# Co-Architecting Controllers and DRAM to Enhance DRAM Process Scaling

June 14, 2014

Uksong Kang, Hak-soo Yu, Churoo Park, *Hongzhong Zheng,

**John Halbert, **Kuljit Bains, SeongJin Jang, and Joo Sun Choi

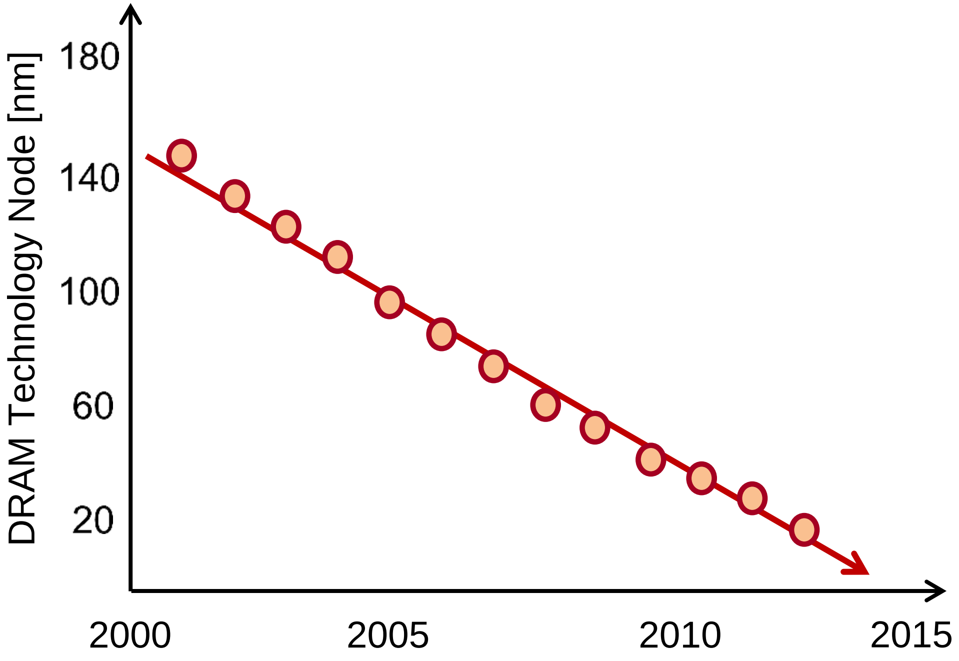
*Samsung Electronics, Korea / *Samsung Electronics, San Jose / **Intel*

# Outline

- Introduction
- DRAM process scaling challenges
- DRAM process scaling enhancing features
  - Sub-array level parallelism with tWR relaxation
  - Temperature compensated tWR
  - In-DRAM ECC with dummy data pre-fetching
- Conclusion

# Introduction

- DRAM process scaling enables continued density and speed increase, power decrease, and bit-cost reduction
- Today’s DRAM process is expected to continue scaling, enabling minimum feature sizes below 10nm
- Main challenges to address are expected to be refresh, write recovery time (tWR), and variable retention time (VRT) parameters



# DRAM Process Scaling Challenges

## ❖ Refresh

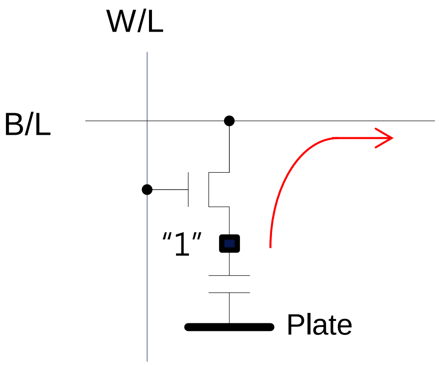
- Difficult to build high-aspect ratio cell capacitors decreasing cell capacitance
- Leakage current of cell access transistors increasing

