | 10 PB |
| Node Performance | 125 GF | 200 GF | 400 GF | 1-10 TF |
| Node Memory BW | 25 GB/s | 40 GB/s | 100 GB/s | 200-400 GB/s |
| Node Concurrency | 12 | 32 | O(100) | O(1000) |
| Interconnect BW | 1.5 GB/s | 10 GB/s | 25 GB/s | 50 GB/s |
| System Size (Nodes) | 18,700 | 100,000 | 500,000 | O(Million) |
| Total Concurrency | 225,000 | 3 Million | 50 Million | O(Billion) |
| Storage | 15 PB | 30 PB | 150 PB | 300 PB |
| I/O | 0.2 TB/s | 2 TB/s | 10 TB/s | 20 TB/s |
| MTTI | Days | Days | Days | O(1Day) |
| Power | 6 MW | ~10 MW | ~10 MW | ~20 MW |

Source : SciDAC,
www.scidacreview.org

SK hynix

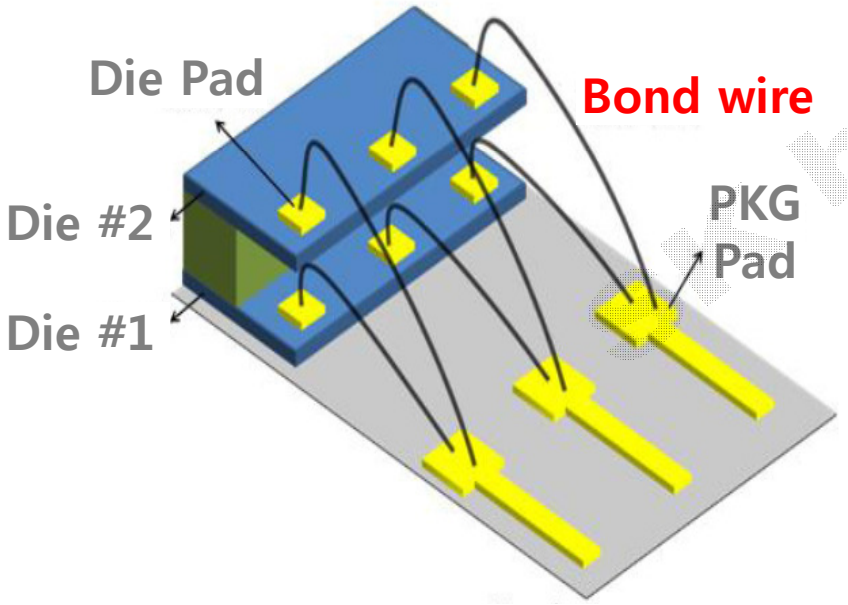
# Memory bottleneck & solution - Speed, Density, Power & SFF

## TSV is a revolutionary technology for overcoming the bottleneck

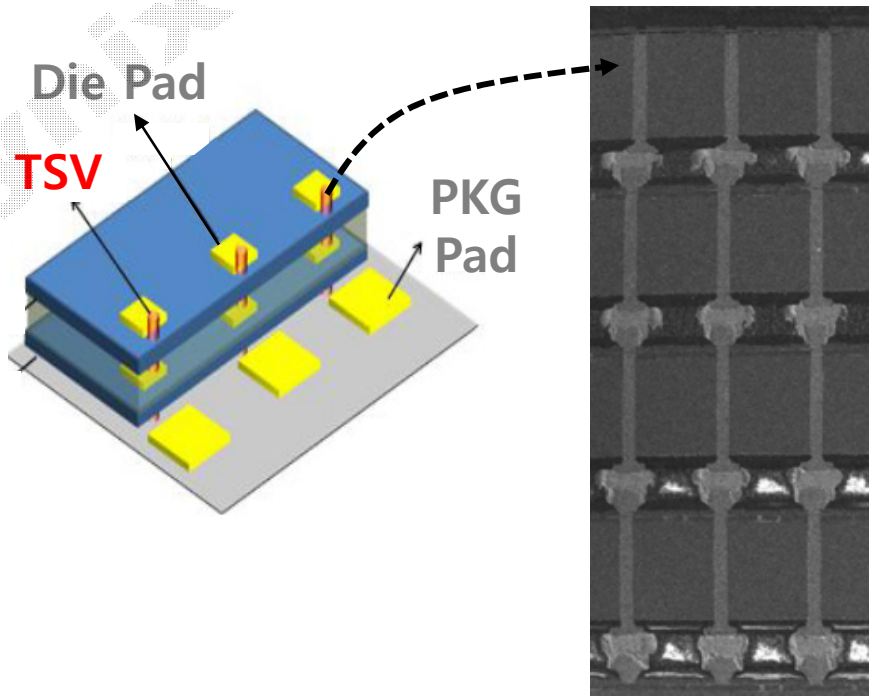


# TSV(Through Silicon via)

## TSV is the technology of 3D Stack
## (High Density / Small size PKG / High speed)

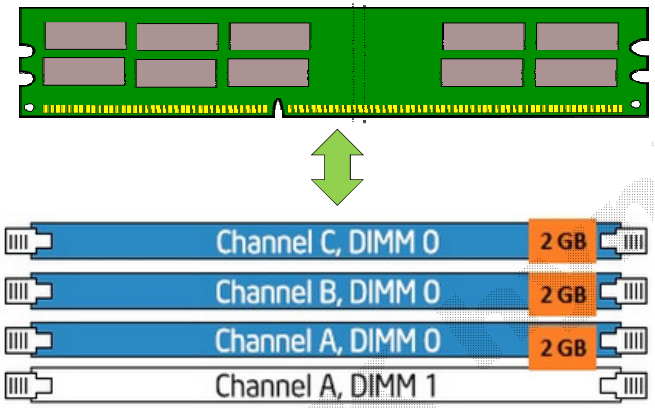
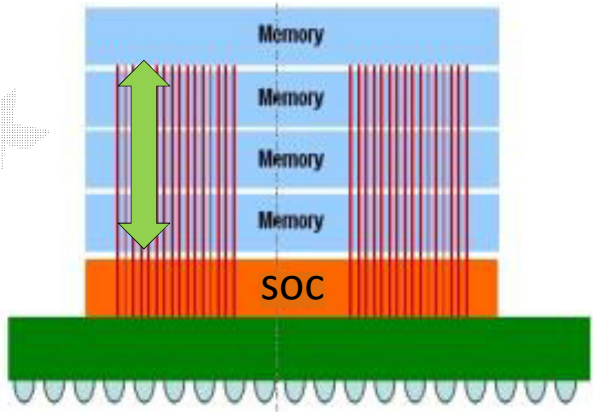



