Chip-Level RAID with Flexible Stripe Size and Parity Placement for Enhanced SSD Reliability

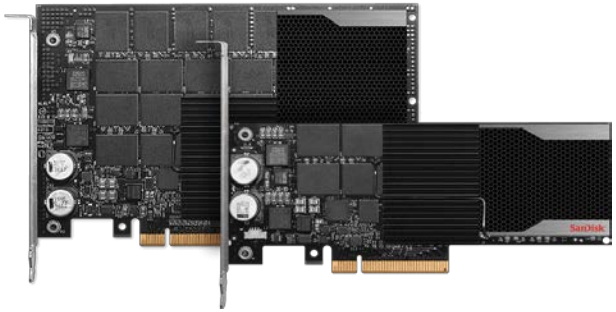
Jaeho Kim (Postdoctoral researcher @ Hongik Univ.)

Donghee Lee (Professor @ Univ. of Seoul)

Sam H. Noh (Professor @ Hongik Univ.)

Introduction & Motivation

- Flash SSD products with RAID-5 like data protection



Fusion-io's ioMemory
(Adaptive Flashback Tech.)

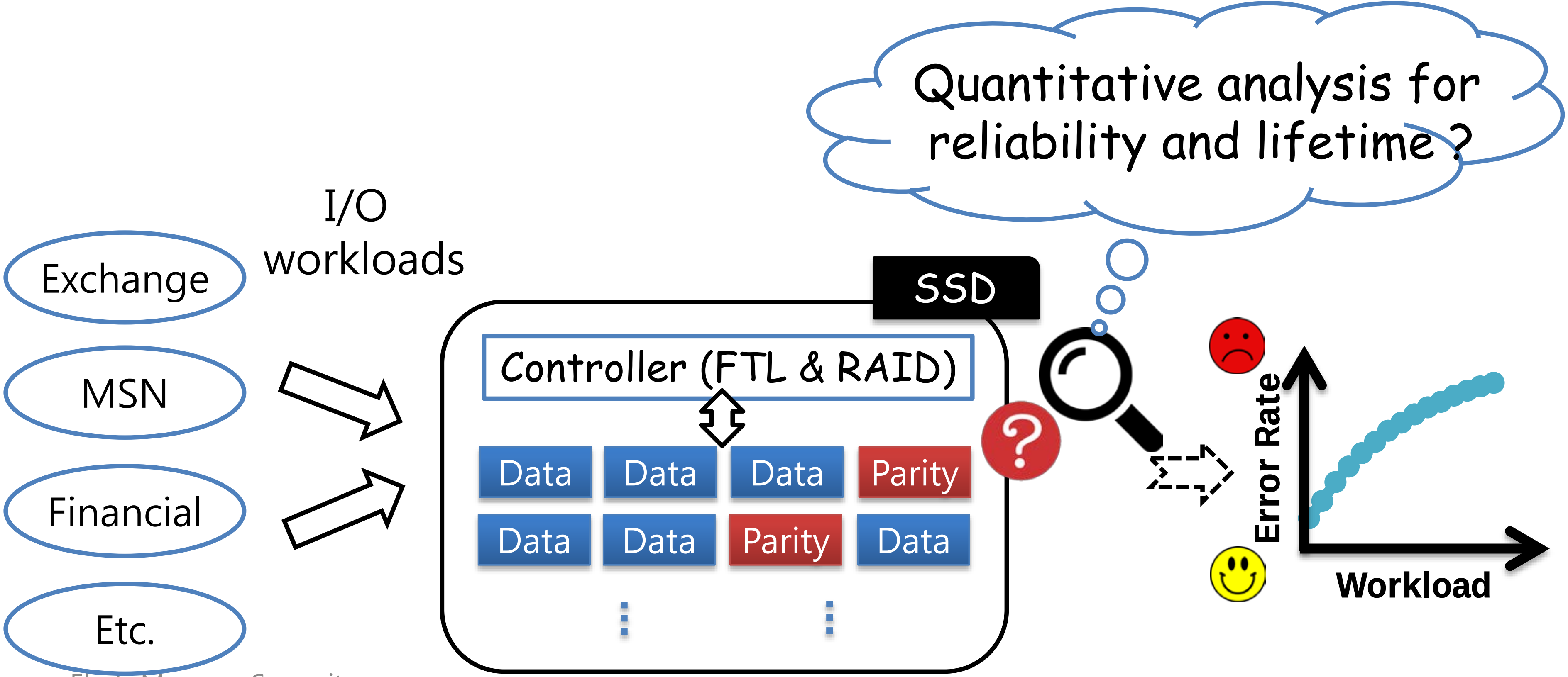
Baidu's Software-Defined Flash

Micron's P420m
(Redundant Array Independent NAND Tech.)

Shannon Systems's Direct-IO

Introduction & Motivation

- Applying RAID-5 into SSD internal
  - Is RAID-5 suitable for flash chips?
  - Is RAID-5 really beneficial for SSD?



Quantitative analysis for reliability and lifetime ?

I/O workloads

Exchange

MSN

Financial

Etc.

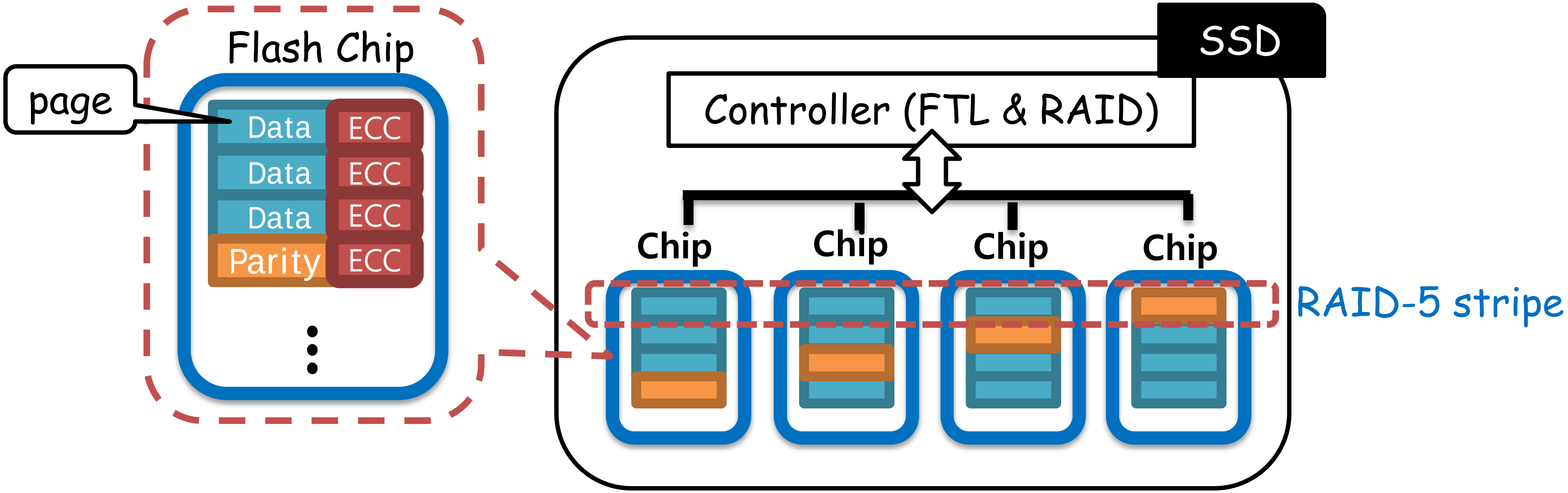
SSD

Controller (FTL & RAID)

