AiF: Accelerating On-Device LLM Inference Using In-Flash Processing

Jaeyong Lee[1], Hyeunjoo Kim[1], Sanghun Oh[1],
Myoungjun Chun[2], Myungsuk Kim[3], and Jihong Kim[1]

[1]Seoul National University
[2]Soongsil University
[3]Kyungpook National University

차세대데이터센터 기술동향 워크샵 (KCC 2025)

(Presented at ISCA 2025)

Talk Outline

001 On-Device LLM Inference

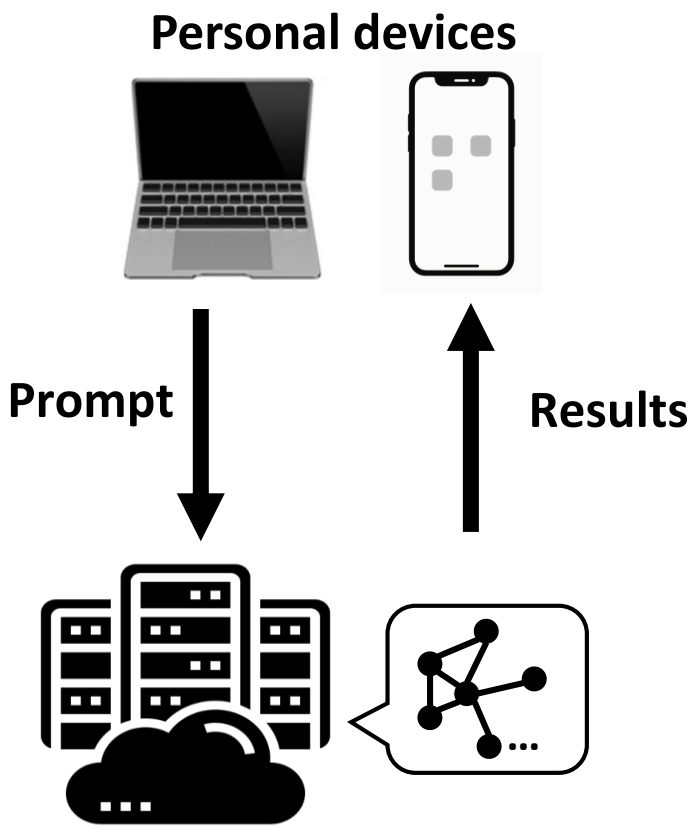
002 SSD-Centric Approaches

003 AiF: Accelerator-in-Flash

004 Evaluation results and Summary

On-Device LLM Inference

- Deploying LLMs directly on edge devices is gaining significant attention

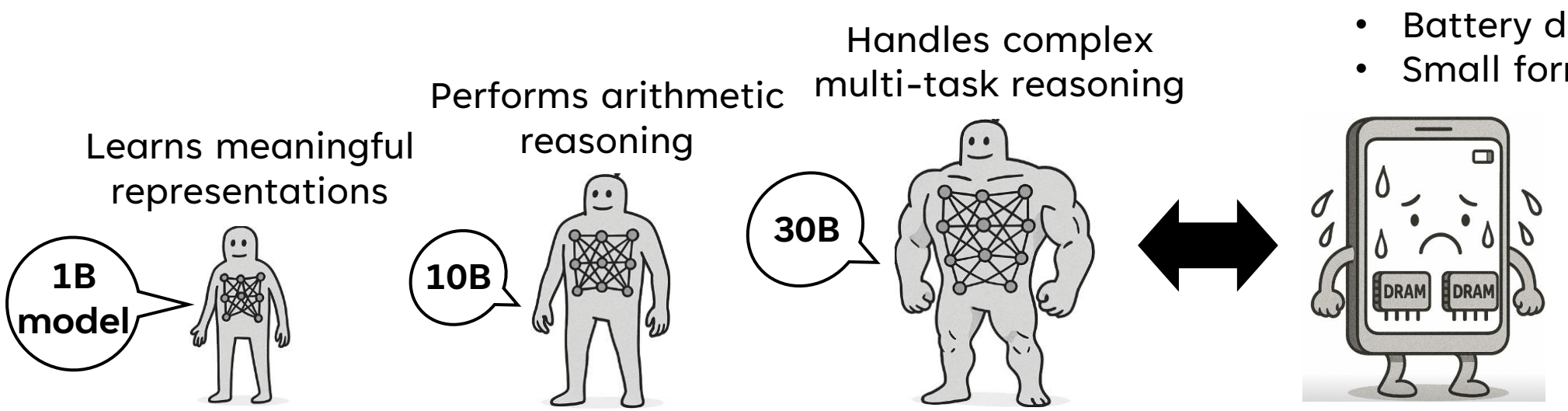


<Cloud-based LLM>

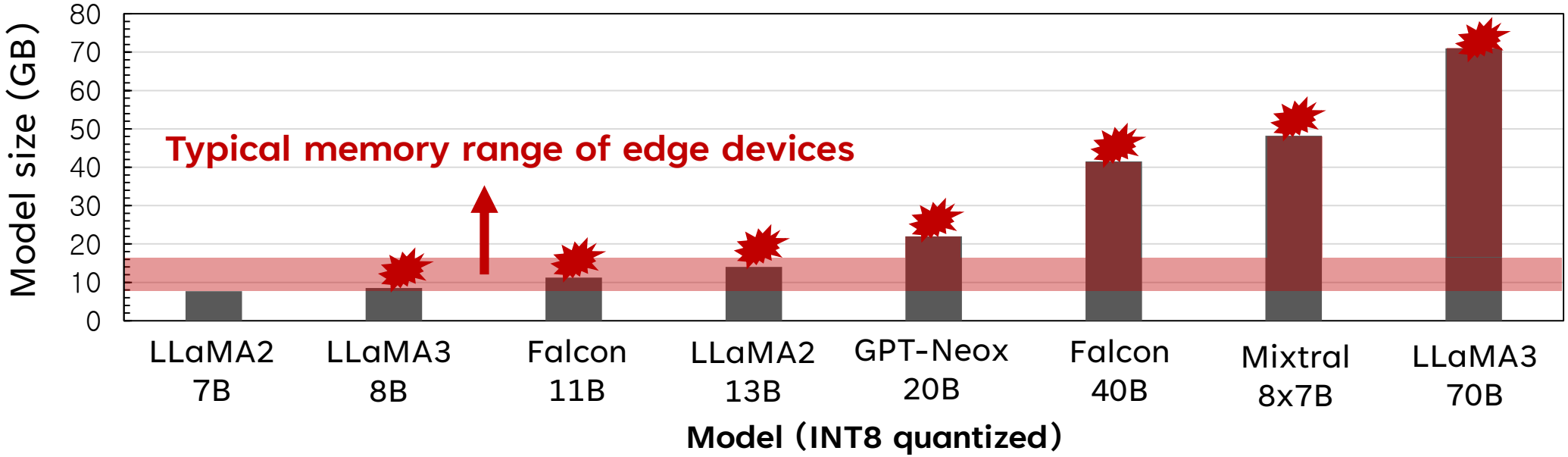
- **Privacy concerns**
- **Network requirement**
- **Personalization**
- **Latency**
- **…**

