Agile-DRAM

Agile Trade-Offs in Memory Capacity, Latency, and Energy for Data Centers

Jaeyoon Lee†, Wonyeong Jung†, Dongwhee Kim†,
Daero Kim*, Junseung Lee† , and Jungrae Kim†

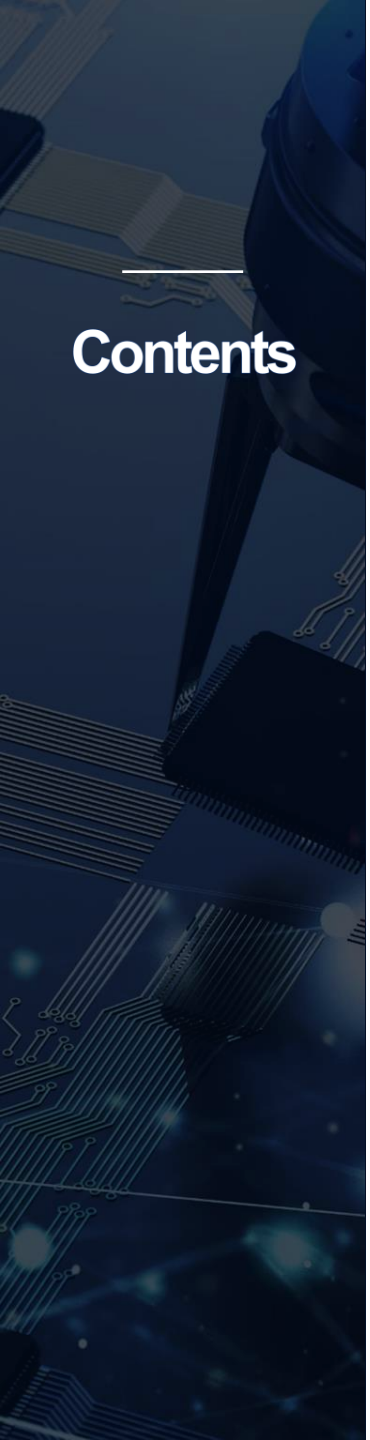
†Sungkyunkwan University, *Samsung Electronics

2024.03.06

SUNG KYUN KWAN UNIVERSITY(SKKU)

SAMSUNG

High Performance Computer Architecture



Contents

☑One of the most expensive components

2x Intel Sapphire Rapids Server	
Component	AI Server
CPU	$ 1,850
8 GPU + 4 NVSwitch Baseboard	$ -
Memory	$ 3,930
Storage	$ 1,536
SmartNIC	$ 654
Chassis (Case, backplanes, cabling)	$ 395
Motherboard	$ 350
Cooling (Heatsinks+fans)	$ 275
Power Supply	$ 300
Assembly and Test	$ 495
Markup	$ 689
Total Cost	$ 10,474
DRAM BOM %	37.5%
NAND BOM %	14.7%
Memory BOM %	52.2%

<Bill Of Materials (BOM) of a data center server>

Source: https://www.semianalysis.com/p/ai-server-cost-analysis-memory-is

DRAM Under-utilization

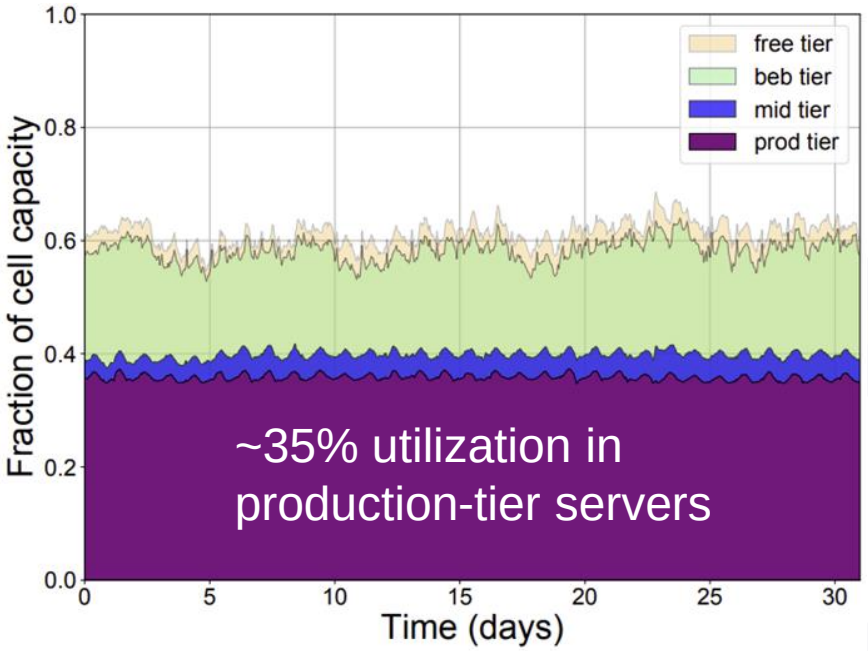
☑ Data centers under-utilize DRAM

- Due to factors like
  - infrastructure over-provisioning
  - sub-optimal workload scheduling
  - fixed system configurations

Metric	Statistics of all jobs			
	Median	Mean	Max	Std Dev
	CPU Jobs			
Allocated nodes	1	6.51	1713	37.83
Job duration (hours)	0.16	1.40	90.09	3.21
CPU util (%)	35.0	39.98	100.0	34.60
DRAM util (%)	13.29	22.79	98.62	23.65

13% (median) utilization in supercomputers

<Resource utilization in Perlmutter supercomputer, Li+, “Analyzing resource utilization in an HPC system: a case study of NERSC’s Perlmutter," High Performance Computing, 2023>

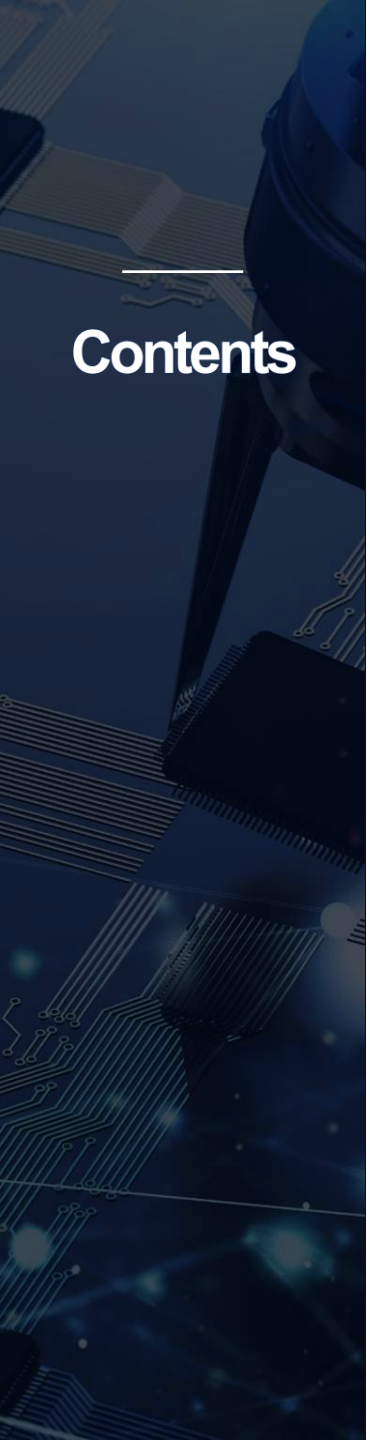


<Memory utilization in Google data centers, Tirmazi+, "Borg: the Next Generation," EuroSys’20>

