Architectural Design of Low-latency and High-bandwidth 3D NAND with Split-Stacked Row Decoder for LLM Inference

Seoul National University of Science and Technology
Department of Electrical and Information Engineering
Wonbo Shim

NVRAMOS 2026
2026. 10. 22.

Information

- Presented at 2026 IEEE International Memory Workshop (IMW), Leuven

Architectural Design of Low-latency and High-bandwidth 3D NAND with Split-Stacked Row Decoder for LLM Inference

Youngjun Park[1#], Chanho Kim[1#], Seunghwan Song[2], Wonbo Shim[1*]

[1]Department of Electrical and Information Engineering, Seoul National University of Science and Technology, Seoul, Republic of Korea

[2]Device Solutions, Samsung Electronics Co., Hwaseong, Republic of Korea

Contents

1. Introduction

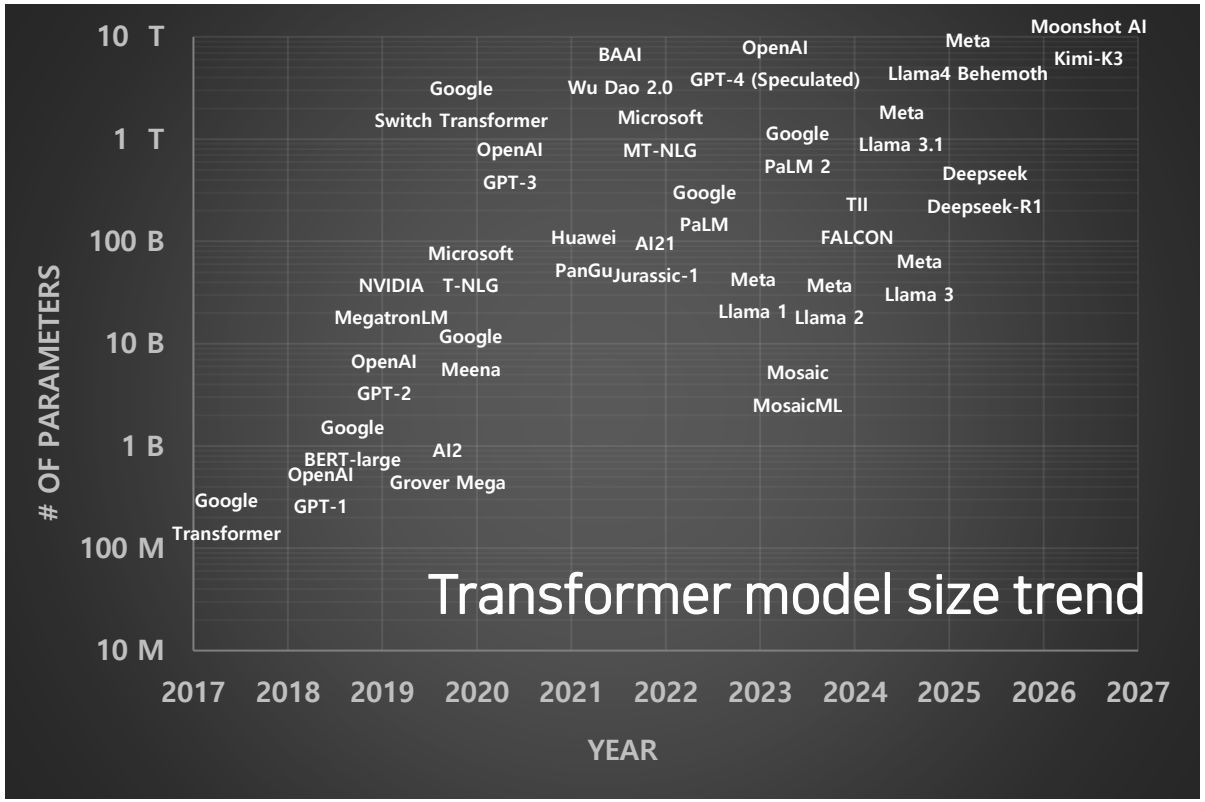
2. Low-latency 3D NAND Array Design

3. Split-Stacked Row Decoder Design using Hybrid Bonding Technology

4. Design Space Exploration of High-bandwidth NAND Stack Architecture

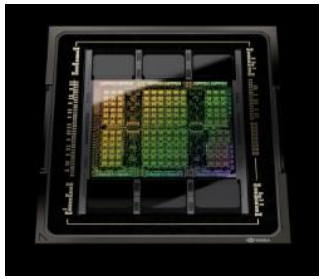
Introduction

- AI demands both high-bandwidth and high-capacity memory.



(Kimi-K3) 2.8T × MXFP4 ≈ 1.5 TB

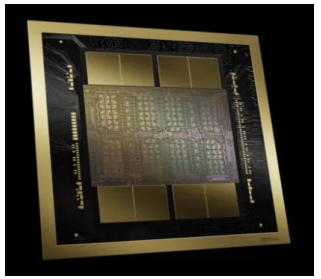
20 ×



H100 (80GB)

or

8 ×

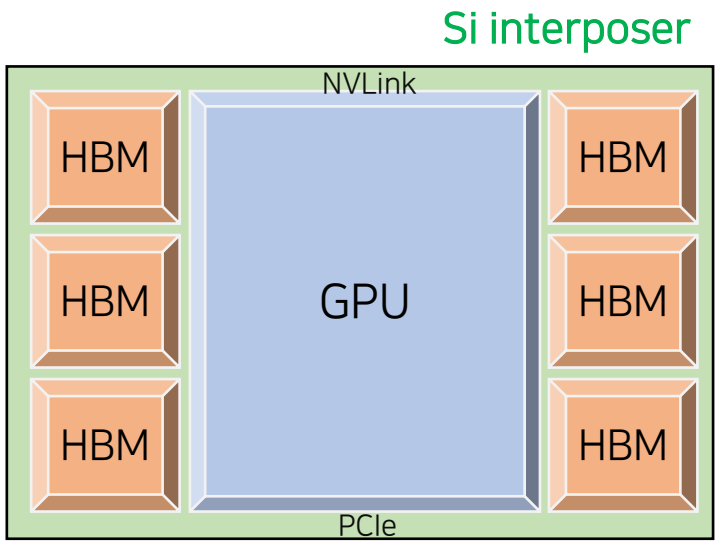


B200 (192GB)

