HotStorage'26

The Host Is Not Idle: Computational SSDs for Host-Free Checkpointing in Multimodal LLM Training

Seungkwan Kang, Hyungseok Ko, Hyun Dong Lee, Myeonghoon Park, Jinseok Kim, Myoungsoo Jung

KAIST CAMEL

1. **Checkpointing in Large Models**

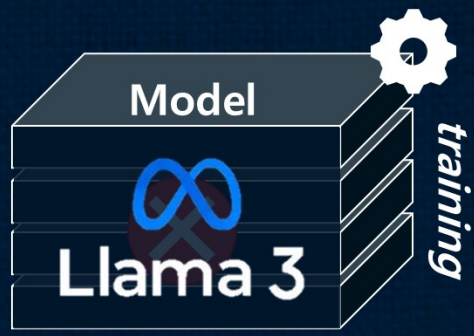
2. Video LLM Training

3. Computational SSD for Host-Free Checkpointing

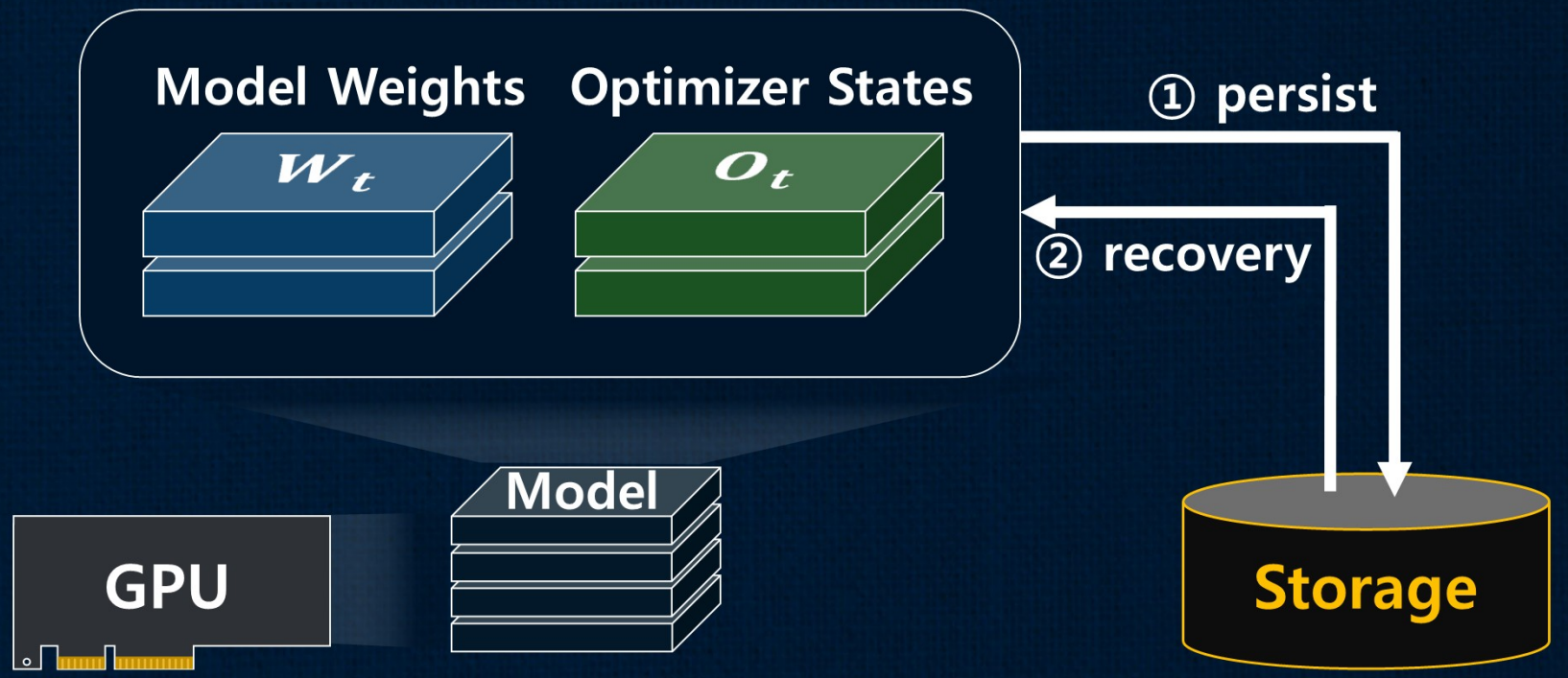
4. Design and Implementation

5. Evaluation Results

Checkpointing



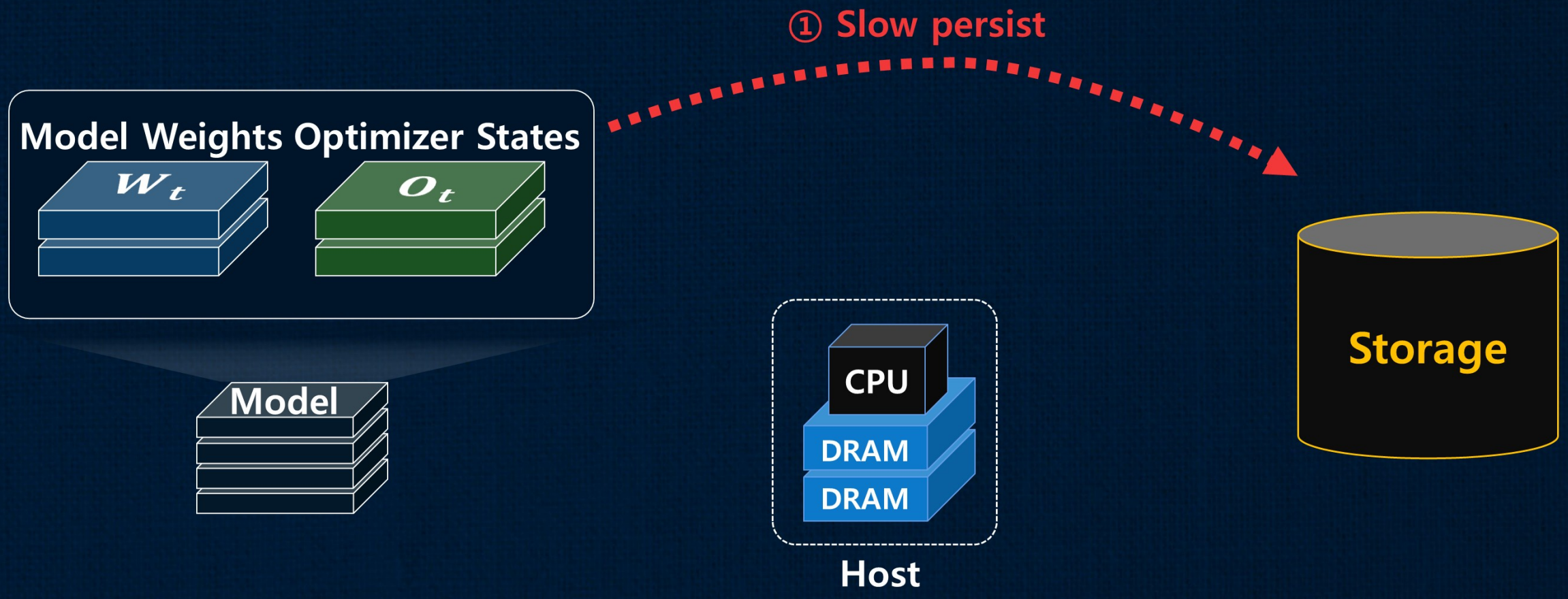
failure per 3 hours



Checkpointing is the standard defense against **hardware failures**.