☑ Memory pooling and disaggregation

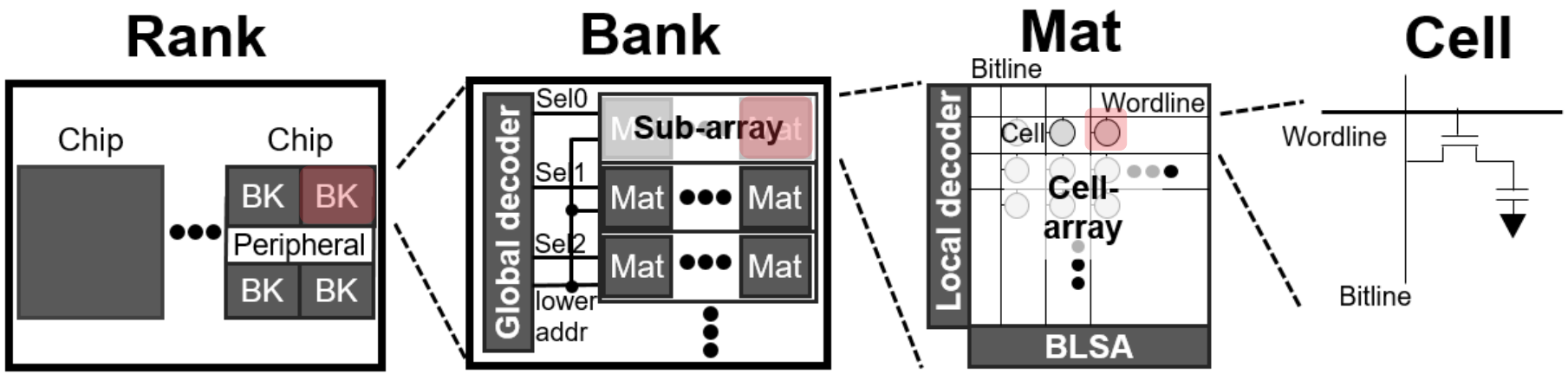
- Share memory to reduce the costs
- But slows down the system due to higher latencies
  - memory pooling: >100ns
  - memory disaggregation: a few us

Instead, we explore an opportunity
to trade under-utilized memory capacity to speed up applications.



Contents

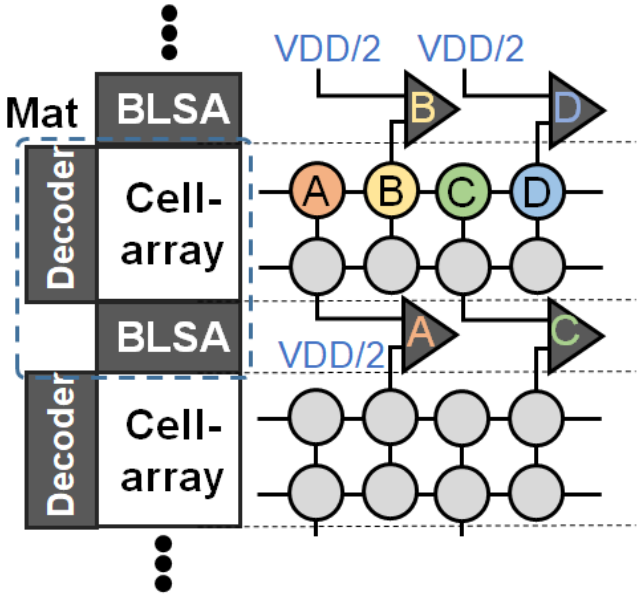
DRAM Hierarchy



Open-bitline Structure

☑ For area efficiency

- Two MATs share a BitLine SenseAmplifier (BLSA)
  - Only one MAT is activated at a time
  - The other is precharged and provides VDD/2 as a reference voltage for sensing



Contents

☑ A new DRAM architecture

- To trade memory capacity for higher performance and lower energy

☑ Support three modes

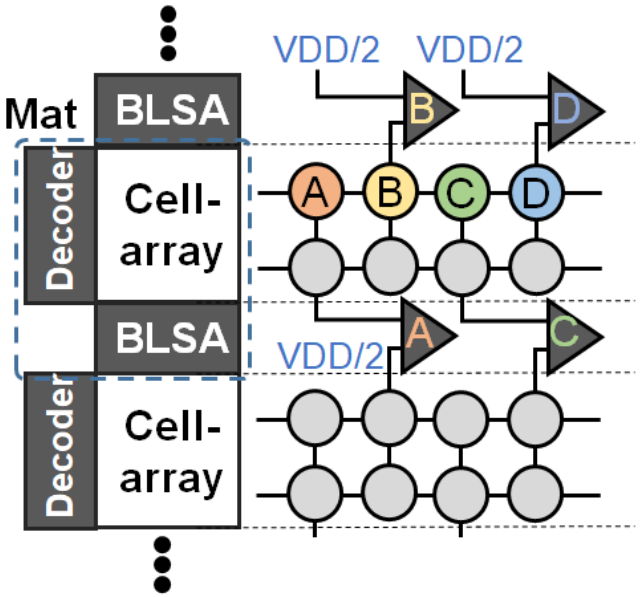
- Max Capacity (MC)
- Low Latency (LL)
- Low Power (LP)

☑ Agile switching between the modes (~300ns transition time)

