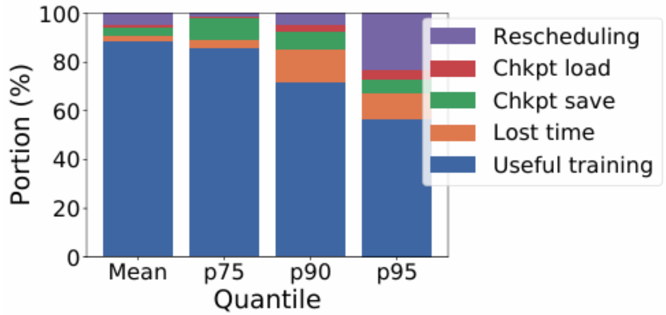
FADU

Sustaining High-Speed LLM Checkpointing with FDP

Heechul (Fletcher) Chae,
Product Planning Lead, FADU

Motivation

- **As LLM model size grow, checkpoints impact overall training time.**
  - Checkpoint operations typically consume around 12% of training time, with worst-case scenarios reaching up to 43%.



Model	parameters	Checkpoint Size (approx.)
GPT	**1T**	13.8TB
PaLM	**540B**	6TB
LLaMA	**544B**	7TB

- **High-speed storage could close the gap.**
  - In **DGX H100** node,

$$4 \times \text{Gen5 NVMe SSD (10 GB/s each)} = \mathbf{40}\ \text{GB/s} \quad \text{write bandwidth}$$

  - If training a GPT model with a checkpoint size of 13.8TB using 8 DGX nodes:

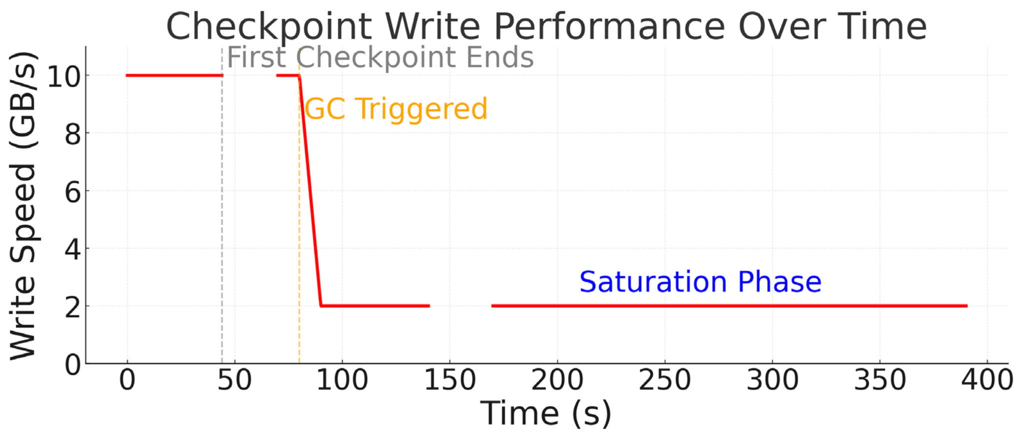
$$\frac{13.8\ \text{TB}}{8 \times 40\ \text{GB/s}} = \mathbf{43.1}\ \text{seconds} \quad \Rightarrow \text{Checkpoint duration} \approx 44\,\text{s}$$

- ***High-speed writes are hard to maintain when mixed data lifecycles collide:***
  - A Portion of training dataset or metadata may be co-located with checkpoints → potential GC trigger
  - Checkpoints with different lifecycles (e.g., latest, best, manual) may be mixed → future fragmentation risk.