< Wire bonding PKG >



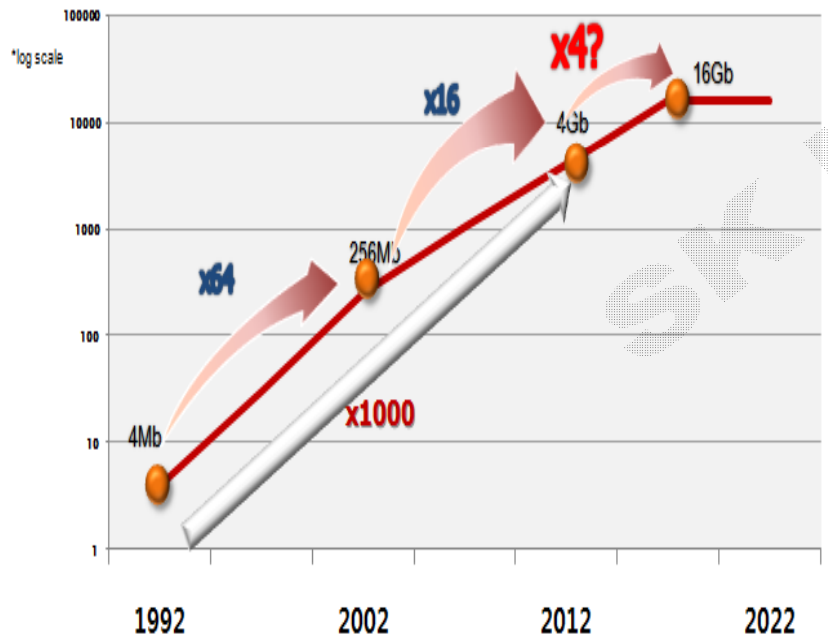

< TSV PKG >

# Bottleneck 1) Bandwidth

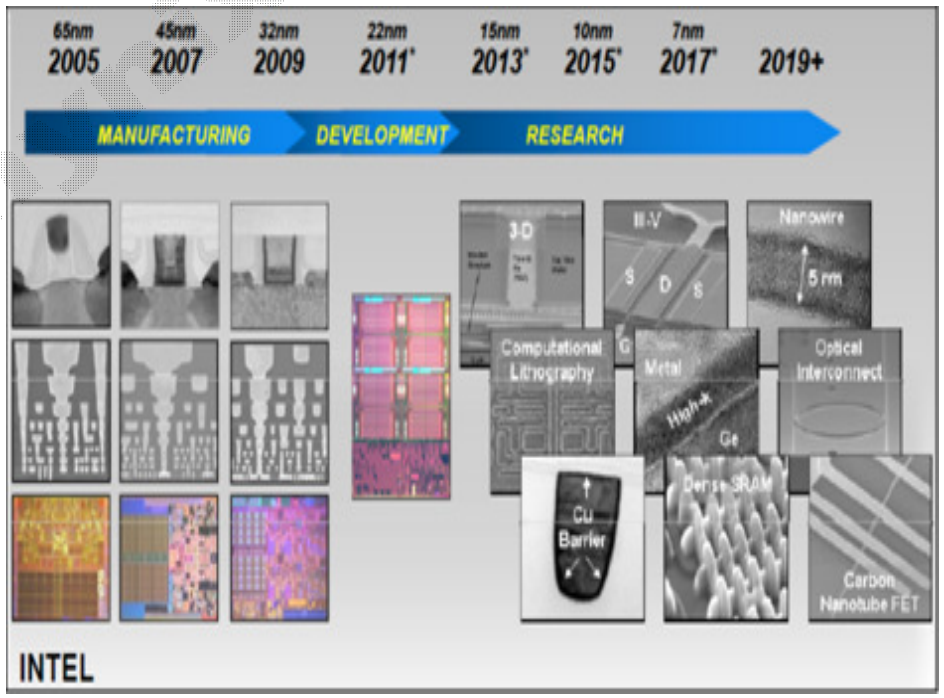
| | DDR3 | TSV(HBM) |
|---|---|---|
| Config. |  |  |
| IO | 64 DQ | 1024 DQ |
| Speed | 1.6G bps | 1~2Gbps |
| Bandwidth | 64 Gbps ➔ 12.8GBps | 1024 Gbps ➔ Max 256GBps |
| Compare | Long Length➔ RLC increase | Short Length➔ RLC decrease |

# Bottleneck 2) Technology Limit

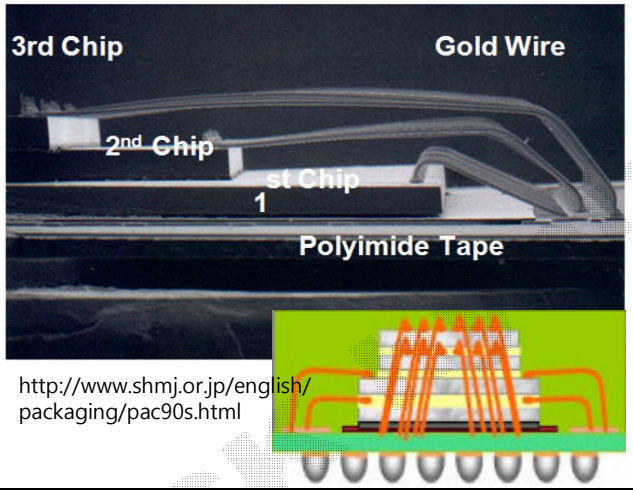
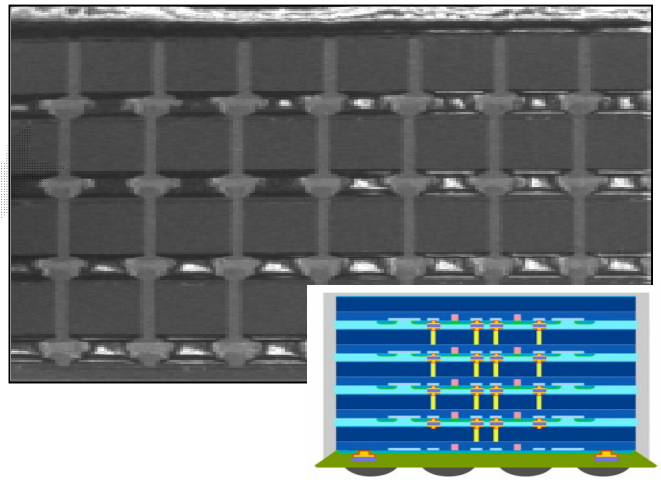
## 1. Capacity Limit




## 2. Scale Down Limit



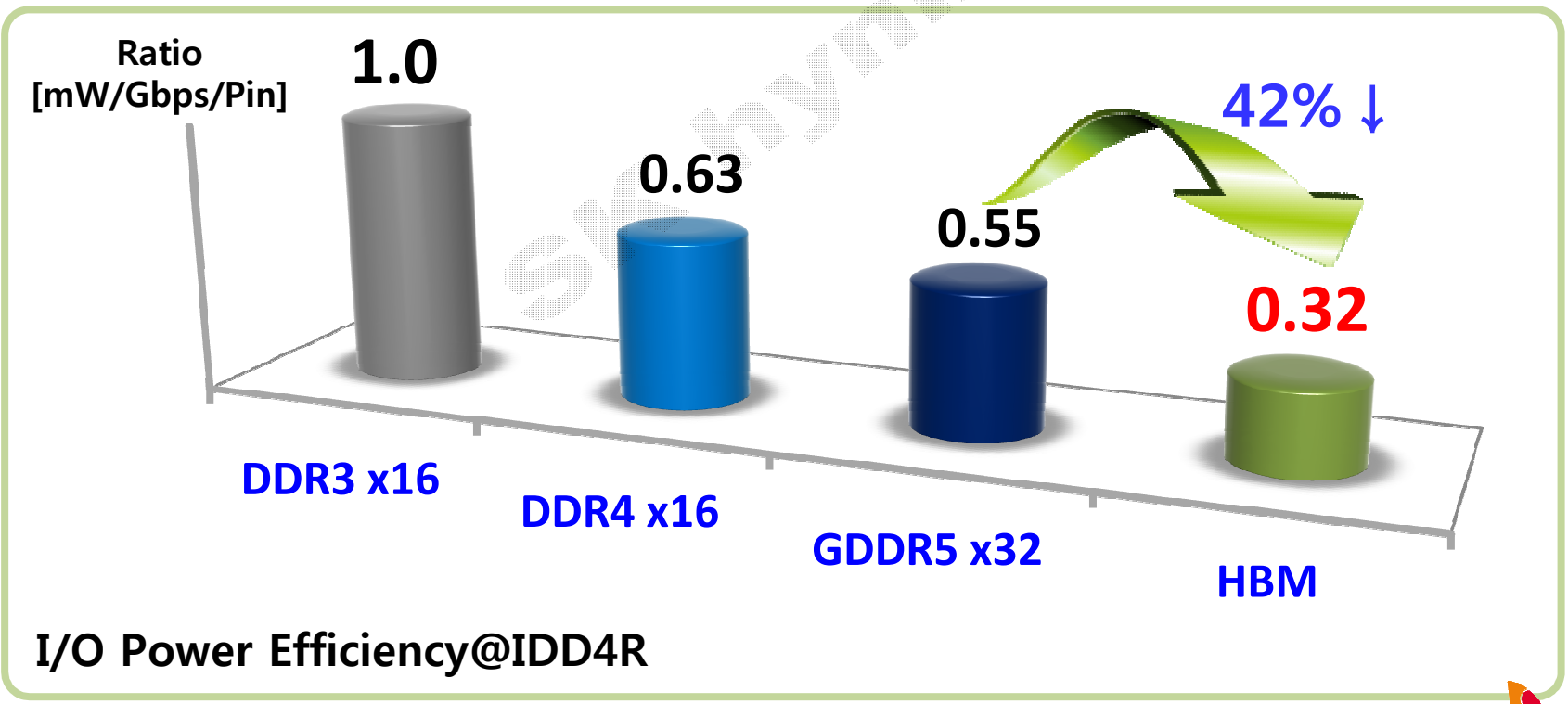
# Bottleneck 3) Small Form Factor

| | Wire bonding - DDR3 | TSV - HBM |
|---|---|---|
| Image |  http://www.shmj.or.jp/english/packaging/pac90s.html |  |
| PKG Size@die | 100% (117mm$^2$) | 36% (42mm$^2$) |
| mm$^2$@128GB/s | 100% (3744mm$^2$) | 11% (42mm$^2$) |
| Power Consumption* @128GB/s | 100% (6.4W) | 51% (3.3W) |