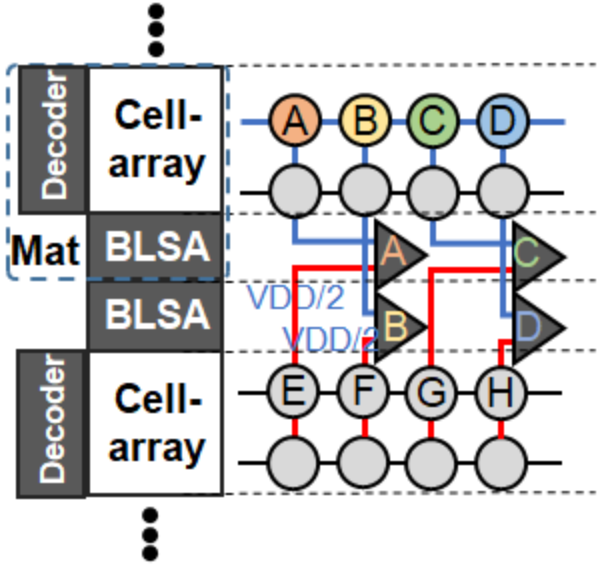
☑ With little hardware overheads

DRAM Structure: Mirrored MAT

☑ A minimal change to DRAM structure



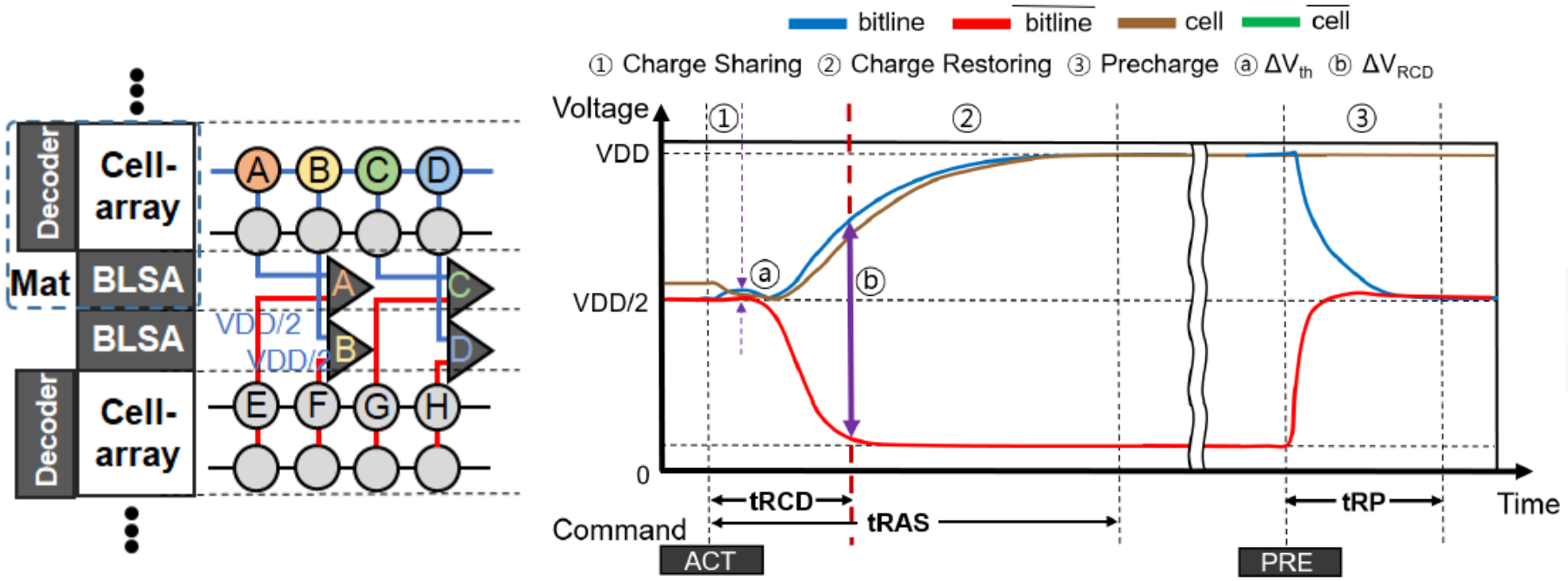
<Conventional open-bitline structure>



<Mirrored MAT structure>

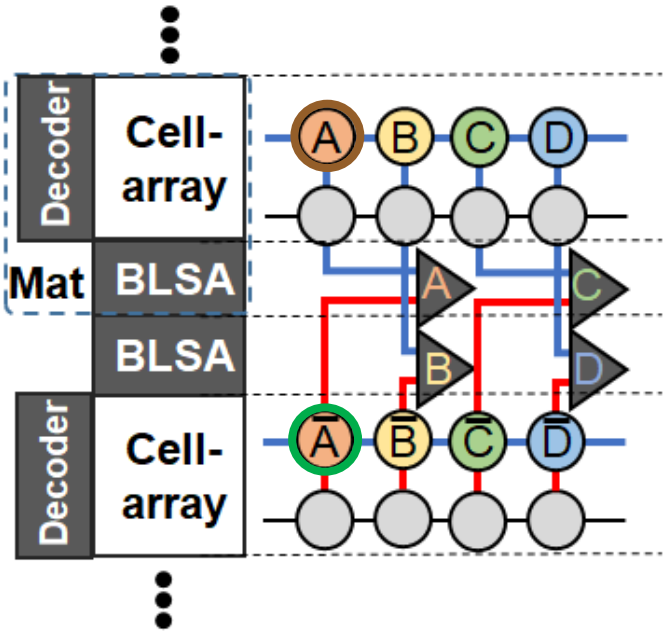
☑ Max-capacity mode (MC)

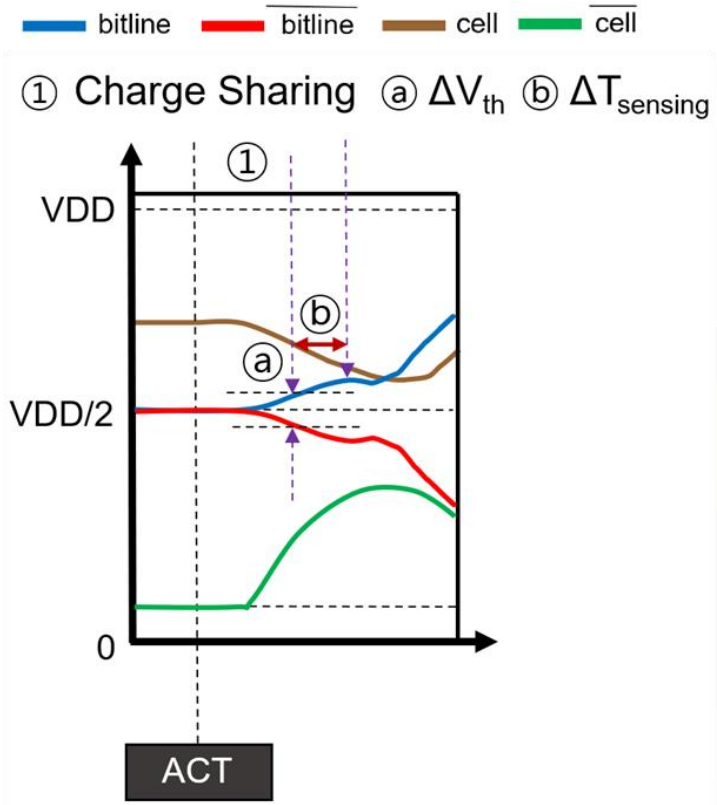
- Operates like conventional DRAM to provide the full capacity



☑ Low-latency mode (LL)

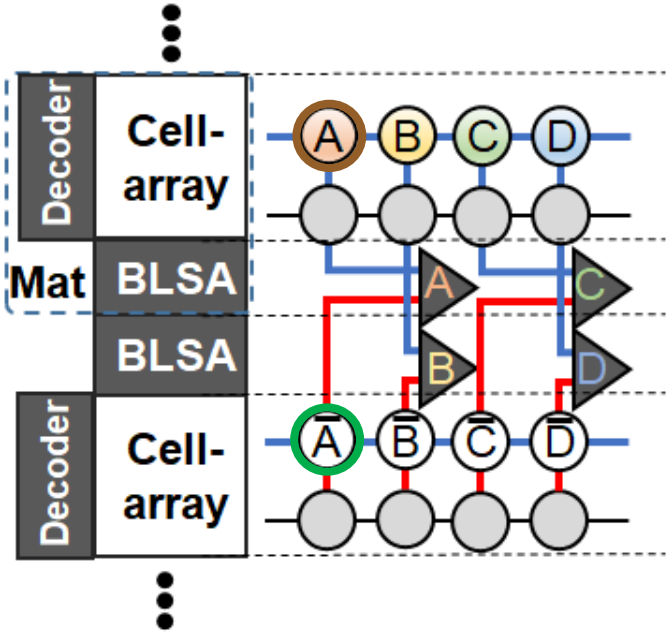
- Senses data more quickly, by storing complementary data

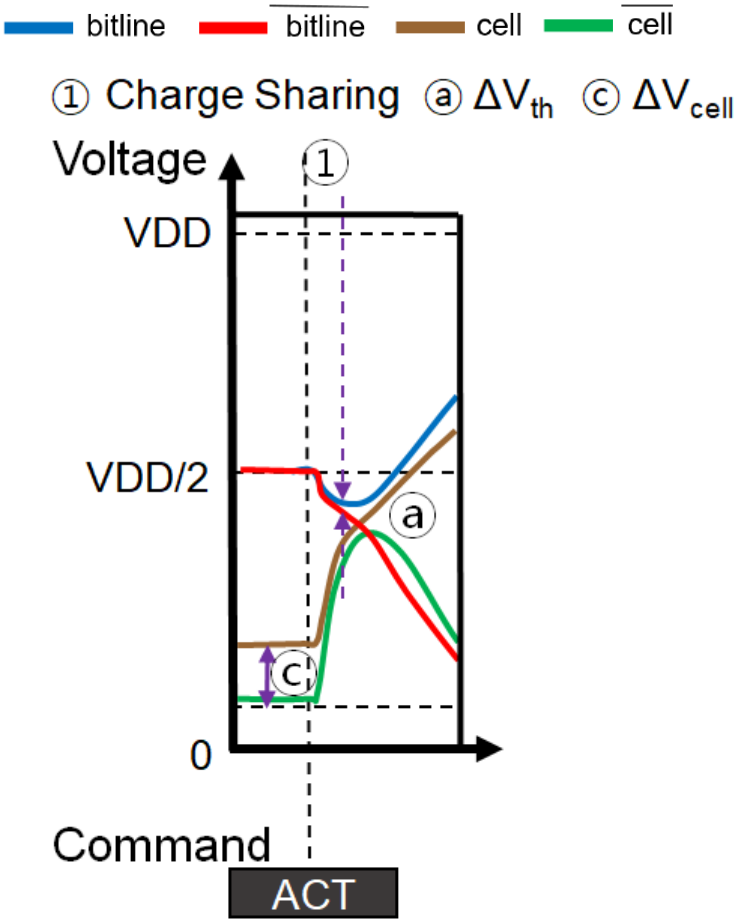




☑ Low-power mode (LP)

