
KubeCon
CloudNativeCon
Europe 2025
LONDON BUS

MUNI Masaryk University

TÉCNICO LISBOA

KubeCon | CloudNativeCon Europe 2025

Efficient Transparent Checkpointing of AI / ML Workloads in Kubernetes

Viktória Spišaková, Radostin Stoyanov, Adrian Reber

Supervisor: Prof. Rodrigo Bruno, Prof. Wes Armour

Red Hat

2022

Improving GPU Utilization using Kubernetes - Maulin Patel & Pradeep Venkatachalam, Google

PAVILION 4, ROOM B | LEVEL 2 | CENTRAL FORUM

2023

Efficient Access to Shared GPU Resources: Mechanisms and Use Cases - Diogo Filipe Tomas Guerra & Diana Gaponcic, CERN

G104-105 | FIRST FLOOR | CONGRESS CENTRE

Increasing GPU Utilisation on K8s Clusters Dedicated for AI/ML Workloads - Maciej Mazur, Canonical / Ubuntu & Andreea Munteanu, Canonical

PAVILION 7 | LEVEL 7.3 | PARIS ROOM

2024

Building Resilience for Large-Scale AI Training: GPU Management, Failure Detection, and Beyond - Ganeshkumar Ashokavardhanan, Microsoft & Ace Eldeib, Cohere

SALT PALACE | LEVEL 1 | 155 E

Maximizing GPU Utilization Over Multi-Cluster: Challenges and Solutions for Cloud-Native AI Platform - William Wang & Hongcai Ren, Huawei

PAVILION 7 | LEVEL 7.3 | PARIS ROOM

Enabling Fault Tolerance for GPU Accelerated AI Workloads in Kubernetes - Arpit Singh & Abhijit Paithankar, NVIDIA

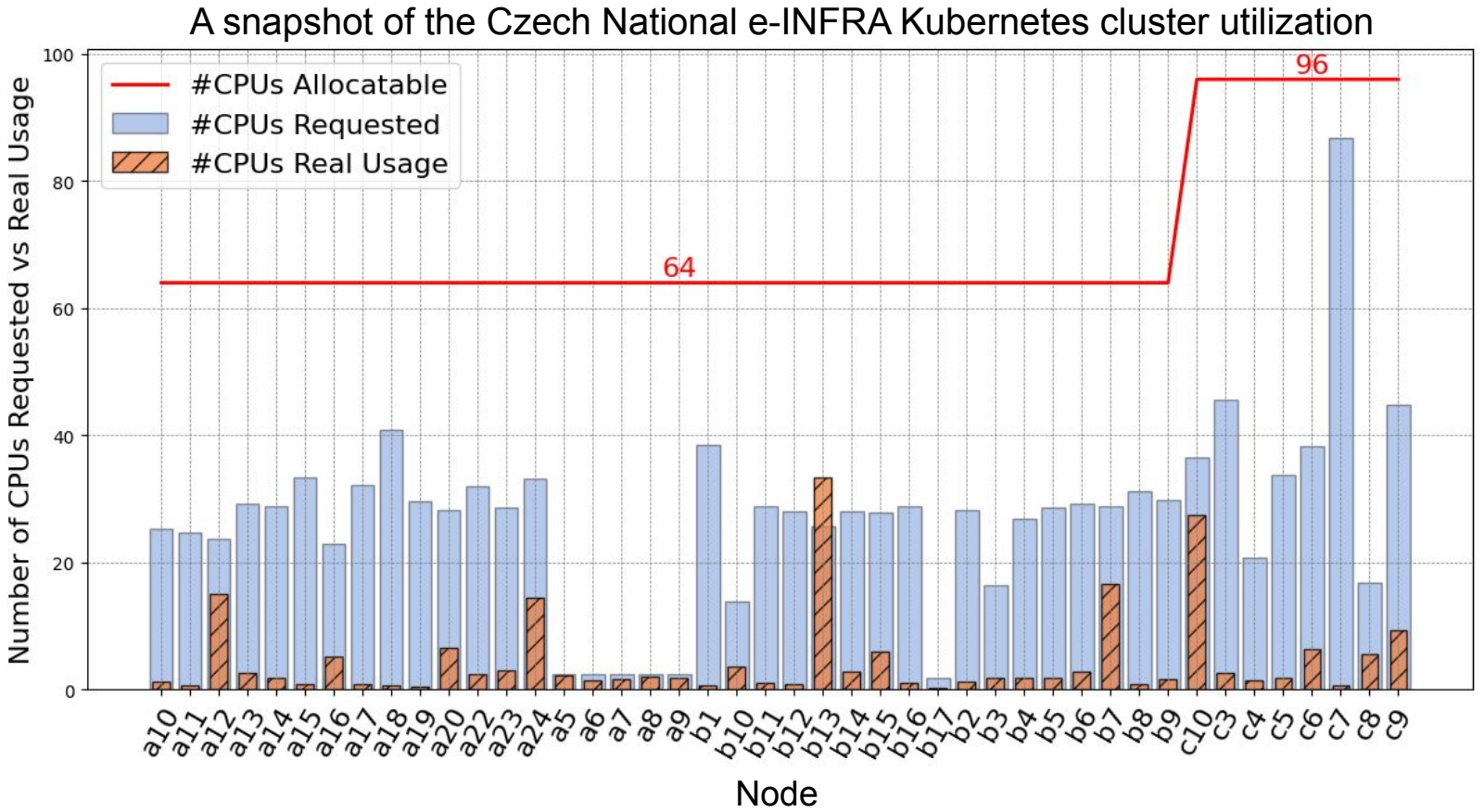
SALT PALACE | LEVEL 2 | 255 E

Real-World Motivations

Managed K8s clusters at e-INFRA CZ

- Czech national e-infrastructure for academics and researchers providing multi-tenant, multi-purpose K8s clusters
  - ~4000 CPU cores, 33 TiB RAM, ~50 GPUs (A10-H100)
  - ~300 active users operating multitude of workload types
    - **Batch workloads**
    - **Interactive workloads**
- Many users not that proficient with the cluster
  - Strive for simplicity
  - No complicated setups or too specific configurations