## ❖ tWR

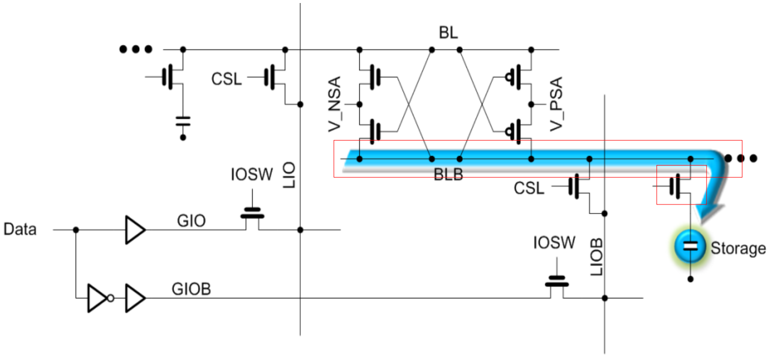
- Contact resistance between the cell capacitor and access transistor increasing
- On-current of the cell access transistor decreasing
- Bit-line resistance increasing

## ❖ VRT

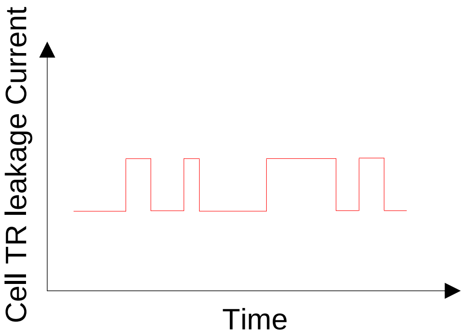
- Occurring more frequently with cell capacitance decreasing



**Refresh**



**tWR**



**VRT**

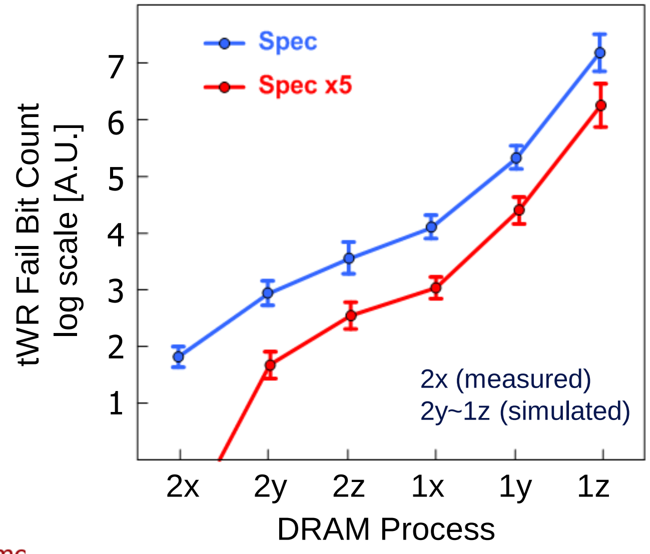
# Sub-array Level Parallelism with tWR Relaxation

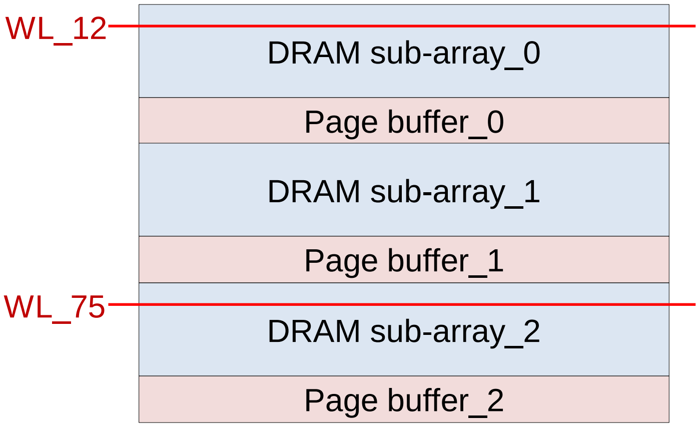
## ❖ tWR relaxation

- Relaxing tWR results in DRAM yield improvement but can degrade performance requiring new compensating features
- By increasing tWR 5X (from 15ns to 75ns), fail bit counts are expected to reduce by 1 to 2 orders of magnitudes