* Power Cal = IMPT

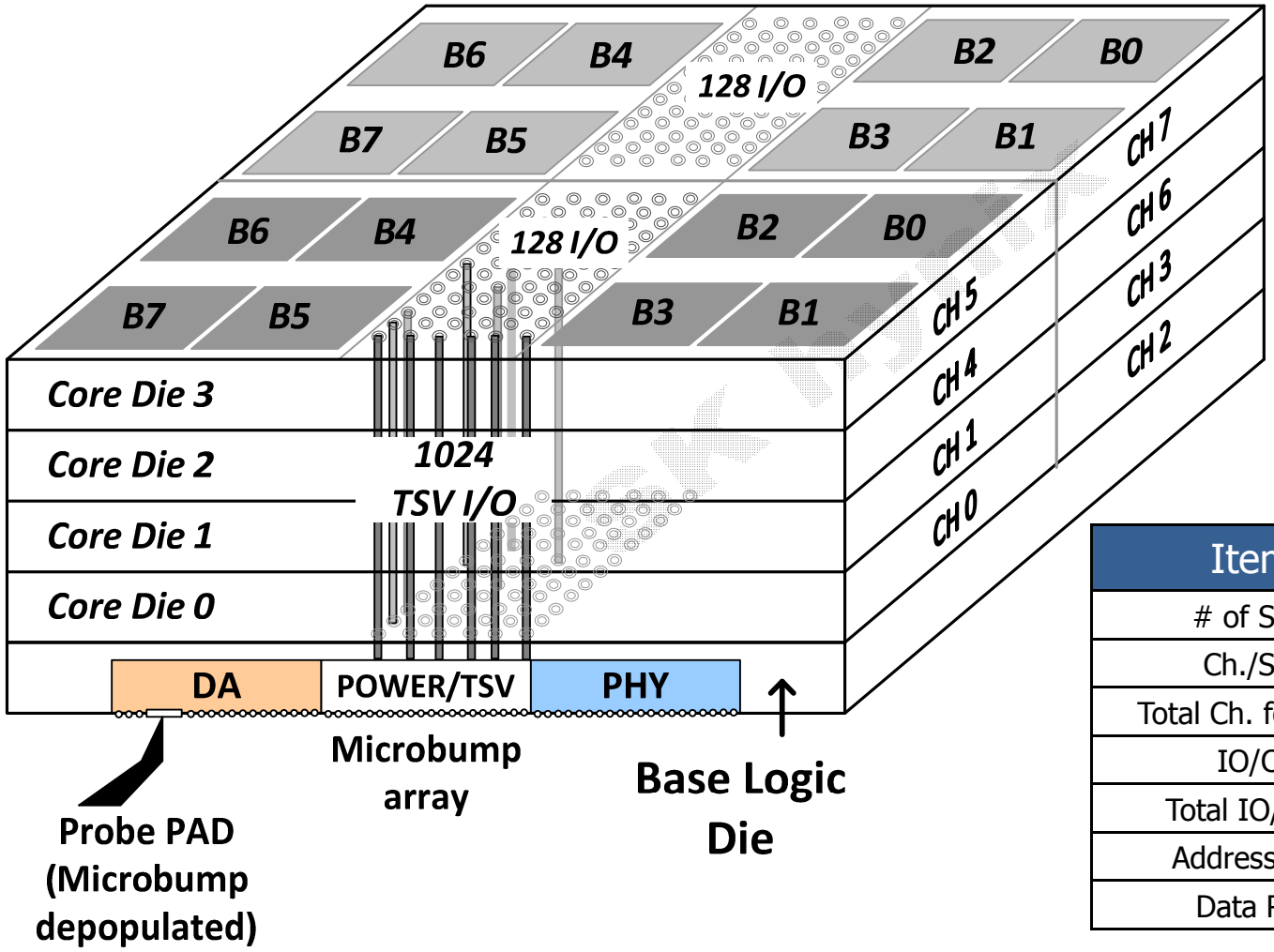
# Bottleneck 4) Low Power

- Lower Speed/pin and x1024 Wide IO ➔ low power consumption per Pin.
- Lower Cio(0.6pF), No Termination ➔ small IO current consumption
- Power consumption is decreased by 42% compared with GDDR5



# HBM Overall Architecture

## 4 Core DRAM + 1 Base logic die (Chip on Wafer)



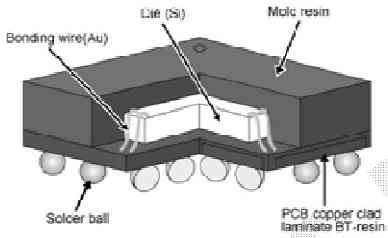

| Items | Target |
|---|---|
| # of Stack | 4(Core) + 1(Base) |
| Ch./Slice | 2 |
| Total Ch. for KGSD | 8 |
| IO/Ch. | 128 |
| Total IO/KGSD | 1024(=128 x 8) |
| Address/CMD | Dual CMD |
| Data Rate | DDR |

[1] D.U Lee, SK hynix, ISSCC 2014

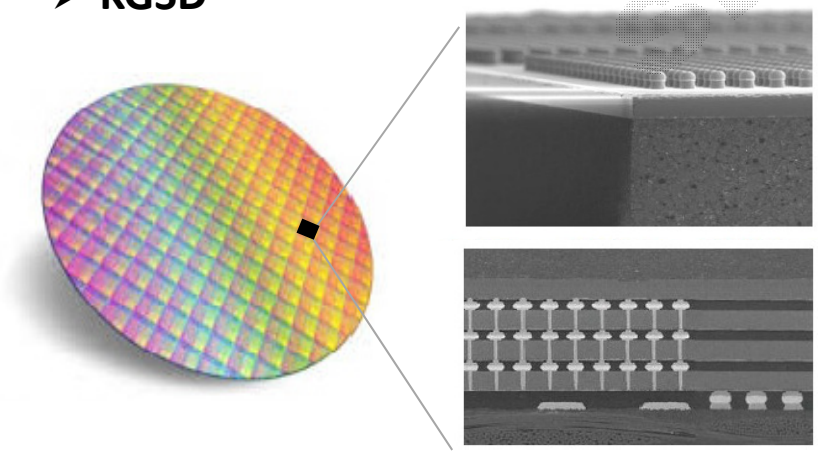
# HBM, What are the differences?

