HotStorage 2025

Revisiting Trim for CXL Memory

Hayan Lee[1], Jungwoo Kim[2], Wookyung Lee[1], Juhyung Park[2], Sanghyuk Jung[3], Jinki Han[3], Bryan S. Kim[4], Sungjin Lee[5], Eunji Lee[1]

[1]Soongsil University, [2]DGIST, [3] Eeum, [4]Syracuse University, [5]POSTECH

Contents

Rising Demand for High-capacity Memory in AI Era

- **Fueled by data-centric and AI/ML applications**

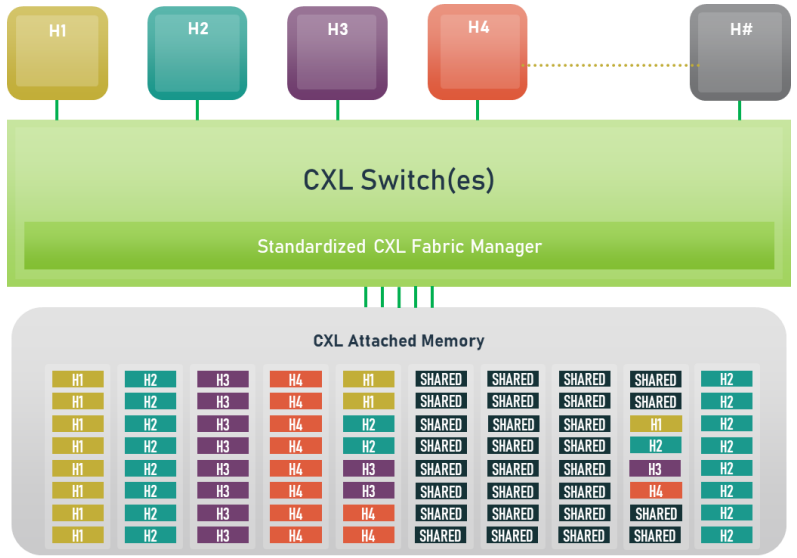
- **Memory disaggregation emerges as a key trend**
  - Realize scalable memory capacity by pooling distributed memory resources
  - Interconnect technologies are attracting significant attention

CXL-Flash

- **Compute Express Link (CXL)**
  - Driven by Intel (2019)
  - Cache-coherent interconnect technology
  - Enables multi hosts and devices to share a common memory space

- **CXL-enabled flash memory**
  - Provide high capacity
  - Hide long latency with intelligent prefetching and data placement
  - E. g. Samsung CMM-H

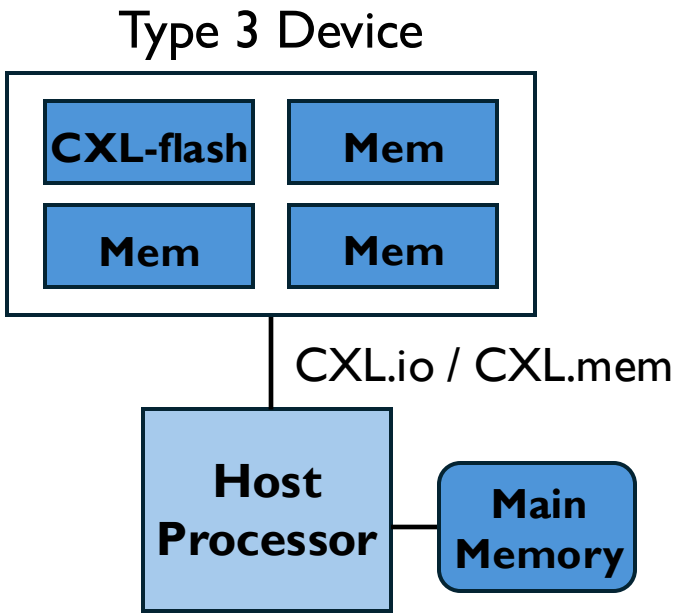
CXL Consortium



https://computeexpresslink.org/blog/explaining-cxl-memory-pooling-and-sharing-1049/

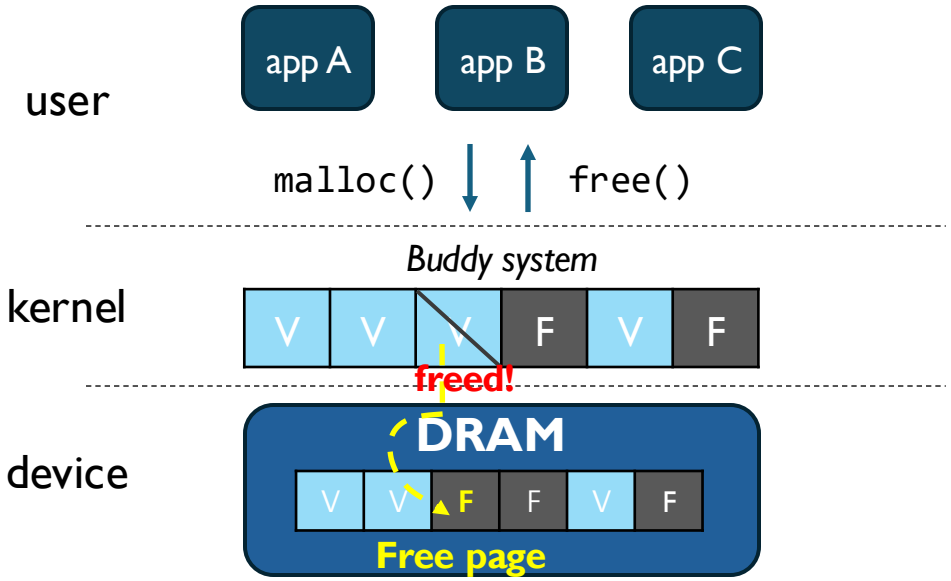
CXL-Flash

- **CXL-Flash**
  - CXL Type 3 Device
  - CXL.io / CXL.mem
  - Accessed in cache-line units (i.e., 64B) by host processor



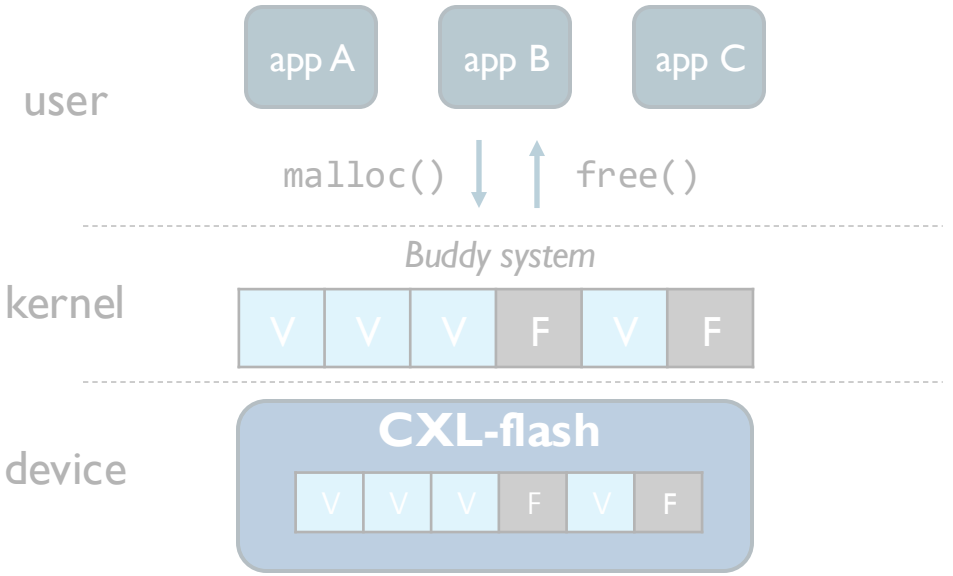
Challenges of CXL-Flash as Memory Module

When DRAM used



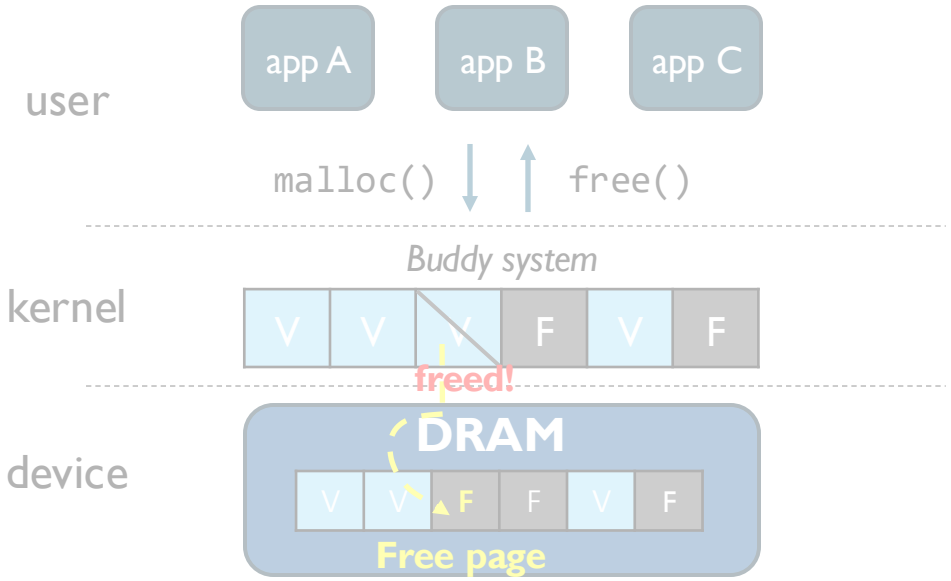
Zero cost for retaining invalid data

When CXL-Flash used



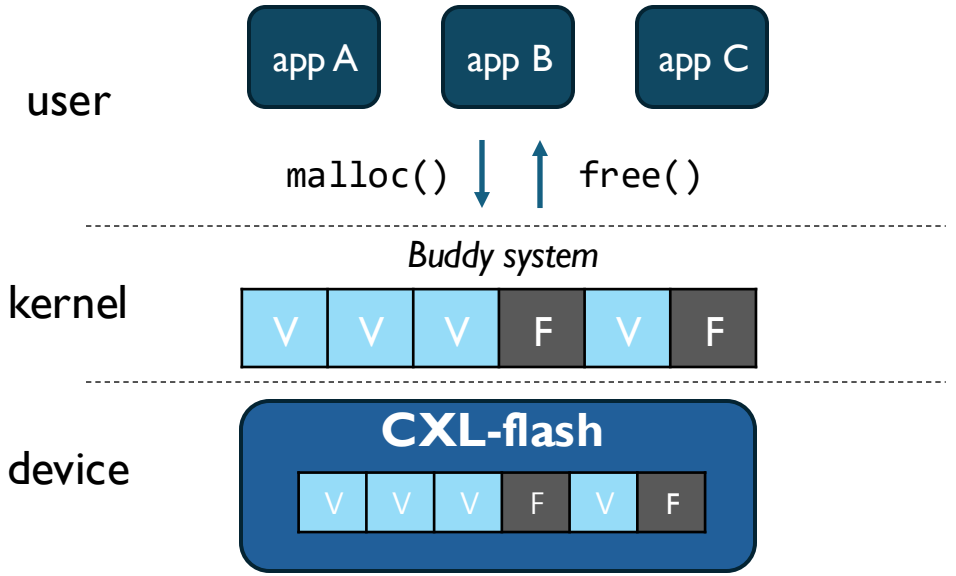
Challenges of CXL-Flash as Memory Module

When DRAM used



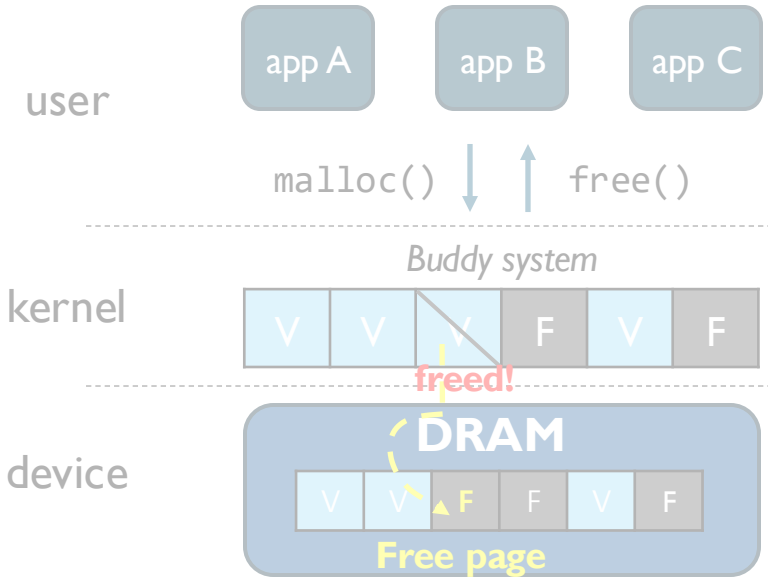
Zero cost for retaining invalid data

When CXL-Flash used



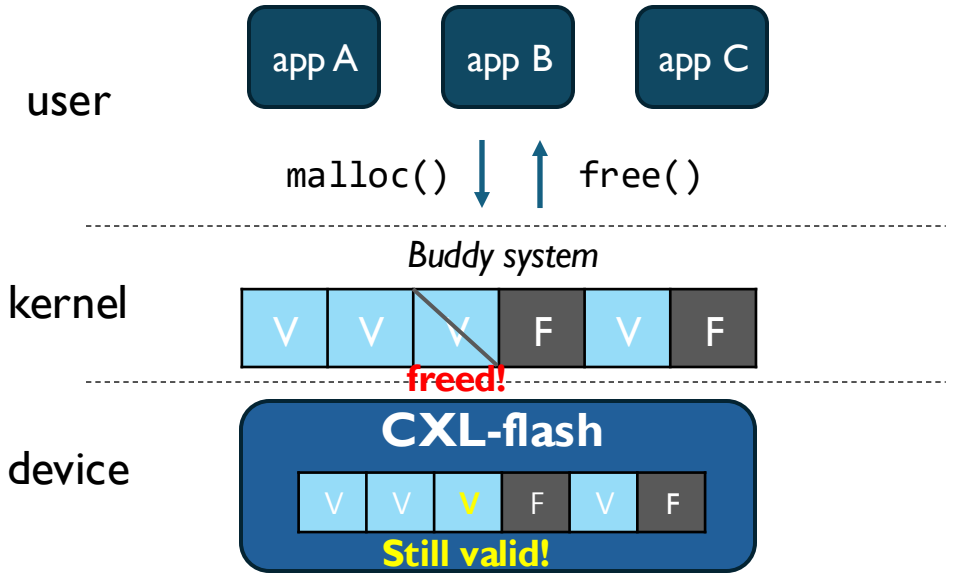
Challenges of CXL-Flash as Memory Module