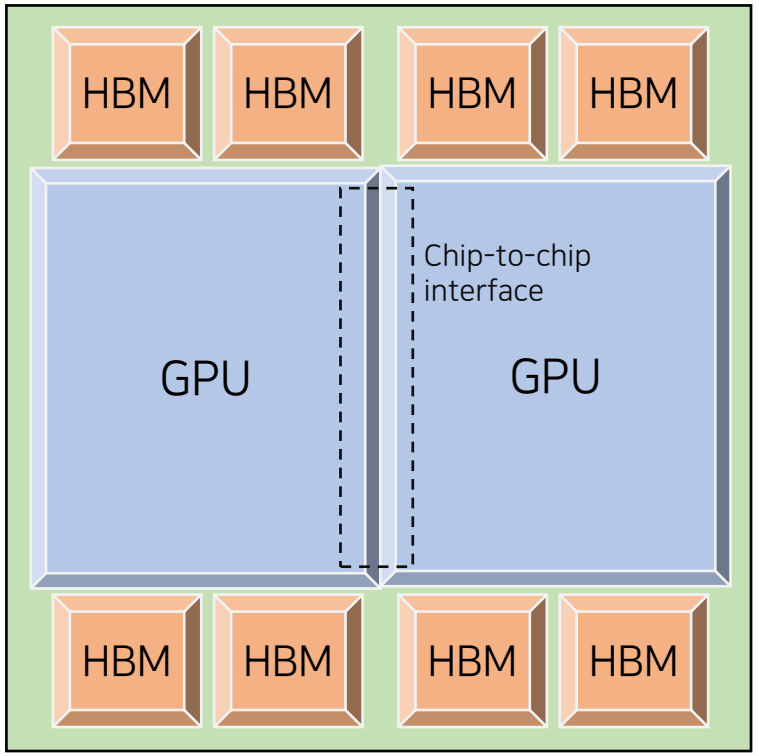
CAPEX, inter-GPU data trasfer ↑

Introduction

- Key barriers to scaling GPU memory capacity



NVIDIA Ampere, Hopper



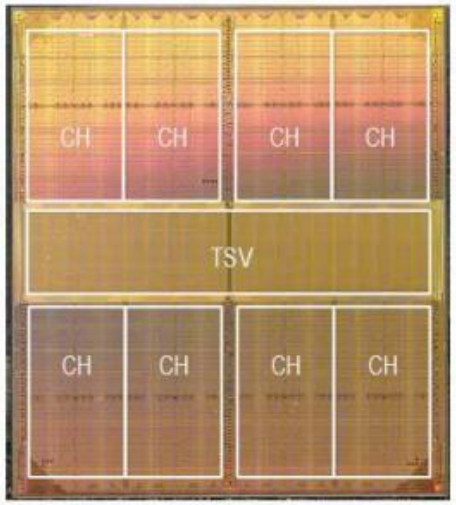
NVIDIA Blackwell, Rubin

- Max GPU die size
  - 26mm × 33 mm
  - Limited shoreline of GPU
  - Limited HBM-GPU interface
- Max GPU die thickness
  - ~720 um (HBM3 gen.)
  - Height limit of HBM die stack

Introduction

- No more space. Higher density required.
  - DRAM cell scaling down, 3D DRAM
  - 3D NAND based High Bandwidth Flash/Storage can be candidate for AI inference.