CAMEL

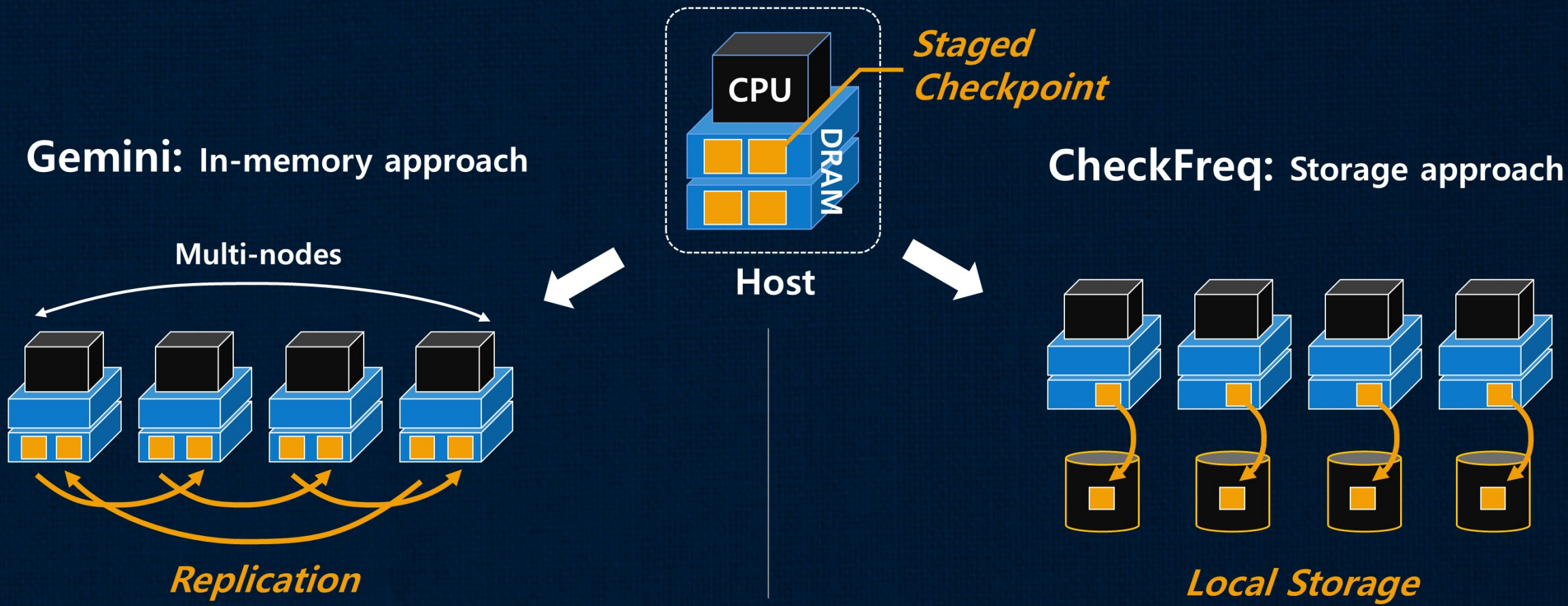
Prior Checkpointing Methods



Existing methods utilizes the host memory as a staging buffer.

CAMEL

Prior Checkpointing Methods



Existing methods differ in the final destination of the checkpoint.