When DRAM used

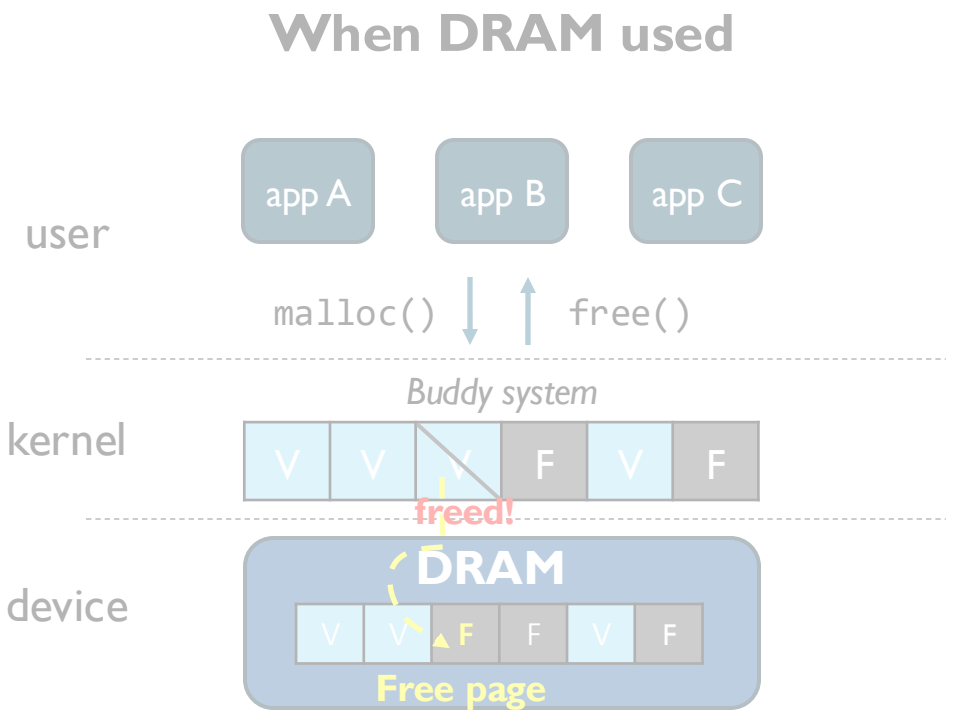


Zero cost for retaining invalid data

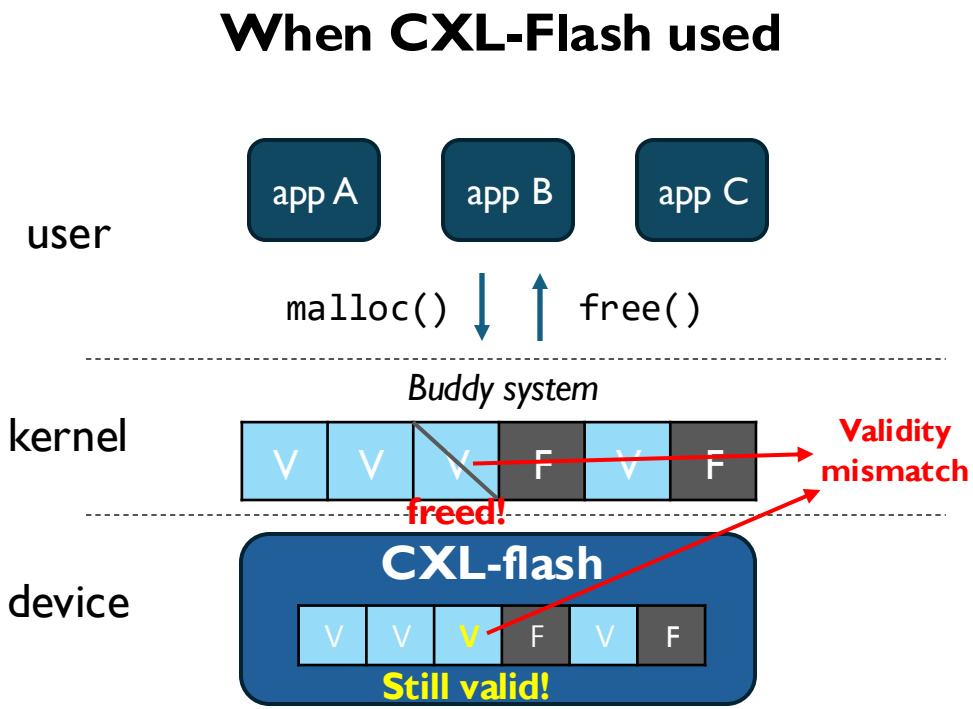
When CXL-Flash used



Challenges of CXL-Flash as Memory Module

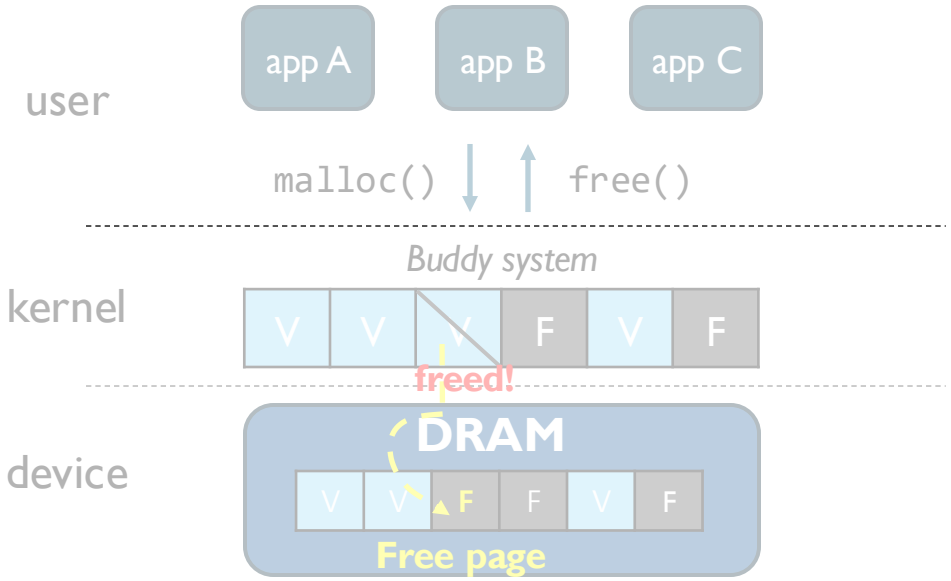


Zero cost for retaining invalid data



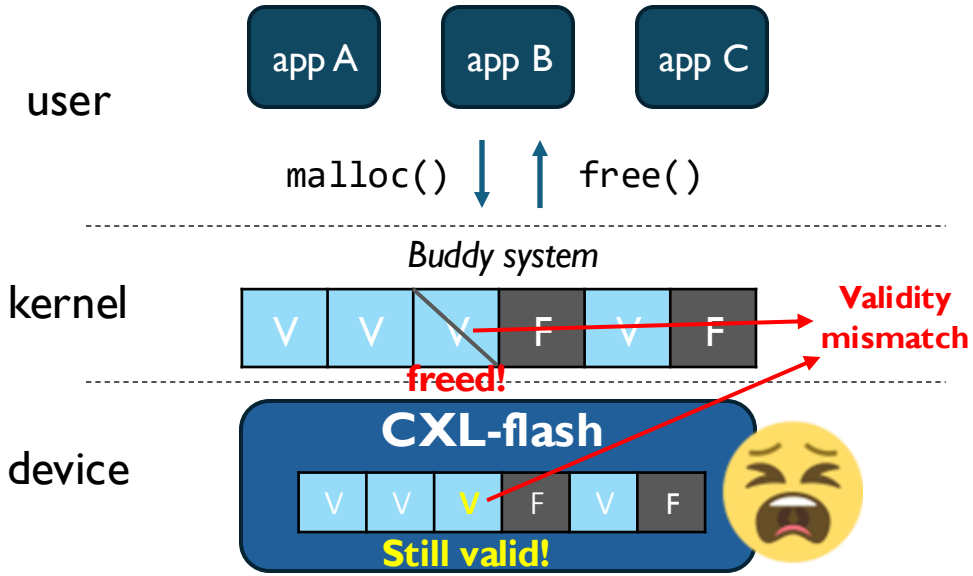
Challenges of CXL-Flash as Memory Module

When DRAM used



Zero cost for retaining invalid data

When CXL-Flash used



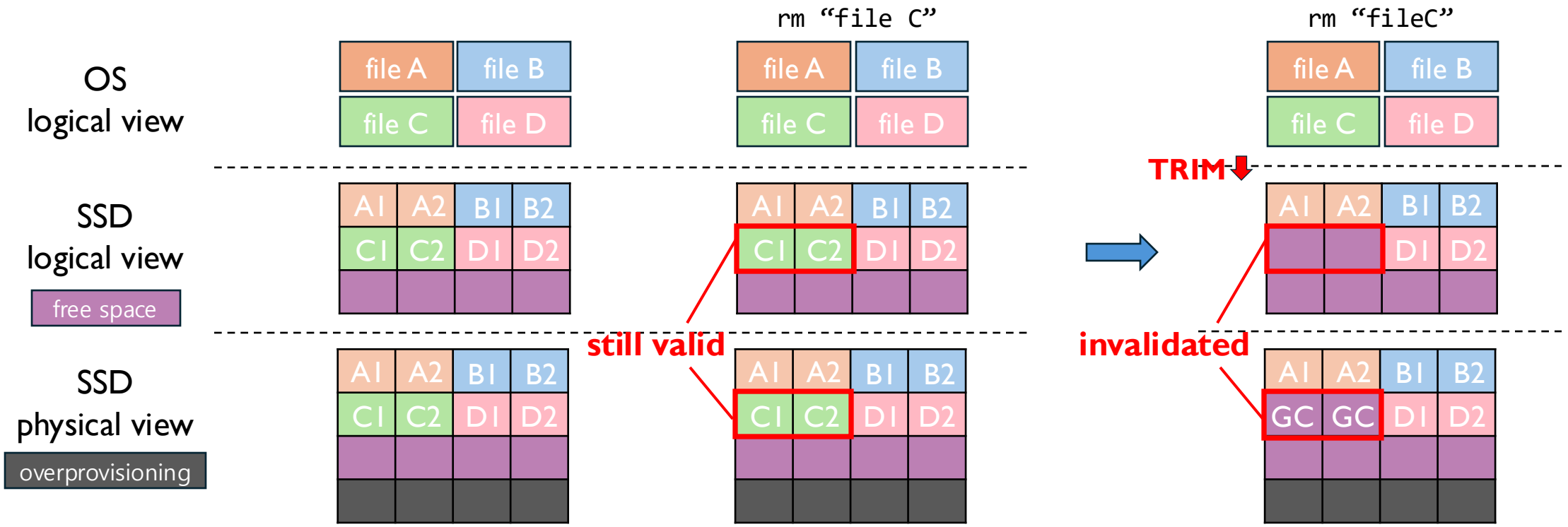
Will be copied during GC !

Need additional cost for invalid data!

