Predicting SSD Performance for Today's Dynamic Workloads

Shirish Bahirat, SSD Engineering System Architecture

Legal Disclaimer

- Presentation material is just conceptual representation of technicalities and for educational purpose only
  - Data represented in this presentation is not attributed to any product or application
  - Material contained in this presentation is copyright of Micron Technology
- Nothing in this presentation is considered as legal advice or official opinion of Micron Technology or presenter
  - Presenter do not assume any legal responsibility or liability arising due to use of any information in this presentation
  - No warranties, or no implied advice, use at your own risk

Contains

- Write amplification – basic terminology
  - Steady state write amplification
- Time domain performance
  - Why steady state write amplification analysis not enough?
- Components impacting dynamic write amplification
  - Logical saturation and degree of randomness – fresh out of box, streams, hot vs cold data, trim/unmap
  - Workload transitions – mix of sequential and random data

Write Amplification

Flash Memory Summit

Workload Visualization

Initial state → Sequential Write

Initial state

B1	B2	B2	B4	B5
B6	B7	B8	B9	B10
B11	B12	B13	B14	B15
B16	B17	B18	B19	B20
B21	B22	B23	B24	B25
OP	OP	OP	OP	OP
SP	SP	SP	SP	SP

NAND filled with sequential data, OP region empty

SP is spare region

Sequential Write

x	x	x	x	x
B6	B7	B8	B9	B10
B11	B12	B13	B14	B15
B16	B17	B18	B19	B20
B21	B22	B23	B24	B25
B1	B2	B3	B4	B5
SP	SP	SP	SP	SP

Incoming sequential data written in OP region, entire block stripe validated

Host Data Written = 1
GC Data Written = 0

Initial state → Random Write

Initial state

B5	B18	B10	B21	B14
B15	B25	B3	B6	B19
B24	B20	B12	B2	B9
B17	B1	B7	B13	B23
B4	B11	B22	B16	B8
OP	OP	OP	OP	OP
SP	SP	SP	SP	SP

Drive filled with random data, OP region empty

Random Write

B5	x	B10	B21	B14
B15	B25	B3	x	B19
B24	B20	B12	B2	B9
x	B1	B7	B13	x
B4	x	B22	B16	B8
B17	B6	B18	B23	B11
SP	SP	SP	SP	SP

Host random data written in OP region, 3 TU's need to be moved to create empty space

Host Data Written = 1
GC Data Written = 0.6

Write Amplification (WA)

Flash Memory Summit

- Physical (Written to NAND) data is more than logical (Written by Host) data
  - NAND Flash memory pages must erased before re-written (Erase Unit - Block)
  - NAND Flash device can be programed only once after erase (Program Unit - Page)
- Lots of research and work to understand WA
  - Generally computed as a function of Overprovisioning (OP)
  - Under steady state (degree of randomness is constant) workloads, WA stabilizes (close to a) fixed number

$$WA = \frac{1}{2}\left(\frac{1 + \sigma}{\sigma}\right)$$

Where σ denotes Overprovisioning factor

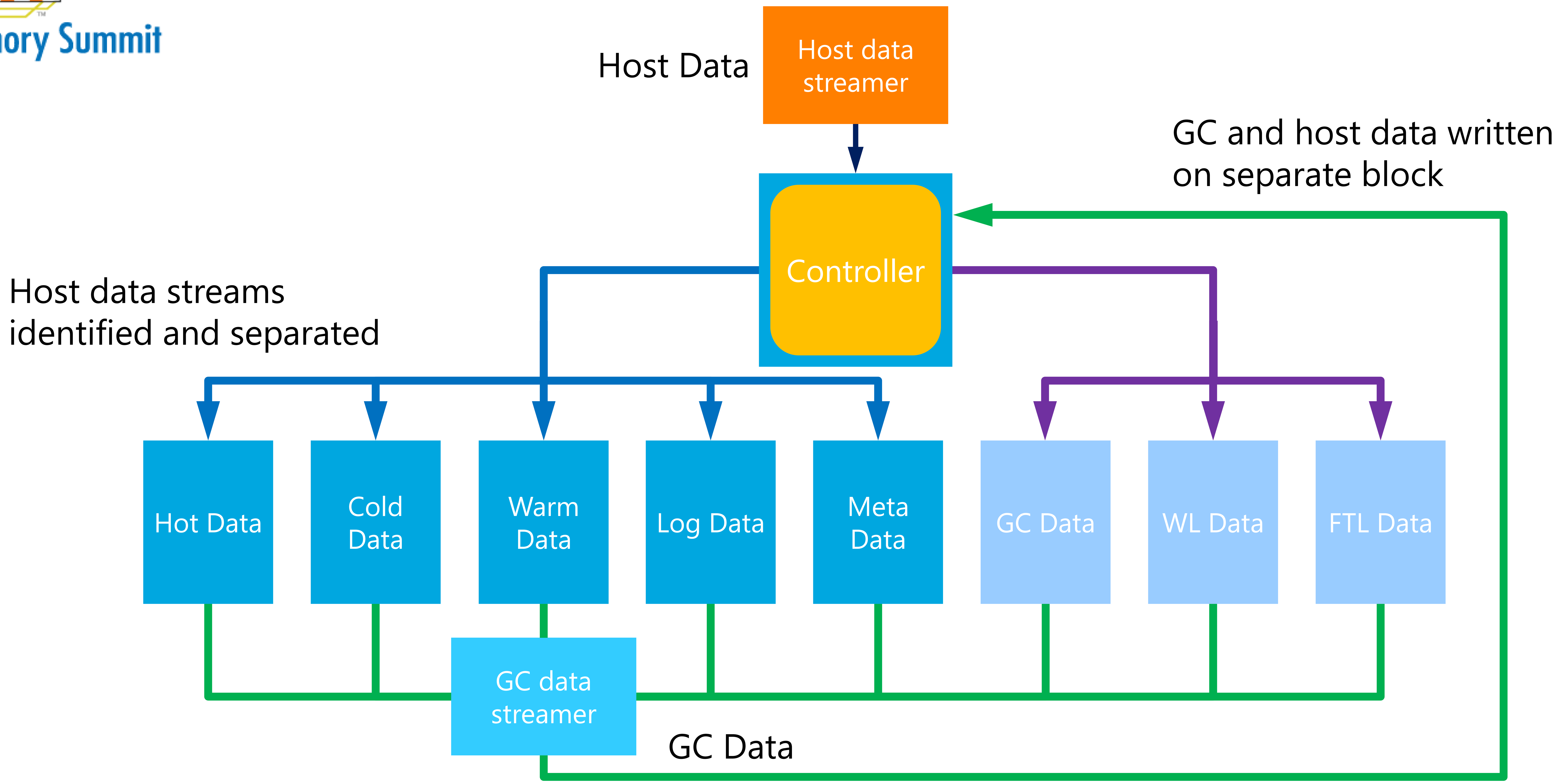
[Ref] Rajiv Agarwal and Marcus Marrow, "A closed-form expression for write amplification in NAND flash", in IEEE Globecom 2010 workshop on Applicat. of Commun. Theory of Emerging Memory Technologies, pp. 1908-1912 - also there is another paper on improved form of this equation

Real life workloads may not be steady state (or hard to predict)
Published equations fall apart under real life working conditions

