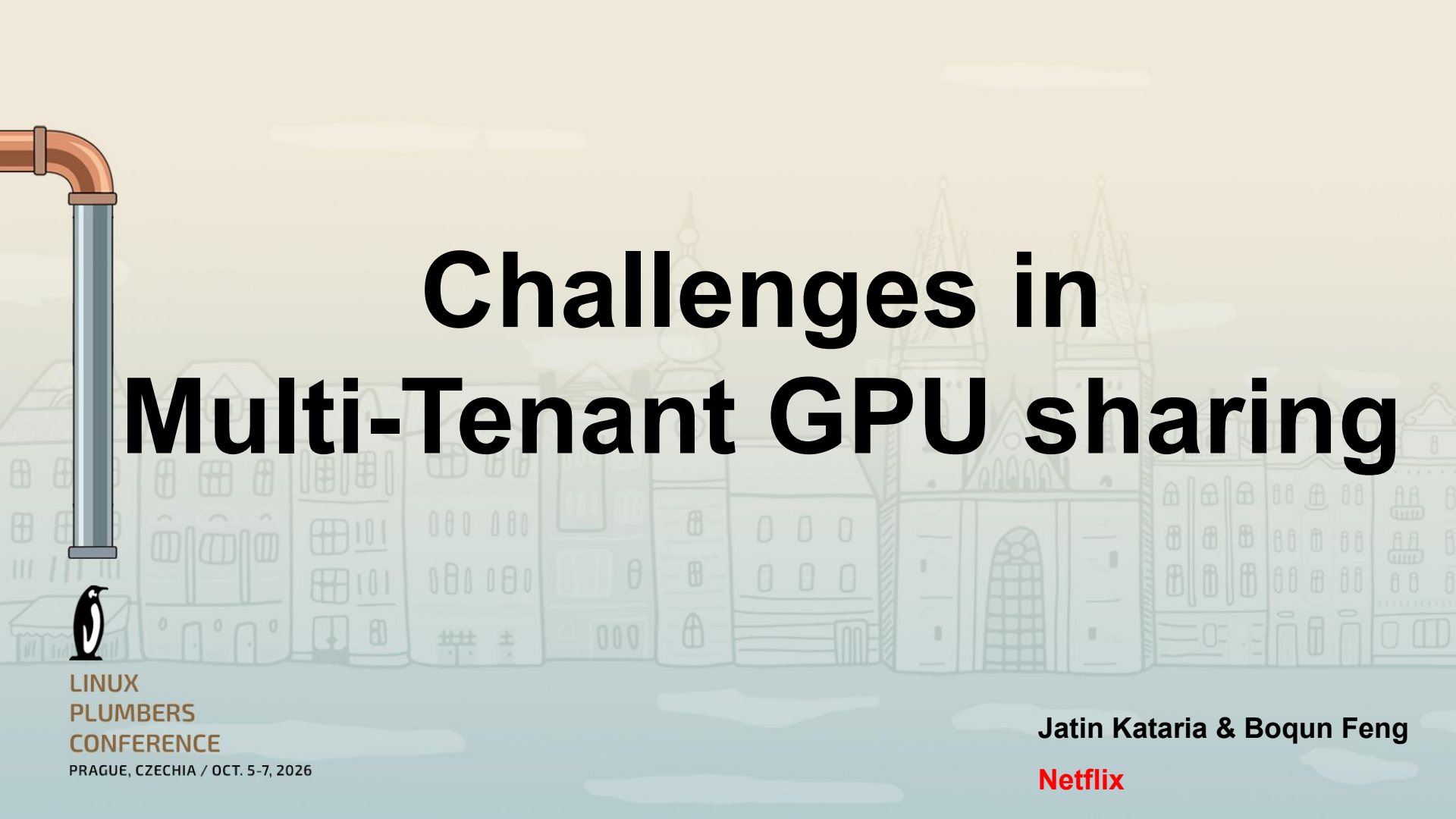




Challenges in Multi-Tenant GPU sharing



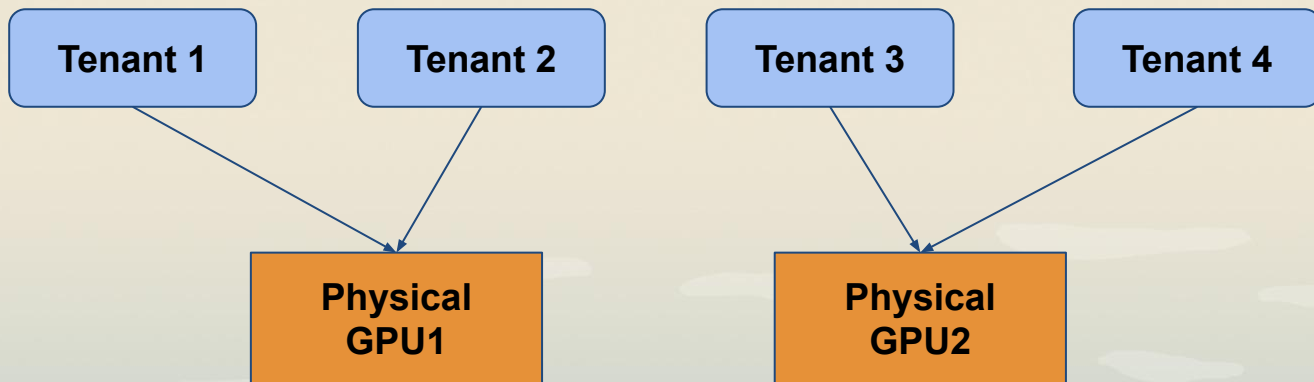
LINUX
PLUMBERS
CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