- **Write amplification problem**
- **Fatal to lifetime-limited flash**

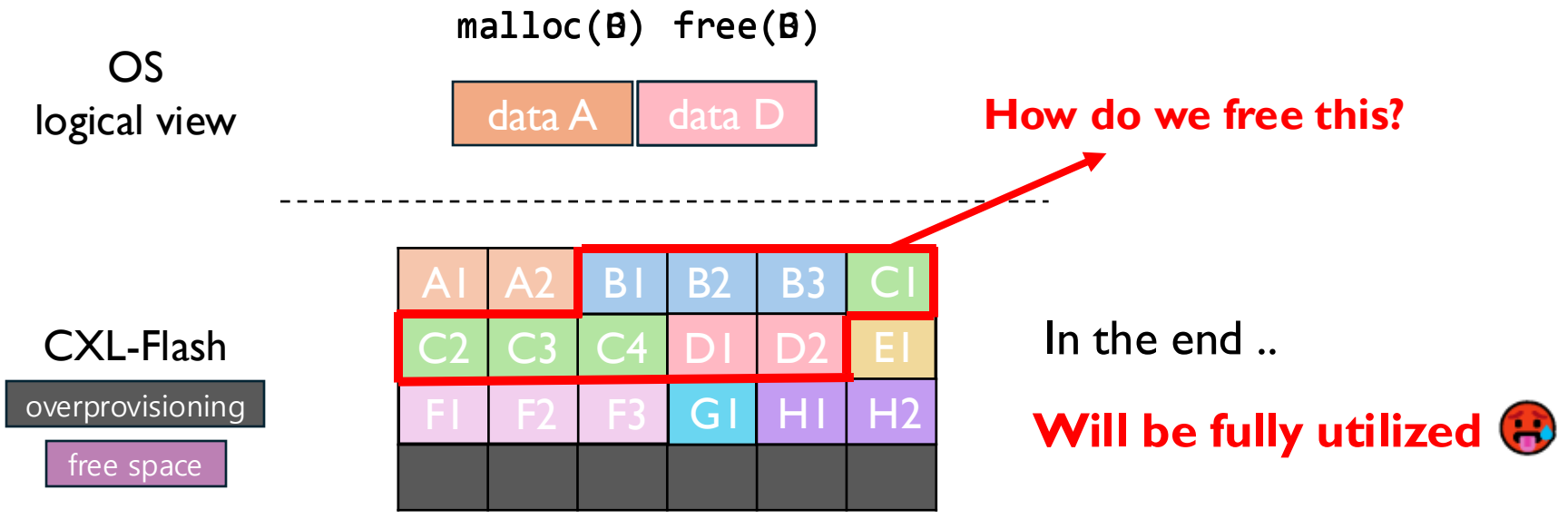
TRIM for Conventional SSD

- **TRIM command**
  - Handle mismatch between file data and actual data on SSD
  - Share validity information of data with the underlying device



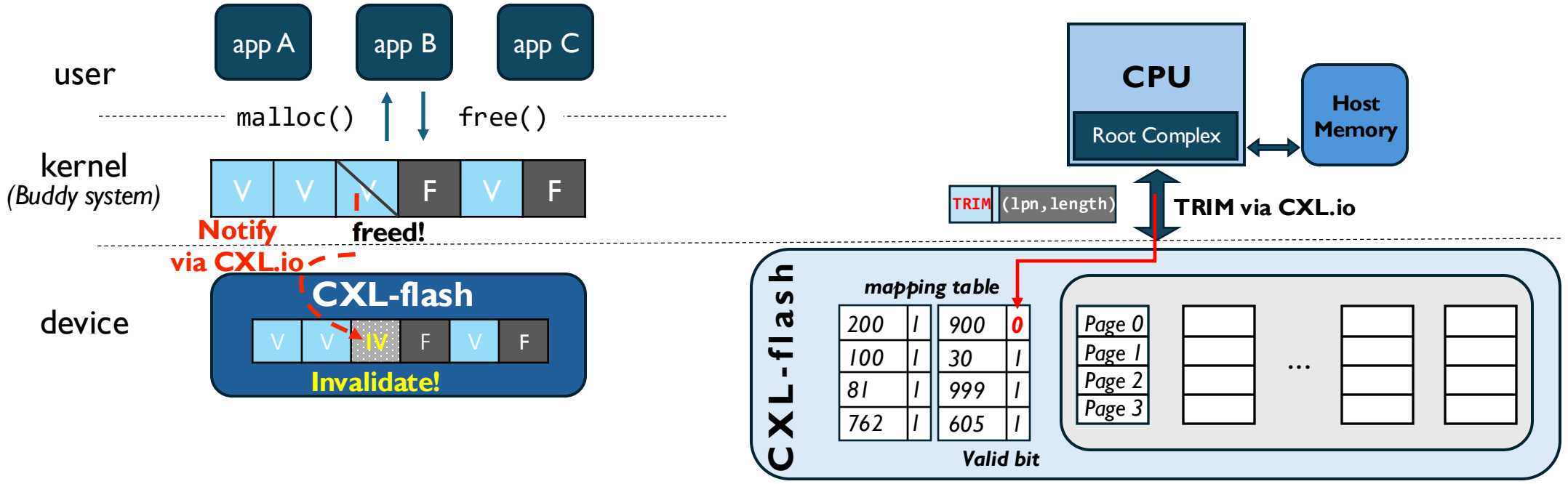
TRIM is Necessary for CXL-Flash too!

- **Invalid data will get accumulated within CXL-Flash**
- **TRIM enables the kernel to inform the underlying device of data validity**
  - No copying of freed data during GC



Designing TRIM for CXL-Flash

- **TRIM triggered by buddy system upon reclamation**
  - Privileged operation
  - Asynchronously via CXL.io protocol



Difficulties in Assessing TRIM Impact

- **Simulation/emulation are not suitable for CXL-Flash**
  - Hard to capture traces during meaningful time window
    - Excessive memory access
    - # of Memory access >>> # of Storage access
  - Emulators suffer from limited functionality or slow execution
    - OpenCIS, Flight Simulator, CXLMemSim
    - NUMA emulation
- **Hard to conduct measurement study**
  - CXL-compliant devices have limited availability
  - Profiling TRIM-related operations is challenging

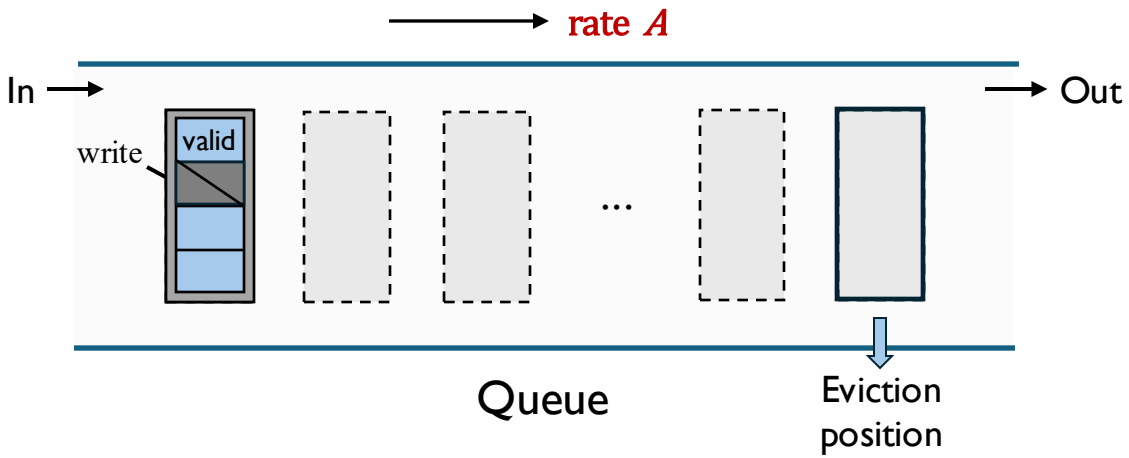
OpenCIS

MemVerge

CMM-H

Analytic Model for CXL-flash

- **Previous work: Analytic Modeling of SSD Write Performance (SYSTOR' 12)**
  - Modeling write amplification based on overprovisioning ratio with LRU cleaning policy
  - No considerations on TRIM command
- $A = Write\ Amplification\ Factor,\ \alpha = \frac{\#\ of\ physical\ block}{\#\ of\ logical\ block}$



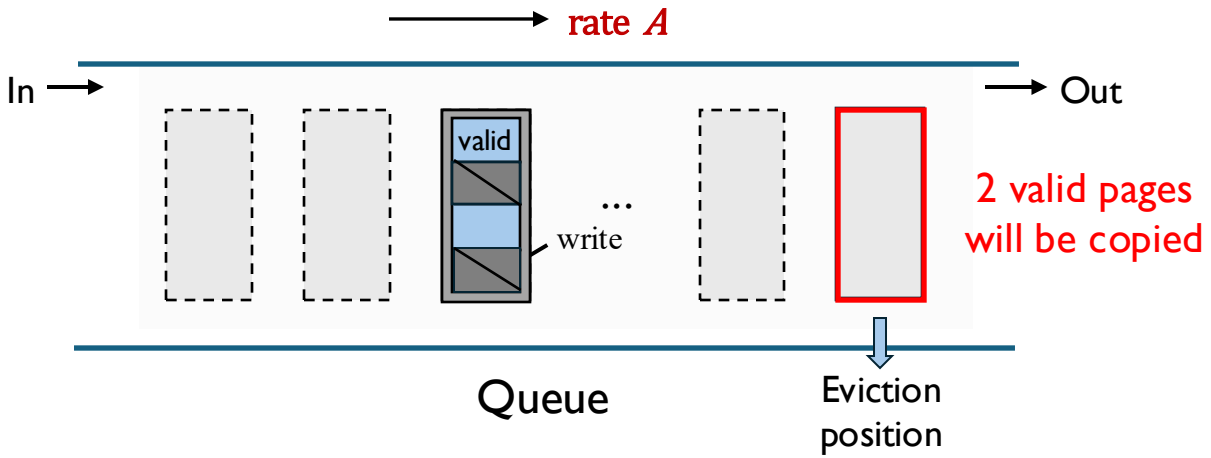
Original model:

$$A(WAF) = \frac{\alpha}{\alpha + W(-\alpha e^{-\alpha})}$$

(overprovisiong ratio α)

Analytic Model for CXL-flash

- **Previous work: Analytic Modeling of SSD Write Performance (SYSTOR' 12)**
  - Modeling write amplification based on overprovisioning ratio with LRU cleaning policy
  - No considerations on TRIM command
- $A = Write\ Amplification\ Factor,\ \alpha = \frac{\#\ of\ physical\ block}{\#\ of\ logical\ block}$



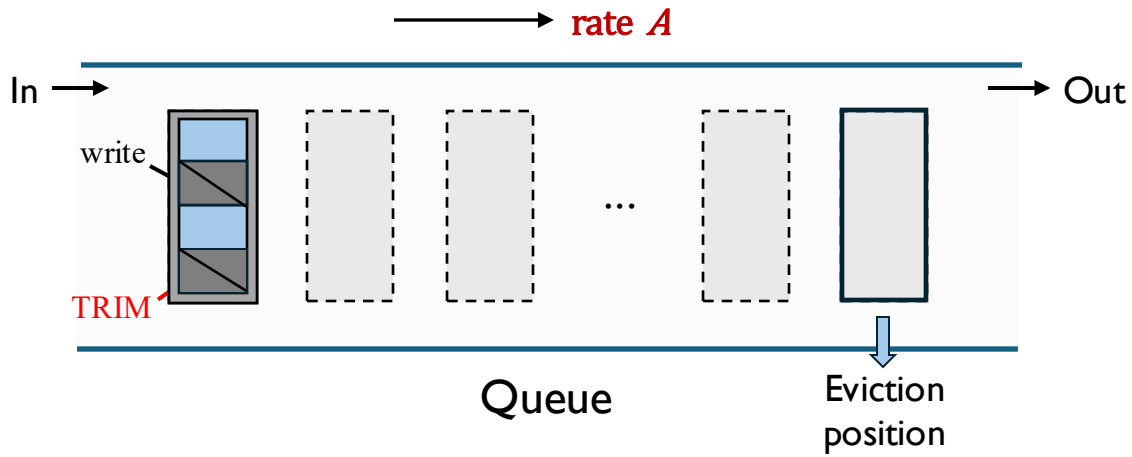
Original model:

$$A(WAF) = \frac{\alpha}{\alpha + W(-\alpha e^{-\alpha})}$$

(overprovisiong ratio α)

Our Analytic Model for CXL-flash

- **Extend analytical model to support TRIM**
  - New terminology D_r: ratio of Trim traffic to write traffic
  - Write amplification(A) prediction using # of valid pages in block



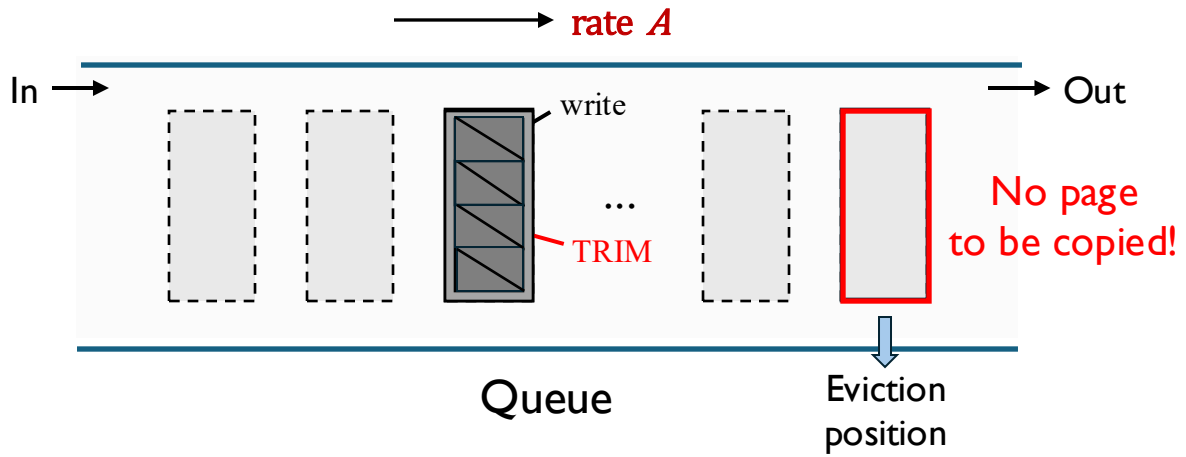
Our model:

$$A(WAF) = \frac{\alpha}{\alpha + \frac{W(-\alpha(1+D_r)e^{-\alpha(1+D_r)})}{1+D_r}}$$