- Allows extended refresh intervals

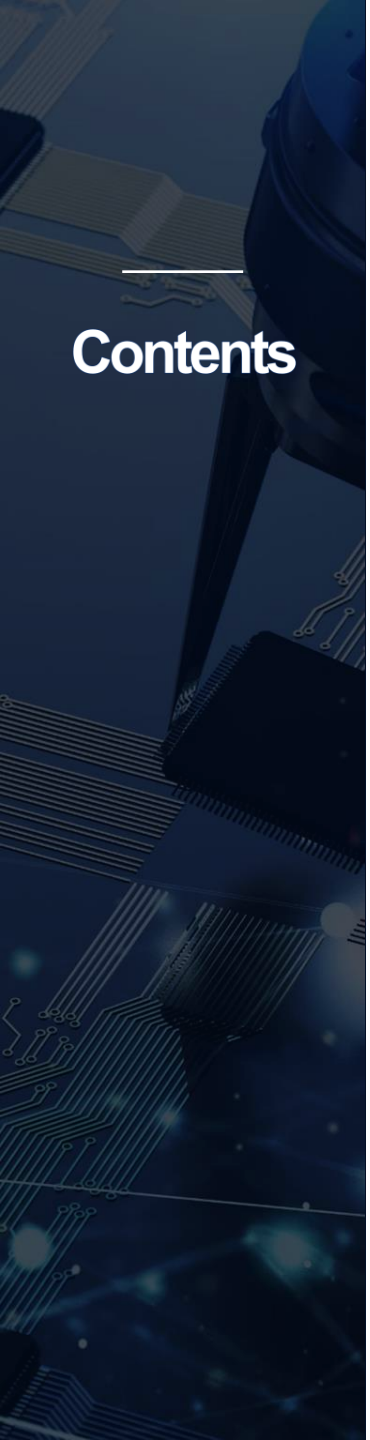




☑ DRAM timing parameters

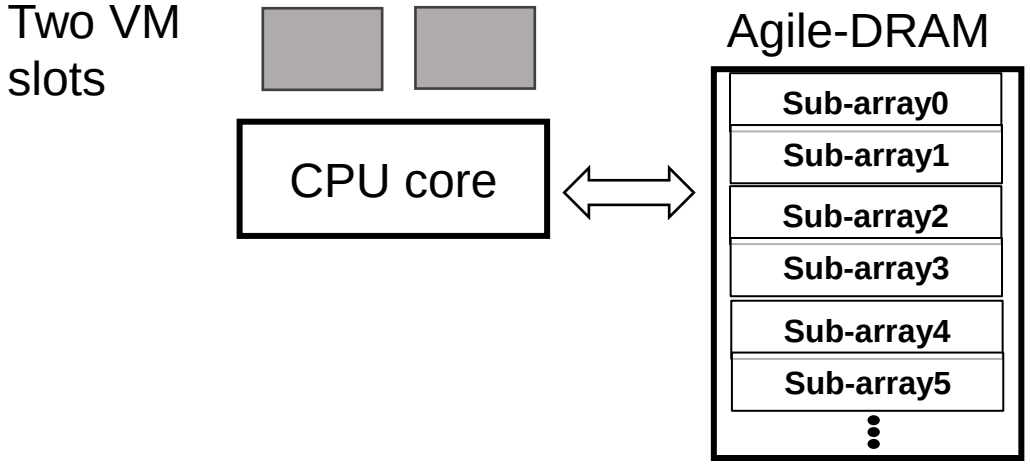
- LL (Low-latency): lower access latency (tRCD, tRAS)
- LP (Low-power): longer refresh interval (tREFI)

Mode		tRCD (ns)	tRAS (ns)	tRP (ns)	tWR (ns)	tRFC (ns)	tREFI (µs)
Baseline / Agile-MC		13.8	39.4	15.5	12.5	260	7.8
Agile-DRAM	LL	9.8	18.7	15.4	7.2	161	7.8
	LP	12.3	21.4	15.5	7.2	175	54.6

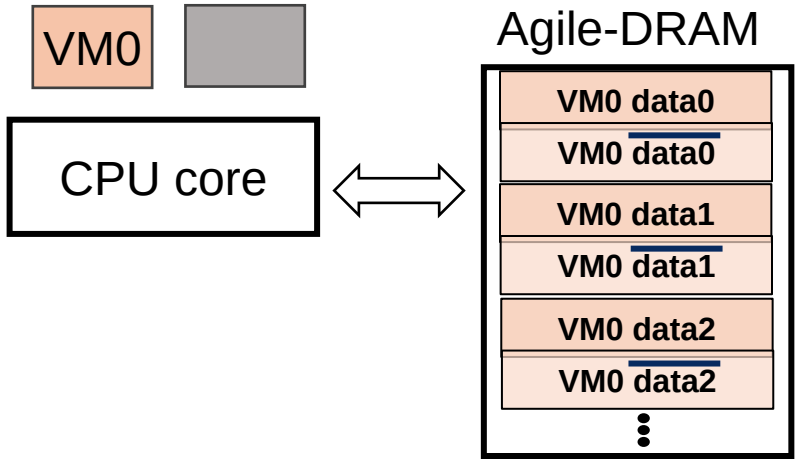


Contents

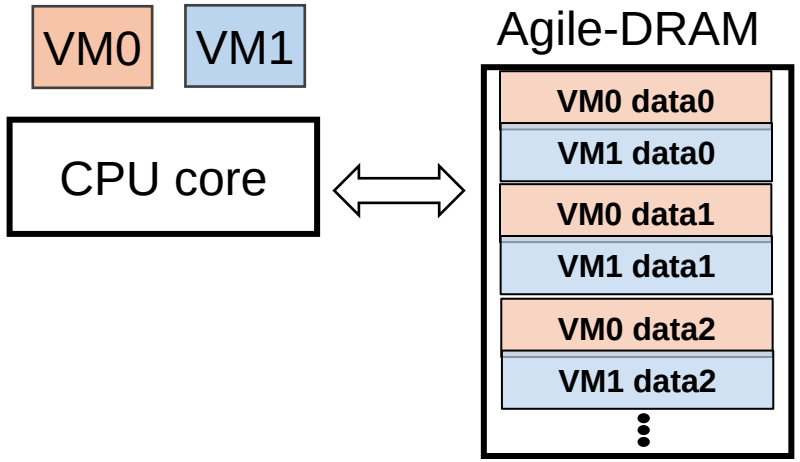
An Infrastructure-as-a-Service (IaaS) server