CAMEL

1. Checkpointing in Large Models

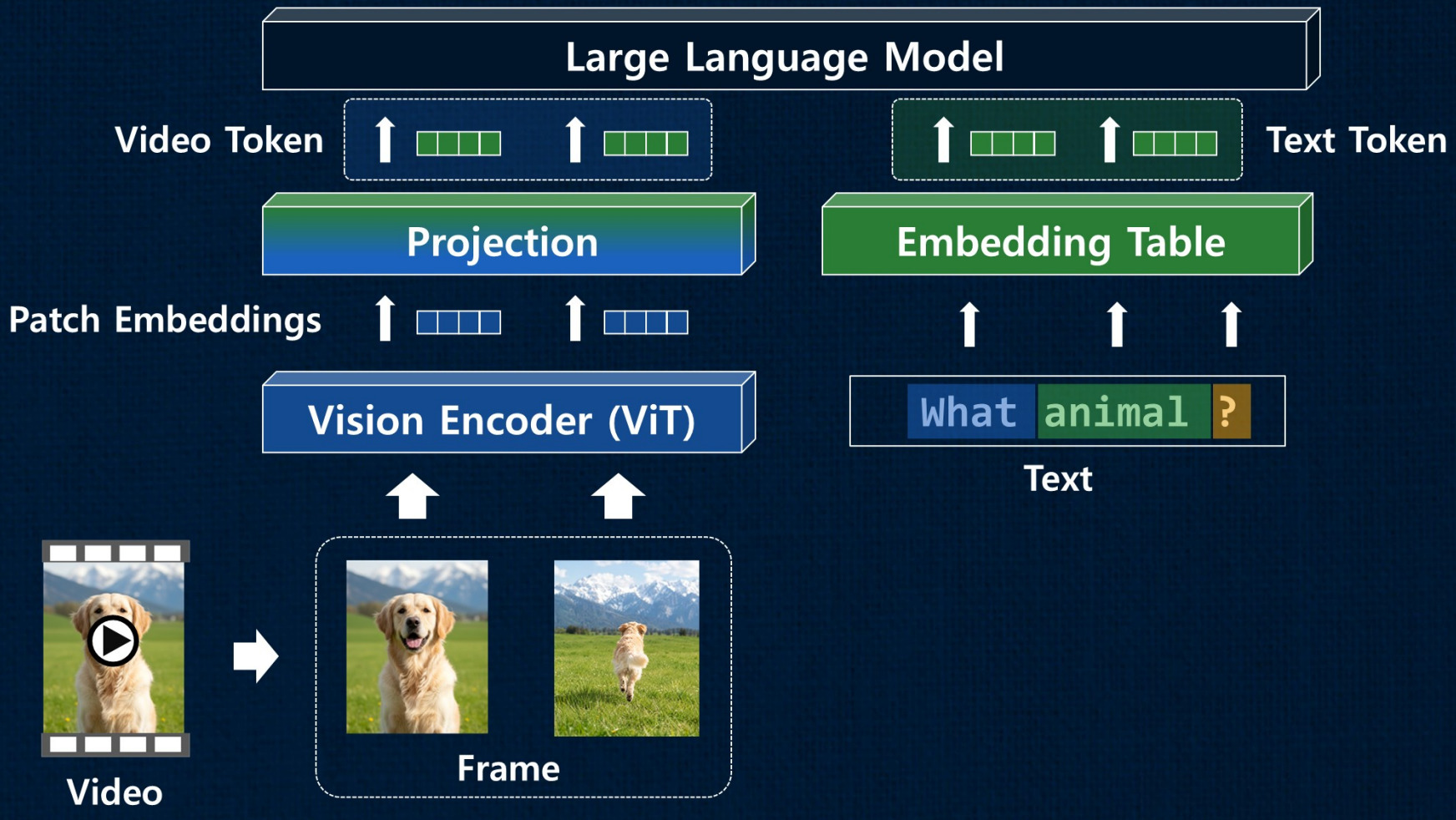
2. Video LLM Training

3. Computational SSD for Host-Free Checkpointing

4. Design and Implementation

5. Evaluation Results

Video LLM

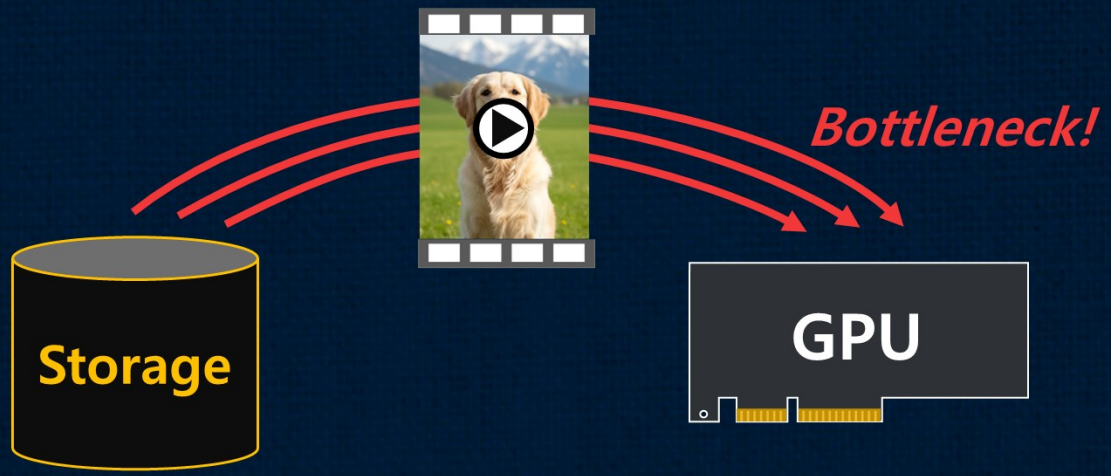


VideoLLMs treat video frames as "tokens".

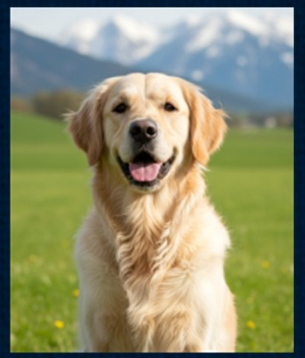
CAMEL

Video LLM's Problems

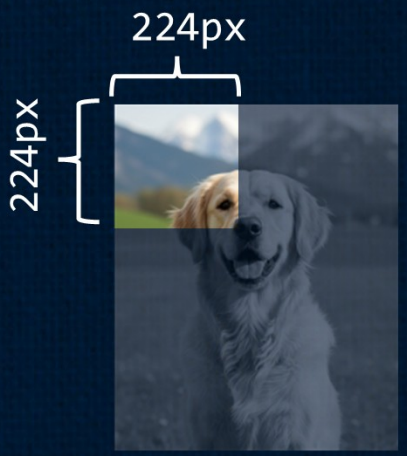
① Raw frames induce massive I/O



② ViT only accepts fixed-size inputs



Raw: 1080p

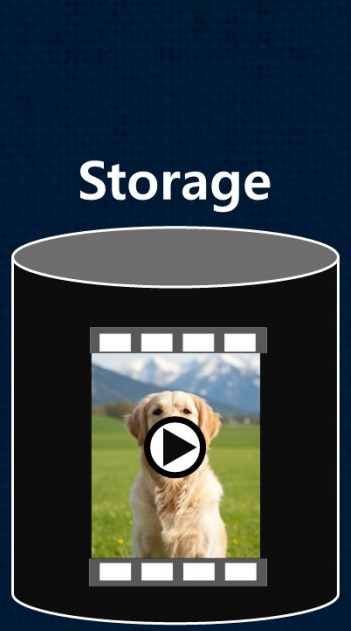


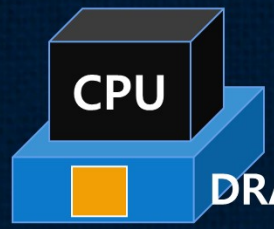
Vit input

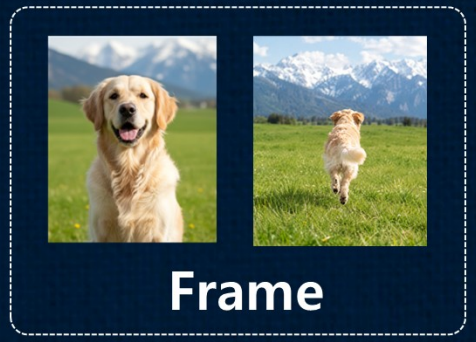
VideoLLMs face two problems compared to text LLMs.