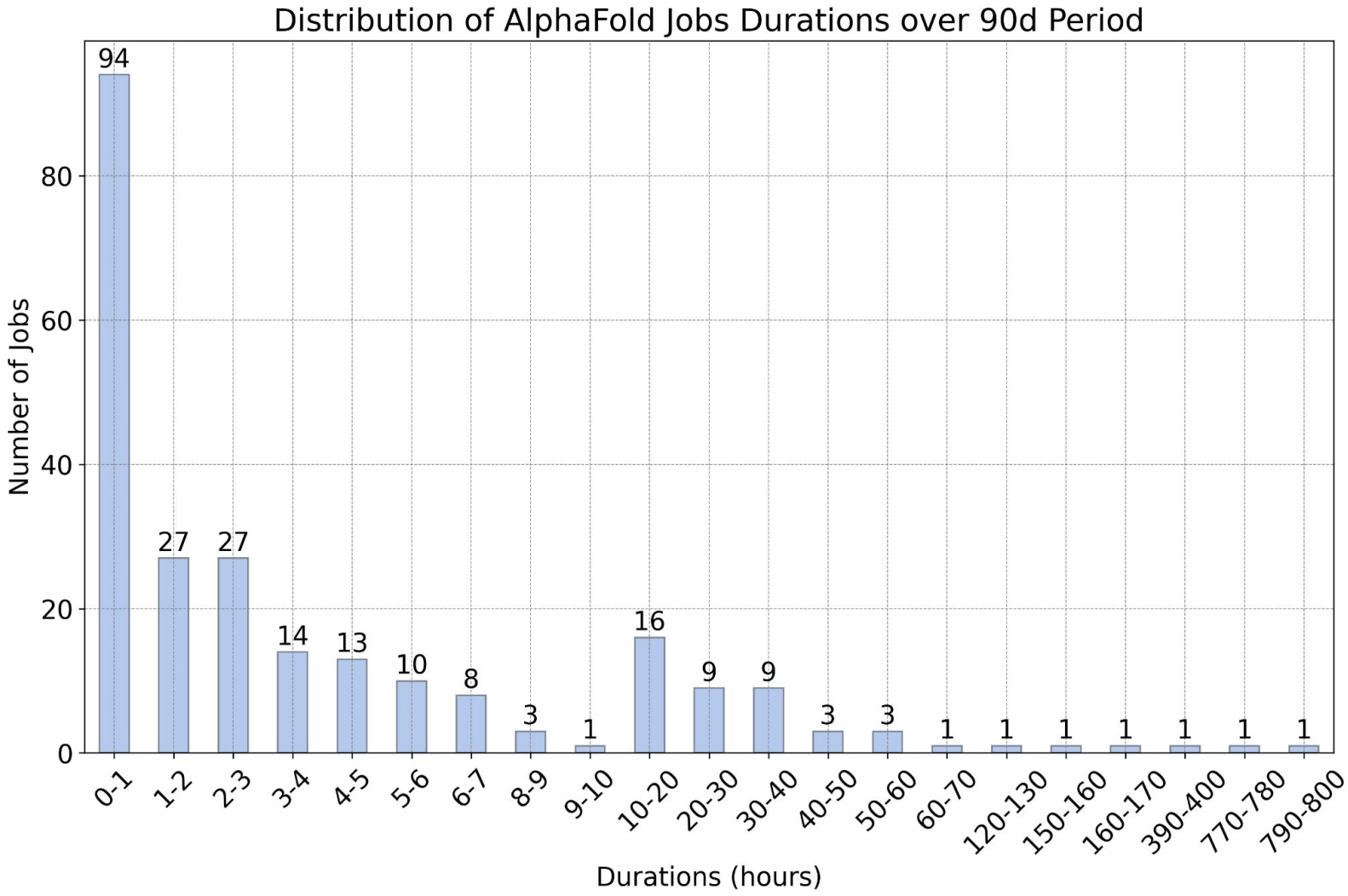
Low Cluster Utilization



Overprovisioning is necessary

- Low resource utilization
- Redundancy & elasticity
- Provision for sudden spikes