(overprovisiong ratio α)

Our Analytic Model for CXL-flash

- **Extend analytical model to support TRIM**
  - New terminology D_r: ratio of Trim traffic to write traffic
  - Write amplification(A) prediction using # of valid pages in block



Higher chance that pages in block get invalidated than the original model

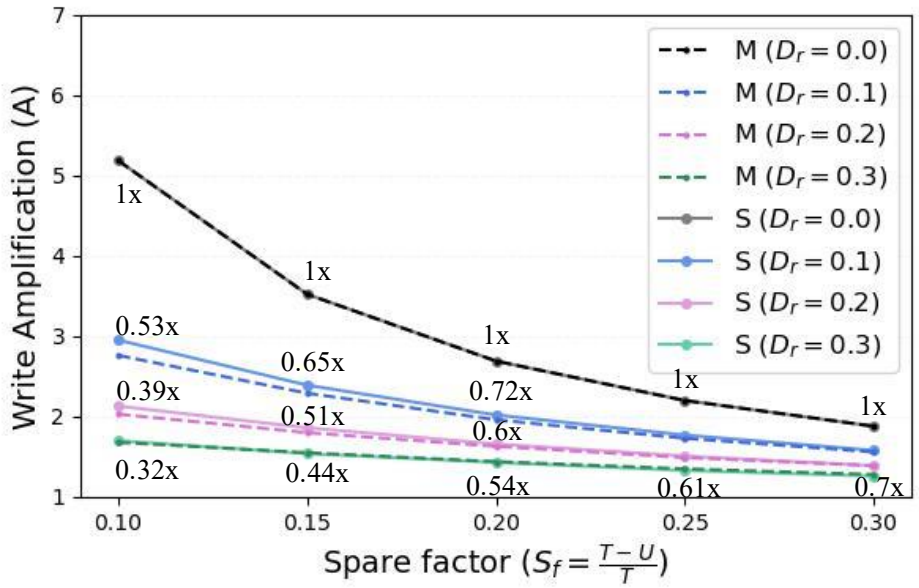
Our model:

$$A(WAF) = \frac{\alpha}{\alpha + \frac{W(-\alpha(1+D_r)e^{-\alpha(1+D_r)})}{1+D_r}}$$

(overprovisiong ratio α)

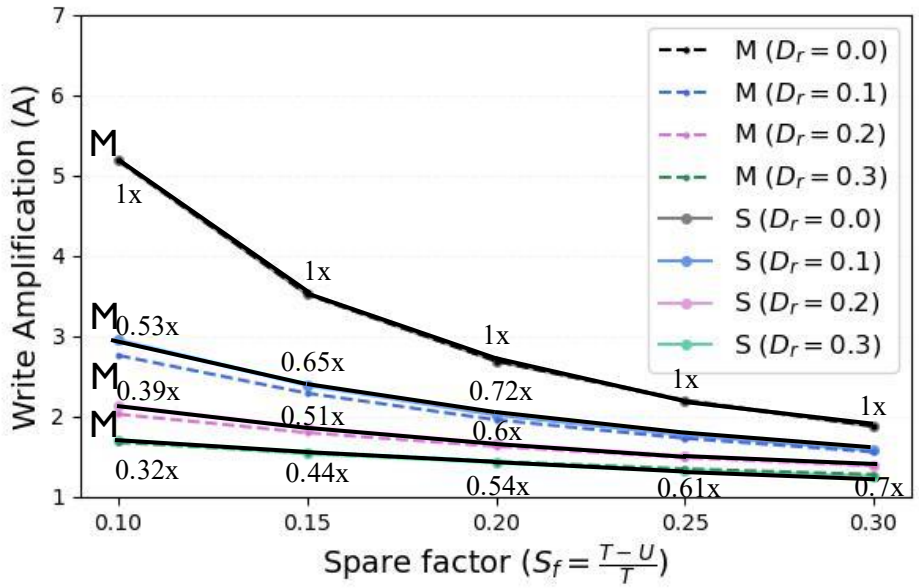
Validation of Analytical Model

- **Comparison between modeled(M) and simulated(S) WAF**
- **Methodology**
  - **Implement TRIM command in FTLSim** (SYSTOR' 12)
  - 10^6 logical blocks, 128 pages per block
  - Synthetic workloads
    - (1) Write entire logical space for warm-up
    - (2) Uniformly distributed writes (1/3 of total capacity)
      + TRIM issued every 10 writes (D_r)



Validation of Analytical Model

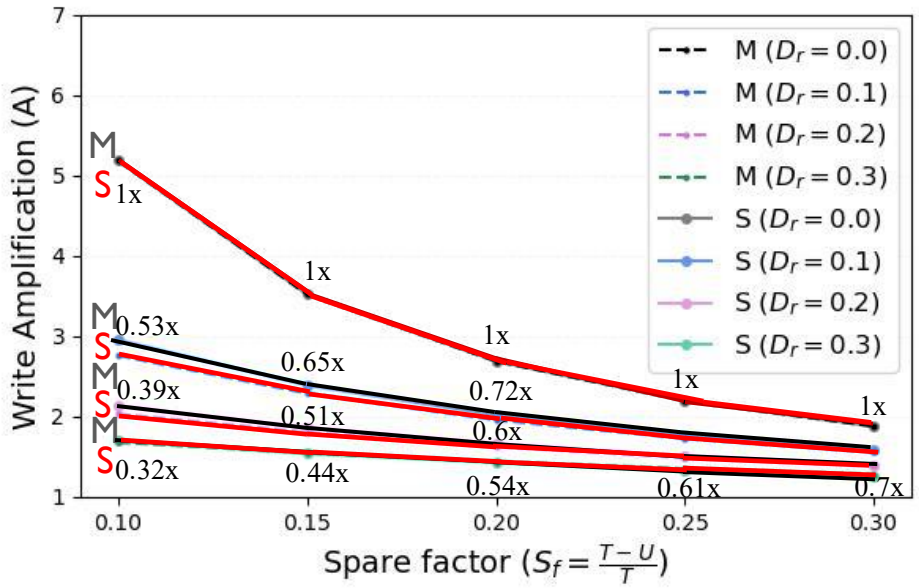
- **Comparison between modeled(M) and simulated(S) WAF**
- **Methodology**
  - **Implement TRIM command in FTLSim** (SYSTOR' 12)
  - 10^6 logical blocks, 128 pages per block
  - Synthetic workloads
    - (1) Write entire logical space for warm-up
    - (2) Uniformly distributed writes (1/3 of total capacity)
      + TRIM issued every 10 writes (D_r)



Validation of Analytical Model

- **Comparison between modeled(M) and simulated(S) WAF**
- **Methodology**
  - **Implement TRIM command in FTLSim** (SYSTOR' 12)
  - 10^6 logical blocks, 128 pages per block
  - Synthetic workloads
    - (1) Write entire logical space for warm-up
    - (2) Uniformly distributed writes (1/3 of total capacity)

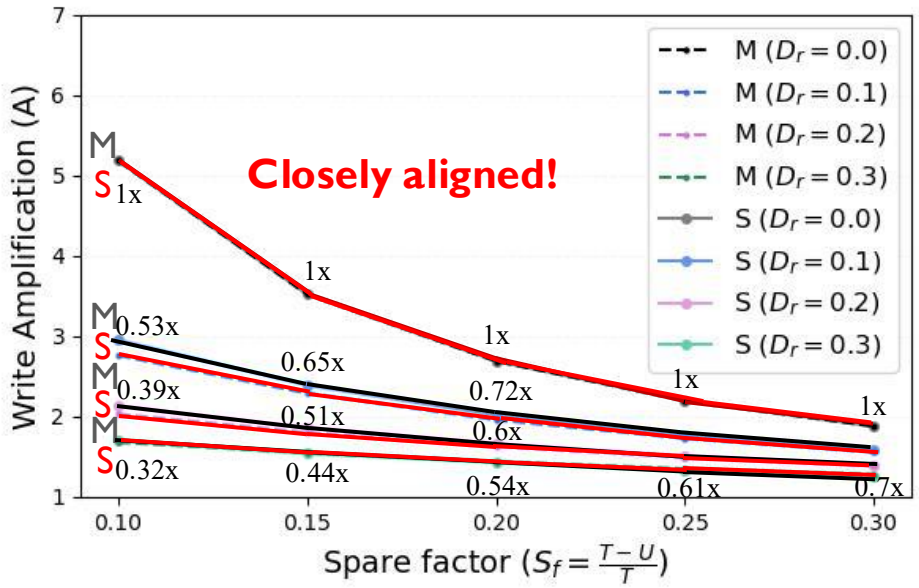
      + TRIM issued every 10 writes (D_r)



Validation of Analytical Model

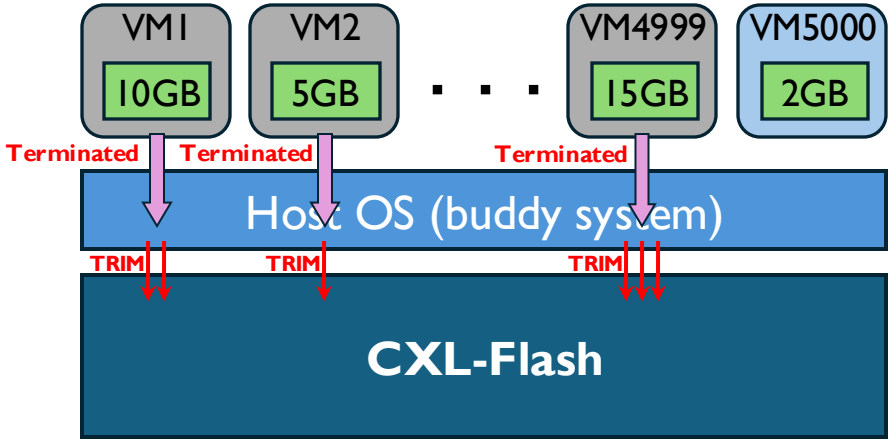
- **Comparison between modeled(M) and simulated(S) WAF**
- **Methodology**
  - **Implement TRIM command in FTLSim** (SYSTOR' 12)
  - 10^6 logical blocks, 128 pages per block
  - Synthetic workloads
    - (1) Write entire logical space for warm-up
    - (2) Uniformly distributed writes (1/3 of total capacity)

      + TRIM issued every 10 writes (D_r)



Estimation of TRIM's Effects in Real World

- **Data centers running numerous VMs the need for CXL-Flash stands out most.**
- **Microsoft Azure VM Traces 5000 VMs sampled**
  - Each VM's Lifetime, Memory (GB)
  - YCSB (A-F) / DLRM (Train/Infer)
- **TRIM is issued upon VM termination**

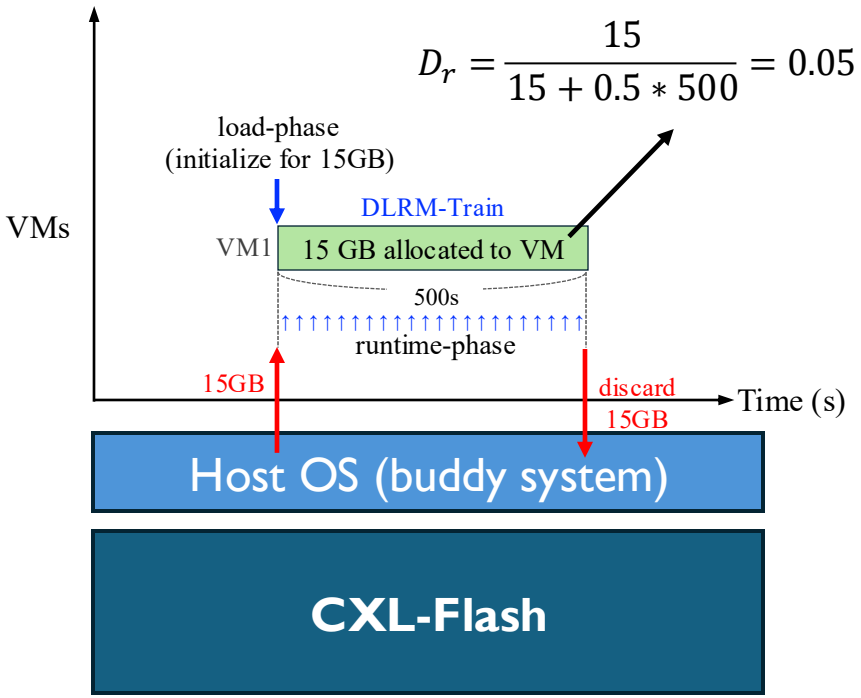


Estimation of TRIM's Effects in Real World (cont')