Data	Data	Data	Parity
Data	Data	Parity	Data

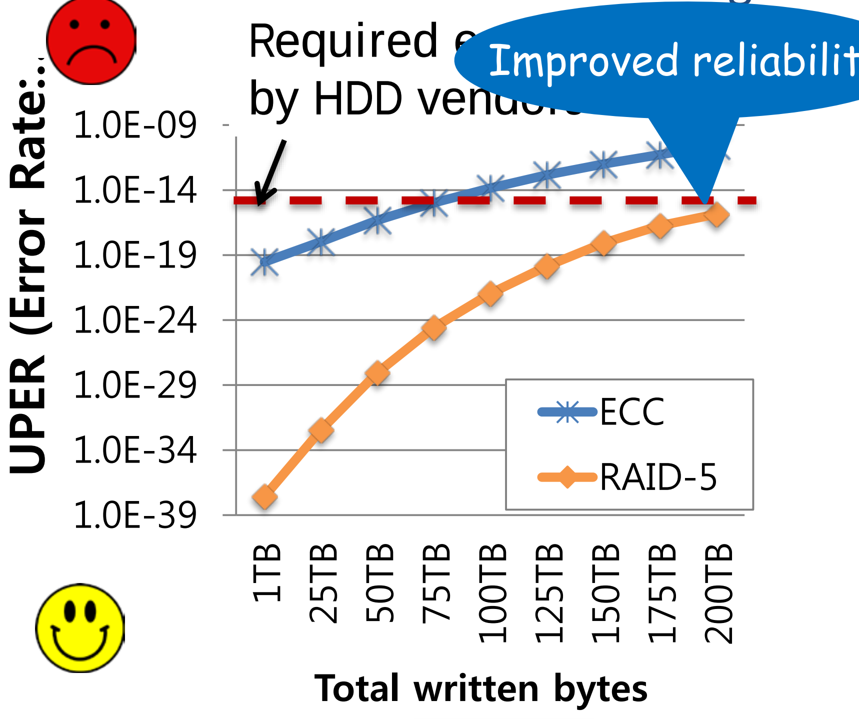
Applying RAID-5 into SSD

- Apply **RAID-5** configuration to chips comprising the **SSD** device



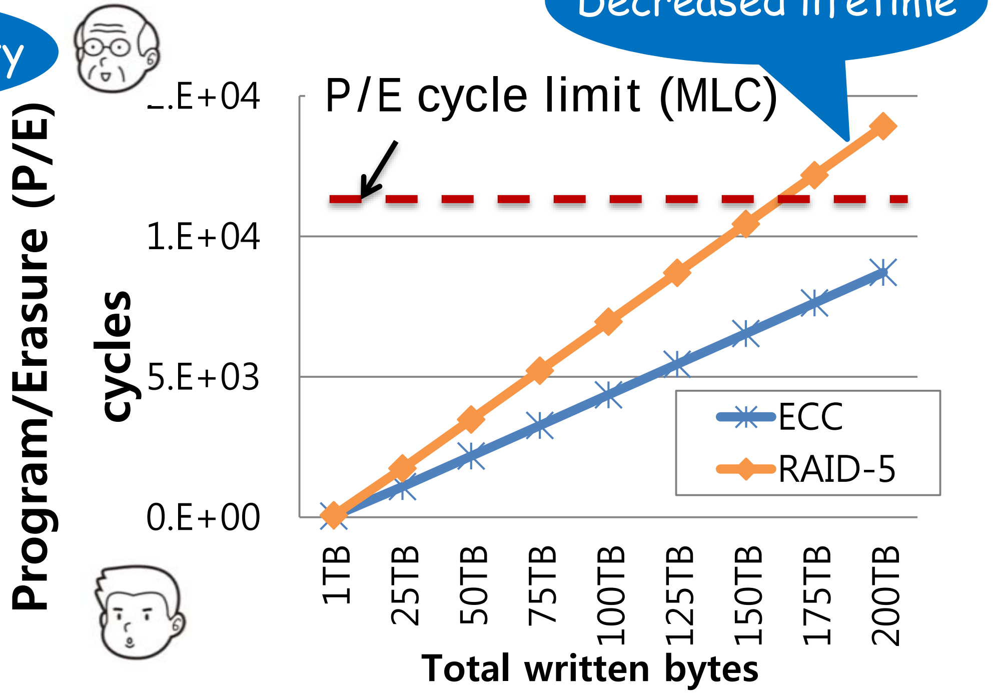
Pros & Cons of RAID-5 in SSD

- Assumptions of quantitative analysis
  - Conventional SSD (denoted “ECC”): MLC 64GB with BCH code (4bit/512bytes) for ECC
  - RAID-5 SSD (denoted “RAID-5”): MLC 64GB applying RAID-5 without parity cache
  - Workload: Financial
- Pros of RAID-5: Improving reliability of SSD
- Cons of RAID-5: Decreasing lifetime of SSD



Improved reliability

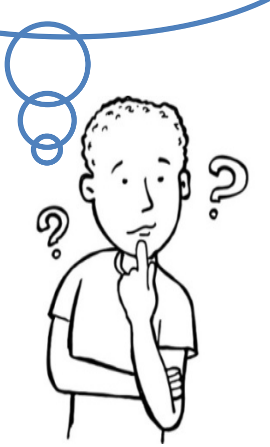
Decreased lifetime



Pros

Cons

How can we improve the reliability while prolong the lifespan of the SSD?

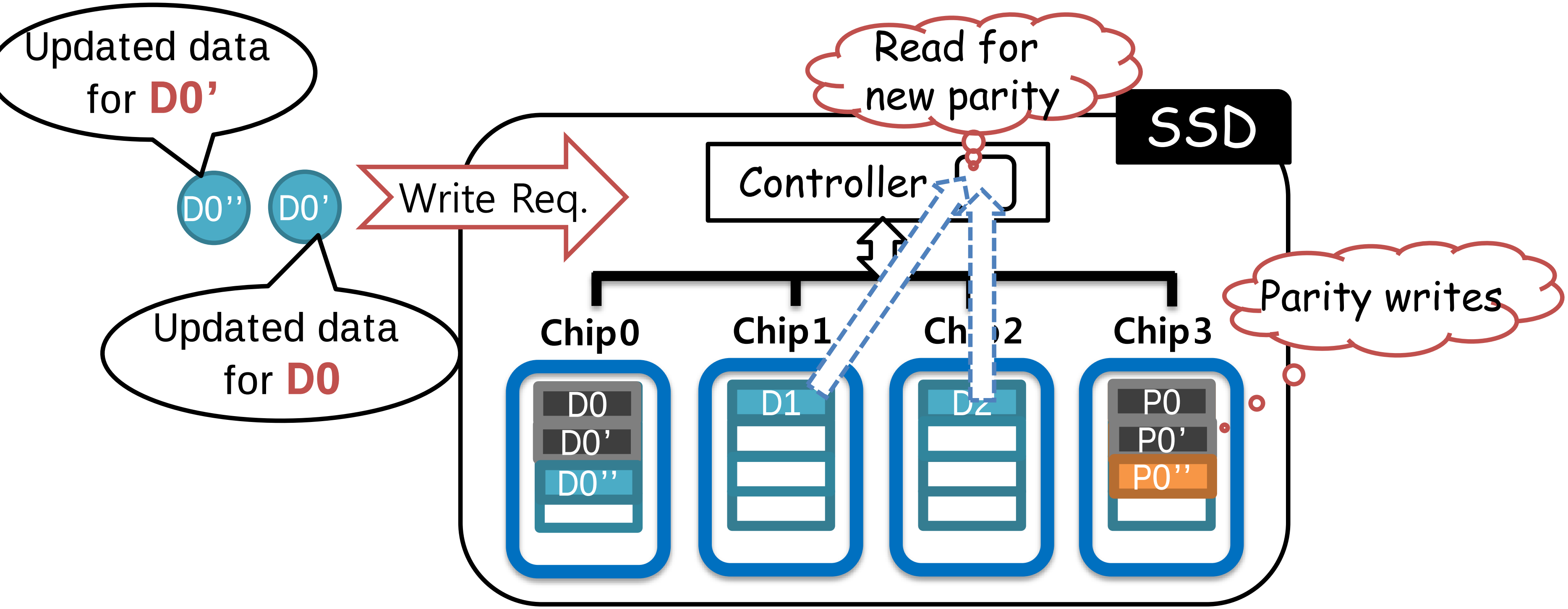


Outlines

- Introduction & Motivation
- Challenges of RAID-5
- Our Solution: eSAP-RAID
- Evaluations
- Analytic Models of RAID Schemes
- Conclusion

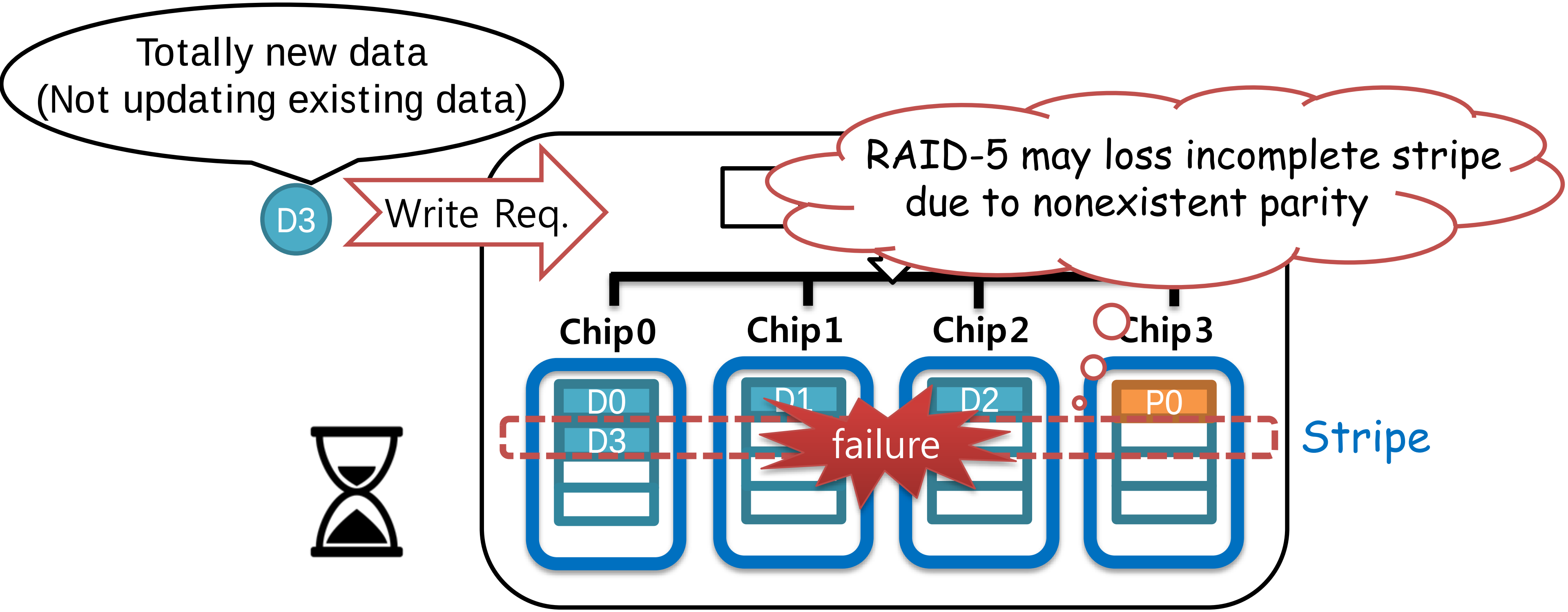
Applying RAID-5 into SSD: Challenges #1,2