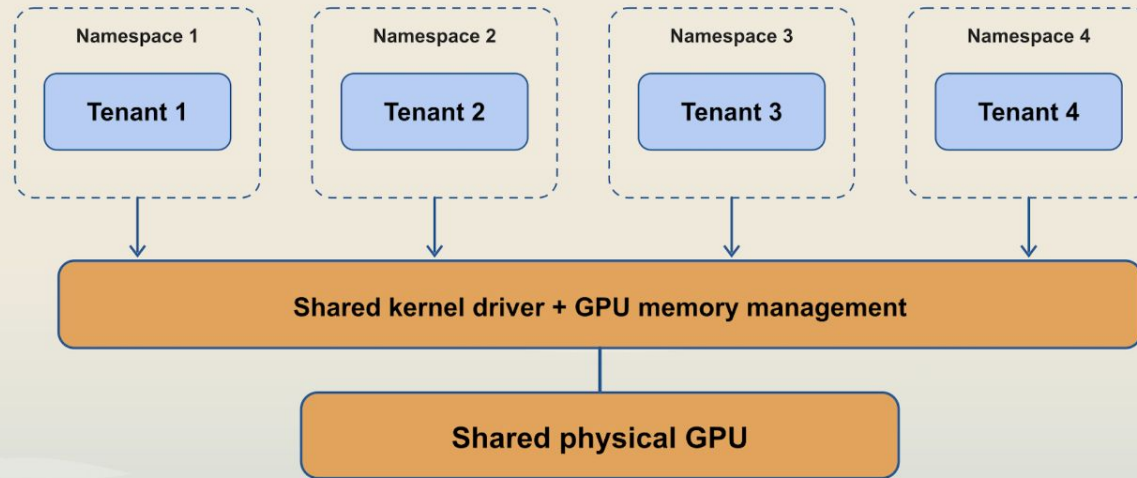
Jatin Kataria & Boqun Feng

Netflix

Context: More Tenants, Fewer GPUs



Tension: Shared GPU Stack



GPU Sharing Challenges

Isolation

Resource Accounting

Debuggability

Fault Attribution/Impact

Performance

Memory path Latency

Contention

Drive lock



LINUX
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026



GPU Sharing in Practice

Who caused GPU reset?

Fault Attribution

Can one tenant stall all GPUs?

Nvidia Driver Lock

Can one tenant bypass its resource limit?

TTM Quicksand

Why does GTT latency grow?