## KGSD* Memory supporting for System in package

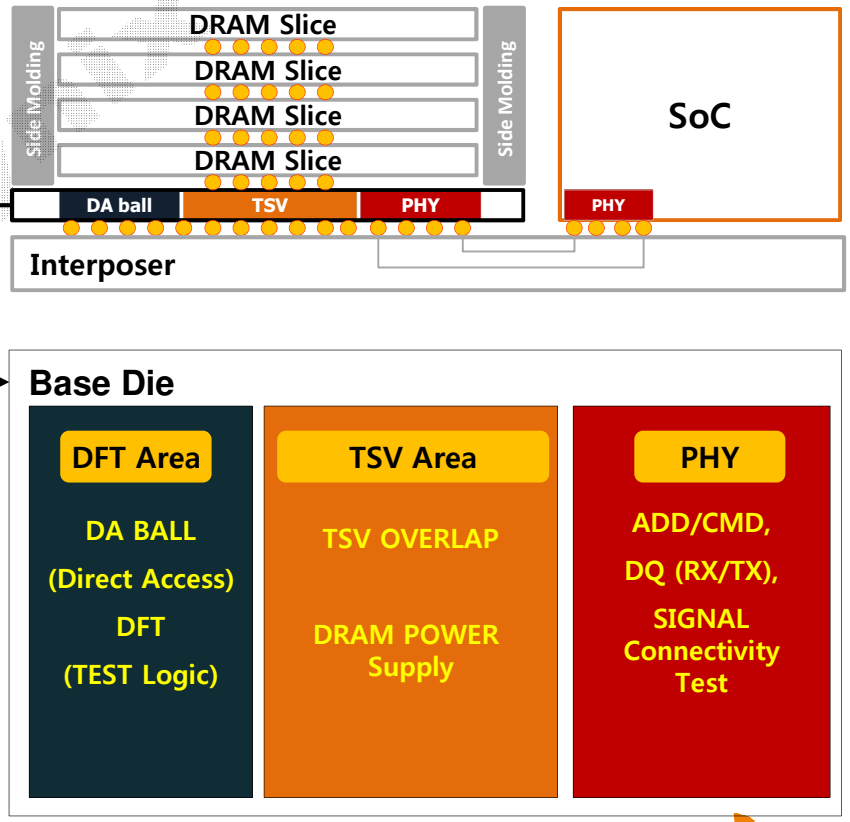
➢ **FBGA**


➢ **KGSD**



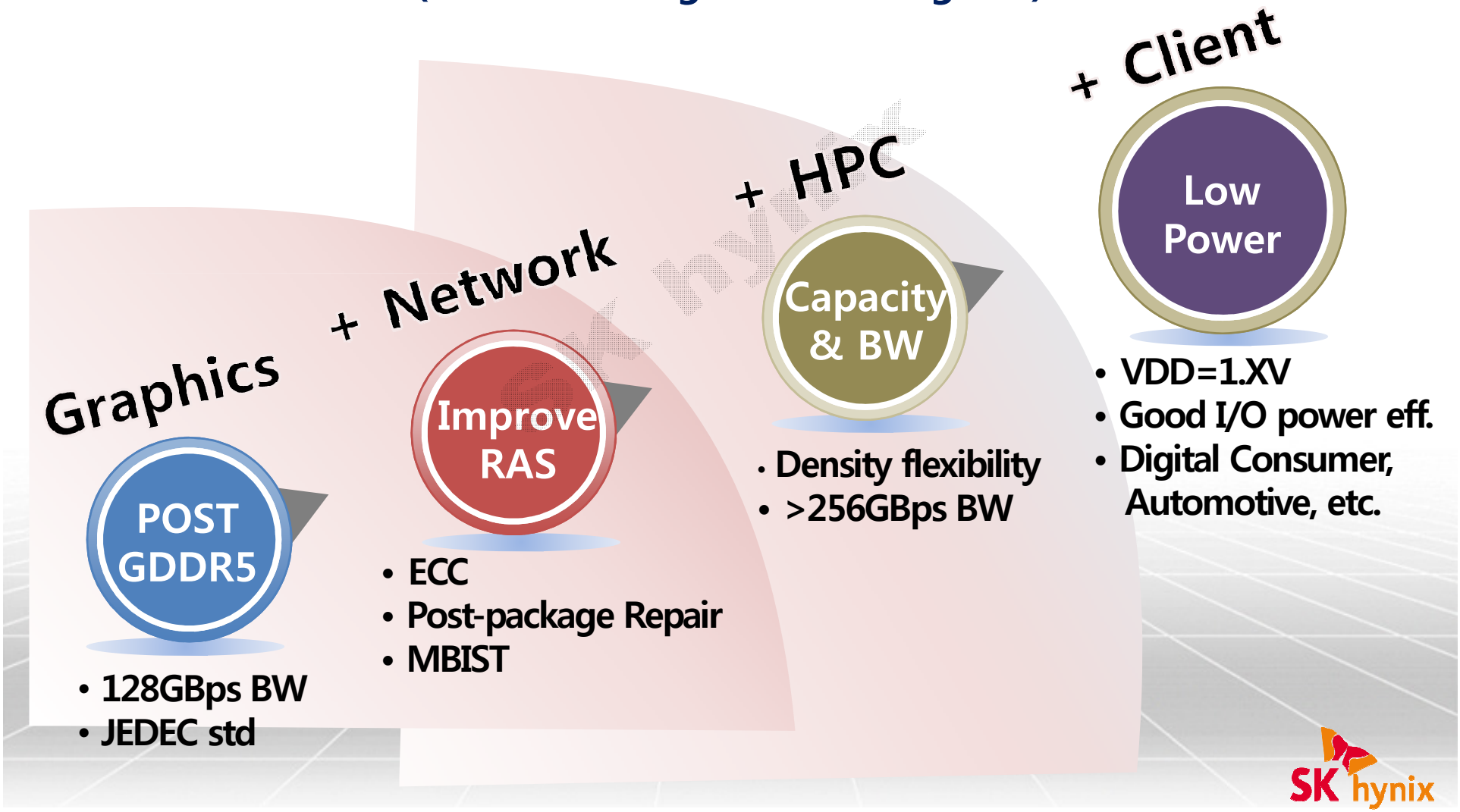
➢ **HBM in 2.5D SiP**




* KGSD (Known Good Stacked Die)

# HBM Market Segments

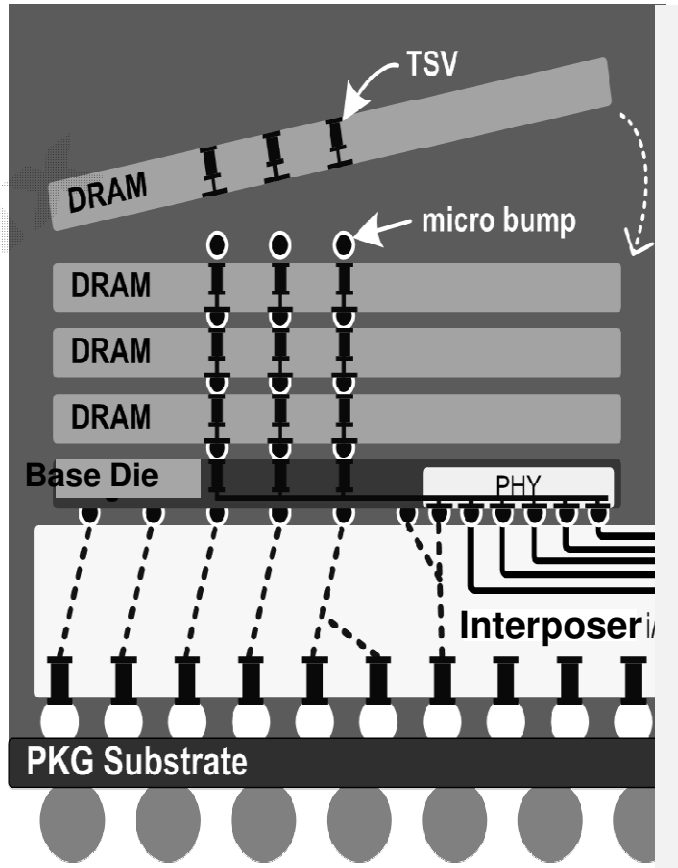
## HBM market will scale-out to various segments (Over 21 Design-Ins In Progress)



# HBM Overall specification

## ➢1nd Gen HBM

- 2Gb per DRAM die
- 1Gbps speed /pin
- 128GB/s Bandwidth
- 4 Hi Stack (1GB)
- x1024 IO
- 1.2V VDD
- KGSD w/ µBump



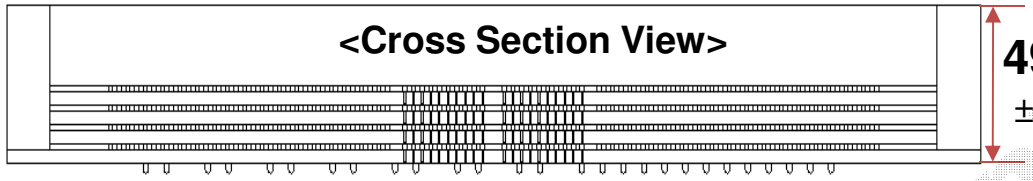

## ➢2nd Gen HBM

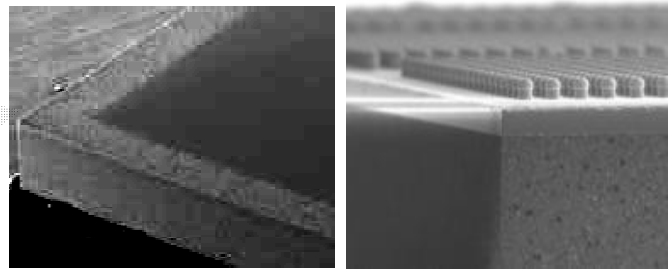
- 8Gb per DRAM die
- 2Gbps speed/pin
- 256GBps Bandwidth/Stack
- 4/8 Hi Stack (4GB/8GB)