Flash Memory Summit

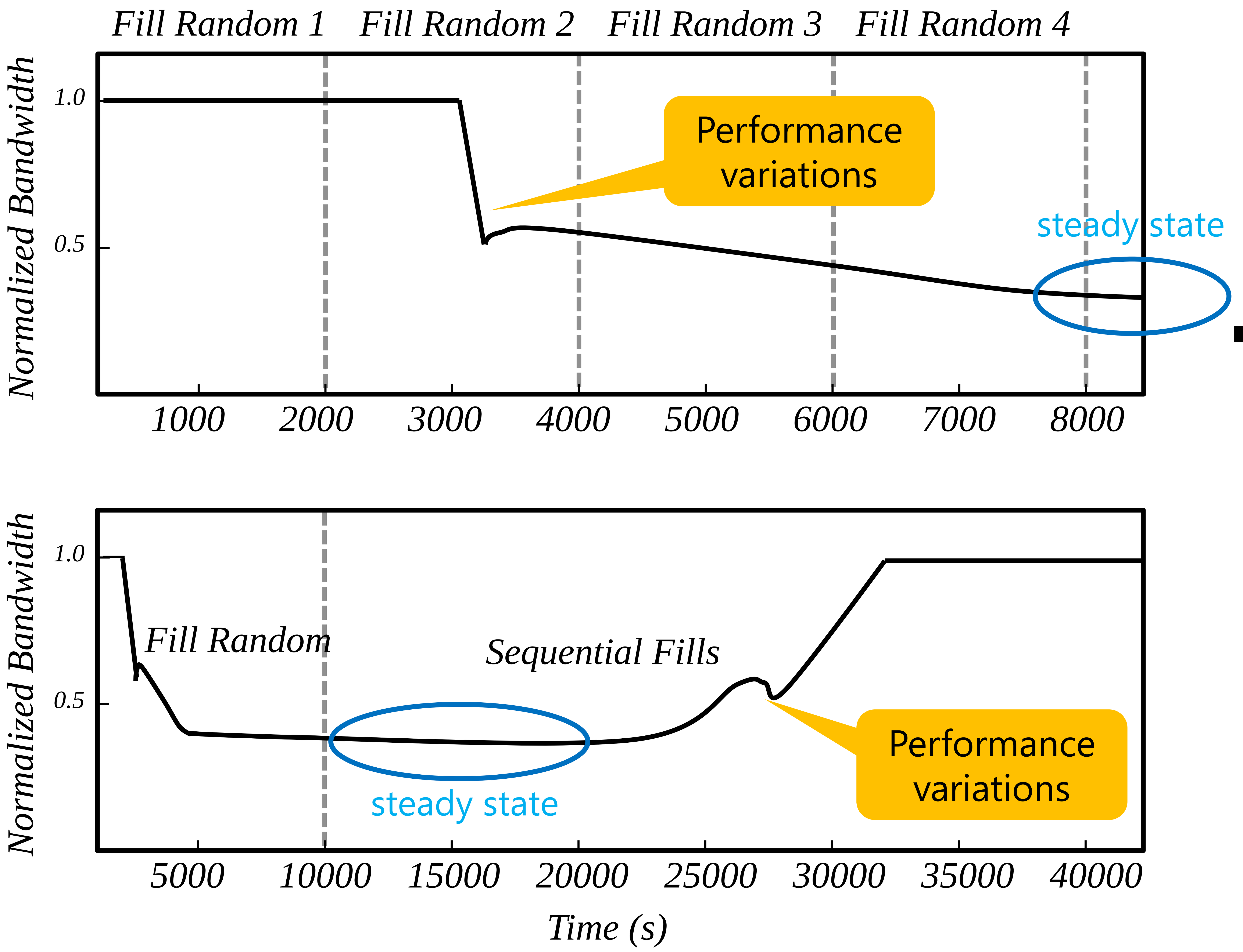
Garbage Collection Process (GC)

Separating incoming data based on expected life



Time Domain Performance

Typical SSD Performance – Bandwidth over time



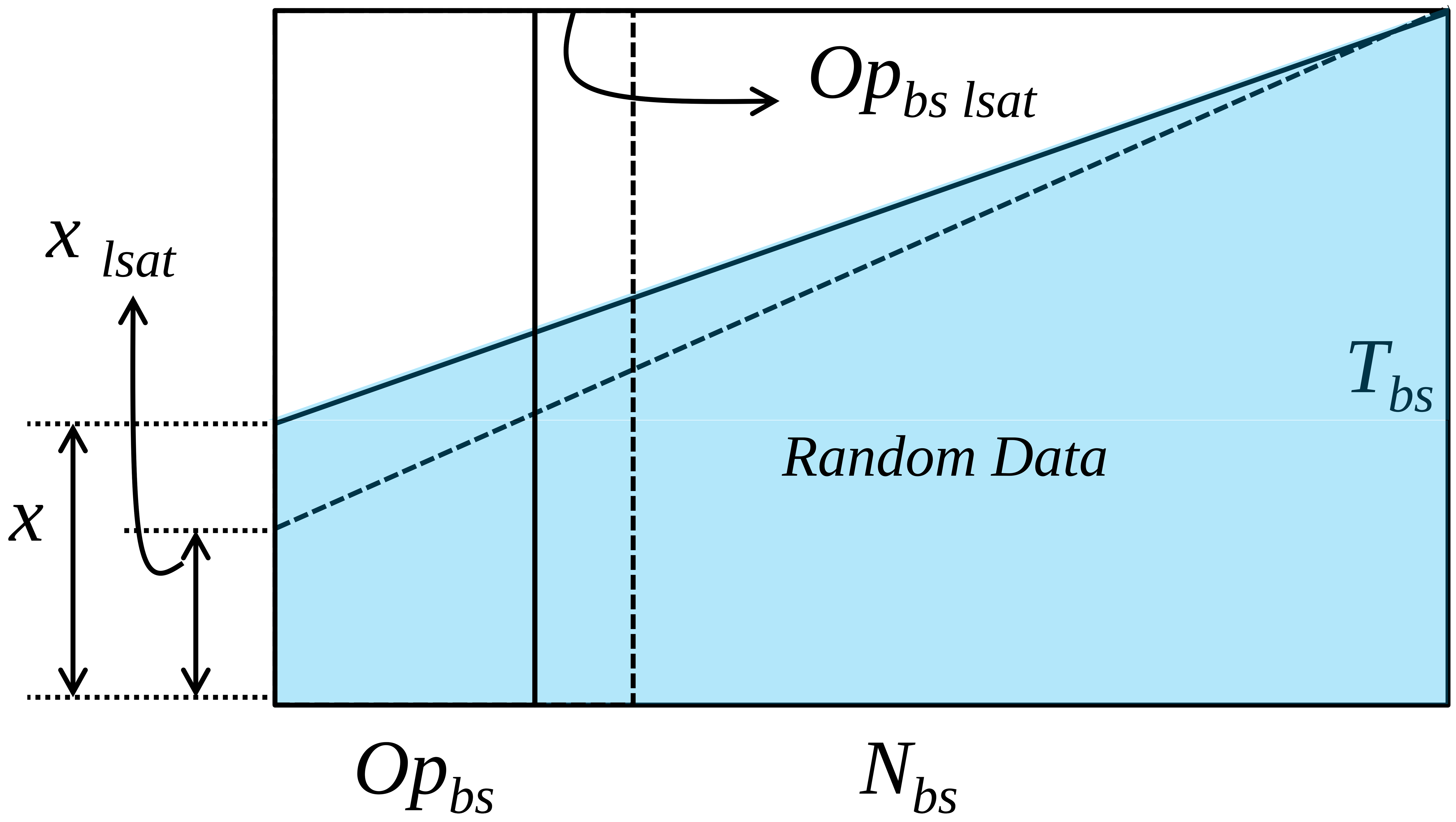
- Bandwidth changes over time
  - No WA when drive is Fresh Out of Box
  - As drive is filled WA kicks in, dropping performance
- Steady state nature of WA is fully understood
  - Little efforts understanding performance consistency
  - Mechanism that impacts write amp under dynamic workload conditions

We will look at 2 key aspects for performance variance
1. logical saturation
2. random to sequential workload transition

Logical Saturation and Degree of Randomness

Flash Memory Summit

Logical Saturation – through simple geometrical mapping



Overprovisioned space ≈ Invalided TU generated free space

- Logical Saturation
  - The portion of user logical block addresses (LBAs) that contain data
- Physical Saturation
  - The portion of physical NAND locations that contain data