VRAM vs GTT



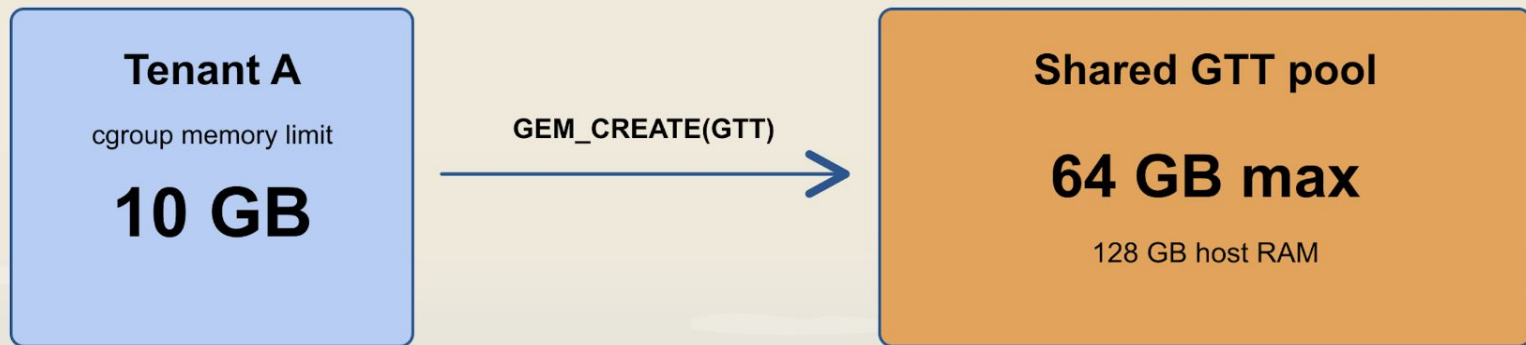
LINUX

PLUMBERS CONFERENCE

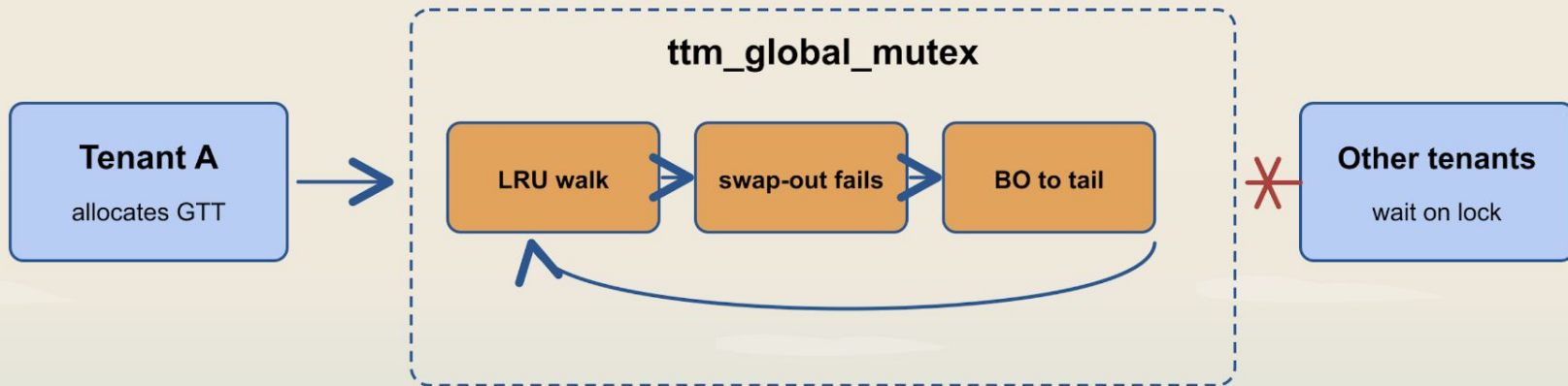
PRAGUE, CZECHIA / OCT. 5-7, 2026



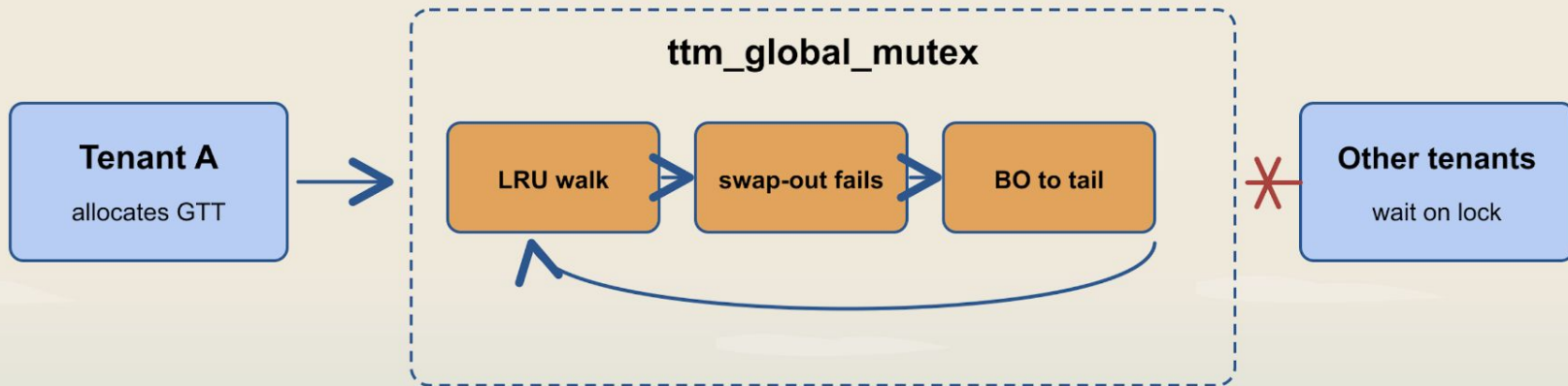
Case Study: Isolation



TTM Quicksand



TTM Quicksand : Fixed



b2ed01e7ad3d ("drm/ttm: Fix ttm_bo_swapout() infinite LRU walk on swapout failure")

