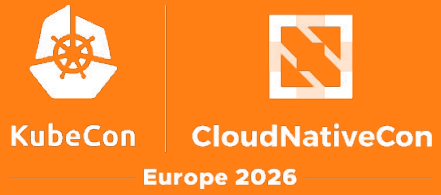


Lightning Talk: The $100K GPU Mystery - Why your AI training dies at 99%

Michael Ifeanyi, Google

Talking Points

KubeCon CloudNativeCon Europe 2026

1. **The Problem**
   When your monitoring lies and training jobs crash at 99%
2. **Understanding Contiguous Memory**
   Why 10 GB free does not mean you can allocate 7 GB
3. **The Root Cause: GPU Memory Fragmentation**
   Real test results from Tesla T4 on GKE
4. **The Solution: Monitor & Mitigate**
   Metrics that matter and practical fixes
5. **Key Takeaways & Resources**
   What you can do today

The Blind Spot in your GPU monitoring

KubeCon CloudNativeCon Europe 2026

What nvidia-smi Shows

57%
Utilized

6.71 GB
Free Memory

"Plenty of room for your job!"

What Actually Happened

```
CUDA out of memory.
Tried to allocate 7.54 GiB
6.08 GiB free
reserved >> allocated
```

Job crashed at 99% complete

How can a job fail when monitoring shows plenty of free memory?

nvidia-smi reports TOTAL free memory — not CONTIGUOUS free memory

That single distinction is the root of the problem.

What Does "Contiguous" Mean?

Contiguous Memory — One Solid Block

10 GB FREE — All in one block

Your 7 GB allocation fits easily ✓

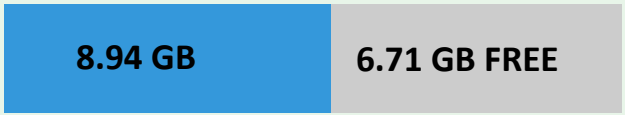
Fragmented Memory — Scattered Pieces

Total Free: 2.5 + 3 + 4.5 = 10 GB ✓

7 GB allocation FAILS — largest single block is only 4.5 GB ✗

Root Cause: Memory Fragmentation

nvidia-smi reports:



"57% utilized - plenty of room"

VS

PyTorch actually sees:

Peak fragmentation: 62.5% (5.37 GB)

The Allocation Attempt

ACTUAL GPU MEMORY LAYOUT:

Needs 7.54 GB contiguous → FAILS!

The Solution: Monitor & Mitigate

KubeCon CloudNativeCon Europe 2026

1 Monitor the RIGHT Metrics

- Use torch.cuda.memory_stats() API to track fragmentation ratio
- Track: (reserved − allocated) / reserved
- Export to Prometheus for visibility

2 Reduce Memory Pressure

- Smaller batch sizes (always works)
- Gradient checkpointing (~20% speed trade)
- Mixed precision (FP16/BF16)

3 Kubernetes: Custom Fragmentation Metric

- Deploy DaemonSet to collect fragmentation on all GPU nodes
- Export to Prometheus, alert when threshold exceeded
- Drain nodes + checkpointing enables recovery (no data loss)

Key Takeaways

KubeCon CloudNativeCon Europe 2026

1. nvidia-smi has a blind spot — it shows total free memory, not usable contiguous blocks

2. Fragmentation accumulates during training- we hit 62.5% peak, preventing allocation even with apparent free memory

3. Monitor fragmentation as a first-class metric, reduce memory pressure with proven techniques, and deploy DaemonSet + Prometheus for monitoring

THANK YOU