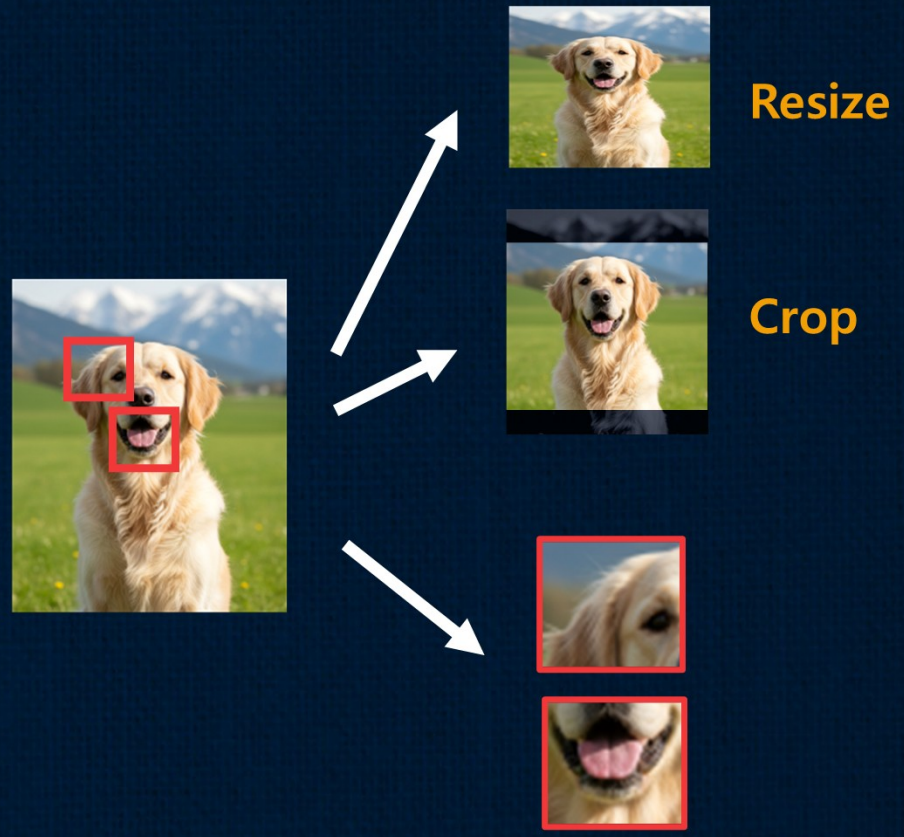
CAMEL

Video LLM Preprocessing





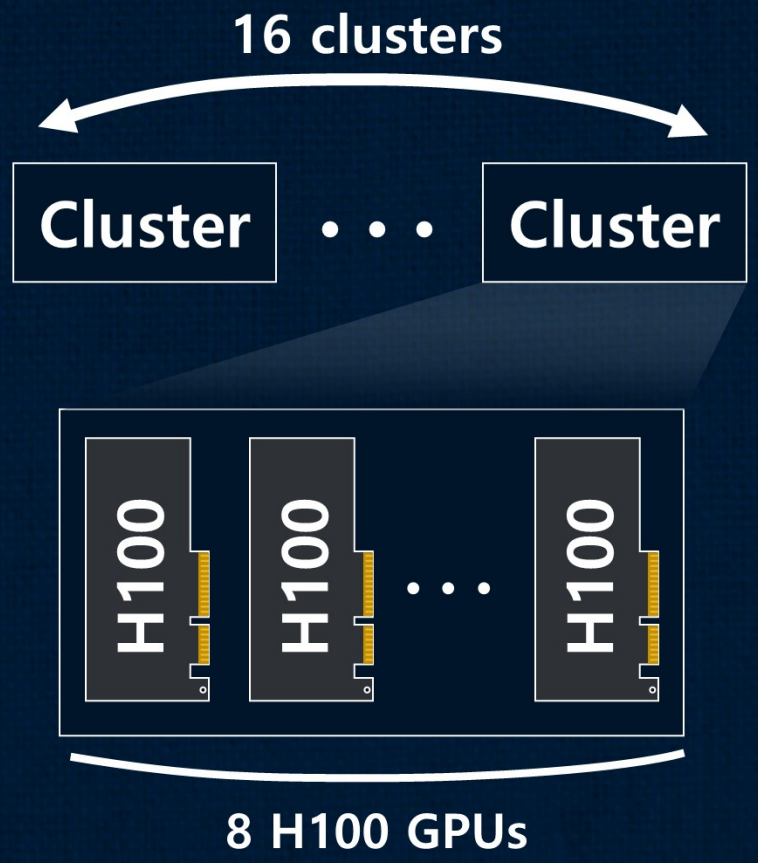




VideoLLMs face two problems compared to text LLMs.

CAMEL

Motivation – Evaluation Setup

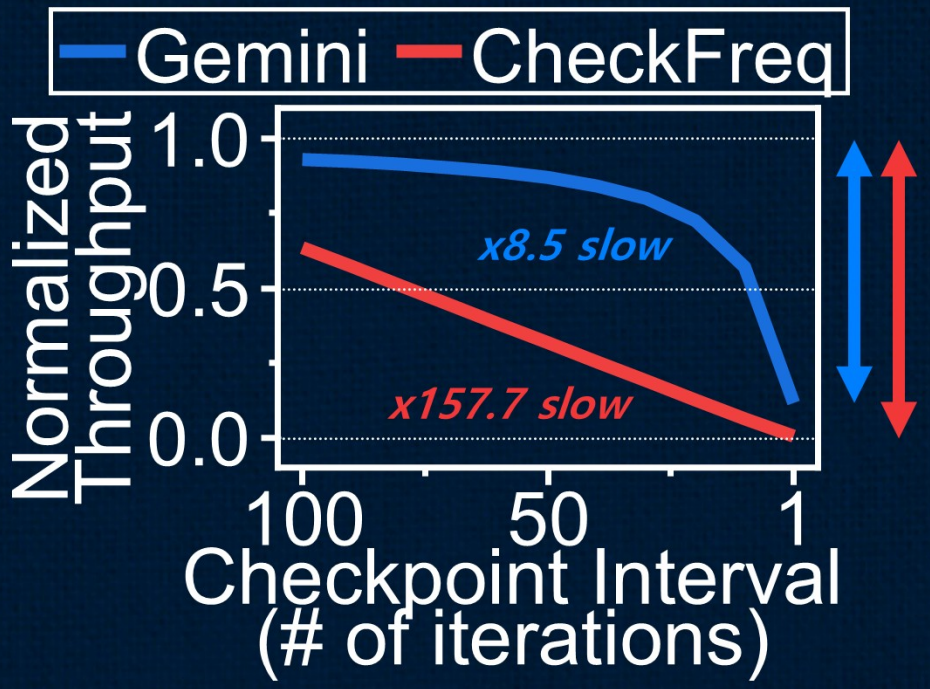


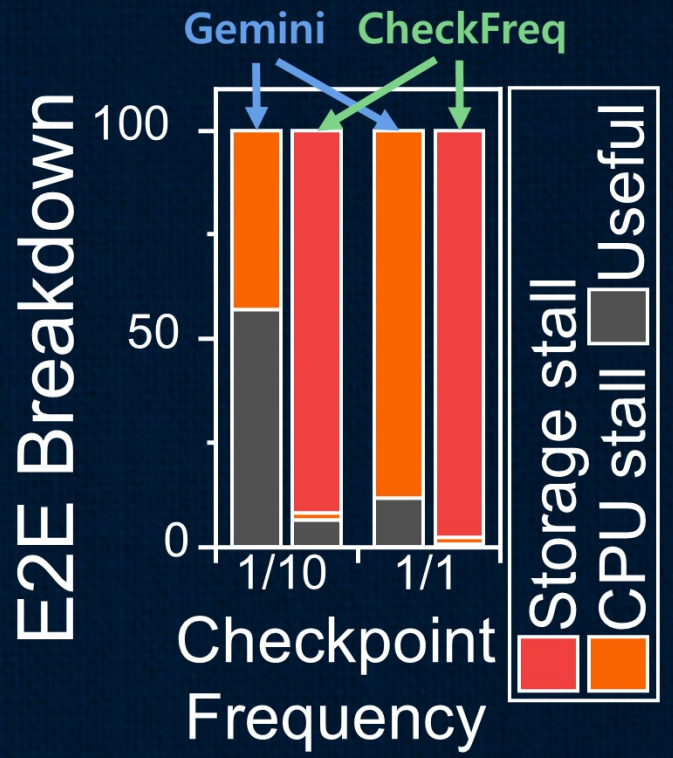
Model Name	Parameter		Experts (act./tot.)
	Total	Active	
LLaVA-Video	7B	7B	Dense
InternVL	78B	78B	Dense
Kimi-VL-A3B	16B	2.8B	8 / 384
Llama-4-Scout	109B	17B	1 /16
Qwen3-VL-30	30B	3B	8 / 128
Qwen3-VL-235	235B	22B	8 / 128
Llama-4-Maverick	400B	17B	1/128

We evaluated seven models on a system based on 128 H100 GPUs.

CAMEL

Motivation – Performance Degradation





Host memory becomes the new bottleneck due to contention.