<On-Device LLM>

Problem: Memory Constraints in On-Device LLMs



- Battery drain
- Small form factor



Talk Outline

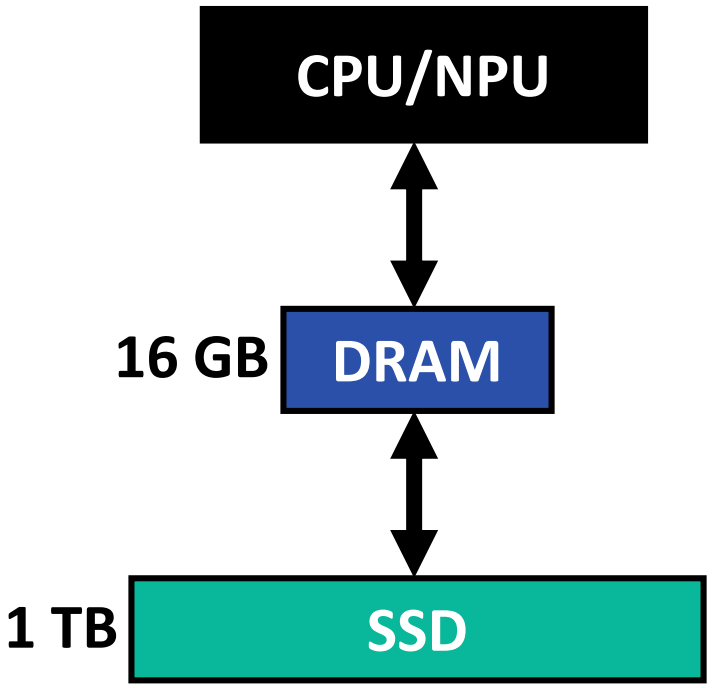
001 On-Device LLM Inference

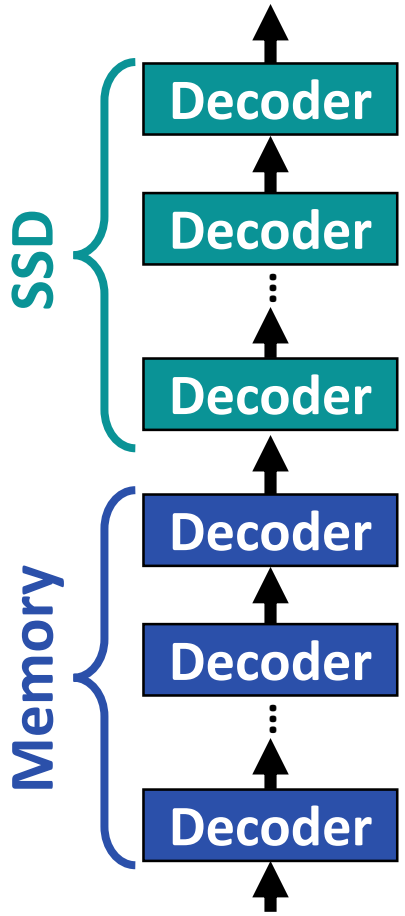
002 SSD-Centric Approaches

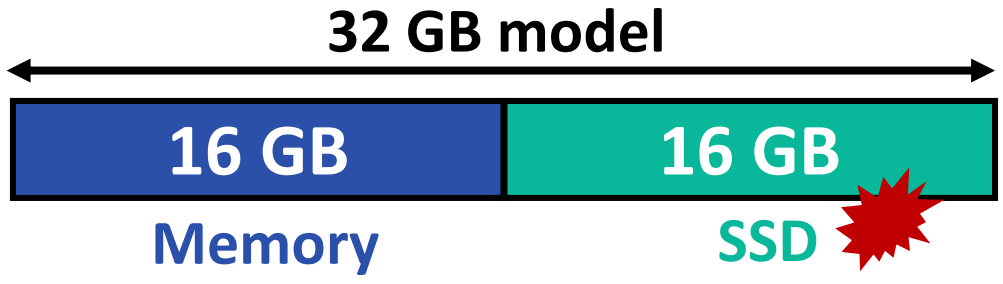
003 AiF: Accelerator-in-Flash

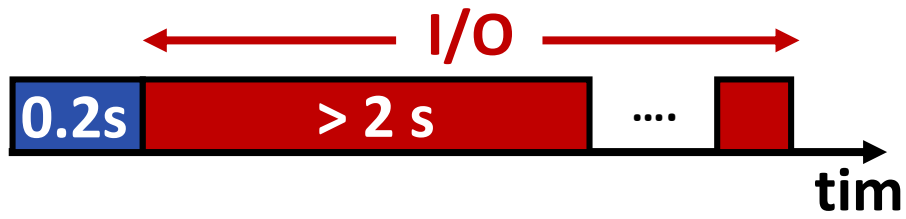
004 Evaluation results and Summary

Offloading Model Parameters to SSD







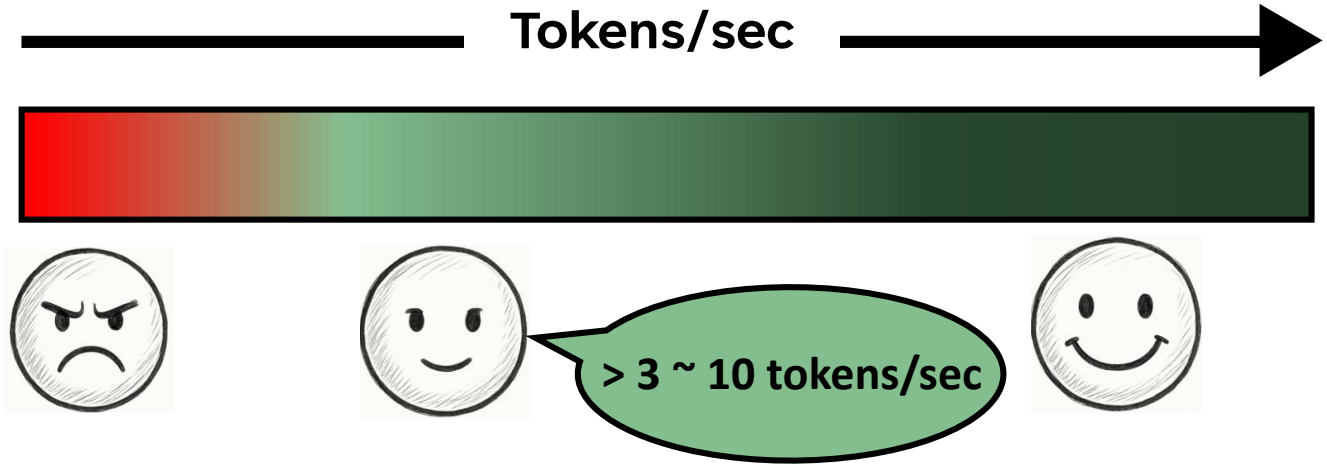


- **Per-token generation latency: > 2 s**
- **Generation throughput: < 0.5 token/sec**

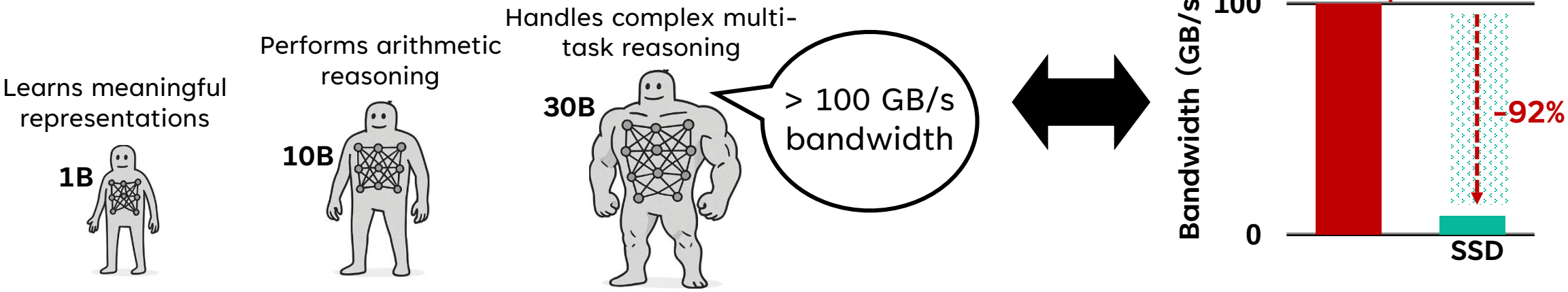
The **limited SSD bandwidth** can significantly degrade the performance

Performance Requirements

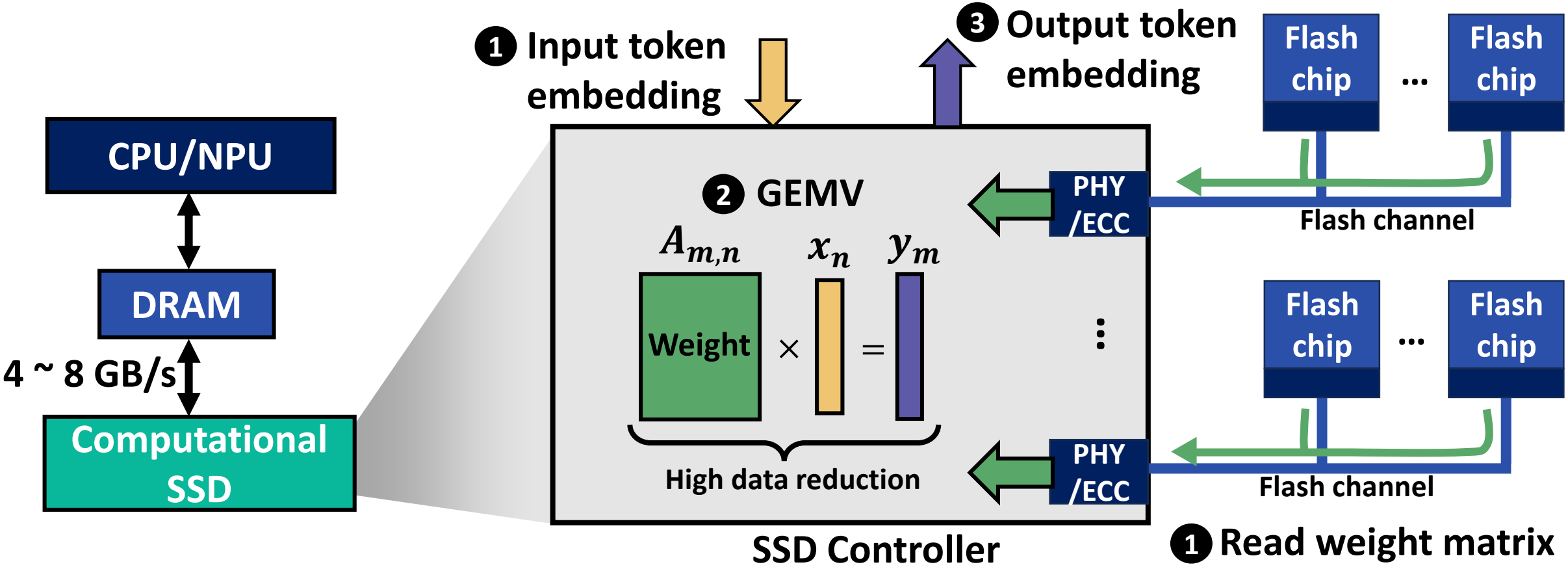
- Minimum token generation rates to ensure a seamless user experience

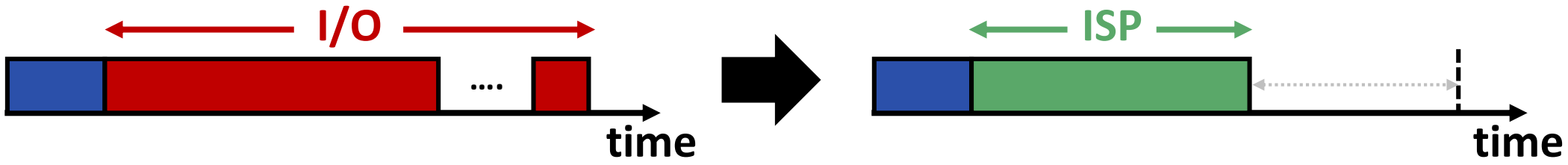


- To support complex models (e.g., > 30B),



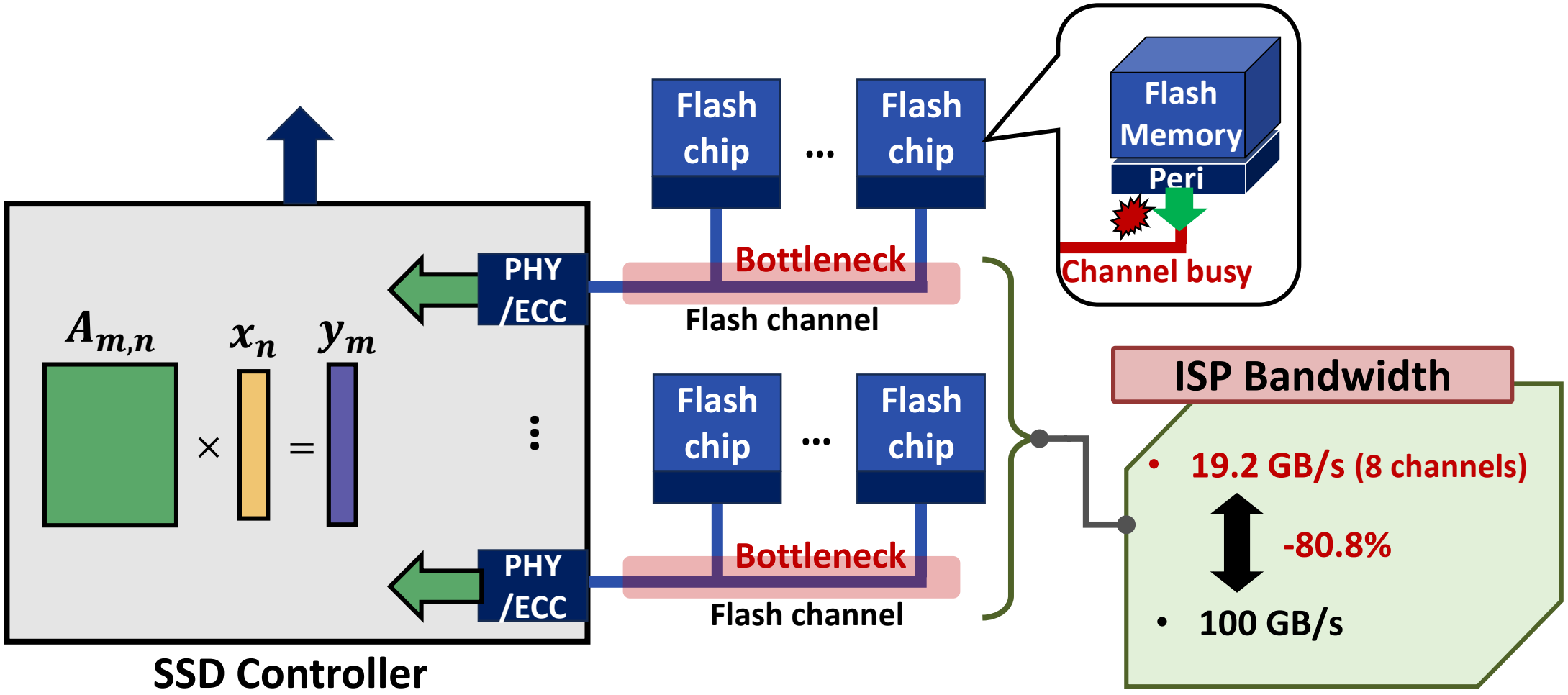
In-Storage Processing (ISP)