## ❖ Sub-array level parallelism (SALP)

- Allows a page in another sub-array in the same bank to be opened in parallel with the currently activated sub-array
- Results in performance gain by increasing the row access parallelism within a bank
  ⇒ Used to compensate for the performance loss caused by tWR relaxation




WL_12

DRAM sub-array_0

Page buffer_0

DRAM sub-array_1

Page buffer_1

WL_75

DRAM sub-array_2

Page buffer_2

Single bank with multiple sub-arrays

# Sub-array Level Parallelism Timing Gain

## ❖ Read

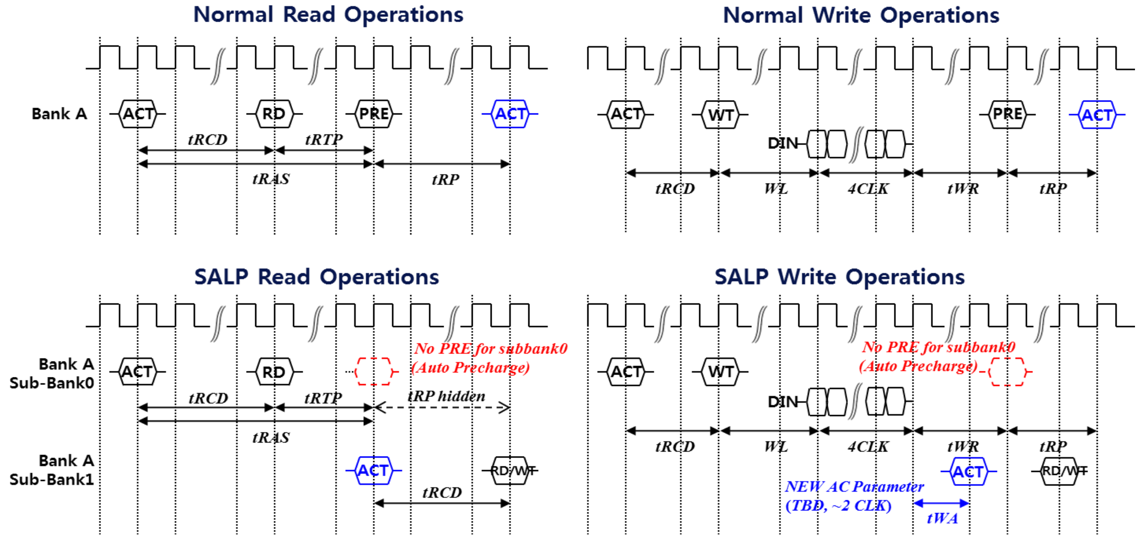
- Next activate command to a different sub-array in the same bank can come after tRAS ⇒ Pulled in by tRP compared to the normal case

## ❖ Write

- Next activate command to a different sub-array can come after tRCD+WL+4tck+tWA ⇒ Pulled in by tRP+tWR-tWA compared to the normal case

## ❖ Mutually exclusive sub-array activation

- All open pages in other sub-arrays are closed automatically once a new activate request to a closed sub-array page is made ⇒ simplifies operation