CAMEL

1. Checkpointing in Large Models

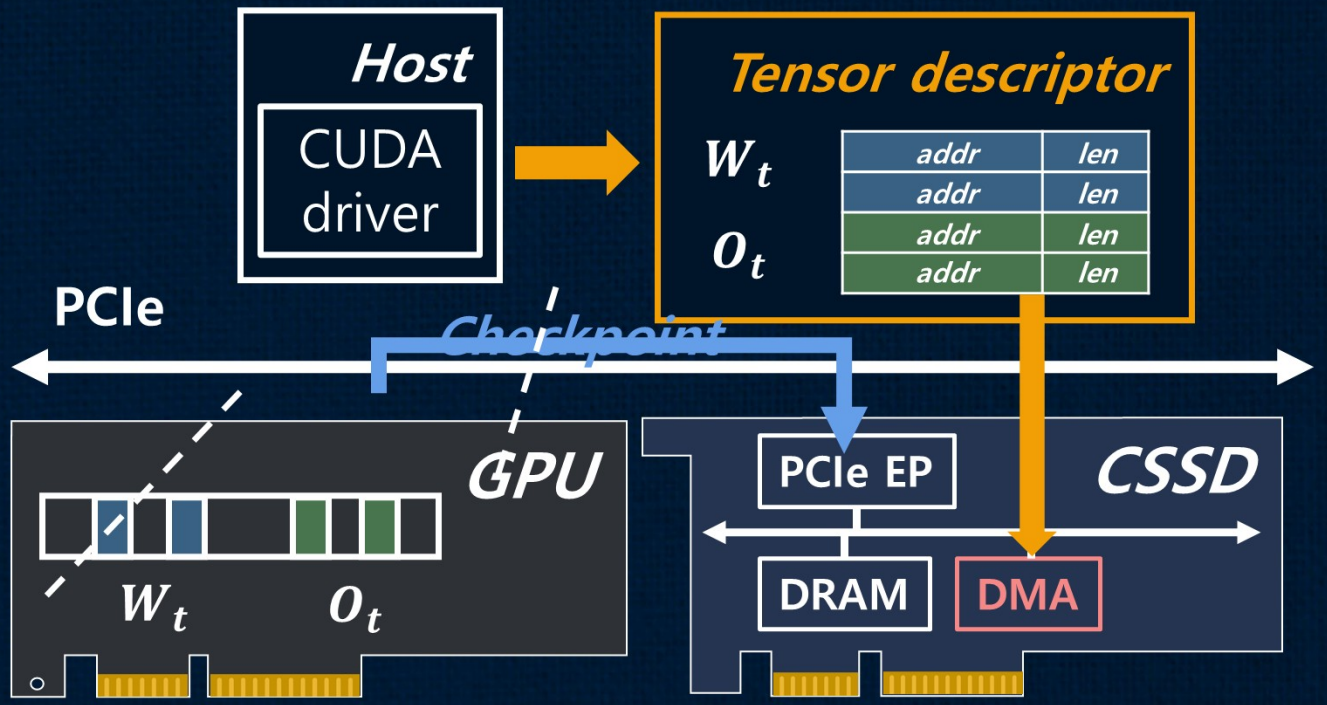
2. Video LLM Training

3. Computational SSD for Host-Free Checkpointing

4. Design and Implementation

5. Evaluation Results

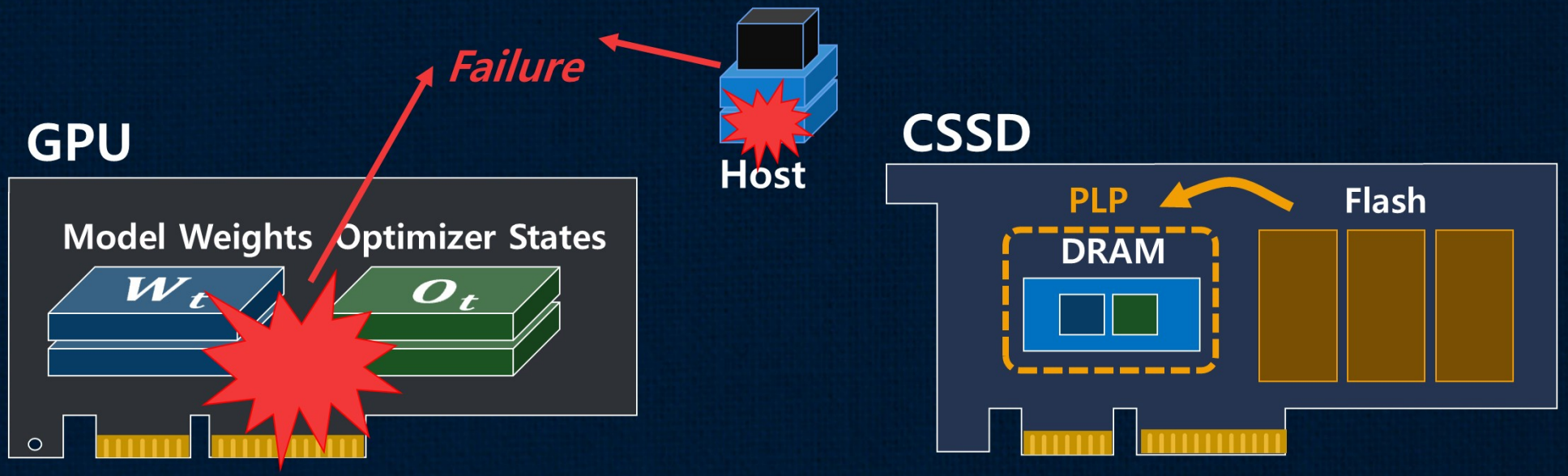
GPU-SSD Direct Checkpointing



We design a tensor descriptor table to enable P2P transfer.

CAMEL

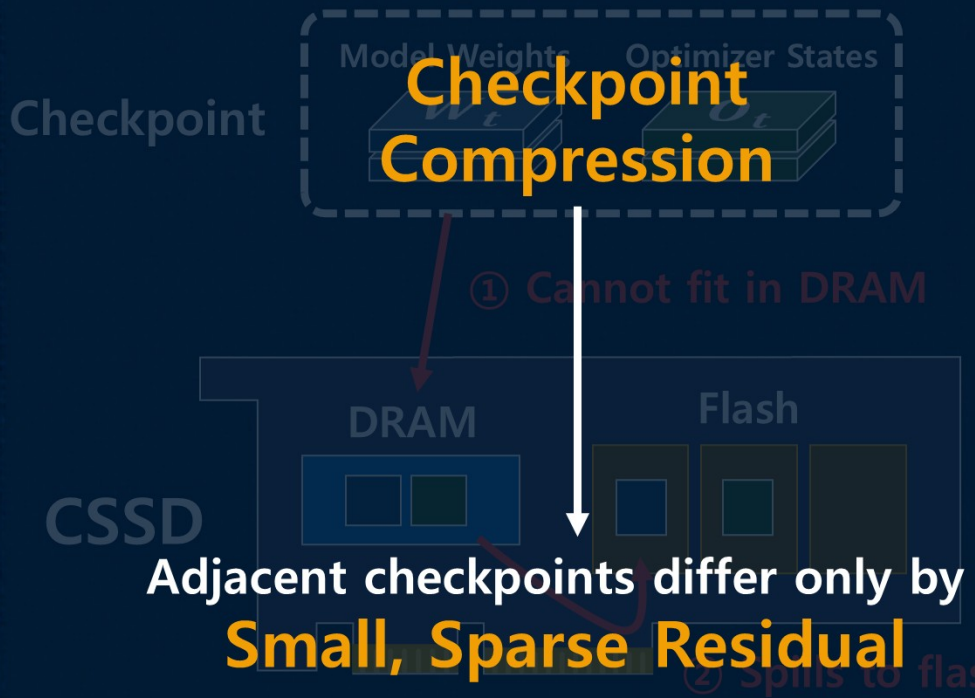
Internal DRAM As Commit Point



We set the internal DRAM as a commit point to allow early return.