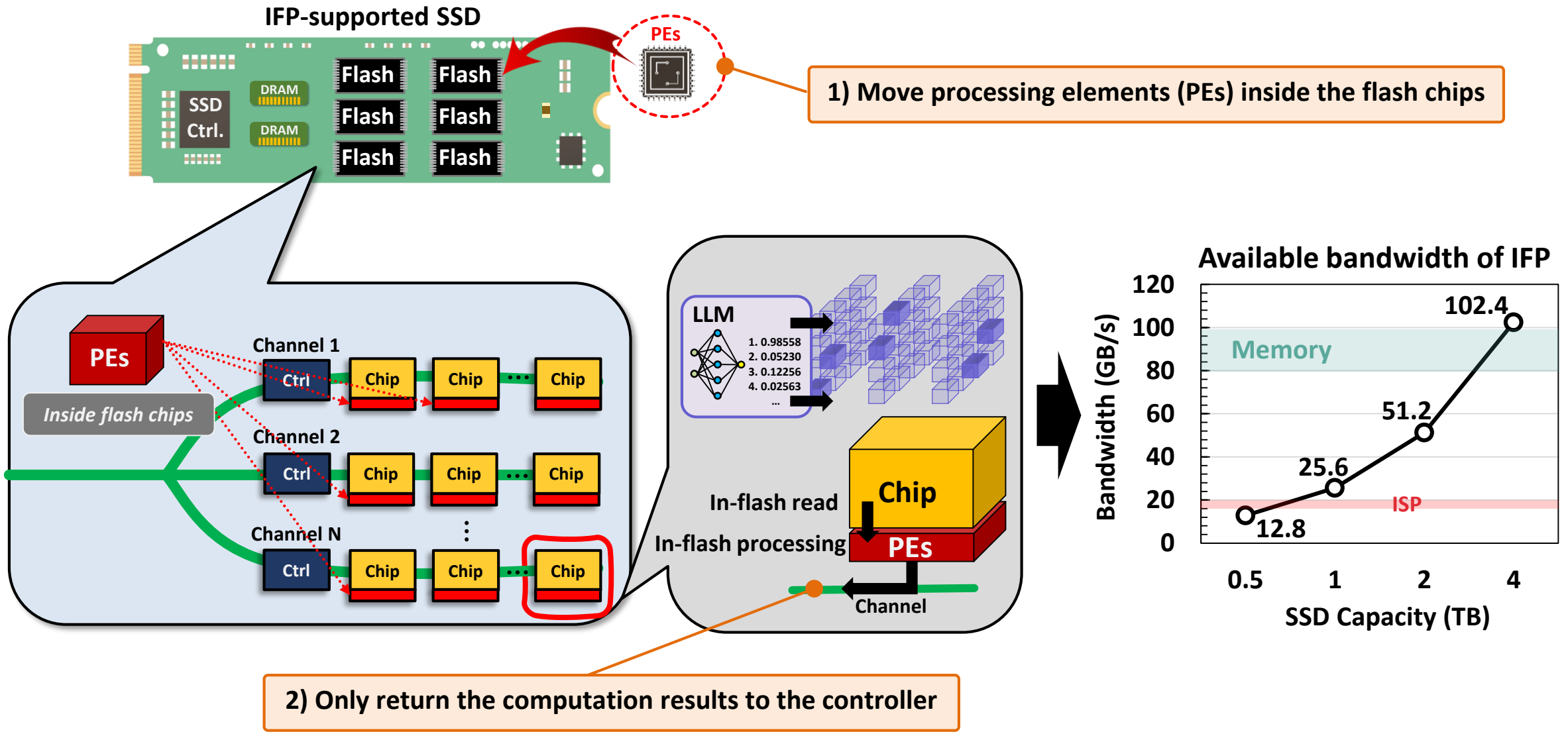


Limitations of ISP

- The performance gain of ISP is constrained by the limited flash channel bandwidth



Our Approach: In-Flash Processing (IFP)



Talk Outline

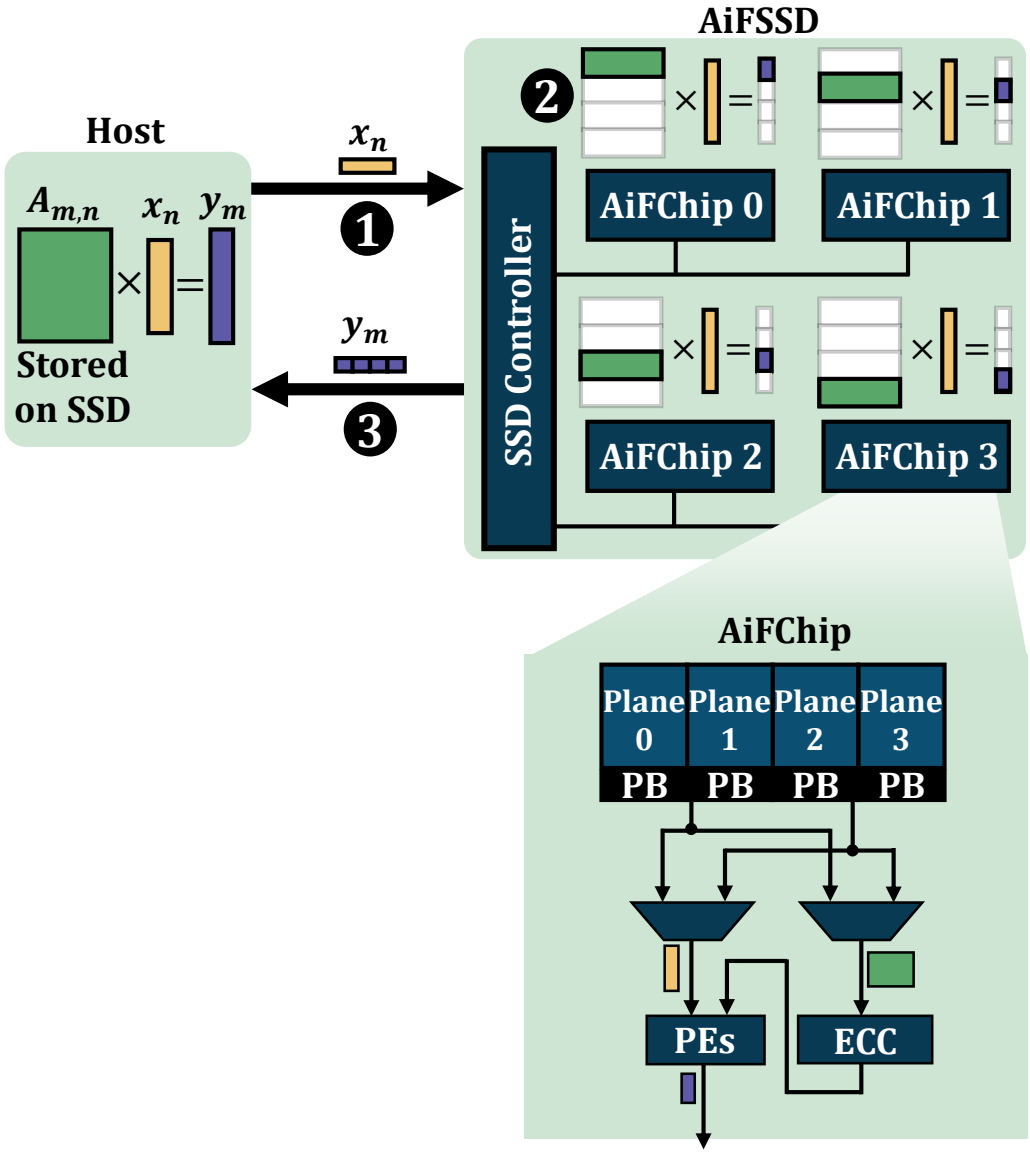
001 On-Device LLM Inference

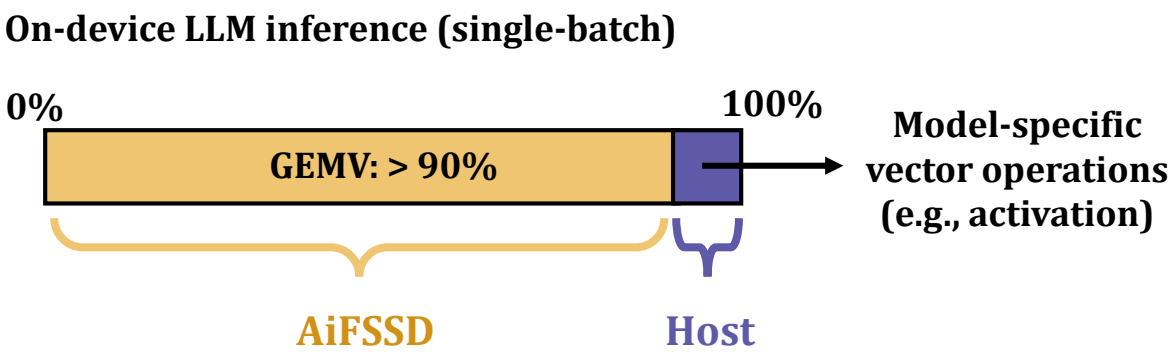
002 SSD-Centric Approaches

003 AiF: Accelerator-in-Flash

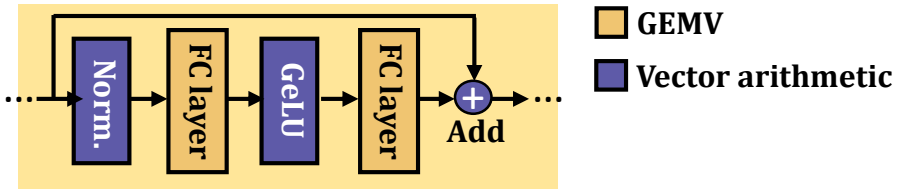
004 Evaluation results and Summary

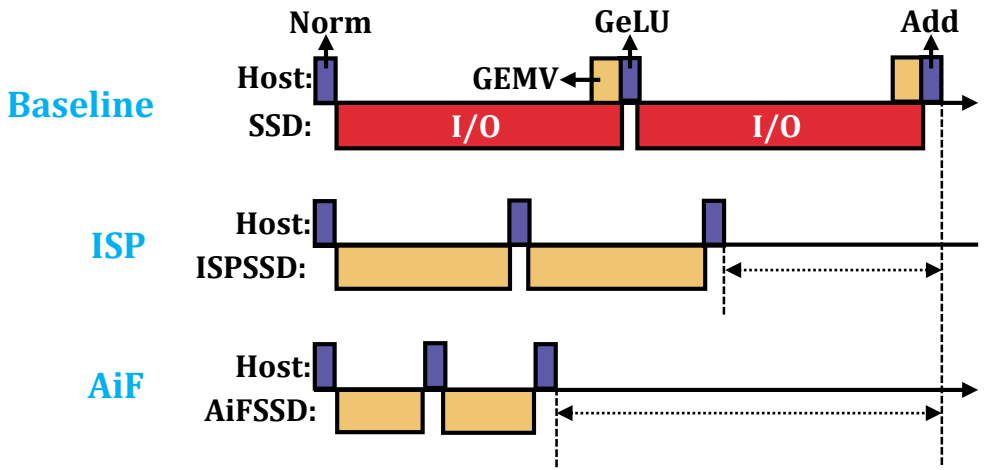
Accelerator-in-Flash (AiF)