Normal Read Operations

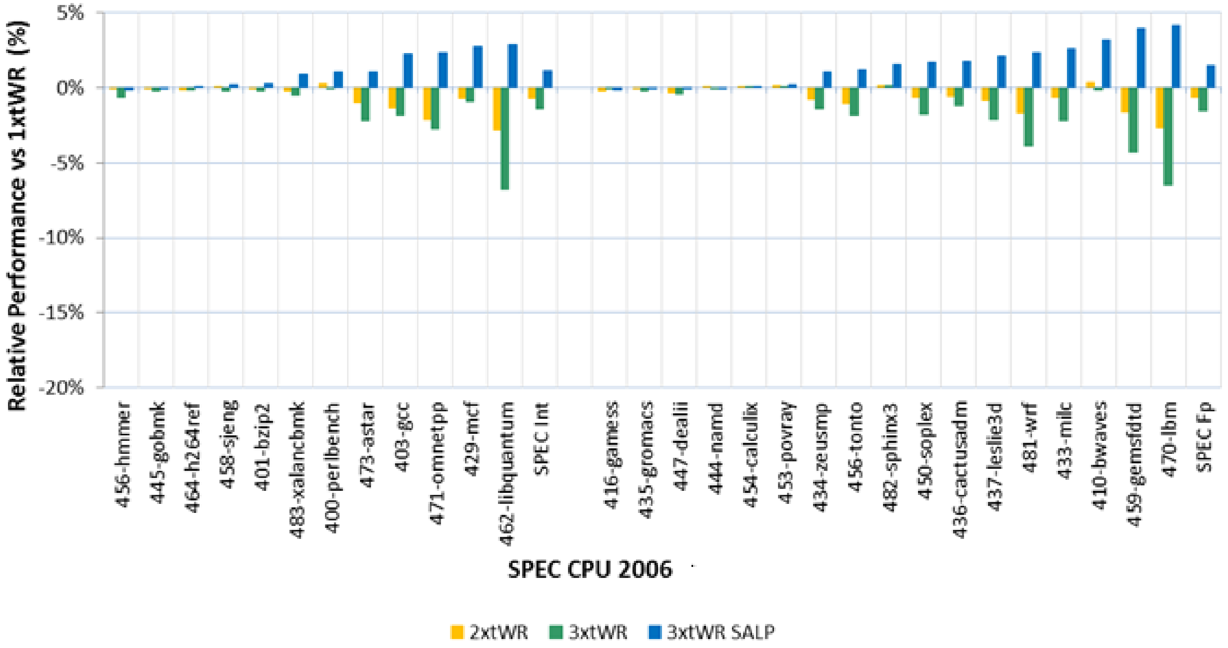
Normal Write Operations

SALP Read Operations

SALP Write Operations

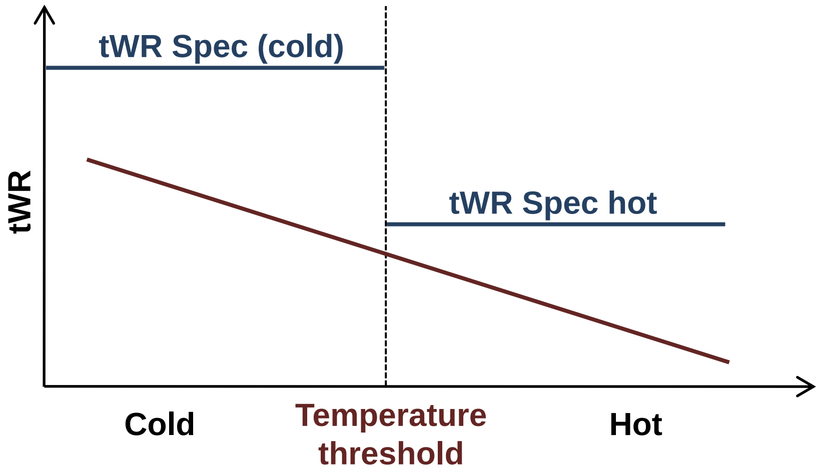
# Performance Impact of SALP and tWR relaxation

- Performance simulations run for various workloads when tWR is relaxed by 2X and 3X, and when SALP is applied with 2 sub-banks
- Results show that performance is reduced by ~5% and ~2% in average if tWR is relaxed by 3X and 2X, respectively
- Results also show that performance is compensated, and even improved to up to ~3% in average when SALP is applied, even with tWR relaxed by 3X