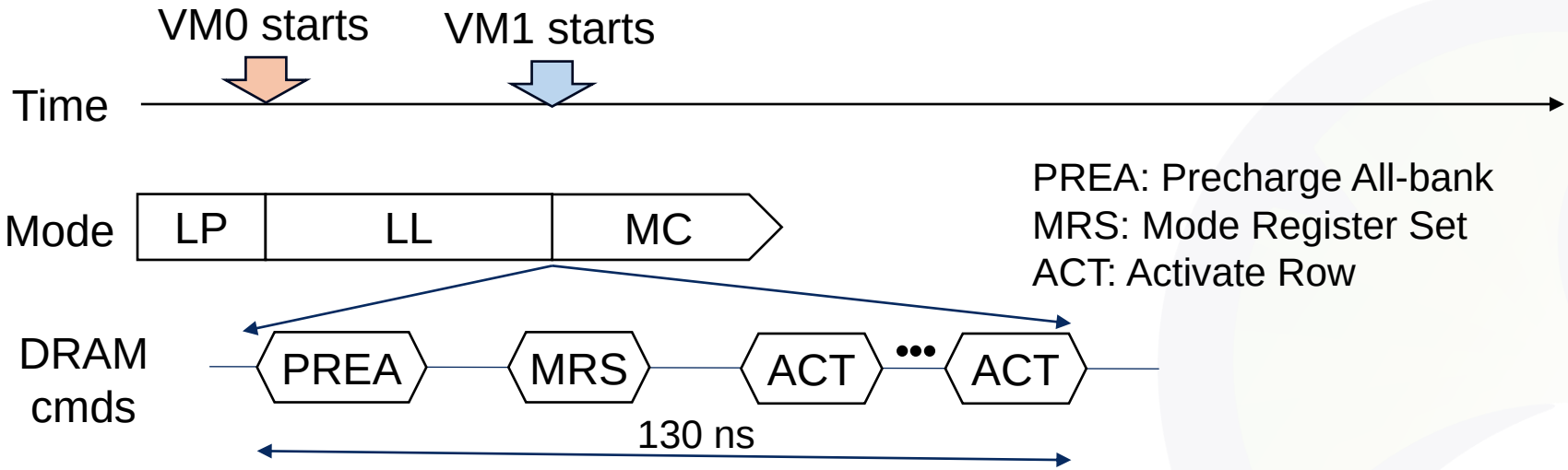


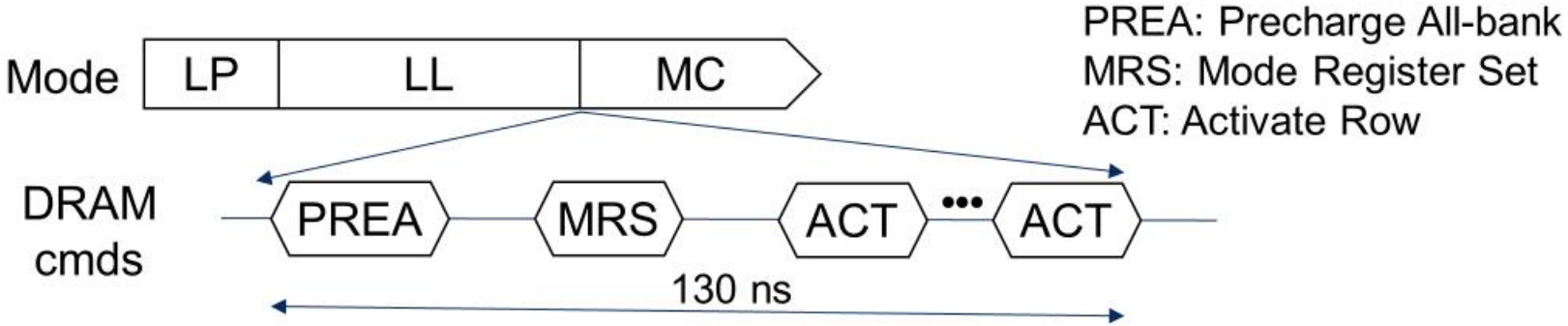
☑ A VM starts in the LL mode

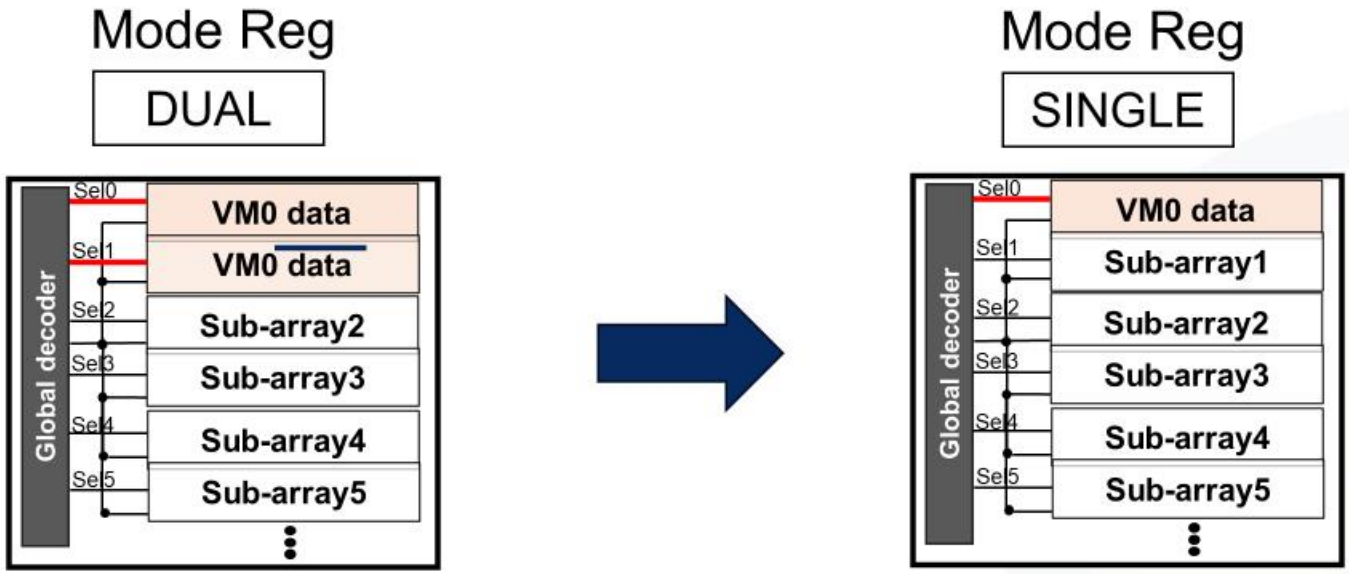


☑ Another VM starts

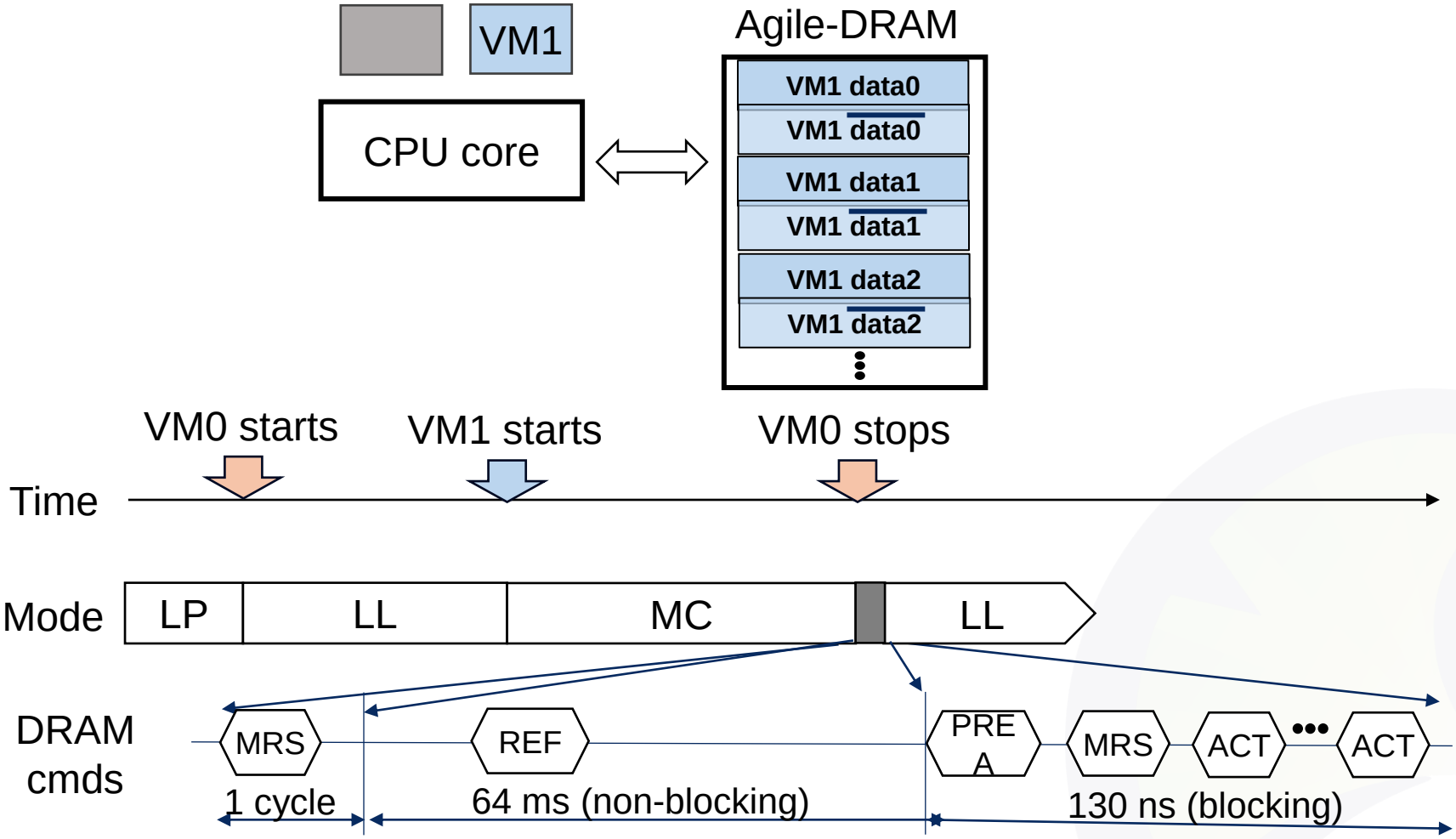


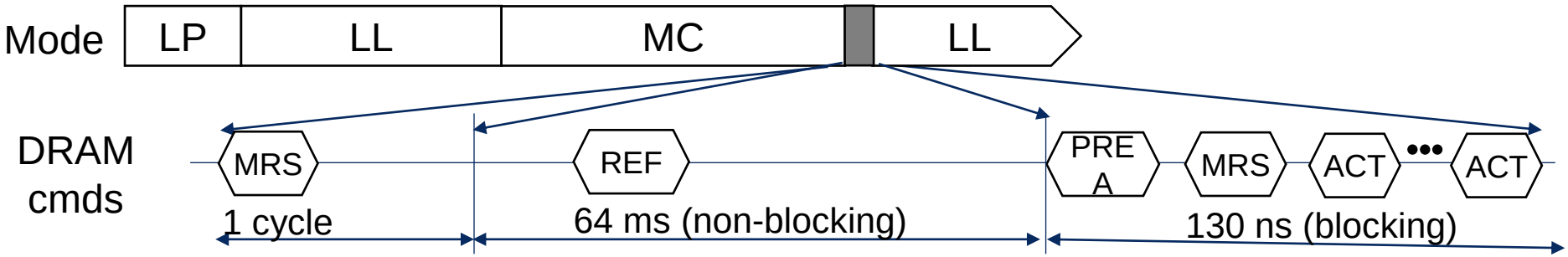



Mode
LP
LL
MC
DRAM cmds
PREA
MRS
ACT
ACT
130 ns
PREA: Precharge All-bank
MRS: Mode Register Set
ACT: Activate Row

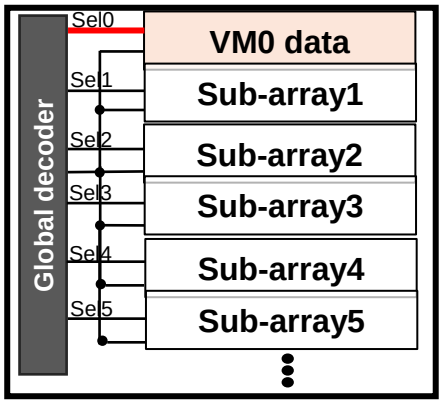

Mode Reg
DUAL
Global decoder
Sel0
Sel1
Sel2
Sel3
Sel4
Sel5
VM0 data
VM0 data
Sub-array2
Sub-array3
Sub-array4
Sub-array5
Mode Reg
SINGLE