Batch Jobs – ML Training, HPC Workloads

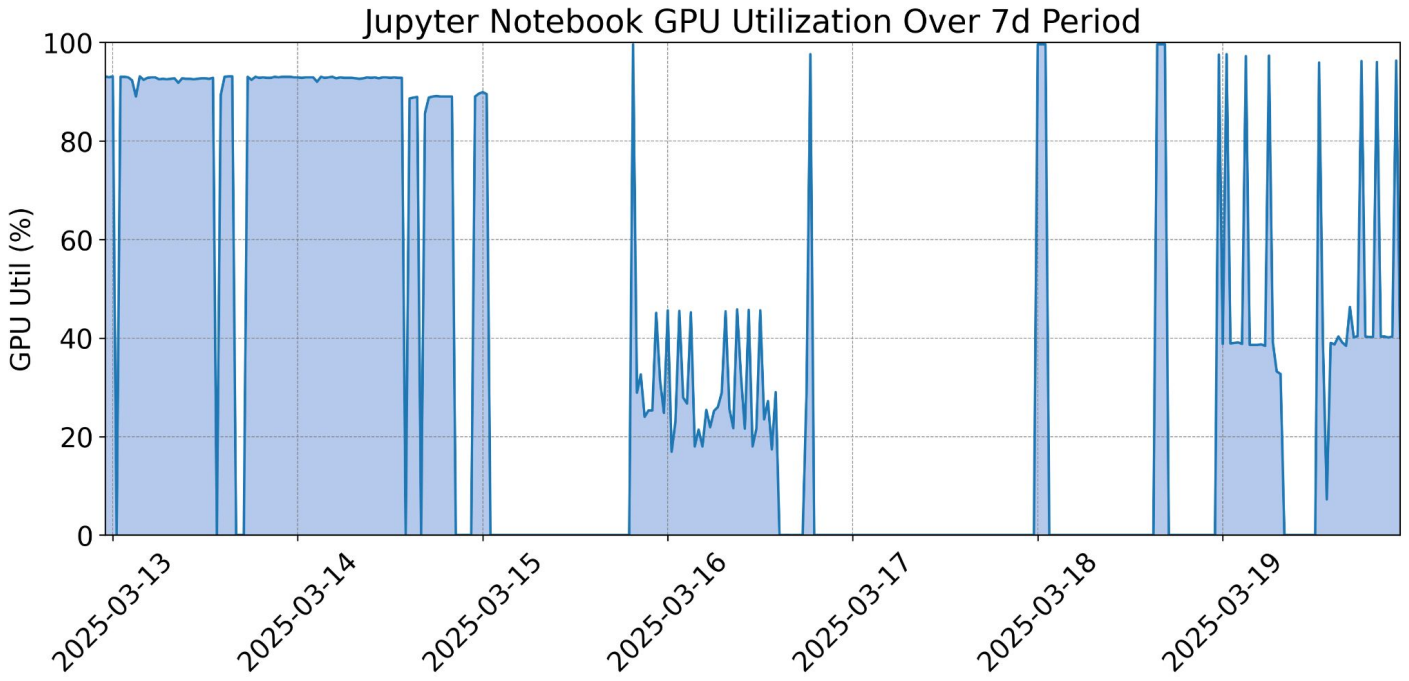


Long-running workloads

- Resource intensive
- Accuracy is important
- Require **fault-tolerance**

Jobs	Avg GPU Util Last Day	Running
Job 1	**91.91%**	2.57 days
Job 2	**85.92%**	2.56 days
Job 3	**98.45%**	1.16 days

Interactive Workloads
(Chatbots, Jupyter Notebooks)



Dynamic and unpredictable workloads

- Services must be always available
  - Cold-start problem affects availability
- Humans are unpredictable
- Low GPU utilization
- Inactive sessions & bursts of activity

Wishlist for GPU Workloads

- Ensure fault-tolerance
  - Failures/errors happen all the time
  - Useful for long-running apps - trainings, HPC applications
- Achieve efficient utilization of resources
  - Useful for interactive apps - model inference, Jupyter Notebooks, visualizations, gaming, remote SW access

Existing Methods for Improving Utilization

- Mitigating over-provisioning through **auto-scaling** (vertical, horizontal, in-place)
  - Requires application awareness / developer involvement
- GPU sharing - allocating **partial** GPUs (e.g., MIG) leads to fragmentation
- Co-location of workloads (time slicing) - worse performance
- New scheduling strategies - workload-specific
- Another versatile tool into the toolbox could be useful - **GPU Checkpointing**