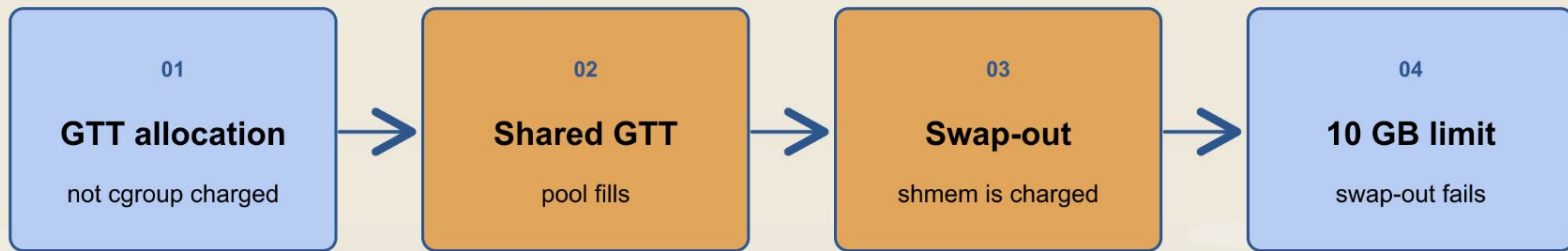


LINUX
PLUMBERS CONFERENCE

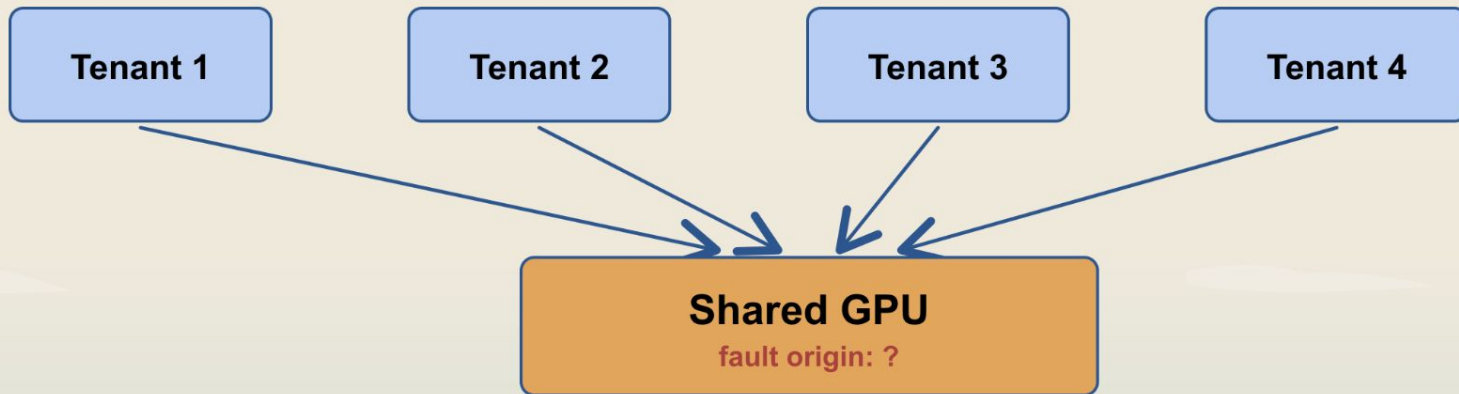
PRAGUE, CZECHIA / OCT. 5-7, 2026



Why allocation doesn't stop?



Debuggability: One reset which Tenant?



All tenants on this GPU affected



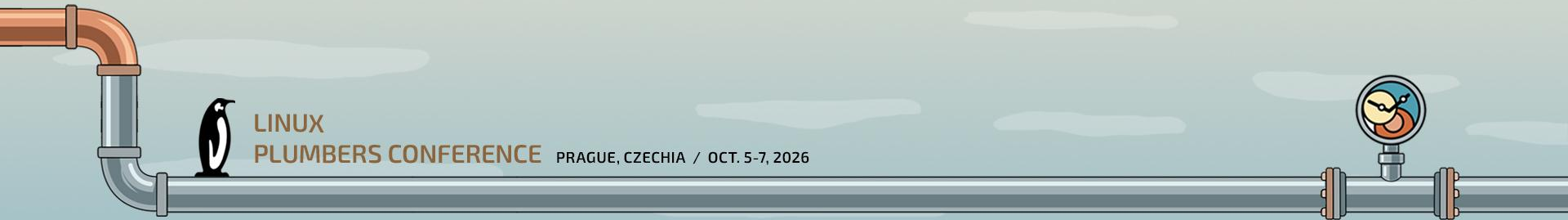
LINUX
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026



Debuggability: Current State of the Art

- Register snapshots in GPU dump
- Some level of issue categorization
- There are debugging features provided by RADV/MESA
 - Can't be run in production
 - Reproducibility is not 100%.
- Rest we are unsure
 - It's probably not enough as problems were not being root-caused in an acceptable time!