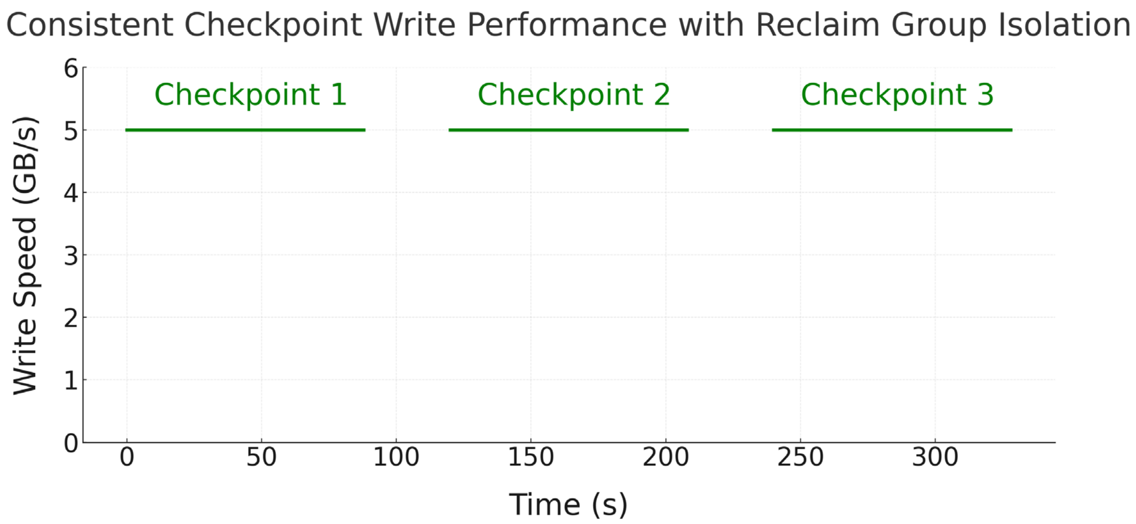
Challenges in Maintaining Consistent Write Performance

- **Problems**
  - Mixed data types(Partial dataset, Metadata, checkpoint data) are physically co-located on SSD
  - Garbage Collection (GC) is triggered more frequently, it leads to **inconsistent write throughput**
  - ***Not only does it slow checkpoint speed, but it also leads to unpredictable training latency and poor QoS.***
- **Solutions with FDP reclaim group**
  - 2 of 4 reclaim groups were dedicated to checkpointing, while random writes ran on the remaining 2 without impacting performance.
  - As a result, we observed that ***checkpoint performance remained consistent, thanks to effective isolation by the reclaim groups.***

AS-IS



TO-BE

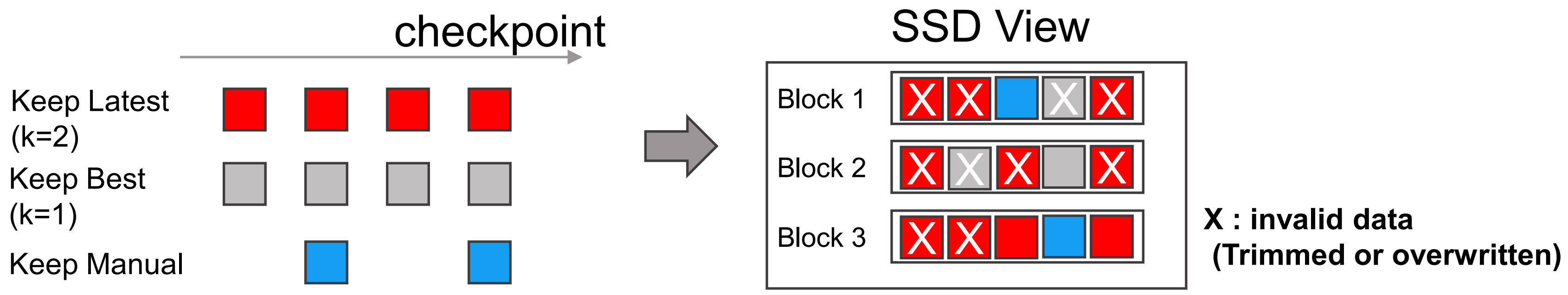


Not All Checkpoints Live the Same

- **Checkpoint lifecycles differ by training configuration.**
  - Some are short-lived (e.g., rotated, overwritten)
  - Others persist long-term (e.g., best checkpoints, manual saves)

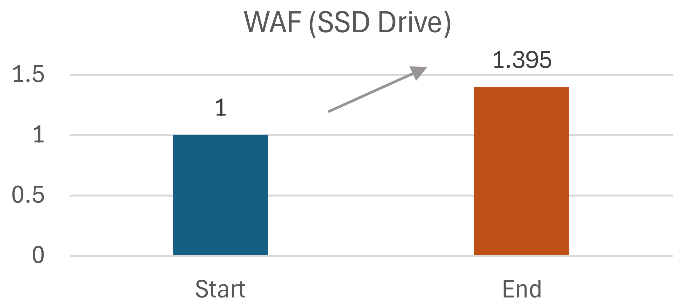
Type	Purpose	Characteristics	Example Use Cases
Keep Latest	Resume training, crash recovery	**Short-lived**, frequently overwritten	Large-scale pretraining, unstable infra
Keep Best	Deployment, analysis, fine-tuning	**Relatively Long-lived**, metric-driven	Model serving, performance tracking
Keep Manual	Milestone checkpoints, versioning	**Long-lived**, irregular saving	Experiment logging, manual inspection