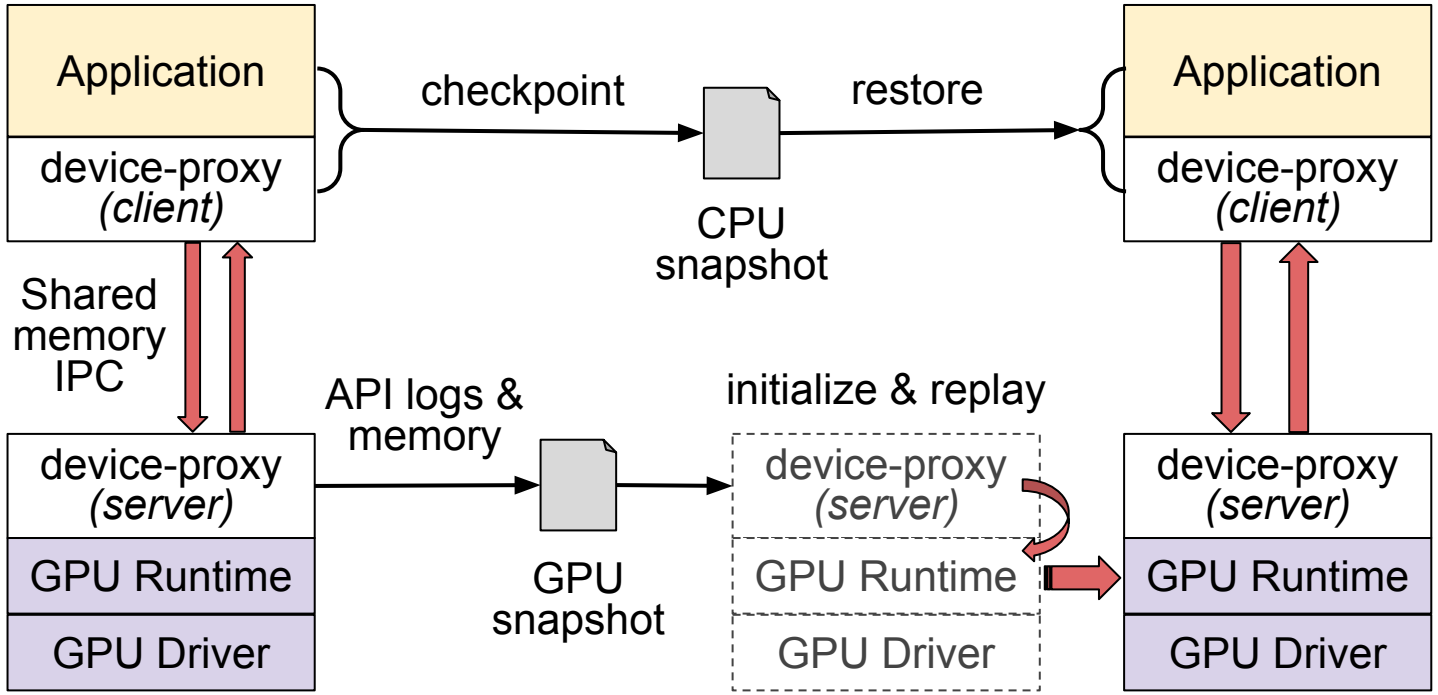
Transparent GPU Checkpointing

Overview of GPU Checkpointing Methods

Enabling Coordinated Checkpointing for Distributed HPC Applications.
Radostin Stoyanov and Adrian Reber. KubeCon Europe 2024

Transparent GPU Checkpointing

Existing Methods - API Interception



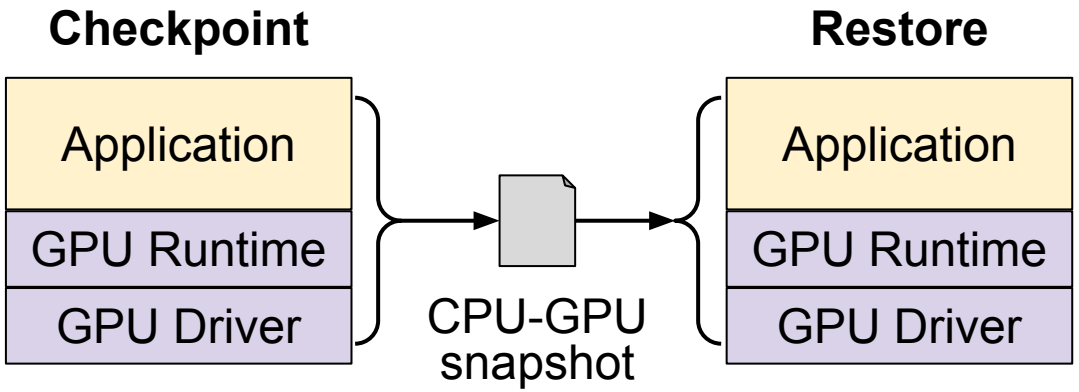
API Interception Challenges

- Difficult to implement & maintain
- Adds performance overhead
- Tracking & logging memory transfers
- GPU arch-specific implementation
- Requires dynamic linking
  - e.g., rebuilding PyTorch

Just-In-Time Checkpointing: Low Cost Error Recovery from Deep Learning Training Failures. Tanmaey Gupta, et. al., 2024

Singularity: Planet-Scale, Preemptive and Elastic Scheduling of AI Workloads. Dharma Shukla, et. al., 2022