Example: Feedforward Network of LLaMA-2

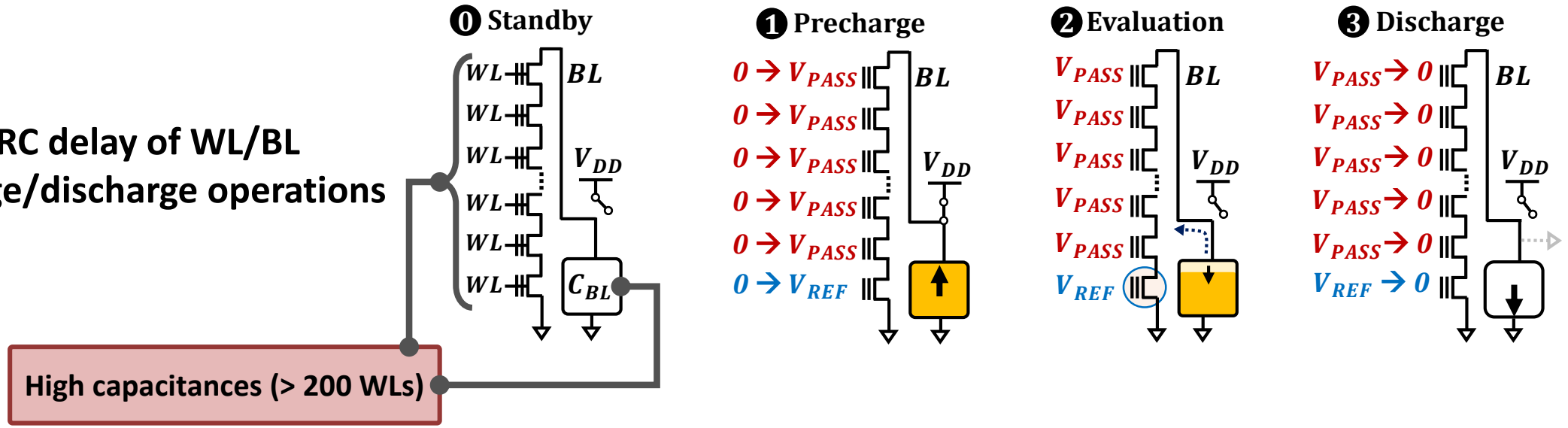




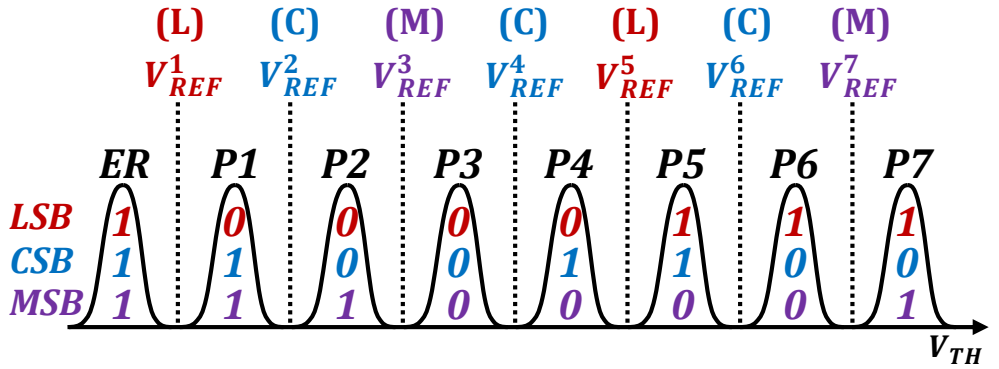
Challenge 1: High Read Latency

- A 1-TiB SSD (16 flash chips) can only reach **an internal read bandwidth of 25.6 GB/s (-74.4 %)**, due to the **high latency of page read.**

1) High RC delay of WL/BL precharge/discharge operations



2) Multiple read operations

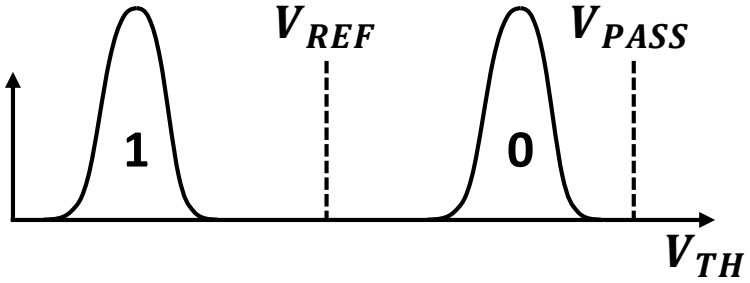


LSB page: 2 reads

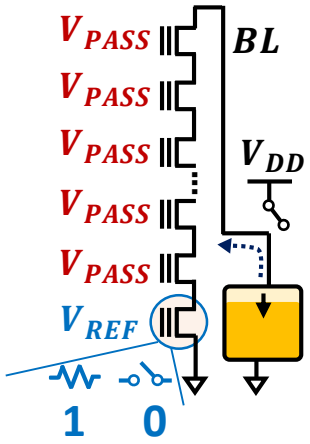
CSB page: 3 reads

MSB page: 2 reads

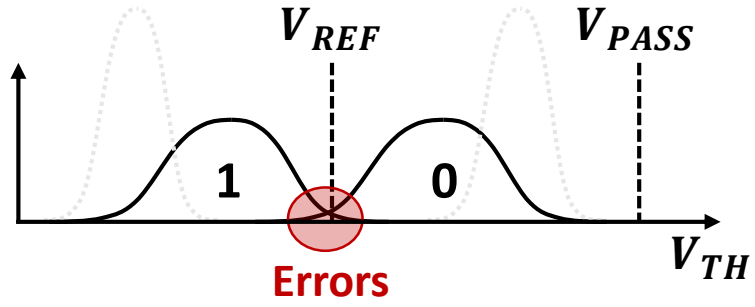
Challenge 2: High Error Rate of Flash Memory

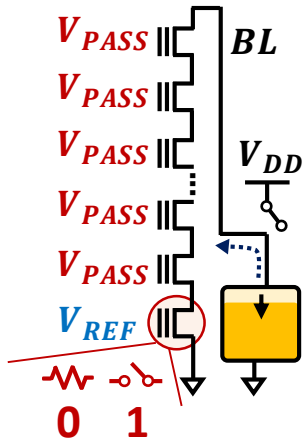


Flash read operation

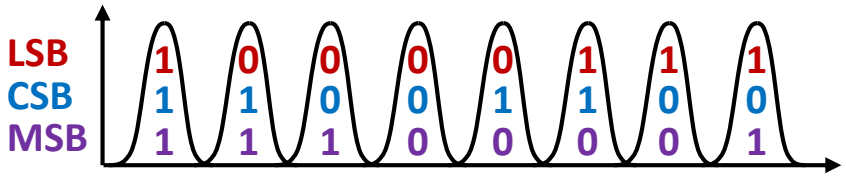


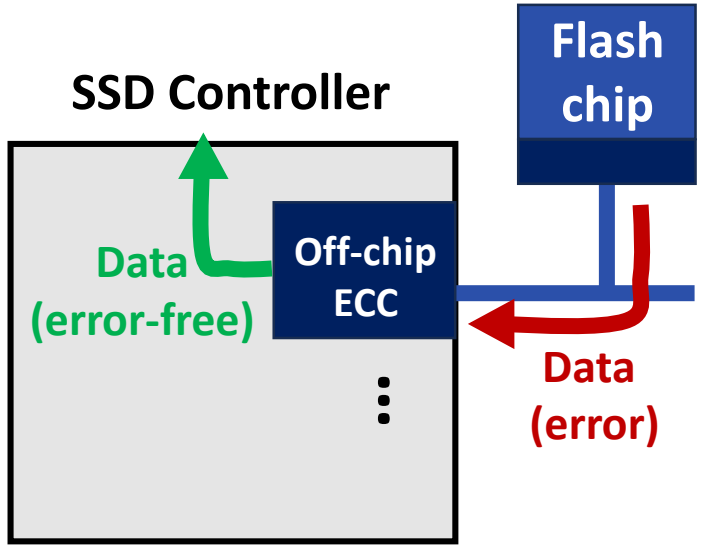
Retention/Endurance





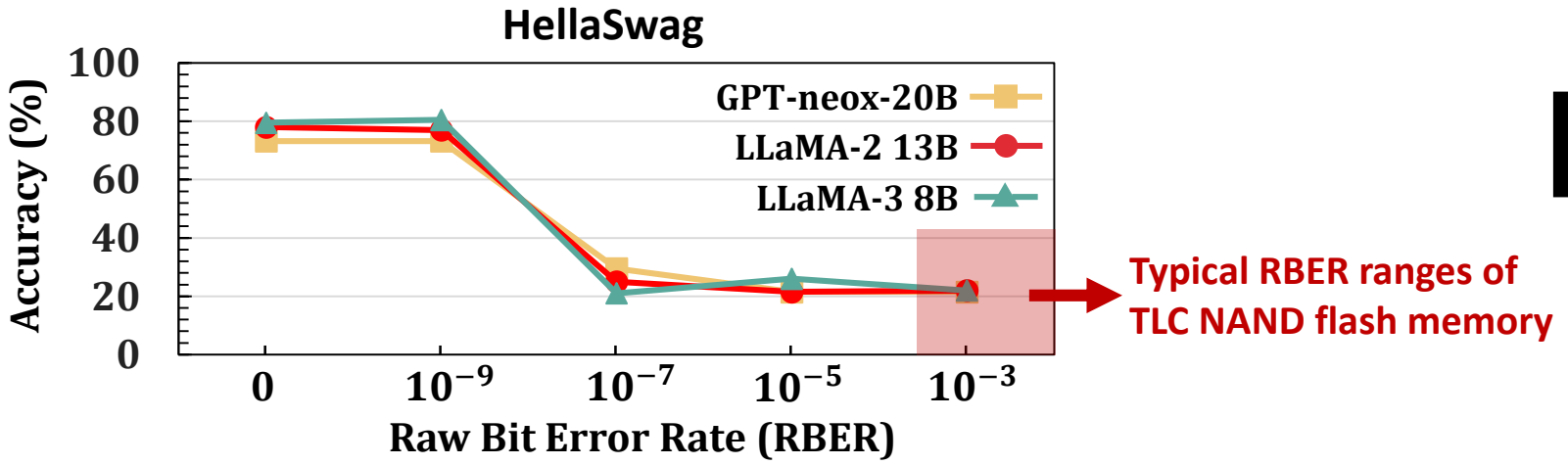
Modern TLC (3b/cell) memory

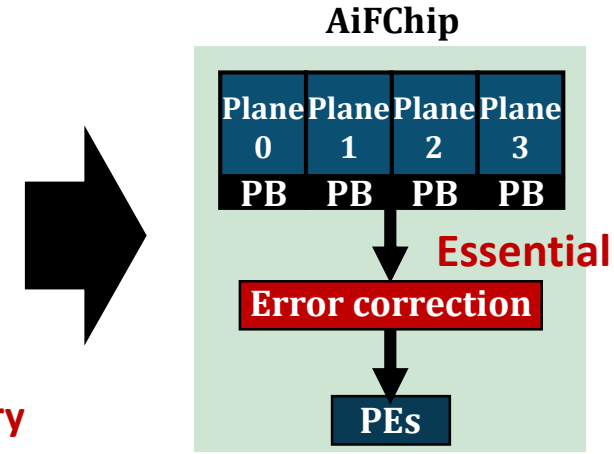




Challenge 2: High Error Rate (cont'd)