- Out-of-place update property of flash memory
  - Parity writes increase write amplification (WA) in SSD
- LBN (Logical Block Number) based striping feature of RAID-5
  - Parity update overhead (Read-modify-write) for small write requests
  - Data are written to specific chip depending on the LBN of data



Applying RAID-5 into SSD: Challenge #3

- Open a window of vulnerability (for totally new data)
  - Small writes must wait until stripe fills up to write parity


Totally new data
(Not updating existing data)
D3
Write Req.
RAID-5 may loss incomplete stripe
due to nonexistent parity
Chip0
Chip1
Chip2
Chip3
D0
D3
D1
D2
P0
failure
Stripe

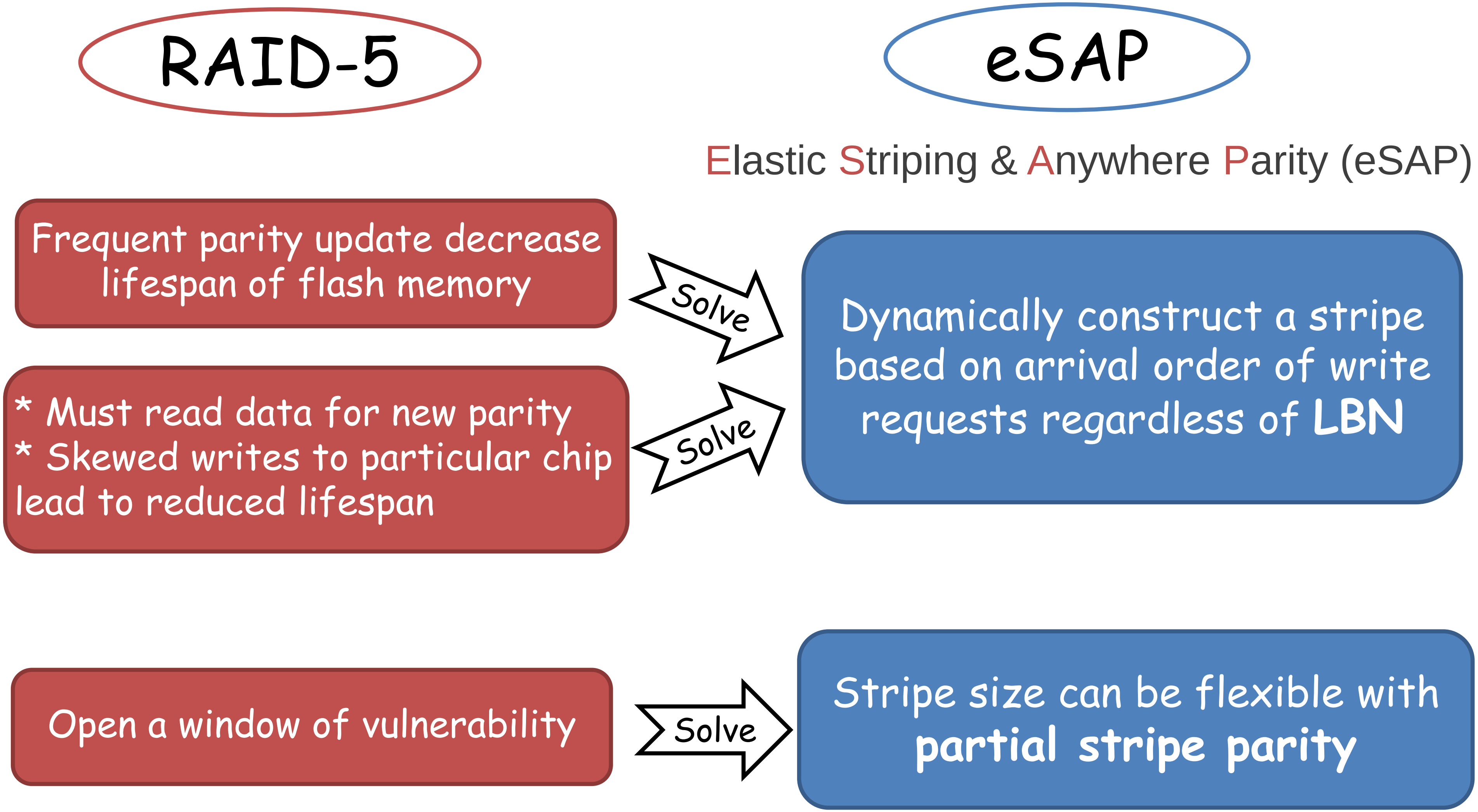
Summary of the Challenges

- **Out-of-place update** property of flash memory
  - Parity update may decrease lifespan of flash memory
- **LBN based striping** feature of RAID-5
  - Must read old data or old parity for parity calculation
  - Data is written to specific chip depending on the LBN of data
- Open a window of vulnerability
  - Small writes must wait until stripe fills up to calculate parity

Outlines

- Introduction & Motivation
- Challenges of RAID-5
- Our Solution: eSAP-RAID
- Evaluations
- Analytic Models of RAID Schemes
- Conclusion

Our Solution: eSAP-RAID



RAID-5

eSAP

Elastic Striping & Anywhere Parity (eSAP)

Frequent parity update decrease lifespan of flash memory

* Must read data for new parity
* Skewed writes to particular chip lead to reduced lifespan

Solve

Solve

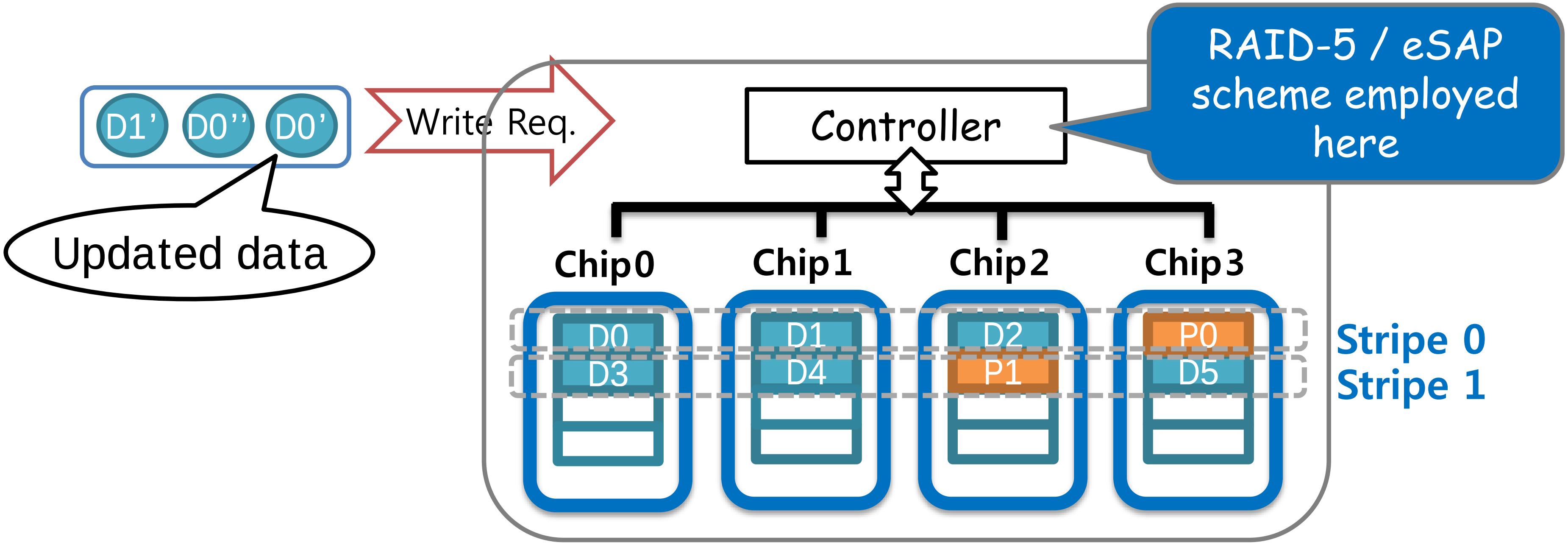
Dynamically construct a stripe based on arrival order of write requests regardless of **LBN**