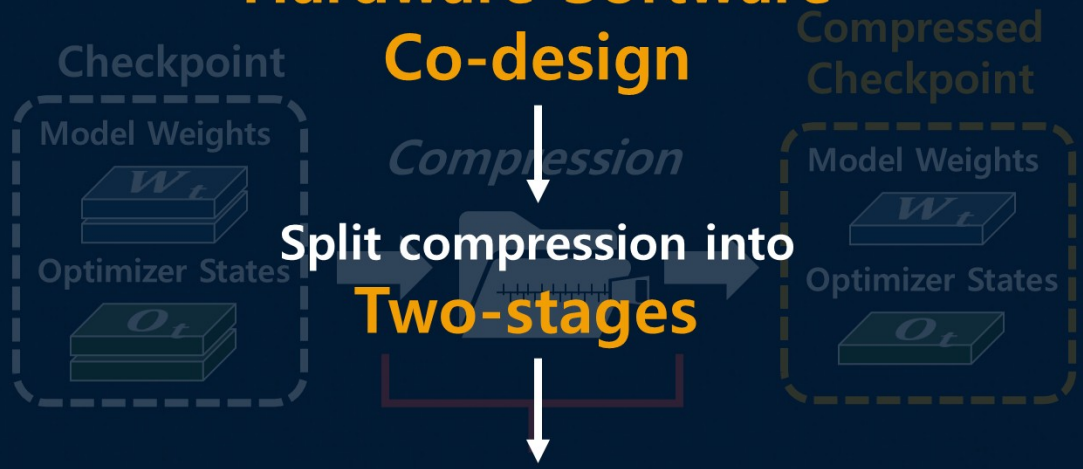
CAMEL

Sub-challenges



Adjacent checkpoints differ only by a
Small, Sparse Residual

Hardware-Software Co-design



Split compression into
Two-stages

Accelerate each stages with
Dedicated Hardware

We propose **HW/SW co-design** to effectively hide the flash latency.

CAMEL

1. Checkpointing in Large Models

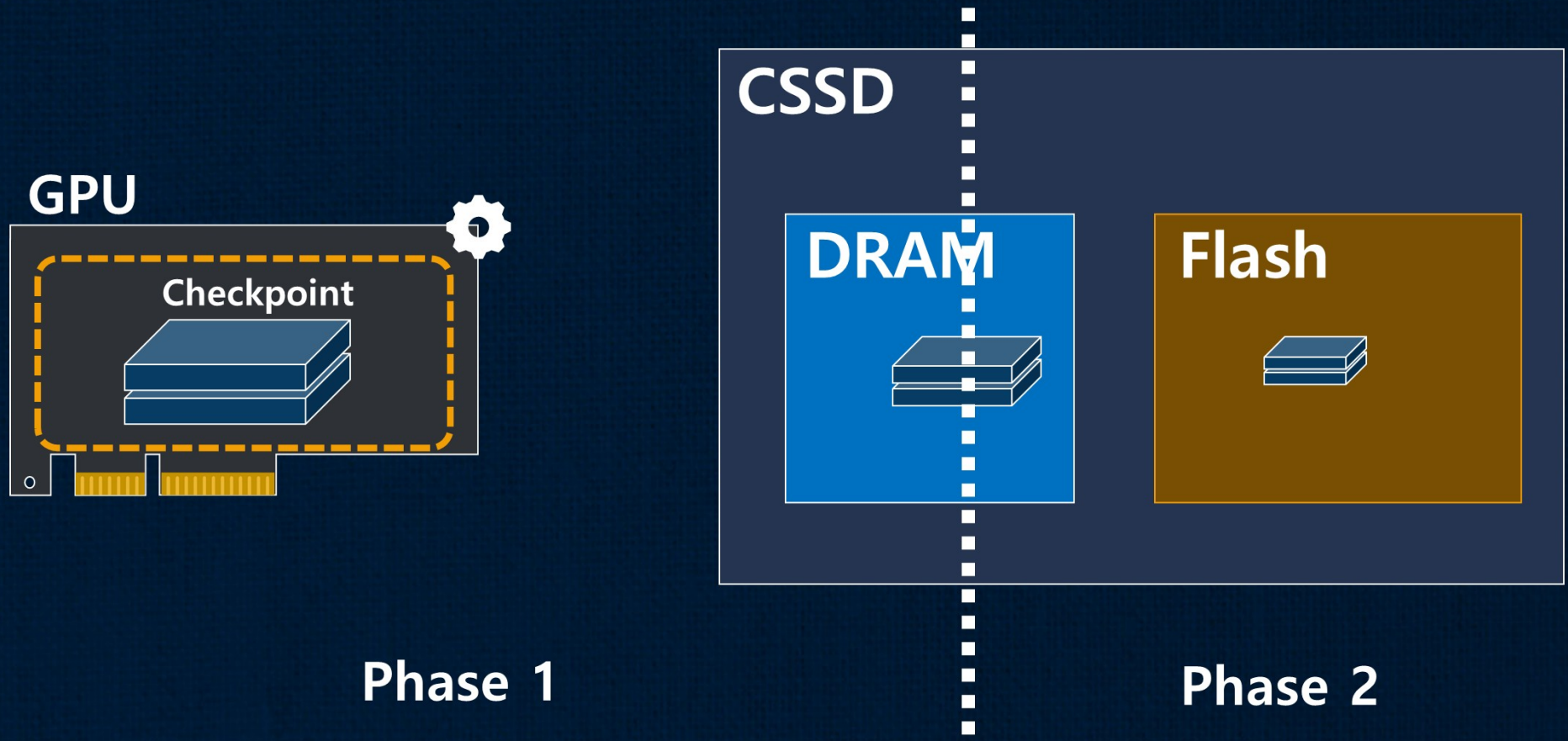
2. Video LLM Training

3. Computational SSD for Host-Free Checkpointing

4. Design and Implementation

5. Evaluation Results

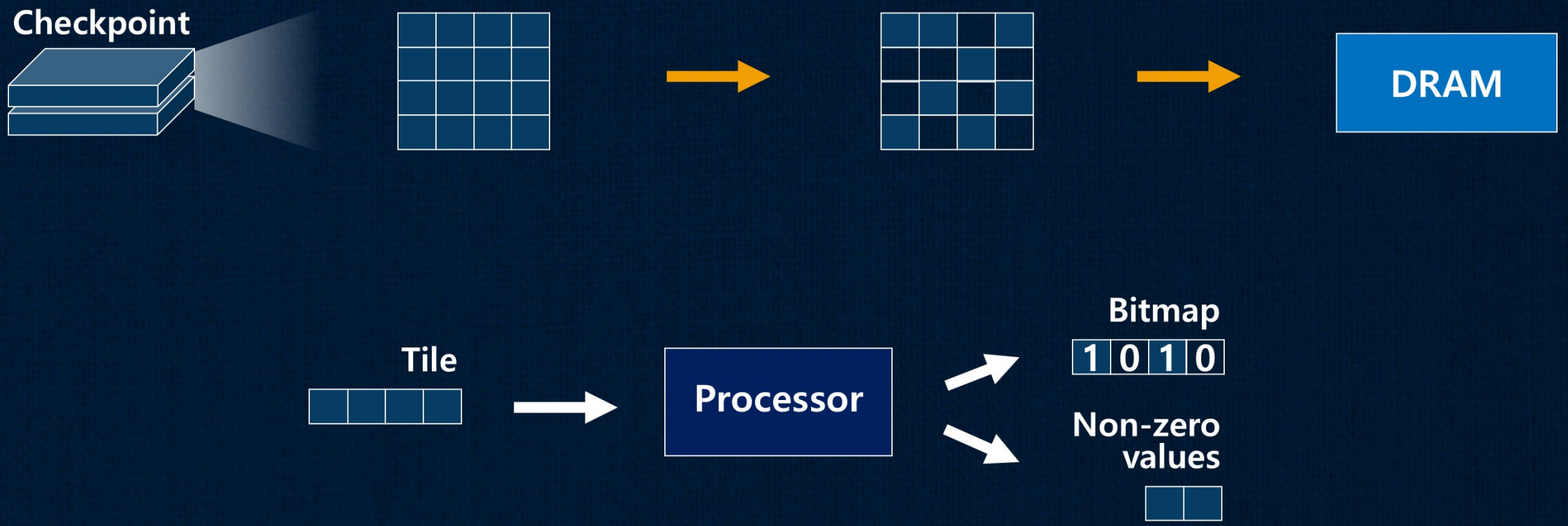
Two-Phase Checkpointing - Overview



We propose **two-phase compression** matching storage hierarchy.