DRAM

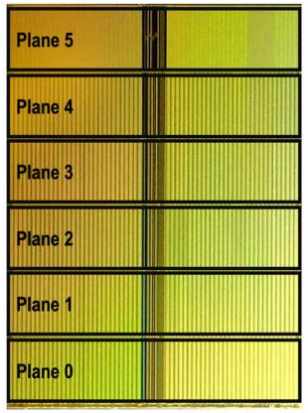


ISSCC 2026, Samsung

HBM4 core die
24Gb @ 12.8mm × 11mm

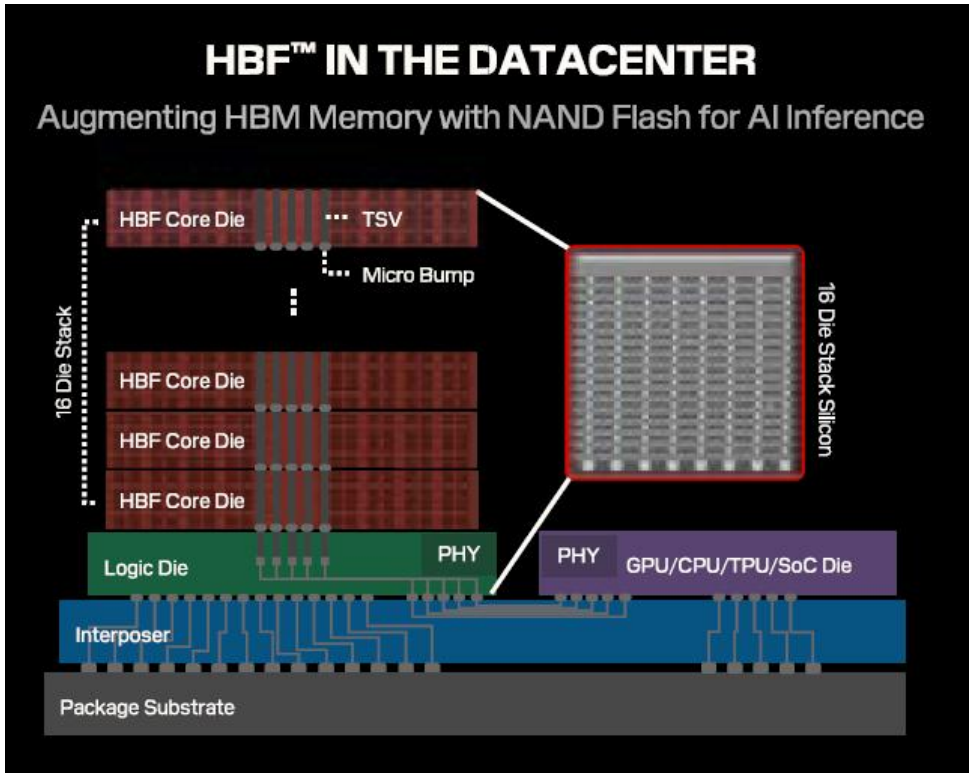
VS

NAND



ISSCC 2026, Sandisk

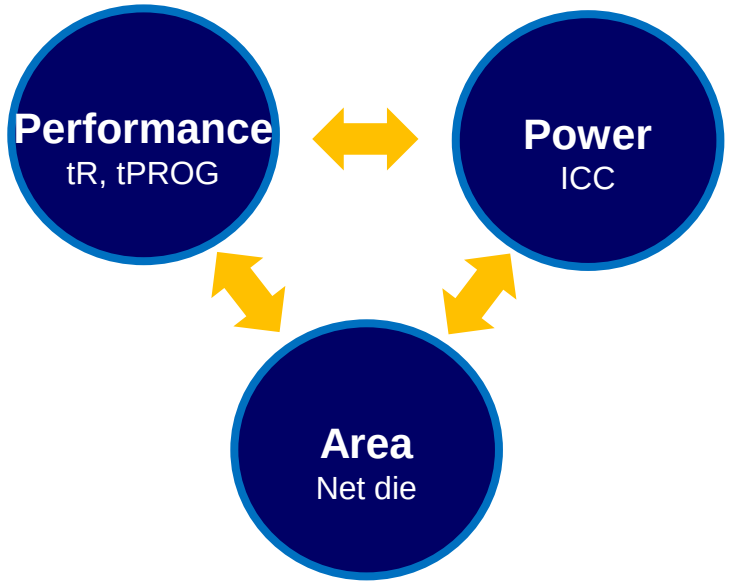
332-layer 3D NAND die
2Tb @ 54mm²



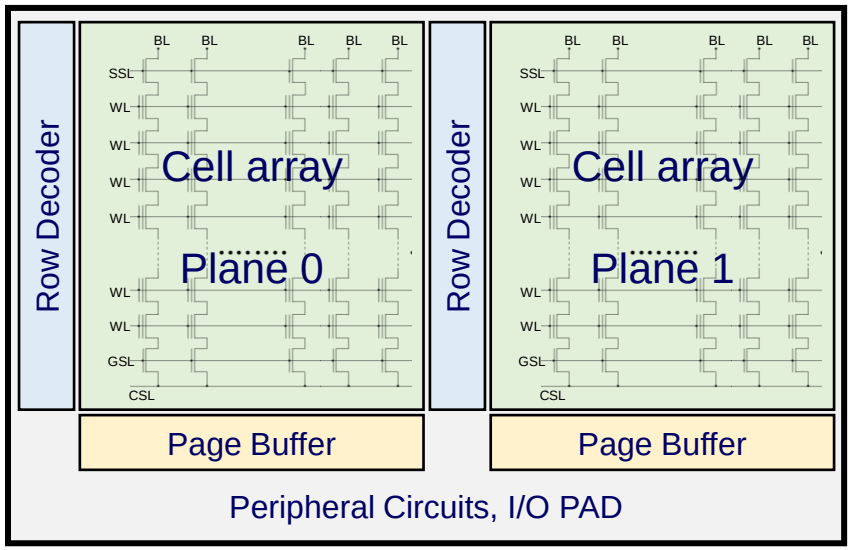
Source: Sandisk

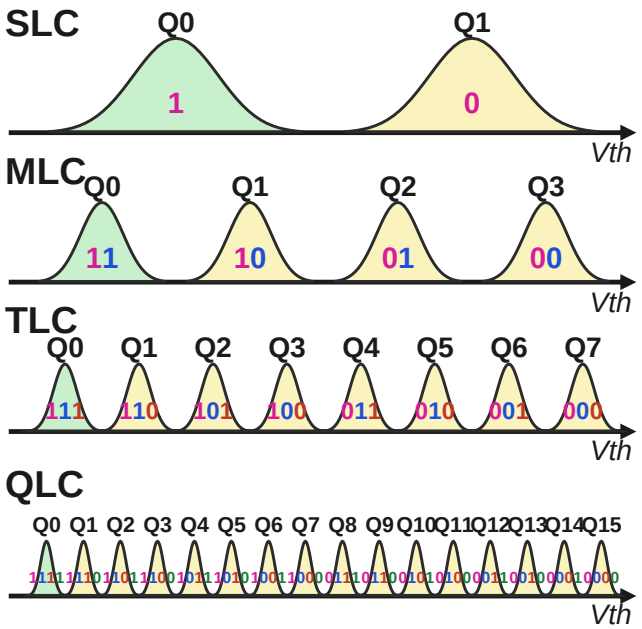
Introduction

- NAND is slow, because…
  - For conventional storage applications, NAND vendors focus primarily on cost reduction.
  - Large array (plane) size (8~16kB WL size, ~mm BL length) → WL/BL RC ↑ → tR ↑
  - Multi-level cell (TLC, QLC) → # reads → tR ↑
  - System-level performance = f(tR, #channel, plane, scheduling…)



NAND die architecture





* Samsung Z-NAND, Kioxia XL-Flash has relatively small array size to achieve 3-4μs tR.

Contents

1. Introduction

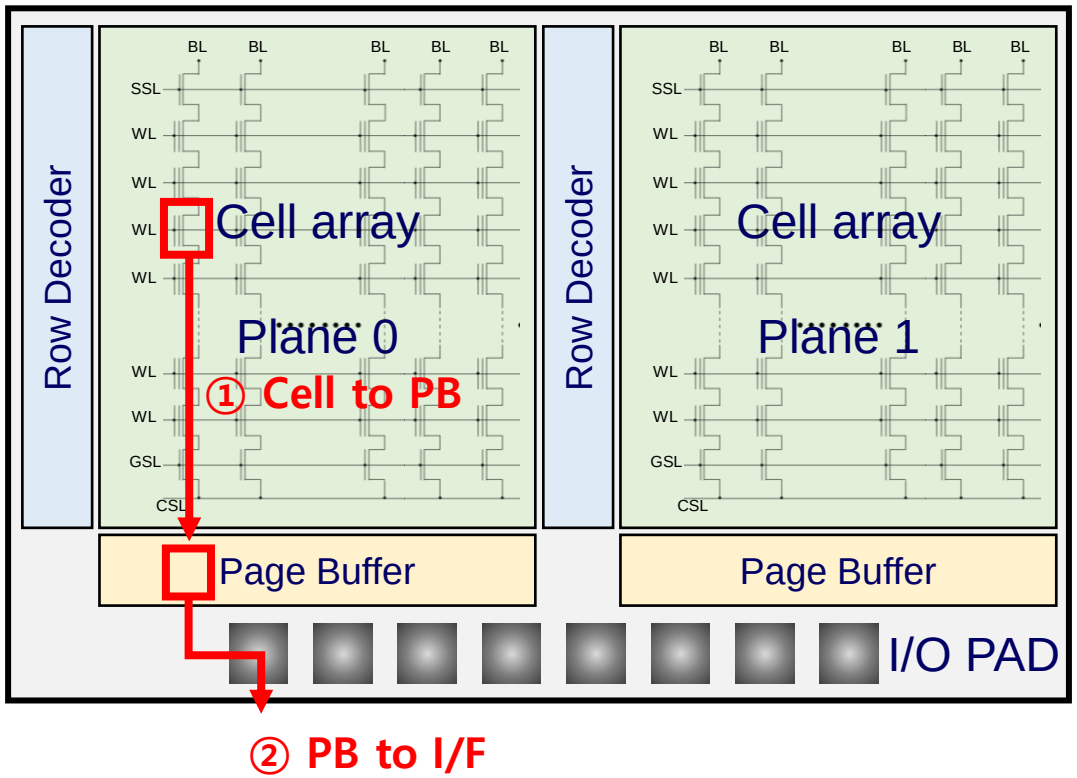
2. Low-latency 3D NAND Array Design

3. Split-Stacked Row Decoder Design using Hybrid Bonding Technology

4. Design Space Exploration of High-bandwidth NAND Stack Architecture

Low-latency 3D NAND Array Design

- Needs for low-latency 3D NAND
  - tR must be reduced for high-bandwidth read operation.



Conventional NAND read throughput (per die)

	Speed	Parallelism	Bandwidth
① Cell to PB	tR 40us (3bit/cell)	PB 64kB (4-plane)	1.6 GB/s
② PB to I/F	3.2 Gbps	8 bit I/O	3.2 GB/s

High-bandwidth NAND read throughput (per die)

	Speed	Parallelism	Bandwidth
① Cell to PB	?	?	?
② PB to I/F	~10 Gbps	256 bit*	320 GB/s

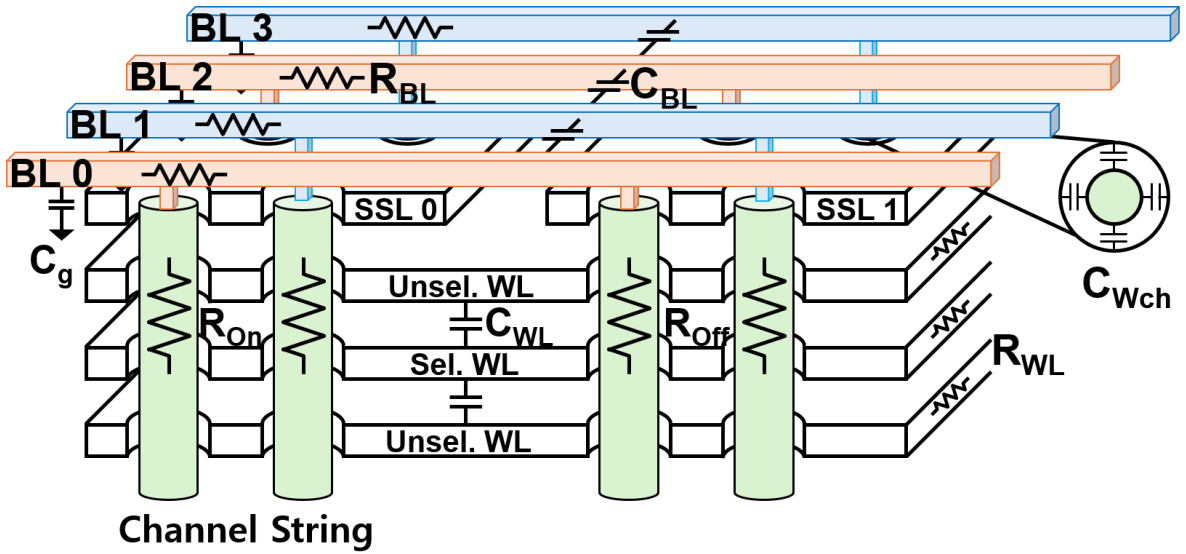
* 16 channel × 64bit ÷ 4 channel per die (OCP HBF spec)

Of course, sometimes interleaving may help hide tR and maintain high bandwidth.