Global decoder
Sel0
Sel1
Sel2
Sel3
Sel4
Sel5
VM0 data
Sub-array1
Sub-array2
Sub-array3
Sub-array4
Sub-array5

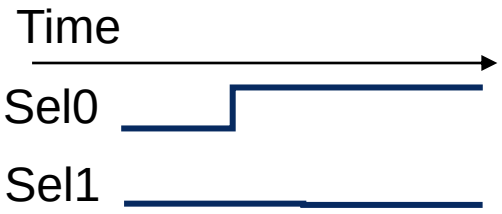
☑ The first VM stops



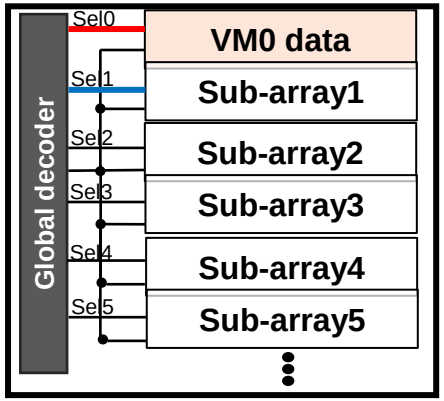


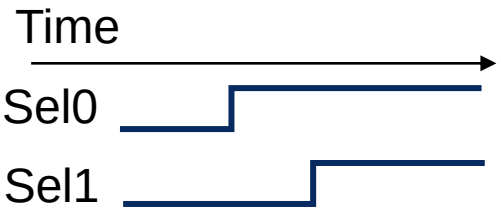
Mode Reg
SINGLE



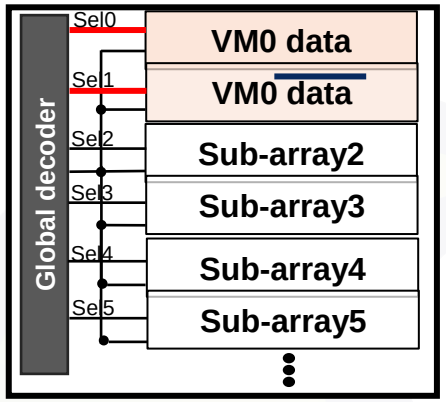


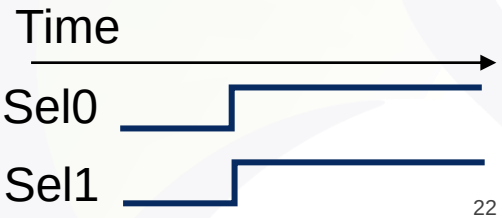
Mode Reg
TRANS

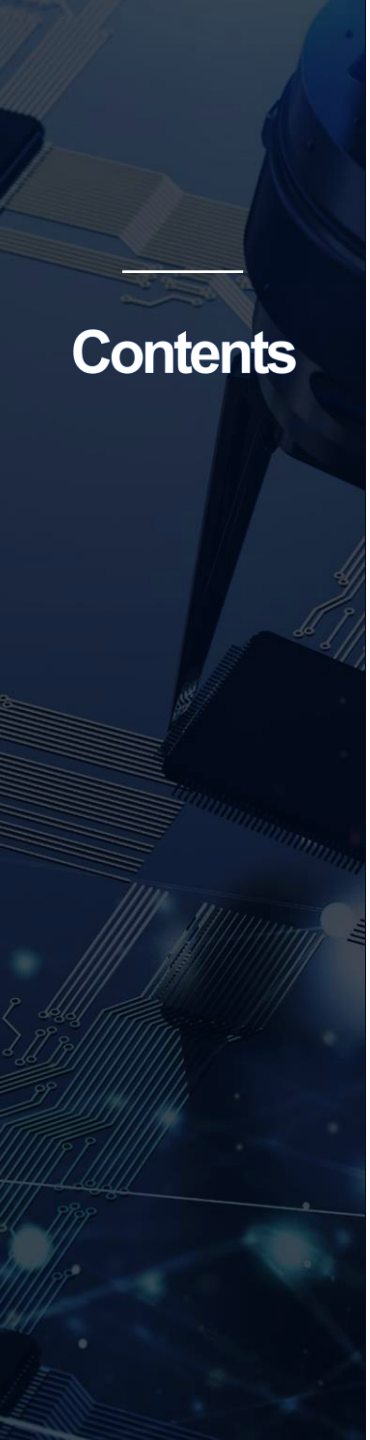




Mode Reg
DUAL







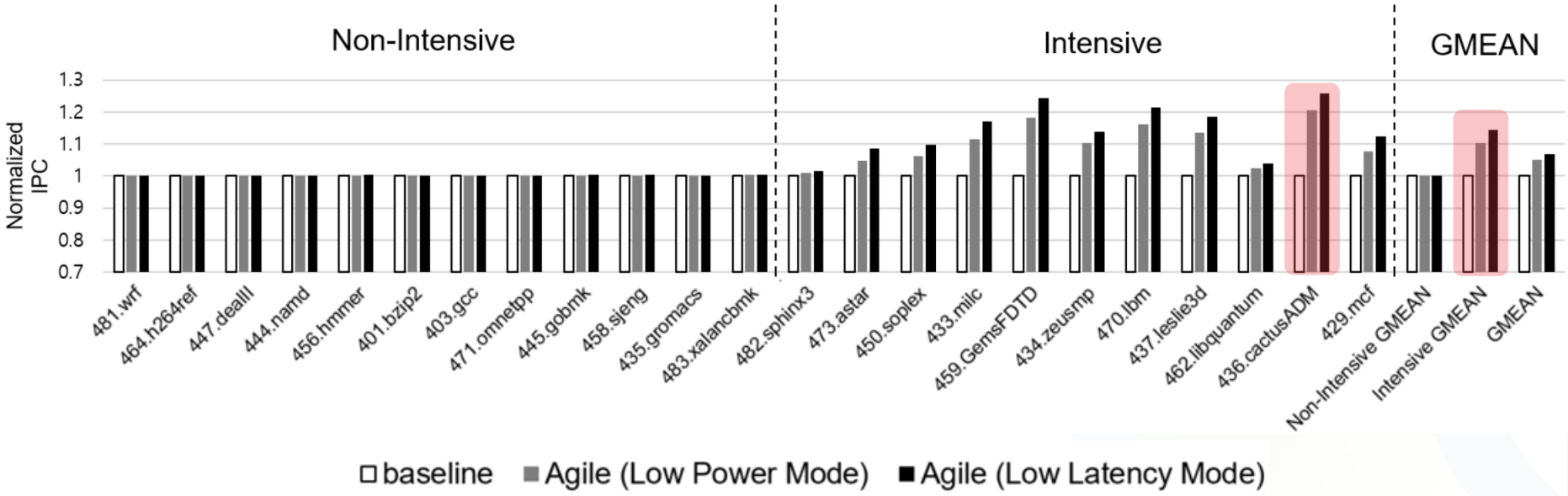
Contents

Single-Core Performance

☑ LL's speed-up over the baseline (or MC)

