Design Considerations of HBM Stacked DRAM and the Memory Architecture Extension

Dong Uk Lee, Kang Seol Lee, Yongwoo Lee, Kyung Whan Kim,
Jong Ho Kang, Jaejin Lee, Jun Hyun Chun

Outline

- **Introduction**
- **HBM stacked memory**
  - Heterogeneous chip structure
- **Design considerations of HBM**
  - I/O test, TSV scan and repair, power distribution
- **Memory architecture**
  - Pseudo channel architecture
  - Multi-rank architecture
- **Conclusion**

Applications of HBM

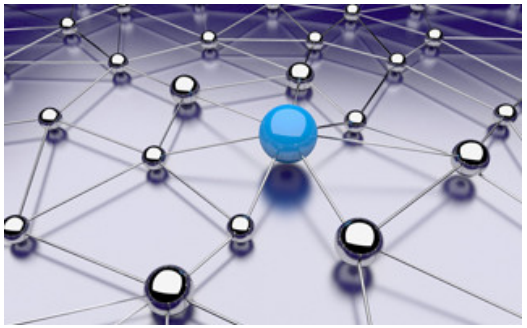
Graphics

HBM

128GB~1TB/s

(1-4 total memory)

HPC

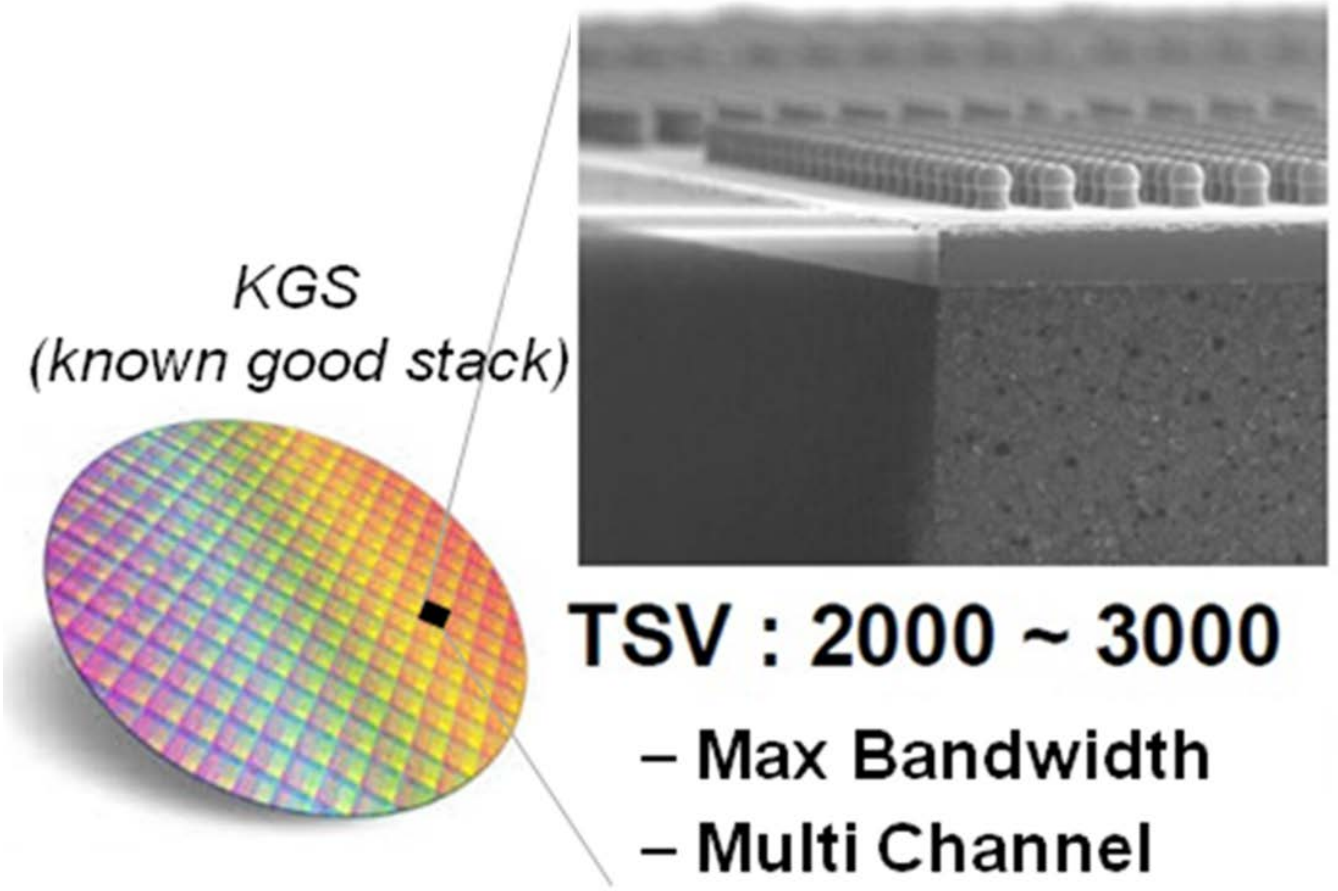


Network

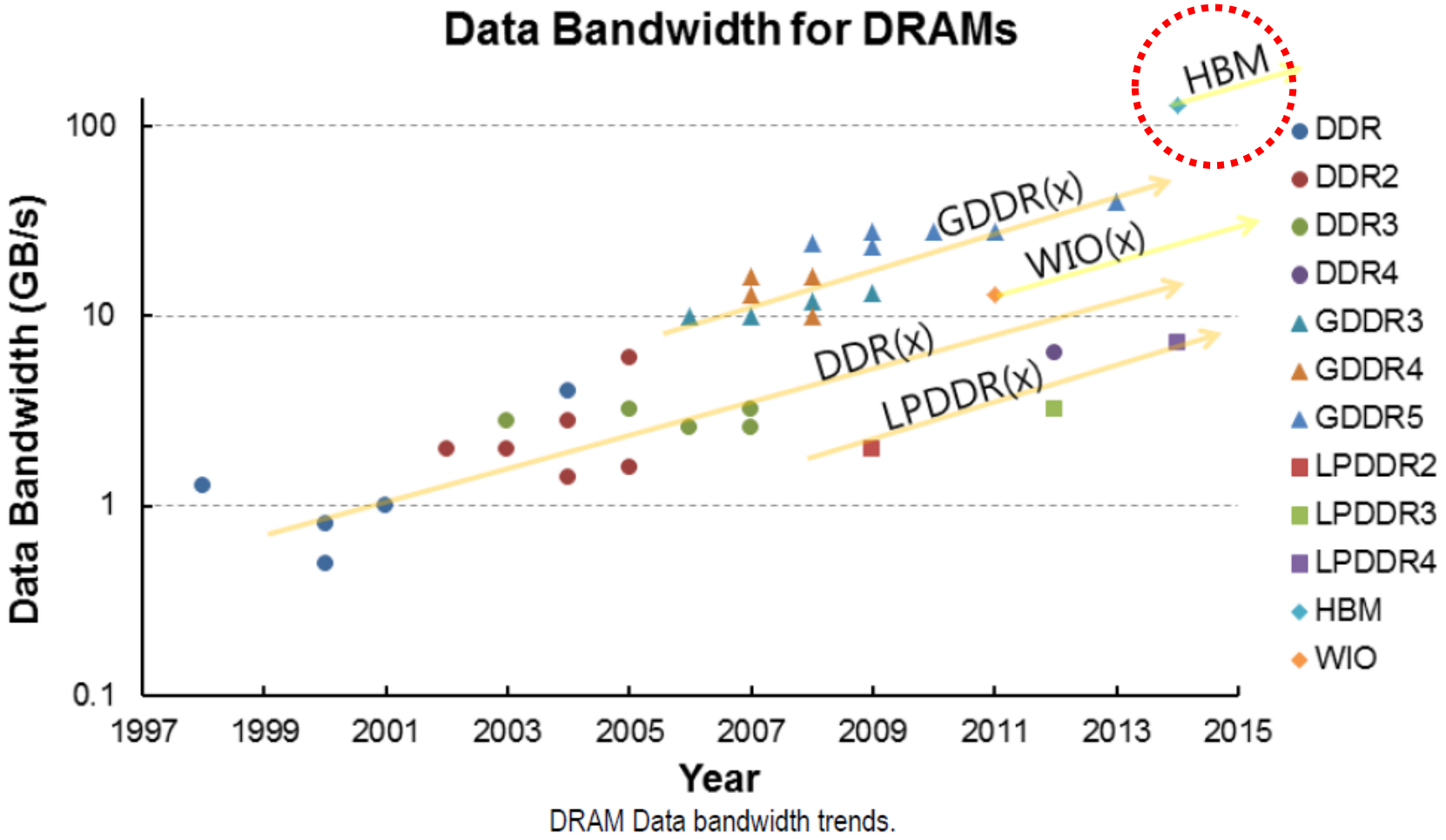
High-end Client

Known Good Stack

- **Known good stack is selected after Chip-on-Wafer processing and testing**



Chip Performance



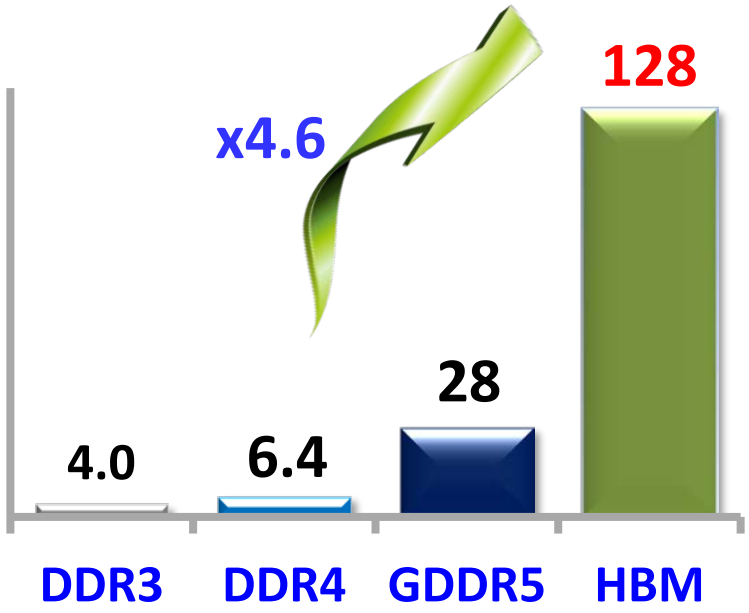
DRAM Data bandwidth trends.

Source : ISSCC 2014 Trends

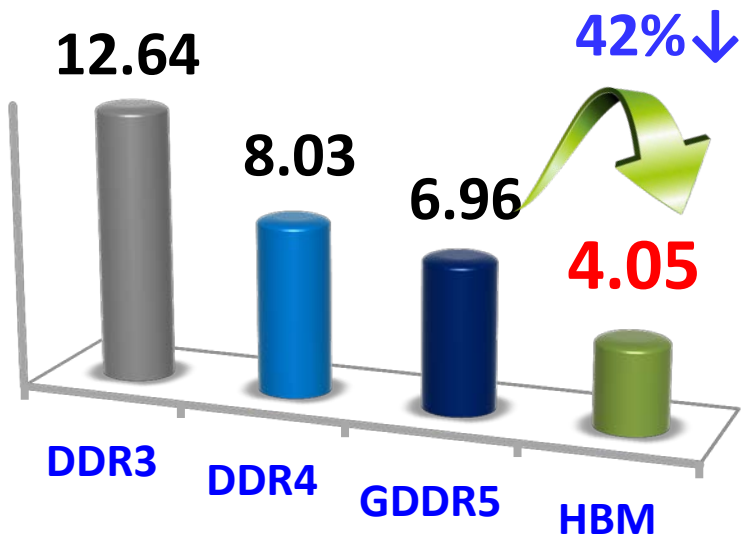
Energy Efficiency

- There is the limitation in system cooling
- Energy efficiency is important in high-bandwidth

Bandwidth per chip



Energy Efficiency@IDD4R

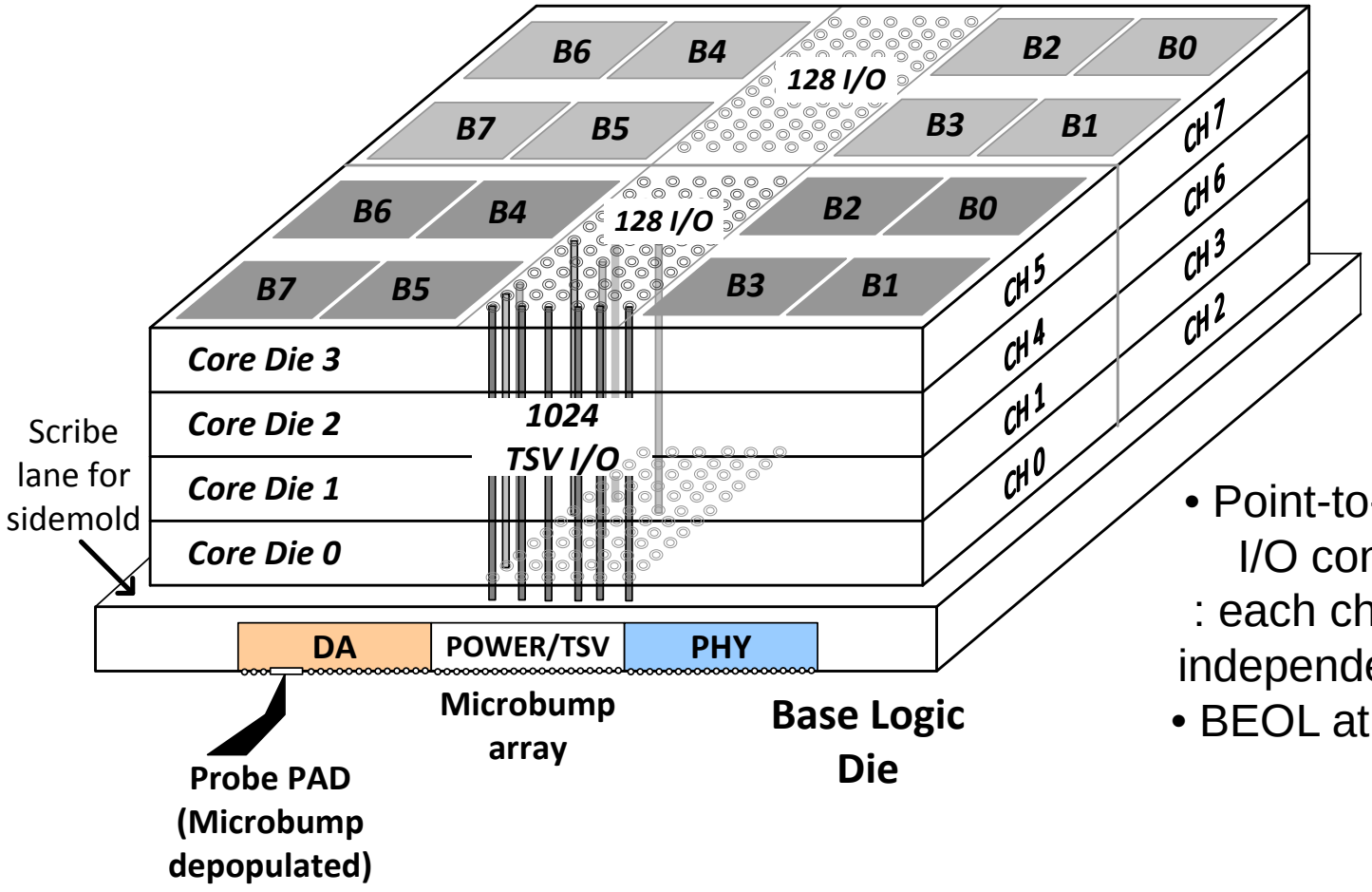


HBM Stacked Memory

- **High Bandwidth Memory(HBM) Features**
  - Stacked DRAM architecture
  - Multi channel (4-8)
  - Dual row/column command interface
  - Chip-on-Wafer process
  - Microbump interface
  - Test method for production(Loopback, MBIST, Scan..)
- **Chip Performance (1st gen.)**
  - 8 channel, 64 banks
  - 1024 I/O, 128GB/s

High Bandwidth Memory Architecture

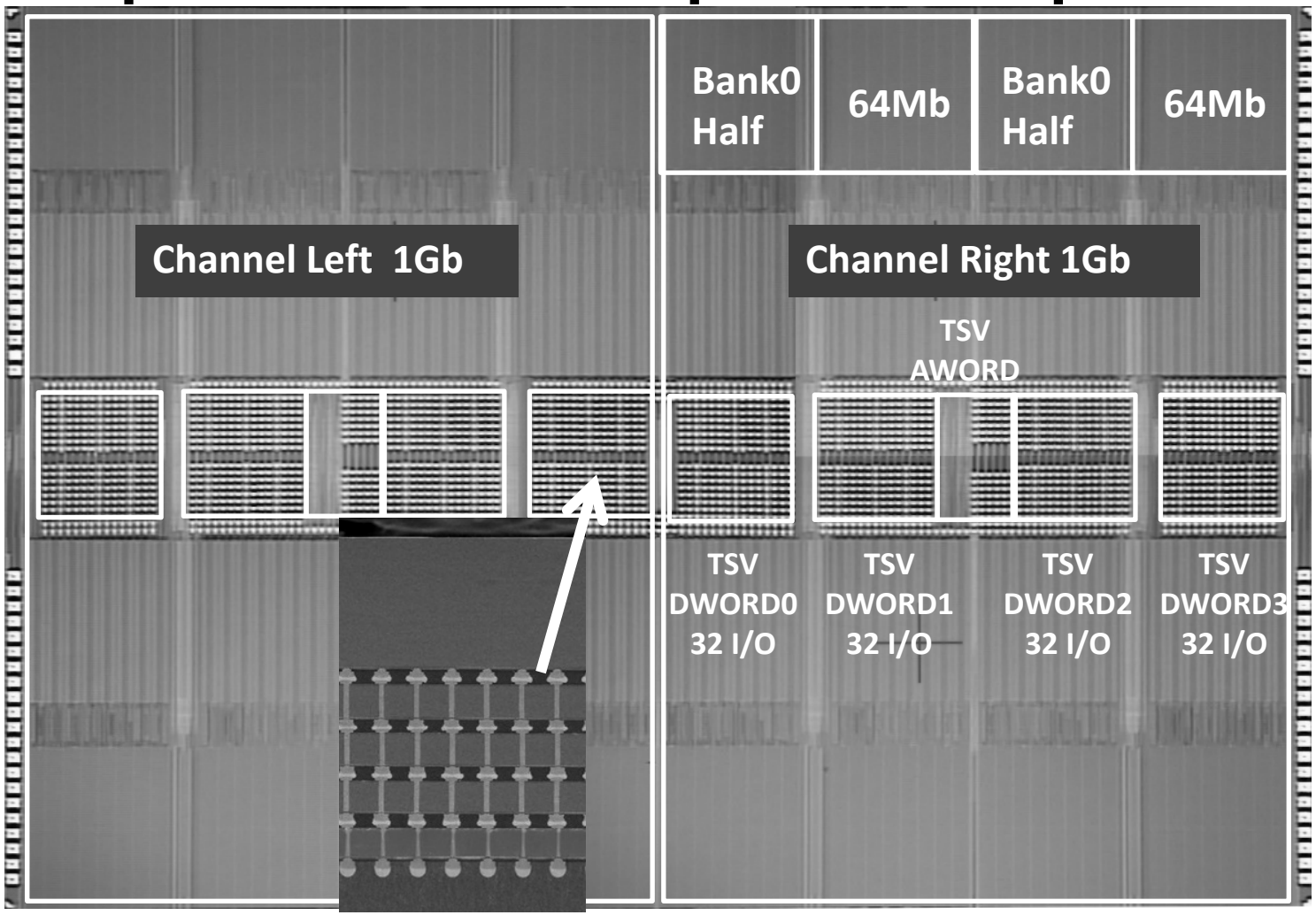
- **Total 8 channels : 128 I/O x 8 ch = 1024 I/O**
  - 4 core DRAM die + 1 base logic die



- Point-to-point TSV I/O connection : each channel has independent 128 I/O
- BEOL at the bottom

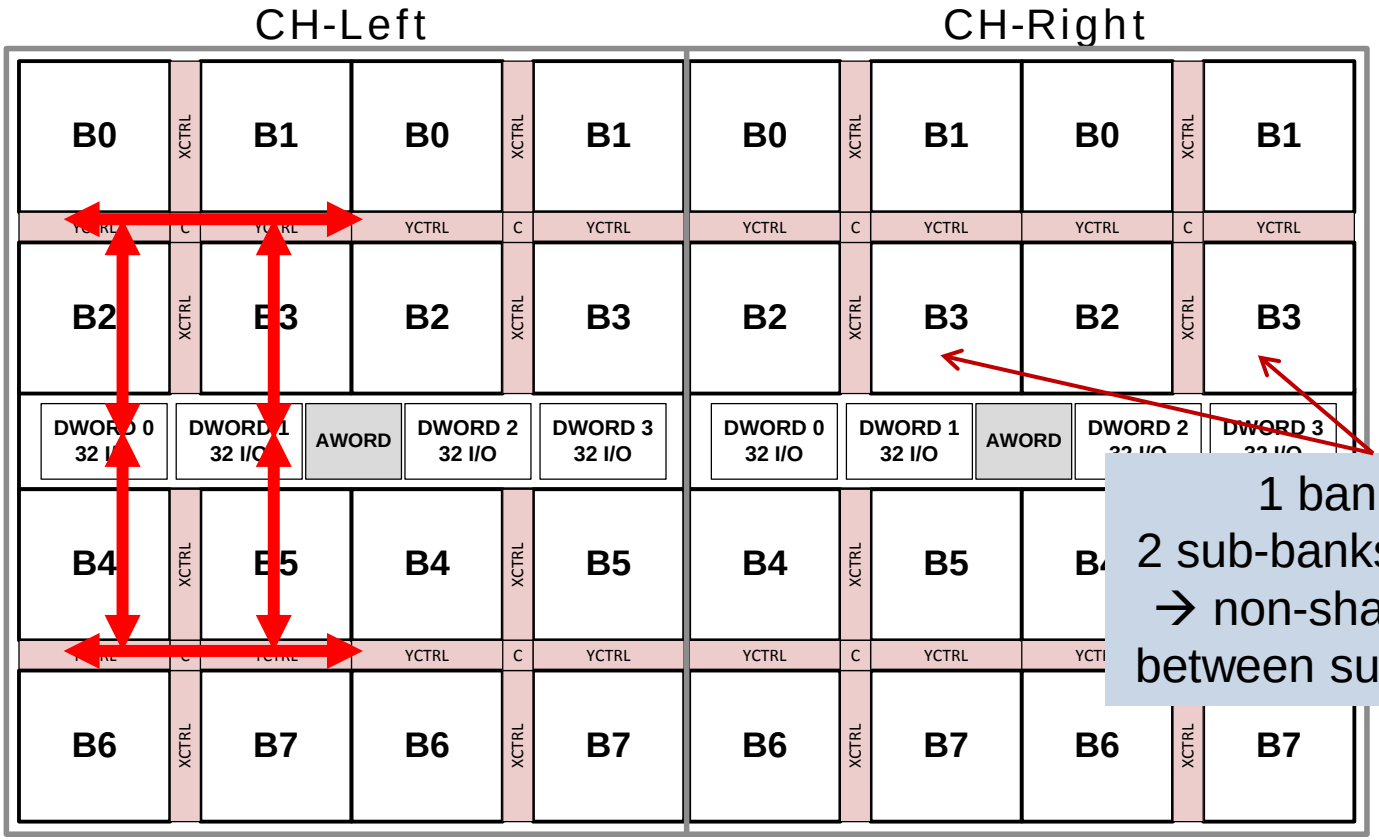
HBM Chip Micrograph (Core)