T_{bs} - *Translation units (TU) per blocks*

N_{bs} - *Logical capacity blocks*

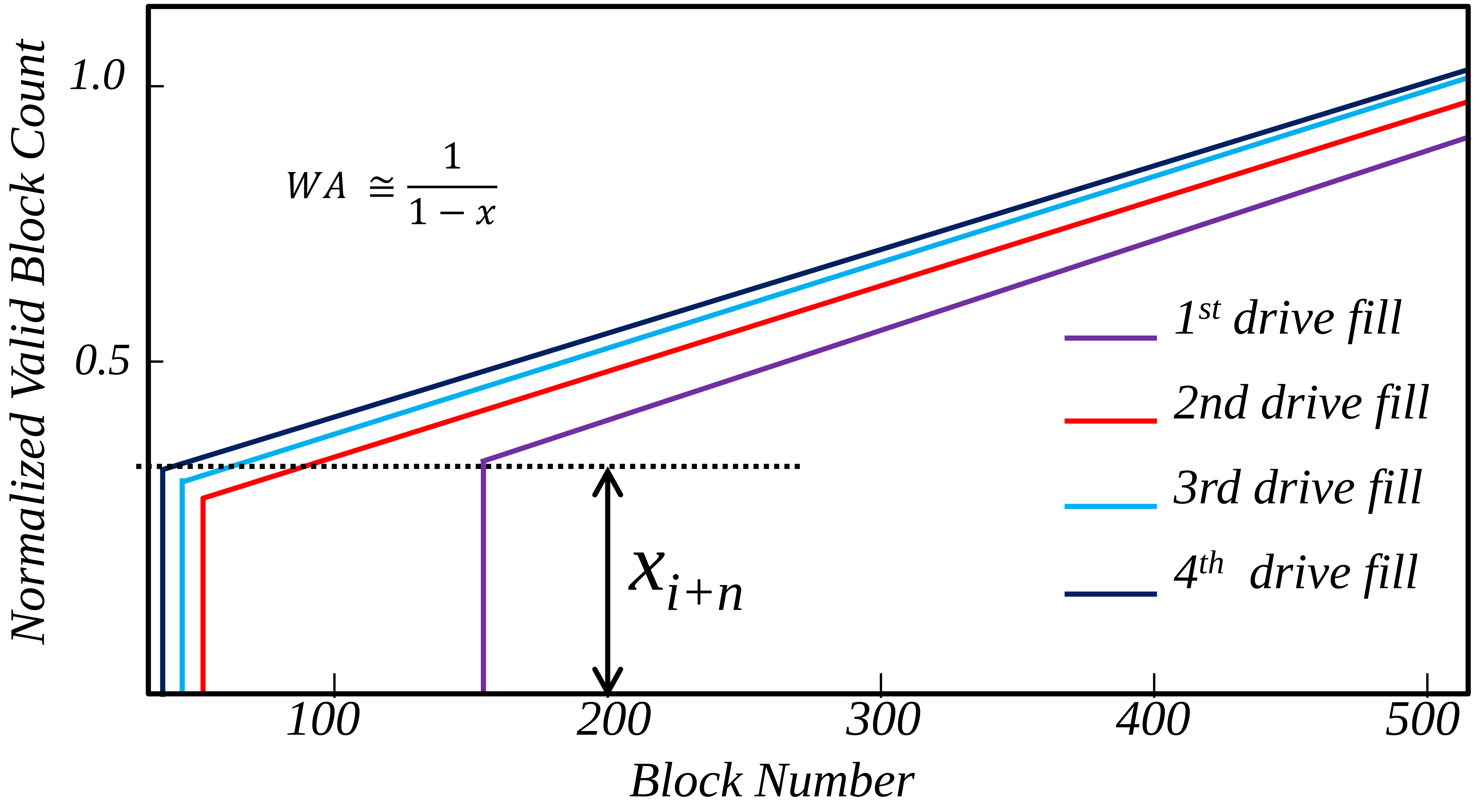
Op_{bs} – *Blocks providing over provisioning*

x – *valid Translation unit count on GC block*

$Op_{bs\ lsat}$ – *Blocks available due to lower logical saturation*

x_{lsat} – *valid TU count due to lower logical saturation*

Projecting Valid TU Count as a Function of Logical Saturation



- Projecting x as function of new blocks written

$$x_{i+1} = \left[1 - \frac{1}{Tu_{Logical}}\right]^{(Tb_i - Tb_{i-1}) * Gc_j}$$

  - x_{i+1} - Valid TU count on the GC victim block
  - $Tu_{Logical}$ - Logical capacity of the drive
  - $(Tb_i - Tb_{i-1})$ - New TUs written on last block, i.e. TUs per block – GC TUs moved *(x_j)* for *j* the block (Gc)

[Iterative form of] Xiao-Yu Hu , Evangelos Eleftheriou , Robert Haas , Ilias Iliadis , Roman Pletka, Write amplification analysis in flash-based solid state drives, Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference, May 04-April 06, 2009, Haifa, Israel

Enables to projection of write amplification with any block range, any form of logical saturation, basically time domain WA as a function of number of random TUs written

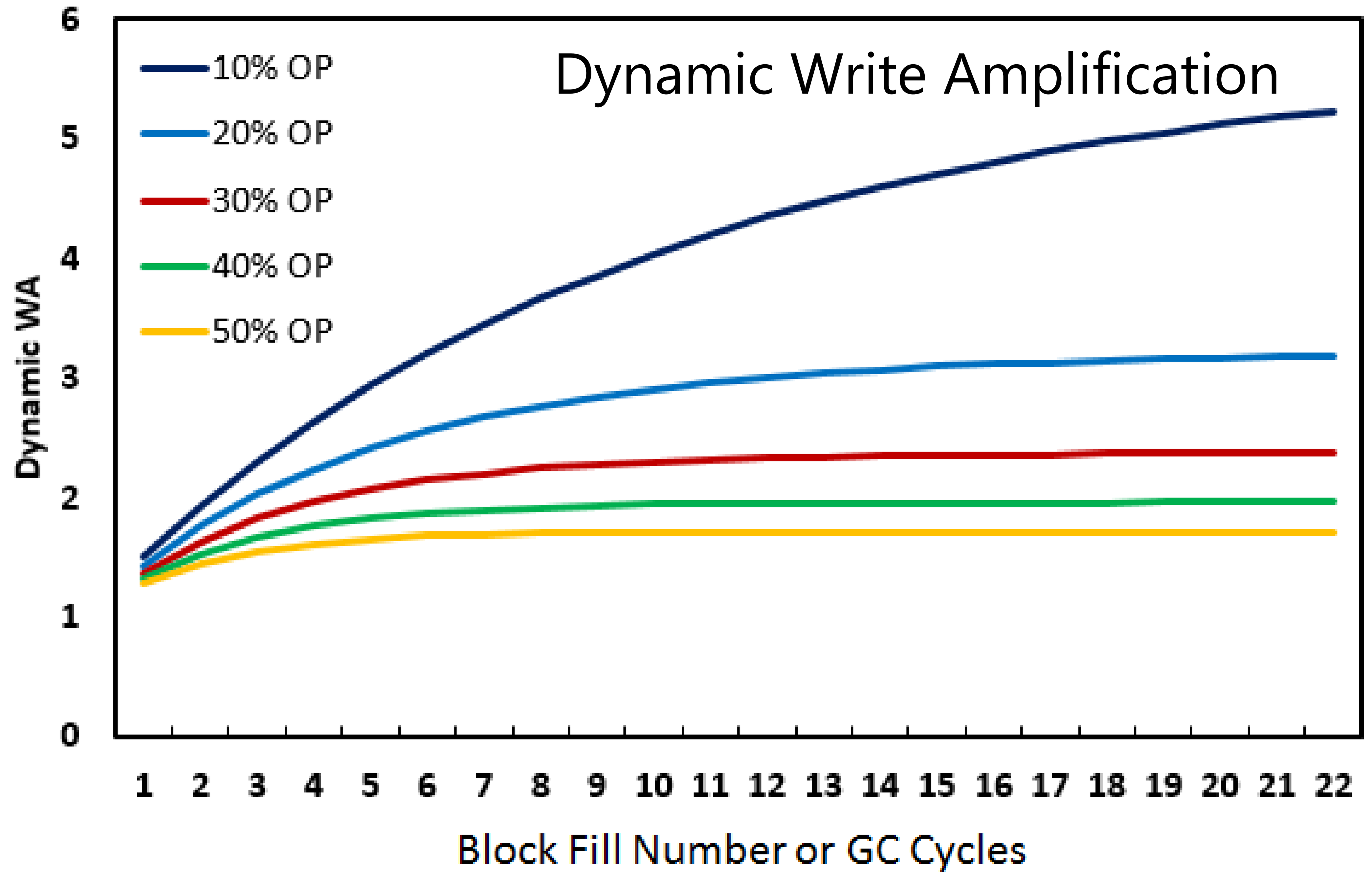
Alternative (simple and less accurate) f