- **Our model:** $A = \frac{\alpha}{\alpha + \frac{W(-\alpha(1+D_r)e^{-\alpha(1+D_r)})}{1+D_r}}$
  - Defined as $D_r = \frac{discarded\ size}{load\ phase\ traffic + runtime\ phase\ traffic}$
  - $discarded\ size$ assumed to be equal to $load\ phase\ traffic$
    - VM's Memory(GB)
  - Profiling `LLC_misses.mem_write` with linux perf tool
    - No data of each VM's memory write traffic
  - Random mapping write traffic to VMs
    - YCSB + DLRM mixed at different ratio 6:4 / 3:7

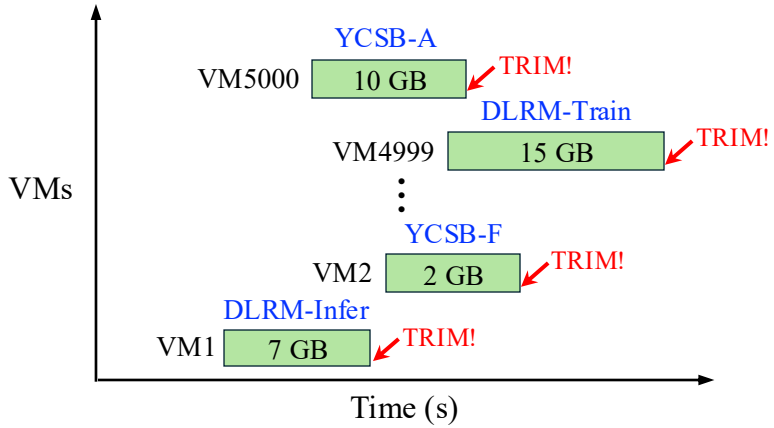
WK	A	B	C	D	E	F	TR	IF
R	262	296	226	524	79	190	1200	380
W	147	154	94	388	40	101	500	267

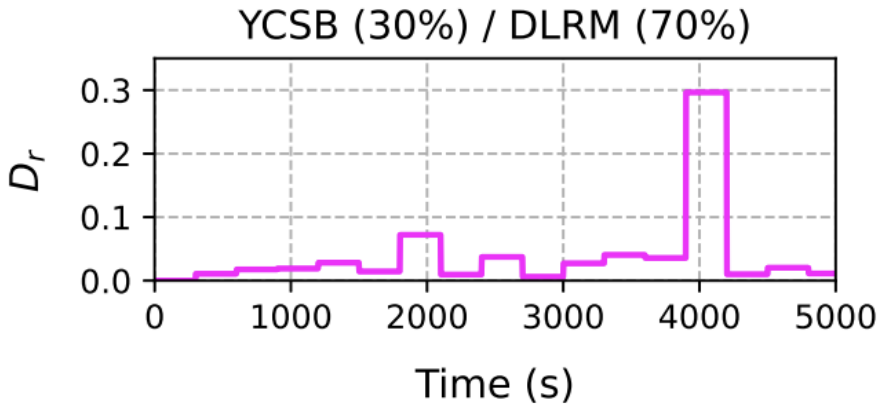
Memory access traffic (MB/s)

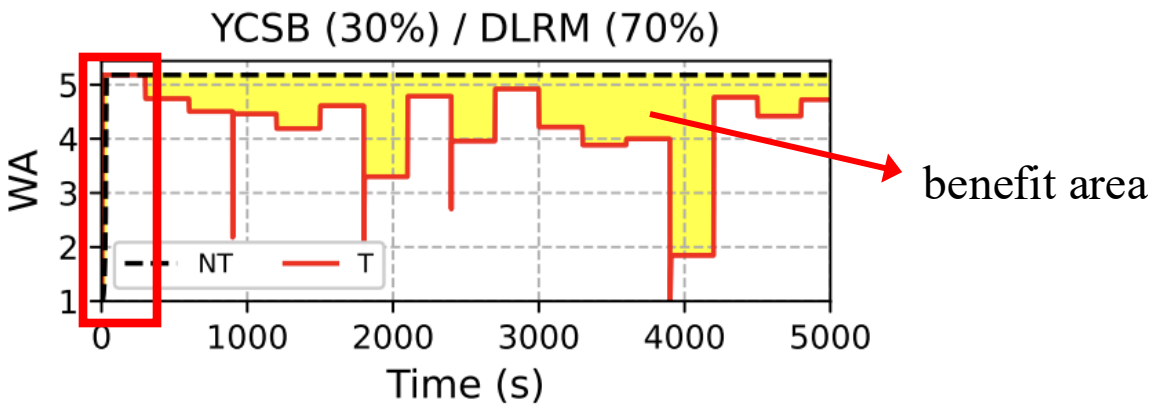


Estimation of TRIM's Effects in Real World (cont')

- **TRIM effect on CXL-Flash running VMs over time**
  - NT(No Trim): reaching a state fully occupied by fake valid data
  - T(Trim): reduction in WAF

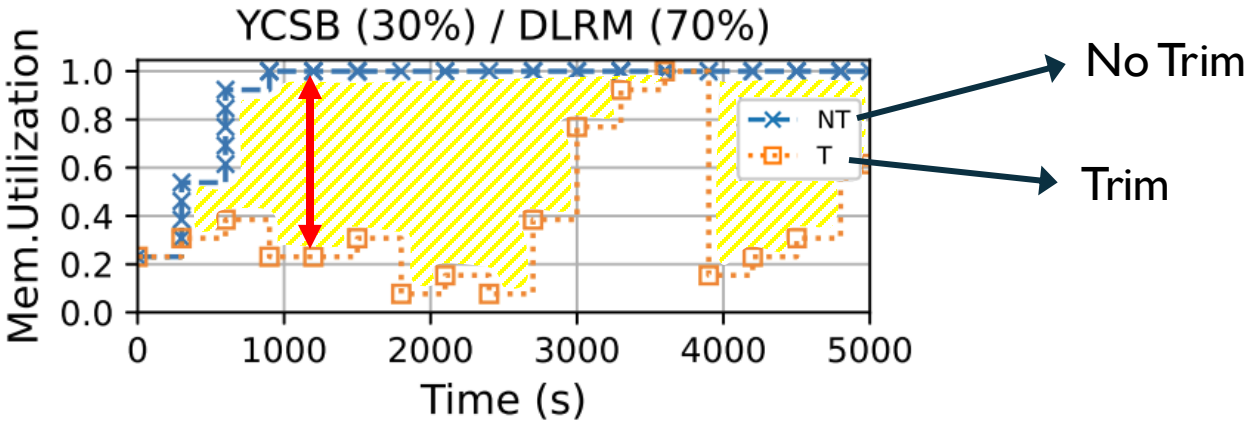






Estimation of TRIM's Effects in Real World (cont')

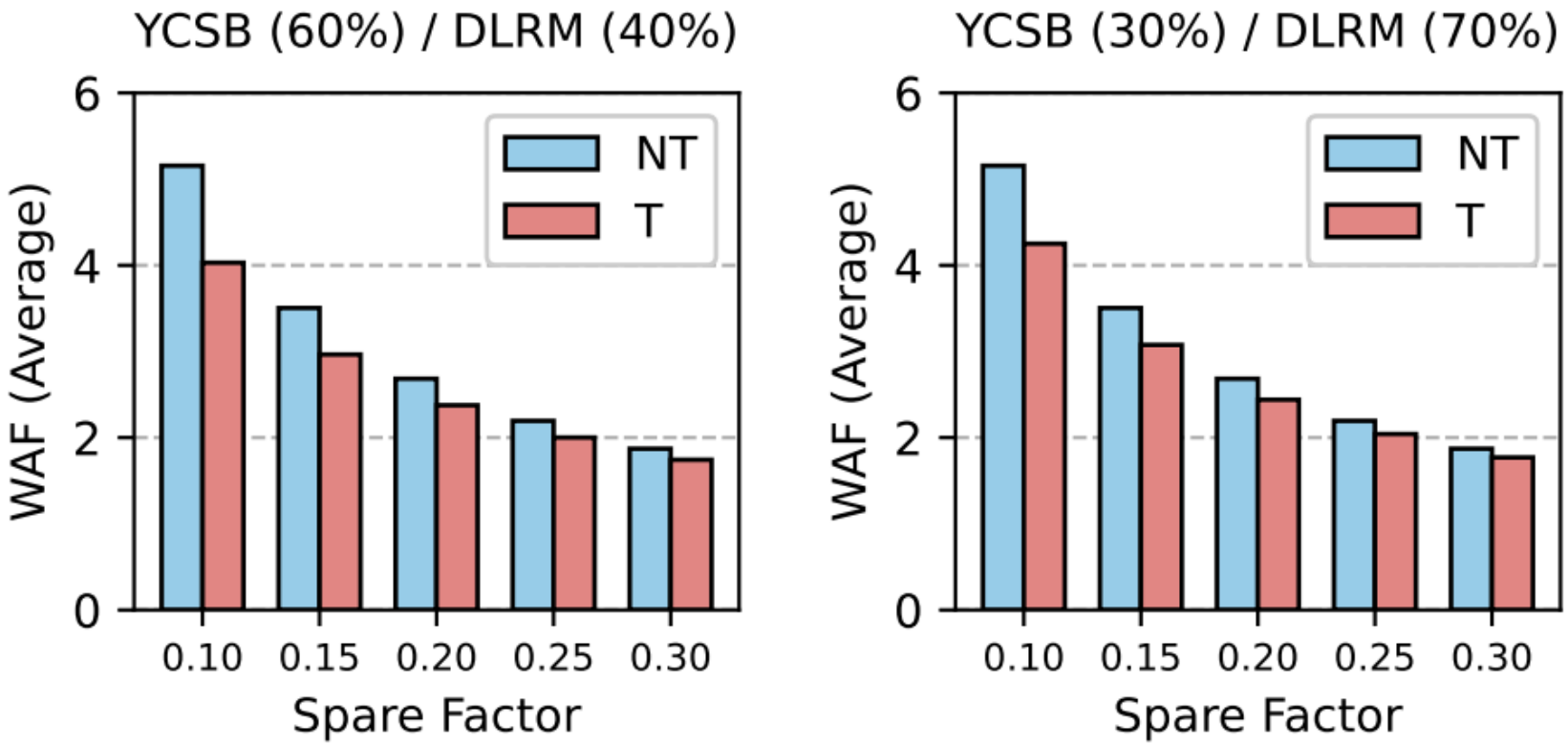
- **Memory Utilization : TRIM (T) vs. No TRIM (NT)**
  - Size of CXL-Flash (the max-memory simultaneously used): 22.75 GB
  - Spare factor of CXL-Flash is set to 0.1
- **CXL-Flash perceives available spaces are fully utilized beyond a certain point**



Memory Utilization over time

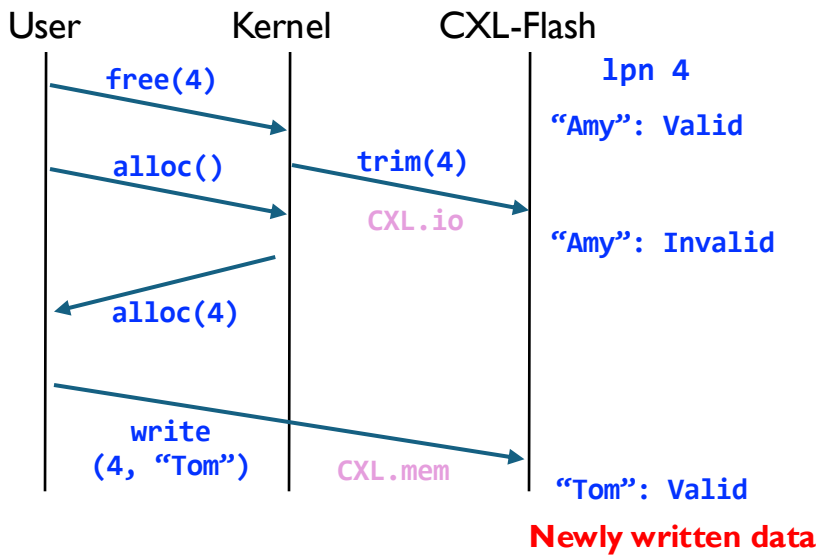
Estimation of TRIM's Effects in Real World (cont')

- **Average WAF in CXL-Flash running VMs**
  - WAF reduced by 11.56% on average, reaching a maximum of 19.69%