# Temperature Compensated tWR

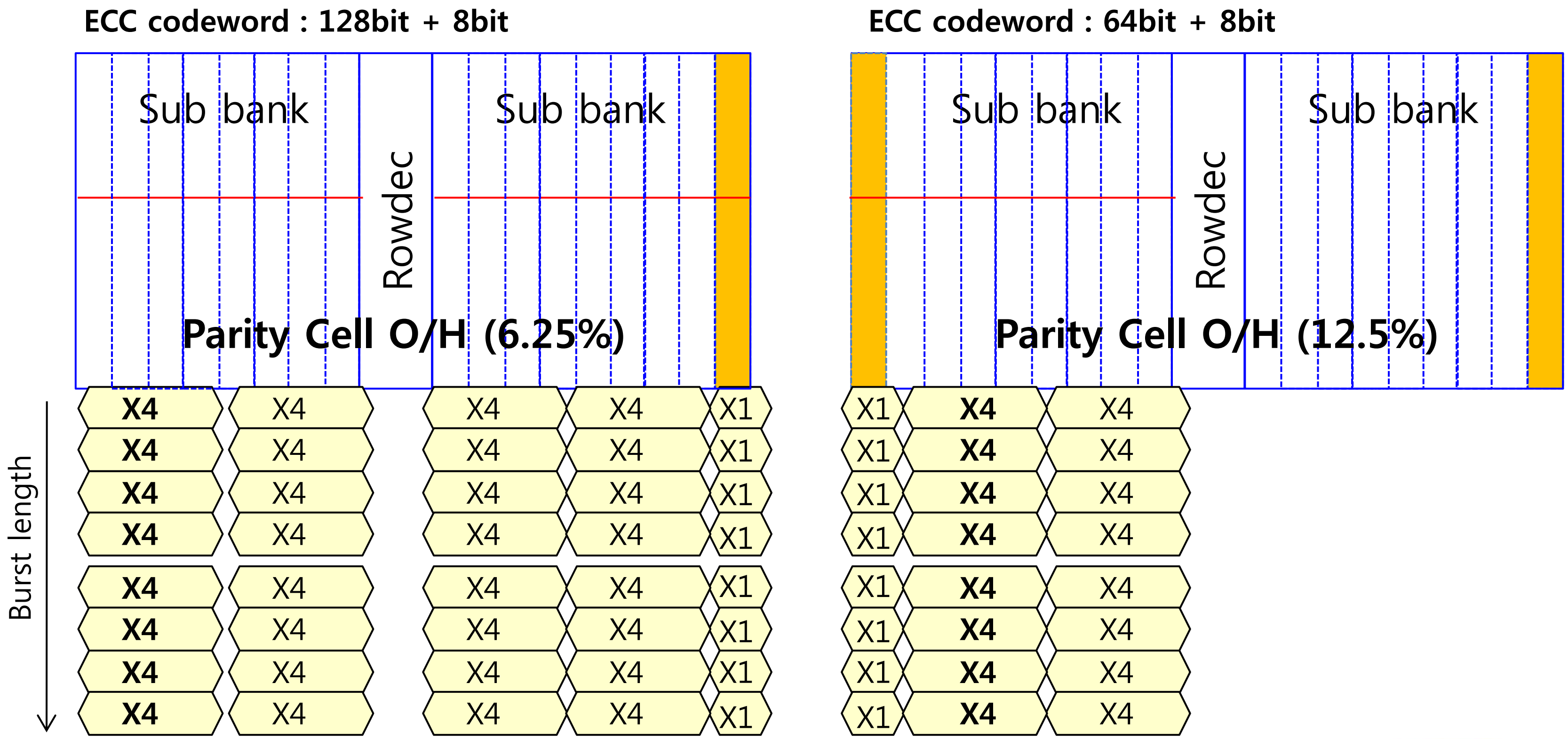
- ❖ Specifies different tWR values below and above a certain given temperature threshold
  - Longer tWR spec at cold temperature since tWR is worse at cold
  - Relaxing tWR spec at cold temperature helps to increase DRAM yield
  - DRAM temperature is checked by the memory controller periodically
- ❖ Combining temperature compensated tWR with SALP can improve performance when ambient system operation temperatures exceed the temperature threshold most of the time



| | <Temp Threshold | >Temp Threshold |
|---|---|---|
| tWR Spec | m+n [ns] | m [ns] |