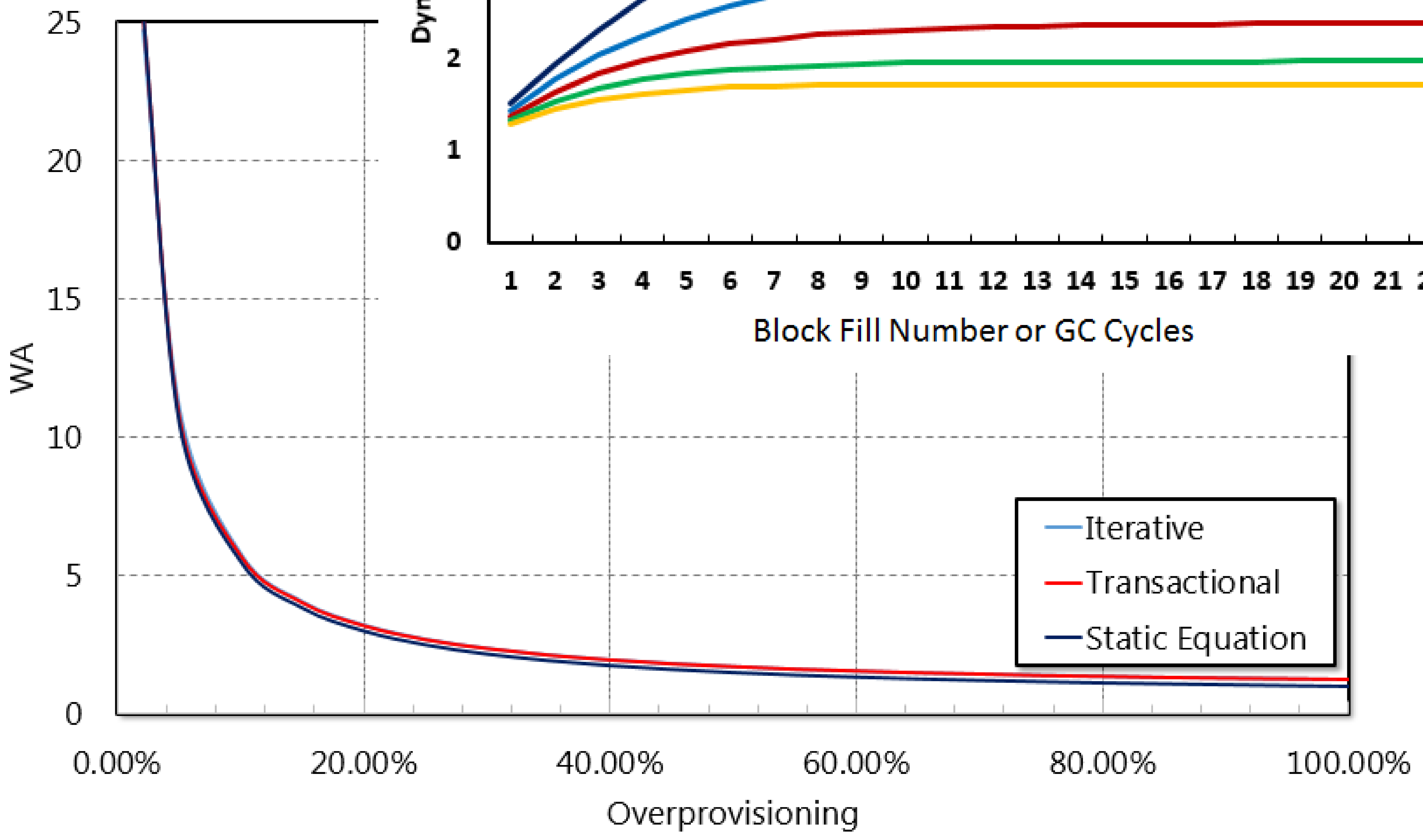
- Function of number of Overprovisioning block stripes and logical saturation – static equation

$$x = T_{bs}\left[\frac{N_{bs} - Op_{bs}}{N_{bs} + Op_{bs}}\right]$$ Ignoring logical saturation

$$x = T_{bs}\left[\frac{(N_{bs}\,(2-l_{sat})) - Op_{bs}}{(l_{sat}\,N_{bs}) + Op_{bs}}\right]$$ Considering logical saturation

- Correlation with this methodology to transactional WA model





Iterative form we derived (1st of its kind) correlates WA within 1% to transactional model
Simpler form correlates within 5% at lower OP and within 20% to higher op (similar to some published work before)

Intermixing of Sequential and Random Content

Flash Memory Summit

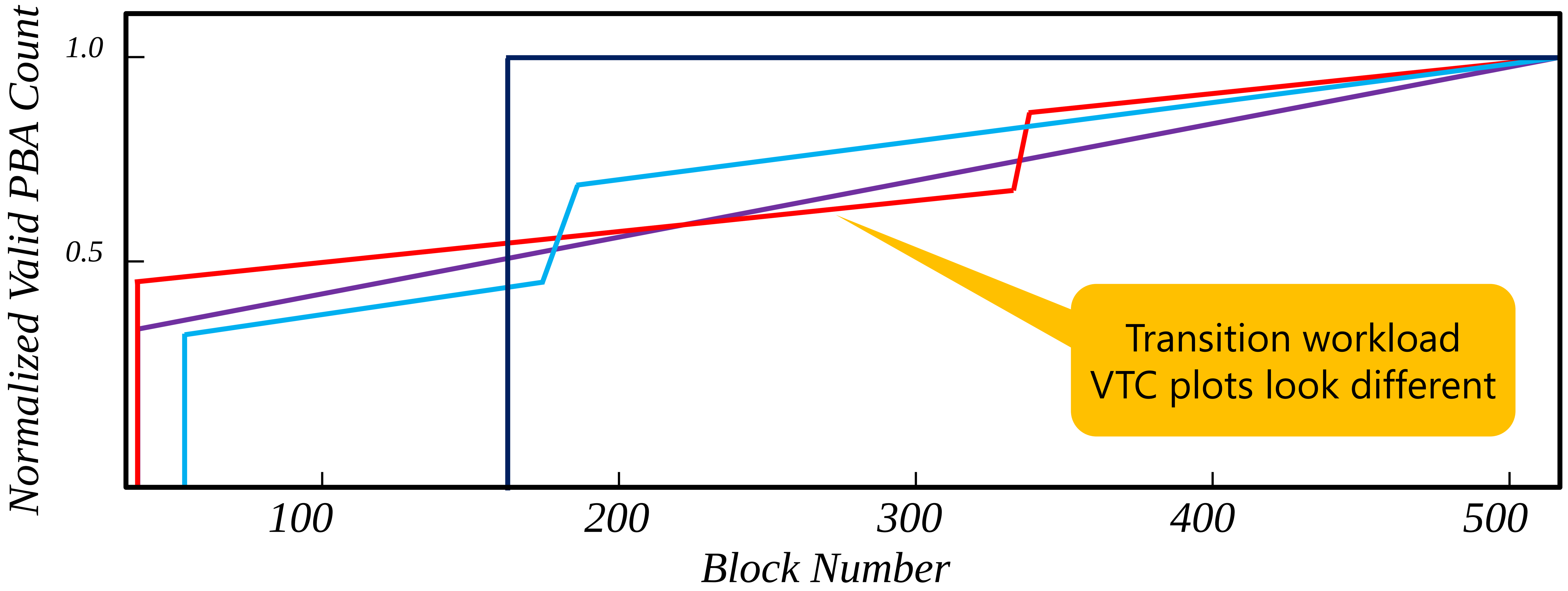
Dynamic WA during Workload Transition

- After drive preconditioned with random data, subsequent sequential fills does not return the bandwidth to nominal state quickly
  - Nominal bandwidth correspond to sequentially preconditioned drive, executing sequential IOs