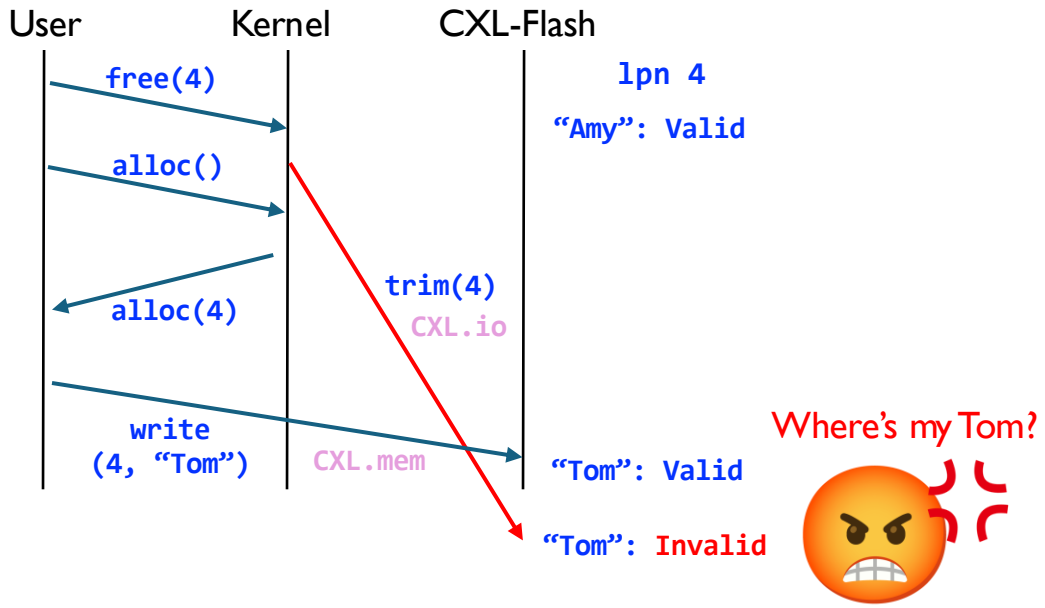


Discussion

- **Proper coordination is needed to prevent data loss**
  - Kernel must defer reallocating the TRIM-pending area until ACK is received
  - Device must notify the kernel upon TRIM completion



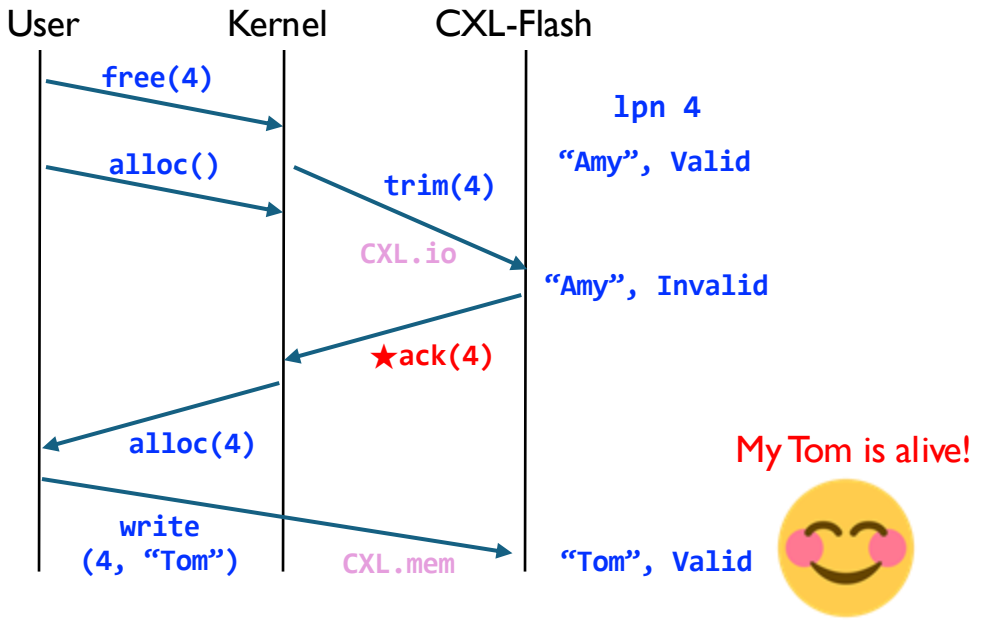
In-order execution



Out-of-order execution

Discussion

- **Proper coordination is needed to prevent data loss**
  - Kernel must defer reallocating the TRIM-pending area until ACK is received
  - Device must notify the kernel upon TRIM completion



Coordinated out-of-order execution

Conclusion

- **Summary**
  - TRIM via CXL.io to invalidate fake valid data in CXL-Flash
  - Near-exact extended analytical model for TRIM effect evaluation
  - Demonstration of the necessity of TRIM on real-world scenarios using our model
  - TRIM-like mechanism is necessary for CXL-Flash!
    - Synchronization issues should be considered

- **Future work**
  - Identification of optimal host-side TRIM initiator and analysis of TRIM overhead
  - Analysis of long-term impact of TRIM on memory performance and endurance in CXL-Flash
  - Estimation of TRIM effectiveness via Workload- and GC policy-aware analytical model in real-world scenario

Our paper!

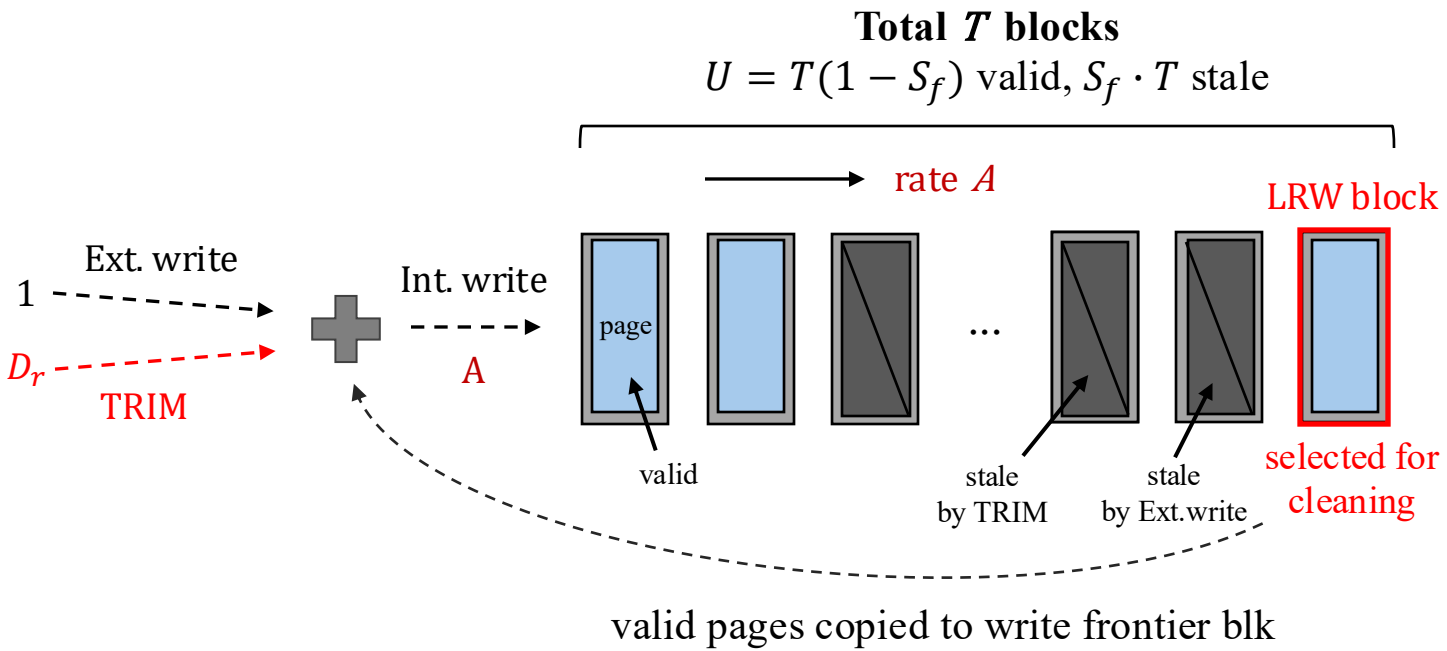
Hayan Lee (hayany19@gmail.com)

Back-ups

- **Design for TRIM in user-level: system call for user apps invoking TRIM**
  - Memory allocator (e.g. Jemalloc, TCMalloc, Mimalloc)
  - JVM

Back-ups: Extended Analytic Model

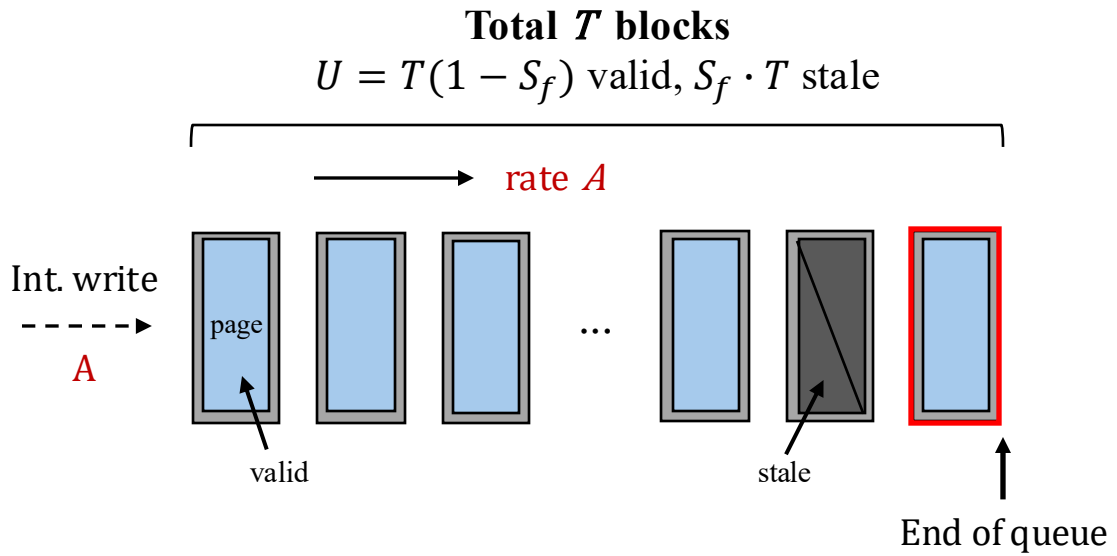
- **Extend analytical model to support TRIM**
  - D_r: ratio of discard (trimmed) traffic to write traffic



$$N_p = 1$$

$$A = \frac{Int.write}{Ext.write}$$

Back-ups: Extended Analytic Model

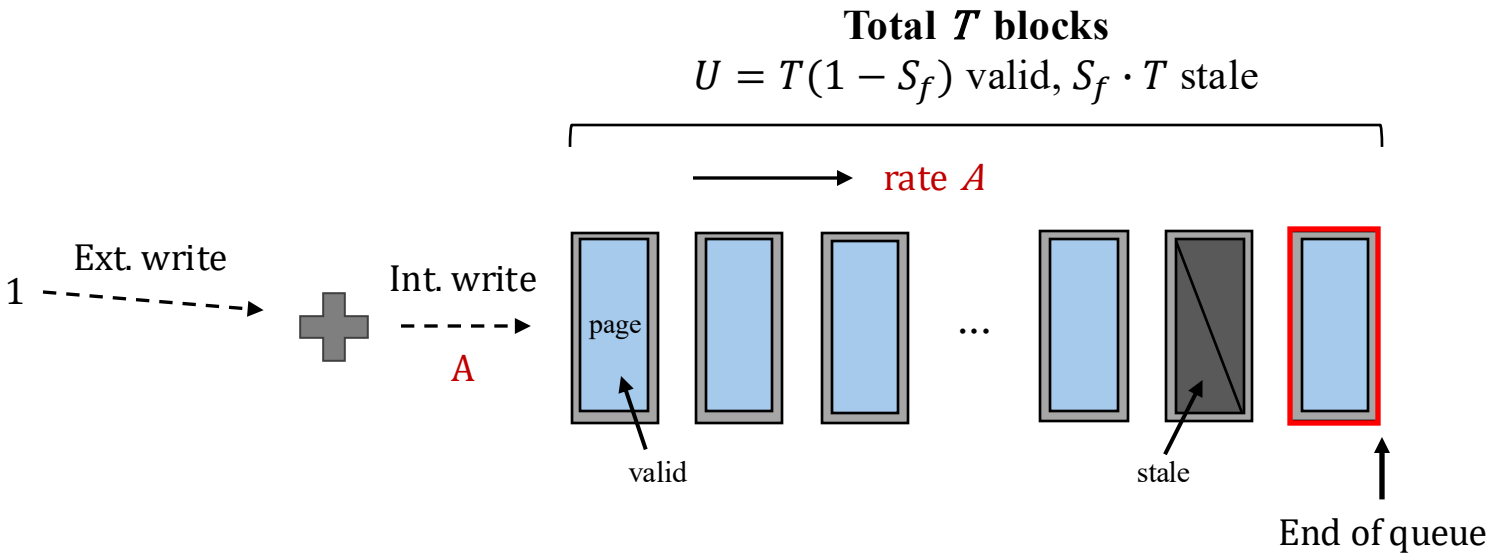


$$N_p = 1$$

$$A = \frac{Int.write}{Ext.write}$$

$$\boldsymbol{S}\ (Survival\ rate) = \left(\mathbf{1} - \left(\frac{\mathbf{1}}{\boldsymbol{UN_p}} + \frac{\boldsymbol{D_r}}{\boldsymbol{UN_p}}\right)\right)^{\frac{\boldsymbol{TN_p}}{\boldsymbol{A}}} \text{ at end of queue}$$