- 1Gb per channel : independent IO per channel



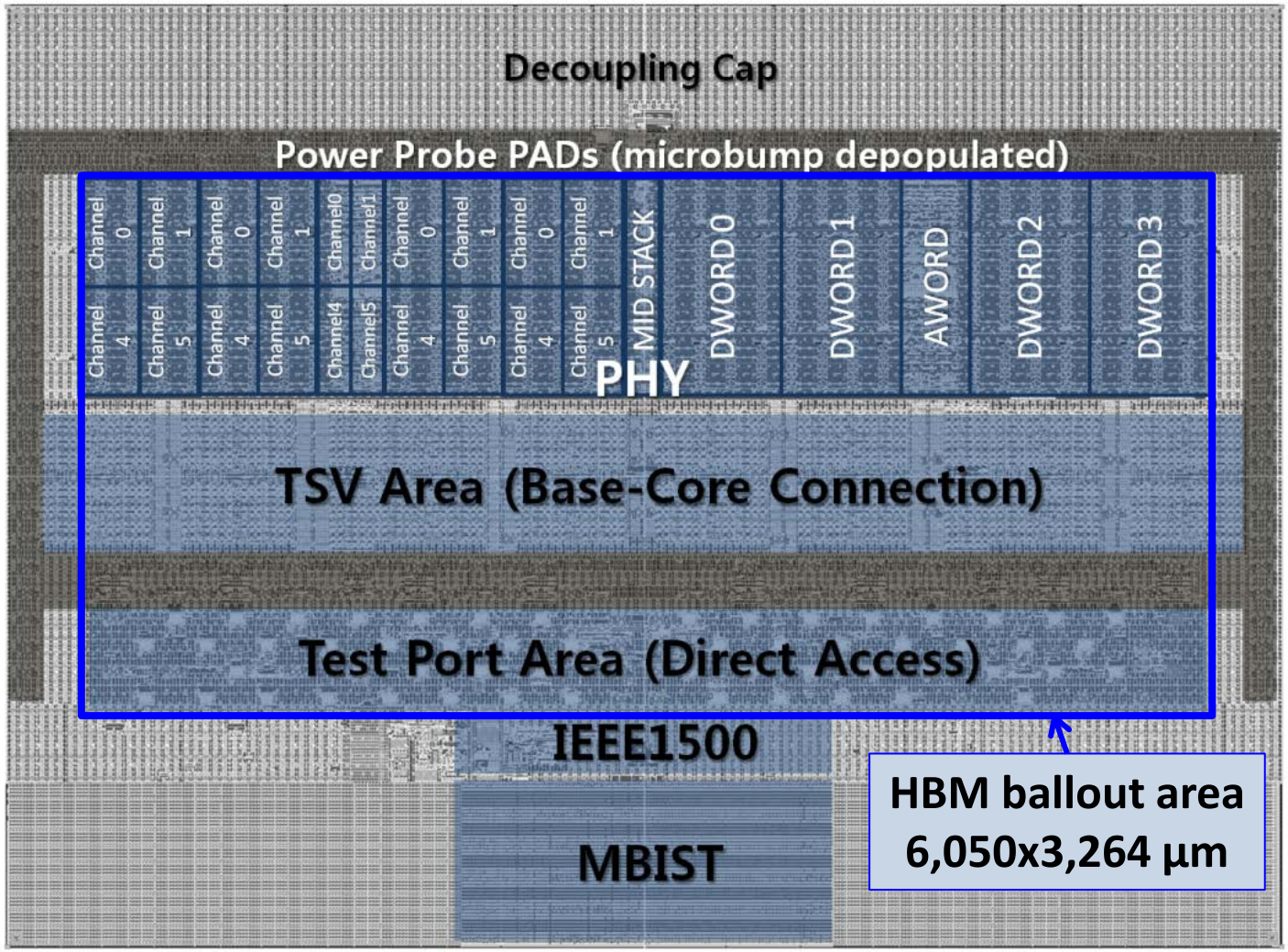
Core Architecture

- **1 slice has 2 channels, a channel consists of**
  - 8 bank(16-sub bank), 128 TSV I/O, 2*n*-prefetch
  - 256 global I/O → 32B access granularity



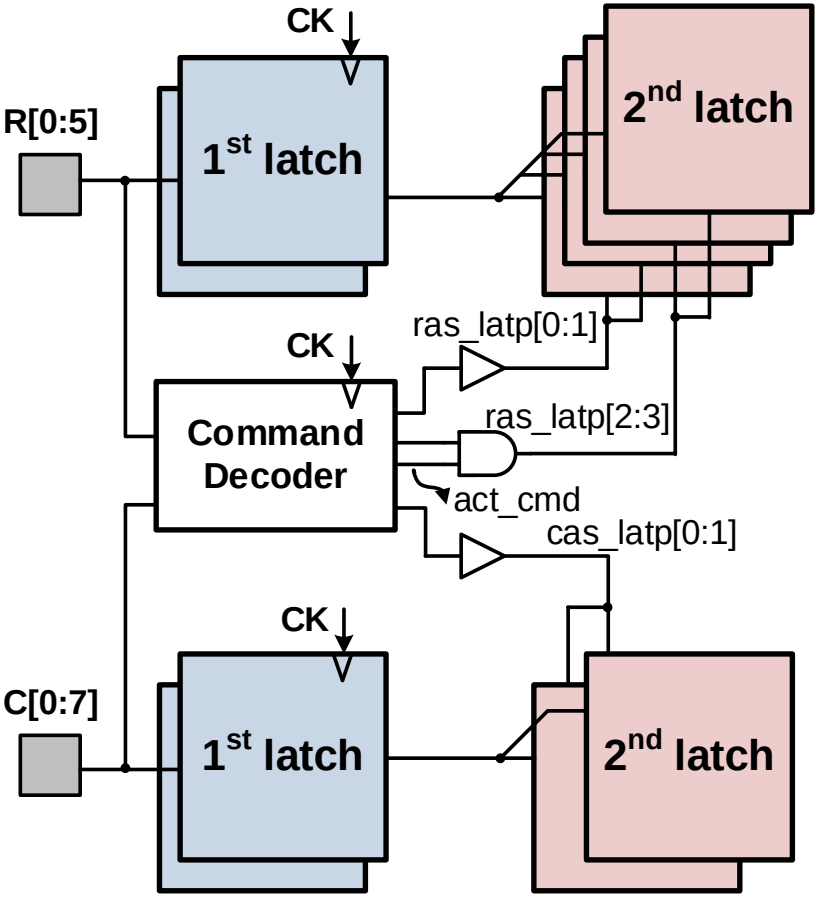
Small Form Factor (Base)

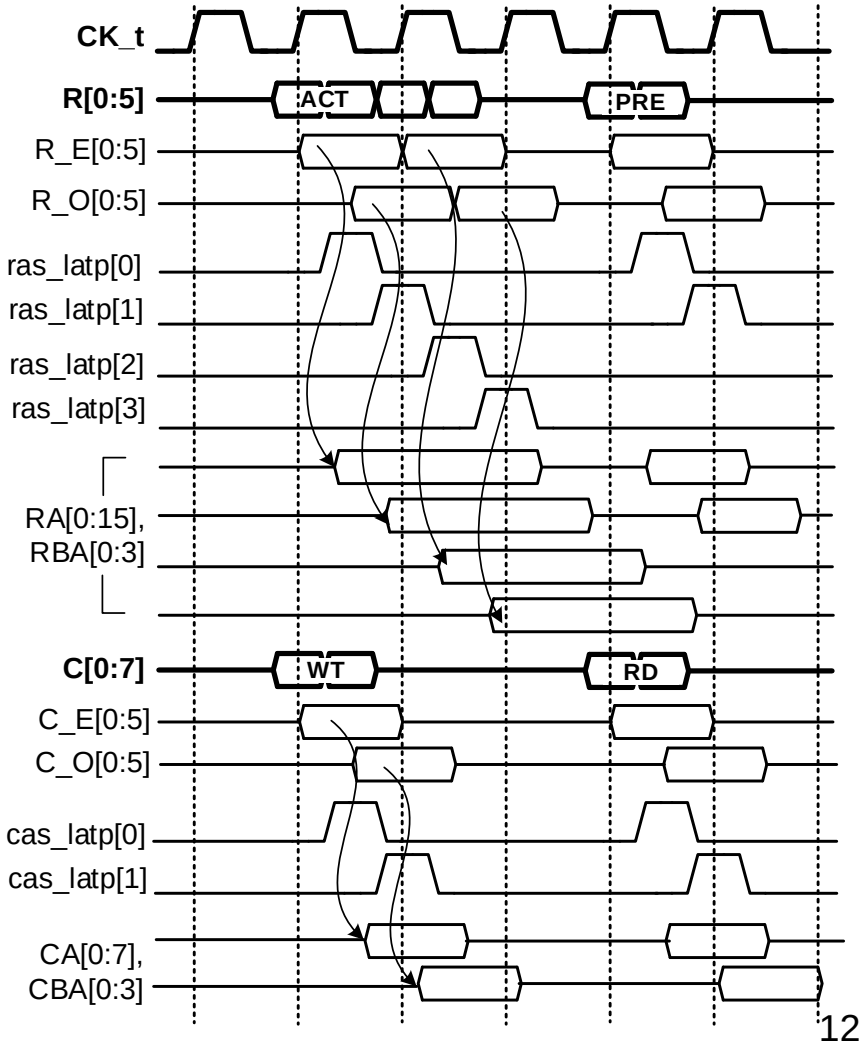
- **PHY has 1024 I/O interface(32DWORD x 32IO)**



Dual Command Interface

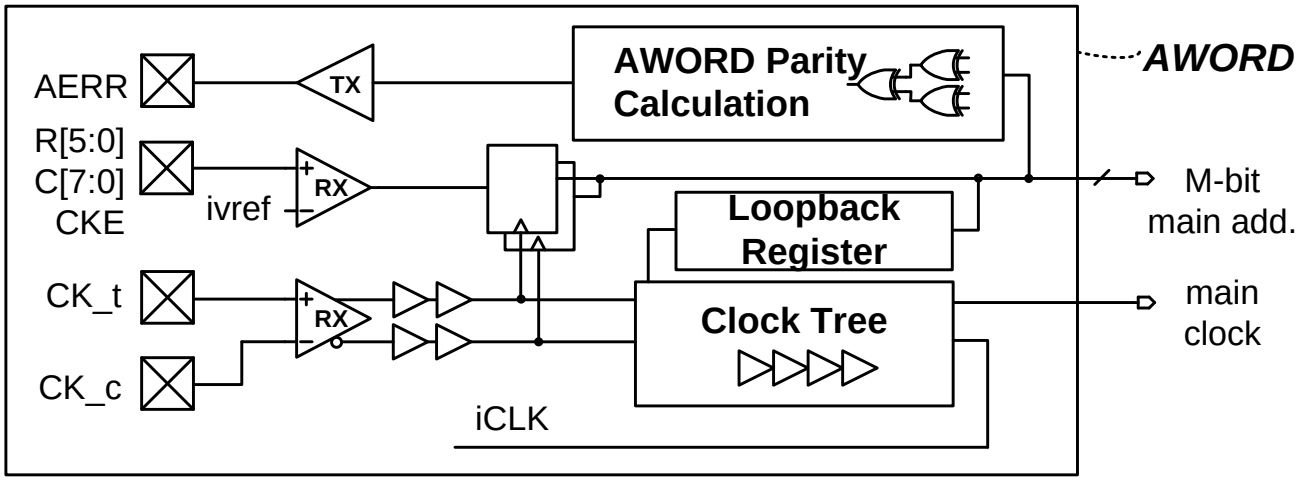
- **Row/column input through different pin**

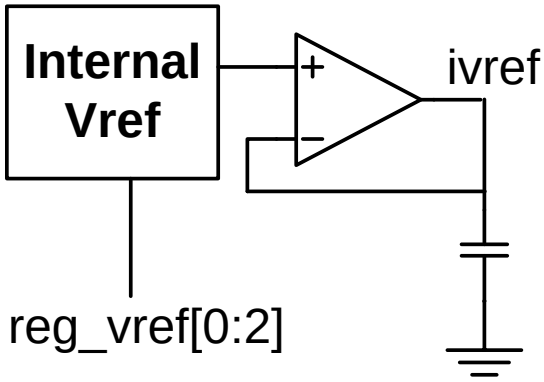




PHY Structure of HBM (AWORD)

- **Differential clock & DDR addressing**
  - AWORD(address buffer) with parity and loopback



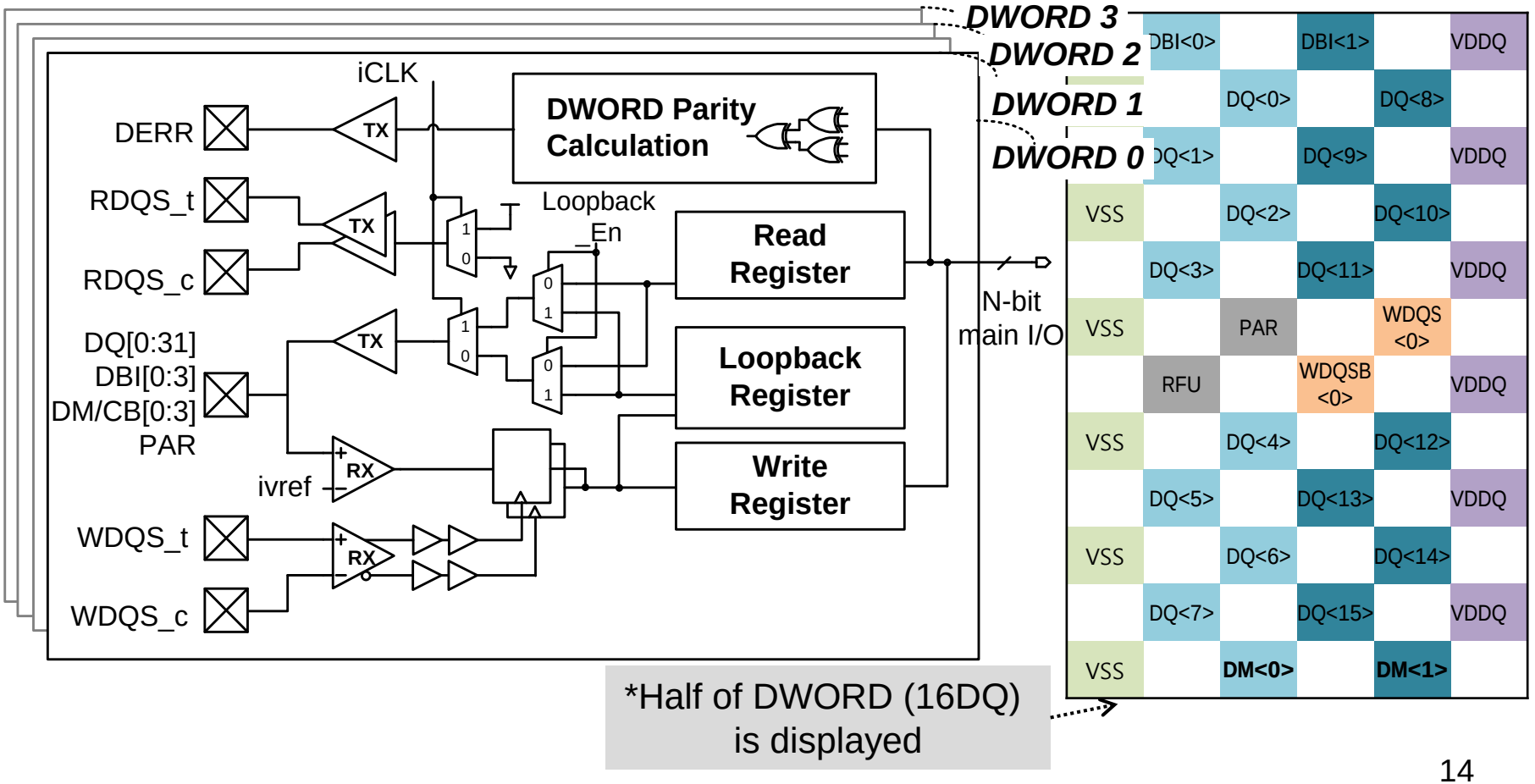


* programmable internal Vref → All RX for small VIH/VIL

	RFU		RFU		VDDQ
VSS		C<0>		R<0>	
	C<1>		R<1>		VDDQ
VSS		C<2>		R<2>	
	C<3>		R<3>		VDDQ
VSS		RFU		CK	
	CKE		CKB		VDDQ
VSS		C<4>		R<4>	
	C<5>		R<5>		VDDQ
VSS		C<6>		RA_RED	
	C<7>		RFU		VDDQ
VSS		CA_RED		RFU	

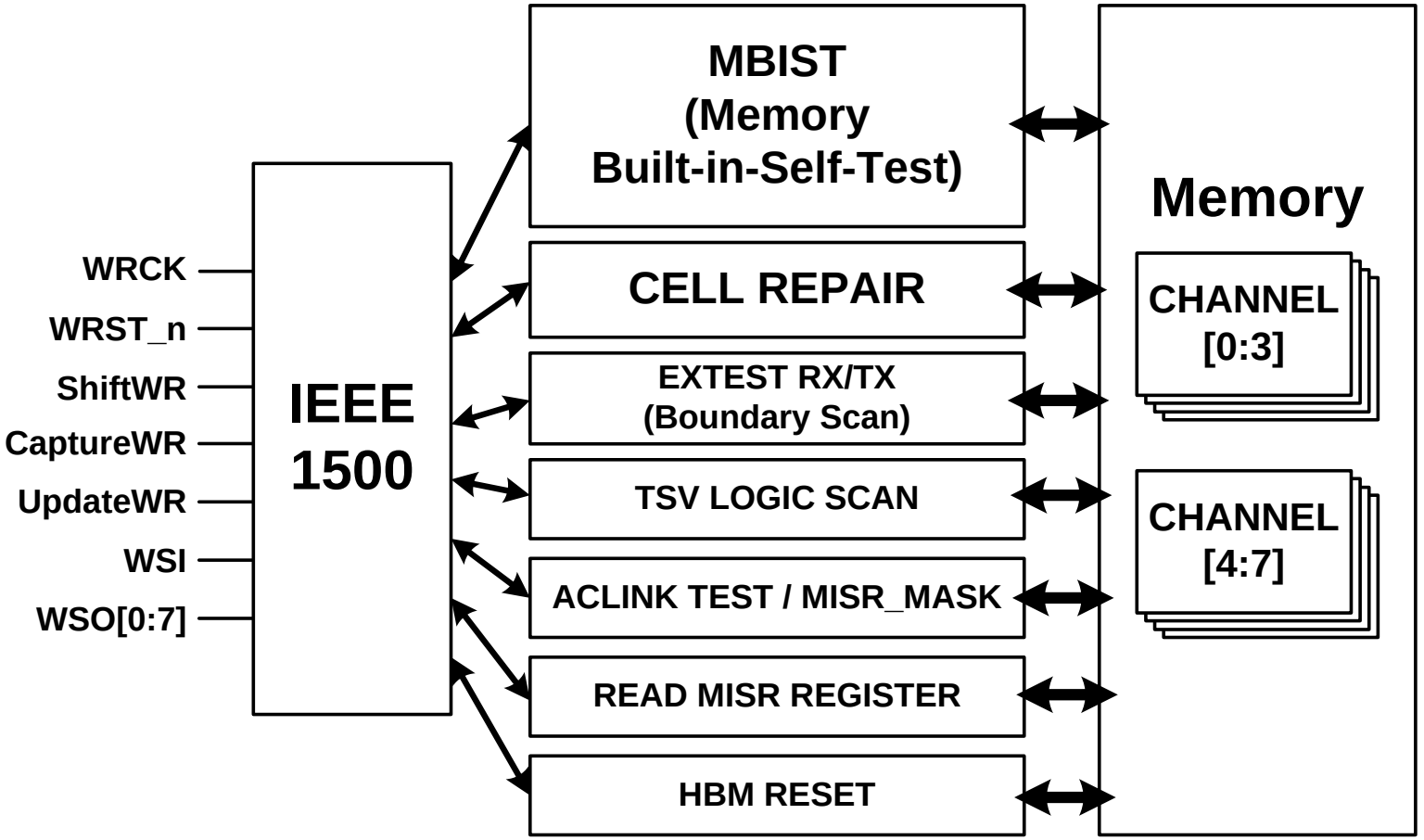
PHY Structure of HBM (DWORD)

- **Unidirectional differential W/R strobes / 32DQ**
  - 4 DWORD(data buffer)/channel (with parity & loopback)