Understand how the performance recovery process works

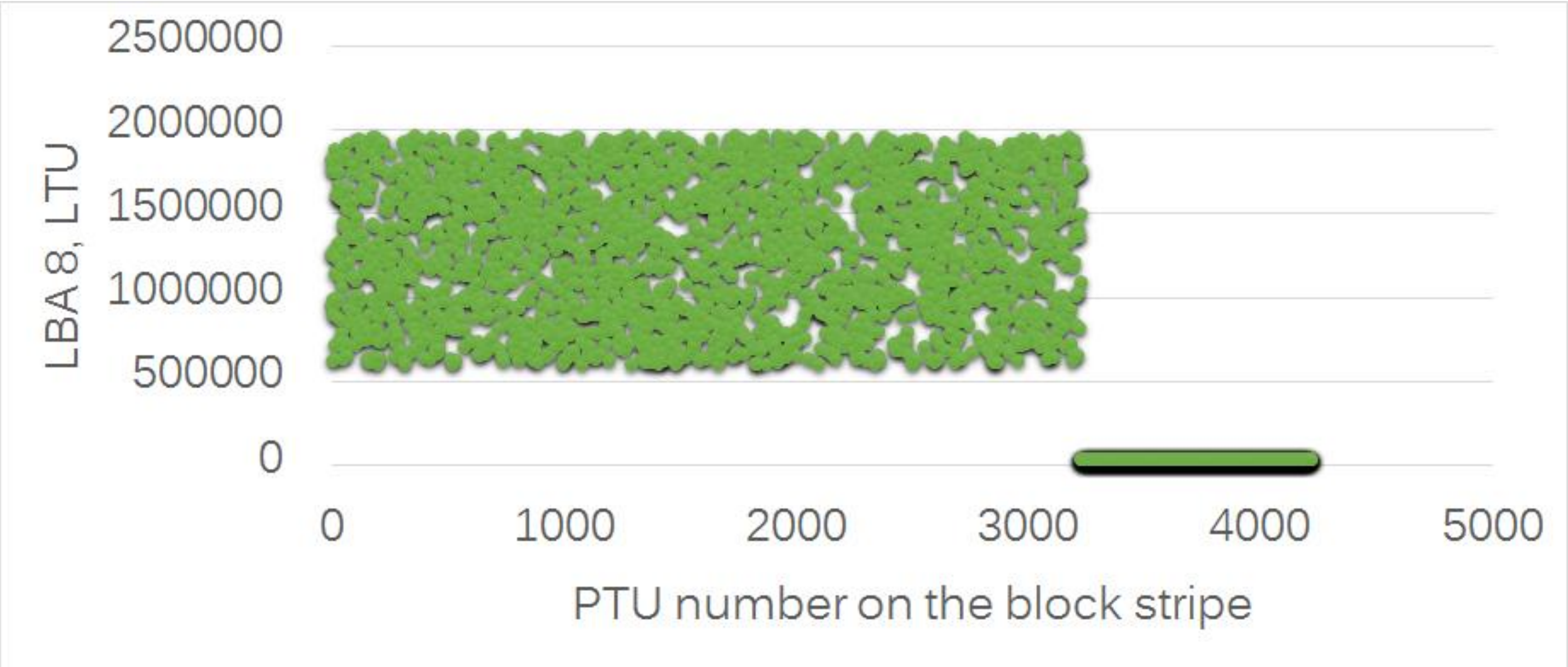
Modeling to Gain Insight – what we learned

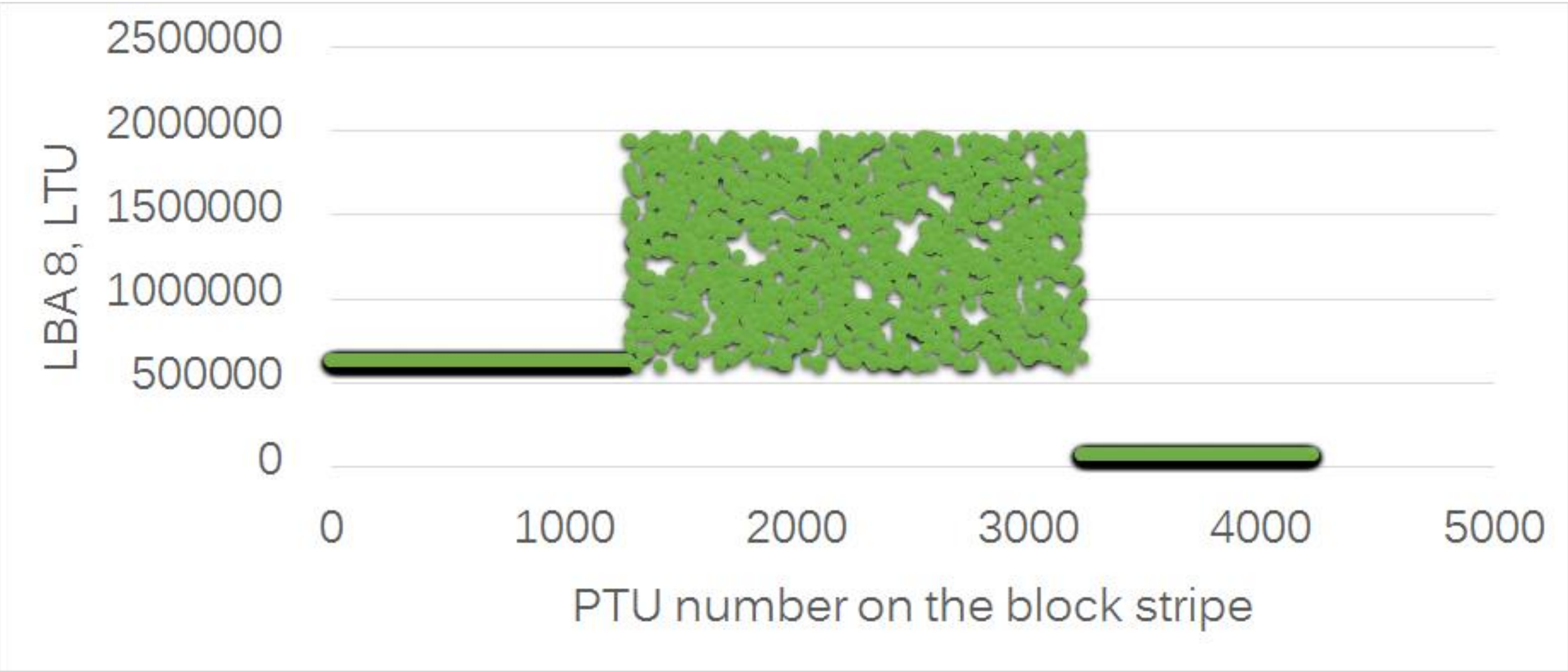
- Sequential BW with Sequential Preconditioning
  - Incoming host data invalidates entire block stripe, no valid data left to move for garbage collection process
- Sequential BW with Random Preconditioning
  - Garbage collection intermix random data and sequential data for a single cursor drive