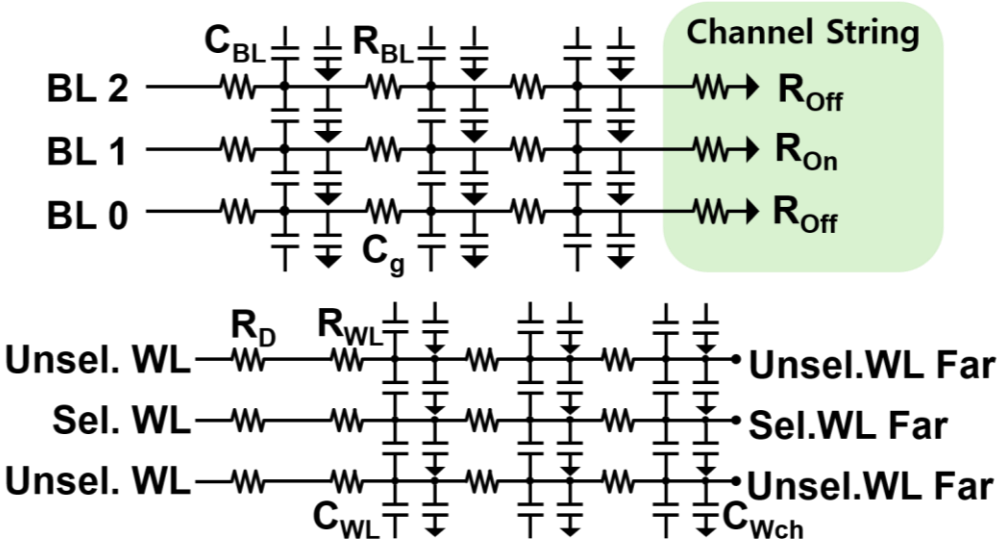
Low-latency 3D NAND Array Design

- Our 3D NAND array model
  - NAND array model is required to estimate read latency (tR) for different array sizes.
  - All RC components were extracted by TCAD simulation or calculated based on the real physical dimension of state-of-the-art 3D NAND array.

3D NAND array model



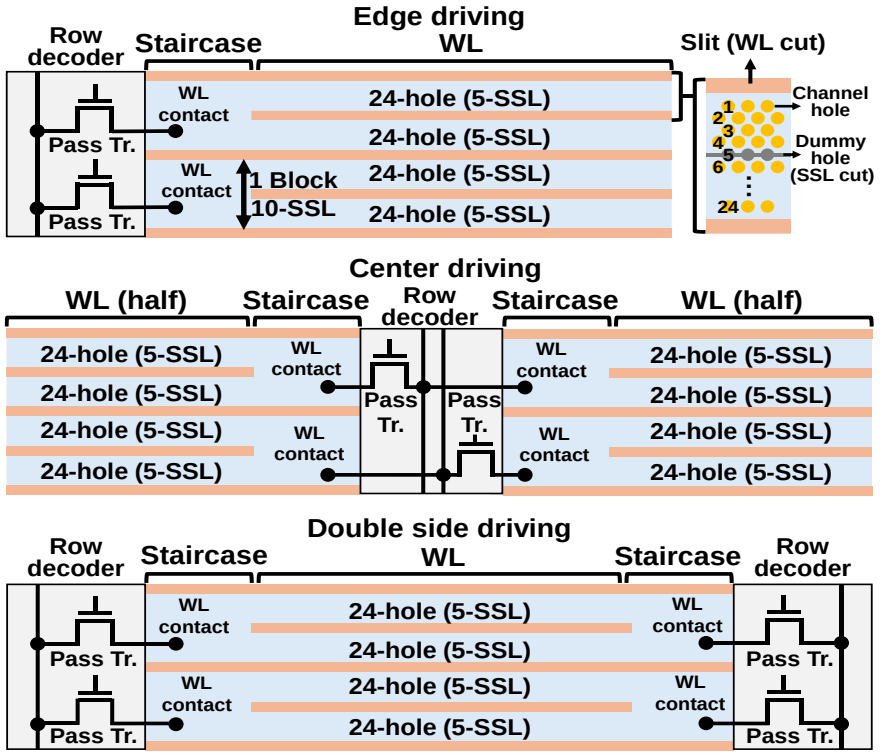
WL and BL RC model



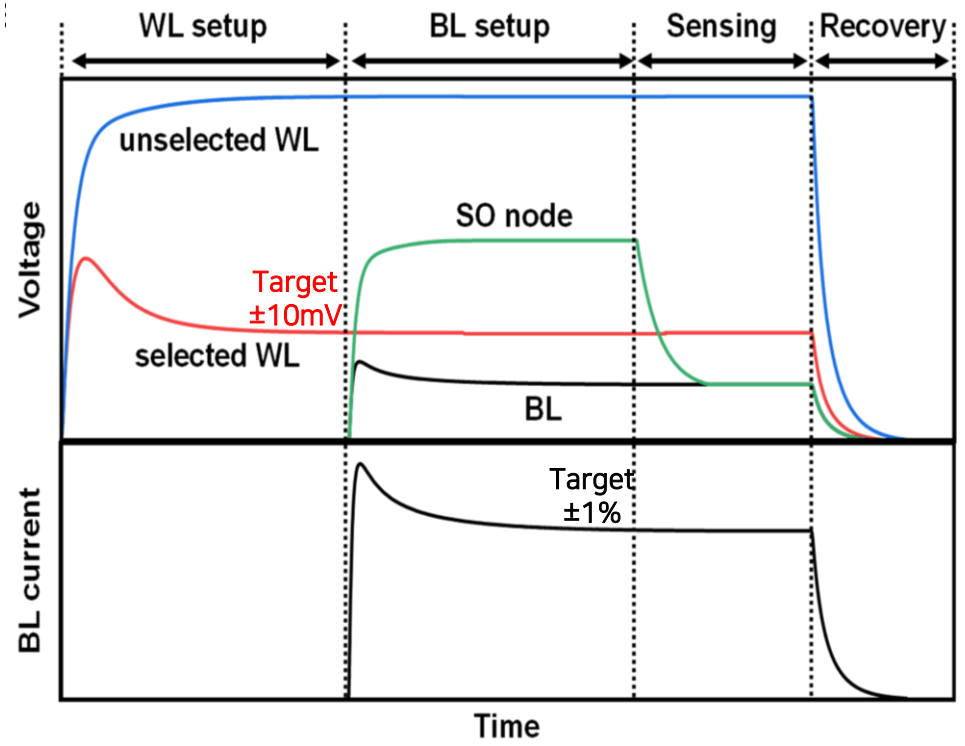
Low-latency 3D NAND Array Design

- tR estimation
  - Key factors: WL/BL RC, NAND string current, WL driving scheme