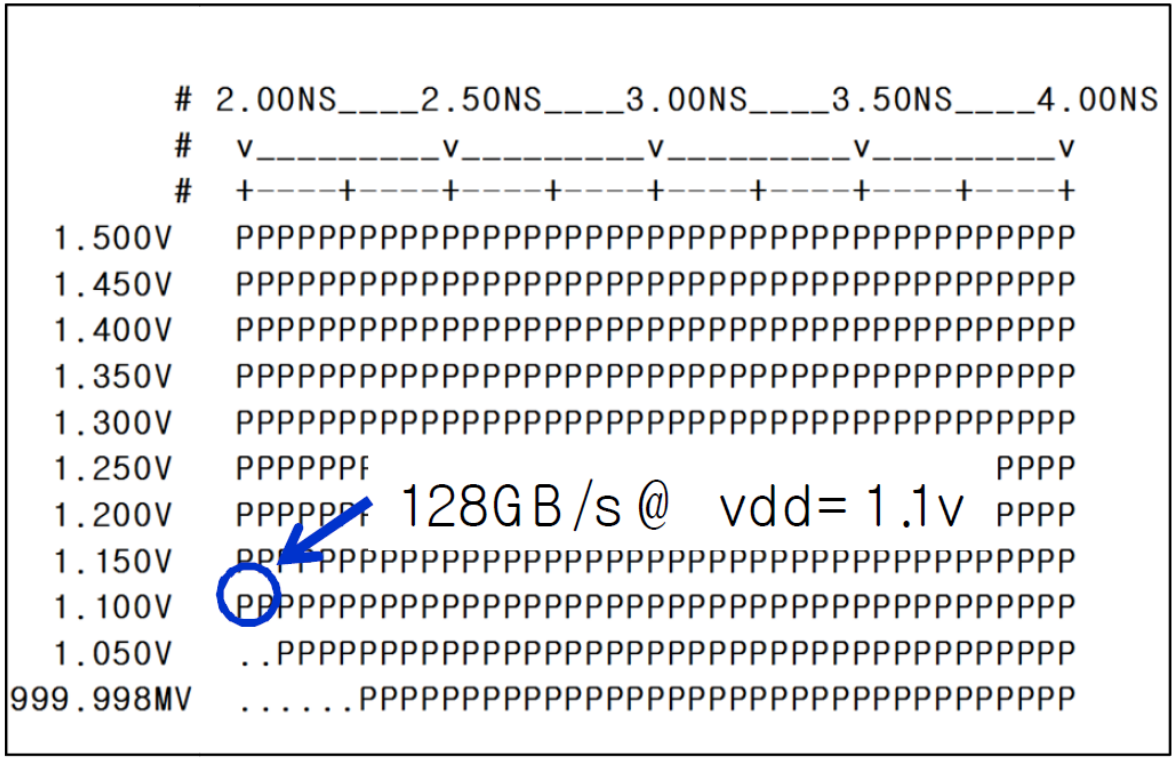
Serial Port Access Function

- **HBM test feature includes test access port through IEEE1500**



Performance (Shmoo Plot)

- **8 channel gapless read, chip-on-wafer test**
  - All channel seamless read operation(tCCD=2ns)

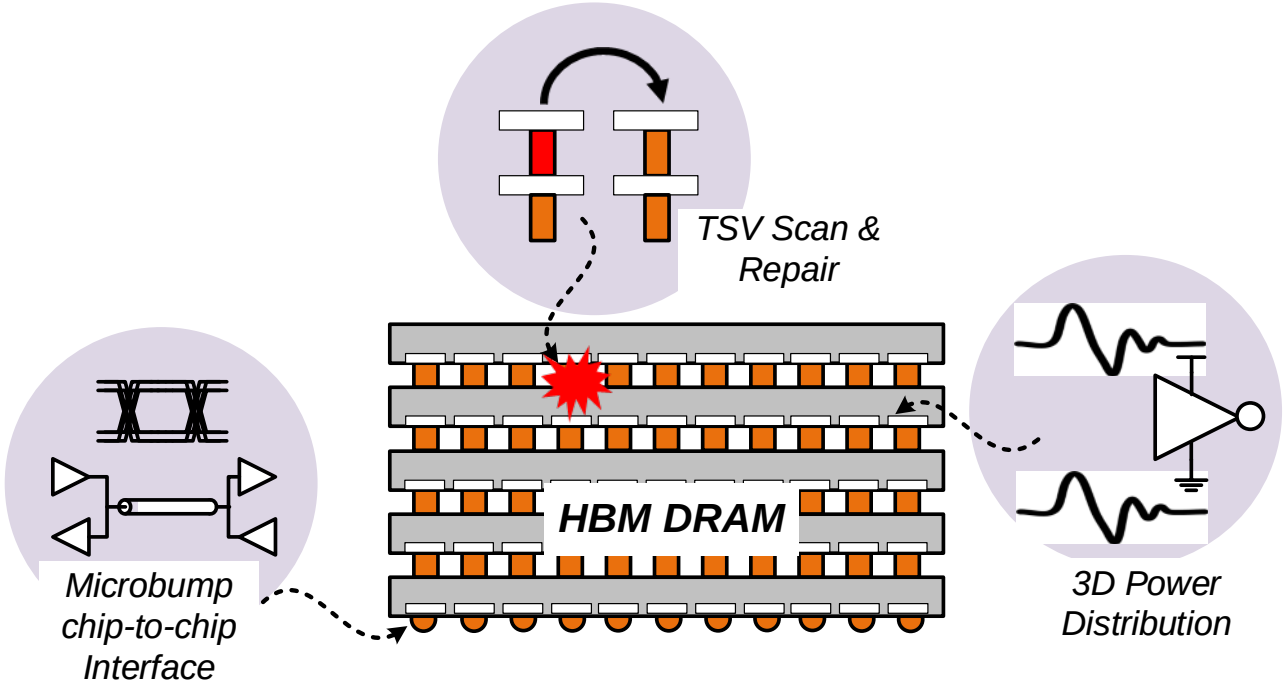


Process	**29nm DRAM process**
Chip Size	**5.10mm x 6.91mm**
Organization	**8 bank x 8 channel x 128 I/O (total 1024 I/O)**
Density	**1Gb / channel**
Microbump pitch (base die)	**48 µm x 55 µm**
Supply Voltage	**VDD=1.2v, VPP=2.5v**
Refresh	**8k / 32ms**
Page Size	**2KB**
Data Rate	**1.0 Gbps (128GB/s)**
C_{IO}	**0.4pF**

Design Considerations of HBM

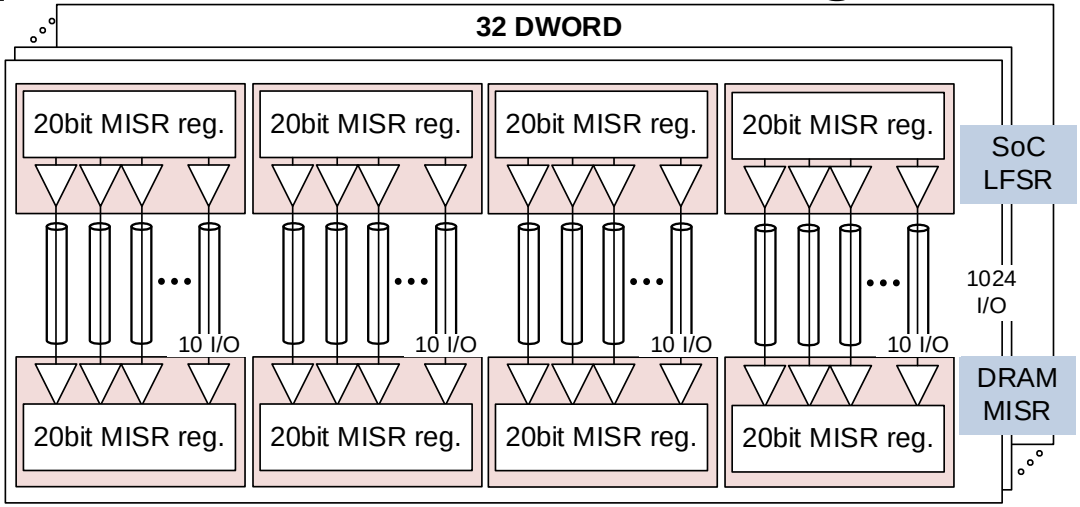
- Microbump chip-to-chip interface
- TSV scan and repair
- Power distribution analysis

→ Major environmental change in 3D design



Loopback Function

- **Loopback for I/O link testing and training**

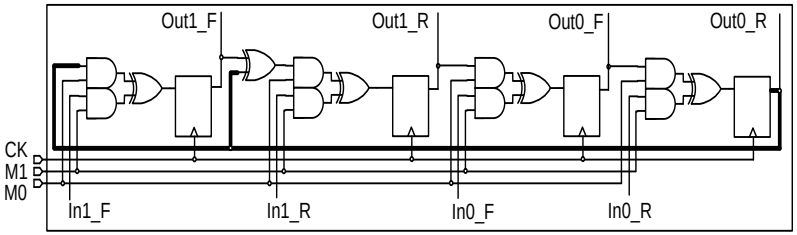


- **DWORD polynomial: $f(x)=x^{20}+x^{17}+1$(for 10 I/O)**

DWORD 20BIT LFSR Sequence

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
# of Cal.	DM	DMF	DQ[0]	DQ[0]F	DQ[1]	DQ[1]F	DQ[2]	DQ[2]F	DQ[3]	DQ[3]F	DQ[4]	DQ[4]F	DQ[5]	DQ[5]F	DQ[6]	DQ[6]F	DQ[7]	DQ[7]F	DBI	DBIF
Initial	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
2	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	1
3	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	1	0
4	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	1	1	1	0	1
5	1	0	1	0	1	0	1	0	1	0	1	0	1	1	1	1	1	0	1	0
6	0	1	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	1	0	1
7	1	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	1	0	1	0
8	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	0	1
9	1	0	1	0	1	0	1	0	1	1										
10	0	1	0	1	0	1	0	1	1	1										
11	1	0	1	0	1	0	1	1	1	1										
12	0	1	0	1	0	1	1	1	1	1										
13	1	0	1	0	1	1	1	1	1	1										
14	0	1	0	1	1	1	1	1	1	1										
15	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0

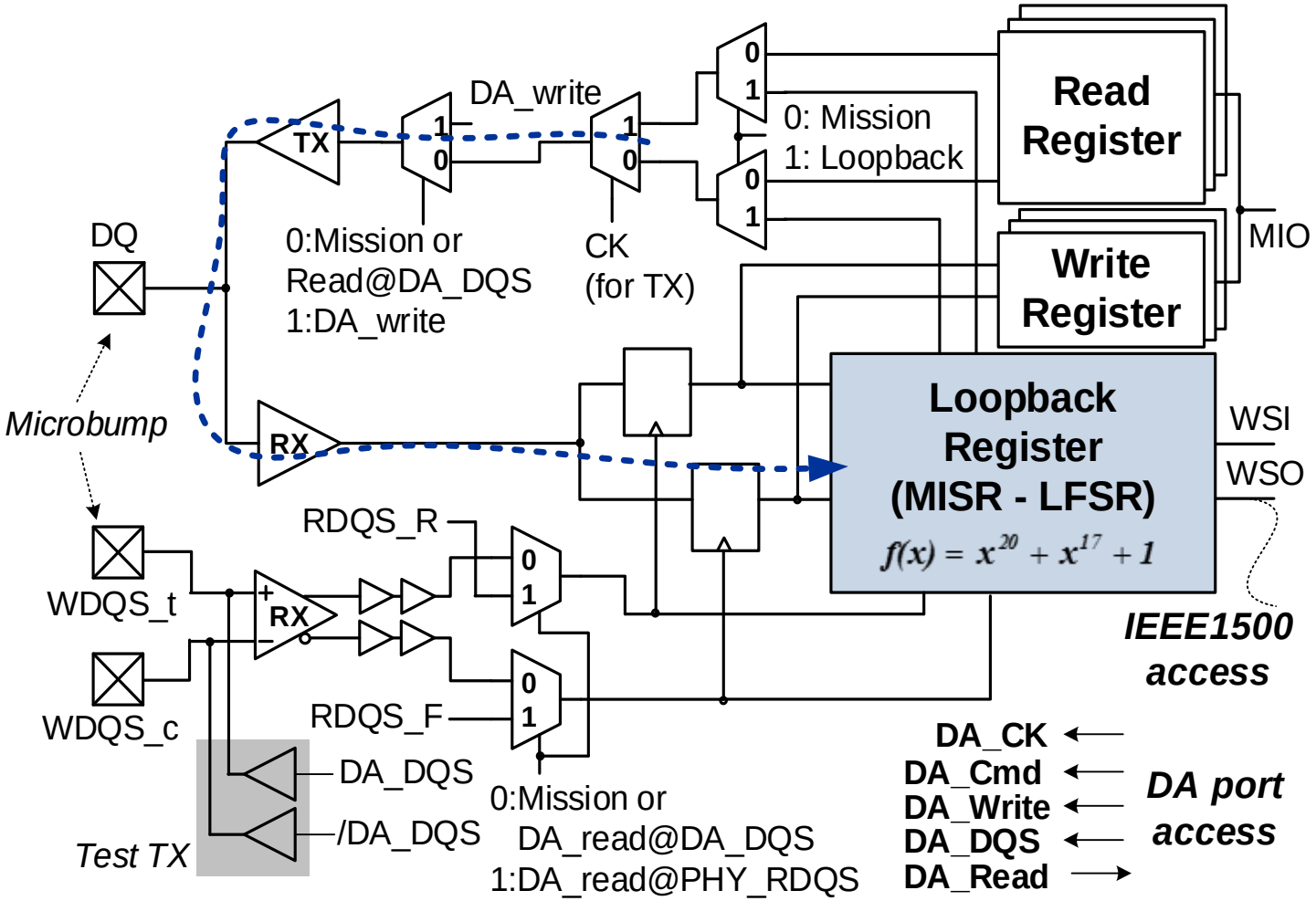
LFSR calculation results :initial all AAAAAh



4bit MISR Register

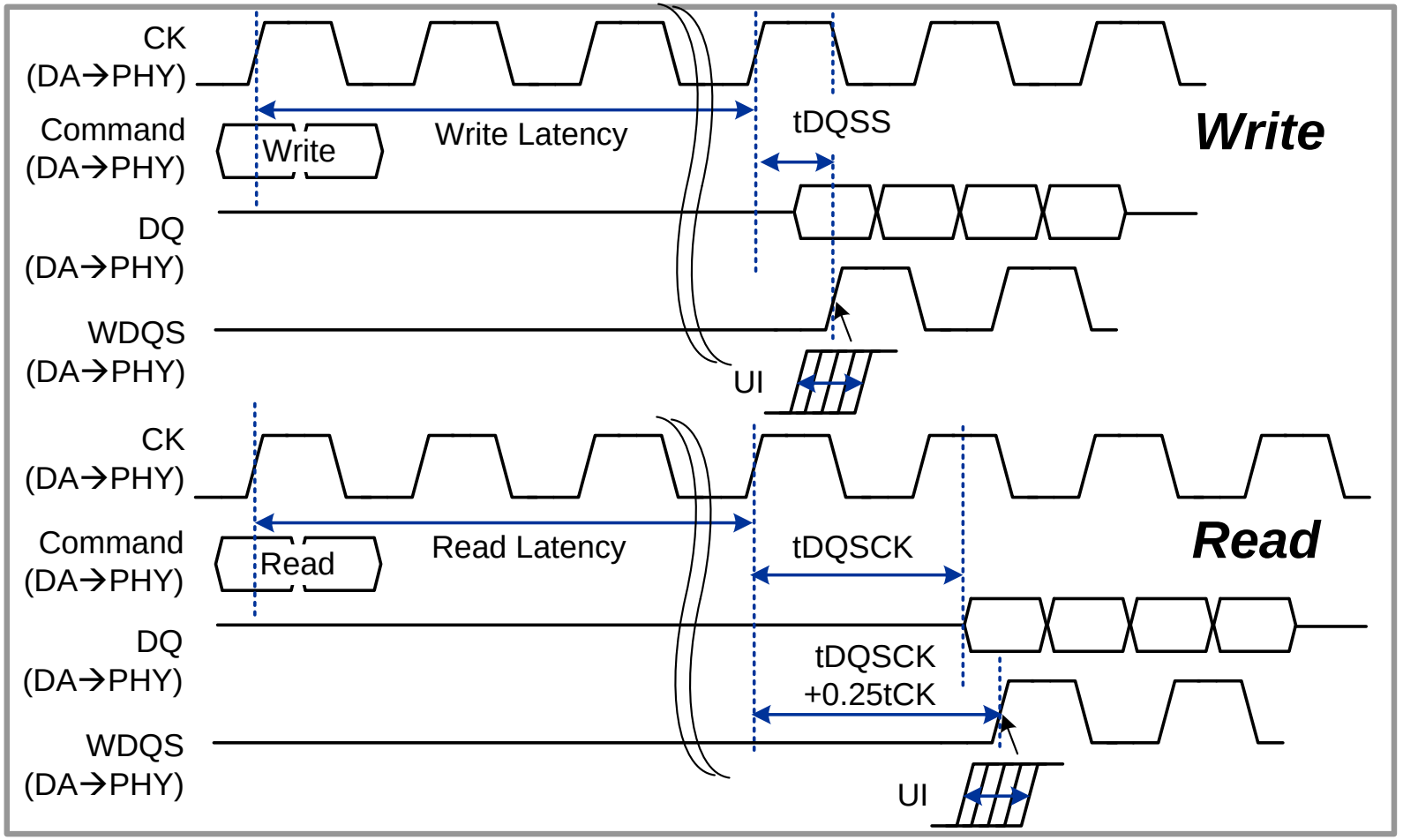
Microbump I/O AC Measure

- **KGS test before assembly**



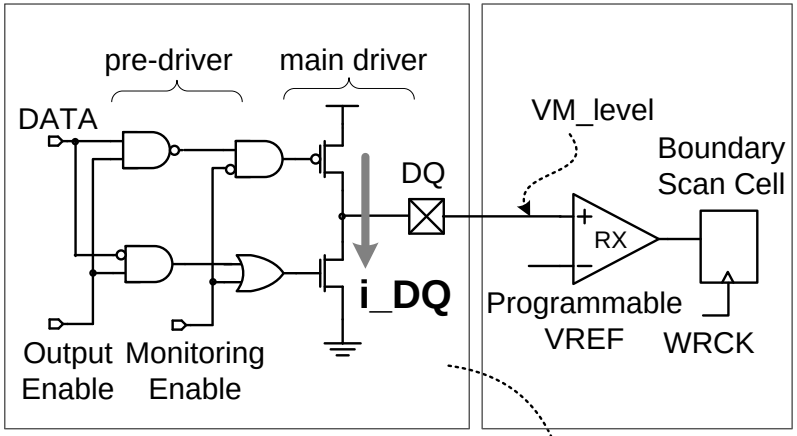
Microbump I/O AC Measure

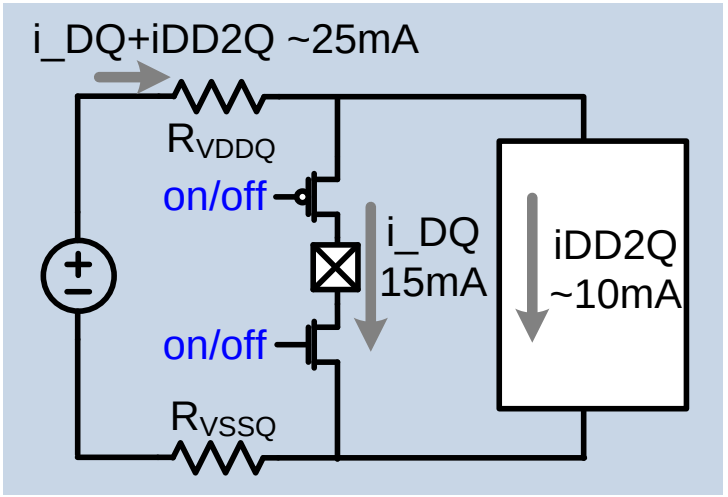
- **Write / read data is captured into loopback registers during KGS test**

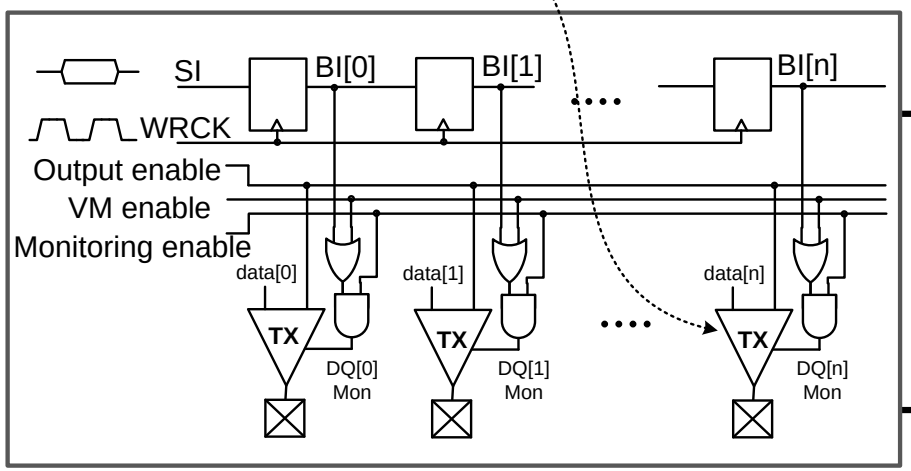


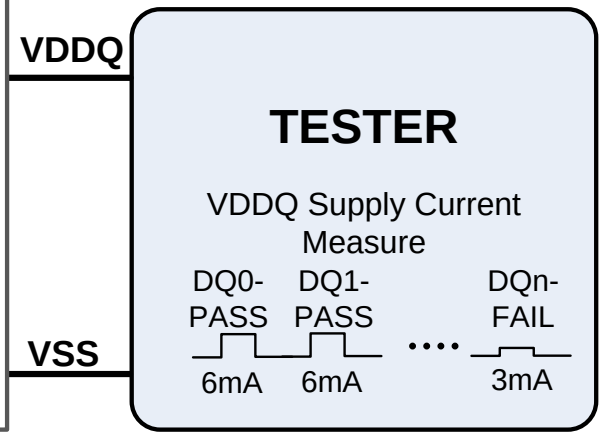
Microbump I/O DC Measure

- **Impedance calculation by i_DQ current measure without touching PAD**



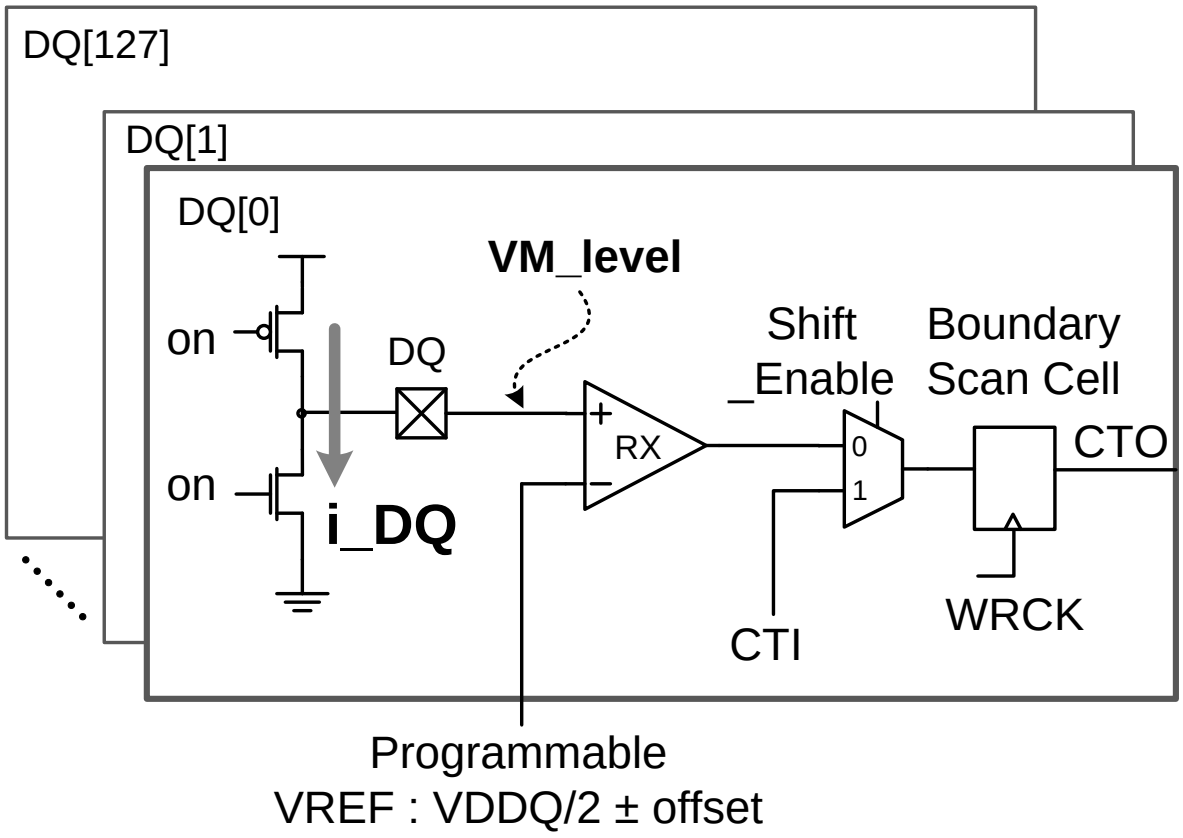






Microbump I/O DC Measure

- **Impedance mismatch by VM_level measuring**
  - Method of detecting pullup/pulldown impedance mismatch