Case Study: Practical GPU Reset

**User
Bug**

**Driver
Bug**

**Firmware/HW
Bug**

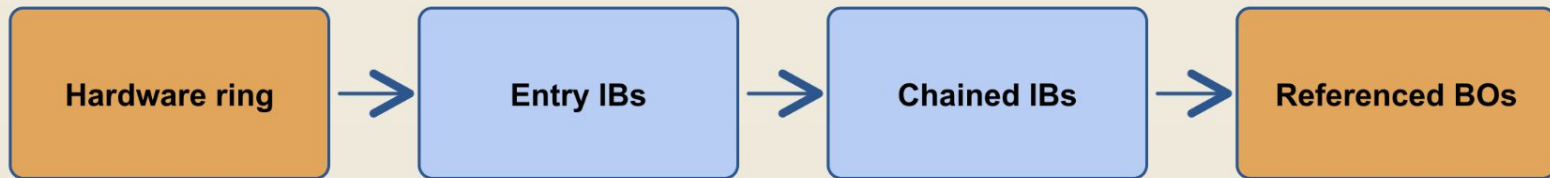


**LINUX
PLUMBERS CONFERENCE**

PRAGUE, CZECHIA / OCT. 5-7, 2026



CBD: Command Buffer Decoder



Bounded graph walk

Capture selected bytes; mark incomplete dumps

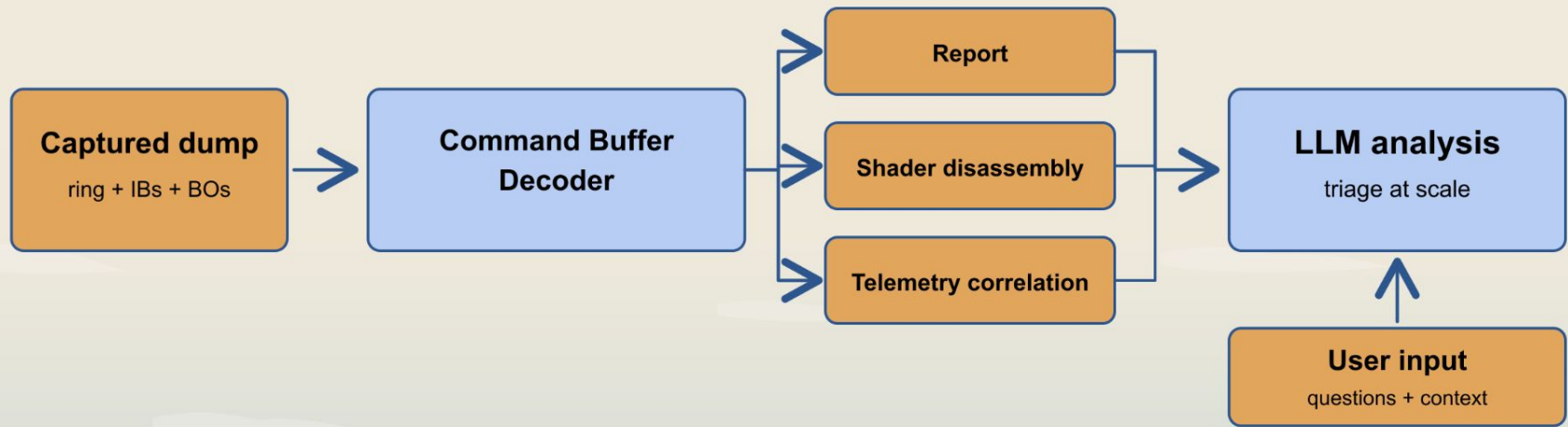


LINUX
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026



Debuggability: Automate for Scalability

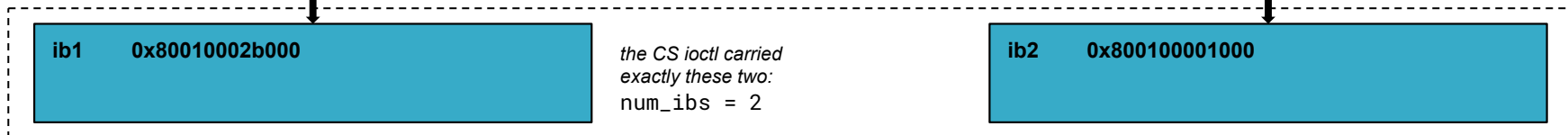


CBD gives Ring Buffer Anatomy

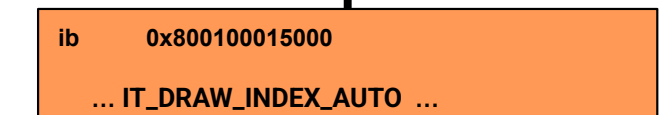
1 · HARDWARE QUEUE



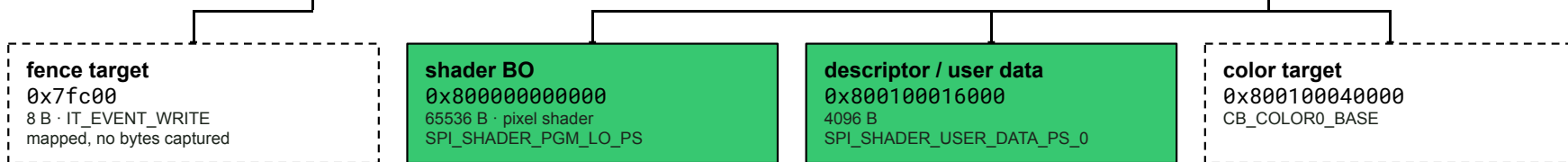
2 · ENTRY IBs



3 · CHAINED IB



4 · REFERENCED MEMORY

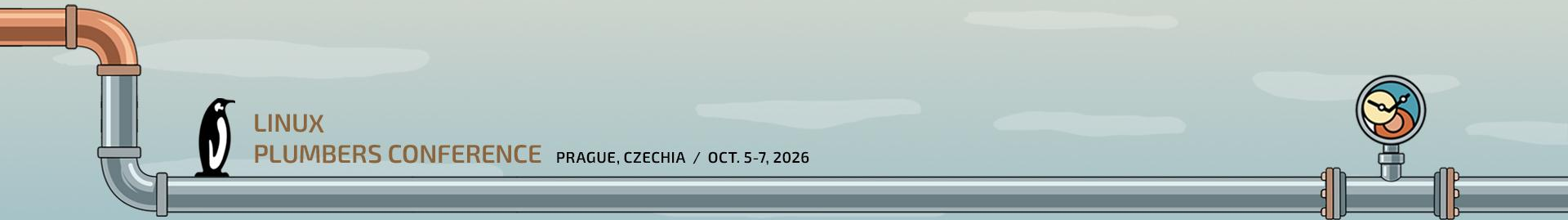


ring entry IB (in the array) chained IB (decoded out of a packet) non-IB memory, captured [] referenced / optionally resolved at hang time

OS/Driver Contention

Case Studies:

- Multi-GPU Lock Contention in Nvidia GPU driver



Multi-GPU Lock Contention

Problem: the more isn't always the merrier.

A 3-GPU platform

- Each tenant can use at most one GPU.
- Total throughput with 1 GPU enabled vs 3 GPUs enabled is far from 1:3
 - No other system resource was the bottleneck.