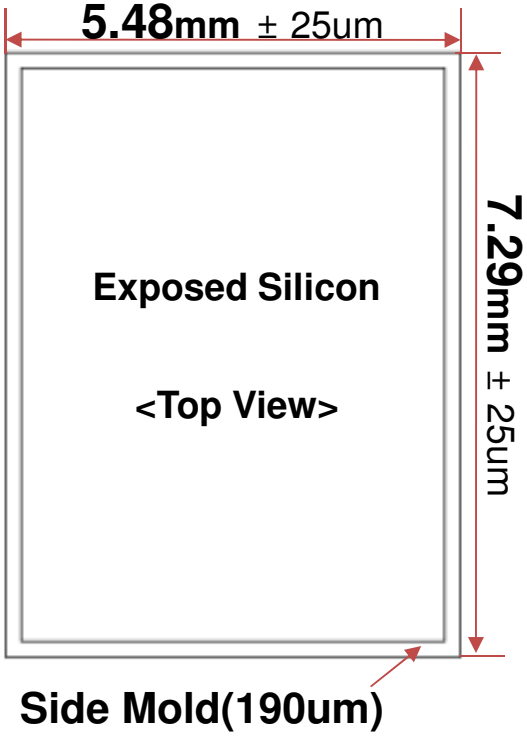
# HBM Gen1 5mKGSD Structure

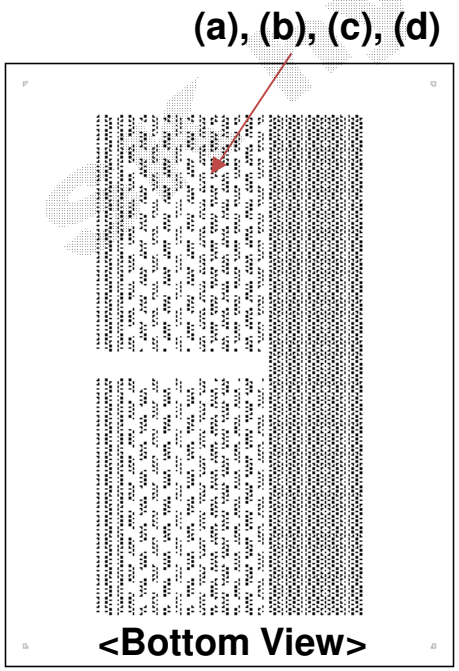
## 5mKGSD (molded Known Good Stacked Die)

**- 1 Base + 4 Core (DRAM) with Side Mold -**






| | Item | Value | Remark |
|---|---|---|---|
| (a) | uBump Diameter | **25um** (± 3 um) | |
| (b) | uBump Height | **35um** (± 3.5 um) | Cu/Ni/SnAG **(17/3/15um)** |
| (c) | uBump Pitch | **55um** | |
| (d) | uBump Array (MPGA) | JEDEC | JC11-2.883 JC11-4.884 |

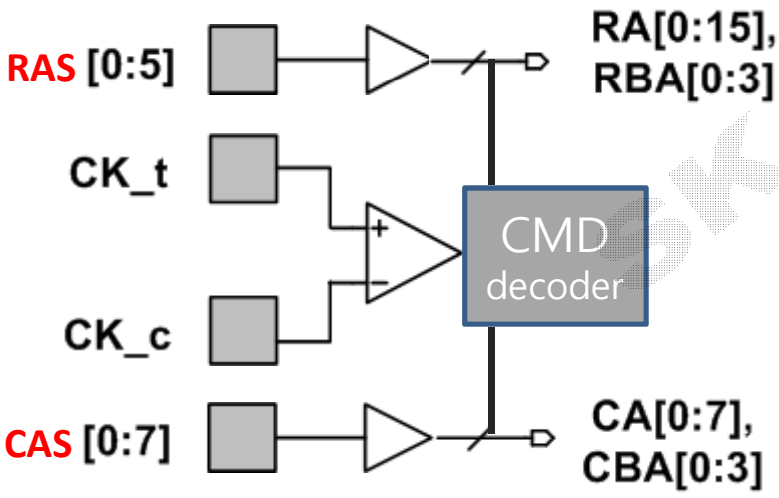
# Comparison of HBM and other DRAMs

| Item | DDR3 (x8) | GDDR5 (x32) | 4-Hi HBM (x1024) |
|---|---|---|---|
| **I/O** | **8** | **32** | **1024** |
| Prefetch (Per IO) | 8 | 8 | 2 |
| Access Granularity (=I/O x Prefetch) | 8Byte | 32Byte | 256Byte |
| **Max. Bandwidth** | **2GB/s** | **28GB/s** | **128~256GB/s** |
| tRC | 40~48ns | 40ns(=1.6v, 1.5v)<br>48ns(=1.35v) | 40~48ns |
| tCCD | 4ns (=4tCK) | 2ns (=4tCK) | 2ns (=1tCK) |
| VPP | Internal VPP | Internal VPP,<br>(Opt. Ext. VPP) | Ext. VPP |
| VDD | 1.5, 1.35 | 1.6, 1.5, 1.35 | 1.2 |
| **CMD Input** | **Single CMD** | **Single CMD** | **Dual CMD** |
| **Refresh Single Bank** | **X** | **X** | **O** |
| DBI mode | X | O (DBI_DC) | O (DBI_AC) |