VDDQ/2 + offset → Vref

Measure (RX → register)

Shift out (128DQ : all 0)

VDDQ/2 - offset → Vref

Measure (RX → register)

Shift out (128DQ : all 1)

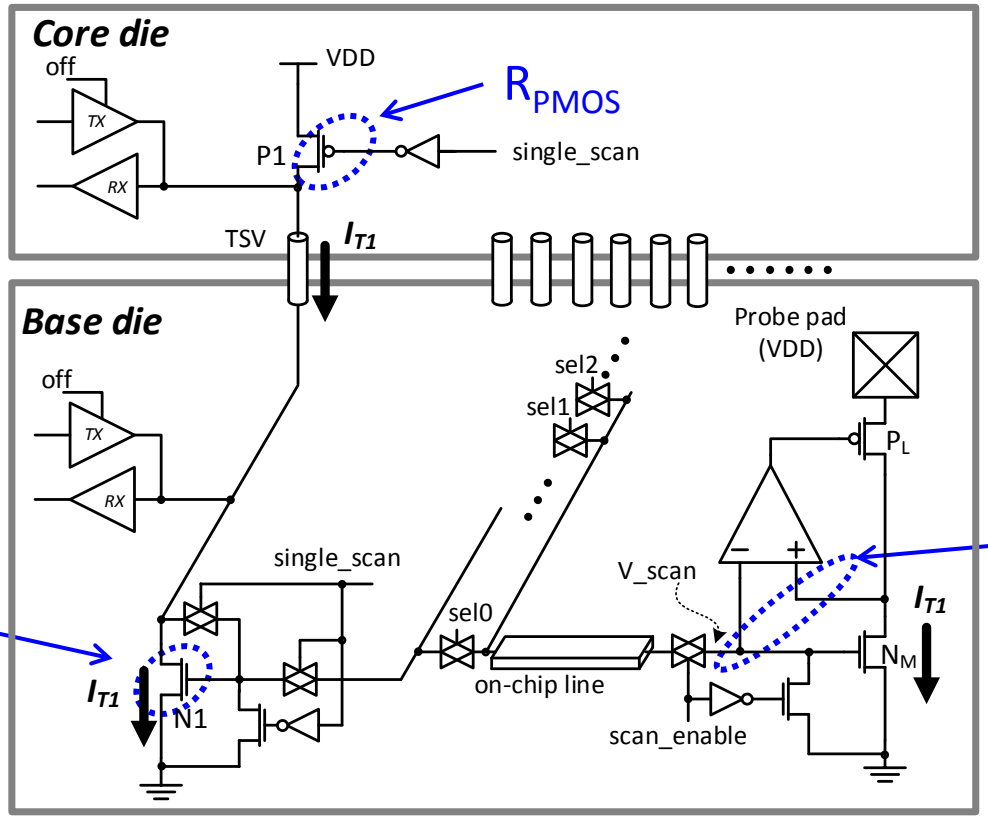
VM(128DQ) :
within VDDQ/2 ±offset

Chip Loss by TSV Failure

- **Yield of single TSV is : Y_{single} (%)**
- **N:1 TSV Repair Scheme :**
  - Repairable Case :
    [Every (N+1) TSV alive] + [N alive with 1 failed TSV]
  - Repairable Group Yield
    $= (Y_{single})^{N+1} + (Y_{single})^{N} ({}_{N+1}C_1)(1- (Y_{single}))$
  - Die yield
    If we assume that total number of groups is m in a die
    $Y_{die} = (Y_{group})^{m}$
- **The number of TSV groups is larger than normal DRAM → Relatively low stack yield**
  - 8 I/O vs 1024 I/O

Current Scan Method (Single Scan)

- **By using current mirror,**
  - On-chip line resistance to test port can be ignored



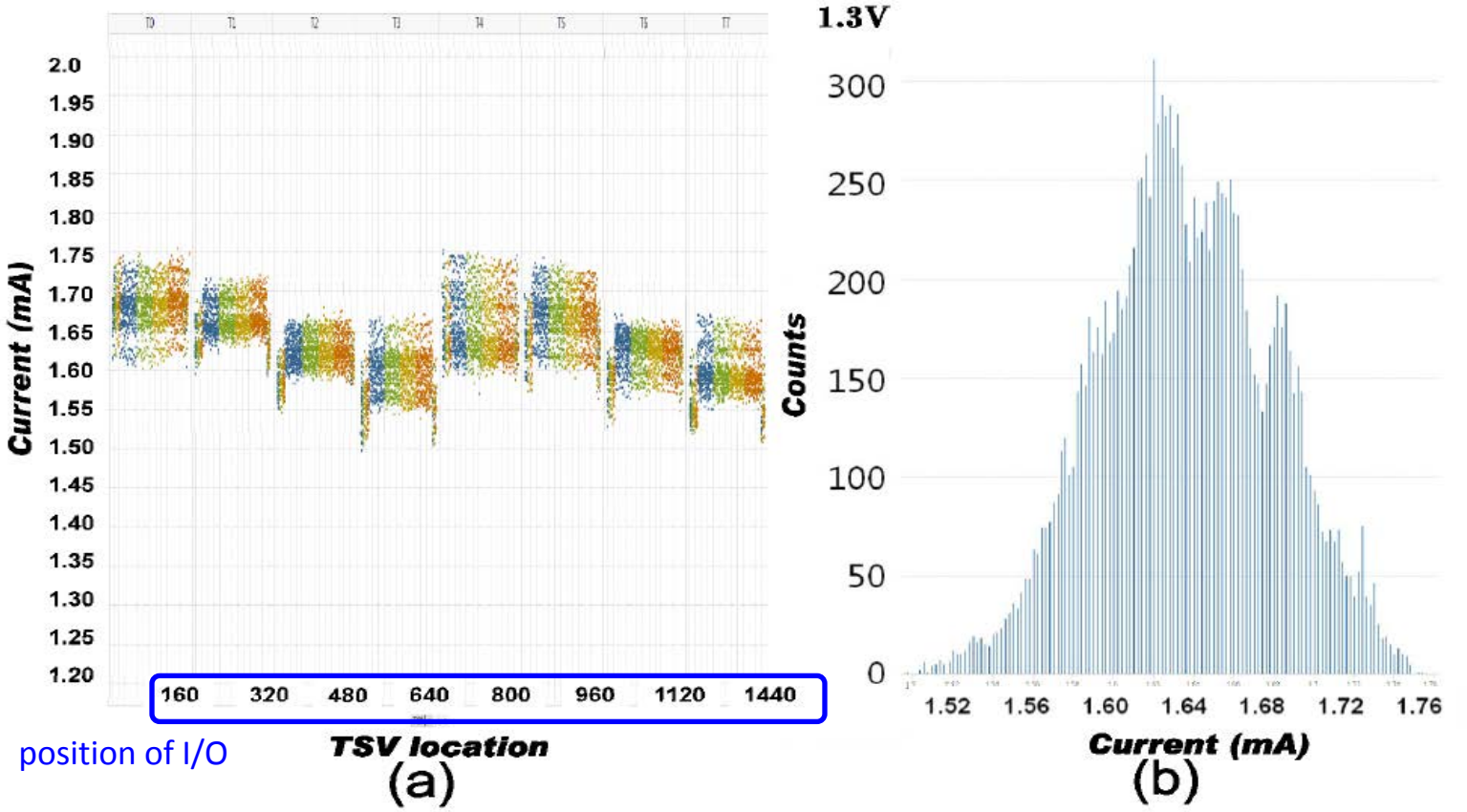
* Scan PMOS can be merged with TX

Virtual short : for minimize V_{DS} difference

Formula : $I_{T1}R_{PMOS} + I_{T1}R_{TSV} + I_{T1}R_{NMOS} = V_{DD}$

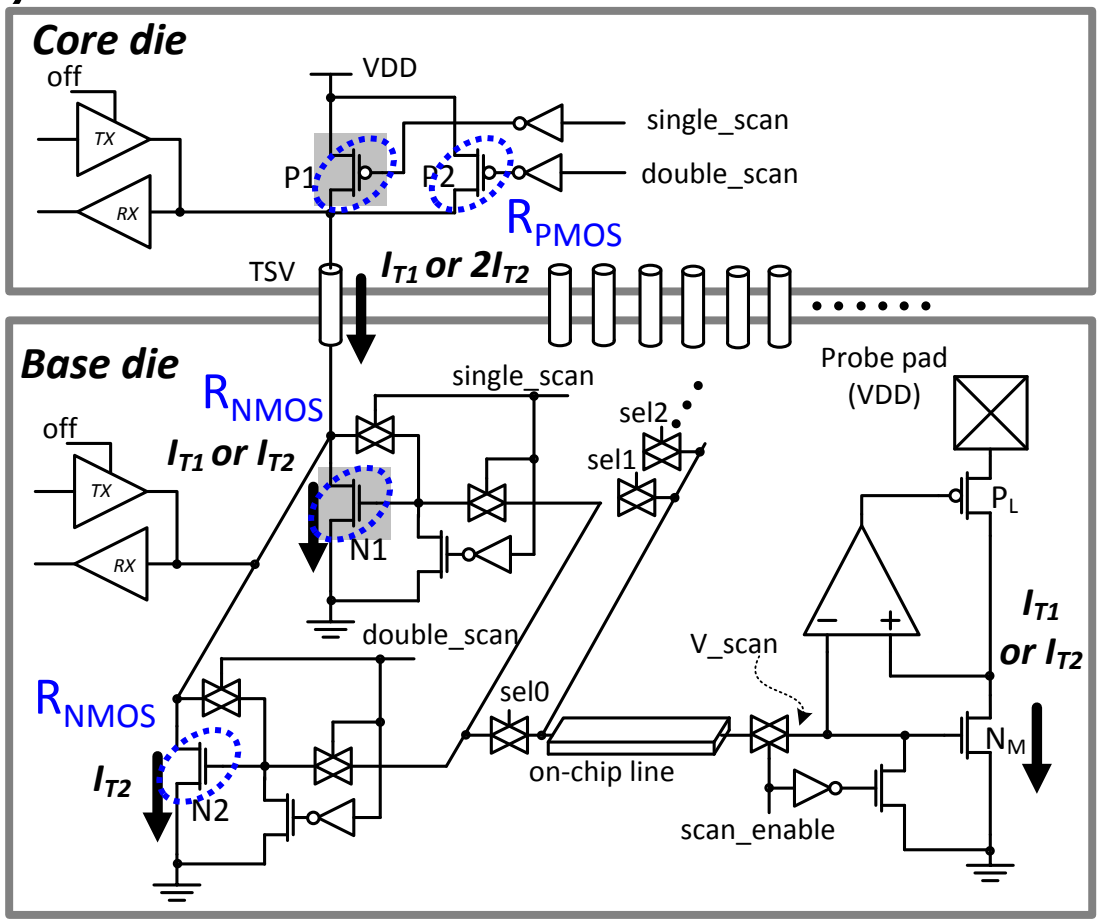
Current Scan Measurement

- **Single current scan**
  - For measuring relative drivability of each TSV driver
  - Depending on NMOS clamp resistance



Current Scan Method (Double Scan)

- **PMOS,NMOS x2 : double current via TSV**



Formula : $$I_{T2}R_{PMOS} + 2I_{T2}R_{TSV} + I_{T2}R_{NMOS} = V_{DD}$$

Formula for TSV Resistance

- **Single scan + Double scan**
  - Correlated double sampling
- **Extracting TSV resistance**
  - TSV resistance is defined by the difference of two measurements $(R_{TSV} << R_{PMOS}, R_{NMOS})$
  - Ignore R_{PMOS}, R_{NMOS} change by V_{DS}

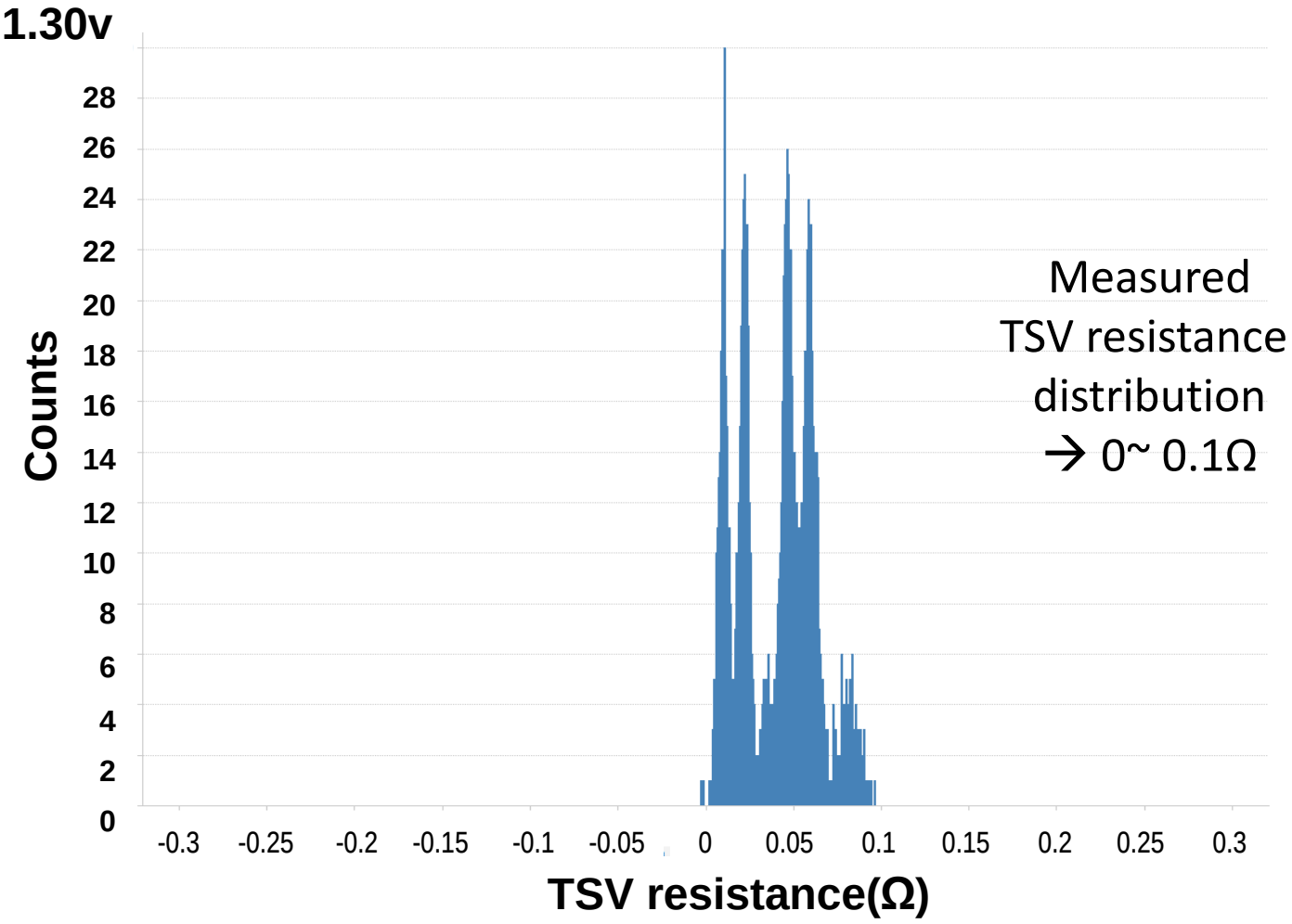
$$I_{T1}R_{PMOS} + I_{T1}R_{TSV} + I_{T1}R_{NMOS} = V_{DD} \quad \text{(Single Scan)}$$

$$I_{T2}R_{PMOS} + 2I_{T2}R_{TSV} + I_{T2}R_{NMOS} = V_{DD} \quad \text{(Double Scan)}$$

$$R_{TSV} = \frac{I_{T1} - I_{T2}}{I_{T1}I_{T2}}V_{DD}$$

TSV Resistance Measurement

- **Resistance using proposed method**

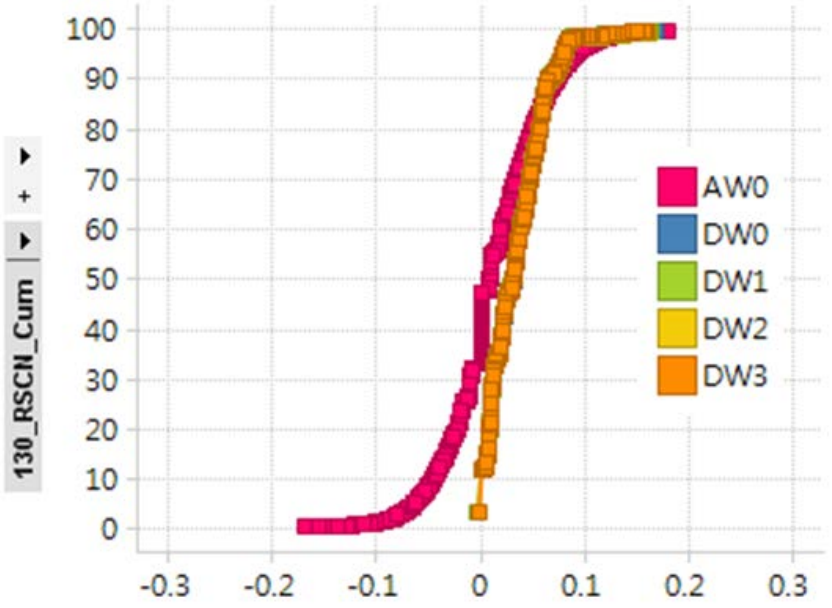


TSV Resistance Measurement

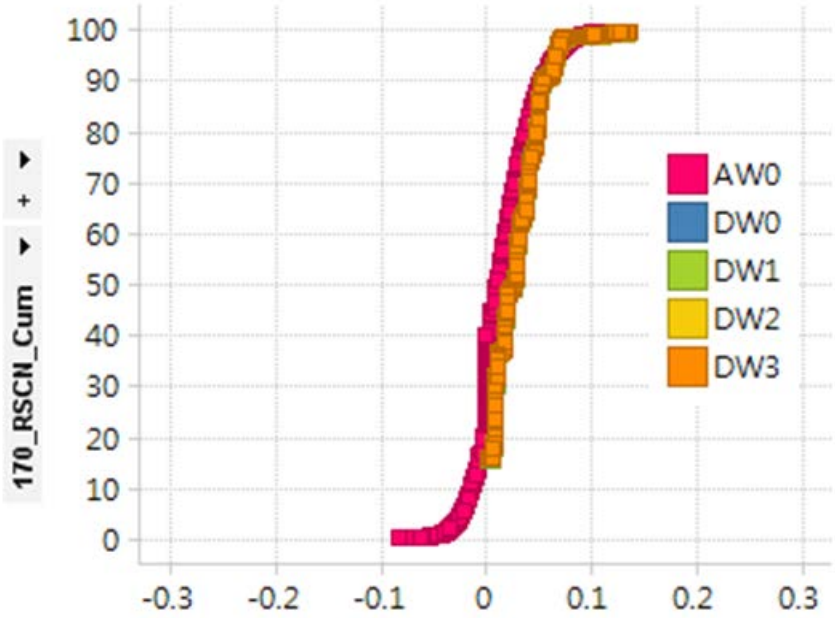
- **Measurement offset reduction**

Adding big scan driver in address TSVs
has impacts on command speed
→ AWORD has small driver, which result in larger offset