Transparent and Unified CPU-GPU Snapshots

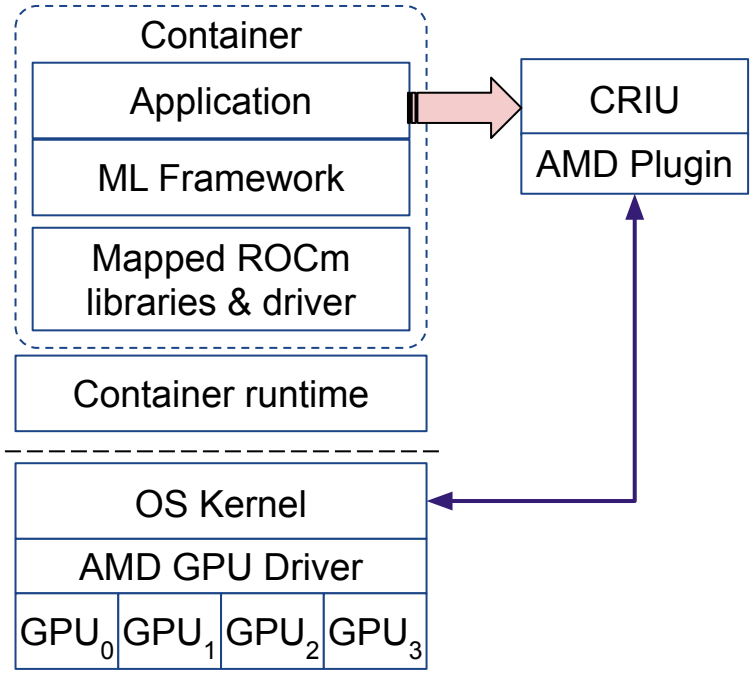


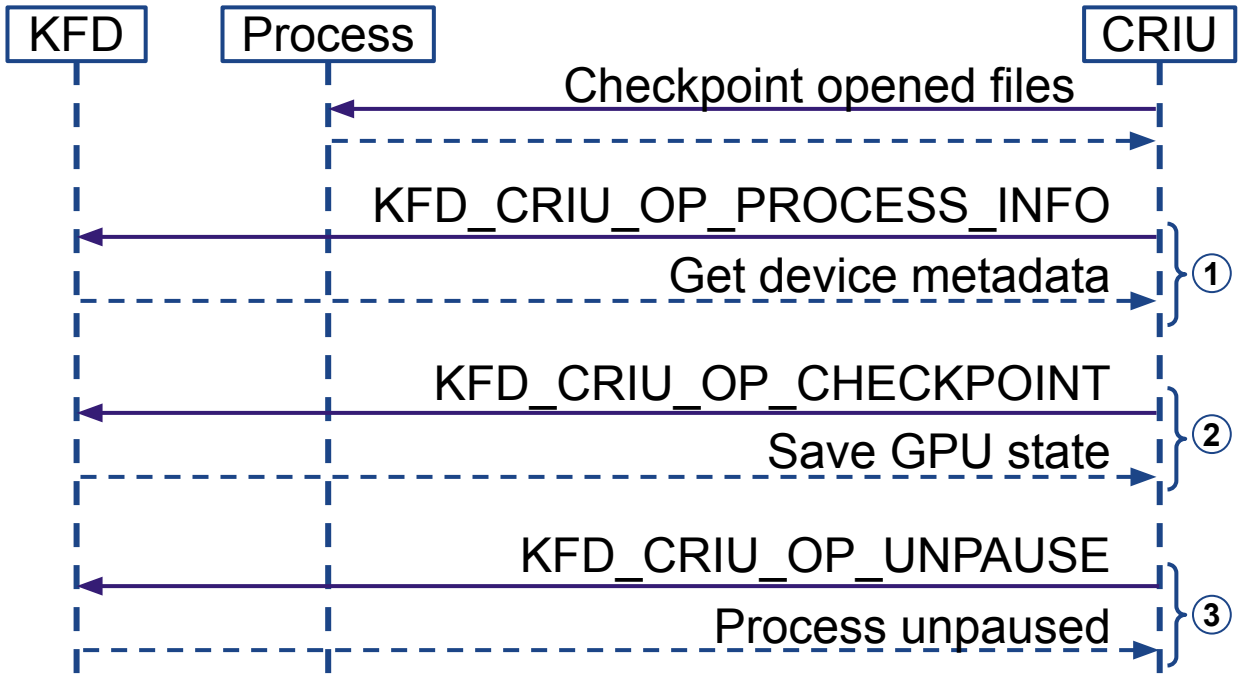
Checkpointing with CRIU + Plugins

- Fully-transparent (no API interception)
- Supports static and dynamic linking
- Out-of-the-box support for Kubernetes
- Supports **AMD** and **NVIDIA** GPUs

CRIUgpu: Transparent Checkpointing of GPU-Accelerated Workloads. Radostin Stoyanov, Viktória Spišaková, Jesús Ramos, Steven Gurfinkel, Andrei Vagin, Adrian Reber, Wesley Armour, Rodrigo Bruno. (2025). https://arxiv.org/abs/2502.16631