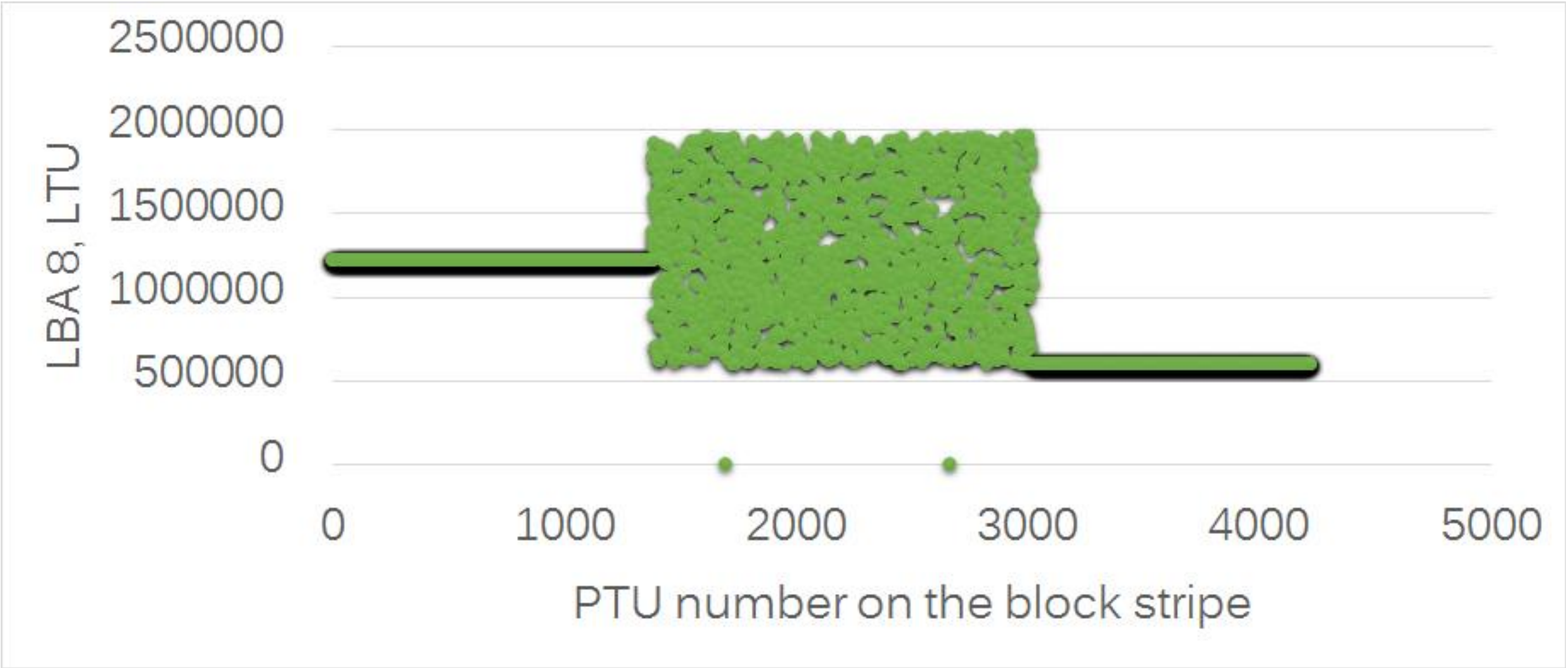


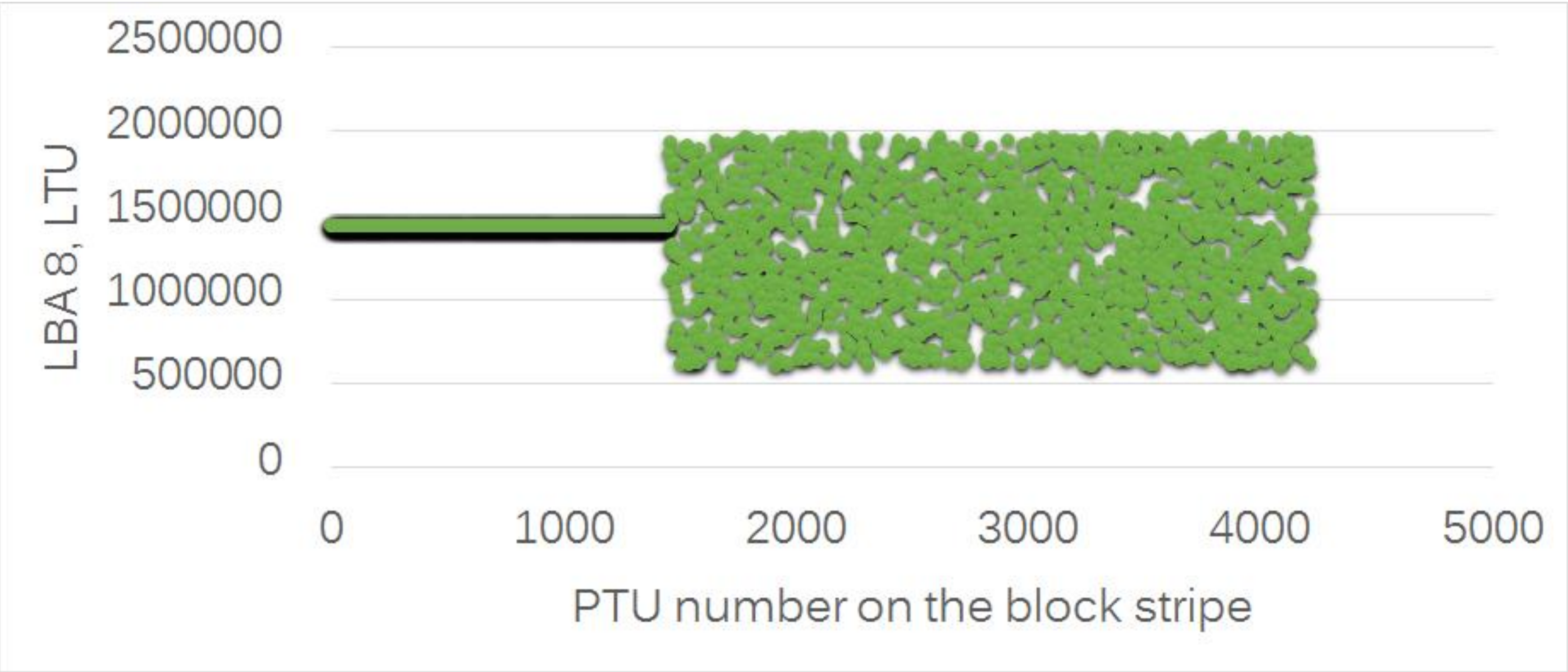
To understand performance during transition we need to understand VTC curves

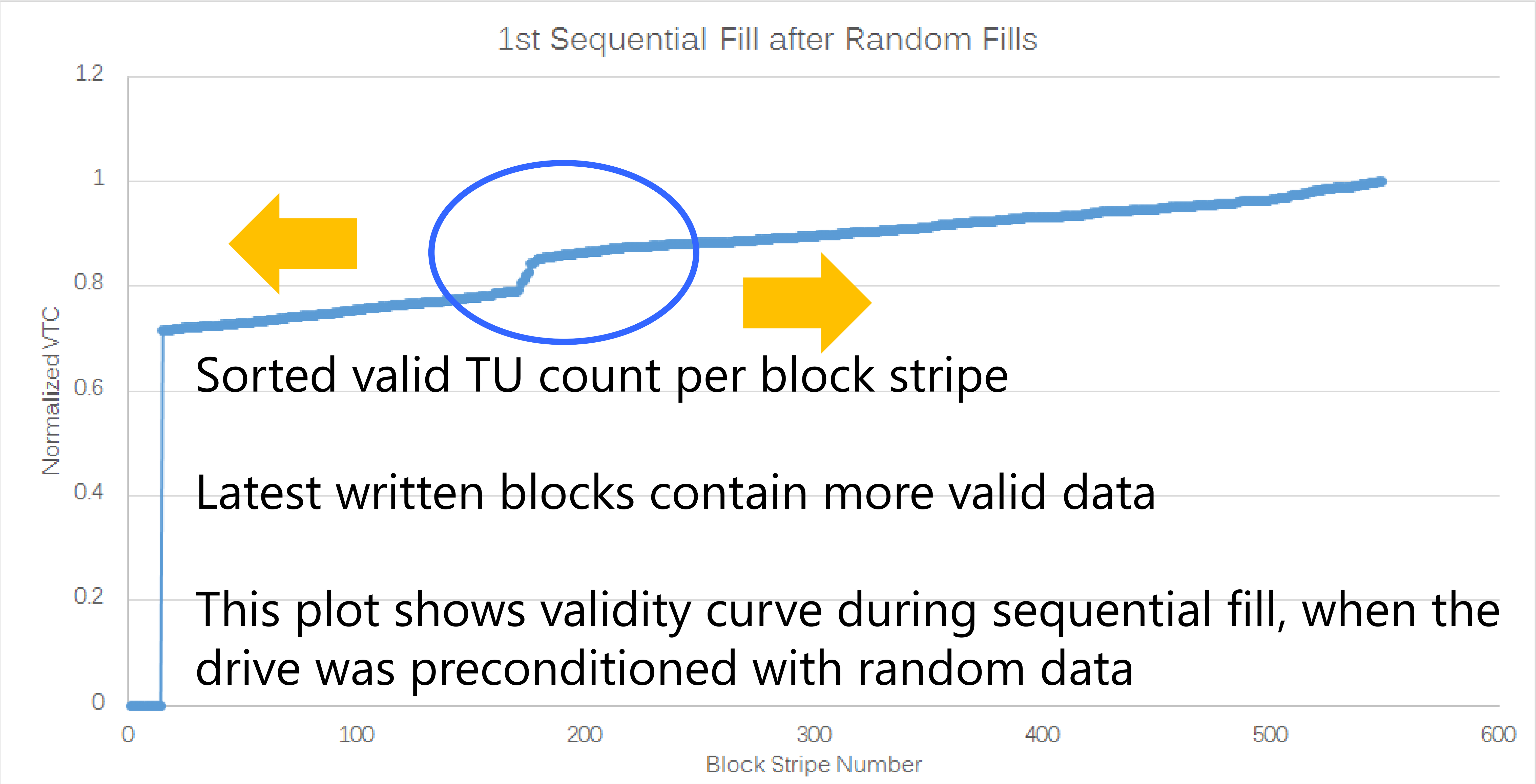
1st Sequential fill after random fills at 25% logical saturation