Back-ups: Extended Analytic Model

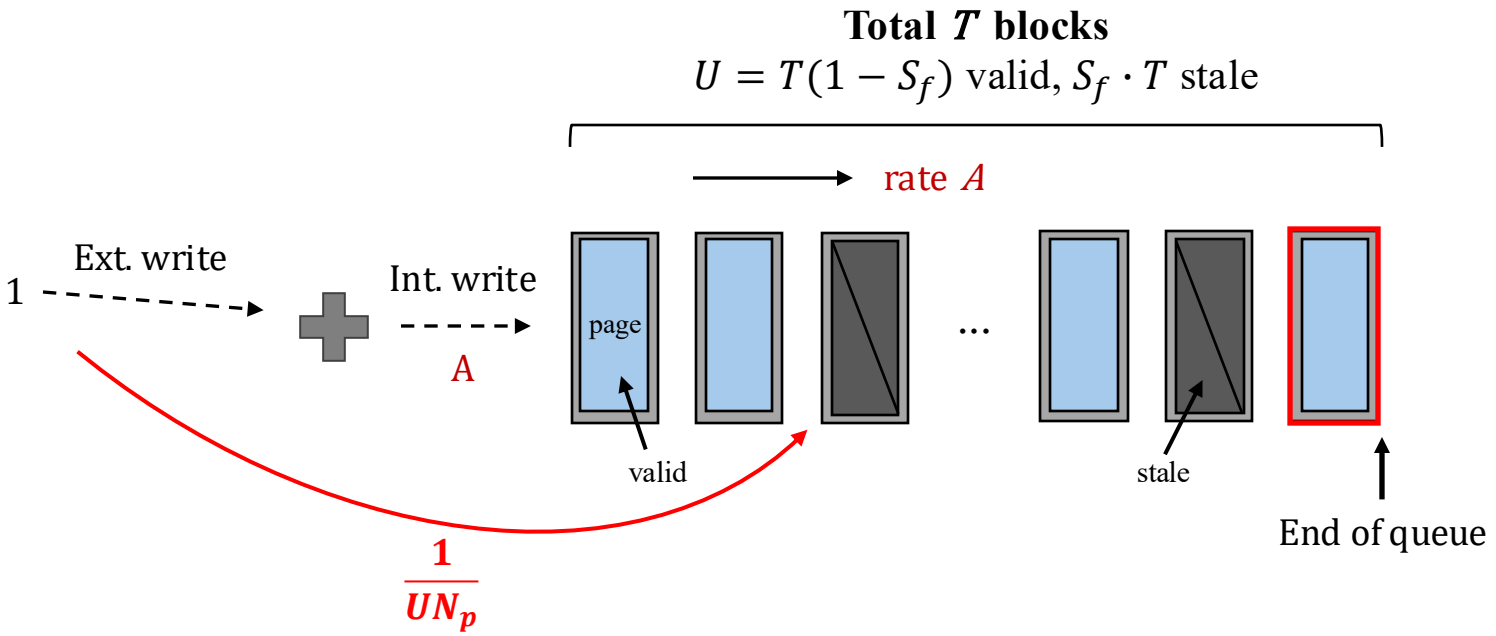


$$N_p = 1$$

$$A = \frac{Int.write}{Ext.write}$$

$$\boldsymbol{S}\ (Survival\ rate) = \left(\mathbf{1} - \left(\frac{\mathbf{1}}{\boldsymbol{UN_p}} + \frac{\boldsymbol{D_r}}{\boldsymbol{UN_p}}\right)\right)^{\frac{\boldsymbol{TN_p}}{\boldsymbol{A}}} \text{ at end of queue}$$

Back-ups: Extended Analytic Model

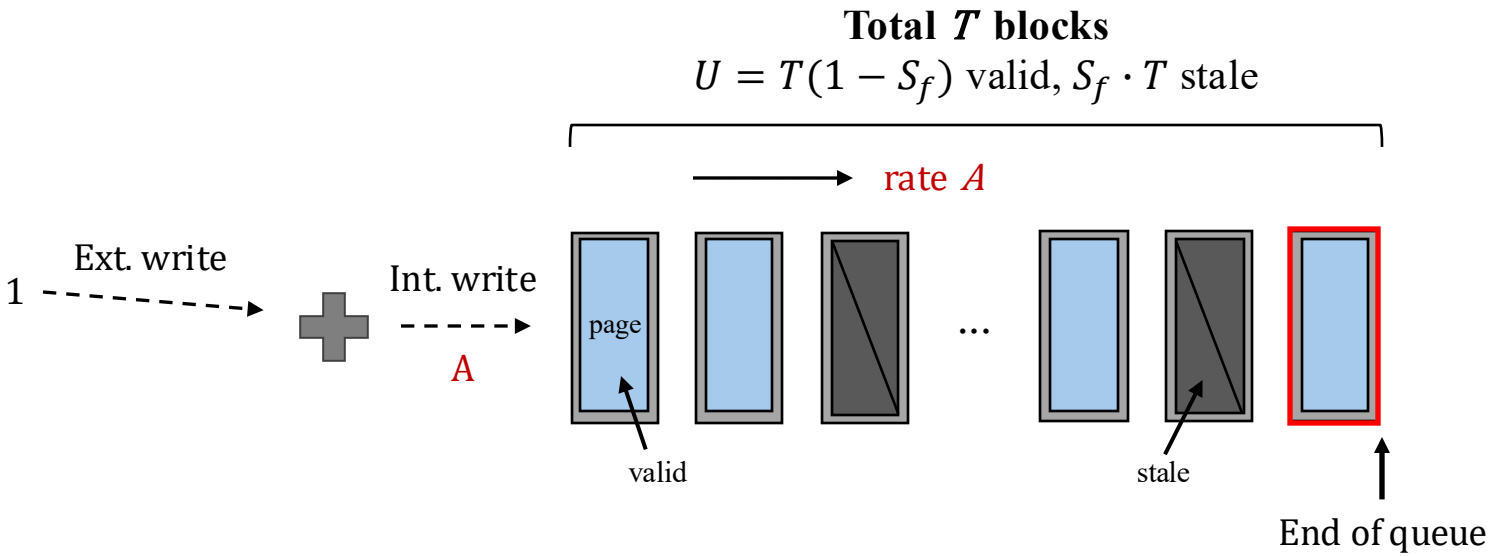


$$N_p = 1$$

$$A = \frac{Int.write}{Ext.write}$$

$$\boldsymbol{S}\ (Survival\ rate) = \left(\mathbf{1} - \left(\frac{\mathbf{1}}{\boldsymbol{UN_p}} + \frac{\boldsymbol{D_r}}{\boldsymbol{UN_p}}\right)\right)^{\frac{\boldsymbol{TN_p}}{\boldsymbol{A}}} \text{ at end of queue}$$

Back-ups: Extended Analytic Model

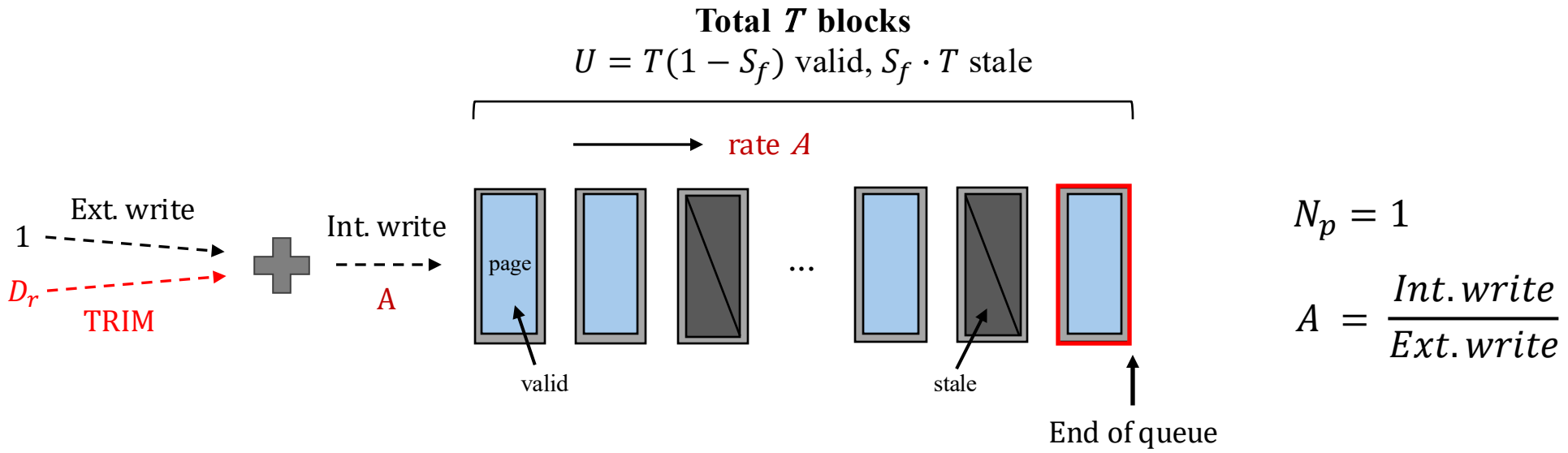


$$N_p = 1$$

$$A = \frac{Int.write}{Ext.write}$$

$$\boldsymbol{S}\ (Survival\ rate) = \left(\mathbf{1} - \left(\frac{\mathbf{1}}{\boldsymbol{UN_p}} + \frac{\boldsymbol{D_r}}{\boldsymbol{UN_p}}\right)\right)^{\frac{\boldsymbol{TN_p}}{\boldsymbol{A}}} \text{ at end of queue}$$

Back-ups: Extended Analytic Model

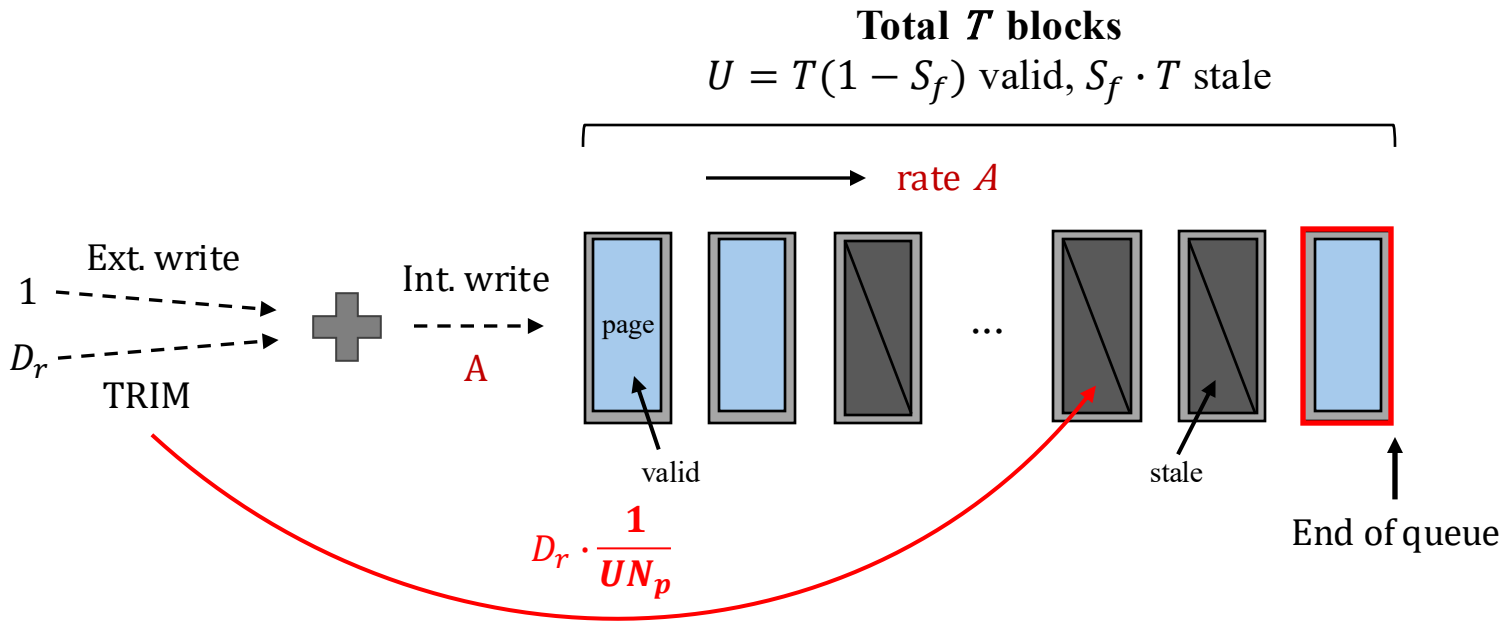


$$N_p = 1$$

$$A = \frac{Int.write}{Ext.write}$$

$$\boldsymbol{S}\ (Survival\ rate) = \left(\mathbf{1} - \left(\frac{\mathbf{1}}{\boldsymbol{UN_p}} + \frac{\boldsymbol{D_r}}{\boldsymbol{UN_p}}\right)\right)^{\frac{\boldsymbol{TN_p}}{\boldsymbol{A}}} \text{ at end of queue}$$