LINUX

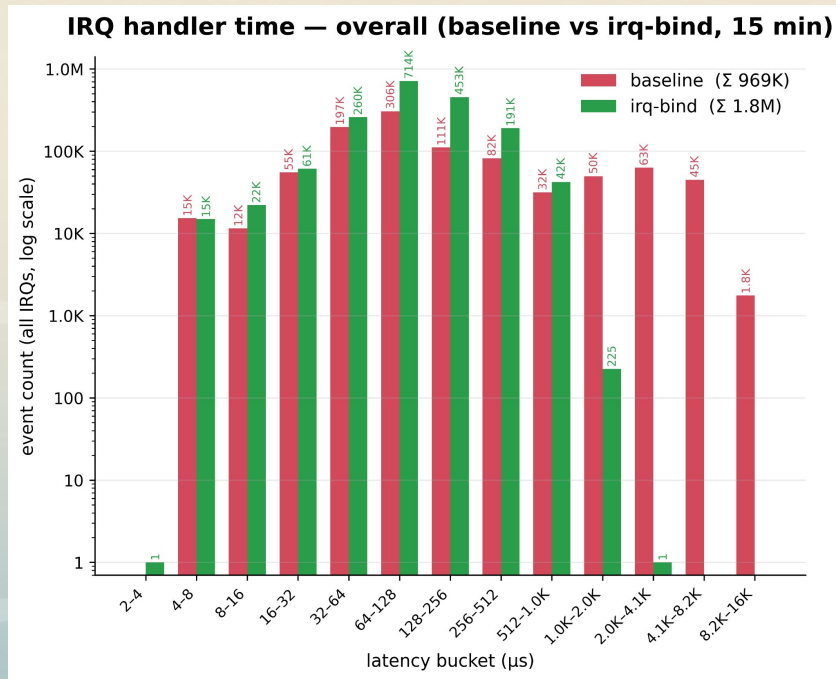
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

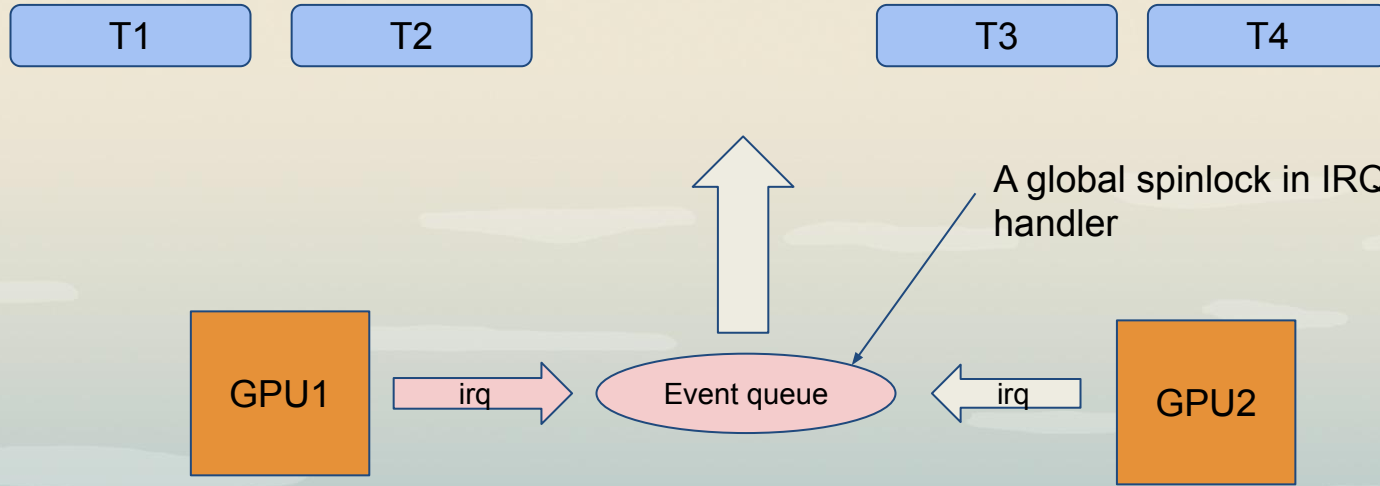


Multi-GPU Lock Contention

- A 3-GPU platform
 - Each tenant can use one GPU
 - Total throughput 1-GPU vs 3-GPU is far from 1:3.
- Binding all GPU interrupts to one CPU cured the issue