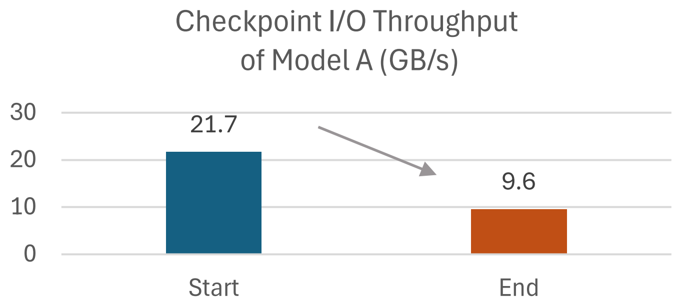
- **Simultaneous checkpointing from multiple models can mix short- and long-lived data, leading to fragmentation and higher GC pressure.**



Reproducing GC with Mixed Checkpoints

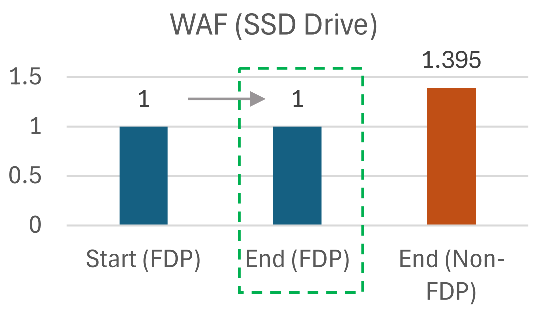
- **Demonstrate how mixed checkpoint types can trigger GC using the DLIO benchmark.**
  - DLIO do not support save policies (save_last, save_top_k, …), so mimicked checkpoint types via save paths and other options.
- **Test Configuration**
  - Model for checkpoint: llama_7b_zero3
  - Emulate three concurrent models writing checkpoints (for 4 hours)
    - Model A (**Keep Latest**): Continuously checkpoints to the **same folder**.
    - Model B (**Keep Best**): Checkpoints to the **same folder**, keeping **only one file** at a time.
    - Model C (**Keep Manual**): Saves each checkpoint to a **new folder** without deletion (**one file per folder**).
  - All models write to the same RAID0 volume (4 × Gen5 FADU NVMe SSDs , ~40 GB/s peak)
- **Test Results (Non-FDP)**
  - WAF rose from 1.0 to 1.395, indicating GC activity. (WAF measured after 4 hours checkpointing)
  - Checkpoint write speed dropped sharply once GC began.

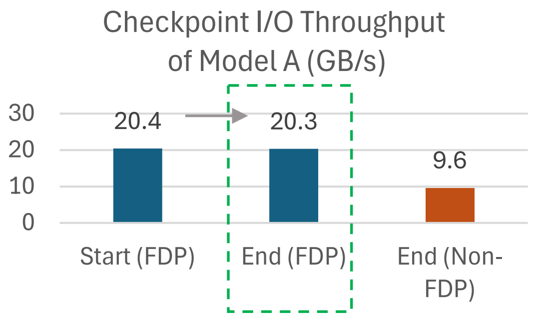




Mitigating GC & Performance w/ FDP

- **Different checkpoint types were isolated using FDP RUH assignments.**
- **Test Configuration (Drive setup for FDP)**
  - DLIO does not support FDP, so it cannot assign a placement ID (Reclaim Unit Handle) to checkpoints.
  - Created **3 namespaces per drive** and assigned **one RUH to each namespace**.
    - Configured RAID0 using the same namespace index across all drives.
  - Performed checkpointing of **three different models on each namespace**.
    - Model A (**Keep Latest**) to NS1 (5% of drive capacity)
    - Model B (**Keep Best**) to NS2 (10% of drive capacity)
    - Model C (**Keep Manual**) to NS3 (85% of drive capacity)
- **Test Results (FDP)**
  - **No GC occurred**, and **no performance drop** was observed. (during the 4-hour test.)





FADU

The SSD Expert