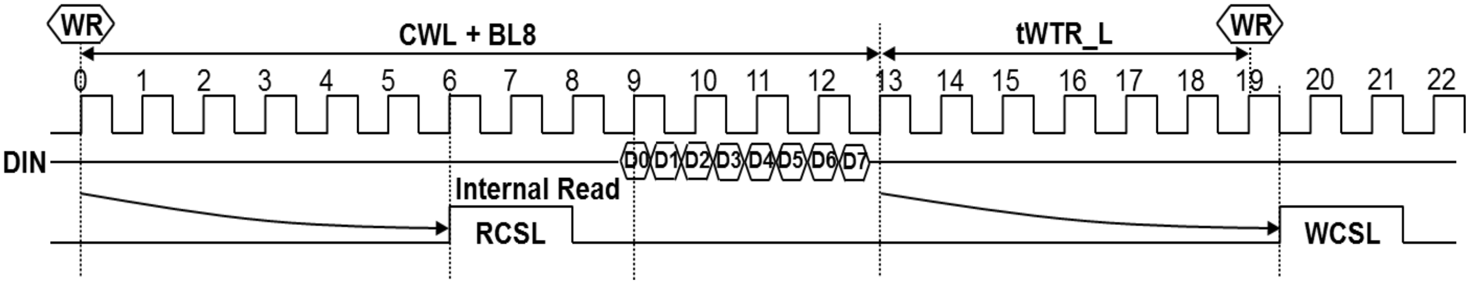
# In-DRAM ECC with Dummy Data Pre-fetching

- In-DRAM ECC is known to have high repair efficiency in repairing single fail bits, but has in general large chip size overhead (~20% for X4 DRAM)
- Proposing a method where extra dummy data bits are pre-fetched internally during reads and writes, resulting in parity bit chip size overhead reduction
  - Ex) External: X4 * BL8 = 32b | Internal: X4 * BL8 * 4 = 128b+8b (Parity O/H: 6.25%)
    Internal: X4 * BL8 * 2 = 64b+8b (Parity O/H: 12.5%)



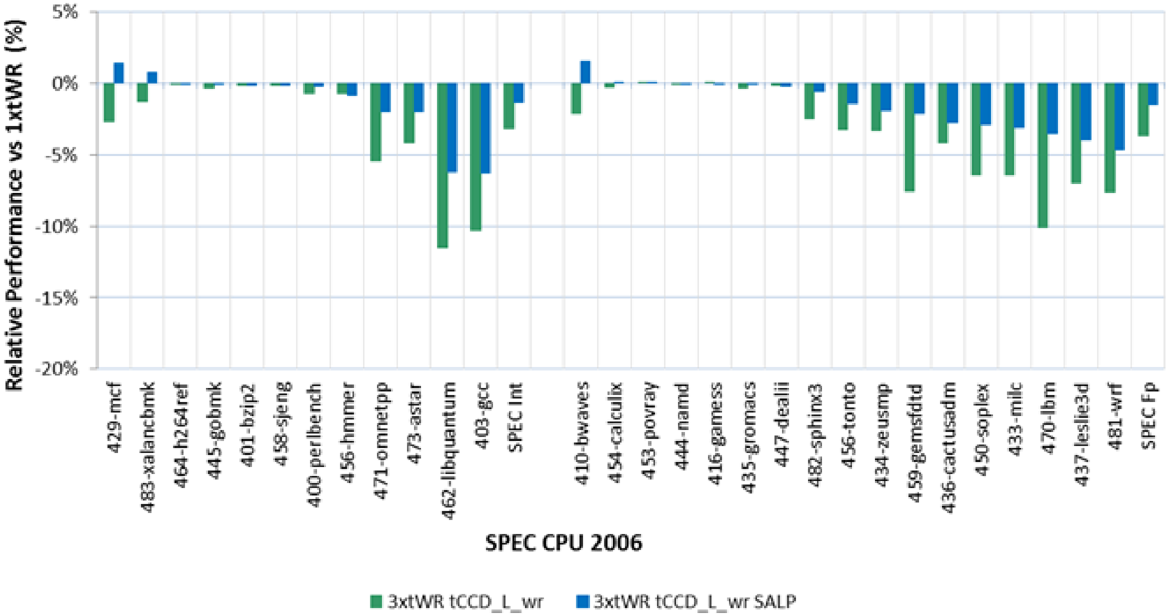
# In-DRAM ECC Timing

- Dummy data bit pre-fetching requires read-modify-write for all writes
  - Only parts of the data code words are overwritten, requiring parity bit update for every write
- Read-modify-write requires increase in tCCD_L_WR by “CWL + BL8 + tWTR_L”
  - tWTR_L needs to be observed from the last write-data burst to the next write command within the same bank group, since the next write actually starts with an internal read
  - tCCD_L_WR needs to be increased from 6~9 clock cycles to 25~29 clock cycles