Open a window of vulnerability

Solve

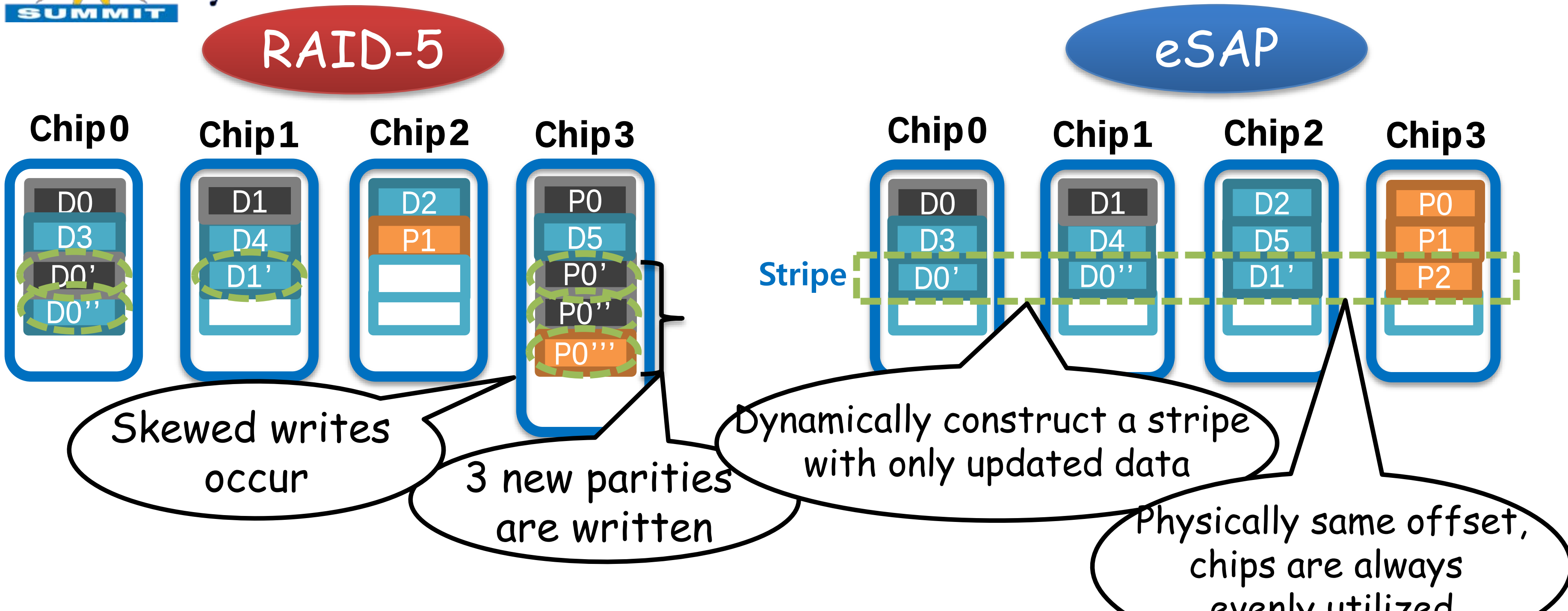
Stripe size can be flexible with **partial stripe parity**

Write Cost: RAID-5 vs. eSAP



- Assumptions
  - Stripe 0 and 1 are already constructed in the SSD
    - Stripe 0: D0, D1, D2, and P0
    - Stripe 1: D3, D4, D5, and P1
  - Updated data separately arrive in **D0'**, **D0''**, and **D1'** order

Write Cost: RAID-5 vs. eSAP



Skewed writes occur

3 new parities are written

Dynamically construct a stripe with only updated data

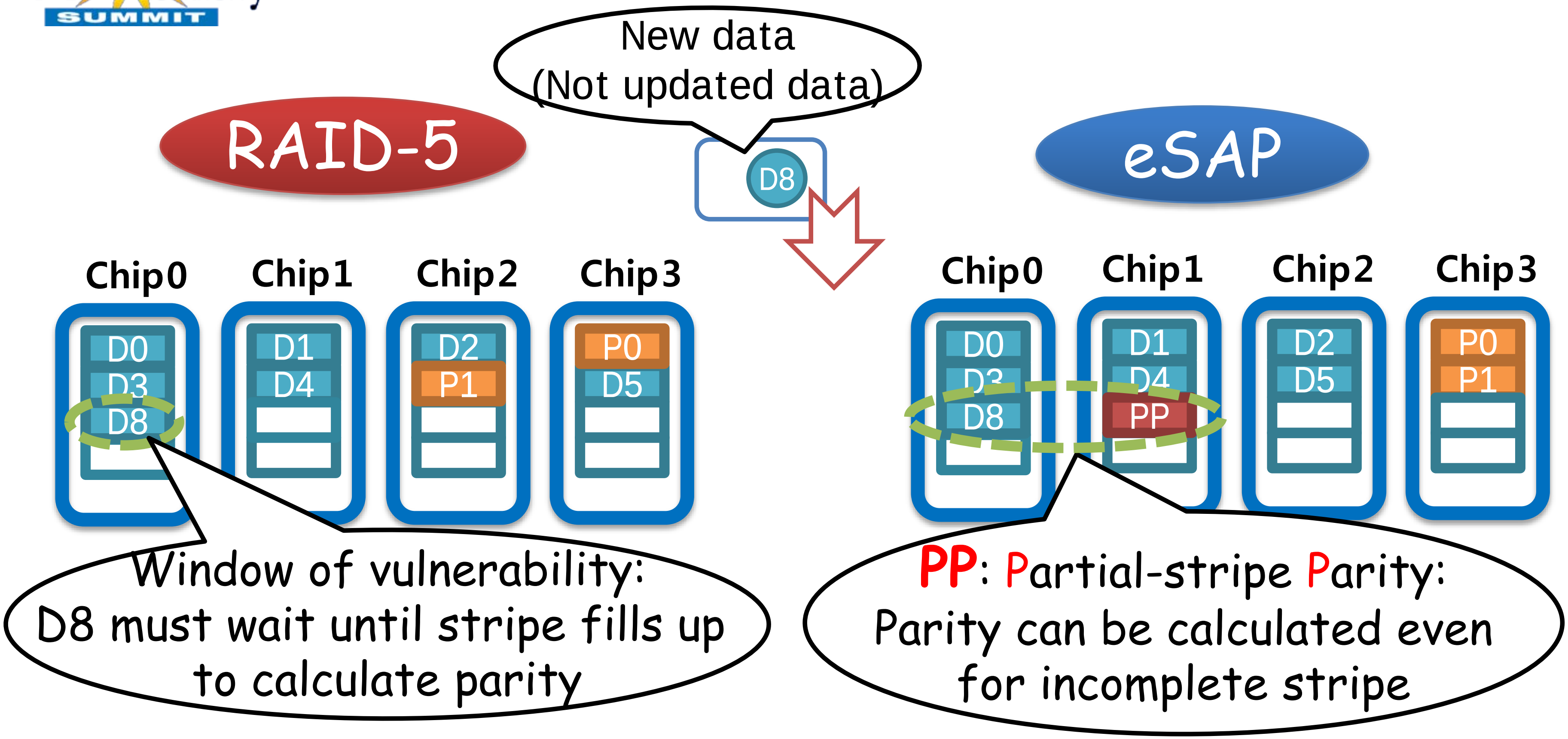
Physically same offset, chips are always evenly utilized

- I/O cost: eSAP reduces the number of read & write

	RAID-5	eSAP
Write updated data	3	3
Read for parity calculation	4	0
Write parity	3	1

Reduced cost of parity overhead

Reliability: RAID-5 vs. eSAP



New data
(Not updated data)

Window of vulnerability:
D8 must wait until stripe fills up
to calculate parity

PP: Partial-stripe Parity:
Parity can be calculated even
for incomplete stripe

- How can we protect **new data ‘D8’** before parity write?
  - RAID-5 **may loss** incomplete stripe due to without parity
  - eSAP **can protect** the new data with partial stripe parity (flexible stripe size)

Outlines

- Introduction & Motivation
- Challenges of RAID-5
- Flash-aware New RAID Architecture
- Evaluations
- Analytic Models of RAID Schemes
- Conclusion

Evaluation Setup

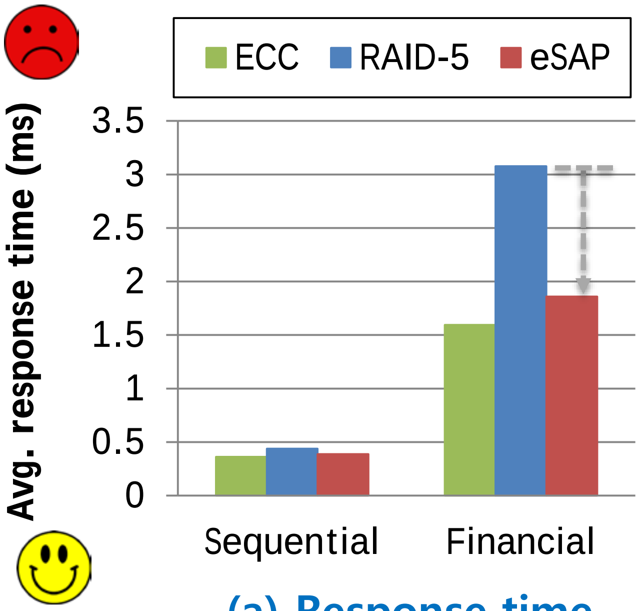
- SSD extension with the DiskSim, which is a simulator for SSD
  - 8 flash memory chips, a stripe consists of 16 pages