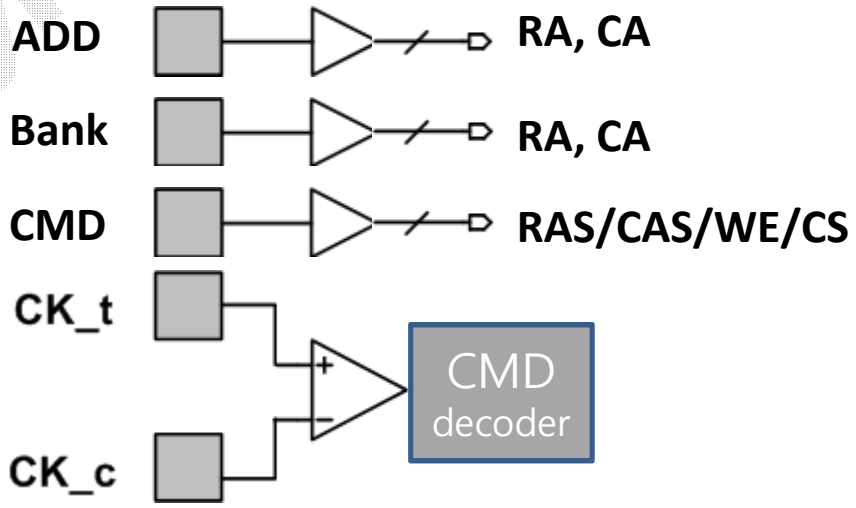
# Dual CMD interface

## CMD efficiency increased by Semi-independent row/column input

**Row/column input through different pins**



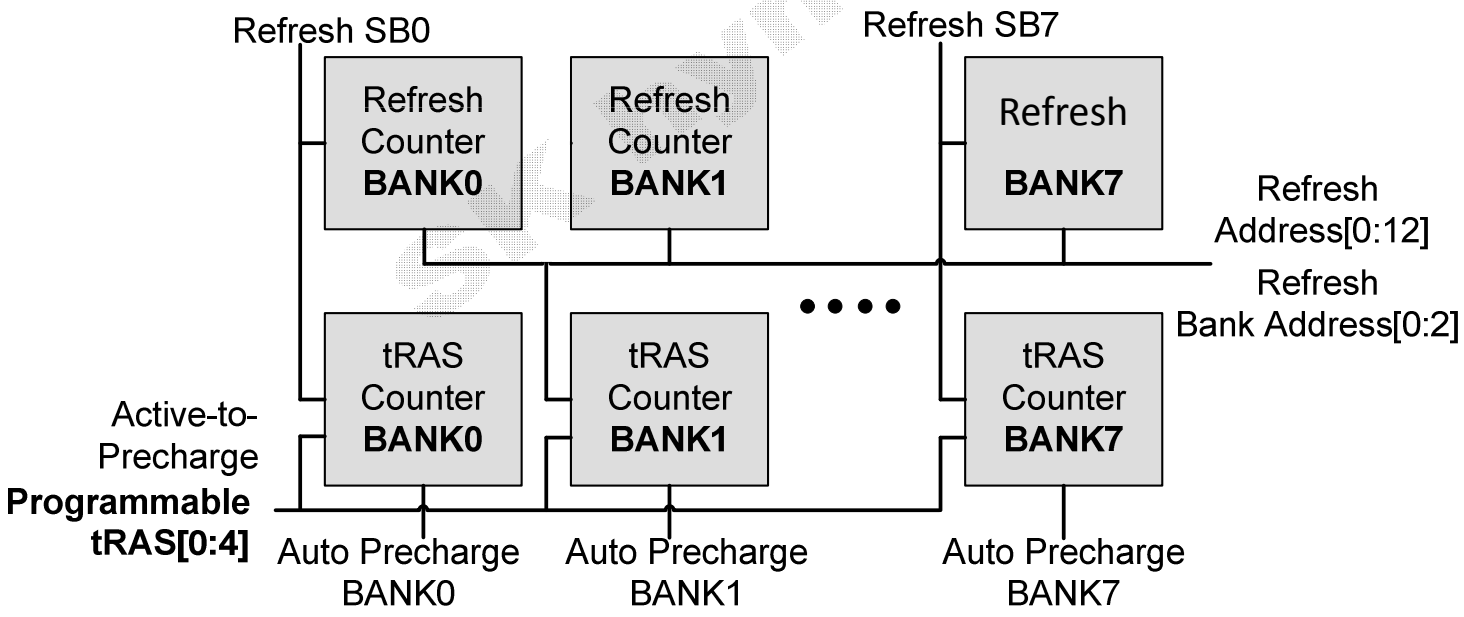

**Conventional drams share RAS/CAS CMD.**



# REF Single Bank

## Single bank refresh and programmable tRAS

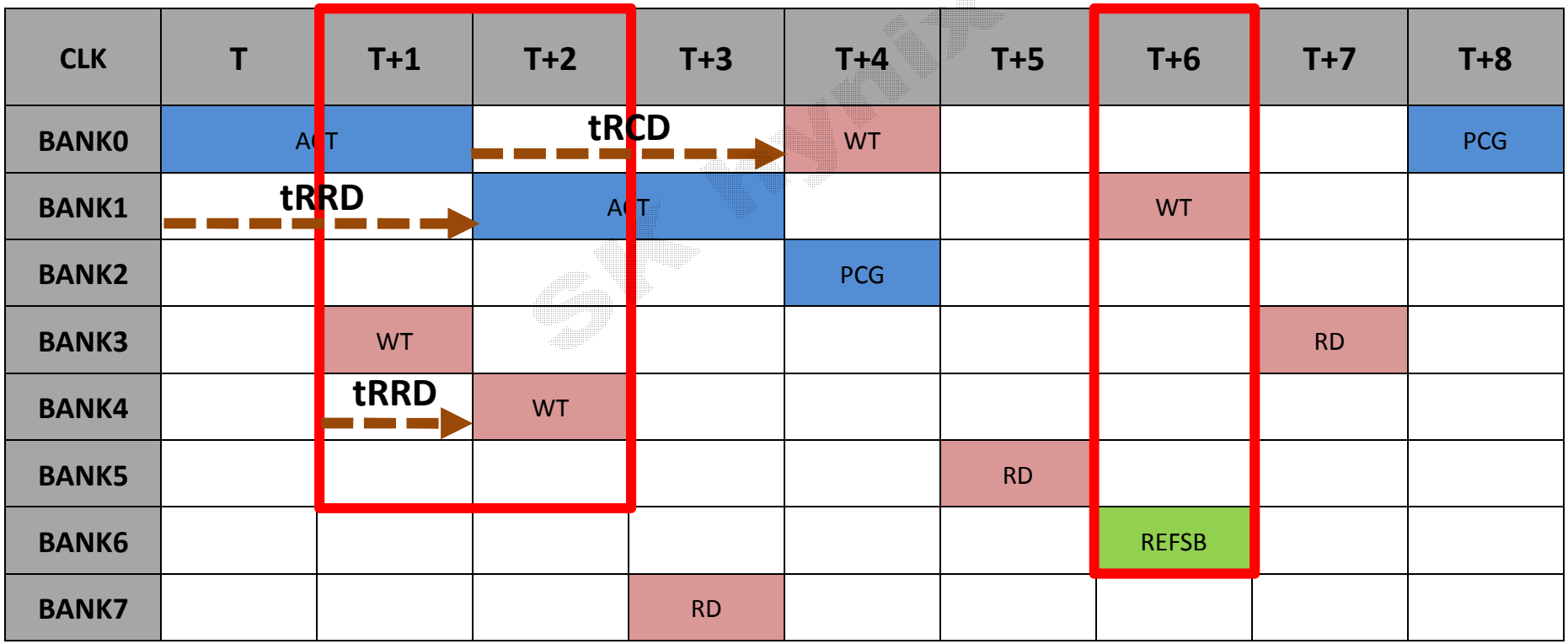
**Concurrent read/write operation with single bank refresh allows data bus to remain active.**



# REF Single Bank

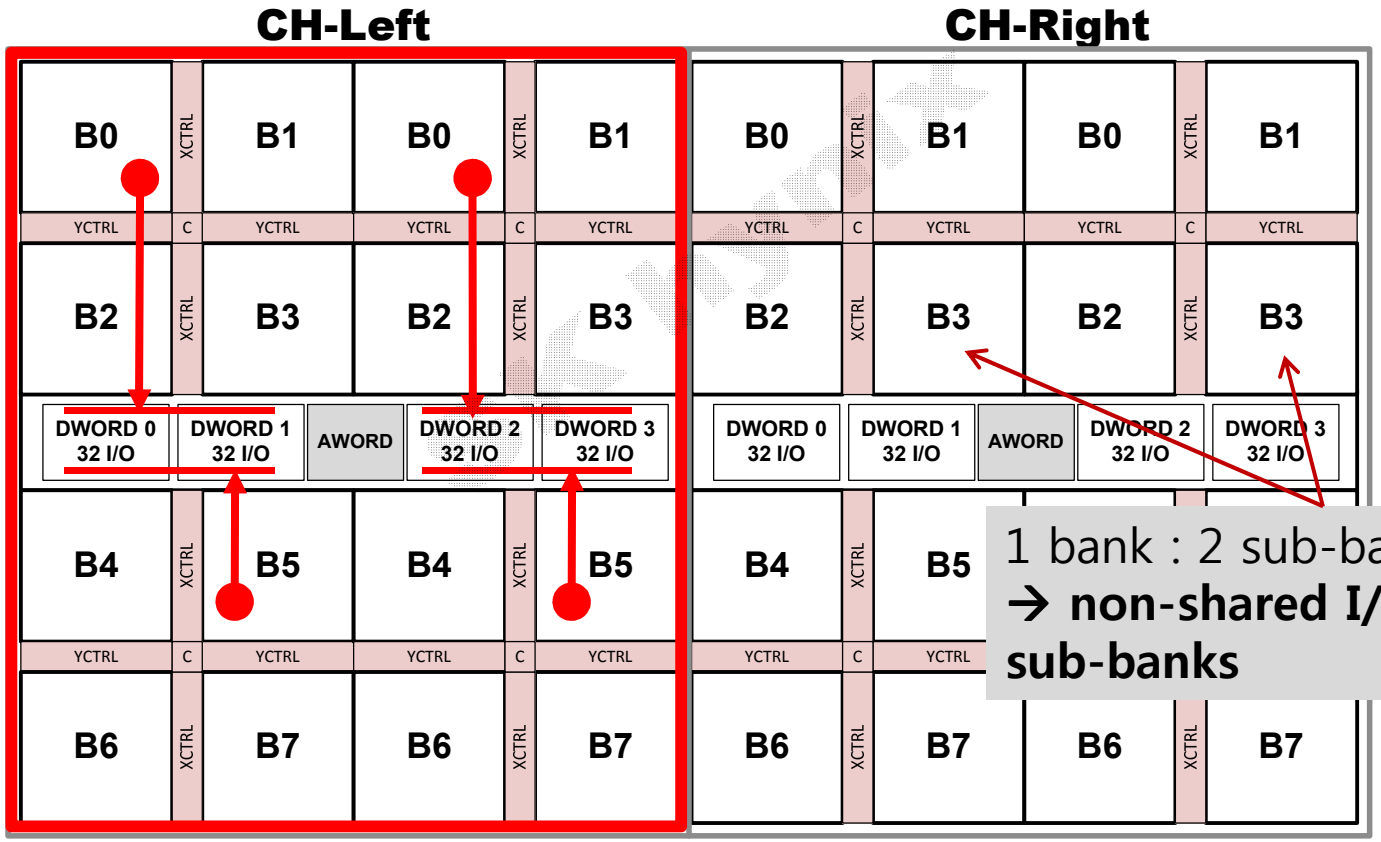
## Command bus efficiency can be maximized