- Error sensitivity analysis of LLM inference





- Huge overheads of on-chip ECC decoder

Required decoding throughput:

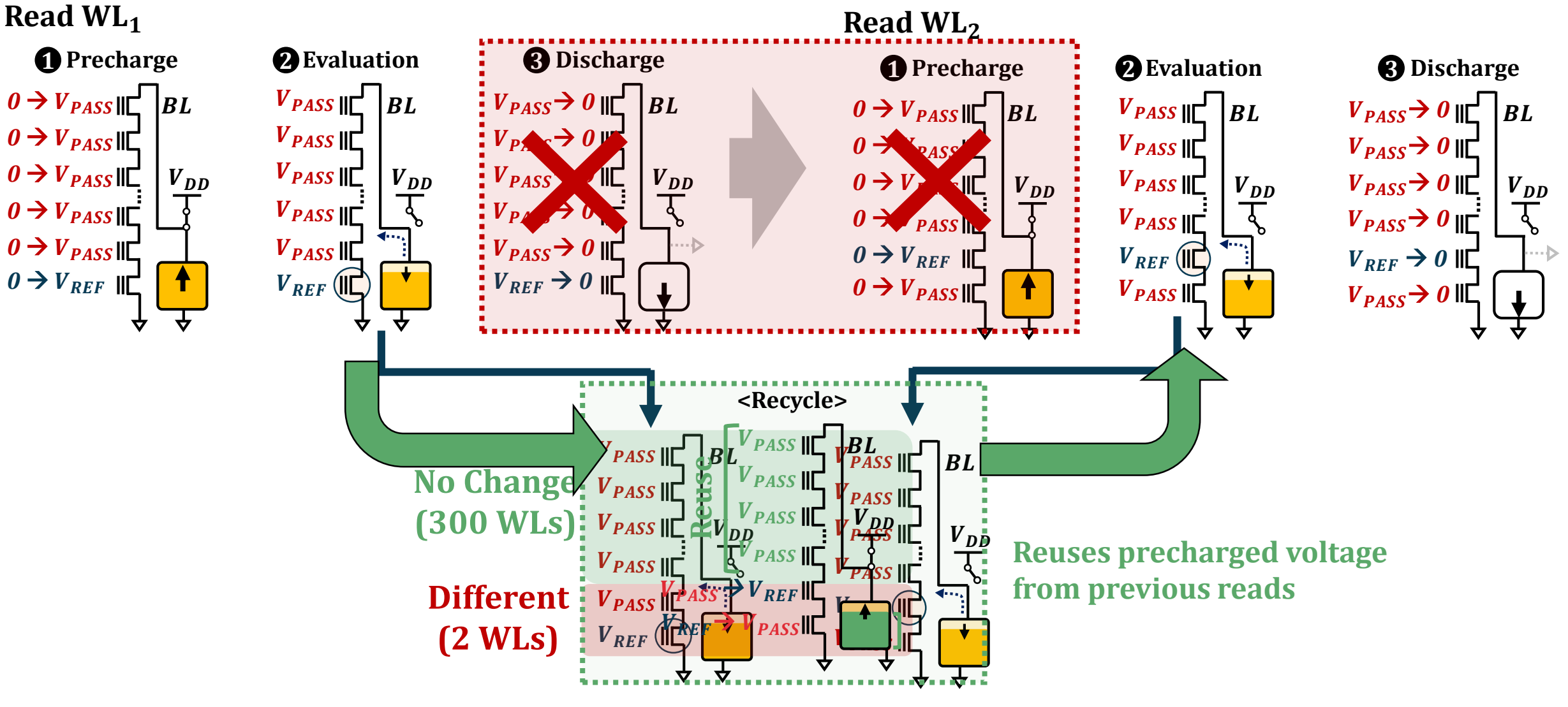
- ✓ **Off-chip ECC:** **Low** external bandwidth (4~8 GB/s)
- ✓ **On-chip ECC:** **High** internal bandwidth

- If internal read bandwidth is 100 GB/s, an overall on-chip ECC decoders can consume
  - ✓ **10.69 W power**
  - ✓ **40.48 mm^2 silicon area**

A lightweight on-chip ECC scheme is required

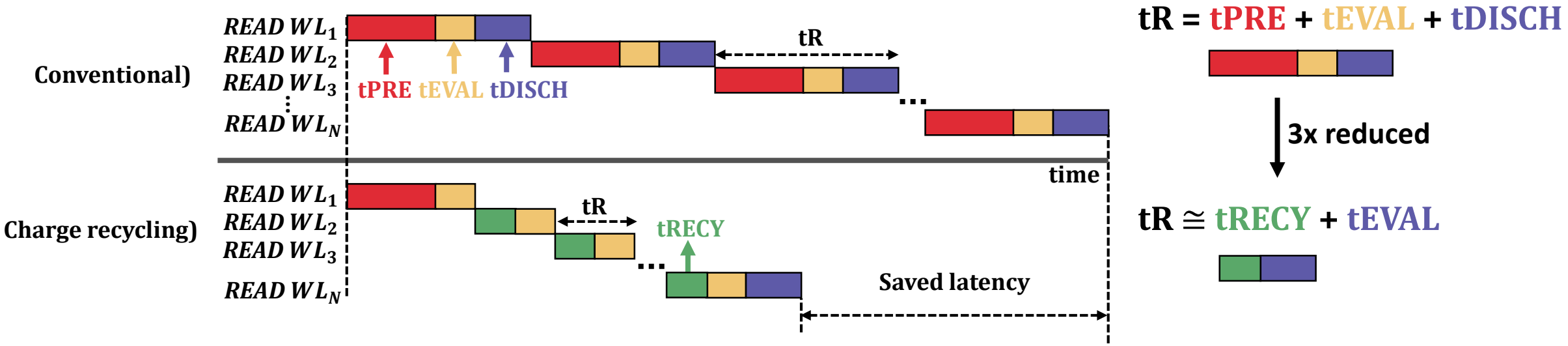
Solution 1: Charge Recycling Read

- Read two consecutive wordlines



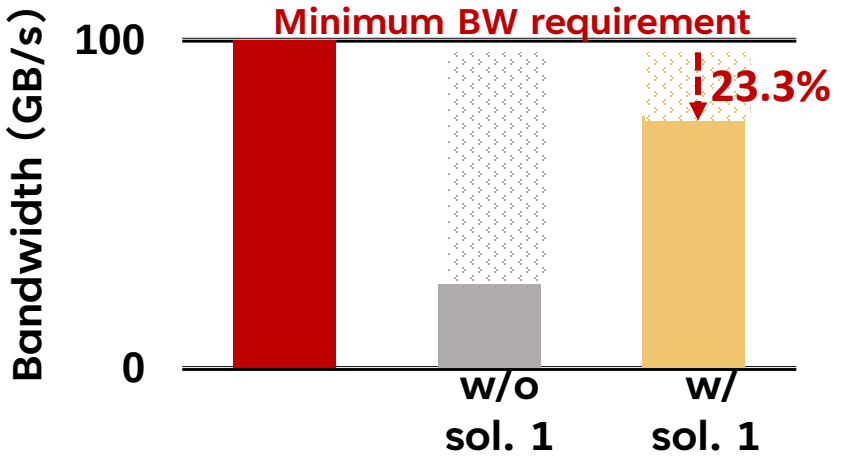
Solution 1: Charge Recycling Read (cont'd)

- Read ***n*** consecutive wordlines



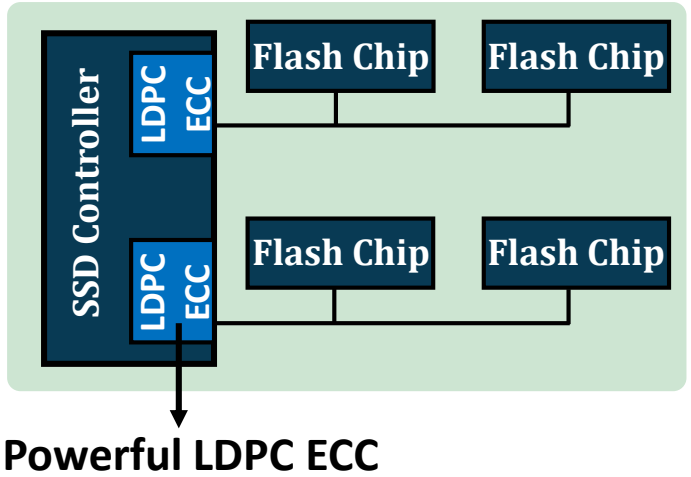
By storing LLM parameters on consecutive WLs,