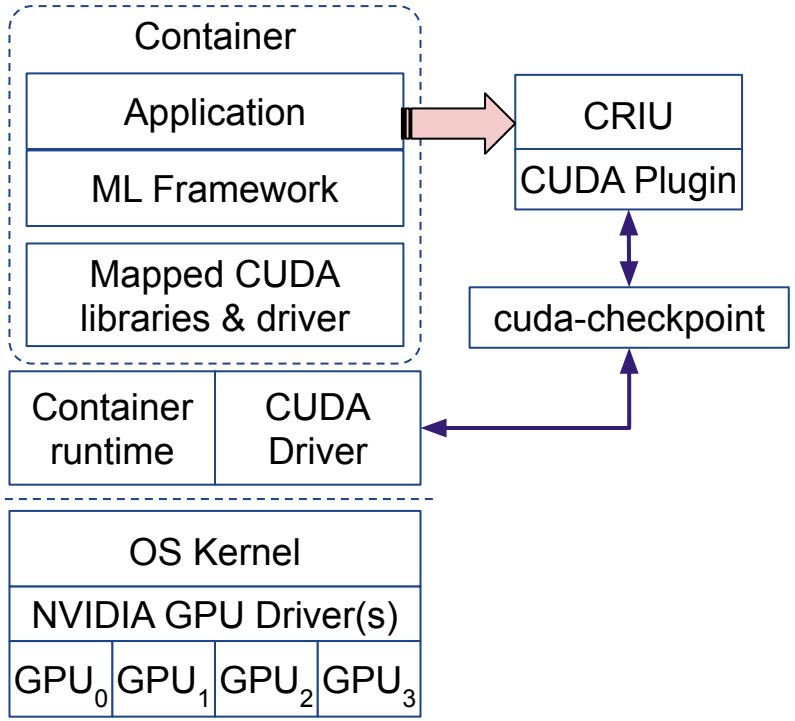
CRIU with AMD GPU Plugin

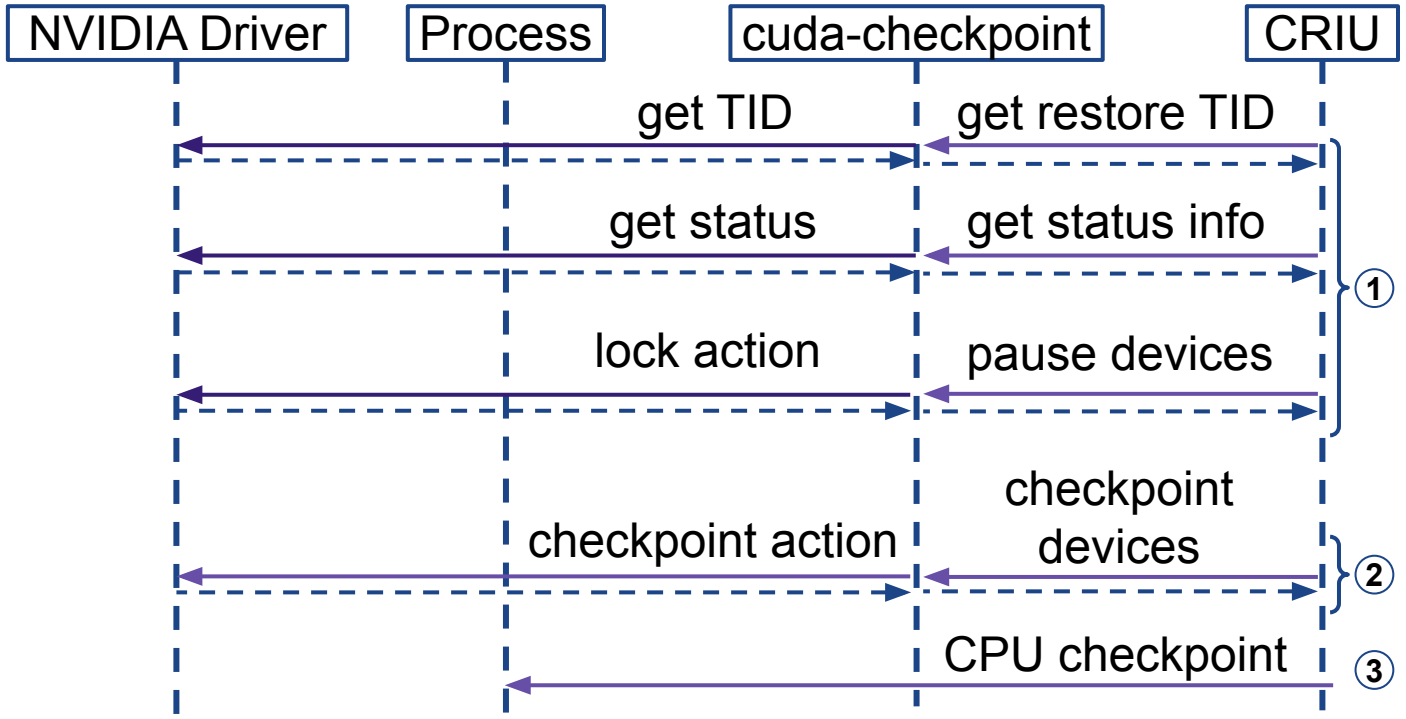




drm/amdkfd: CRIU Introduce Checkpoint-Restore APIs (Linux kernel commit: 36988070)