Back-ups: Extended Analytic Model

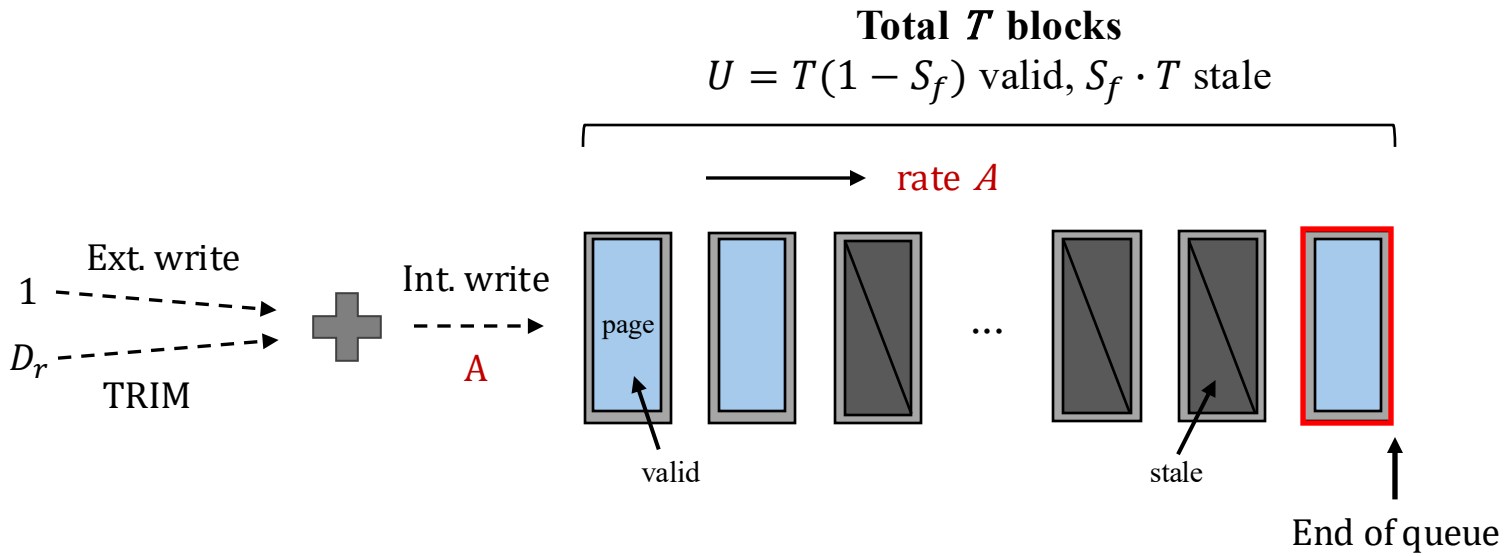


$$N_p = 1$$

$$A = \frac{Int.write}{Ext.write}$$

$$\boldsymbol{S}\ (Survival\ rate) = \left(1 - \left(\frac{1}{UN_p} + \frac{D_r}{UN_p}\right)\right)^{\frac{TN_p}{A}} \text{ at end of queue}$$

Back-ups: Extended Analytic Model



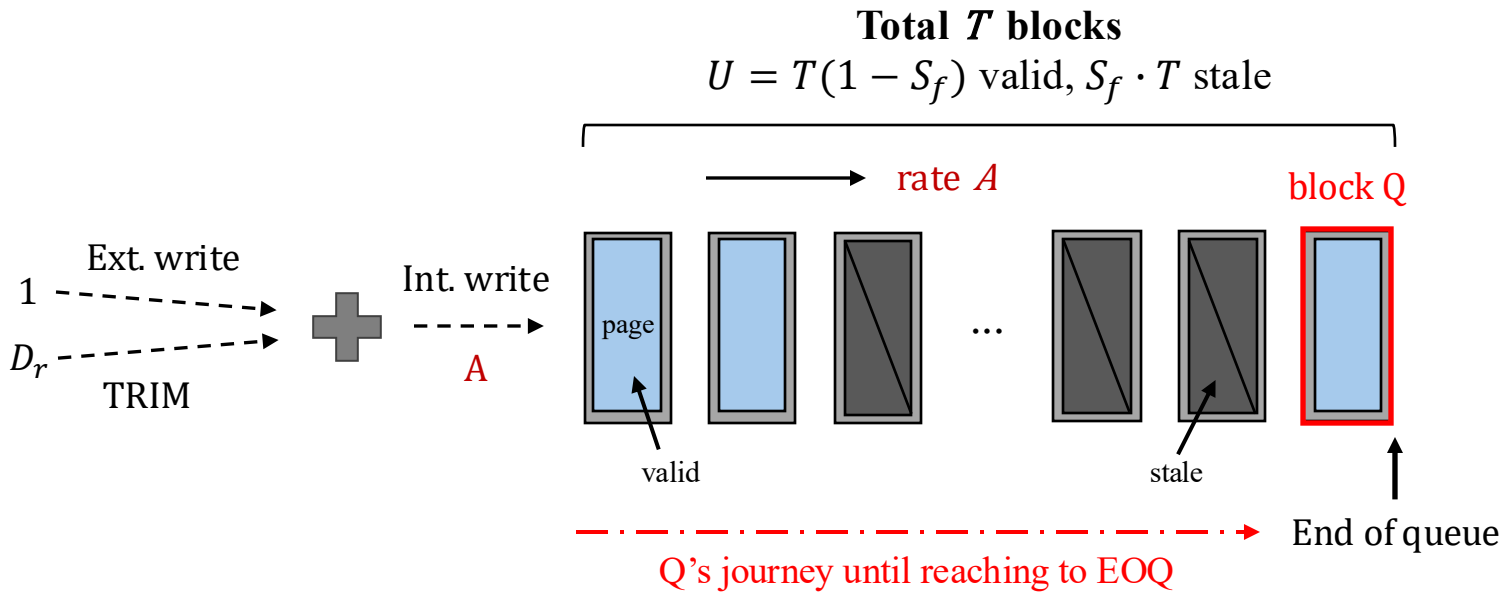
$$N_p = 1$$

$$A = \frac{Int.write}{Ext.write}$$

$$\boldsymbol{S}\ (Survival\ rate) = \underbrace{\left(1 - \left(\frac{1}{UN_p} + \frac{D_r}{UN_p}\right)\right)}^{\frac{TN_p}{A}} \text{ at end of queue}$$

possibility of a page remaining valid

Back-ups: Extended Analytic Model

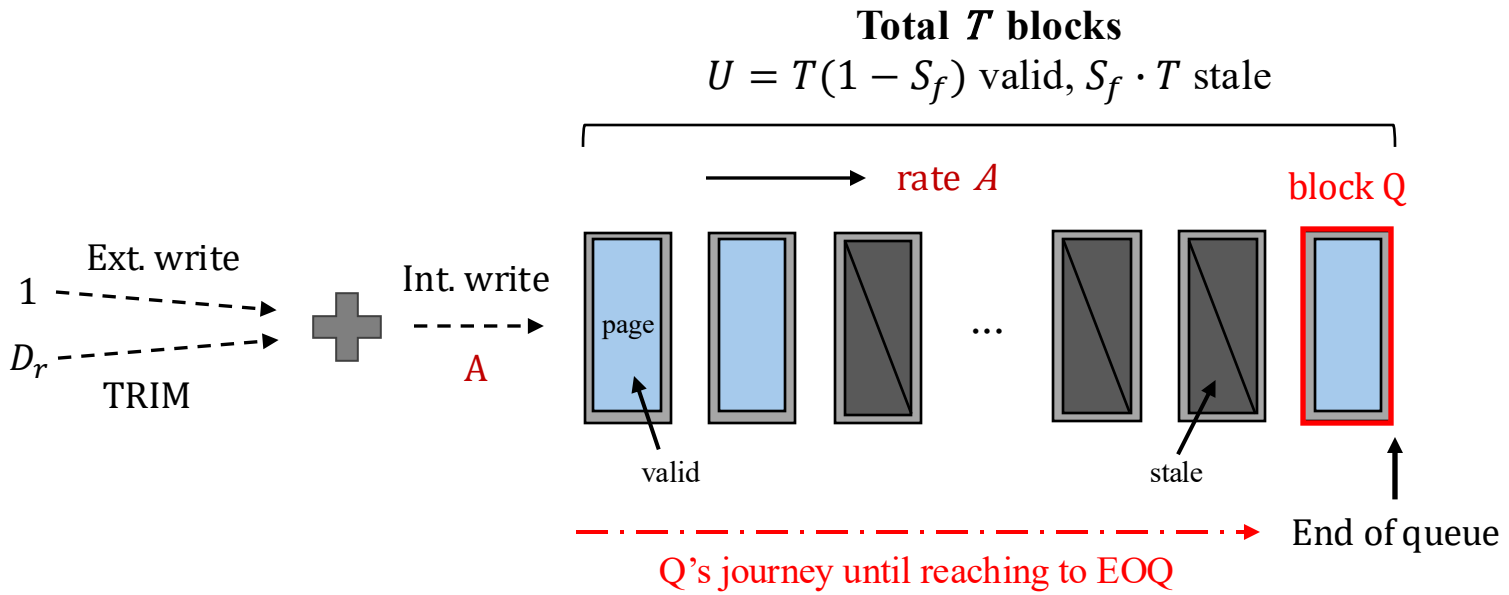


$$N_p = 1$$

$$A = \frac{Int.write}{Ext.write}$$

$$\boldsymbol{S}\ (Survival\ rate) = \left(\mathbf{1} - \left(\frac{\mathbf{1}}{\boldsymbol{UN_p}} + \frac{\boldsymbol{D_r}}{\boldsymbol{UN_p}}\right)\right)^{\frac{\boldsymbol{TN_p}}{\boldsymbol{A}}} \text{ at end of queue}$$

Back-ups: Extended Analytic Model



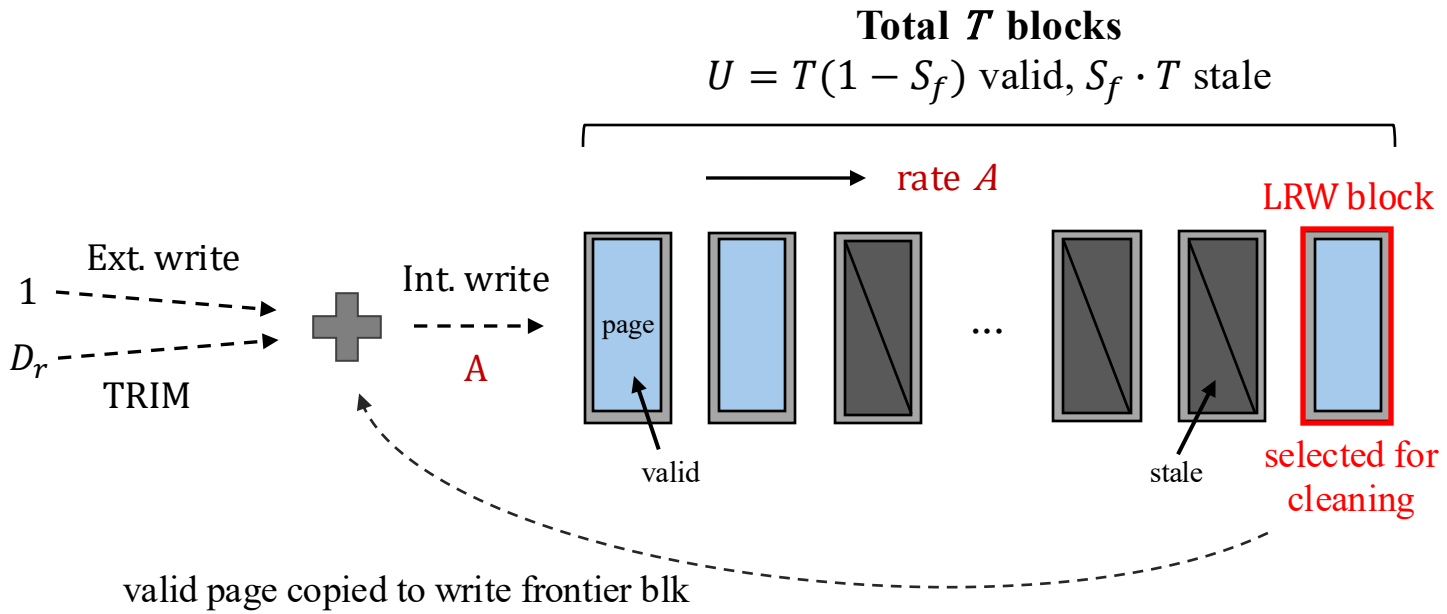
$$N_p = 1$$

$$A = \frac{Int.write}{Ext.write}$$

$$\boldsymbol{S}\ (Survival\ rate) = \left(\mathbf{1} - \left(\frac{\mathbf{1}}{\boldsymbol{UN_p}} + \frac{\boldsymbol{D_r}}{\boldsymbol{UN_p}}\right)\right)^{\frac{\boldsymbol{TN_p}}{\boldsymbol{A}}} \text{ at end of queue}$$

# of expected writes during Q's journey

Back-ups: Extended Analytic Model



$$N_p = 1$$

$$A = \frac{Int.write}{Ext.write}$$

$$\boldsymbol{S}\ (Survival\ rate) = \left(\mathbf{1} - \left(\frac{\mathbf{1}}{\boldsymbol{UN_p}} + \frac{\boldsymbol{D_r}}{\boldsymbol{UN_p}}\right)\right)^{\frac{\boldsymbol{TN_p}}{\boldsymbol{A}}} \text{ at end of queue}$$

$$\boldsymbol{A}\ (Write\ Amplification\ Factor) = \frac{\mathbf{1}}{\mathbf{1}-\boldsymbol{S}}$$