- **25.8%** (max.)
- **14.3%** (memory-intensive geomean)
- 6.9% (total geomean)

- Ramulator
- Single-core
- DDR4-2400R, 4Gb,
- Benchmarks: SPEC CPU2006



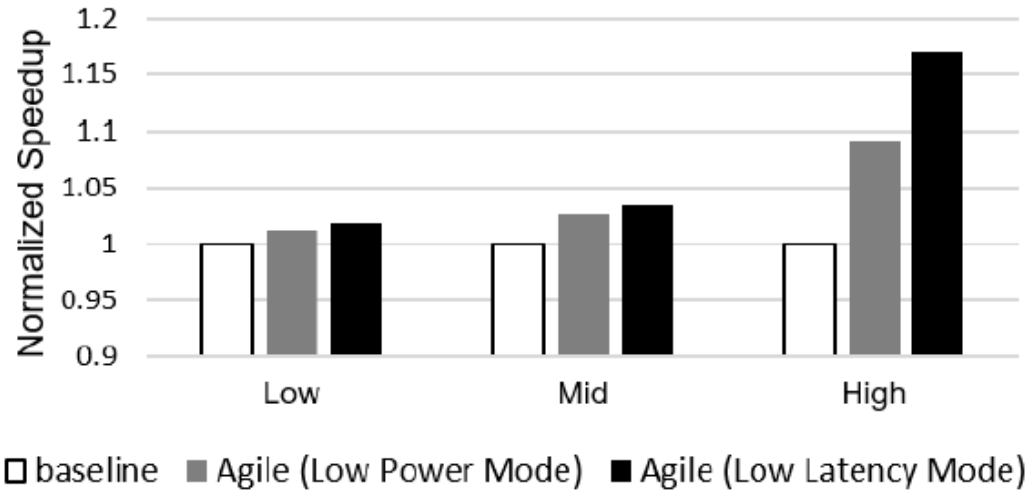
Multi-core Performance

☑ LL's speed-up over the baseline (or MC)

- **17.0%** (max.)
- 7.3% (total geomean)

- Gem5, Ramulator
- 4-core
- DDR4-2400R, 4Gb,
- Benchmarks: SPEC CPU2006

✓ Low: Low memory intensity mix
✓ Mid: Medium memory intensity mix
✓ High: High memory intensity mix

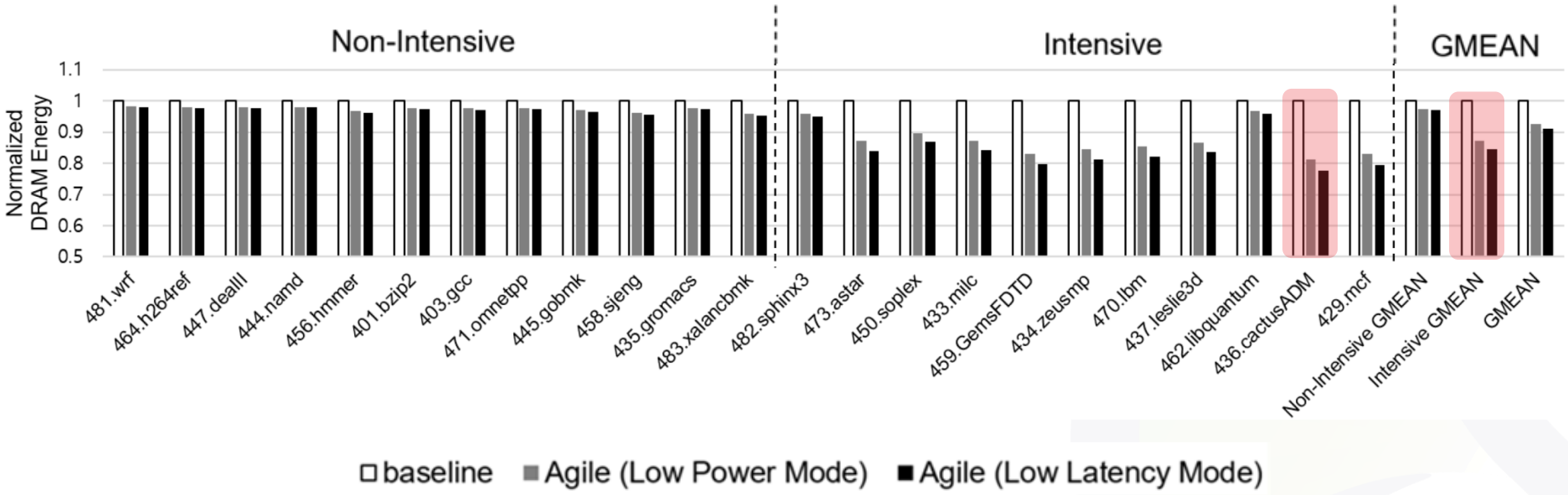


DRAM Energy Consumption

☑ LL's reduction over the baseline (or MC)

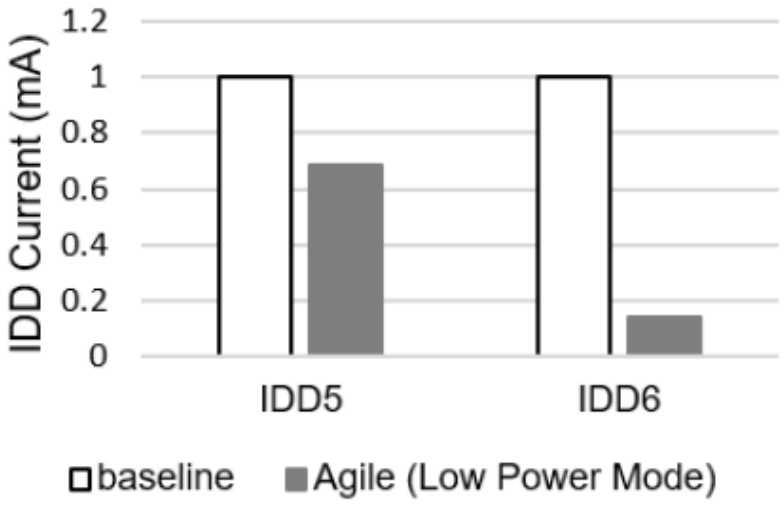
- **22.4%** (max.)
- **15.5%** (memory-intensive geomean)
- 9.0% (total geomean)

- DRAMPower
- Single-core
- DDR4-2400R, 4Gb,
- Benchmarks: SPEC CPU2006



☑ Reduced standby power (LP)

- IDD5 (refresh current): 31.6%
- IDD6 (self-refresh current): 85.7%



Summary

☑ Motivation

- Memory under-utilization

☑ Idea: Agile-DRAM

- Trade memory capacity for higher performance and lower energy with little hardware overheads

☑ Evaluation

- performance (6.9%), energy (9.0%), chip size (≈0%)

Thank you

Q&A

SUNG KYUN KWAN UNIVERSITY(SKKU)

SAL
SCALABLE ARCHITECTURE LAB

Hardware Overheads

☑ Negligible overhead

- an additional area of **75um²** (about 30 NAND2 gates)
  - can be easily accommodated in unused space on a **~100mm² DRAM die**

Simulation Configuration

☑ Simulation configuration

- single-core
- multi-core (4 core)