- Evaluate three configurations
  - ECC: No parity (similar to RAID-0)
  - RAID-5: Conventional RAID-5 scheme
  - eSAP: Elastic Striping and Anywhere Parity-RAID (Proposed scheme)

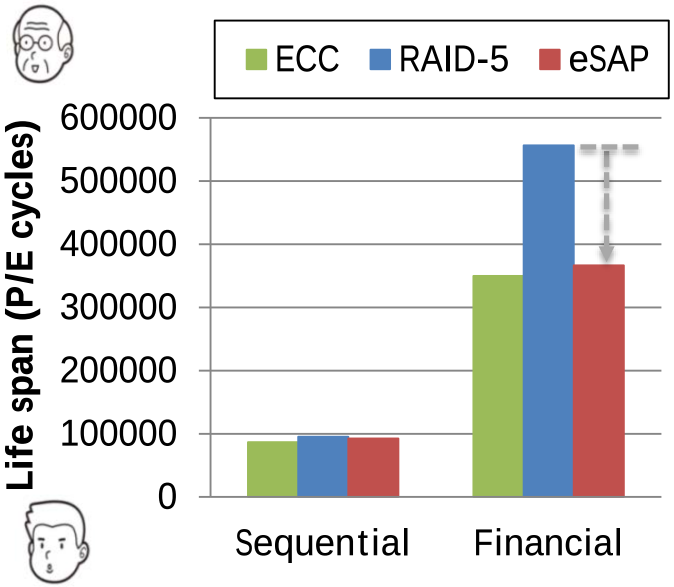
- Characteristics of I/O workloads

Workload	Total Data Req.(GB)	Write Ratio
Sequential	**21.8**	**1.0**
Random	30.2	1.0
Financial	**35.7**	**0.81**
Exchange	101.2	0.46
MSN	29.7	0.96

Response time & Life span



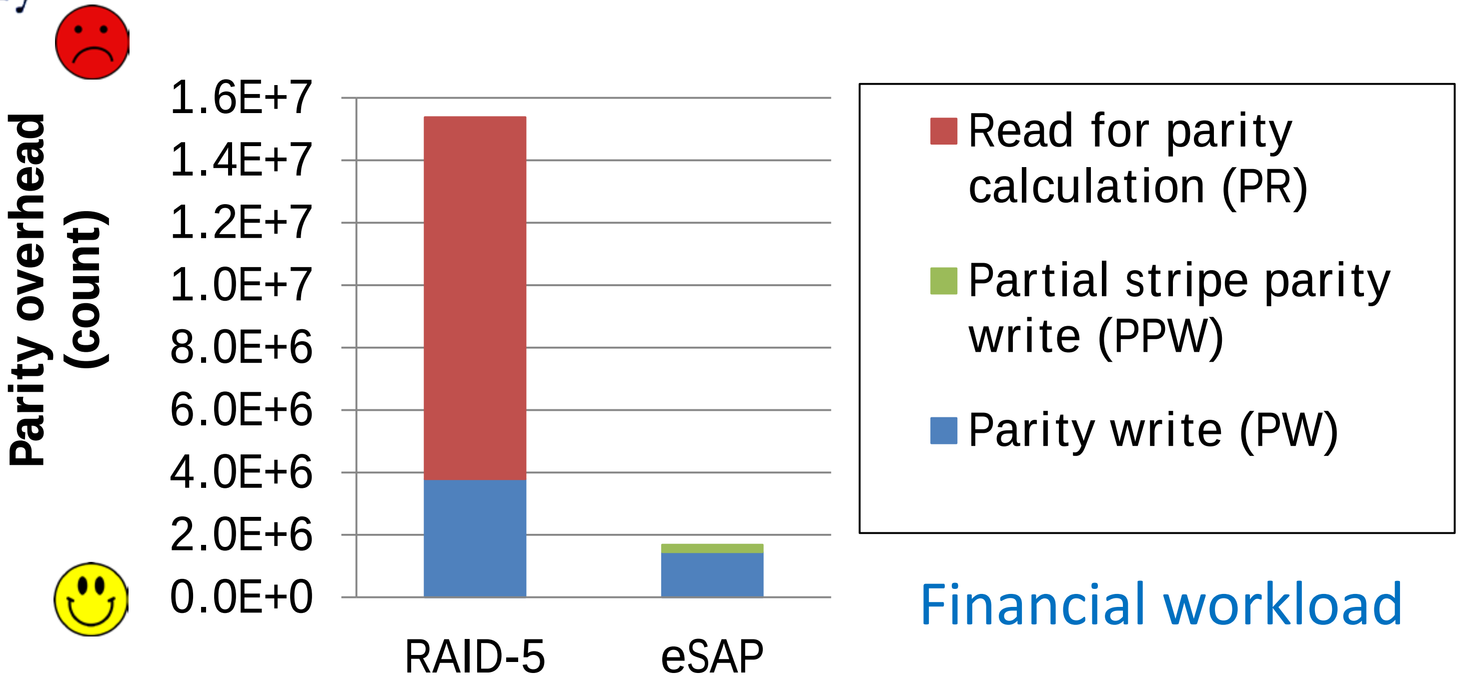
(a) Response time



(b) Life span (P/E cycles)

- eSAP reduces the response time over RAID-5
- eSAP prolongs the life span of SSD over RAID-5
- RAID-5 performs worst, especially for the financial workloads
  - Small writes incur heavy parity overhead

Analysis of Parity Overhead



Financial workload

- Parity overhead of RAID-5
  - Reads for parity calculations **(PR)**
  - Parity writes **(PW)**
- Parity overhead of eSAP
  - Parity writes **(PW)**
  - Partial stripe parity writes **(PPW)** for small write request

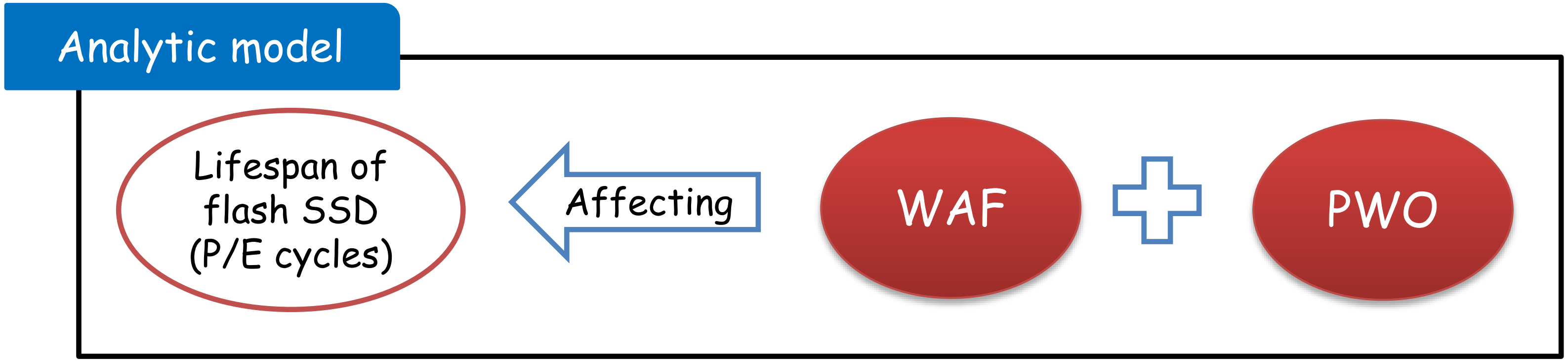
Outlines

- Introduction & Motivation
- Challenges of RAID-5
- Flash-aware New RAID Architecture
- Evaluations
- **Analytic Models of RAID Schemes**
- Conclusion