### - Dual Command & REF single bank -



# HBM Core Architecture

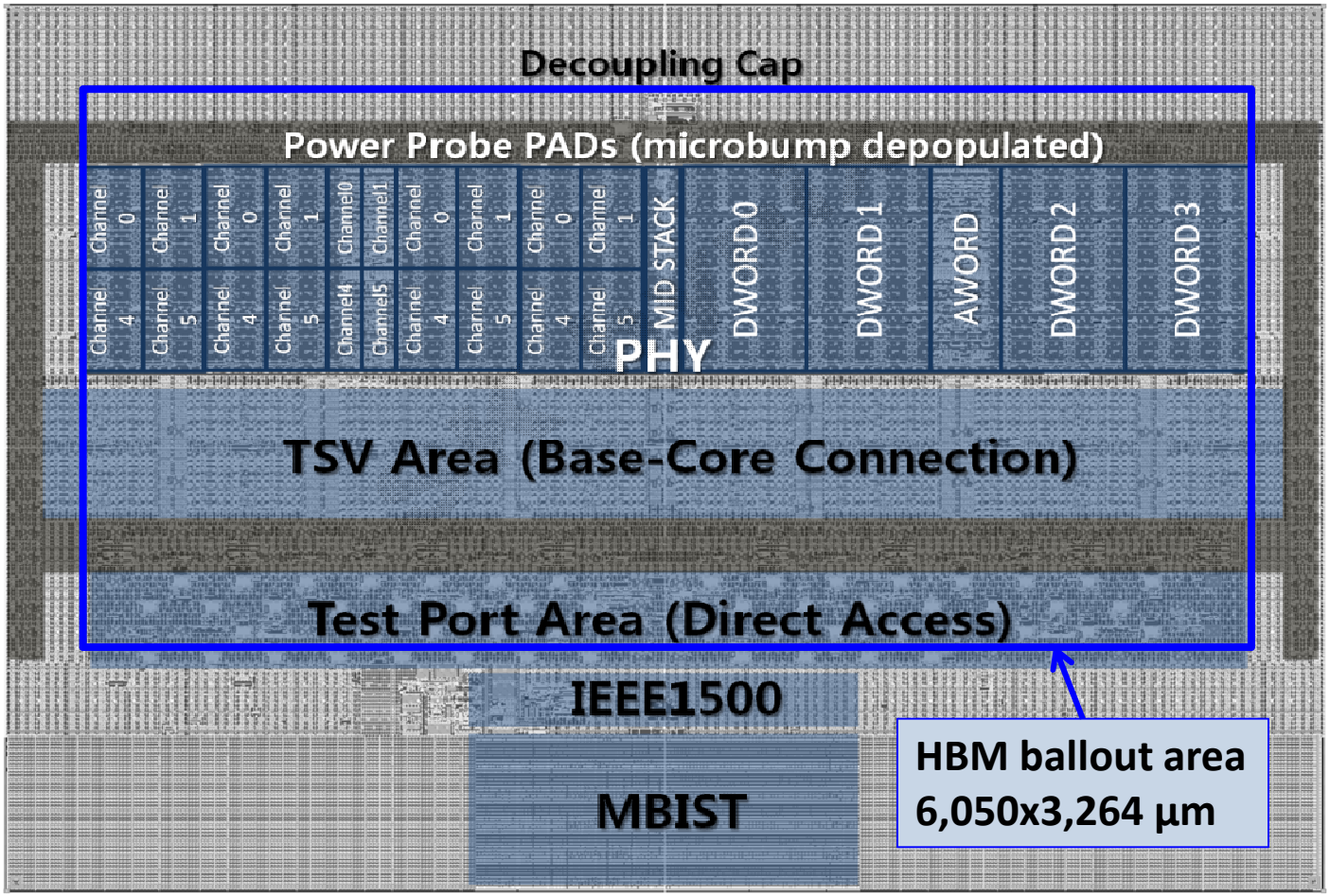
## HBM single die has 2channels
## 1 channel consists of 128 TSV I/O with 2n prefetch




[2] D.U Lee, SK hynix, ISSCC 2014

# HBM Base Die Architecture

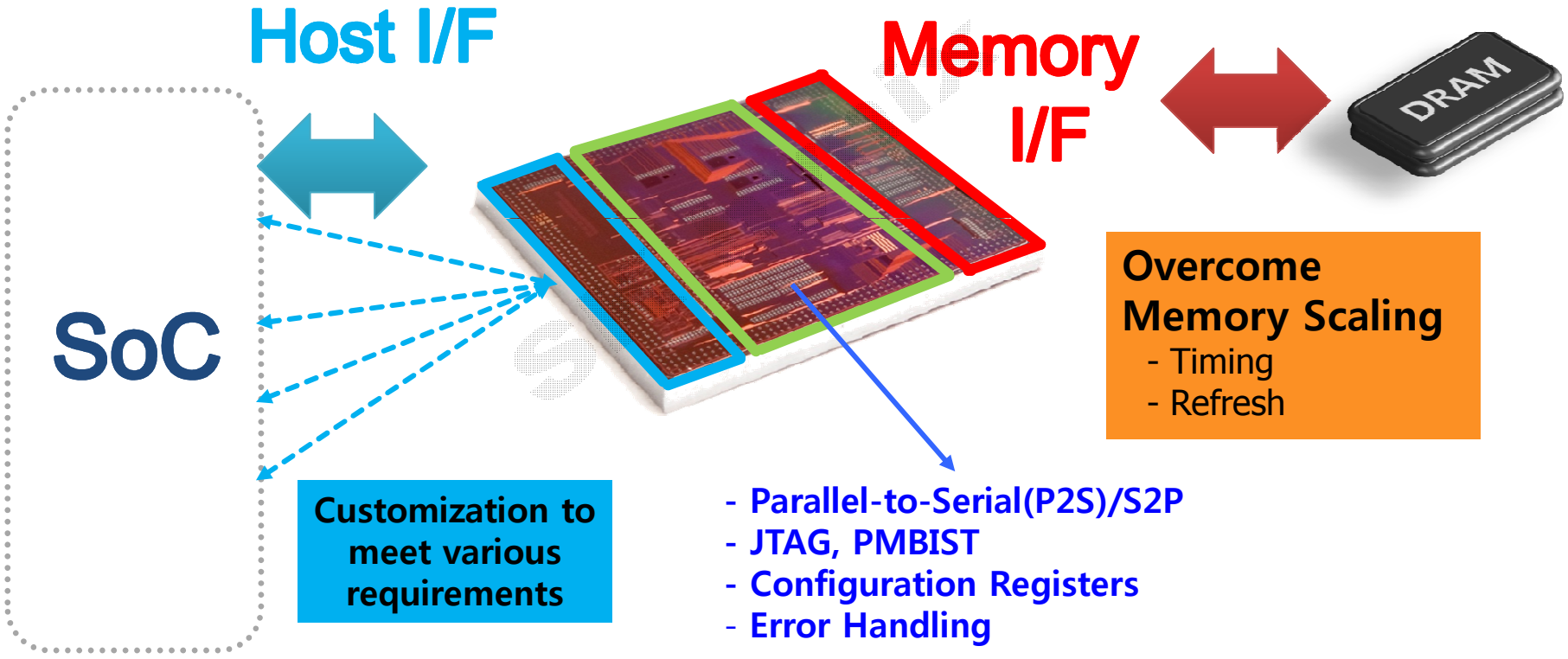
## Base die consists of 3 Areas – PHY, TSV, Test Port Area