CAMEL

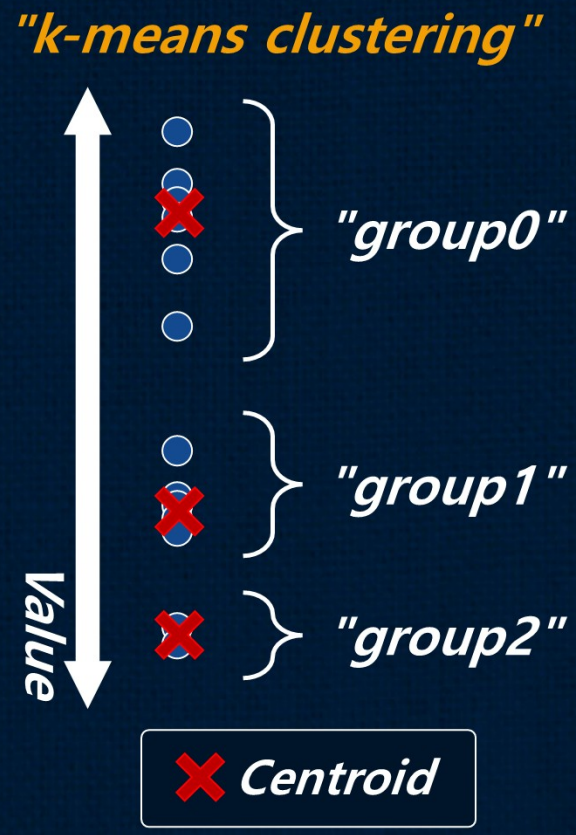
Two-Phase Checkpointing - Phase 1



Lightweight compression via **pruning**.

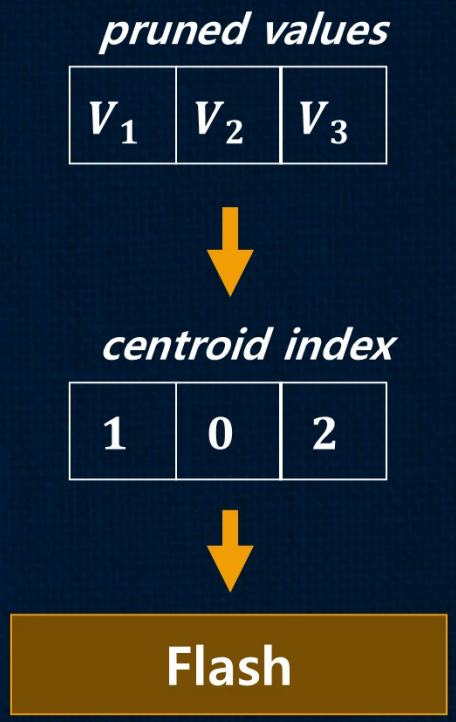
CAMEL

Two-Phase Checkpointing - Phase 2



"Codebook"

RegFile
$Centriod_0$
$Centroid_1$
$Centroid_2$
...

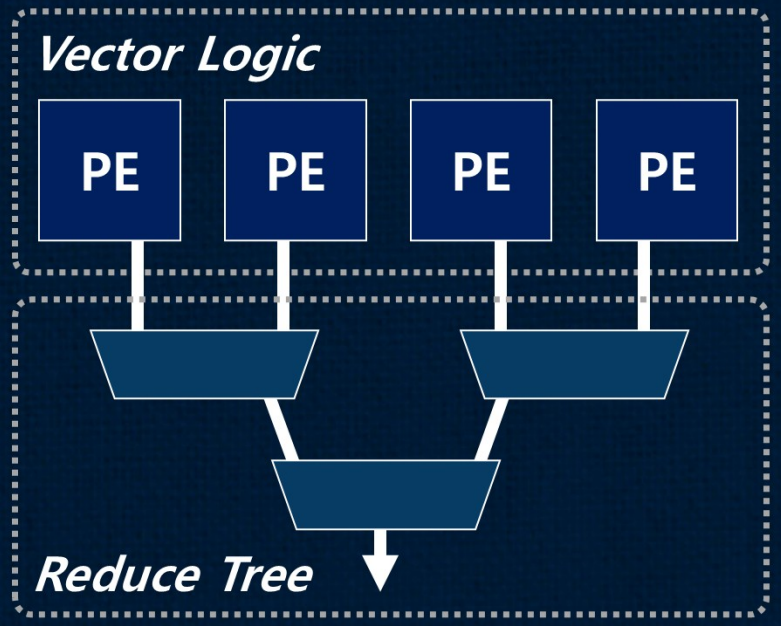


Compute-heavy compression via **quantization**.

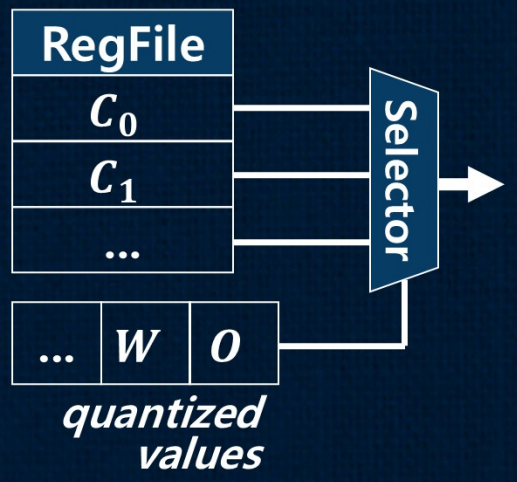
CAMEL

Specialized Accelerators

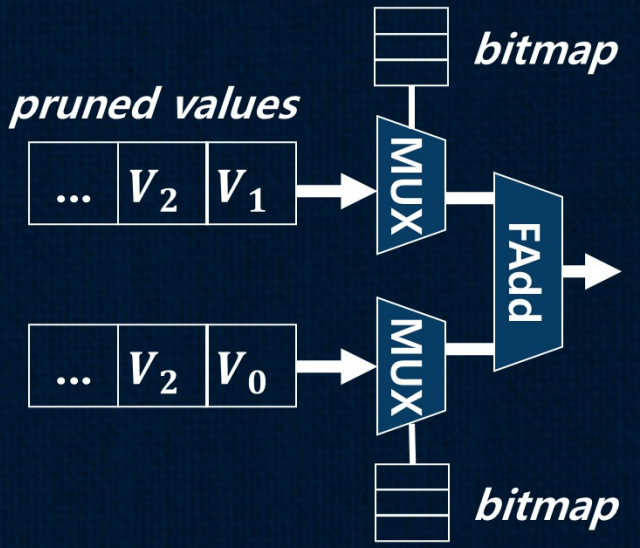
Reduce Tree



Dequantizer



Sparse Adder



Three accelerator modules support fast compression.