Multi-GPU Lock Contention



LINUX
PLUMBERS CONFERENCE

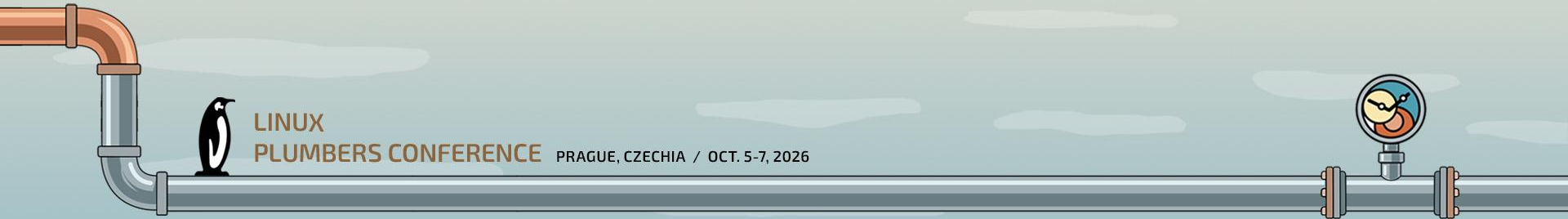
PRAGUE, CZECHIA / OCT. 5-7, 2026



Multi-GPU Lock Contention

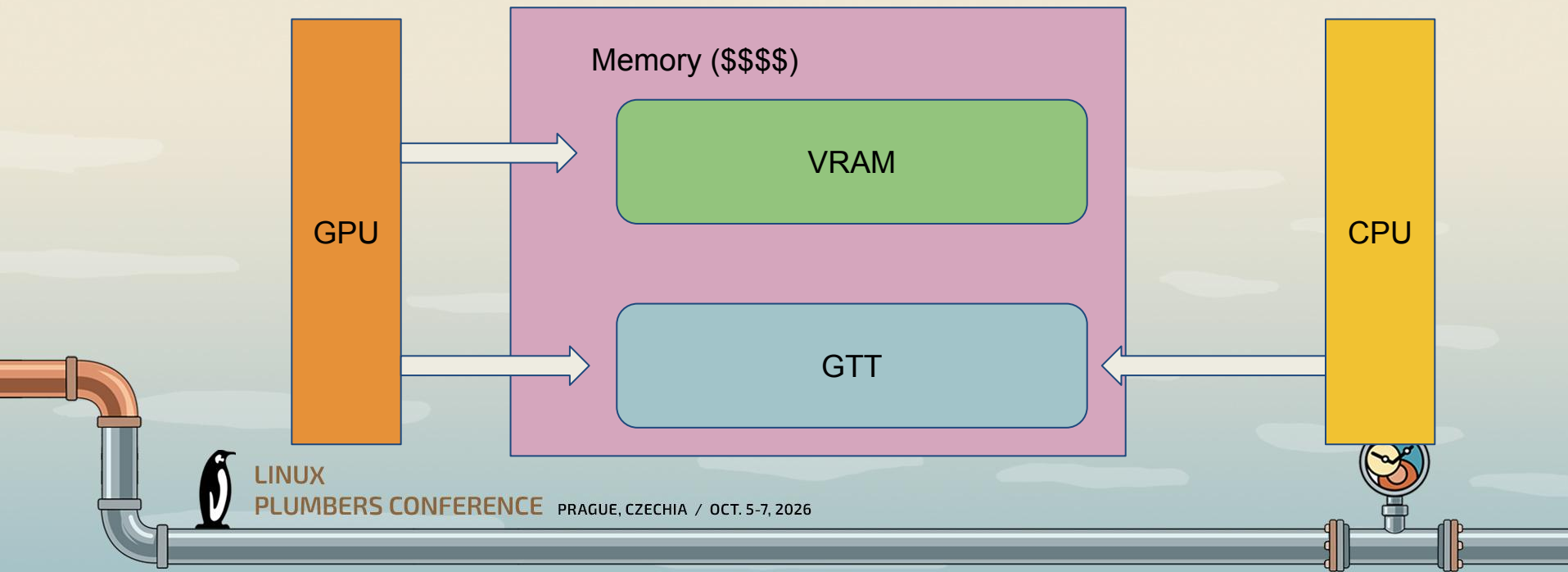
Fix:

- [RmEventNotifyIsrOpt](#) in NVIDIA/open-gpu-kernel-modules



Performance

Case Study: AMD APU VRAM vs GTT latency gap



AMD APU VRAM vs GTT latency gap

GPU access latency in us

Footprint	Workload	VRAM	GTT	Ratio GTT/VRAM
4 MiB	sequential	750	1,188	1.58x
4 MiB	random	1,165	1,191	1.02x
64 MiB	random	5,092	9,872	1.94x
256 MiB	random	8,563	30,273	3.54x