WL driving scheme



SLC read simulation

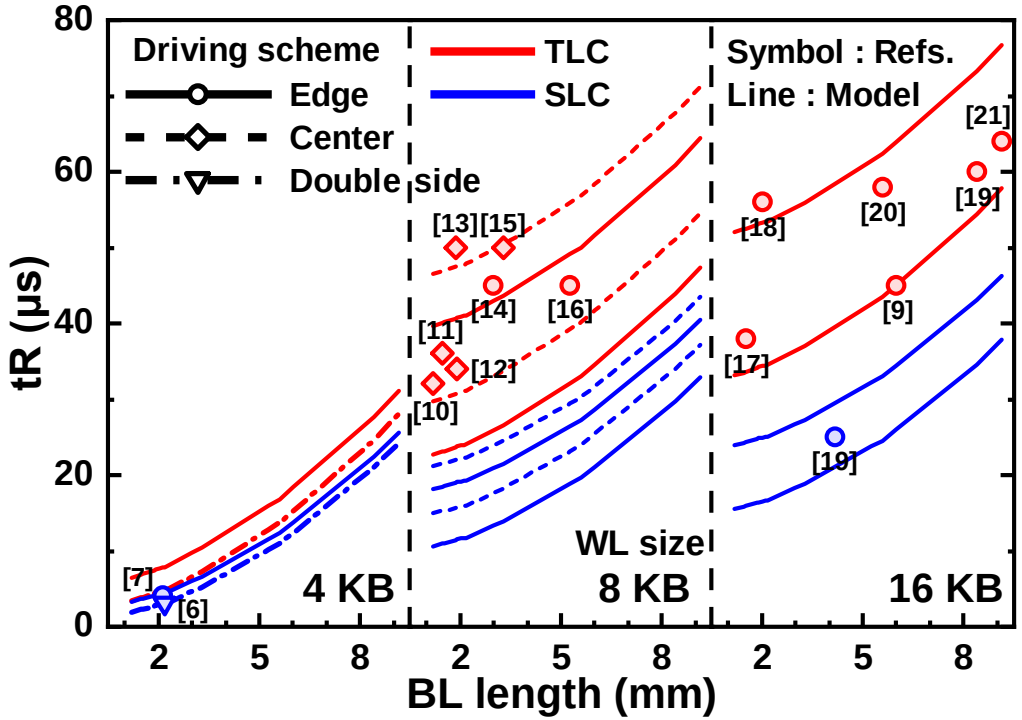


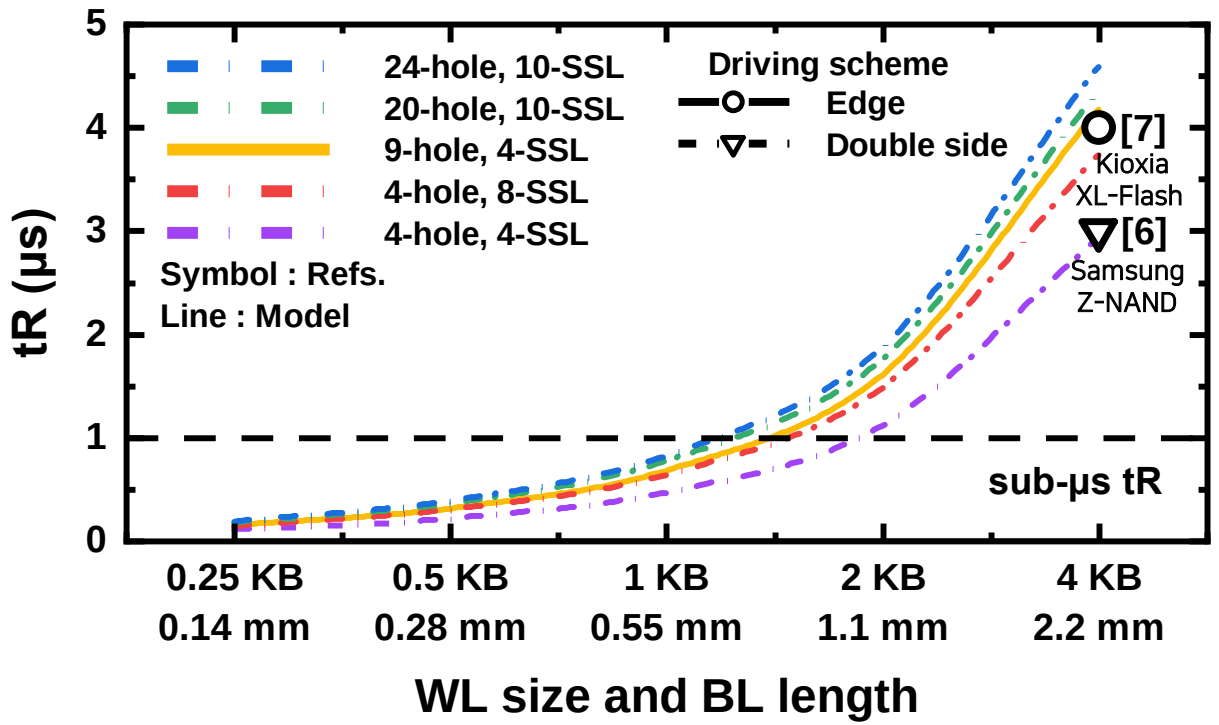
Low-latency 3D NAND Array Design

- tR estimation
  - Validated with industry products with various array sizes and driving schemes.
  - Sub-µs tR is achievable with WL size < 1kB, BL length 0.55mm, released WL pitch 62nm.

(recent product WL pitch ~ 40nm)