Flash Memory Summit

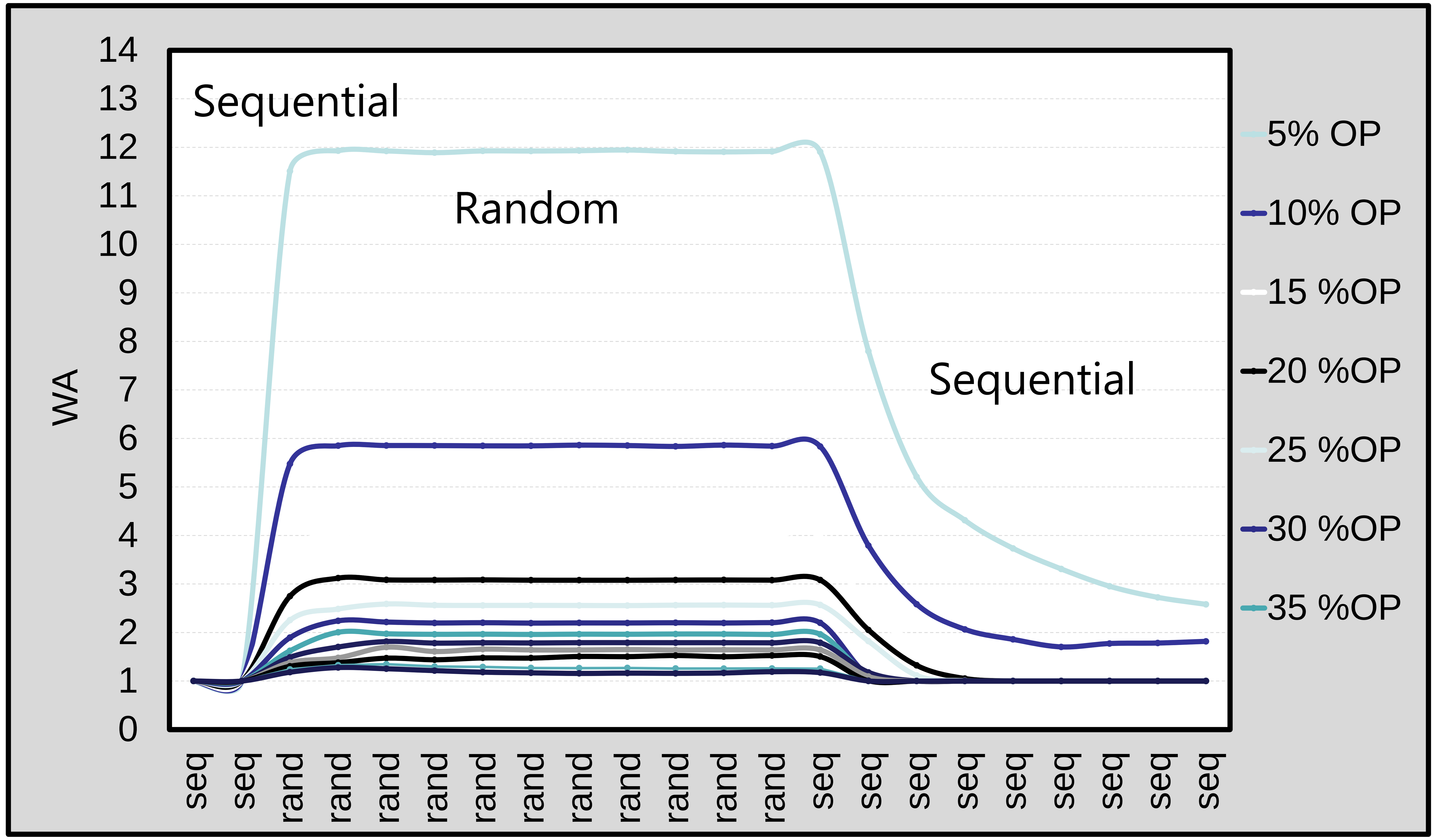
50% sequential fill after random fills

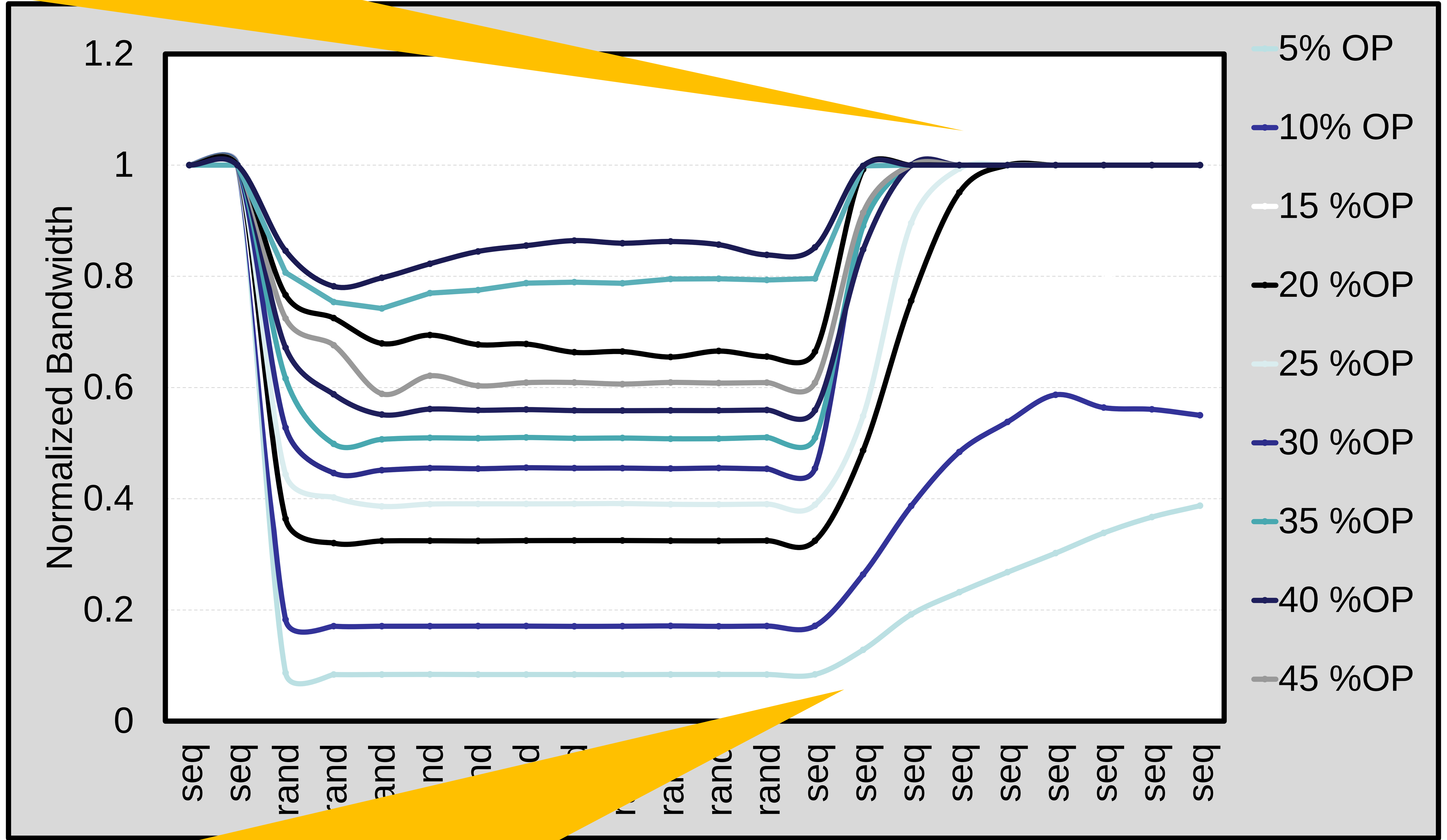
Flash Memory Summit

100% sequential fill after random fills

Bandwidth and Write Amp Vs. OP during workload transitions

On high OP, we get back to the Sequential Write performance after 2 drive fills

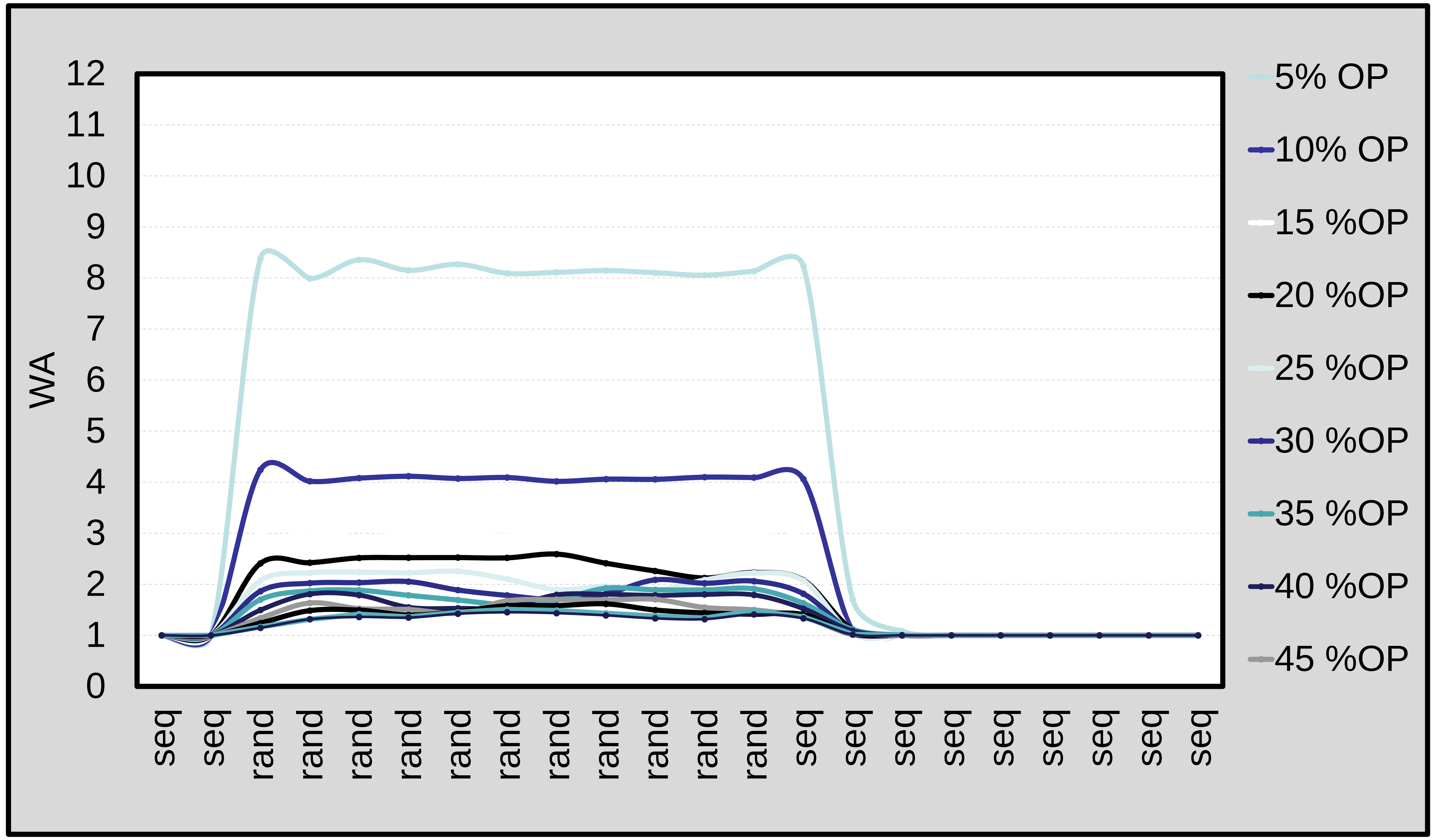




On low OP, it can take as much as 6 drive fills to back to the Sequential Write performance

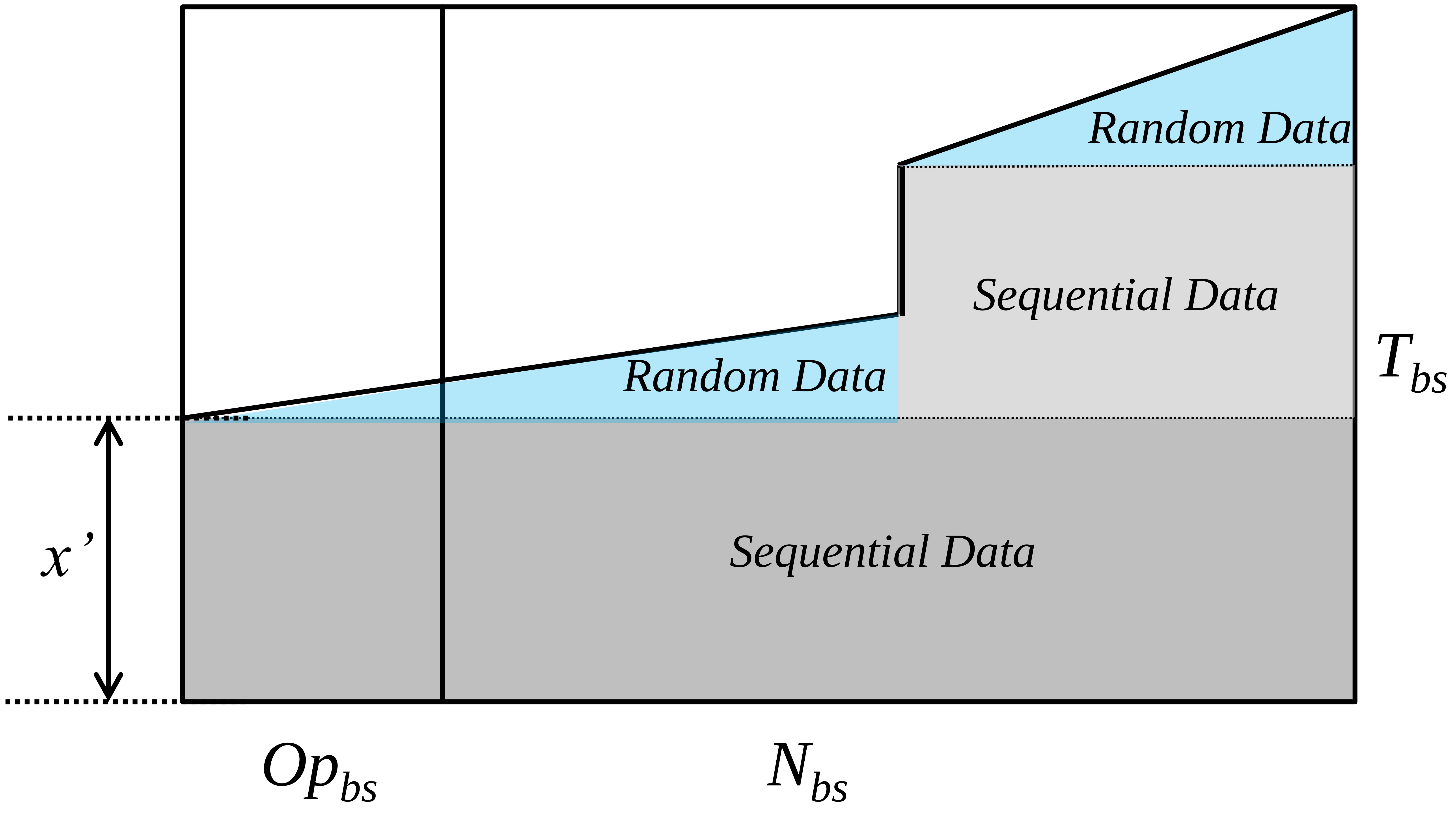
Flash Memory Summit

Multi-cursor Recovery Process



- Drive can recover performance with 1 fill even for low OP configurations
- GC random data and Host sequential data is written to separate block stripes
- Avoiding intermixing of the sequential and random data
- However in real life there is always possibly even host sending intermixed sequential and random data

Mathematical Formulation for Intermixed Workload



- Can be extended form of previous equations presented
- Number of discontinuous sections are function of Overprovisioning and WA
- Slope of VTC curve can be projected using iterative form presented previously

Summary

Flash Memory Summit

Dynamic WA is key to understand performance consistency

- Real life workloads can include number of variables impacting write bandwidth – trim/unmap, sequential/random, hot/cold etc
- Presented 1st of kind work (to best of our knowledge) to better understand how WA impacts performance consistency
- Demonstrated feasibility to characterize dynamic WA

Micron®