Processor	1 or 4 core(s), 4GHz, 4-wide issue, 16 MSHRs per core
LLC	64B cacheline, 8-way associative, 8MB total capacity (2MB/core with 4 cores)
DRAM Controller	FR-FCFS-Cap scheduling, 64-entry read/write request queue
DRAM	1 channel, 1 rank, **DDR4-2400R**, 4Gb ×8 chip
Benchmarks	23 benchmarks from **SPEC CPU2006 [4]**

Performance Comparison

☑ Comparison of Agile-DRAM with prior works

- table provides a comparison detailing the pros and cons of each work

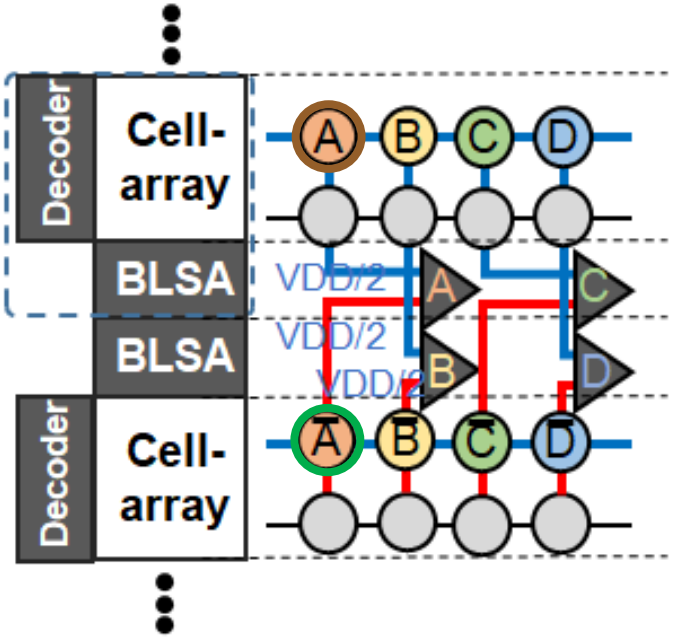
	DRAM Type	MAX Speedup(%)	Agile Switching	Capacity Reduced(%)	Chip Size Penalty(%)
Improving DRAM latency [5]	DDR3 -1066	21	No	0	3
Tiered-Latency DRAM [6]	DDR3 -1600	8.9	Yes	0	3.15
CLR-DRAM [1]	DDR4 -2400	59.8	No	0-50	3.2
Agile-DRAM (LL)	**DDR4 -2400**	**25.8**	**Yes**	**0-50**	**0**

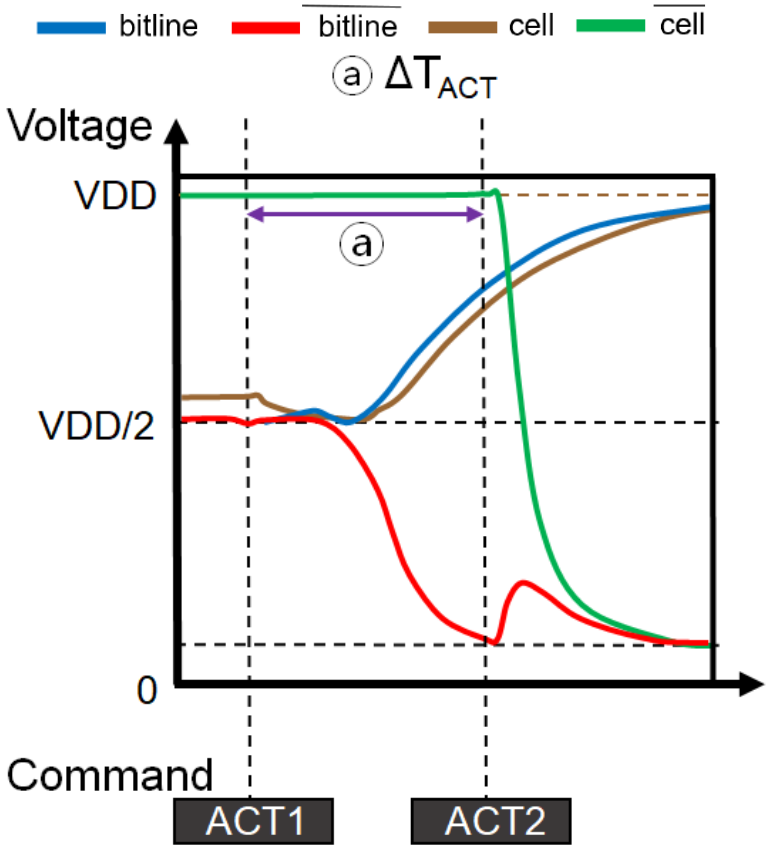
Hardware Overheads

☑ Synthesis setup

- compiler: synopsys design compiler
- logic libraries: synopsys 32nm lvt
- corner: ss
- pvt (32nm, 0.75V, 125'C)

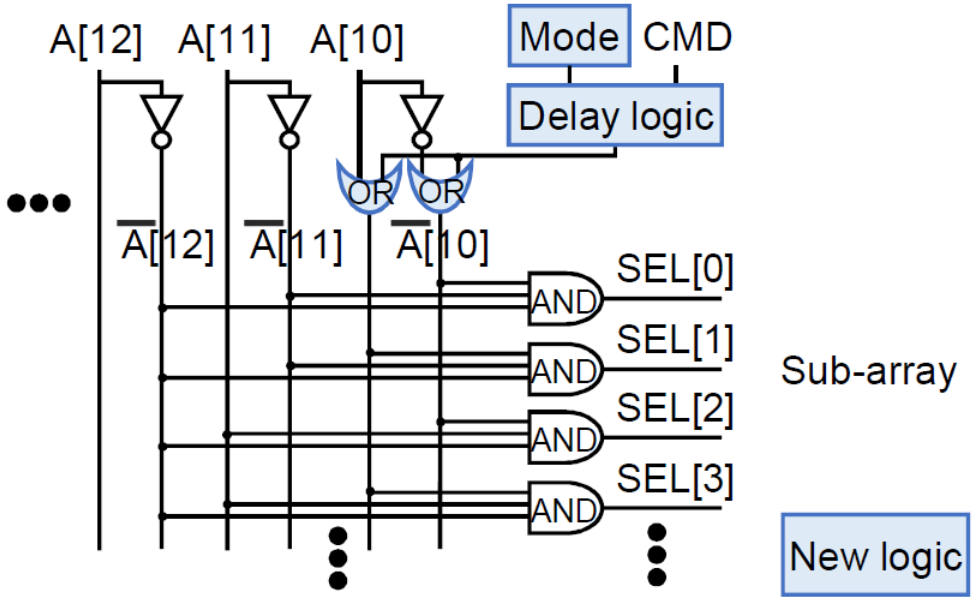
☑ Agile mode switching





☑ Modification in row decoder

- modifies the global row decoder
- allows paired sub-arrays to be selected and activated in parallel during low-latency or low-power modes



References

[1] H. Luo, T. Shahroodi, H. Hassan, M. Patel, A. G. Yaˇglıkc¸ı, L. Orosa, J. Park, and O. Mutlu, “Clr-dram: A low-cost dram architecture enabling dynamic capacity-latency trade-off,” in 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA). IEEE, 2020, pp. 666–679

[2] Y. Kim, W. Yang, and O. Mutlu, “Ramulator: A fast and extensible dram simulator,” IEEE Computer architecture letters, vol. 15, no. 1, pp. 45–49, 2015.

[3] K. Chandrasekar, C.Weis, Y. Li, B. Akesson, N.Wehn, and K. Goossens, “Drampower: Open-source dram power & energy estimation tool,” URL: http://www. drampower. info, vol. 22, 2012.

[4] J. L. Henning, “Spec cpu2006 benchmark descriptions,” pp. 1–17, 2006.

References

[5] S.-L. Lu, Y.-C. Lin, and C.-L. Yang, “Improving dram latency with dynamic asymmetric subarray,” in Proceedings of the 48th International Symposium on Microarchitecture, ser. MICRO-48. New York, NY, USA: Association for Computing Machinery, 2015, p. 255–266. [Online]. Available: https://doi.org/10.1145/2830772.2830827

[6] D. Lee, Y. Kim, V. Seshadri, J. Liu, L. Subramanian, and O. Mutlu, “Tiered-latency dram: A low latency and low cost dram architecture,” in 2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2013, pp. 615–626