1.30V

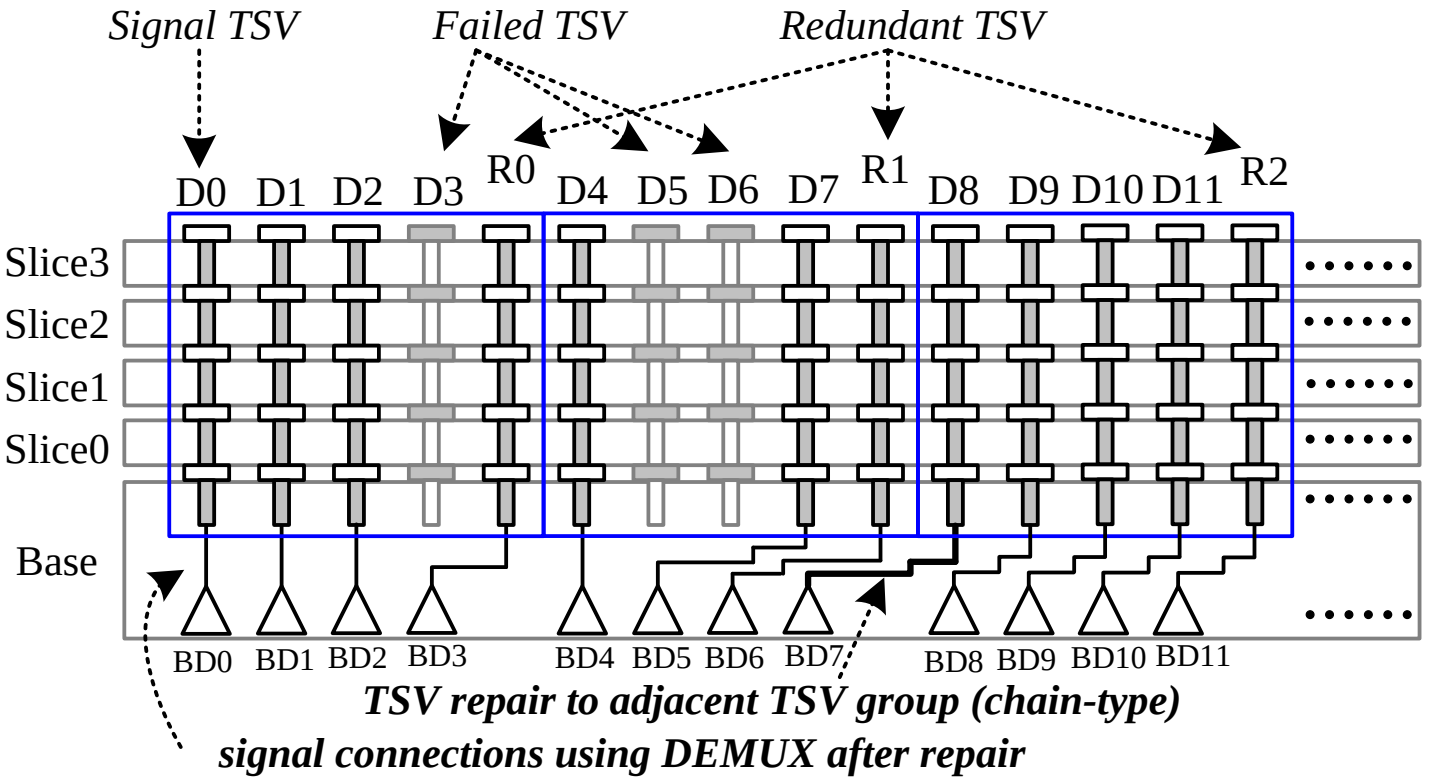


1.70V



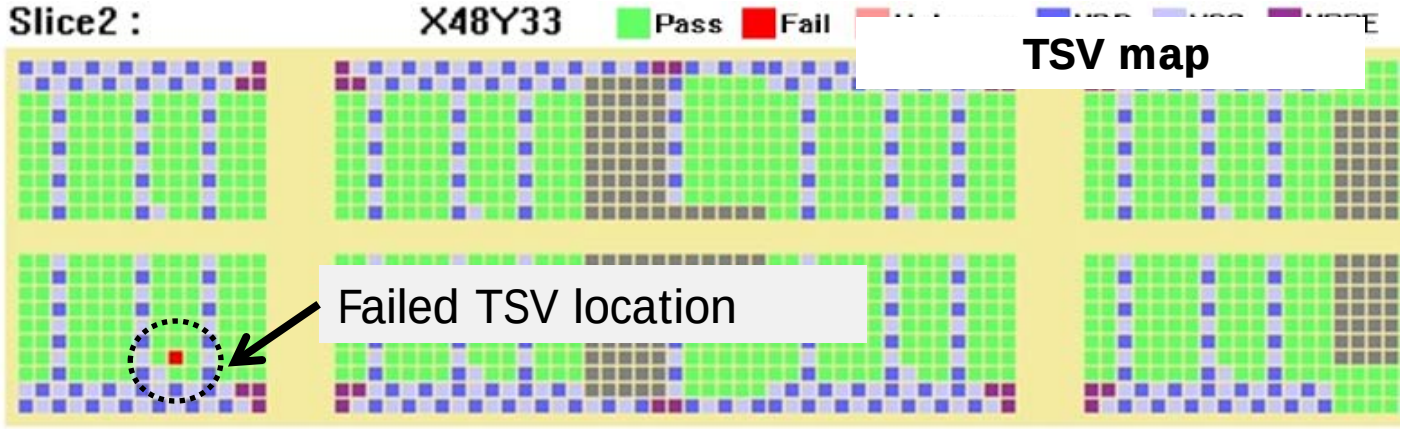
TSV Chained Redundant Array

- **Chain-redundant mapping repairs more than one TSV per group**
  - N:1 redundancy within TSV group only repair 1 TSV

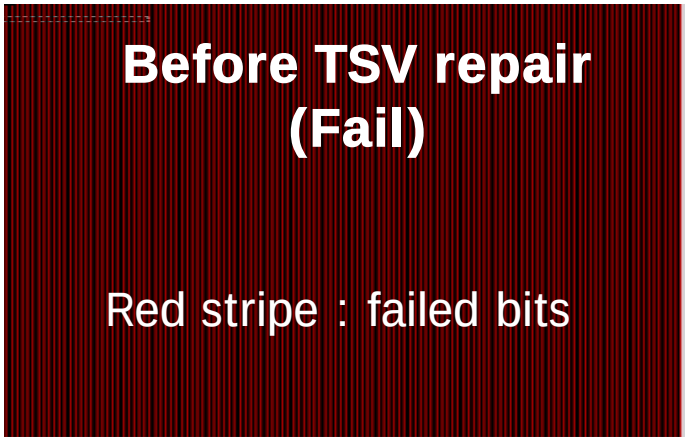


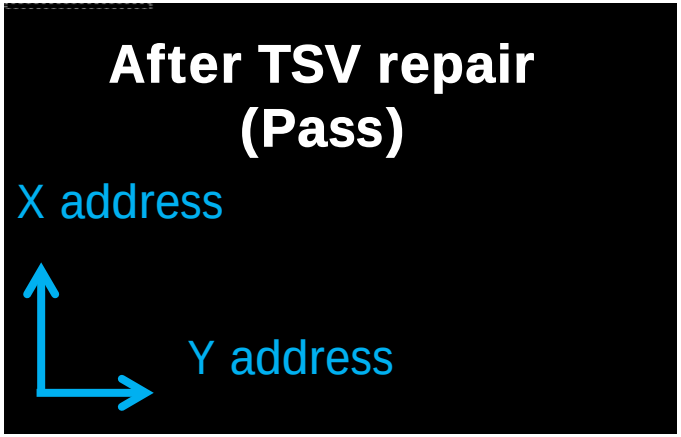
TSV Redundancy Mapping

- **Identify defected TSV by current scan**



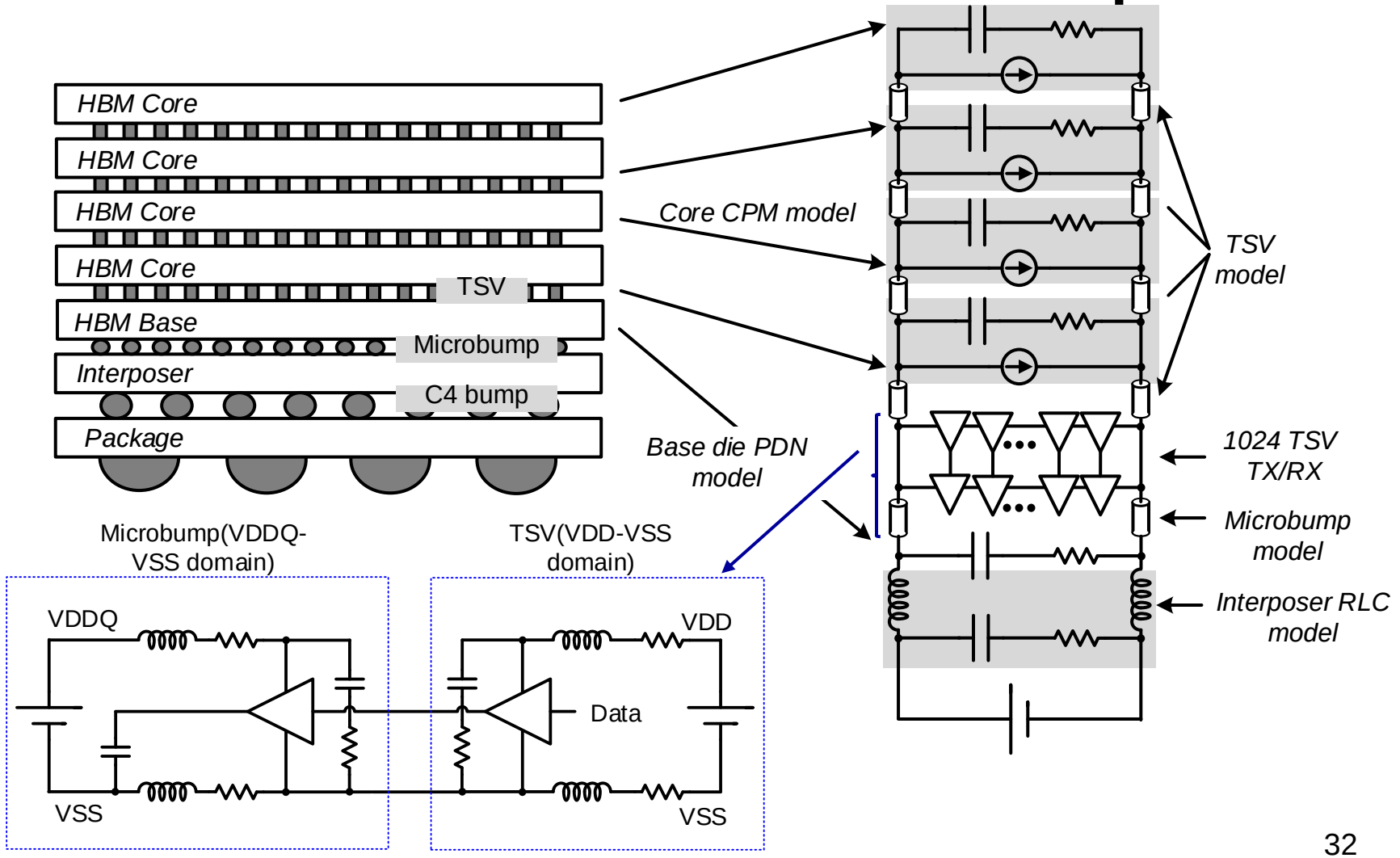
- **Function test results(bitmap) : before / after repair**





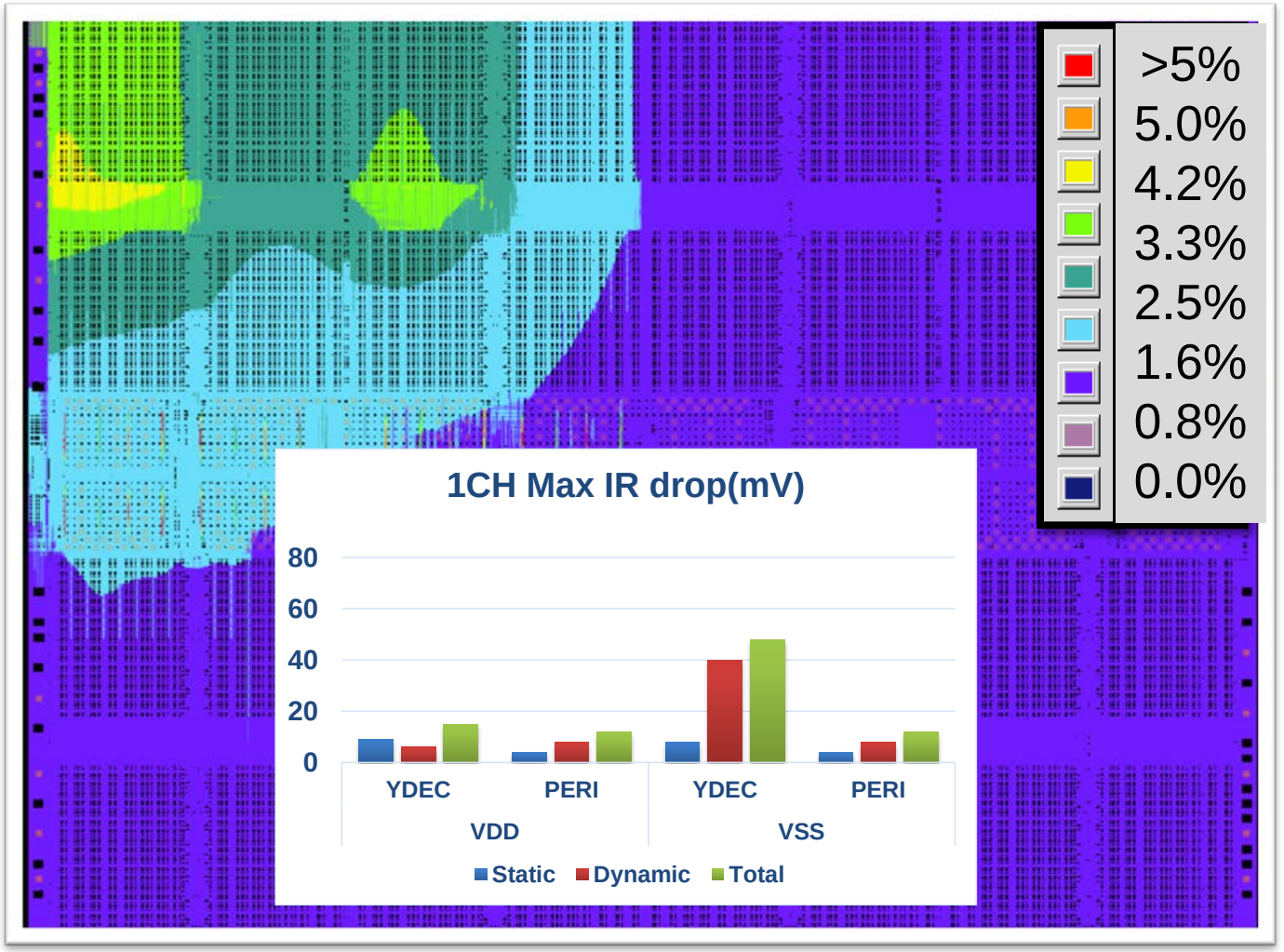
Power Distribution Network Modeling

- **TSV location is limited in stacked chip**



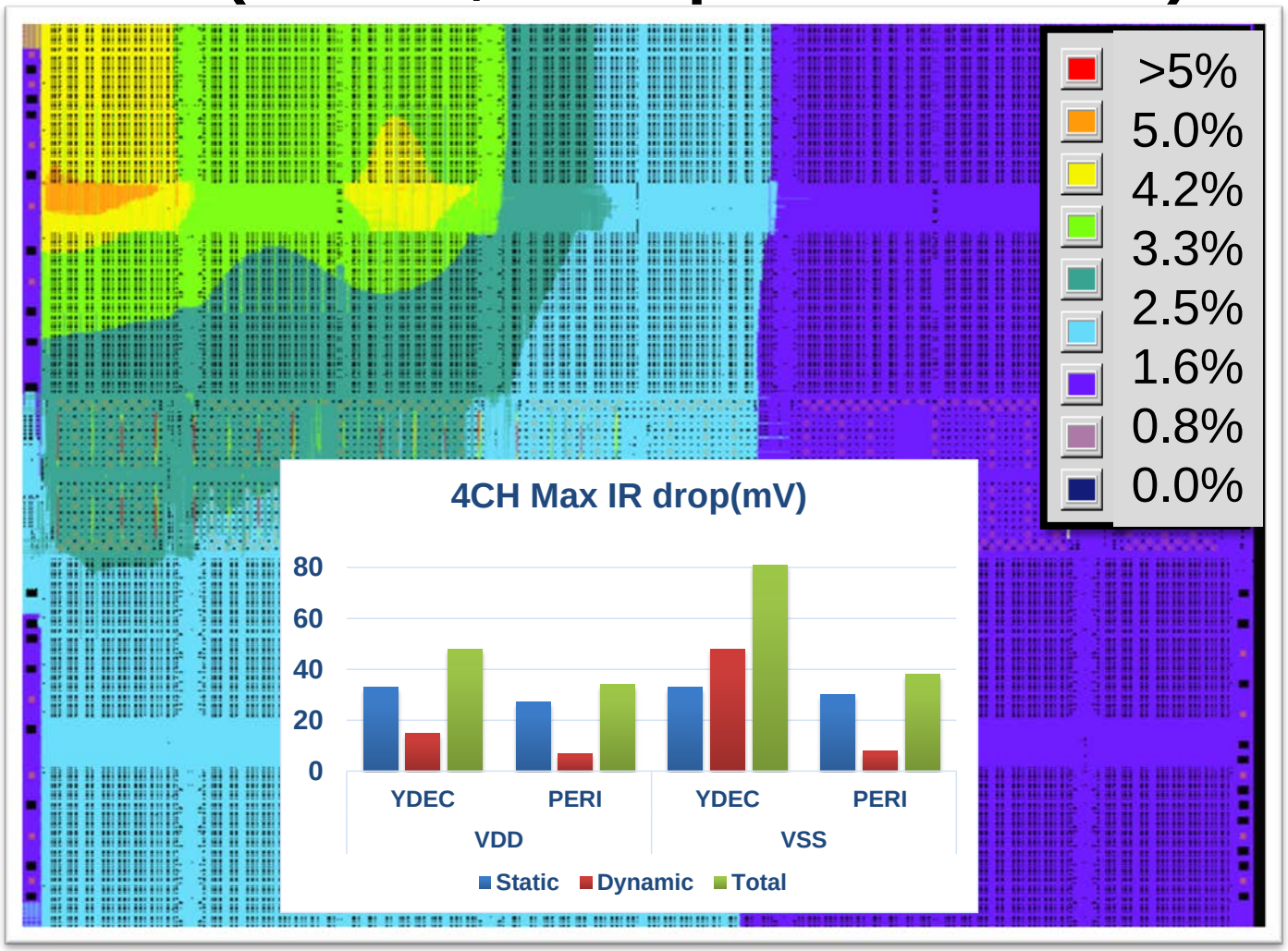
Power Distribution

- **Single slice (IDD4W, 1ch operation - VDD)**



Power Distribution

- **All slice (IDD4W, 4ch operation - VDD)**



Memory Architecture of HBM

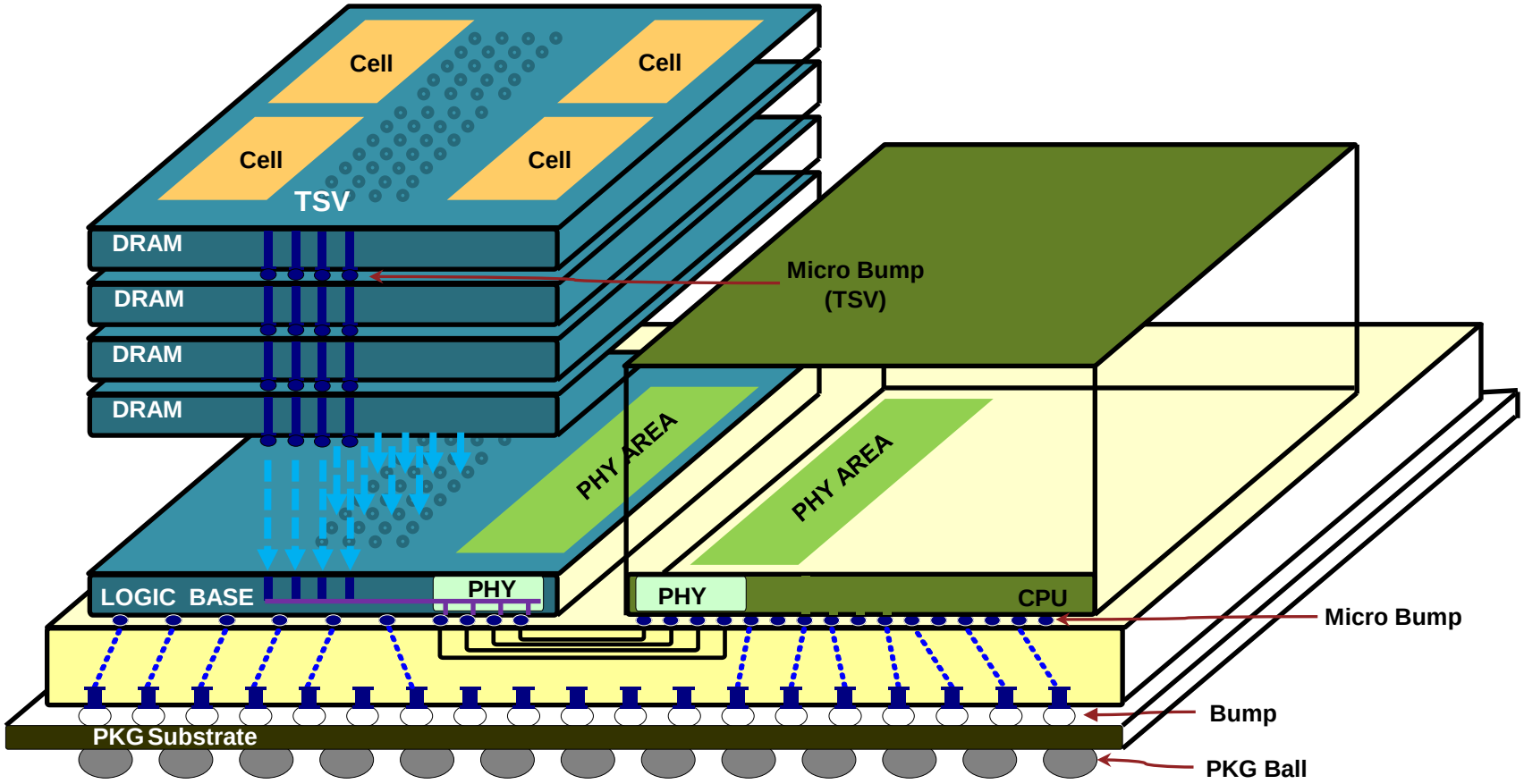
- **Memory requirements in the system**
  - Bandwidth (128GB/s → 256GB/s)
  - tFAW, tRRD reduction, page current reduction
  - More banks/channel

- **Pseudo channel architecture (2nd Gen.)**
  - 128 I/O → 2 x 64 I/O (16 → 32 banks)

- **Multi rank architecture (2nd Gen.)**
  - Twice the number of banks (32 → 64 banks)
  - High density, same bandwidth