CAMEL

1. Checkpointing in Large Models

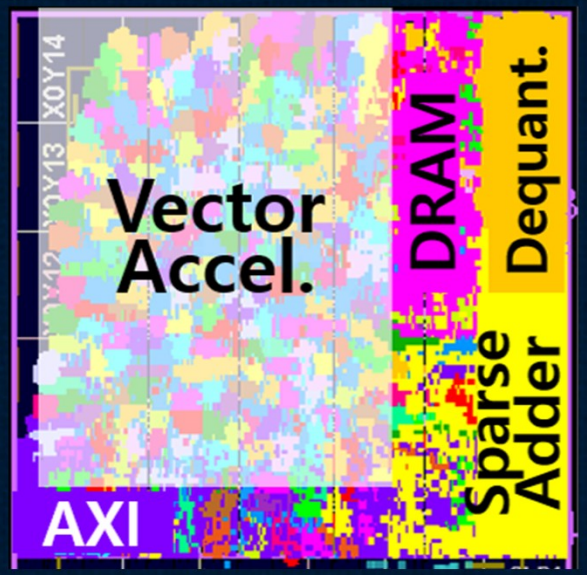
2. Video LLM Training

3. Computational SSD for Host-Free Checkpointing

4. Design and Implementation

5. Evaluation Results

Evaluation Setup

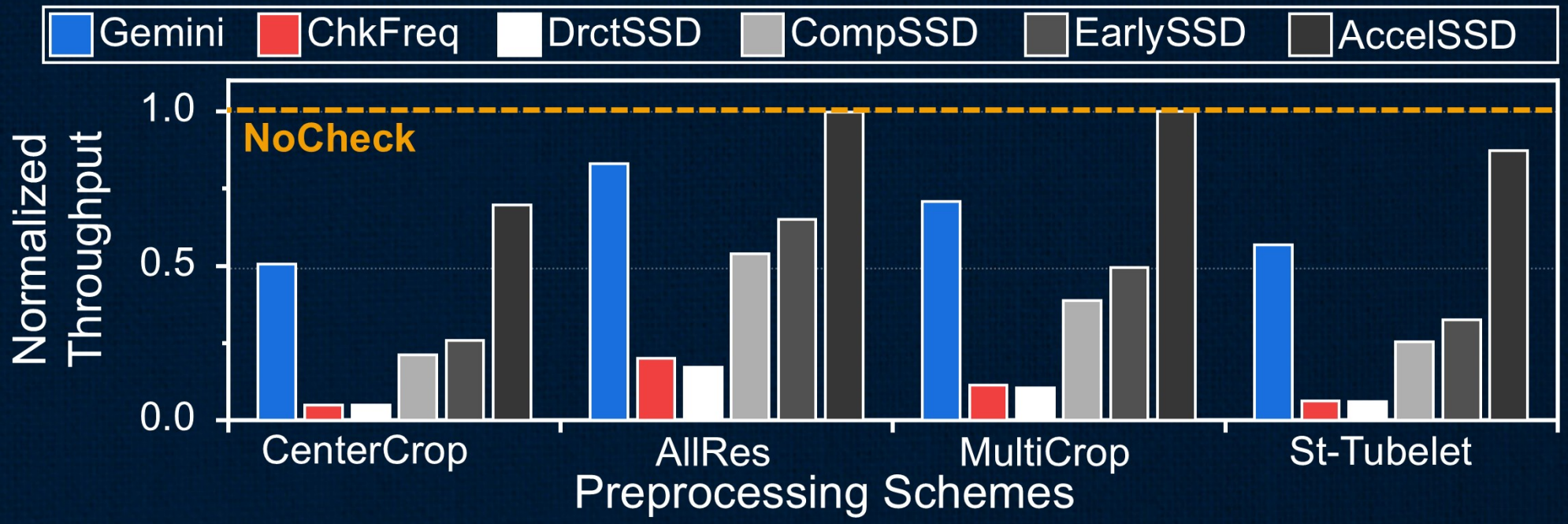


FPGA Floorplan

Component	Specification
Cluster	16 nodes
GPU	8× NVIDIA H100 80 GB
CPU	2× AMD EPYC 7R13
Host DRAM	2,048 GiB DRAM
Storage	8× Local NVMe 3.8 TB
Component	**Version**
Kernel	Linux v5.15
CUDA	v12.6
PyTorch	v2.12.0
FlashAttention	v2.8.3
Parallelism	8×TP + 16×ZeRO-3

CAMEL

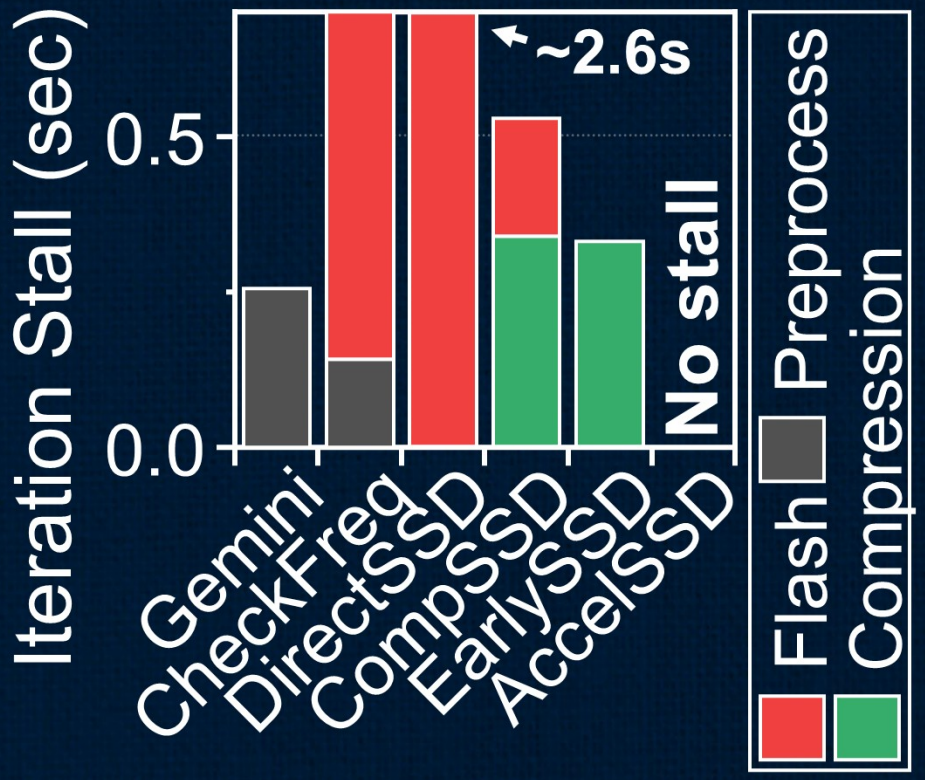
Overall Performance



FlashStash can provide 89.3% of ideal throughput on average.

CAMEL

Stall Time Breakdown



FlashStash advantages from host-free, two-phased compression with specialized HW.

CAMEL

Summary

- We identified that **the host is not idle** in multimodal LLM training.
- The host becomes the new bottleneck, leading to performance degradation.

- We proposed **FlashStash**, a computation SSD-based solution that takes the host off the datapath entirely.

- FlashStash **directly** commits from the GPU to the SSD.
- FlashStash hides the flash write latency through **two-phase checkpointing** and **in-storage compression**.
- FlashStash adopts **specialized hardware** logic to accelerate compression.

- We achieved **89.3% of ideal throughput** on average, even on iterative checkpointing.

Thank You



This work was supported by
Samsung Research Funding & Incubation Center for Future Technology of Samsung Electronics.
(Project Number: SRFC-IT2302-05)

KAIST CAMEL