Analytic Models of RAID Schemes

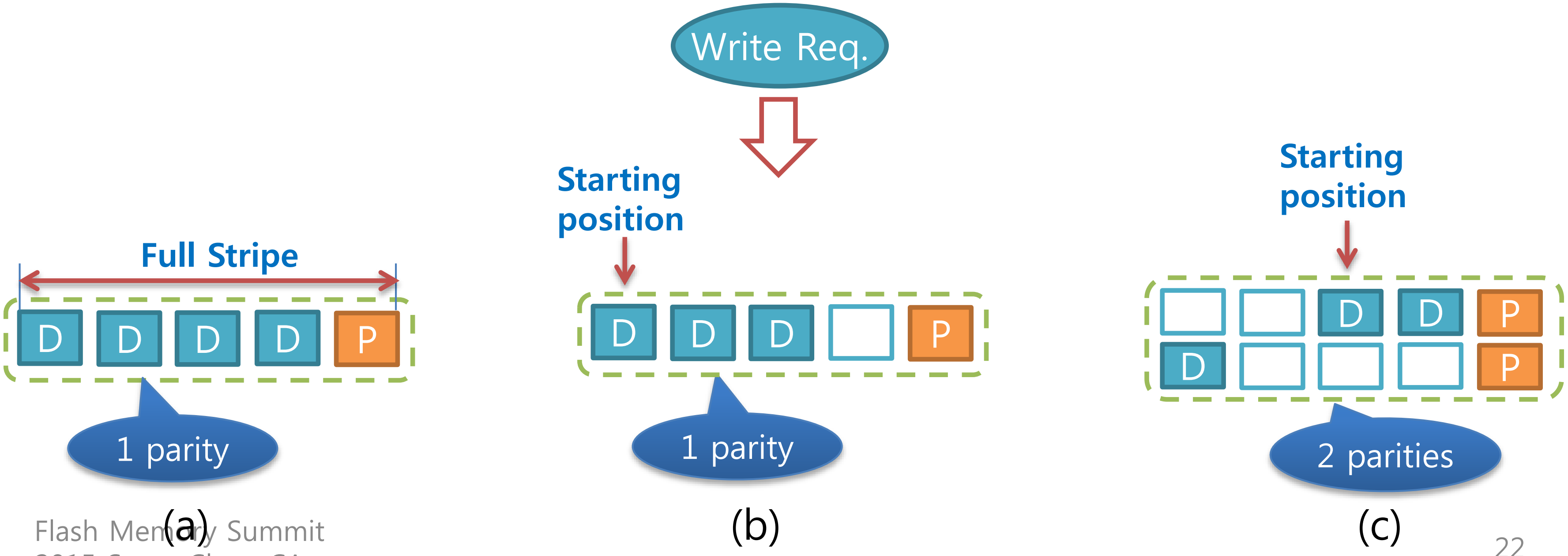
- Goals of analytic models
  - Find expected lifespan (P/E cycles) of SSD with various I/O workloads
  - Project long-term reliability according to the lifespan of SSD

- Two factors affecting lifespan of SSD
  - Write Amplification Factor (WAF)
    - Garbage collection cost
  - Parity Write Overhead (PWO)
    - Term derived from this work (Mathematical model)



Parity Write Overhead (PWO)

- Parity write overhead are determined by
  - 1) Size of RAID stripe
  - 2) Size of write request
  - 3) Starting position of write request within a stripe
- From the PWO,
  - We can estimate the number of page writes and erase operations

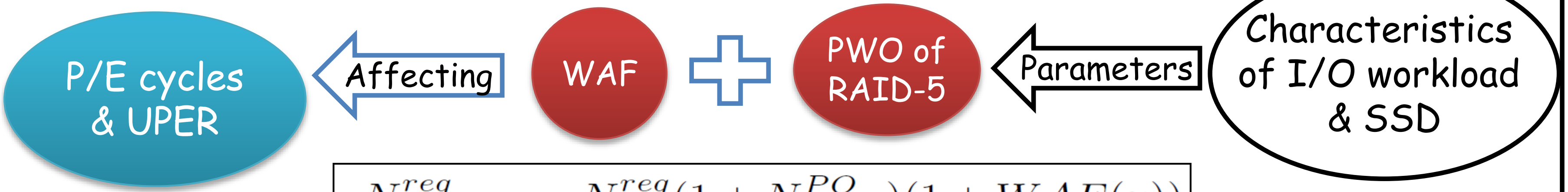


Analytic Models of RAID-5 and eSAP

Please refer to our paper for details!!

- Expected lifespan of SSD with RAID-5

Analytic model of RAID-5


P/E cycles & UPER

Affecting

WAF

PWO of RAID-5

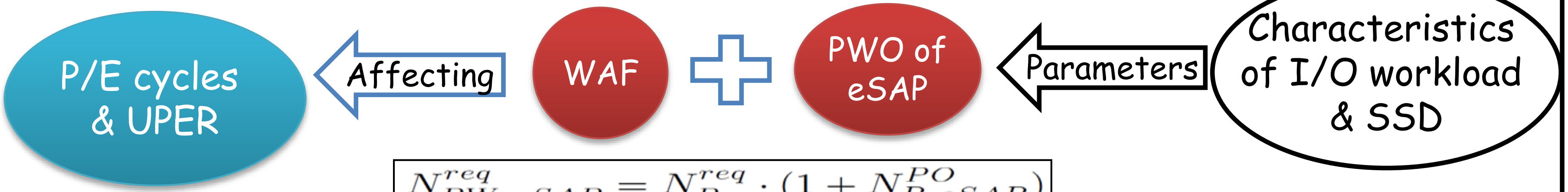
Parameters

Characteristics of I/O workload & SSD

$$N^{req}_{PW\_R5} = N^{req}_{P}(1 + N^{PO}_{P\_R5})(1 + WAF(u))$$

- Expected lifespan of SSD with eSAP

Analytic model of eSAP


P/E cycles & UPER

Affecting

WAF

PWO of eSAP

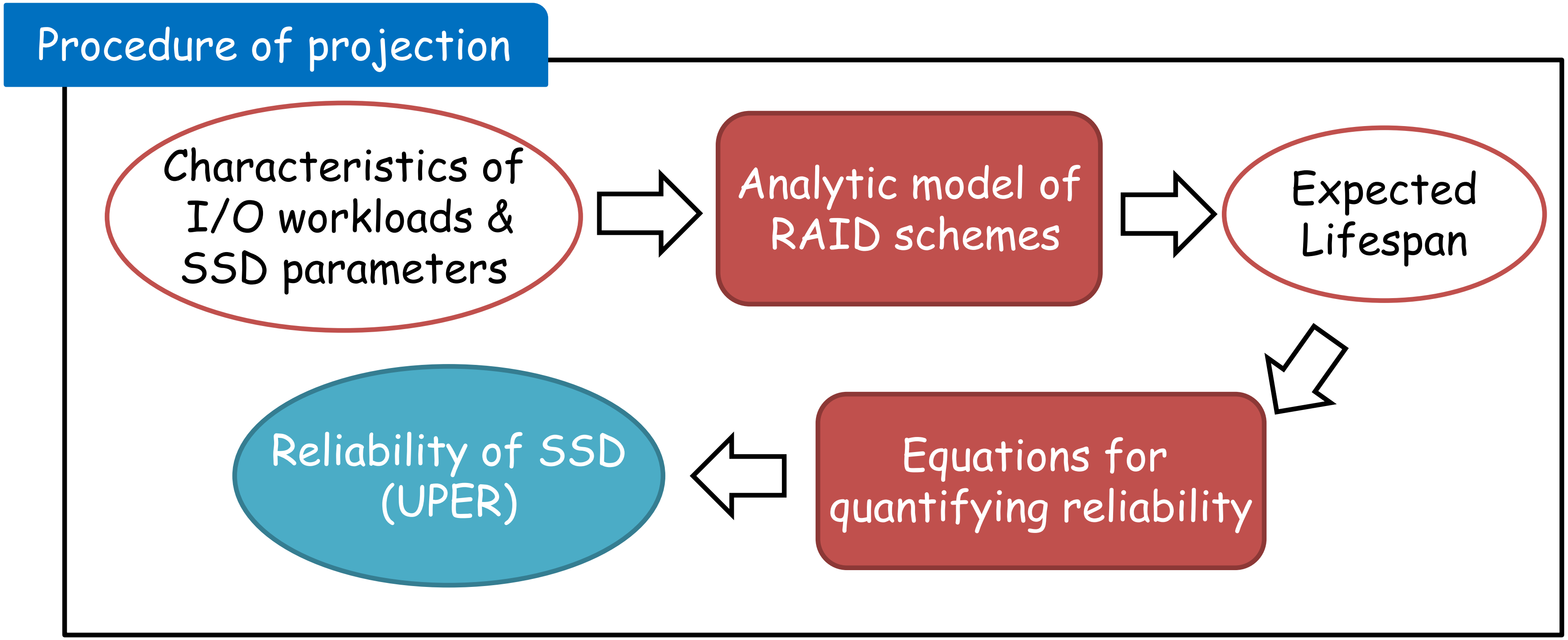
Parameters

Characteristics of I/O workload & SSD

$$N^{req}_{PW\_eSAP} = N^{req}_{P} \cdot (1 + N^{PO}_{P\_eSAP}) + N^{req}_{P} \cdot (1 + N^{PO\_WA}_{P}) \cdot WAF(u)$$