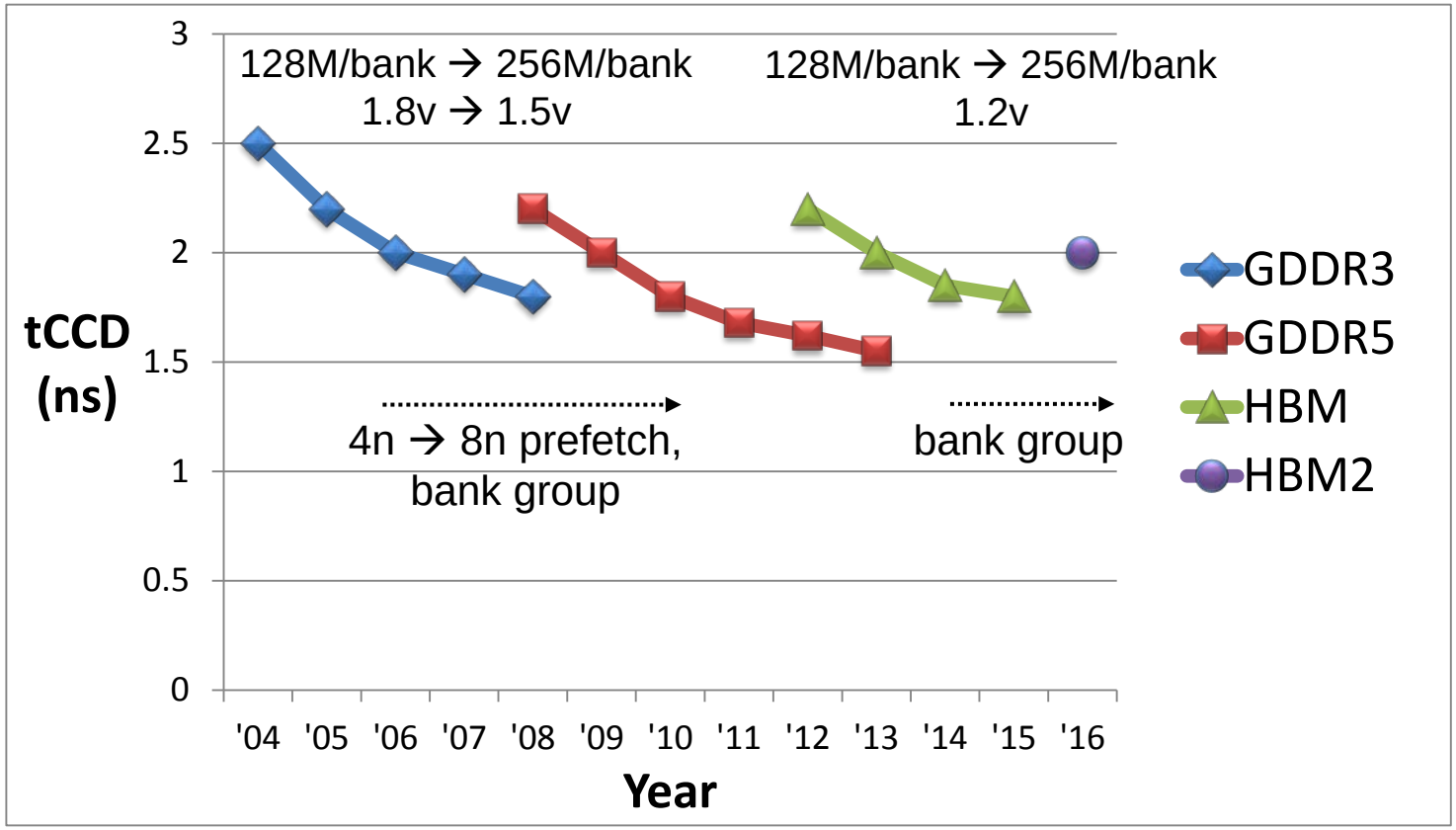
System-in-Package using Interposer

- Logic layer has direct interface with SoC(2.5D)
- Many applications use HBM as Near-Memory



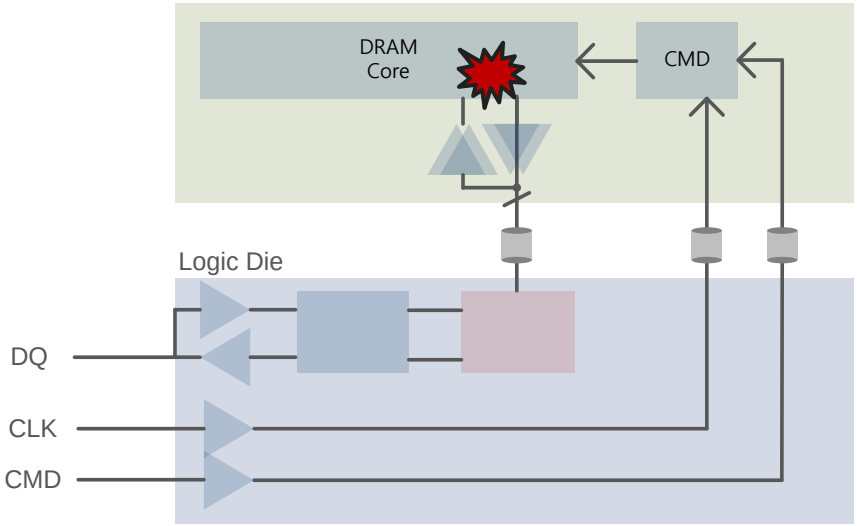
tCCD limitation

- **Year vs. tCCD**
  - Because of the increase of DRAM density and voltage down, tCCD is bounded to around 2ns

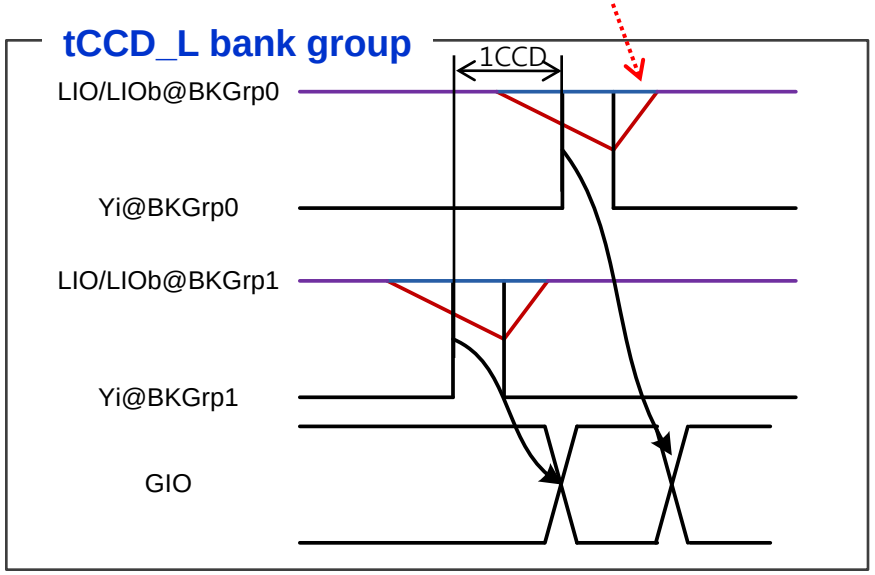


tCCD limitation

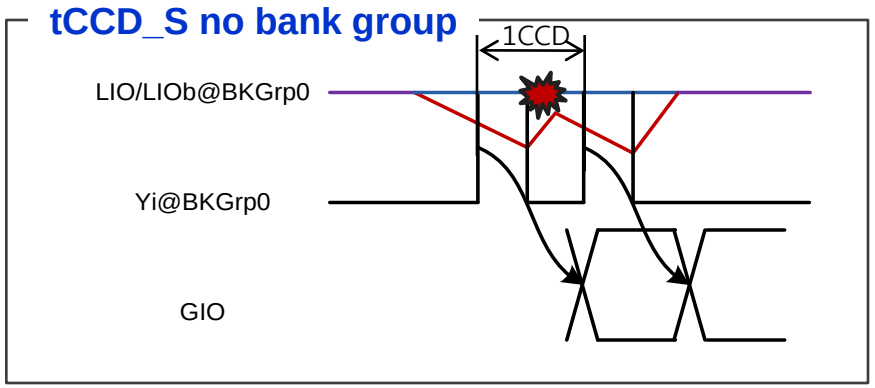
- **Source of tCCD limitation**



Long LIO → small swing

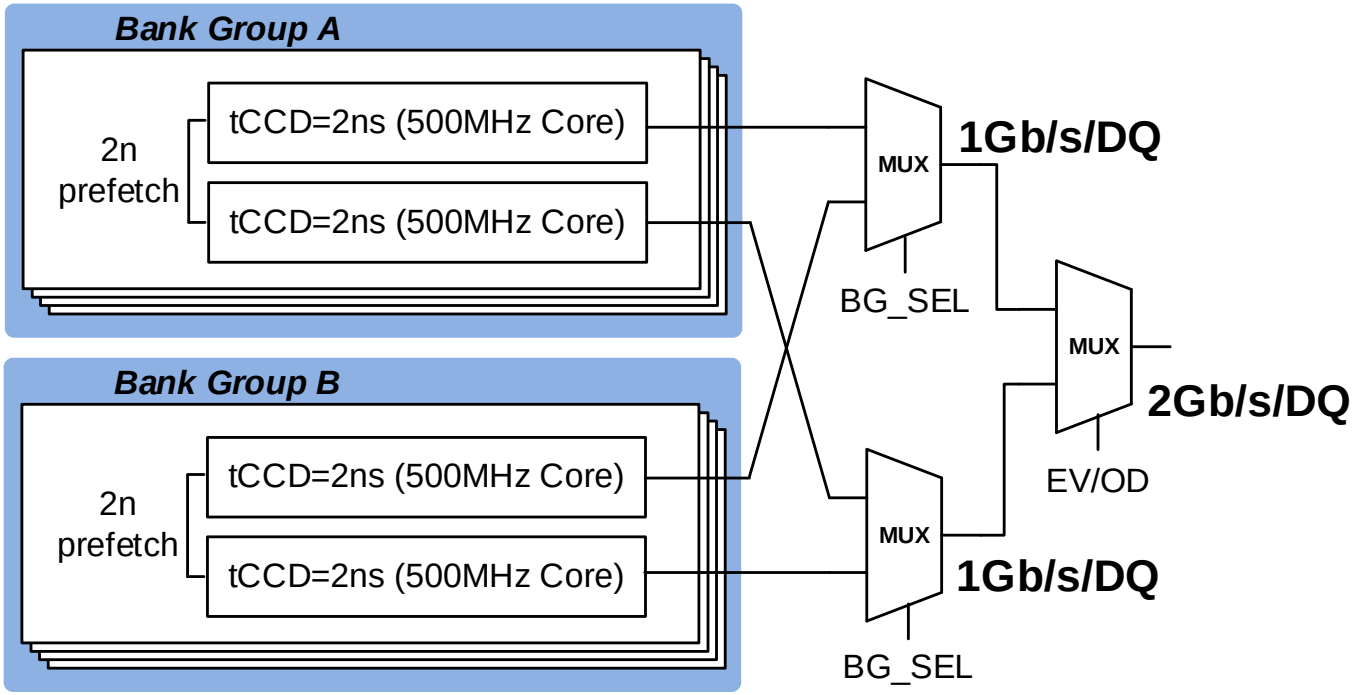


– LIO/LIOb of bank core cannot be fully precharged without bank group



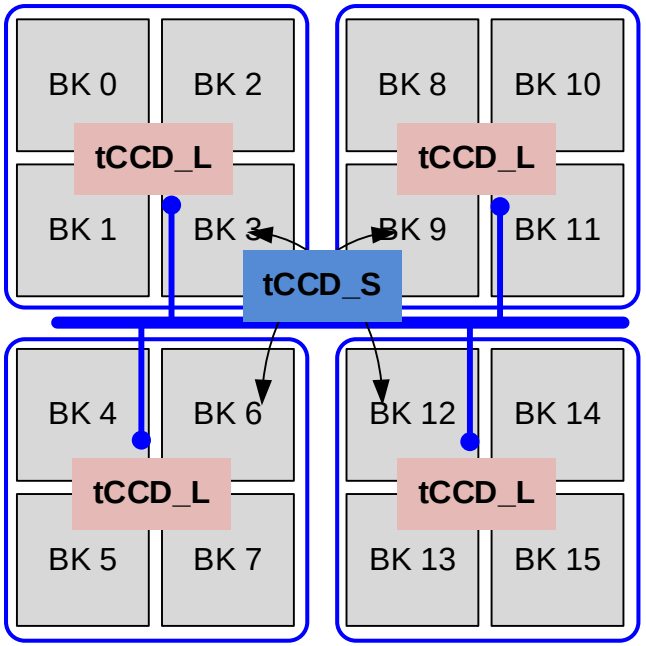
Bank Group

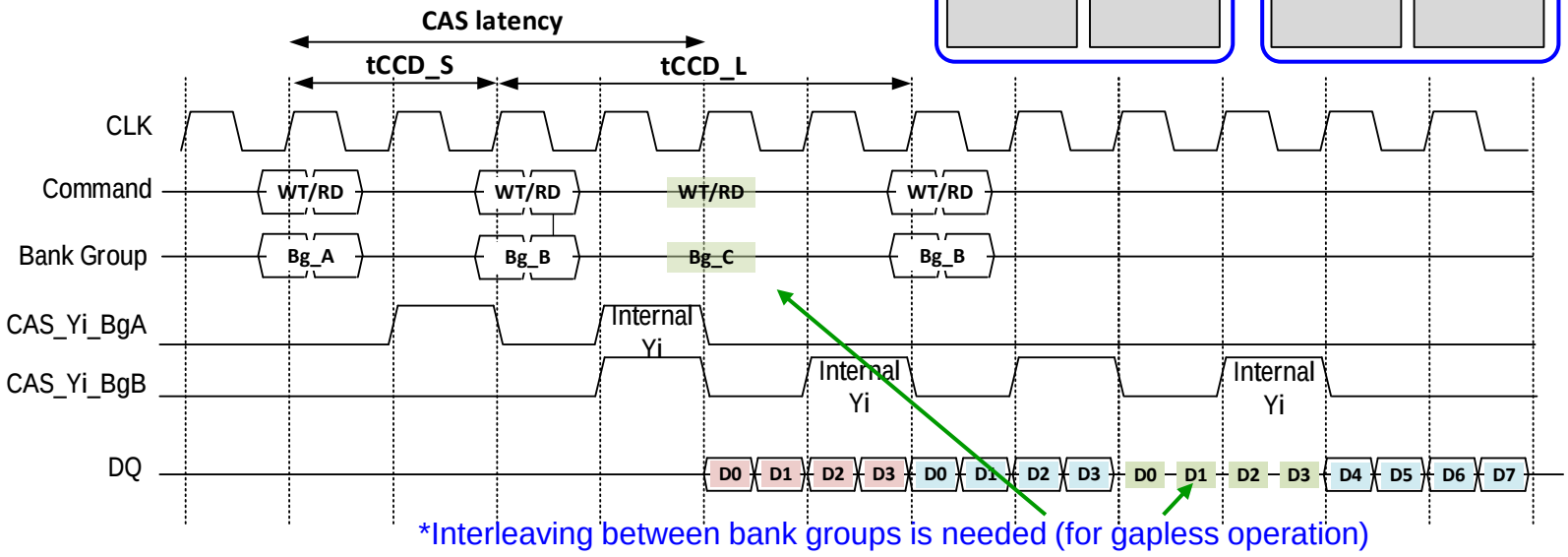
- **Bank group provide increased bandwidth without more prefetch**
- **Restriction between bank groups is needed**
  - 4 banks x 4 groups are implemented in HBM 2nd



Bank Group

- **Case1: Prefetch increase**
  - Large area penalty
  - Increased page size
- **Case2: Bank group**
  - Small area penalty
  - Restriction in bank access

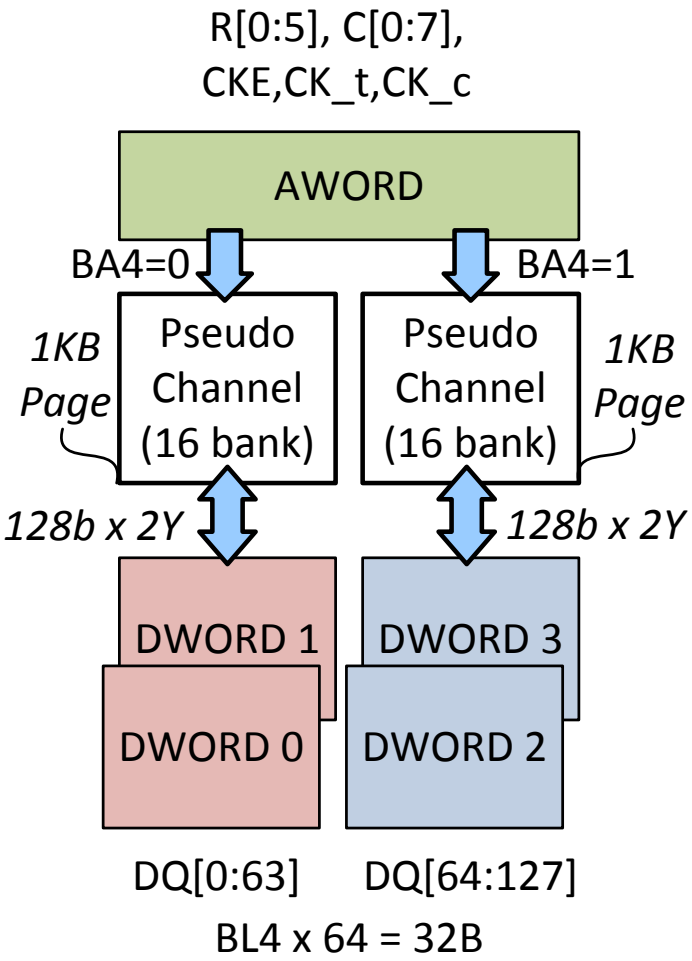




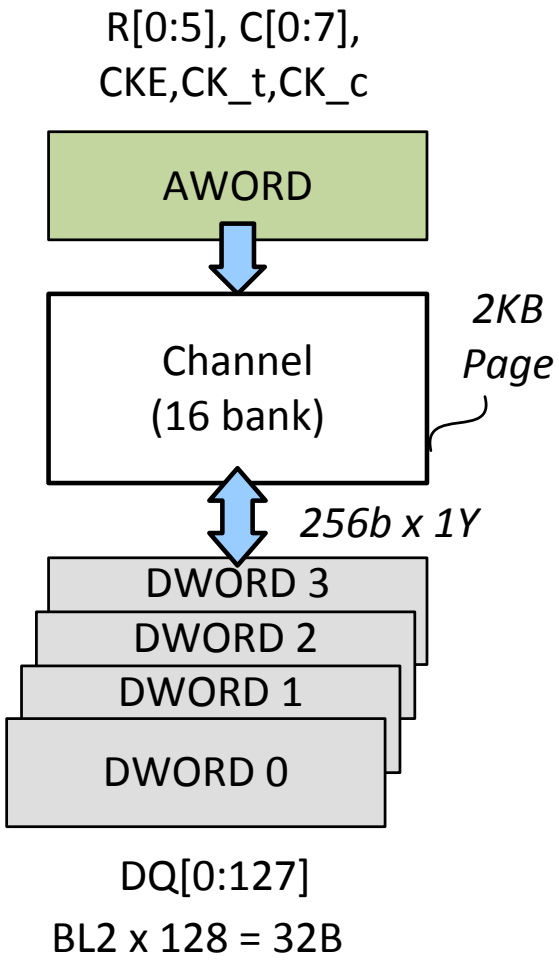
*Interleaving between bank groups is needed (for gapless operation)

Architecture Comparison

Pseudo Channel Mode

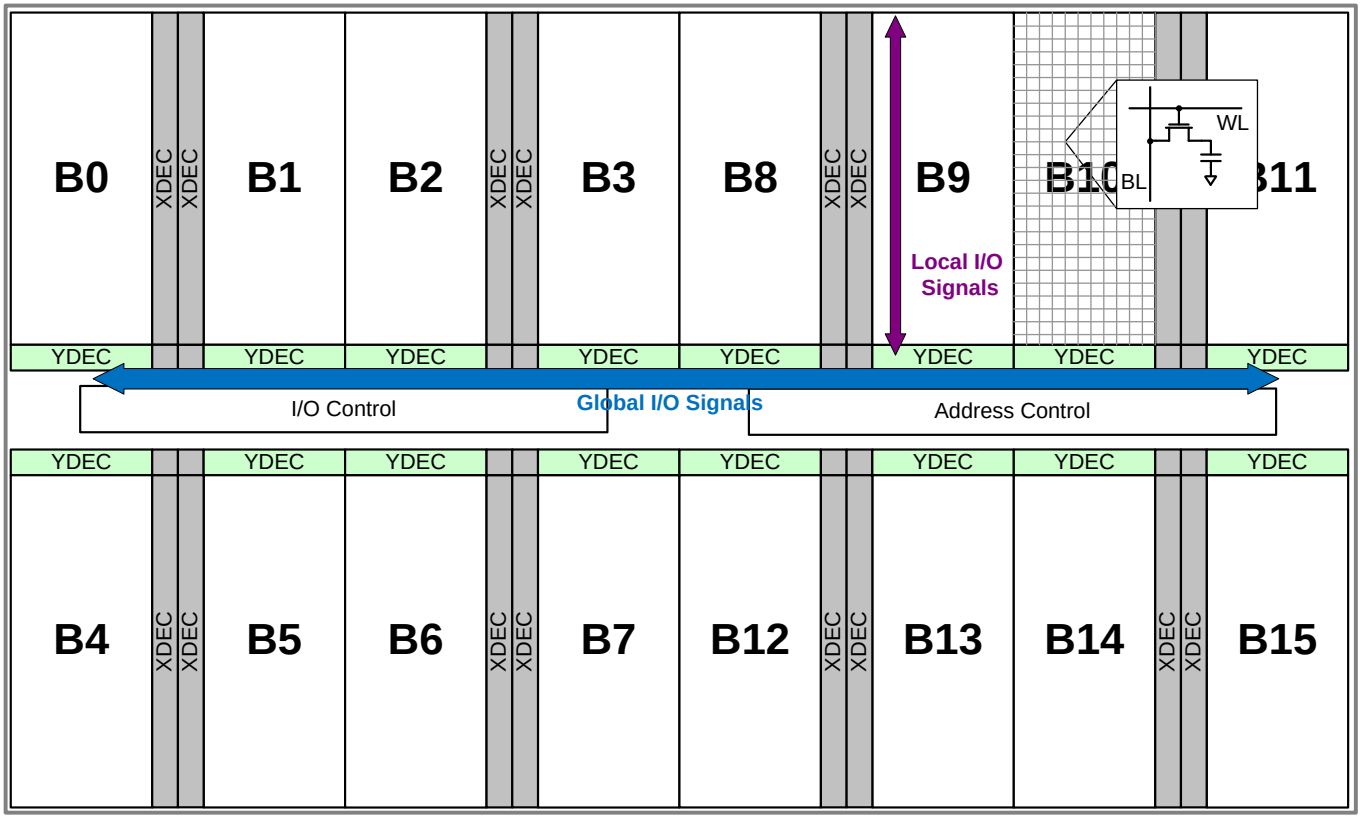


Legacy Mode



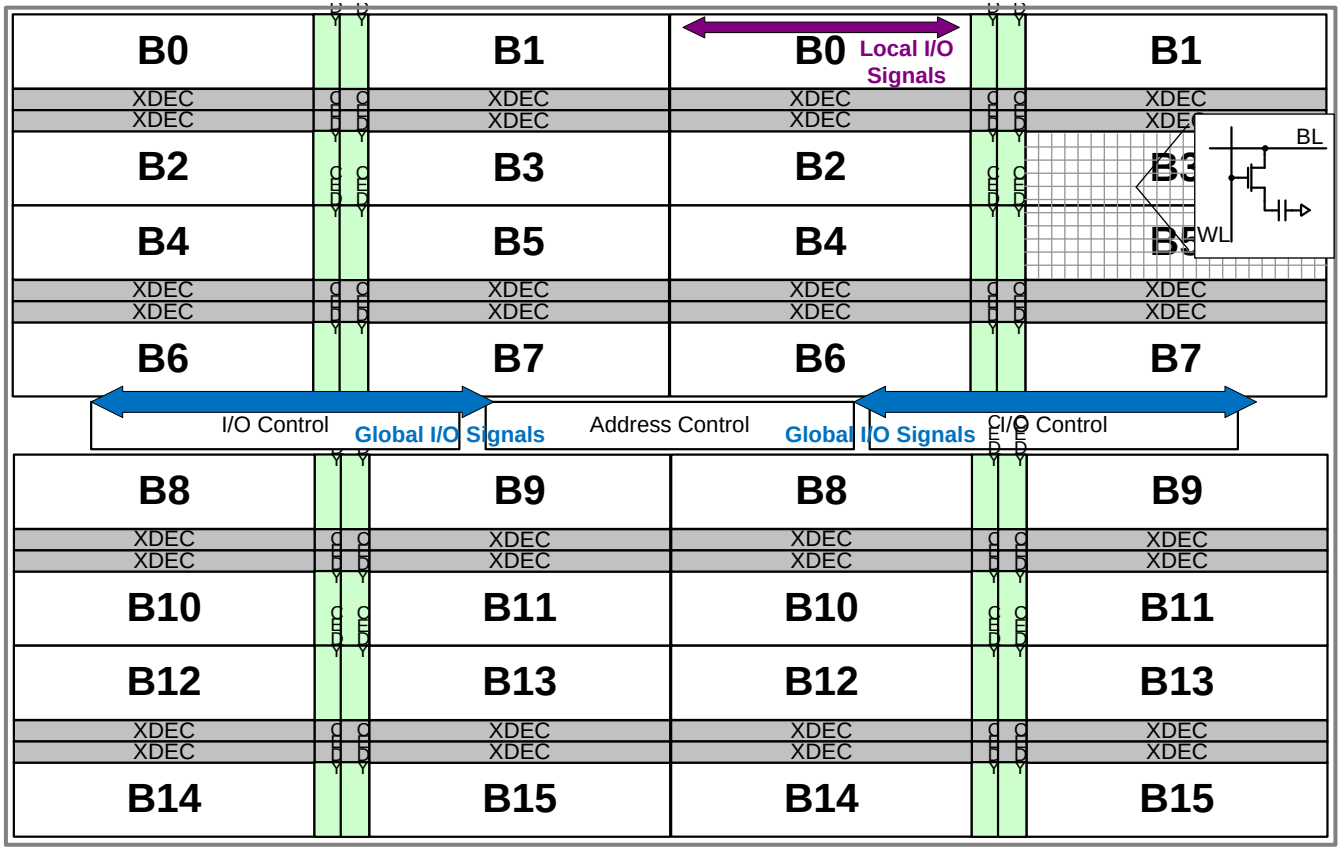
Full Bank Architecture

- **Local I/O, global I/O : Long**
- **Cell efficiency : High (minimum area penalty), Page Size : No increase**