- 3x improved **bandwidth**
- 3.6x improved **energy efficiency**

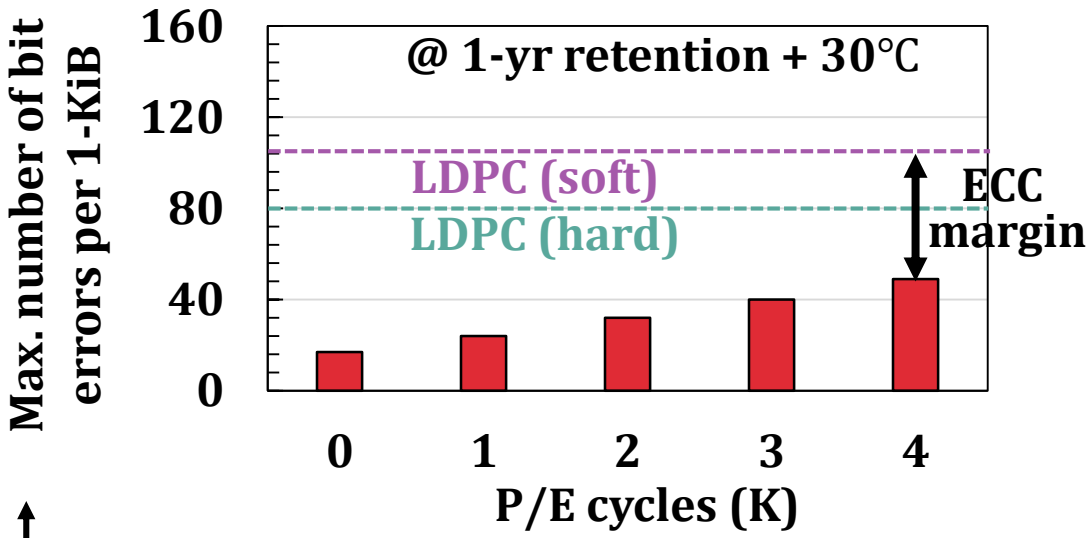


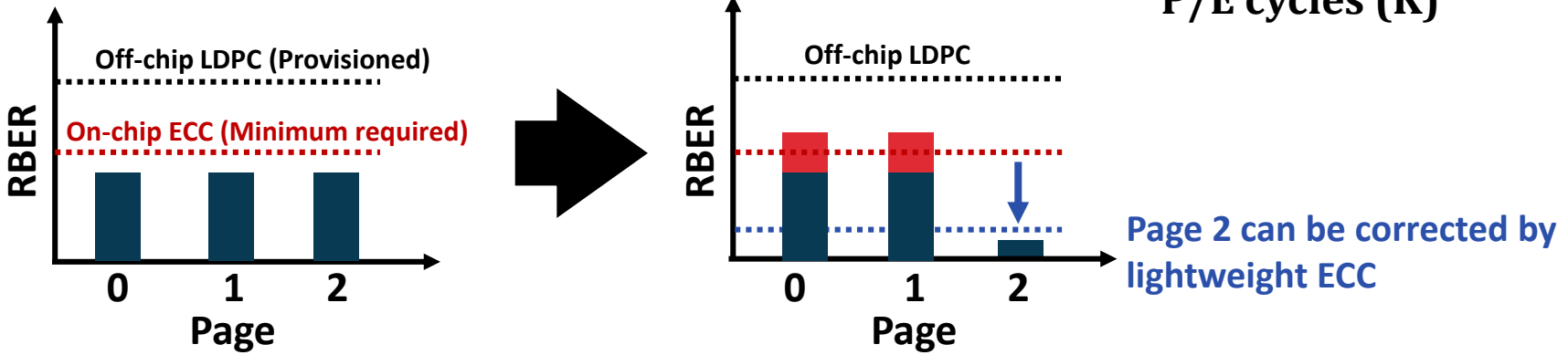
Solution 2: Biased-Error Encoding

- Key Idea
  - ❶ Make errors uneven across pages
  - ❷ Store LLM parameters only on reliable pages

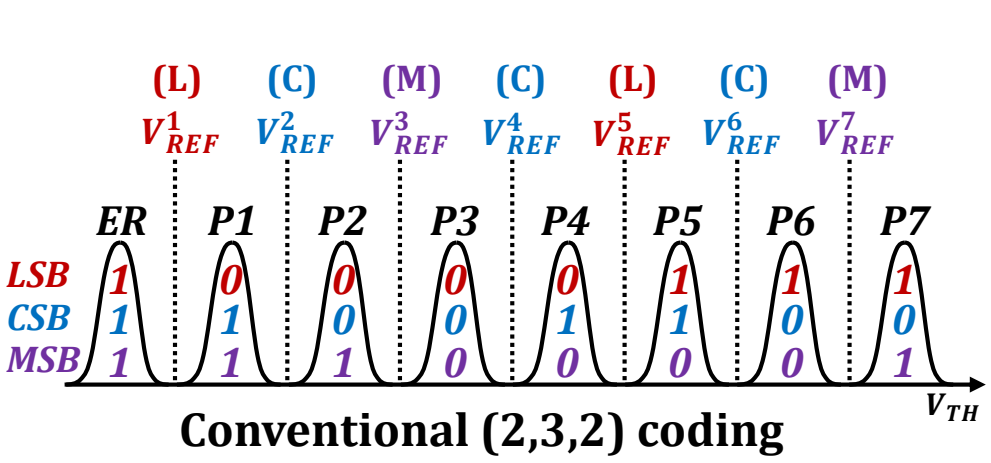


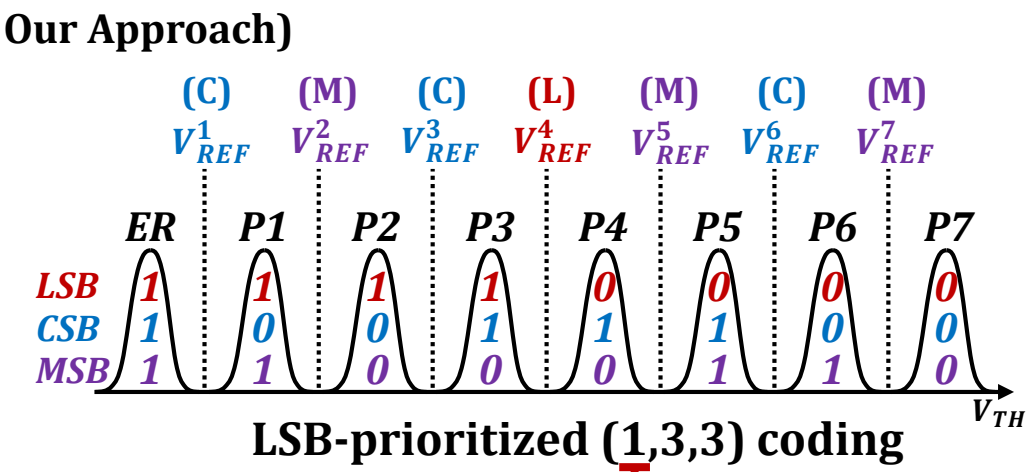
<Characterization results from 160 TLC flash chips>



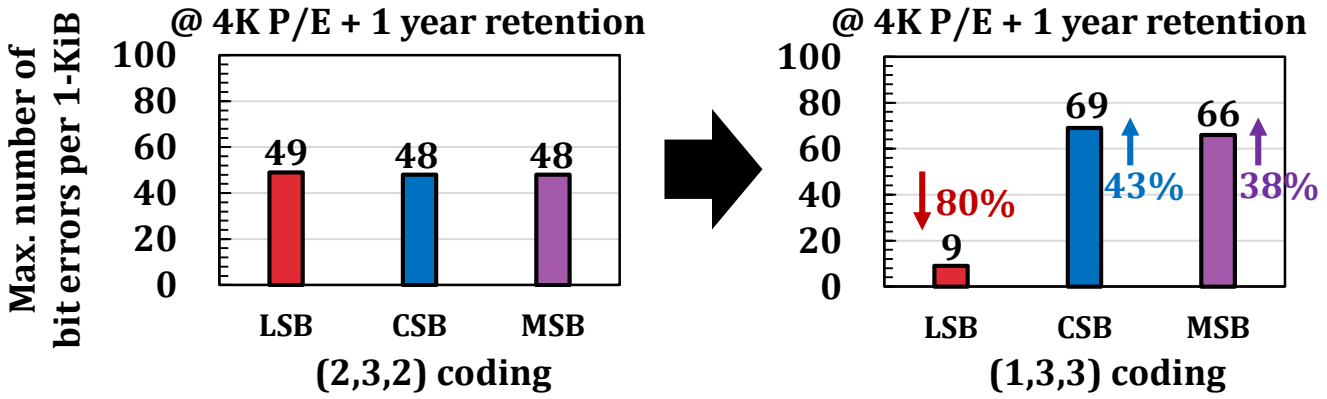


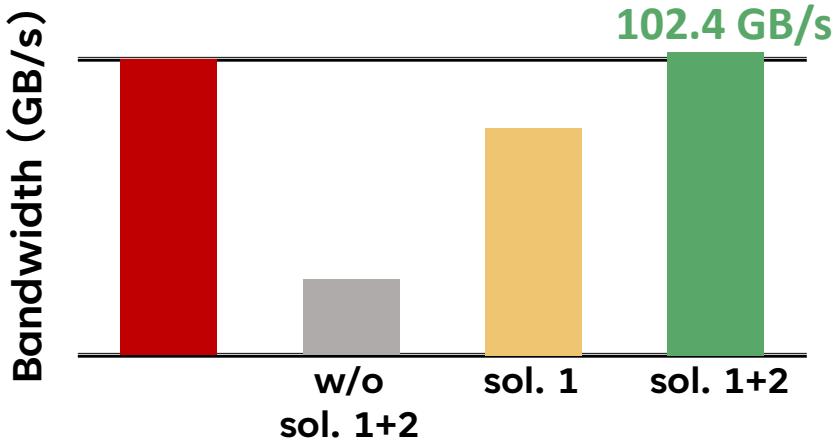
Solution 2: Biased-Error Encoding (cont'd)





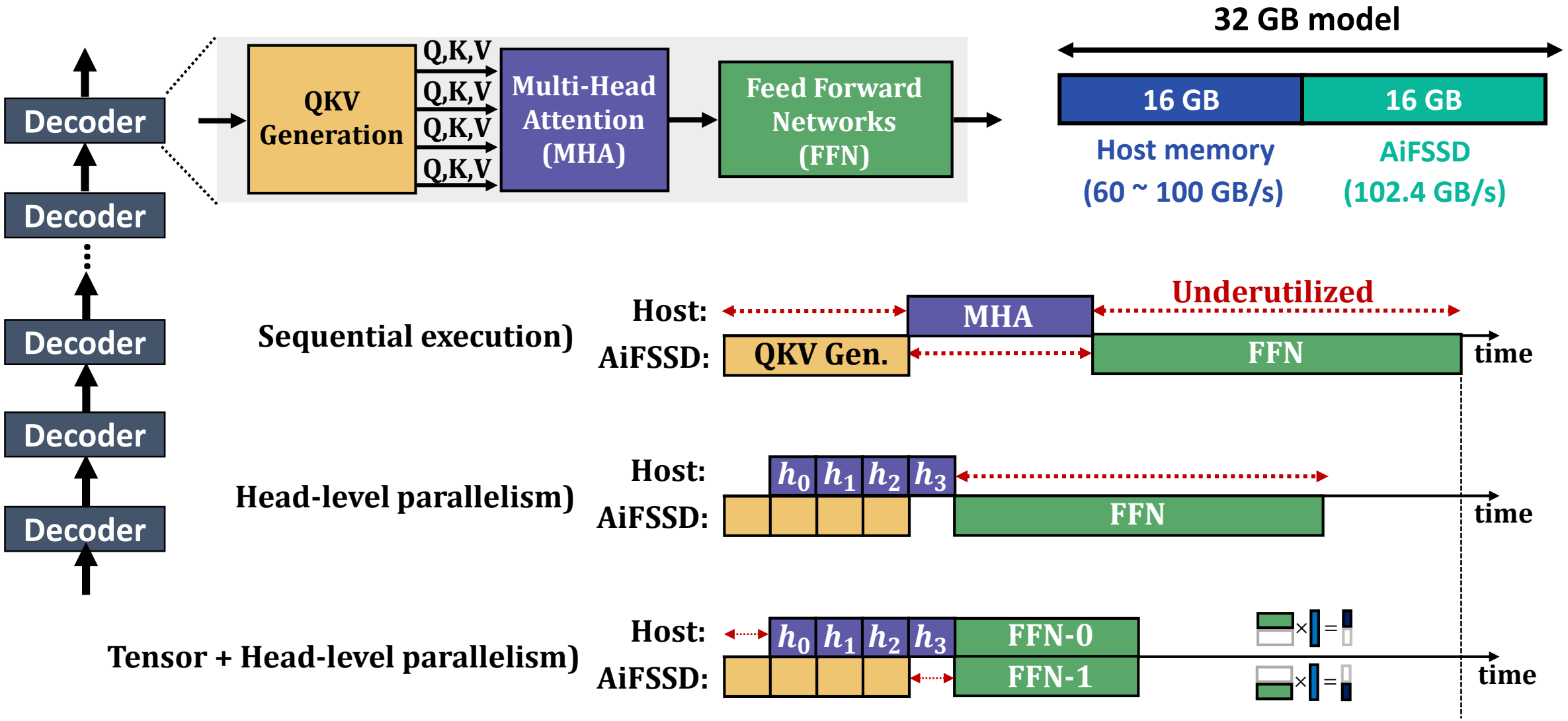
Required sensing count of LSB pages: 2 → 1