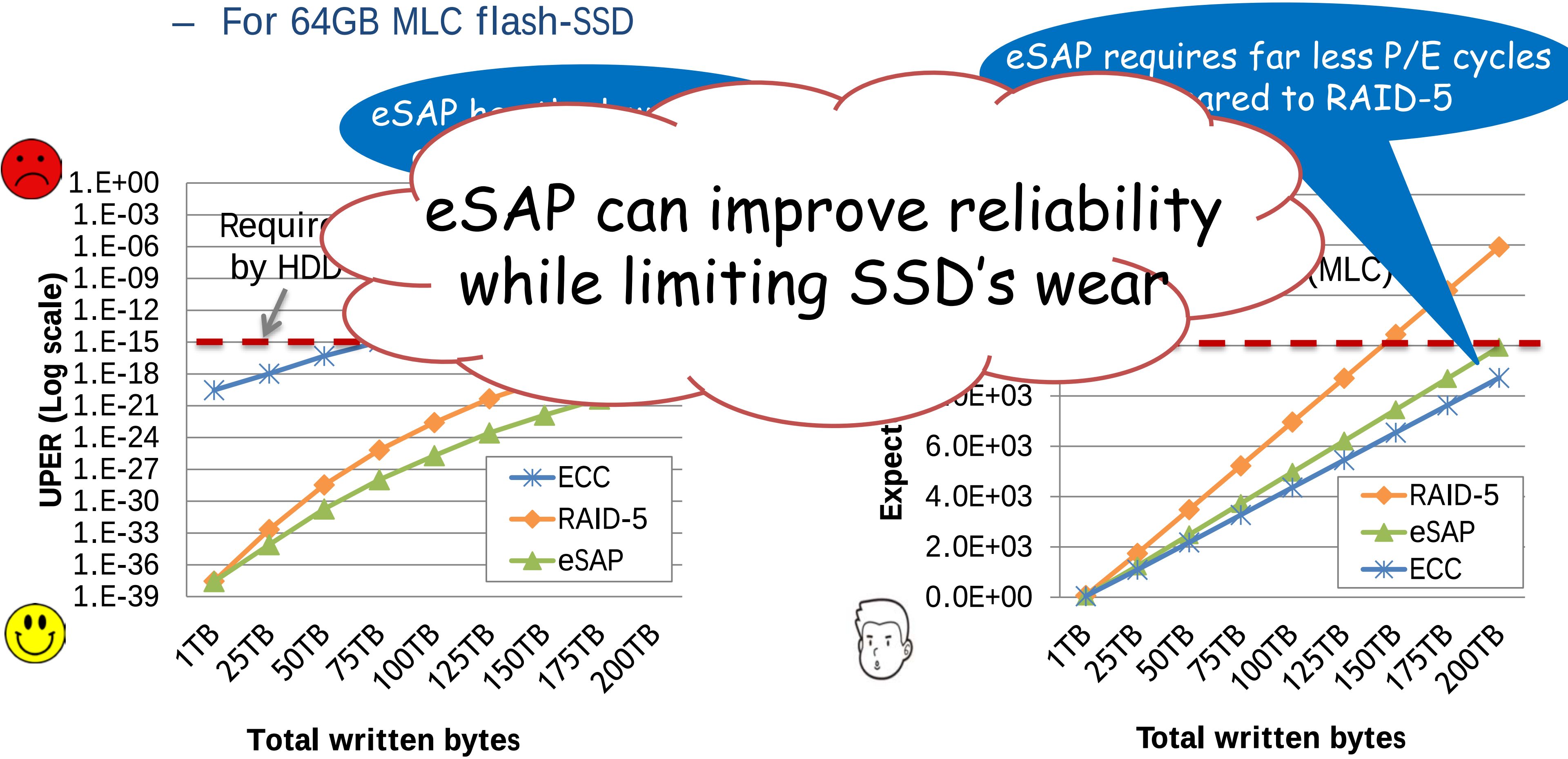
Projecting Long-term Reliability

- Procedure of projection
  1) Extract characteristics of I/O workloads and parameters of SSD
  2) Put extracted values into analytic models to expect lifespan of SSD
  3) Calculate reliability equations with expected lifespan of SSD
  4) Find Uncorrectable Page Error Rate (UPER) from the calculation


Procedure of projection
Characteristics of I/O workloads & SSD parameters
Analytic model of RAID schemes
Expected Lifespan
Equations for quantifying reliability
Reliability of SSD (UPER)

Analysis of Long-term Reliability

- Uncorrectable Page Error Rate (UPER) and life span of SSD
  - Financial workload
  - For 64GB MLC flash-SSD



eSAP requires far less P/E cycles
ared to RAID-5

eSAP can improve reliability while limiting SSD's wear

Conclusion

- Reliability of flash based storage is getting more crucial

- A solution to improve reliability is to apply RAID configuration into SSD
  - Conventional RAID-5 is not suitable
  - eSAP: A novel flash-aware RAID scheme is proposed

- Derive the analytical model of RAID schemes in SSD
  - Derive performance and lifespan models of RAID schemes in SSDs
  - Project long-term reliability of SSDs

Thank you! & Questions?

Please refer to the paper for details,

"Chip-Level RAID with Flexible Stripe Size and Parity Placement for Enhanced SSD Reliability",
IEEE Transactions on Computers, 2015

Jaeho Kim (kjhnet@gmail.com)

Backup Slides

Accuracy of Model

- Accuracy ratio of the model compared to the experimentally obtained
  - Most of the cases, the difference between the model and the experimental results are within 10%

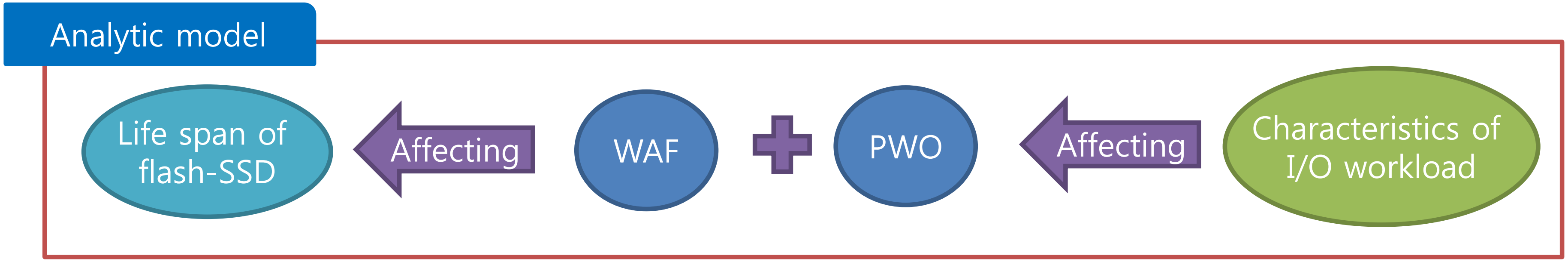
Workloads	RAID-5	eSAP
Sequential	0.92	0.95
Random	0.99	0.91
Financial	0.99	0.93
Exchange	0.93	0.99
MSN	0.98	0.95

What is Determination factor for WAF & PWO ?

- WAF and PWO are determined by characteristics of the I/O workloads

utilization

Workload	Scheme	# of Write req.	Avg. size of Write	Avg. *u* of victim blocks for GC
Sequential	RAID-5	368K	62K	0
	eSAP	184K	124K	0
Financial	RAID-5	3617K	9K	0.66
	eSAP	416K	78K	0.64