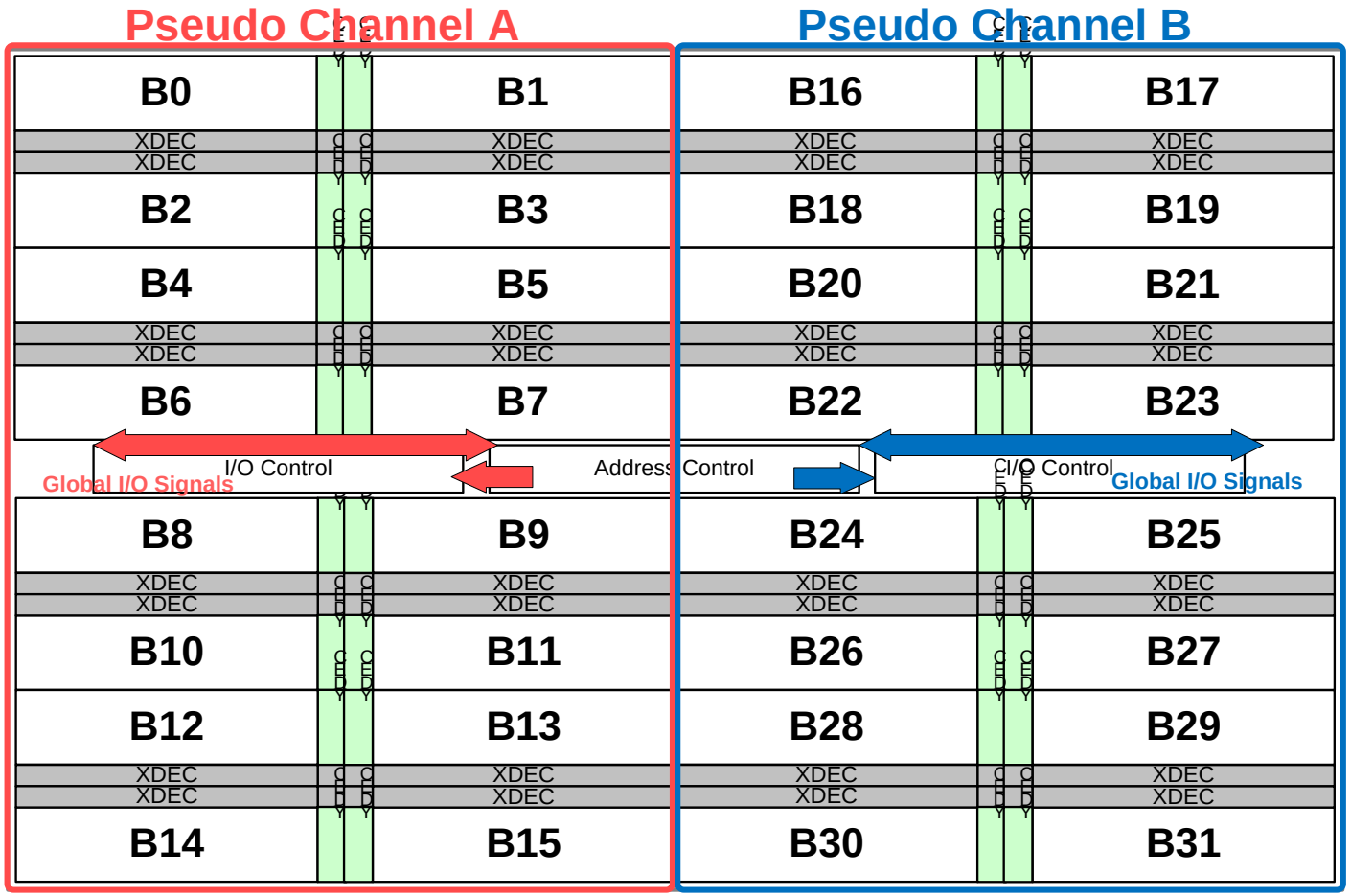
Half Bank Architecture

- **A bank is divided into two sub-banks**
  - Local I/O, Global I/O : short
- **Cell efficiency : Low, Page Size : x2**



Pseudo Channel Architecture

- **Architecture : almost same as half-bank**
- **Page Size : No increase (x2 number of banks)**



Advantages of Pseudo Channel

- **Increased system performance**
  - Number of banks are doubled, IO are separated, page size is reduced
- **Area efficient DRAM architecture**
  - DRAM has optimized bank architecture
  - Any additional functionality would increase cost
  - 0.4% of total chip area increase by AWORD

→Independent I/O control with area efficiency

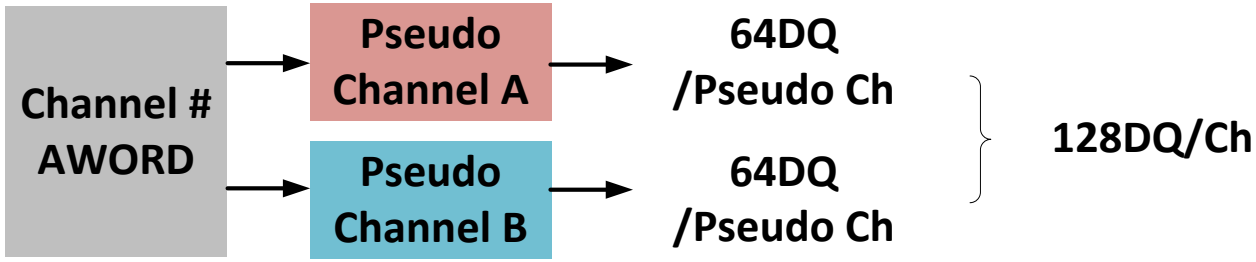
Candidates of Pseudo Channel

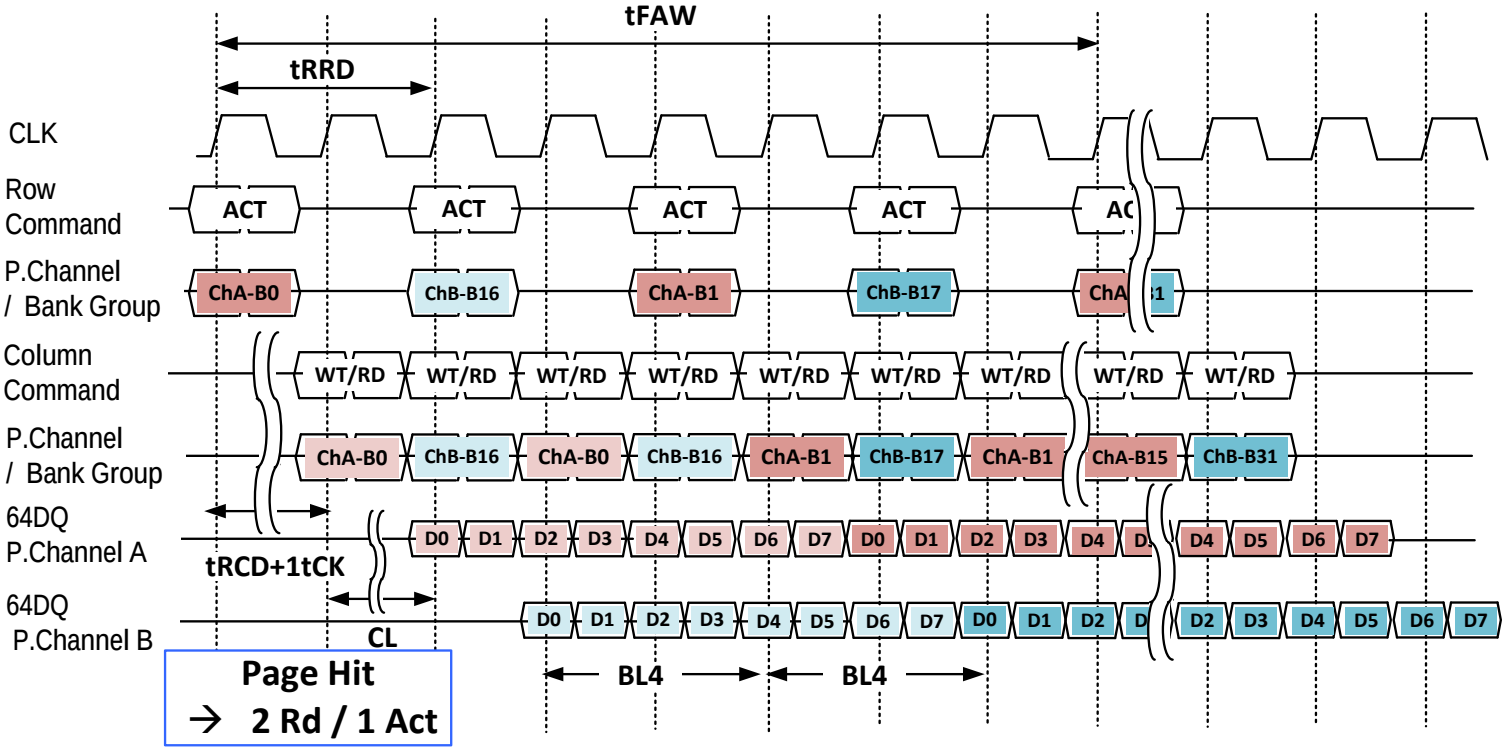
- **Three possible architecture for pseudo channel (4Gb/ch)**

	Double Row	Double Column	Double Bank
Pro	Simple circuitry No area penalty	Prefetch 2 → 4 (tCCD=1ns)	Effective bandwidth increase
Con	Restriction in bank activation	2KB page size per pseudo channel	Area penalty+0.4% Circuit complexity
Addressing (4Gb)	**RA[0:14]** CA[0:5] BA[0:3]	RA[0:13] **CA[0:6]** BA[0:3]	RA[0:13] CA[0:5] **BA[0:4]**
Select P.C.	**BA3**	**BA3**	**BA4**
HBM 2nd adopt	X	X	O

Timing Diagram of Pseudo Channel

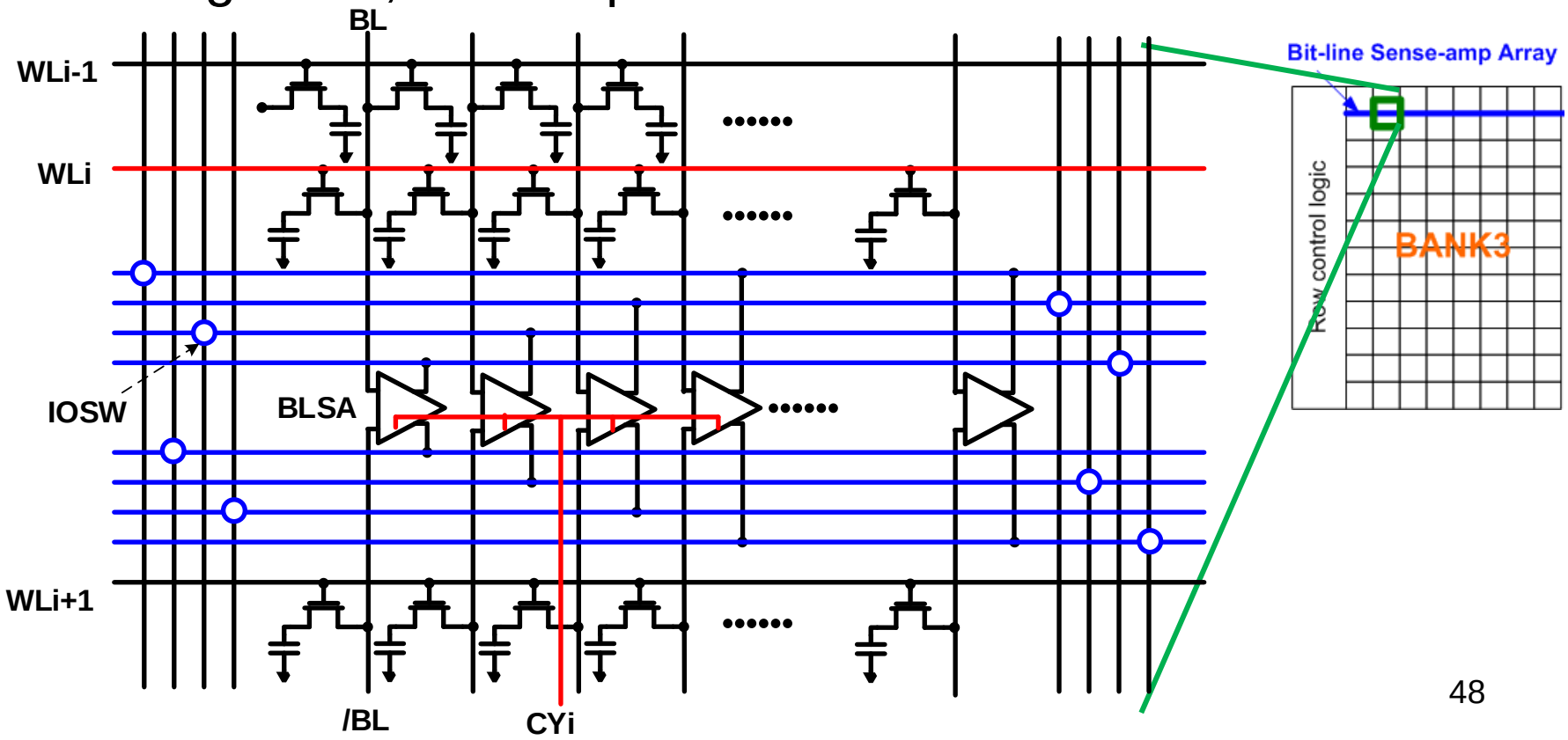
- **Command interleaving, I/O separation**





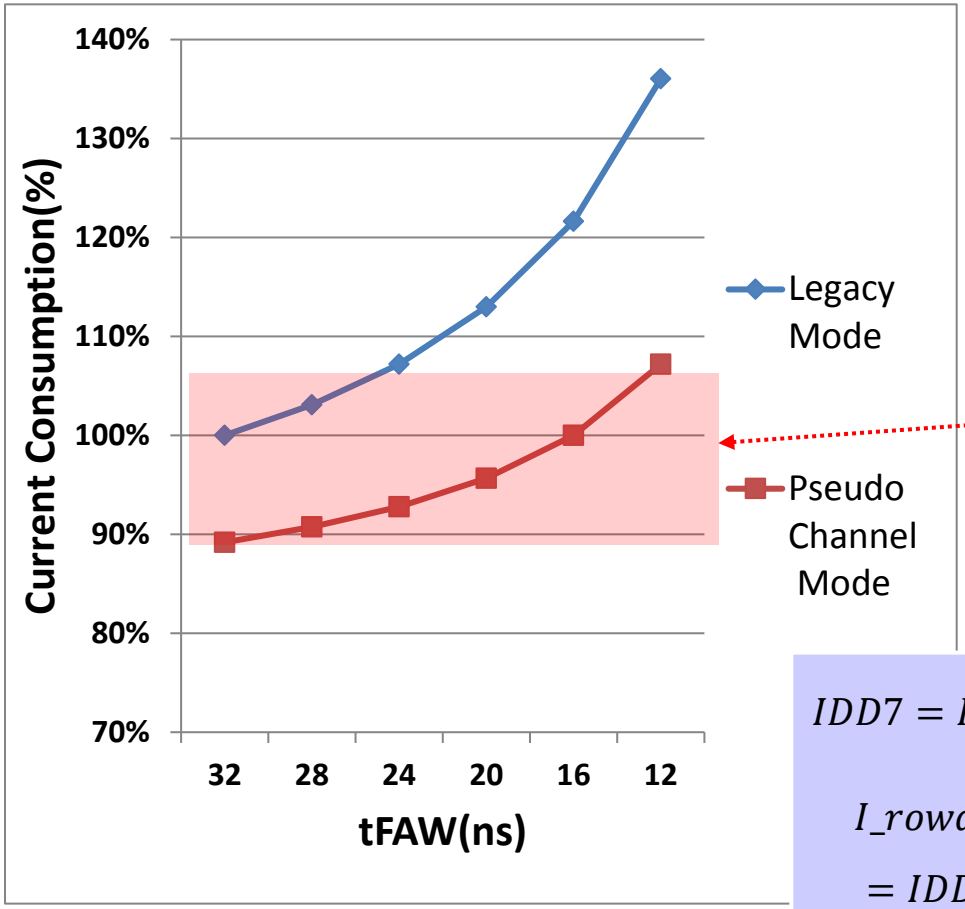
Cell Array for Page Size

- **DRAM page size**
  - The amount of data transferred from DRAM cell into the open page buffer
  - Page size, access period increase → active current



Current Consumption vs. tFAW

- **Max current consumption increase**



** Bus usage is assumed as 80%*

** tFAW(2KB) of conventional DRAM : 30-40ns*

**memory power budget is limited (without additional area penalty)*

$$IDD7 = IDD4R * bus\_usage + \frac{tRC}{tRRD} * I\_rowact$$

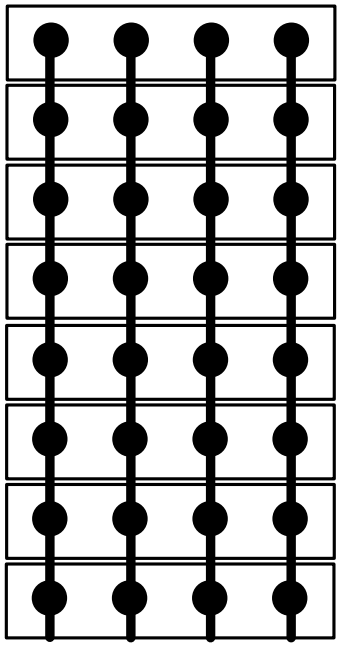
$$I\_rowact = IDD0 - \frac{tRC}{tRAS} * IDD3N - \frac{tRC}{tRP} * IDD2N$$

TSV connection

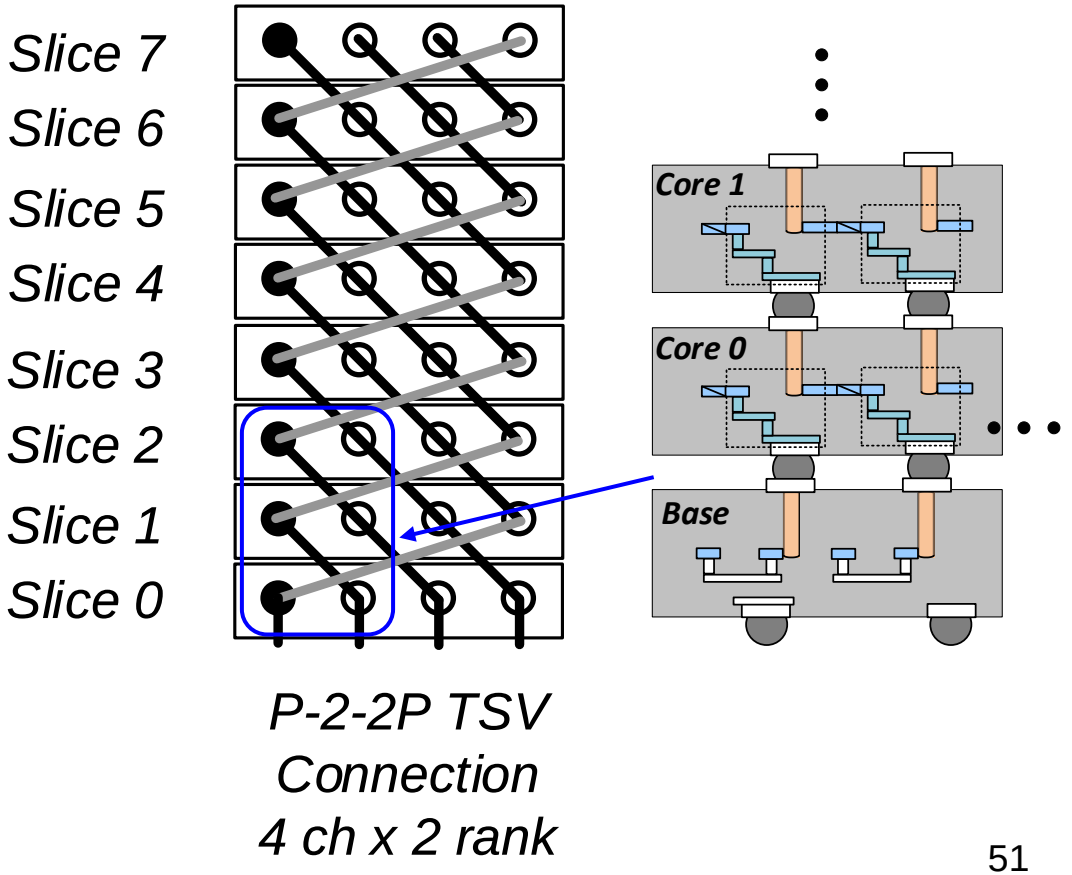
- **Multi-drop connection**
  - I/O capacitance large, via middle, via last, good flexibility and simple TSV structure
- **P2P connection**
  - I/O capacitance small, via middle, limited flexibility and complex TSV structure
- **P-2-2P connection**
  - I/O capacitance moderate, via middle, moderate flexibility and complex TSV structure