Estimated tR validation





Contents

1. Introduction

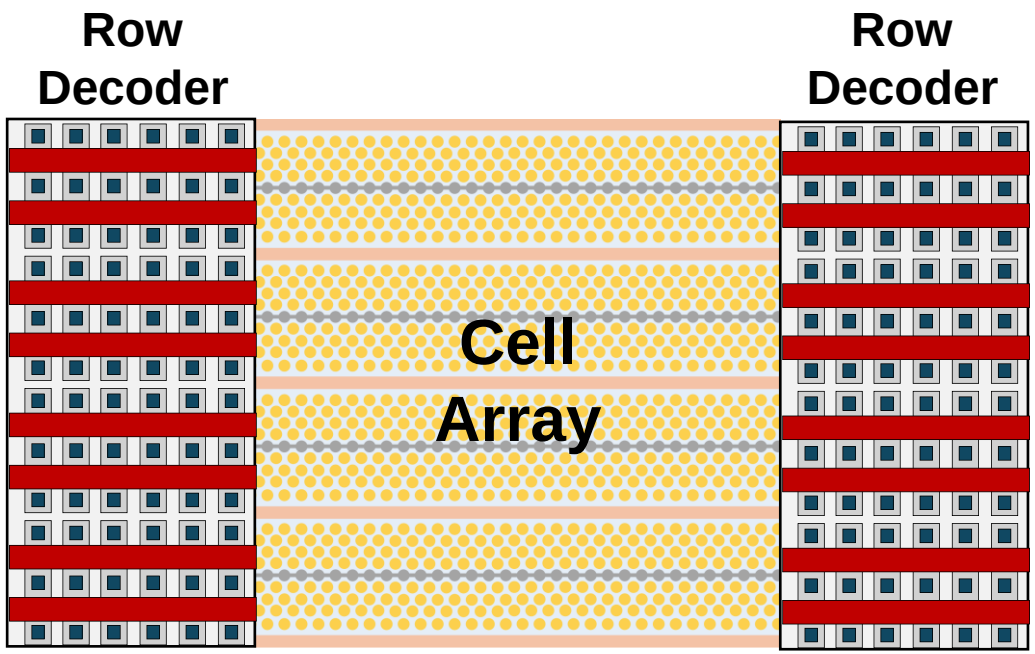
2. Low-latency 3D NAND Array Design

3. Split-Stacked Row Decoder Design using Hybrid Bonding Technology

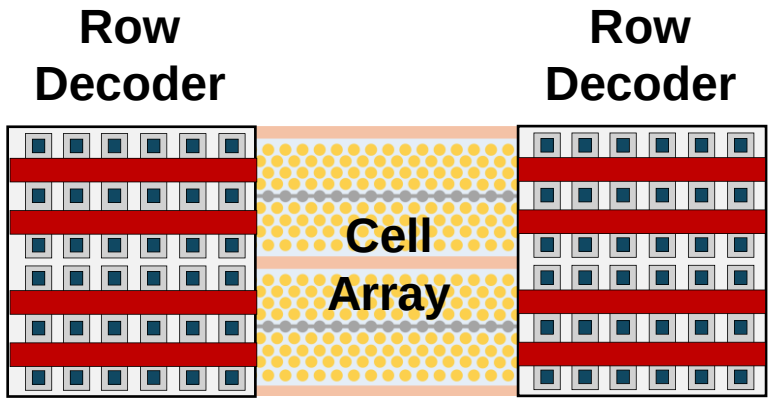
4. Design Space Exploration of High-bandwidth NAND Stack Architecture

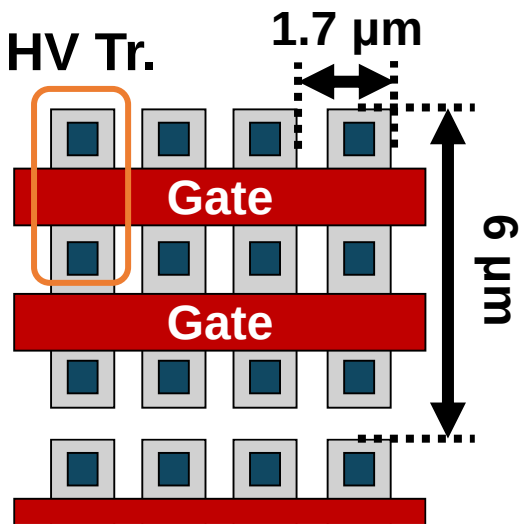
Split-Stacked Row Decoder (SSRD) design using Hybrid Bonding Technology

- Small array size lowers NAND density
  - Plane size ↓ → peripheral circuit area ratio ↑ → density ↓
    (Row Decoder, Page Buffer, WL staircase)
  - Especially, row decoder is not scalable.
    - Pass Tr. W/L ~ 1um because of high WL voltage (~25V)



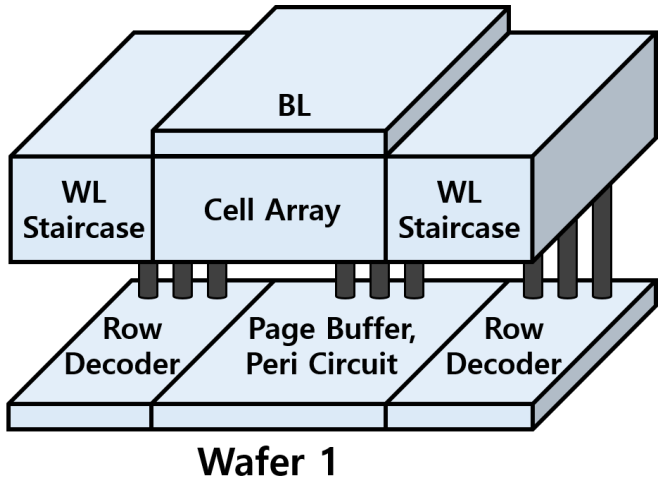
Cell array area **1/4**,
Row decoder area **1/2**



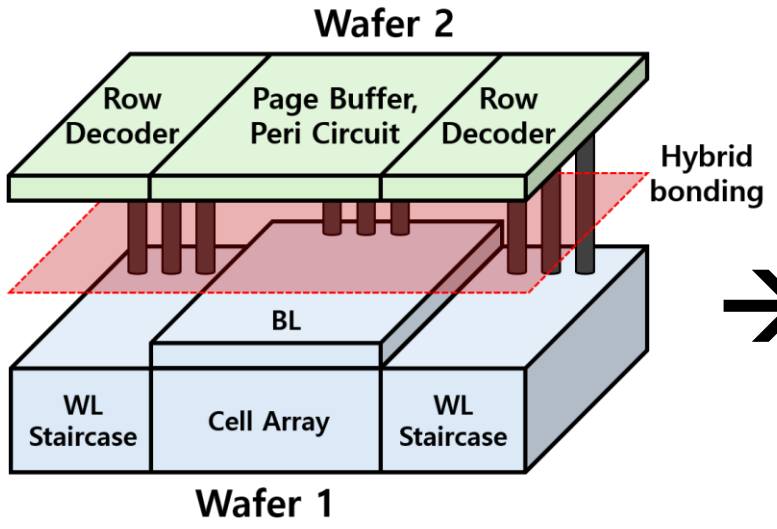


Split-Stacked Row Decoder (SSRD) design using Hybrid Bonding Technology

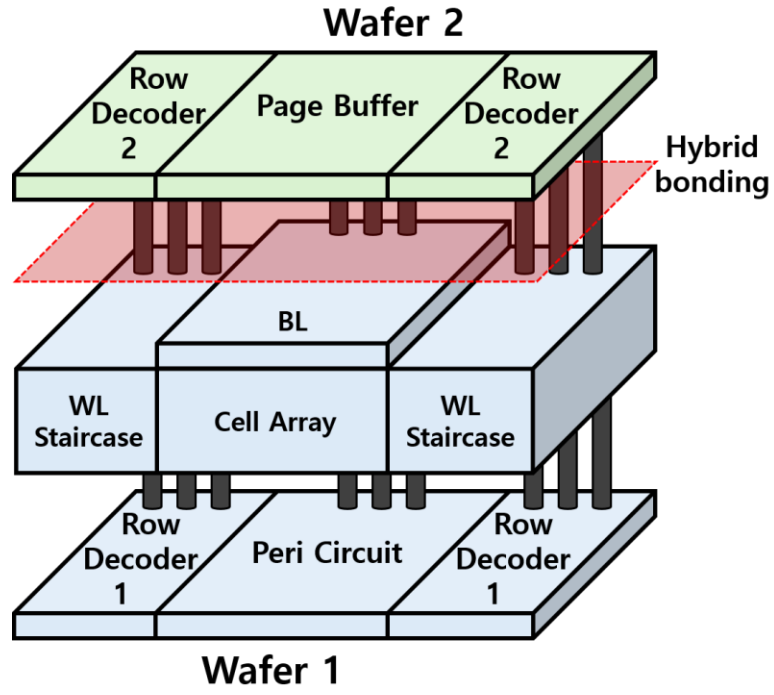
- **Concept of Split-stacked Row Decoder (SSRD) Architecture**
  - Split the row decoder across two wafers : COP style buried routing + CBA style wafer bonding
  - Row decoder area reduction without HV transistor scaling