[3] D.U Lee, SK hynix, ISSCC 2014

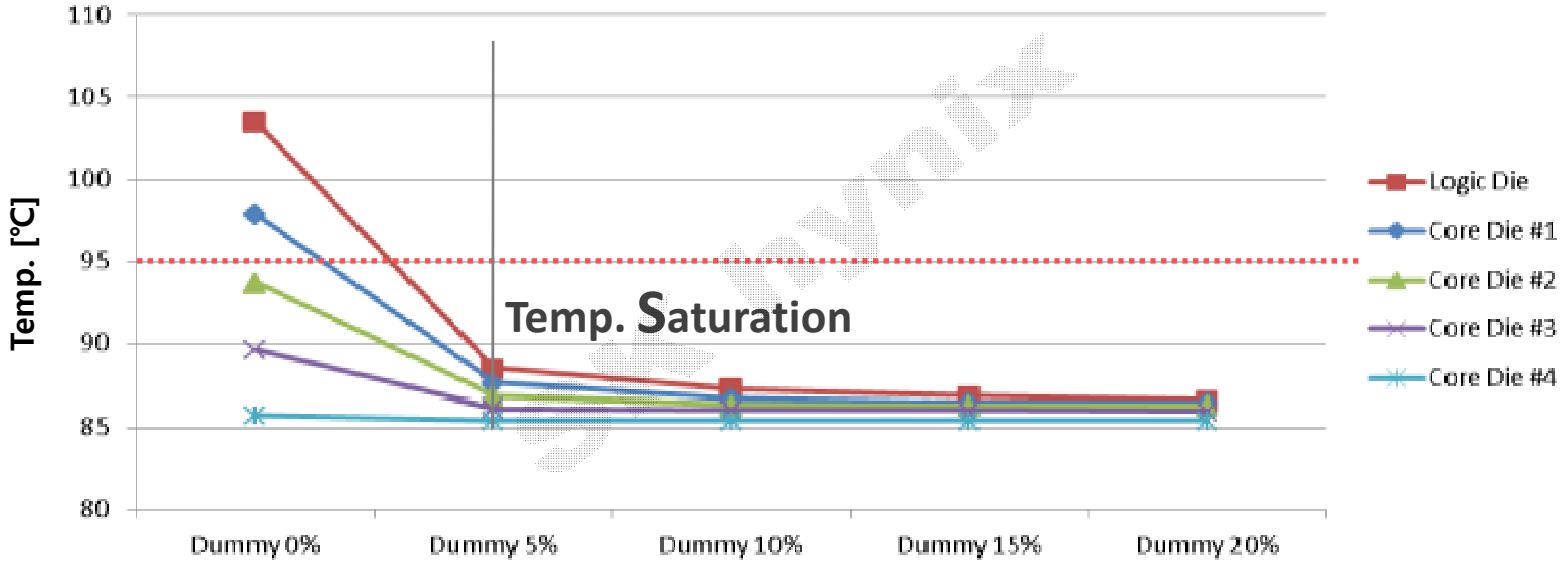
# Base Die Customization – Future HBM Concept

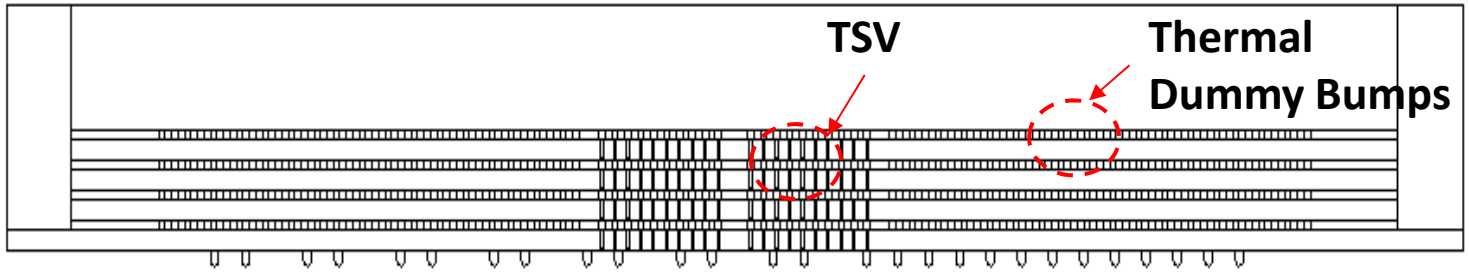
**Logic Layer ➔ Host I/F + Memory I/F + Base Logic/IP Block**



# HBM Thermal Management

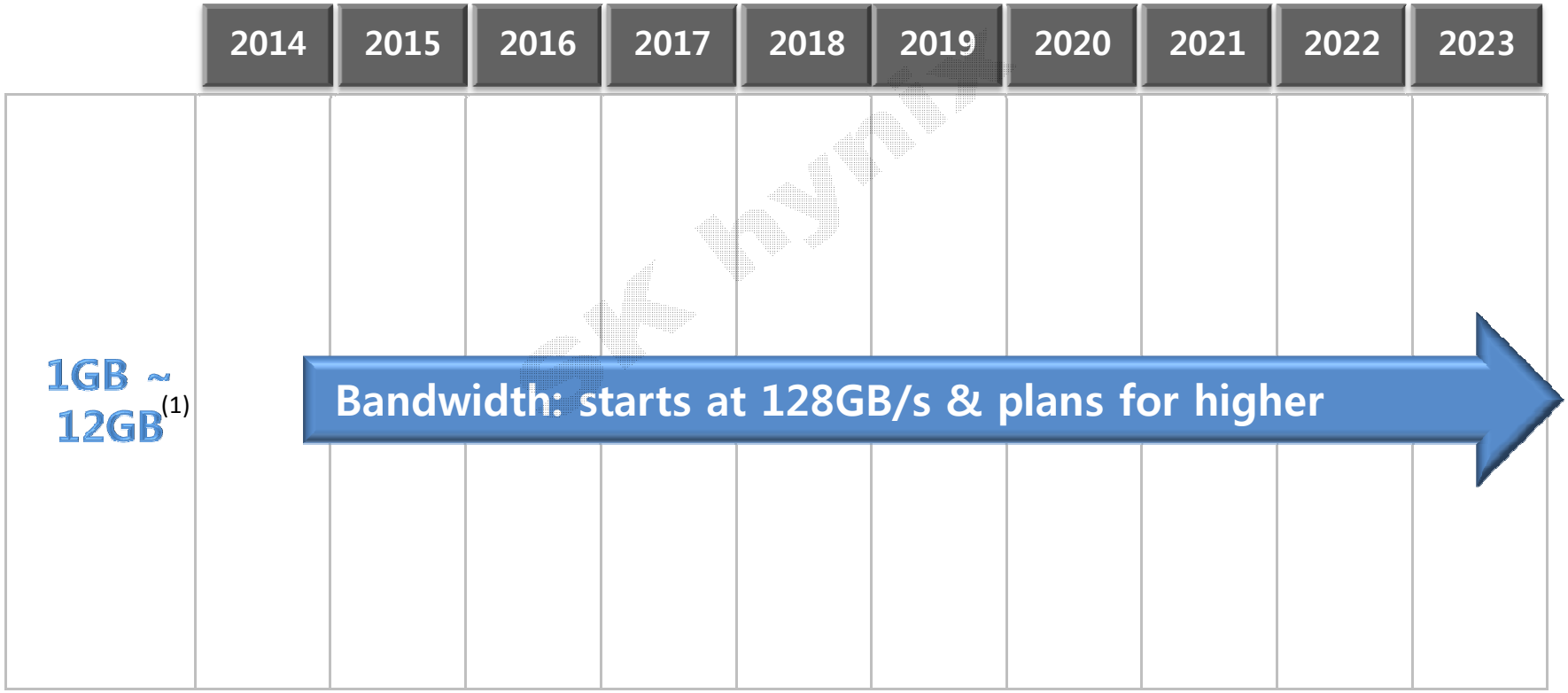
**Thermal dummy bumps as well as well-designed device architecture are helpful for thermal dissipation → No mechanical reliability issues by thermal dummy bumps.**



# HBM Long-term Roadmap(2) (Preliminary)

## HBM product longevity is critical in several applications
## SK hynix plans to address longevity requirement



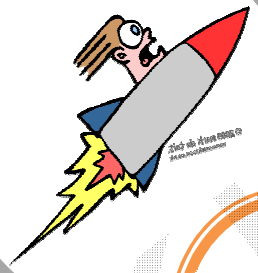

•Note 1 – anticipated future HBM density
•Note 2 – roadmap is subject to changes without prior notifications

# HBM Summary

## Perfect memory solution for various application requirement

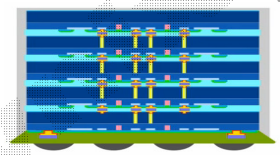
### High Bandwidth

~256GB/s



### High Density PKG

Up to 8GB

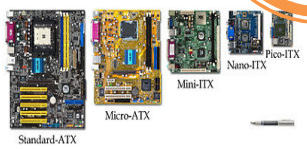


4 ~ 8 Stacks
KGSD

**HBM**

### Smaller Form Factor

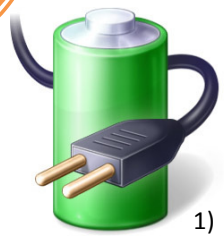
-65%




1 ~ 4 Cubes
per GPU

### Good Power Efficiency

68%



1)

SK hynix

# Thank You !