Store LLM parameters only on reliable & fast LSB pages

Host-AiFSSD Collaborative Inference



Talk Outline

001 On-Device LLM Inference

002 SSD-Centric Approaches

003 AiF: Accelerator-in-Flash

004 Evaluation results and Conclusion

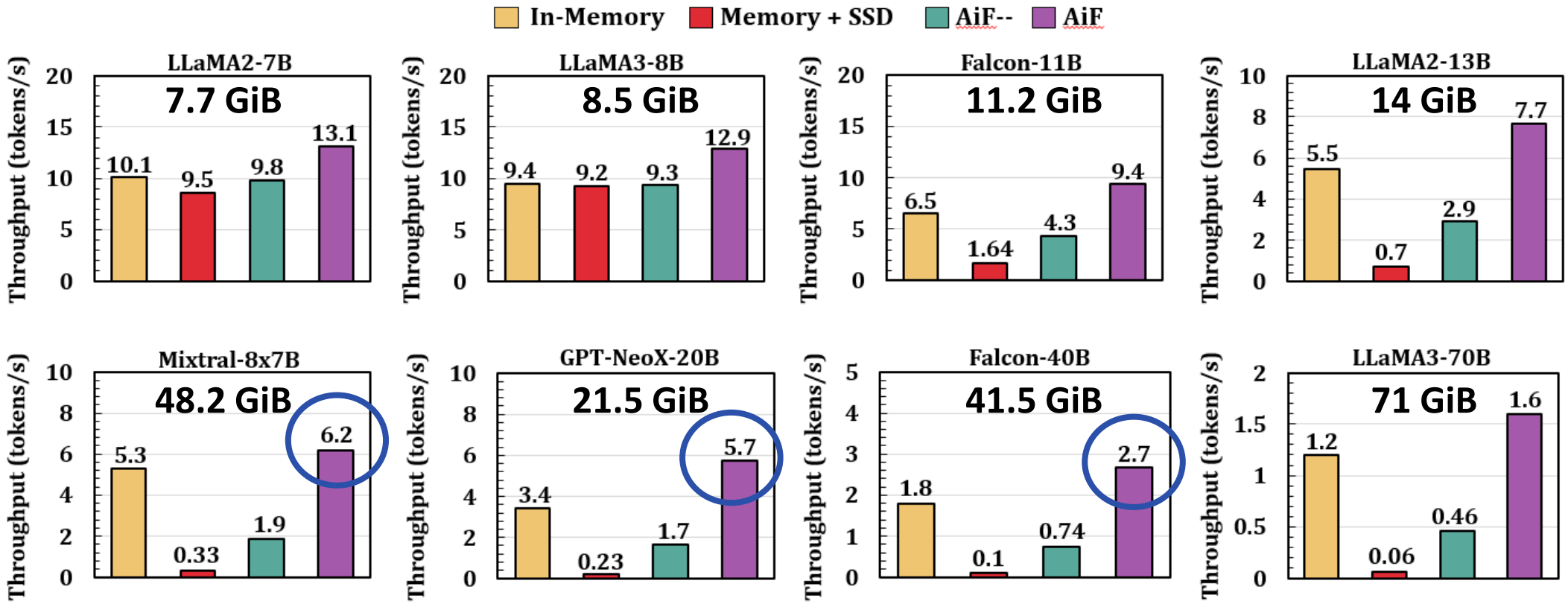
Evaluation Setup

- **LLM Inference Engine:** llama.cpp
- **AiFSSD Emulator:** NVMeVirt

- **Comparison schemes**
  - **1) In-Memory:** An ideal system without memory constraints
  - **2) Memory + SSD:** 8 GiB memory constraint
  - **3) AiF--:** AiF without two optimization techniques
  - **4) AiF**

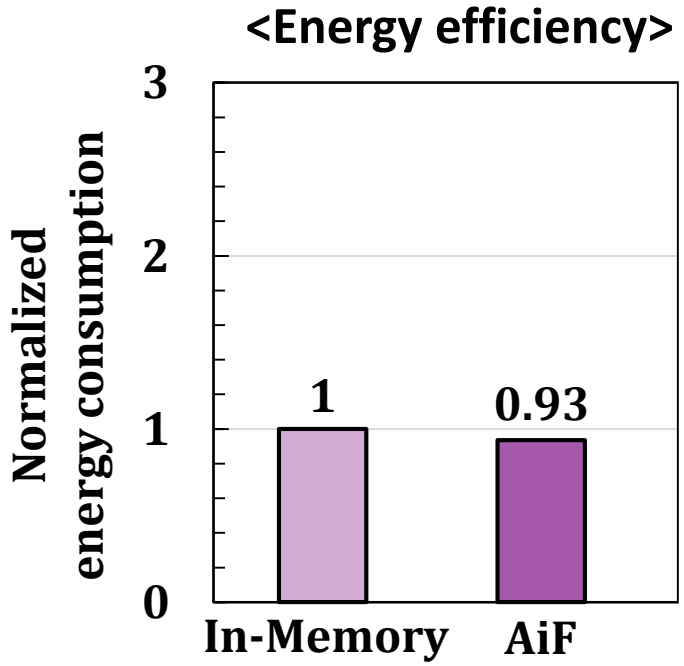
- **Evaluated models**
  - **LLaMA2-7B** (7.7 GiB)
  - **LLaMA3-8B** (8.5 GiB)
  - **Falcon-11B** (11.2 GiB)
  - **LLaMA2-13B** (14 GiB)
  - **Mixtral-8x7B** (48.2 GiB)
  - **GPT-NeoX-20B** (21.5 GiB)
  - **Falcon-40B** (41.5 GiB)
  - **LLaMA3-70B** (71 GiB)