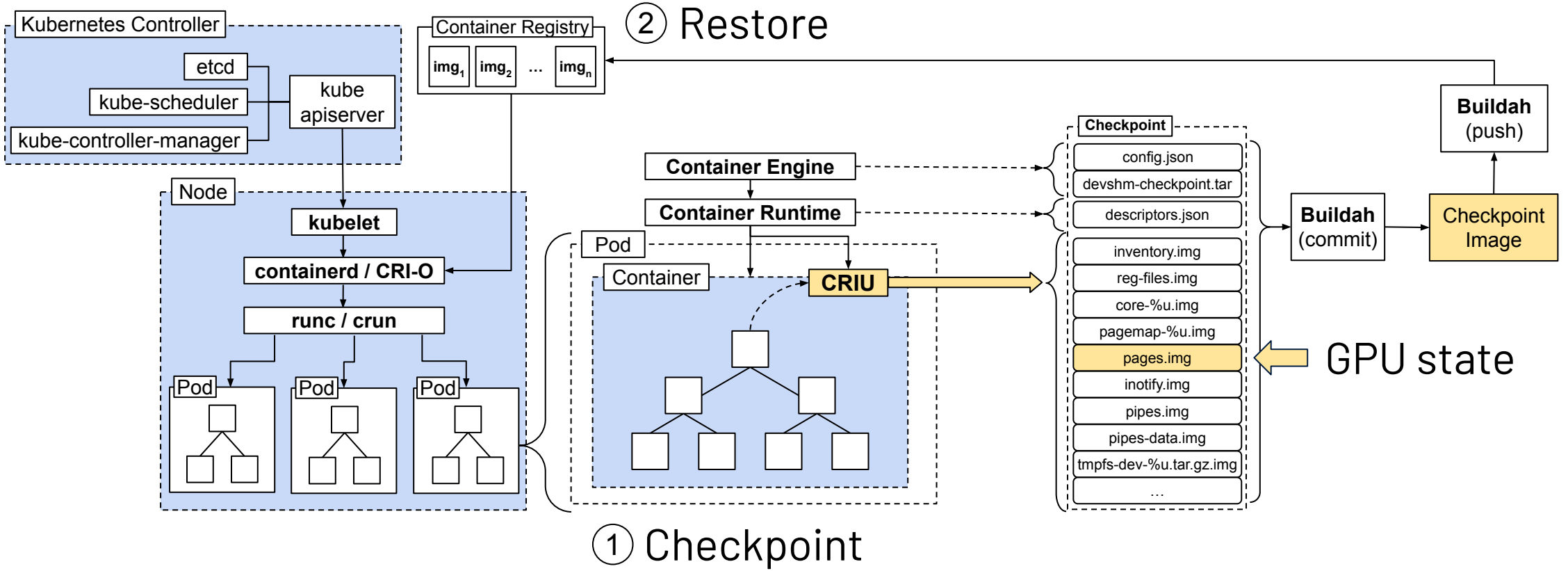
CRIU with CUDA Plugin





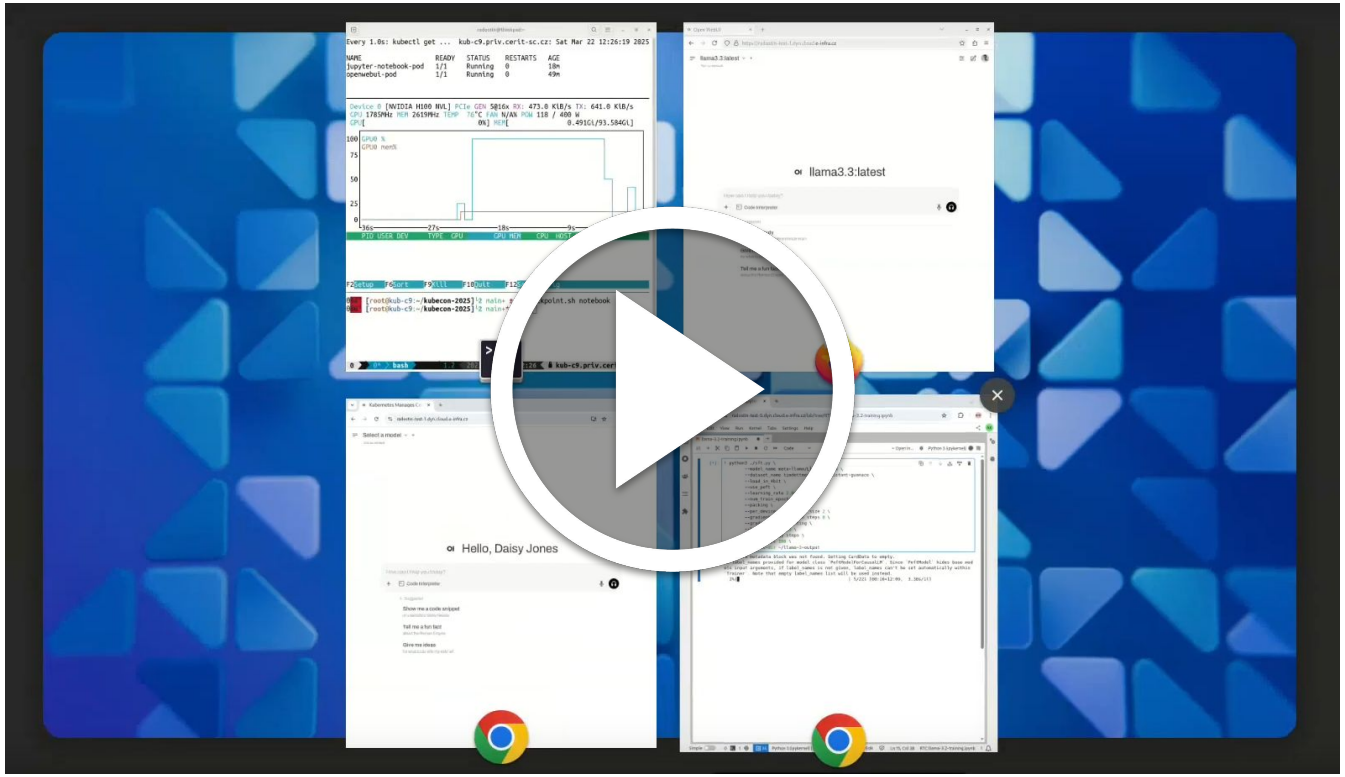
Checkpointing CUDA Applications with CRIU – Steven Gurfinkel. NVIDIA Technical Blog (2024)

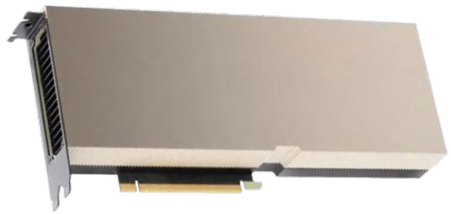
GPU Checkpointing Integration with Kubernetes



Transparent Hot-Swapping

KubeCon | CloudNativeCon Europe 2025

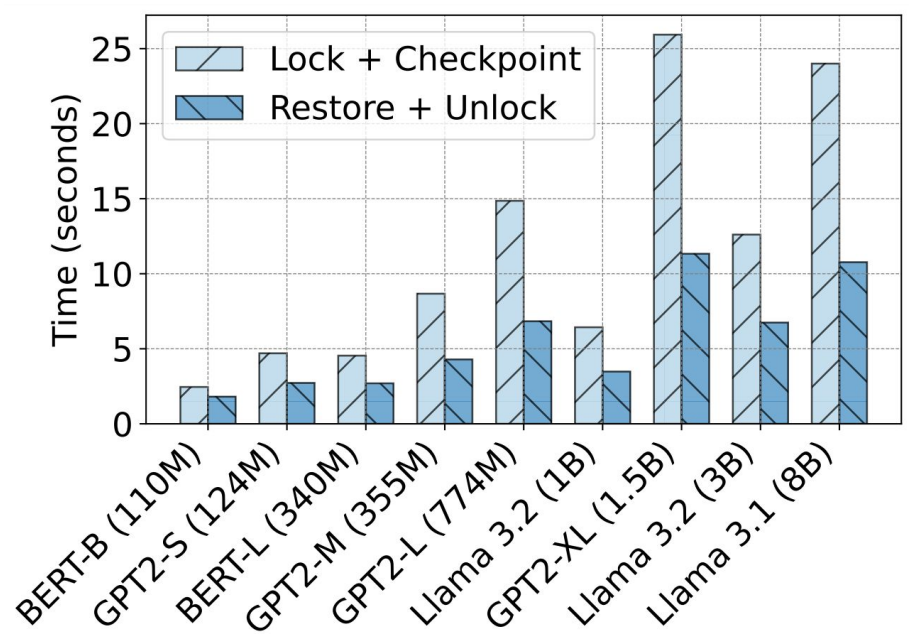




Computational resources were provided by the e-INFRA CZ project (ID:90254),
supported by the Ministry of Education, Youth and Sports of the Czech Republic.