Cell Over Peri (COP*) Structure



CMOS directly Bonded to Array (CBA) Structure

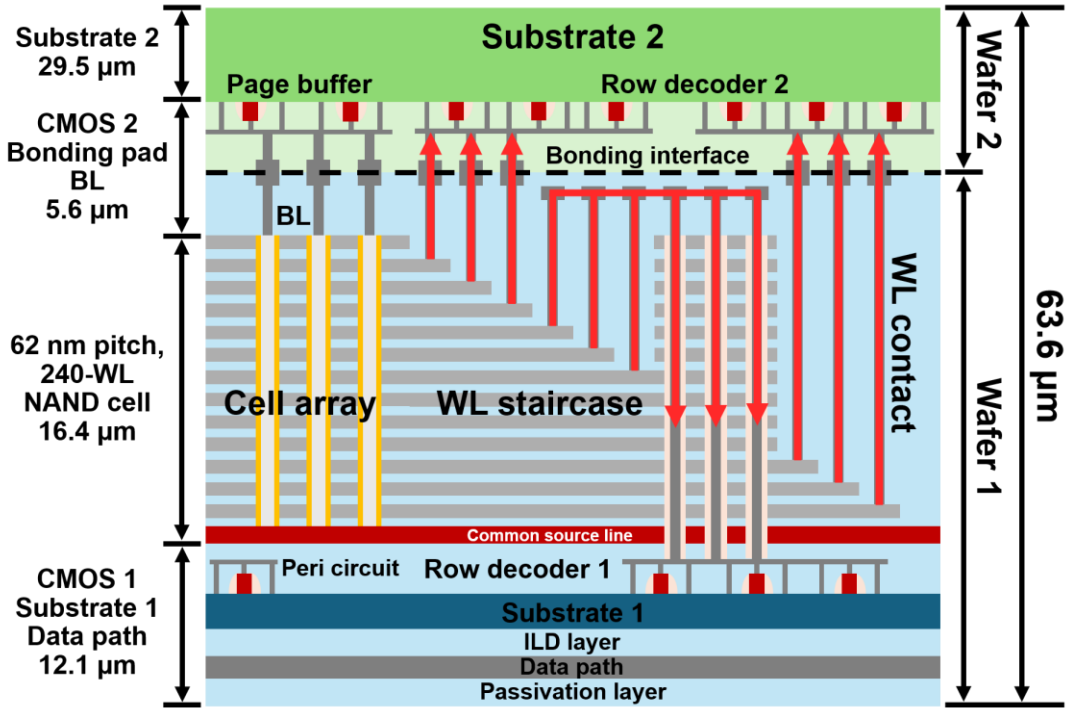


Split Stacked Row Decoder (SSRD) Structure

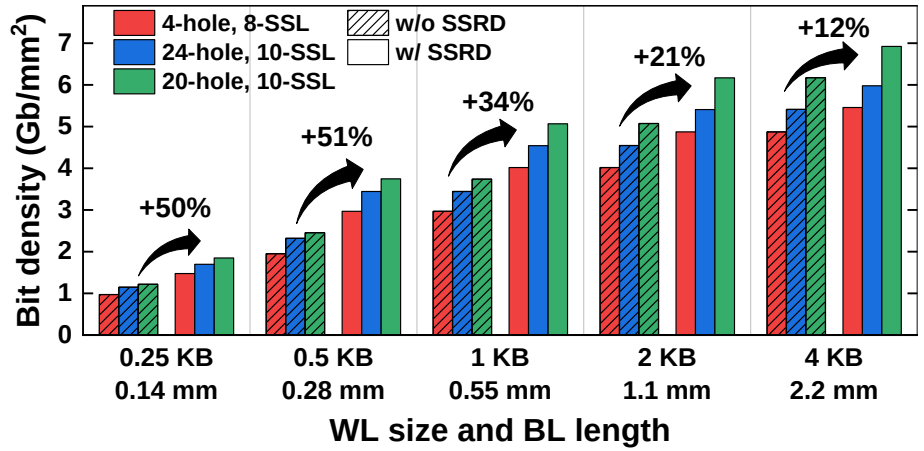
COP* = PUC (Peri Under Cell) = CUA (CMOS under Array)

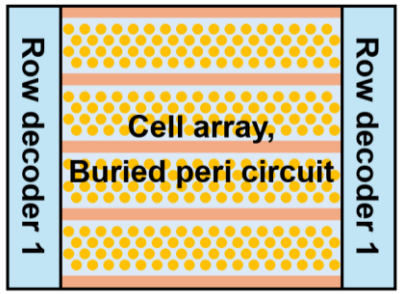
Split-Stacked Row Decoder (SSRD) design using Hybrid Bonding Technology

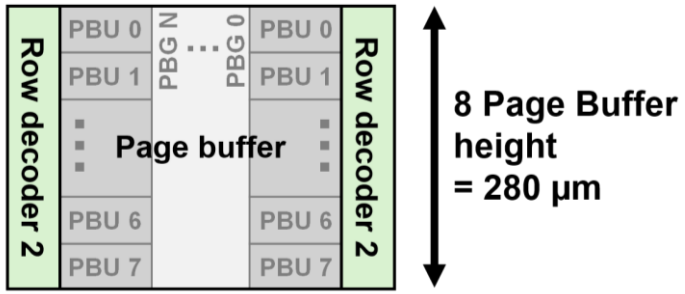
- SSRD structure and bit density increase
  - Place row decoder (pass Tr.) on top & under the WL staircase region.
  - As the array size smaller, bit density improvement of SSRD increases.



Substrate thickness can vary with thinning/bonding/electrical stability.







서울과학기술대학교 SEOULTECH SEOUL NATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY
SDCL Semiconductor Devices and Circuits Laboratory

Contents

1. Introduction

2. Low-latency 3D NAND Array Design

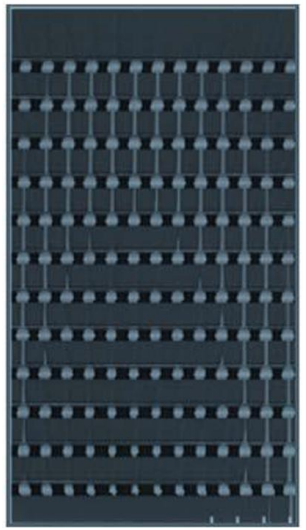
3. Split-Stacked Row Decoder Design using Hybrid Bonding Technology

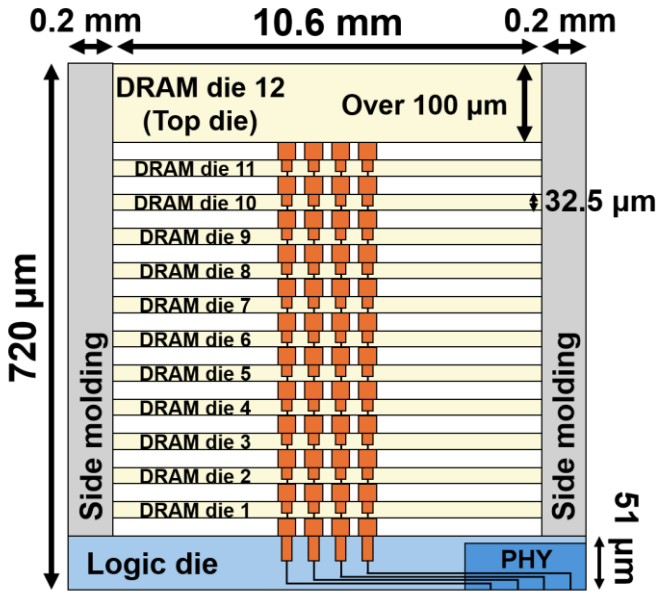
4. Design Space Exploration of High-bandwidth NAND Stack Architecture