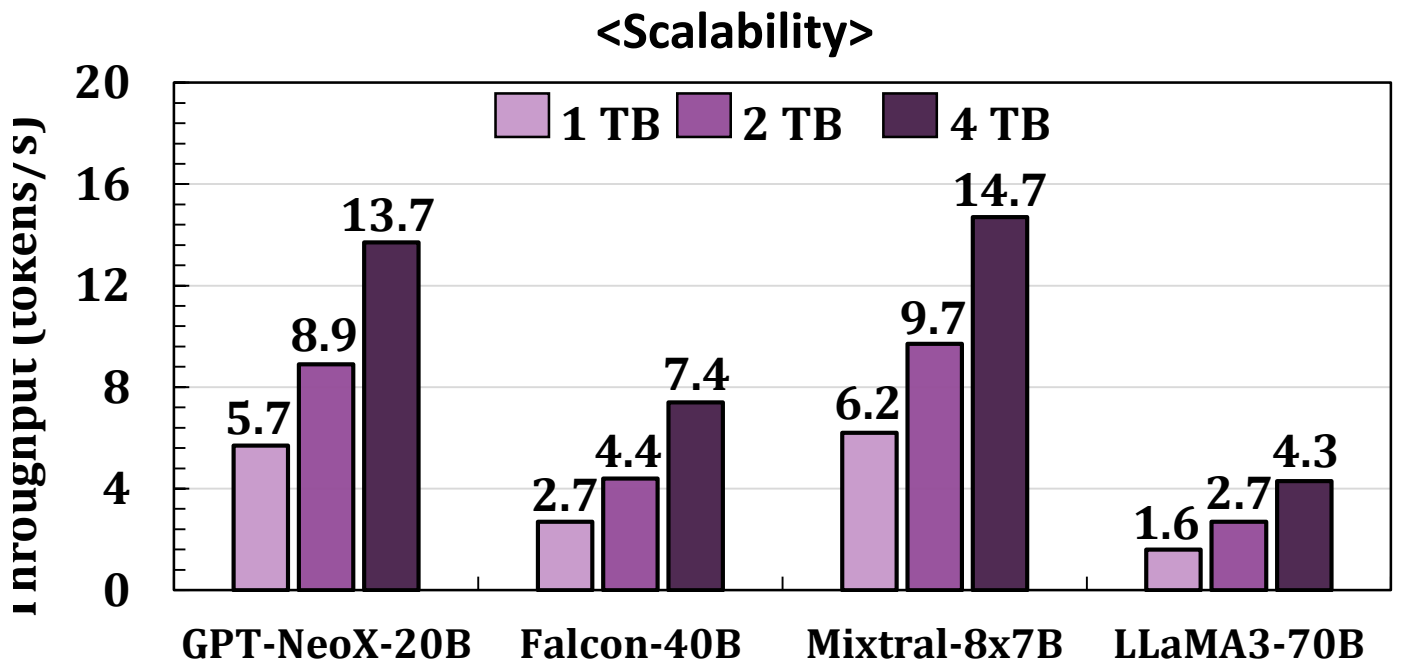
Generation Throughput



AiF delivers 1.4x higher throughput over In-Memory
with 8-GB host memory

Energy Efficiency & Scalability





AiF consumes 7% less energy than In-Memory inference

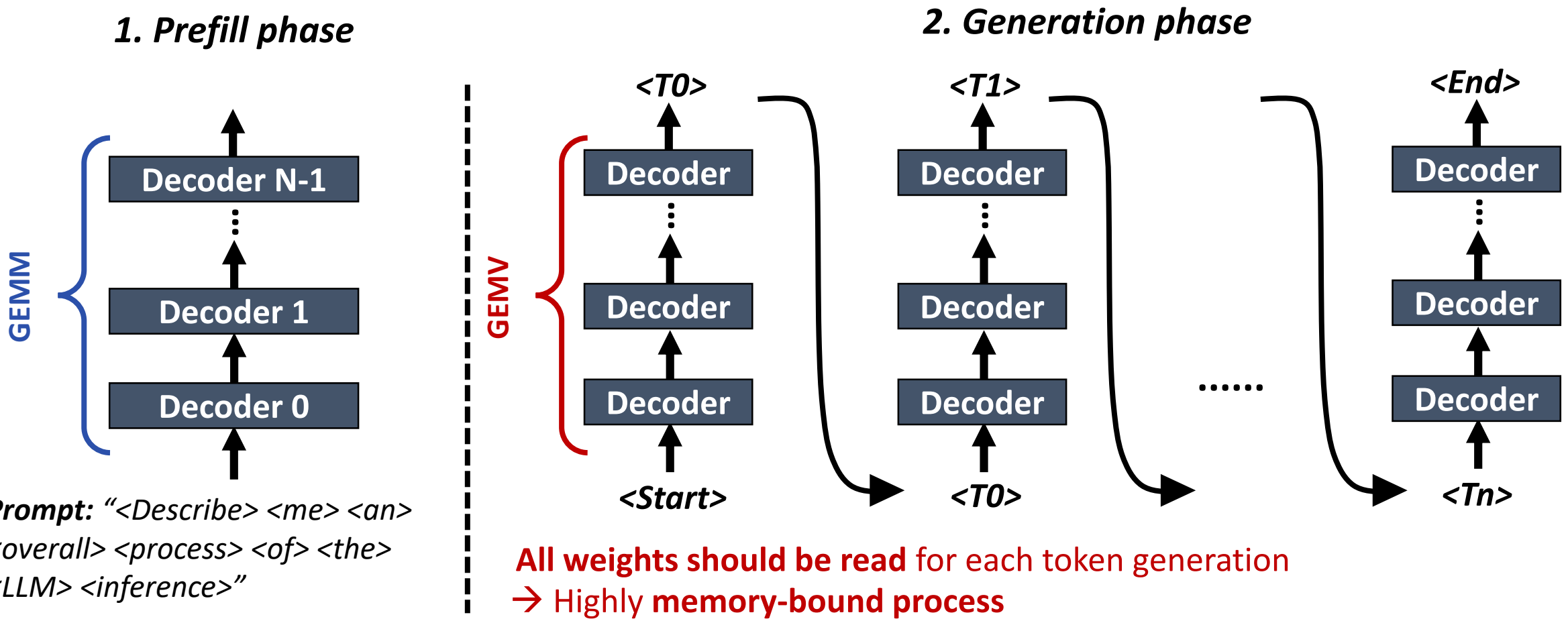
AiF scales well with SSD capacity

Conclusions

- **Proposed Accelerator-in-Flash (AiF) for on-device LLM.**
  - Utilize the internal bandwidth of SSDs by **embedding computation into flash chips**
- **Developed two key techniques.**
  - **Charge recycling read:** Reuses the voltage setup from prior reads to reduce read latency
  - **Biased-error encoding:** Prioritizes reliability and speed for pages storing LLM parameters
- **Improved generation throughout by 14.6x and 1.4x over the baseline SSD and an ideal in-memory scheme.**

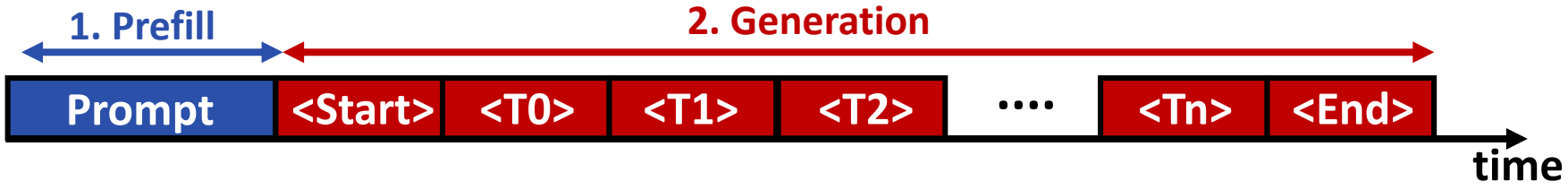
Thanks for Listening!
Any Questions? ☺

Execution Flow of On-Device LLM Inference

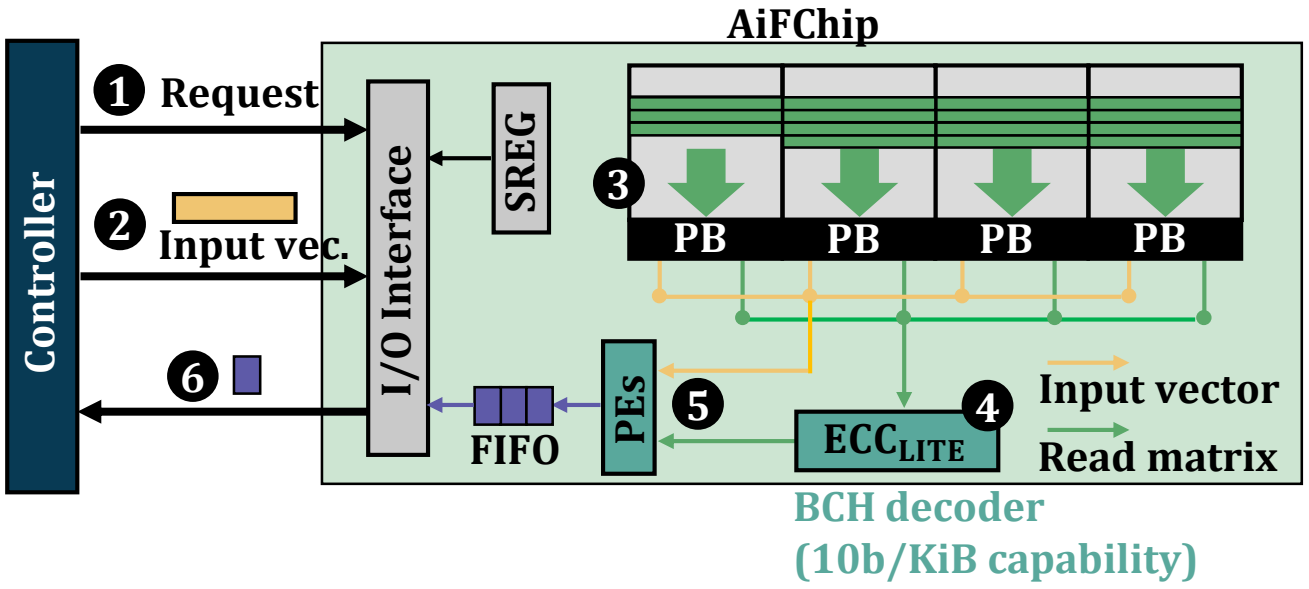


Prompt: *"<Describe> <me> <an> <overall> <process> <of> <the> <LLM> <inference>"*

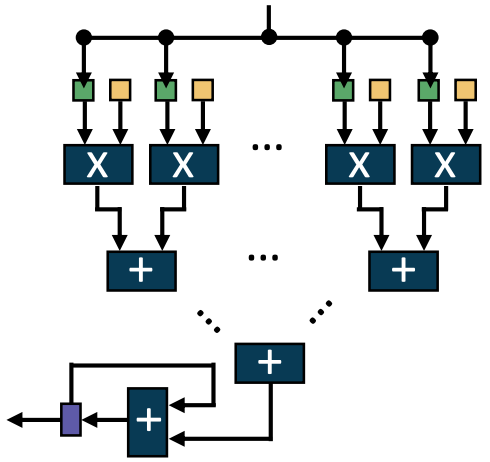
All weights should be read for each token generation
→ Highly **memory-bound process**



On-chip Computation Flow



- **ECC_{LITE}** : BCH decoder (10b/KiB capability)
- **PEs:** Multiplier and Adder tree



@ 200 MHz, 45 nm

Module Name	Area [mm^2]	Avg. Power [mW]
ECC_{LITE}	0.167	45.1
Processing Elements (PEs)	0.026	3.98
Others	0.016	2.6
Total	0.209	51.68

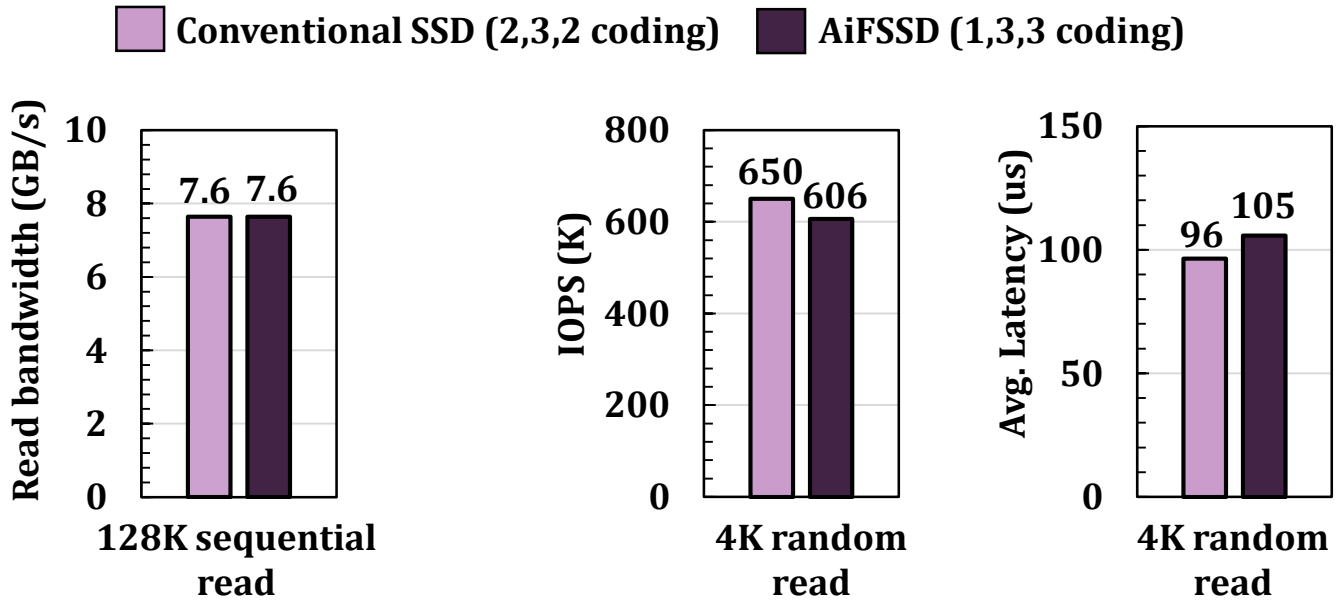
Breakdown of area and power

Evaluated SSD Configurations

Configuration	8 channels; 2 chips/channel; 4 planes/chip; 16-KiB page
tR (2,3,2 coding)	LSB = 37μs; CSB = 46μs; MSB = 37μs;
tR (1,3,3 coding)	LSB = 28μs; CSB = 46μs; MSB = 46μs;
tR (cr-read)	9.7μs;
Bandwidth (PCIe)	8.0 GB/s external I/O bandwidth (PCIe 4.0, 4-lane);
Bandwidth (ONFI)	2.0 GB/s flash channel I/O bandwidth;

Overhead Analysis

tR (2,3,2 coding)	LSB = 37μs; CSB = 46μs; MSB = 37μs;
tR (1,3,3 coding)	LSB = 28μs; CSB = 46μs; MSB = 46μs;



(a) Bandwidth (b) IOPS (c) Latency