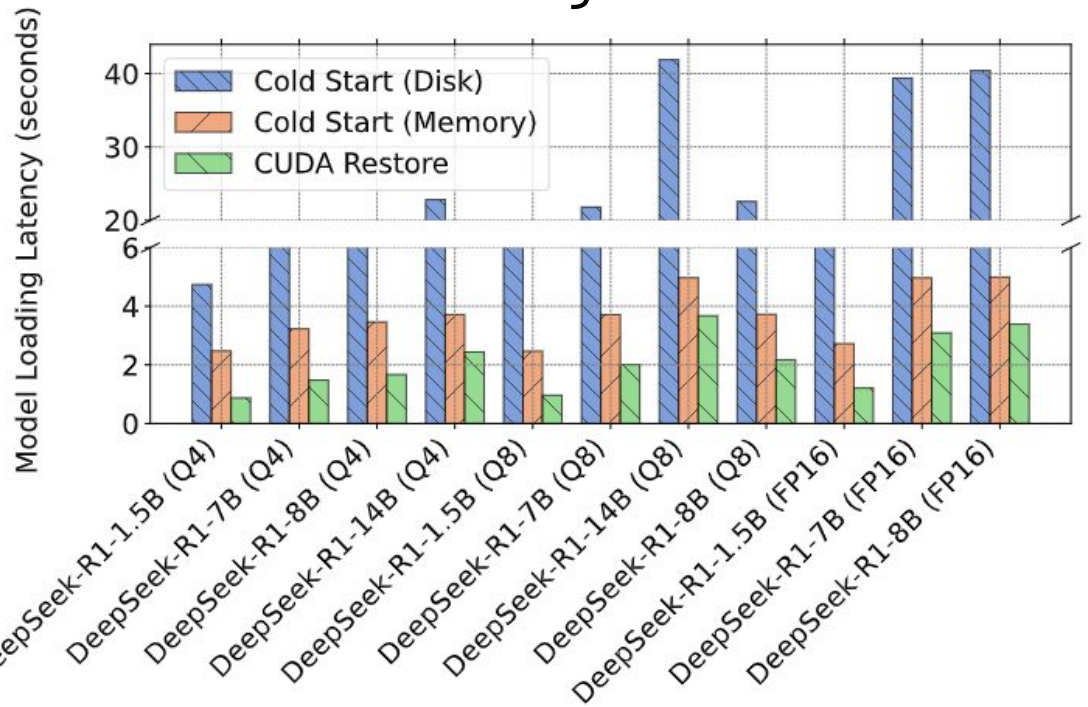
Evaluation Results

Fast Error-Recovery



In-memory GPU Checkpoints of Model Training with Hugging Face Transformers

Accelerating Cold-Starts



Ollama LLM Inference

NVIDIA A100 SXM4 80GB

State of Checkpoint/Restore in Kubernetes

2015: Kubernetes

Ticket discussing container migration

Forensic Container Checkpointing

Kubernetes 1.25 (2022)
Alpha: Forensic Container Checkpointing

Kubernetes 1.30 (2024)
Beta: Forensic Container Checkpointing

Next Steps

Declare checkpointing Stable/GA

Next Steps

`kubectl` support

Admin Panel

GRUS | Weaversoftio

crc

- Pods
- Checkpoints
- Registry
- Secrets
- Users
- Logout

Pods

Search: Name, Namespace, Node

	Name	Namespace	Nodename	No. of Container
⌄	crc-debug-2t9qg	default	crc	1
⌃	grus-6688bc5d5-jvprk	grus	crc	2

Containers

Name	Image	Action
grus-api	docker.io/frenzy669/grus-api:latest	CREATE CHECKPOINT
grus-admin	docker.io/frenzy669/grus-admin:latest	CREATE CHECKPOINT

⌄	restore-pod2	grus	crc	1
⌄	spring-music-deployment-744947bf6...	grus	crc	1
⌄	spring-music-deployment-744947bf6...	grus	crc	1

Task Progress

CLEAR

Cluster Verification

```
Cluster verification initiated with clusterName:
crc


PLAY [Verify CRIU and CRIO Configuration]
*************************************

TASK [The 'enable_criu_support = true' is not
present in '/etc/crio/crio.conf.d/05-enable-criu']
***
ok: [192.168.130.11]

TASK [The 'tcp-established' is not present in
'/etc/criu/criu.conf'] ***********
ok: [192.168.130.11]

TASK [Get 'criu' version]
*************************************************
****
ok: [192.168.130.11]

TASK [Extract 'criu' version number]
*****************************************
ok: [192.168.130.11]

TASK [The version of 'criu' is not correct]
***********************************
ok: [192.168.130.11] => {
```

Summary & Questions

- Fully-transparent GPU checkpointing
- Supports AMD and NVIDIA GPUs
- Out-of-the-box integration with Kubernetes

github.com/checkpoint-restore/criu

github.com/nvidia/cuda-checkpoint