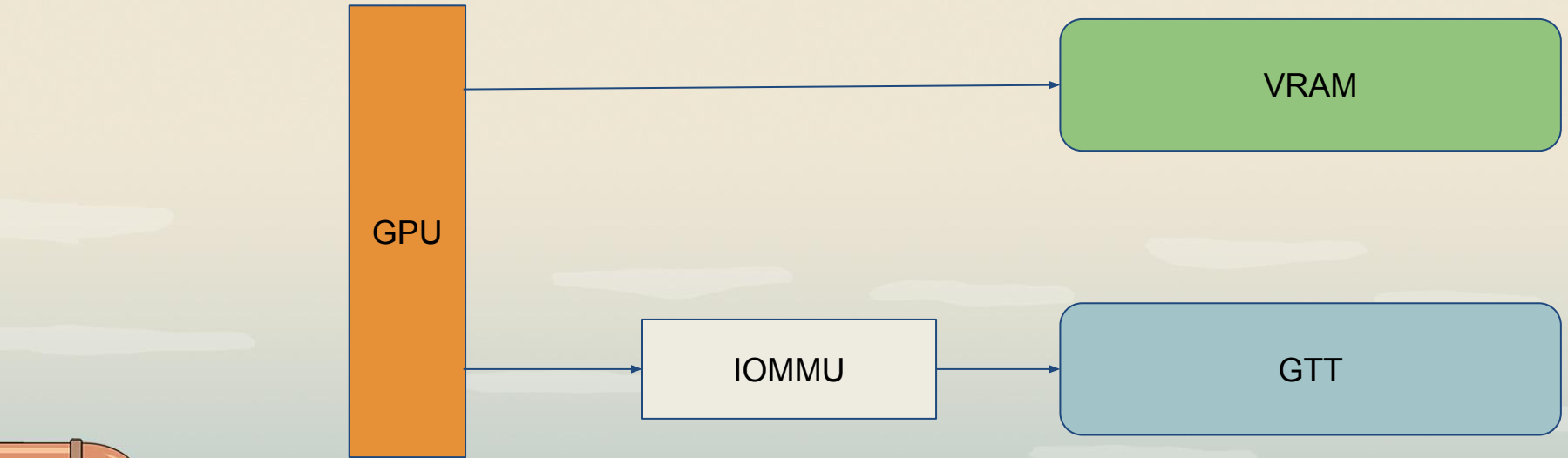


LINUX
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026



VRAM vs GTT latency gap

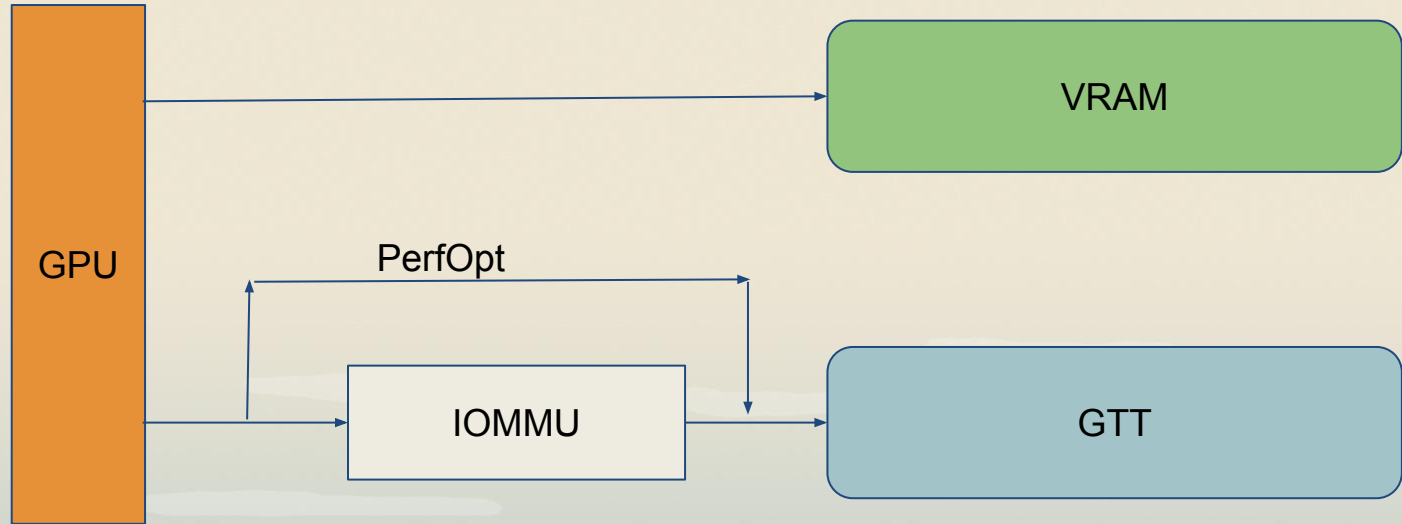


LINUX
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026

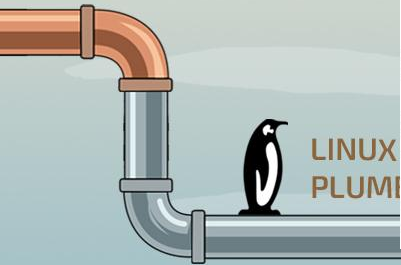


VRAM vs GTT latency gap



MMIO Offset 016Ch IOMMU Performance Optimization Control Register

This register is used by system software to enable the optional IOMMU Performance Optimization feature. This feature allows accesses from privileged (usually integrated) I/O devices such as GPUs to bypass the IOMMU when directly accessing system memory. It is recommended to set this bit, unless it is planned for the system software to use the the host I/O tables for GPA->SPA redirection



VRAM vs GTT latency gap

[PATCH v2 0/2] IOMMU Performance Optimization support

GPU access latency in us

Footprint	Workload	VRAM	GTT	GTT (fixed)	Ratio GTT/VRAM -> fixed/VRAM
4 MiB	sequential	750	1,188	783	1.58x -> 1.04x
4 MiB	random	1,165	1,191	1,171	1.02x -> 1.01x
64 MiB	random	5,092	9,872	5,411	1.94x -> 1.06x
256 MiB	random	8,563	30,273	9,526	3.54x -> 1.11x

Outro

- **Sharing** a GPU isn't the same as **isolating** tenants
- Blast radius for a bug is larger.
- Debugging is hard as reproducibility is harder.
- Of course, we are not planning to solve all the challenges by ourselves ;-)

Isolation

Debuggability

Performance

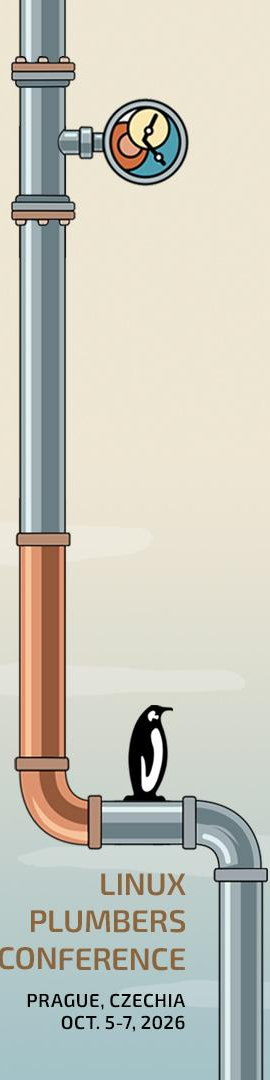
Contention



LINUX
PLUMBERS CONFERENCE

PRAGUE, CZECHIA / OCT. 5-7, 2026



A vertical pipe with a gauge and a penguin on a horizontal section. The pipe is grey with orange-brown joints. A circular gauge with a yellow face and blue needle is attached to the upper part. A penguin is perched on a horizontal section of the pipe near the bottom. The background is a light beige gradient with faint, stylized clouds.

Thanks!

**LINUX
PLUMBERS
CONFERENCE**

PRAGUE, CZECHIA
OCT. 5-7, 2026