Design Space Exploration of High-bandwidth NAND Stack Architecture

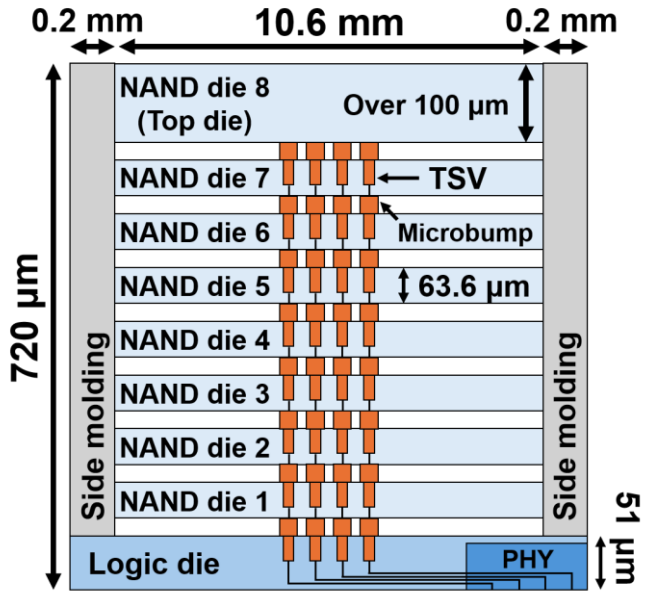
- **NAND Die Stacking**
  - Same core die size, TSV area, die-to-die gap, stack height with HBM3
    (10.6mm × 10.6mm) (MR-MUF 13um) (720um)
  - 8-hi NAND stack (vs 12-hi HBM3) because of 3D NAND cell height, two-wafer bonding.

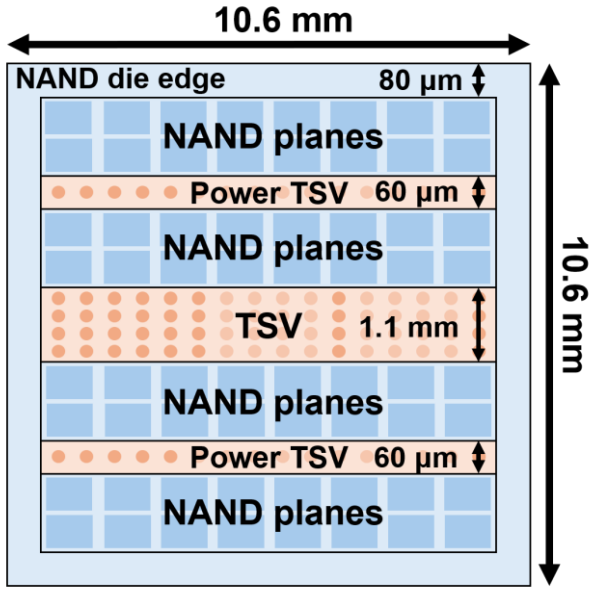
ISSCC 2022, SK Hynix





HBM3 12-hi : 24 GB





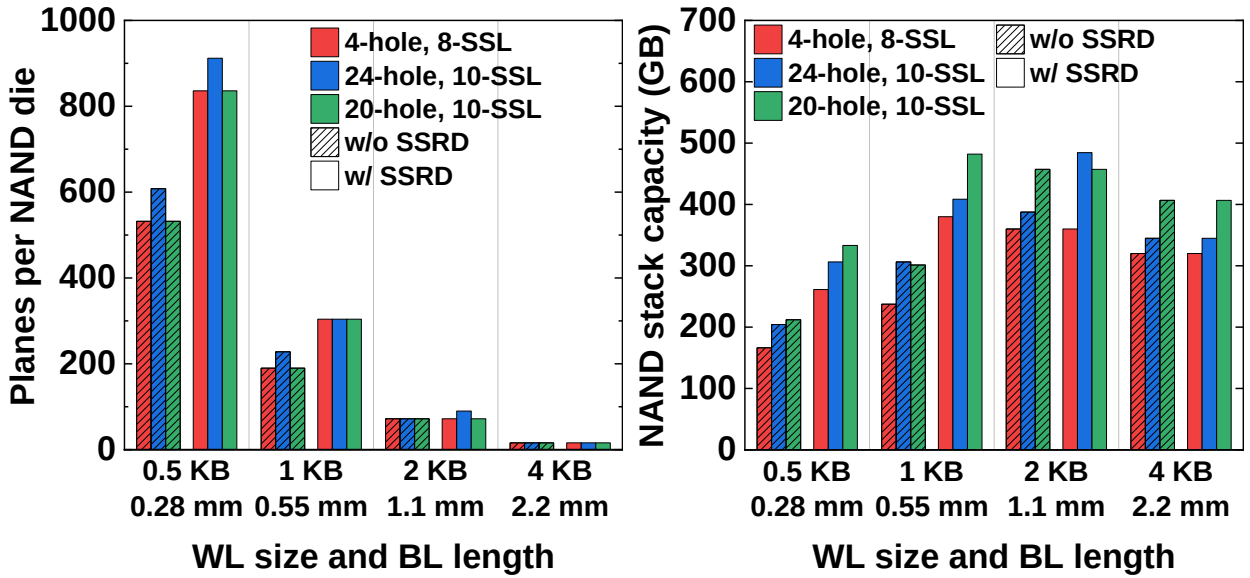
High Bandwidth NAND Stack 8-hi

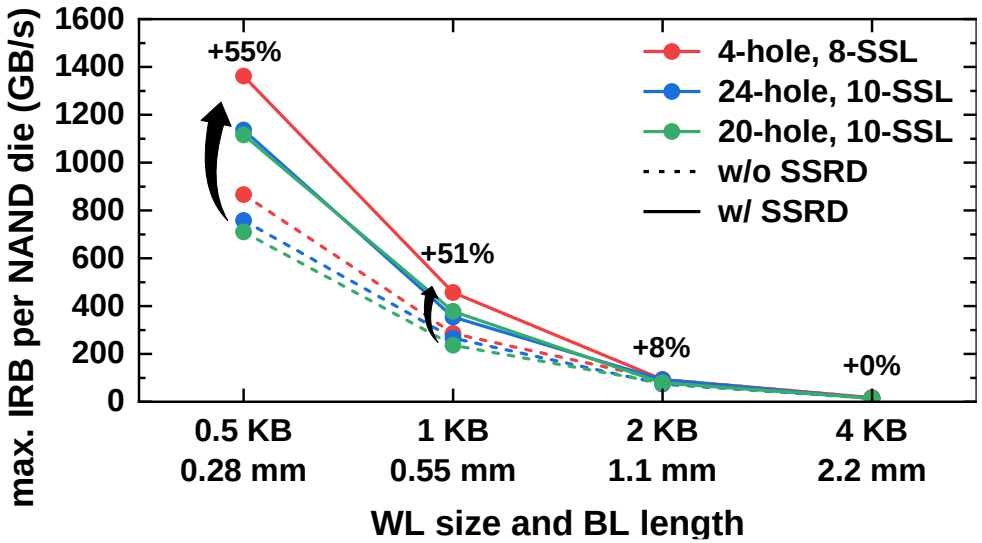
Design Space Exploration of High-bandwidth NAND Stack Architecture

- **Capacity and bandwidth High-bandwidth 3D NAND**
  - Stack capacity : 200~500 GB. Over 10× larger capacity than HBM3 (12-hi 24 GB)
  - Max. IRB (Internal read bandwidth) : 400~1400 GB/s per NAND die

    (1KB WL) (0.5KB WL)

    Max. IRB = (1/tR) × WL Size × number of Planes, assuming all PB/planes in a die working at the same time.
  - SSRD architecture improves max. IRB up to 55%





Conclusion

- **Overall NAND design should be tailored to the application.**
  - Model weights? KV cache?
  - Application-specific cell design, array size, peripheral circuit, wafer integration process are required.

- **Thermal-aware NAND stack design is required.**
  - NAND stack has different thermal conductivity from HBM.
  - NAND cell data retention, cooling of the whole system must be considered.


Thank you