ECC Bit Correction Requirements Trends

- Bit requirements for BCH

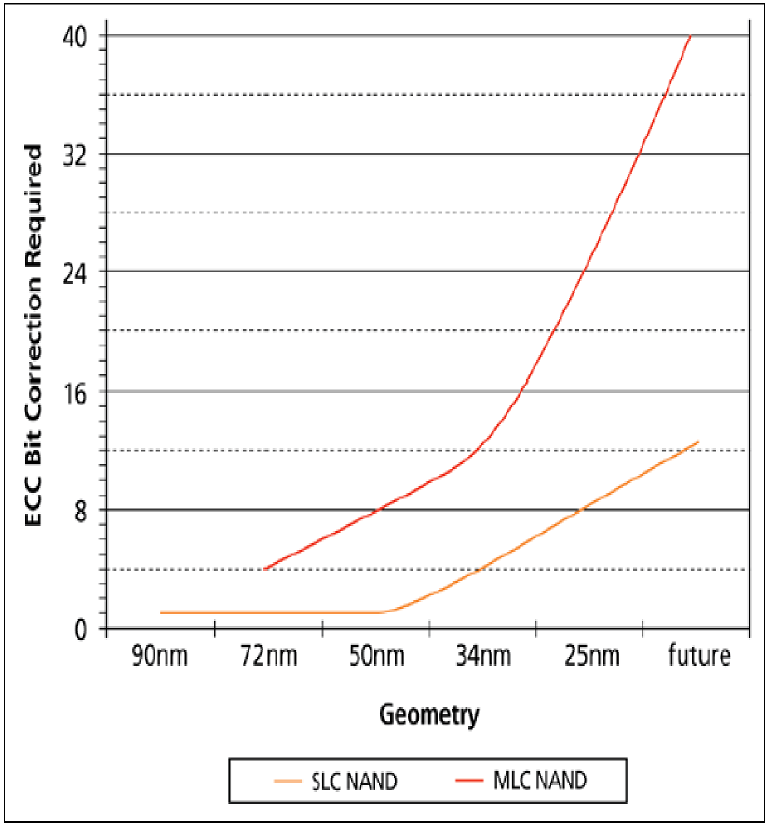
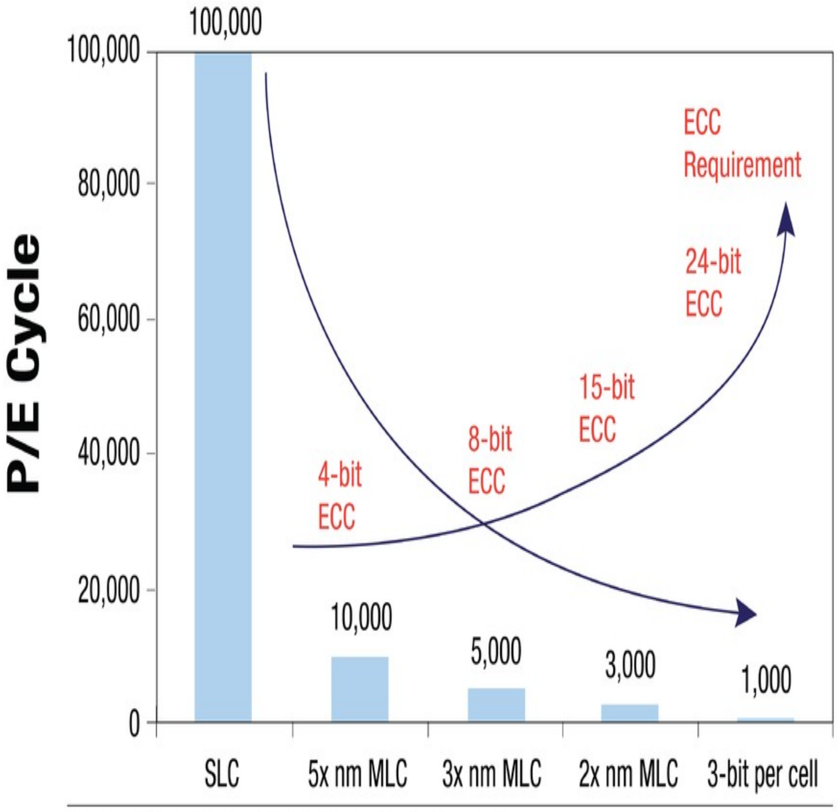


Figure 1. ECC Bit Correction Requirements Trend for SLC and MLC NAND

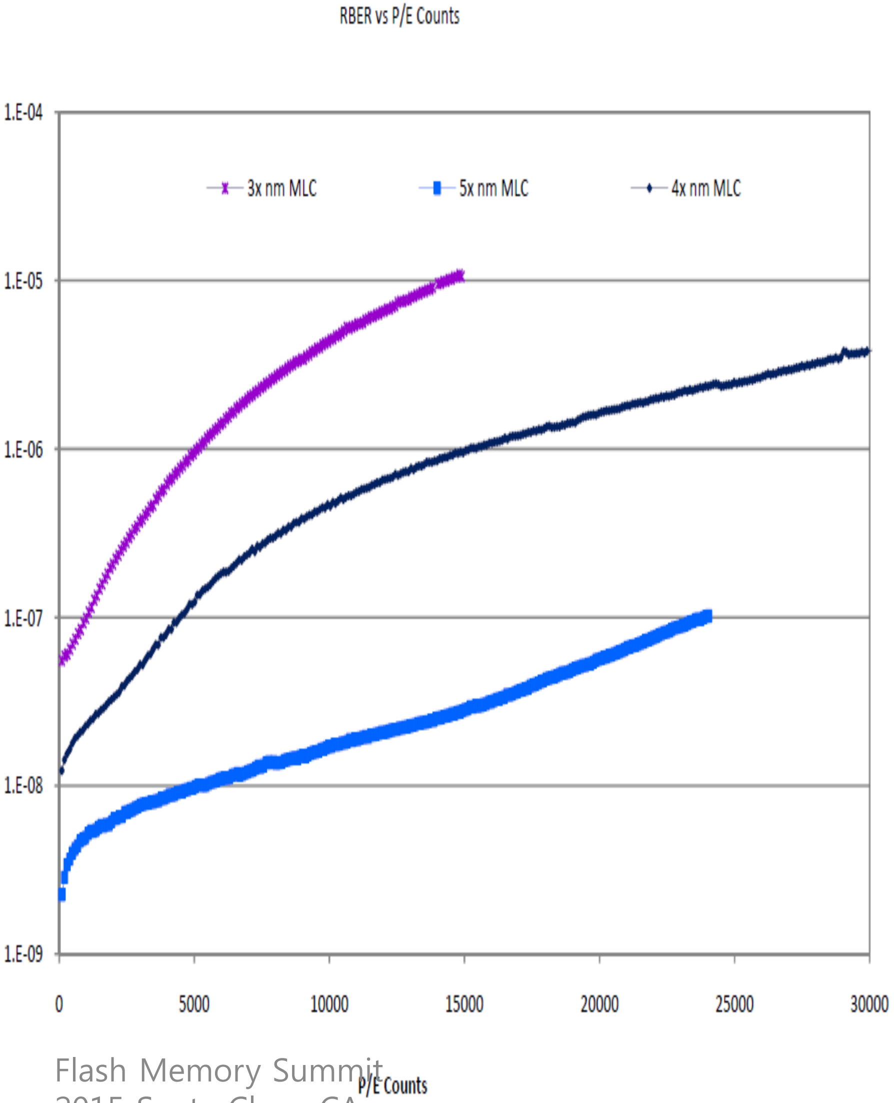


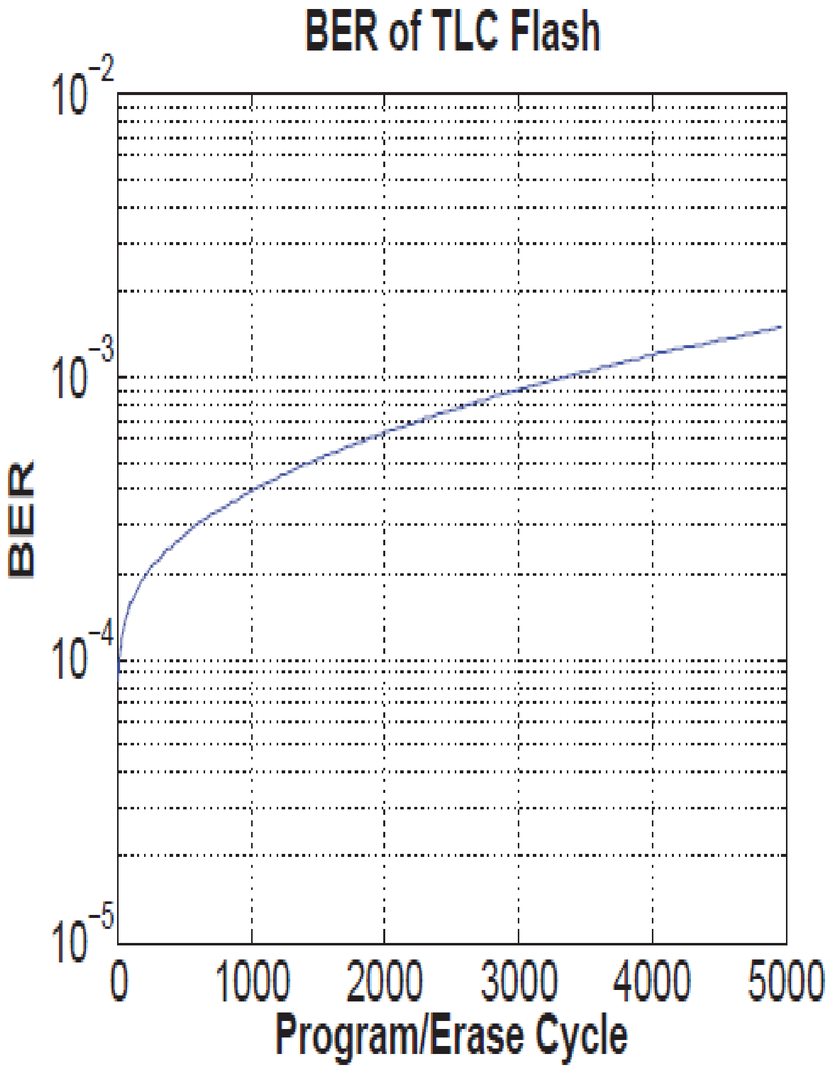
Source: JMicron, Western Digital, Morgan Stanley Research

From: 1) ECC Options for Improving NAND Flash Memory Reliability – Micron, 2012
2) Signal processing and the evolution of NAND flash memory – Anobit, 2010

RBER of MLC vs. TLC

BER of MLC Flash


RBER vs P/E Counts
3x nm MLC
5x nm MLC
4x nm MLC
1.E-04
1.E-05
1.E-06
1.E-07
1.E-08
1.E-09
P/E Counts


BER of TLC Flash
BER
10^{-2}
10^{-3}
10^{-4}
10^{-5}
Program/Erase Cycle