TSV connection

- **Multi-drop：Large I/O load, minimum metal load**
- **P2P：Small I/O load, slightly increased metal load**



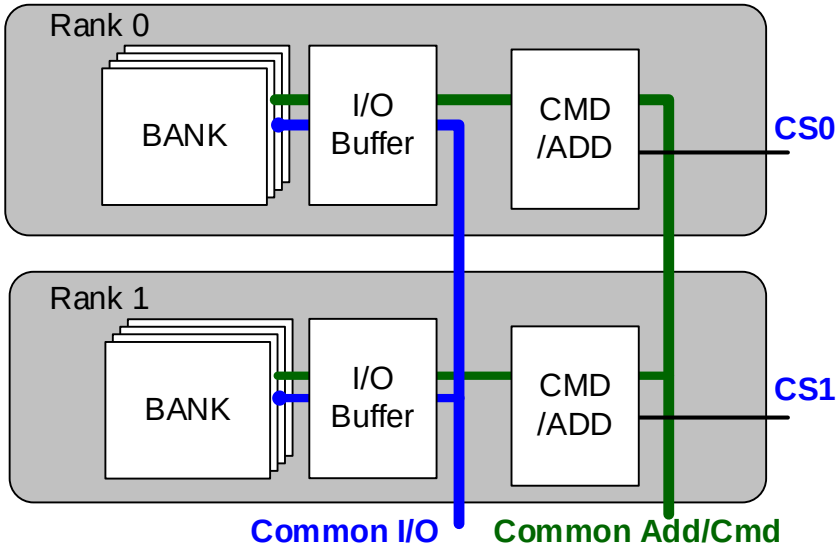
Multi-drop TSV Connection



P-2-2P TSV Connection 4 ch x 2 rank

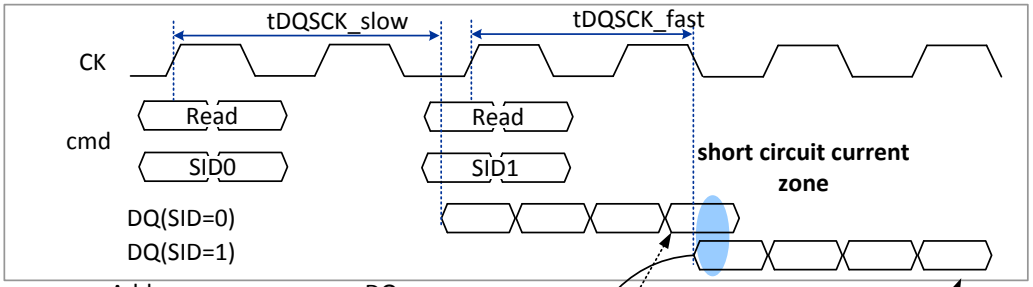
Advantages of Rank Structure

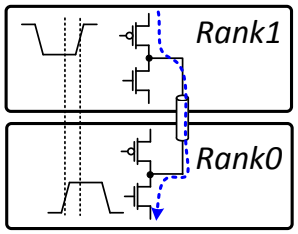
- **Memory Core**
  - More memory banks, density and same page size
- **Memory I/O**
  - Short I/O path, no change in pin count(except for /CS)
- **Flexibility**
  - Ranks are easily added or removed without structure change

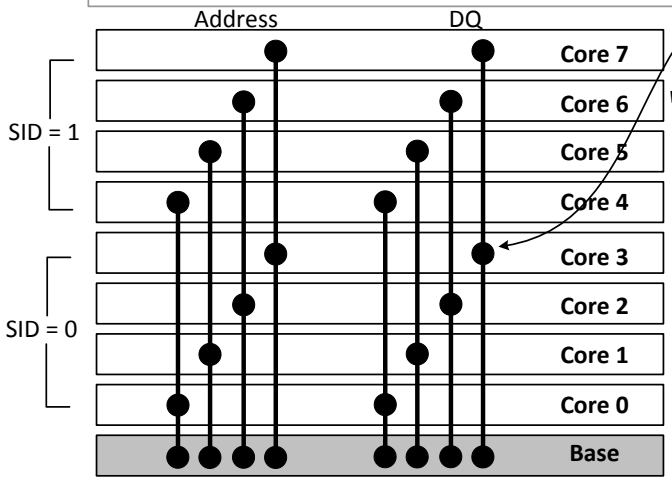


8-Hi Rank Structure

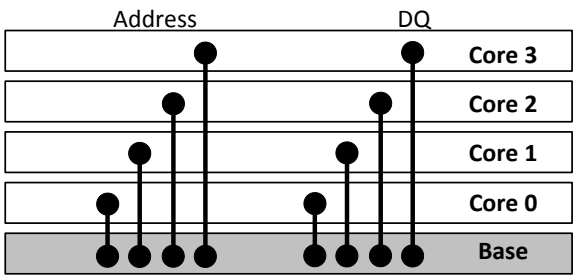
- **Reduced signal routing with chip-to-chip slice skew variation**
  - Large I/O circuit loading are added







8 Hi Structure
Point-to-2Point

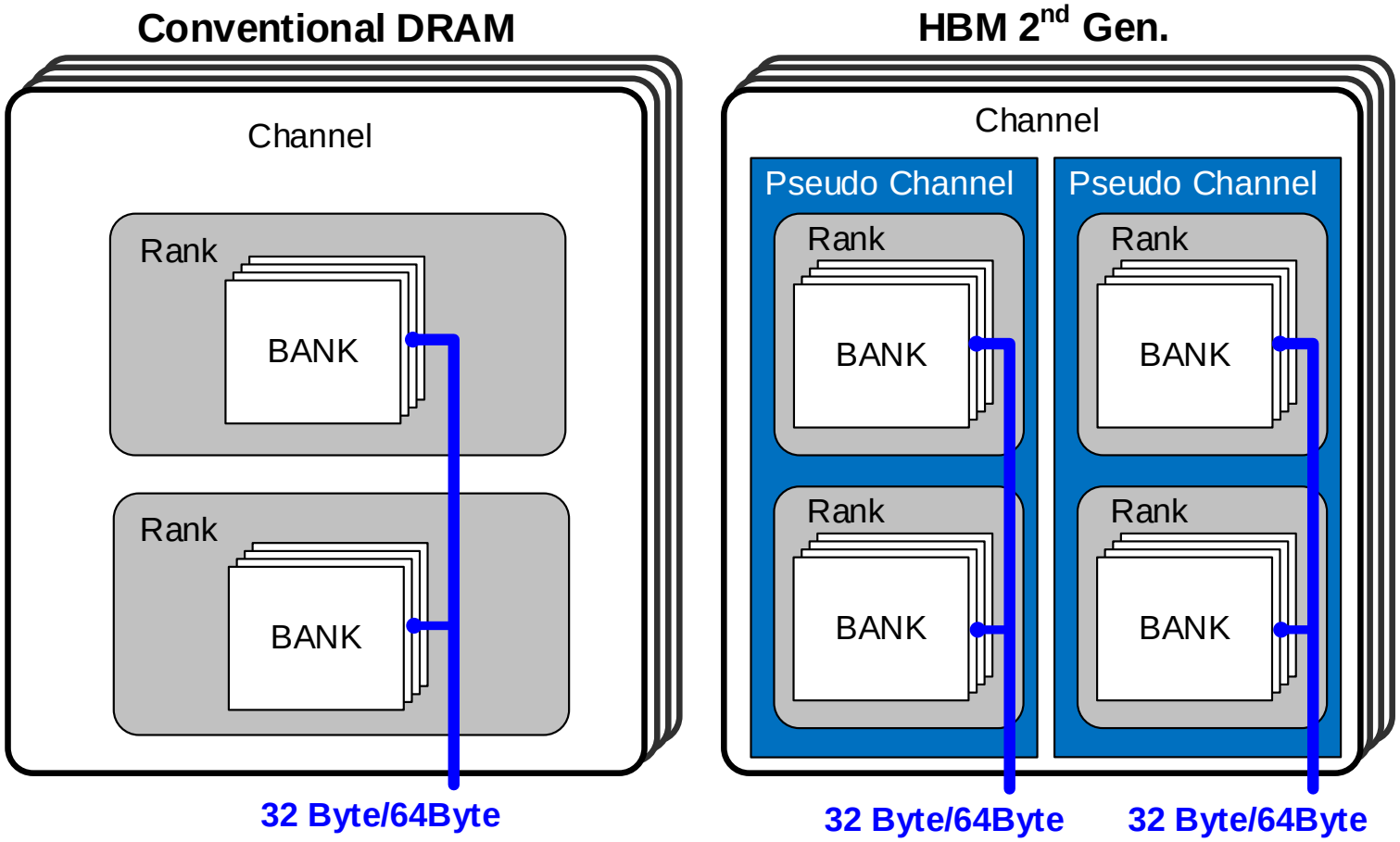


4 Hi Structure
Point-to-Point

tCCDR definition
: gap between ranks

Memory Hierarchy

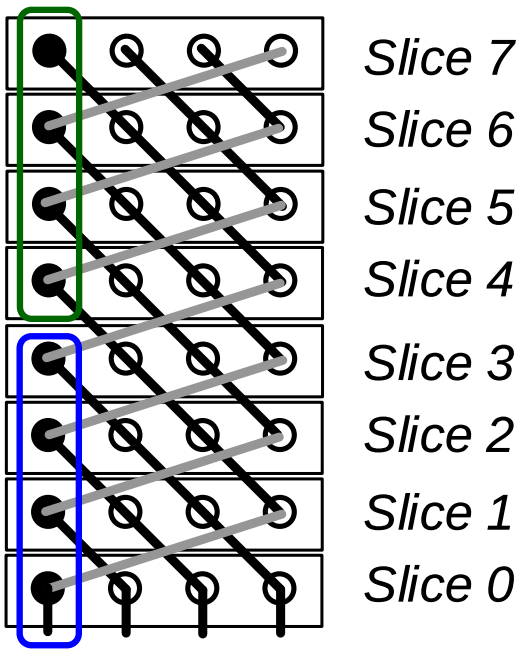
- **A channel supply two independent memory complex**



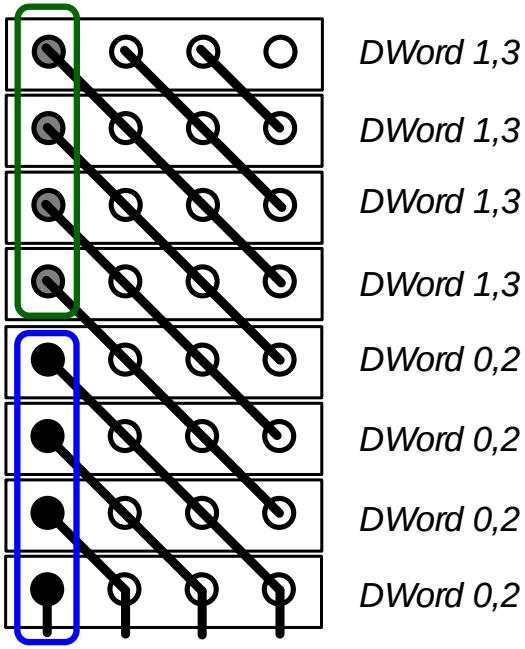
Comparison between 8-Hi Structures

- **8-Hi organized to dual rank (Left) and single rank (Right)**

Each I/O connected *2n prefetch x 2Y*



P-2-2P TSV Connection 4 ch x 2 rank



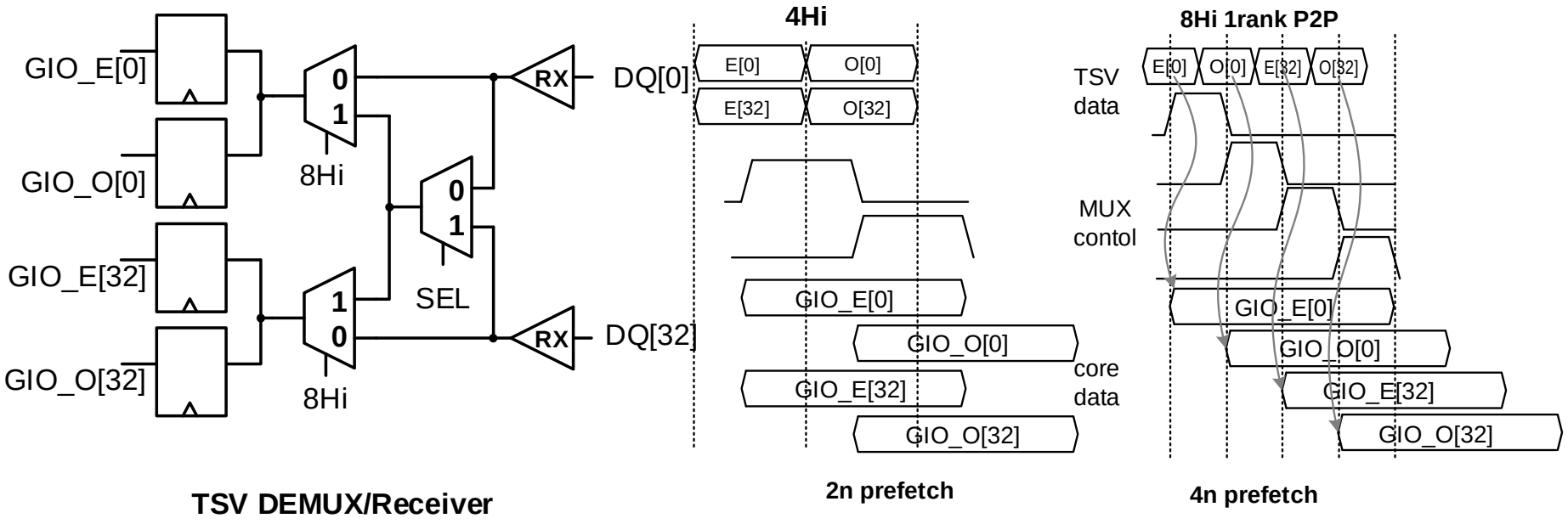
P2P TSV Connection 4 ch x 1 rank

Each I/O separated *4n prefetch x 1Y*

*4ch x 2 rank is adopted in HBM 2nd

Single Rank P2P TSV Deserializer

- **2 x 2-to-1 Deserializer (4-Hi) – 2n prefetch**
- **4-to-1 Deserializer (8-Hi) – 4n prefetch**
  - Vertical connection enables the increased prefetch



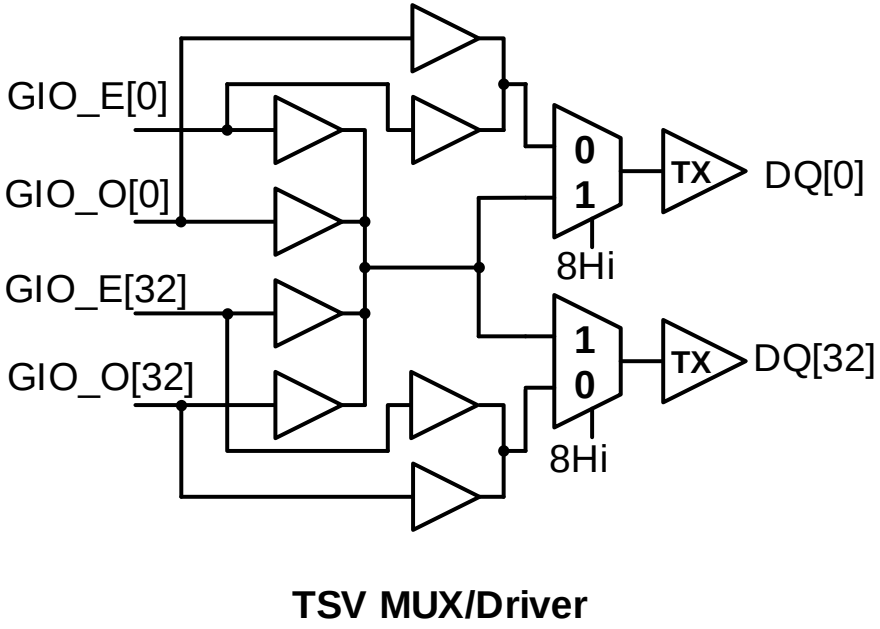
TSV DEMUX/Receiver

2n prefetch

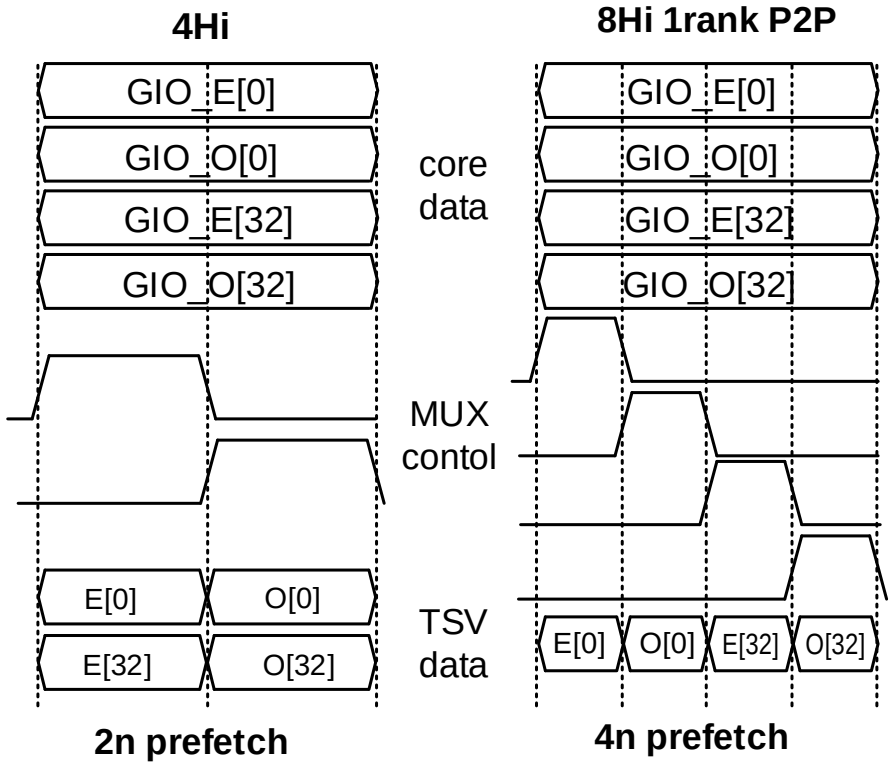
4n prefetch

Single Rank P2P TSV Serializer

- **2 x 2-to-1 Serializer (4-Hi) – 2n prefetch**
- **4-to-1 Serializer (8-Hi) – 4n prefetch**
  - Rank-to-rank I/O skew problem is removed



TSV MUX/Driver

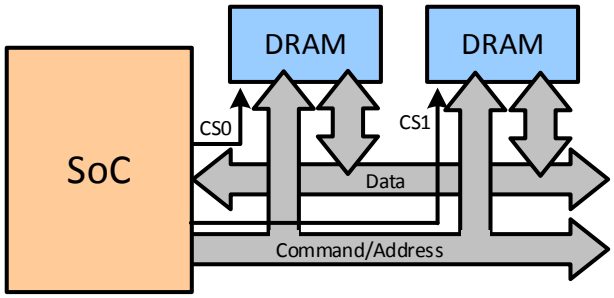


2n prefetch **4n prefetch**

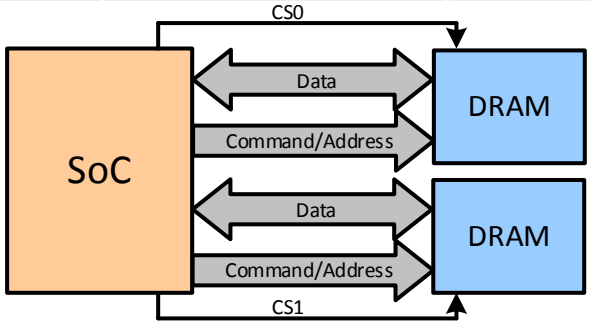
Memory Organization

- **Architecture analogy with off chip I/O**
  - HBM 2nd has three off chip architecture on one chip

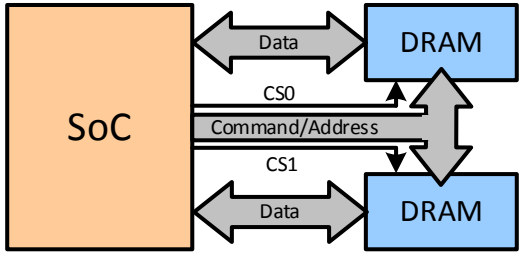
	I/O	Command	Off chip (ID)	HBM 2nd (ID)
Bank	Common	Common	-	-
Rank	Common	Common	CS0, CS1,…	SID
Shared C/A (Pseudo Channel)	Separated	Common	CS0, CS1,…	BA4, /BA4
Channel	Separated	Separated	Independent operation	Independent operation



Multi Rank



Multi Channel



Multi Channel With shared C/A

Target Specification

	HBM 1st Gen.	HBM 2nd Gen.
Density(1 channel / Total)	1Gb / 8 Gb	8Gb / 64Gb
Microbump ballmap	6.05mm x 3.26mm (48x55µm pitch)	
I/O (Per channel / Total)	128 IO / 1024 IO	
Supply(VDD/VDDQ/VPP)	1.2 / 1.2 / 2.5	
Channel / Bank / Page	8ch * 8bank / 2KB	8ch * 64bank/ 1KB(p.c.)
Speed (Per pin / Total)	1Gbps / 128GBps	2Gbps / 256GBps
Output Driver	6mA~12mA	6mA~18mA
Chip stack (Max)	4Core + 1Base (5Hi)	8Core + 1Base (9Hi)
Burst length	2,4	4

Conclusion

- **HBM is developed to get the bandwidth 128GB-256GB/s**
  - Applications want to break though the memory wall
- **Microbump AC/DC test, TSV scan/repair and 3D PDN analysis**
  - Minimize the cost adder caused by stacking
- **Architecture for system performance**
  - Pseudo channel architecture
  - Vertical multi-rank architecture
- **Future of HBM DRAM**