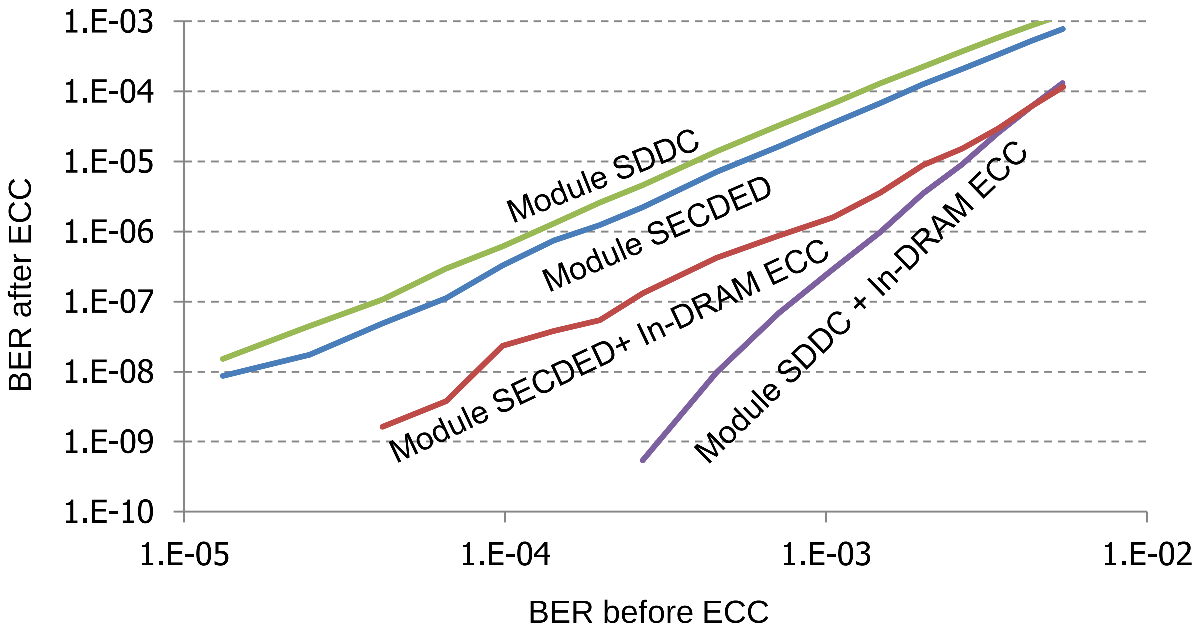
# Performance Impact of In-DRAM ECC

- ❖ Performance simulations run when tWR is relaxed 3X, tCCD_L_WR is increased to 24tck, and SALP is used to compensate for the performance loss
  - Assumption: tCCD_L_WR=24tck, tWR 3X, DDR4-2400, 6 channels, 1 DIMM/channel, 2 rank/DIMM
- ❖ Results show an average performance loss of ~2% with SALP and tWR relaxed 3X
- ❖ By applying in-DRAM ECC only without increasing tWR, performance loss is expected to break even when applying SALP



# Field Reliability and in-DRAM ECC

- Field failure rate probability analysis with and without in-DRAM ECC, and with SECDED and SDDC (chip-kill) were performed
  - Assumption: 4Gb X4 DRAM, single bit failures only, errors randomly scattered
- Results show that the bit error rate decreases by orders of magnitudes when in-DRAM ECC is adopted



# Conclusion

- ❖ To enhance DRAM process scaling, challenging process parameters including refresh, tWR, and VRT need to be addressed

- ❖ Combining SALP with TCWR enables tWR relaxation and compensation of performance loss caused by this parameter relaxation

- ❖ In-DRAM ECC allows efficient repair of fail bits coping with VRT and refresh failures resulting in increased field reliability

- ❖ By enabling the proposed scaling features, future DRAM process scaling is expected to accelerate further beyond 10nm

# Thank you