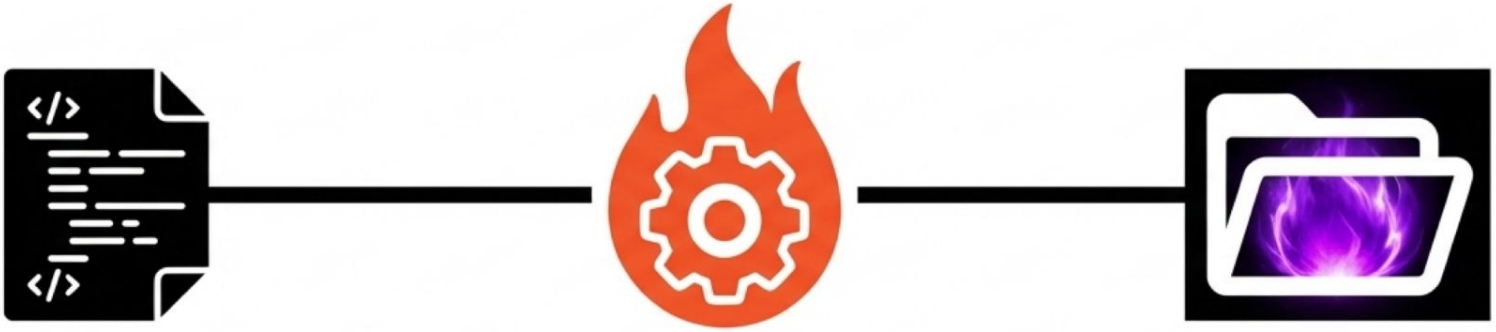

# Slash LLM Cold-Start Times by Pre-distributing GPU Caches

Billy McFall & Maryam Tahhan, Red Hat

# The hidden cost of scale is a severe JIT compilation penalty


| Redundant Compilation | The Scaling Penalty | The Cold Start Reality |
| --- | --- | --- |
| PyTorch workloads repeatedly trigger heavy Triton/NVCC compilation per instance. | Distributed inference wastes expensive GPU cycles rebuilding identical caches across nodes. | This redundancy creates a severe latency penalty, costing literal minutes per pod. |

# Capturing the localized GPU Kernel Cache artifact



**PyTorch Code**

Standard execution triggers compilation.

**JIT Compiler**

Generates highly specialized machine code, causing severe runtime delays.

**Hardware-Specific Binary**

The resulting GPU Kernel Cache directory. A highly valuable, reusable artifact.

# Model Cache Vault standardizes and secures the artifact


### Standardization

Packages GPU Kernel Caches into standard OCI-compliant container images.

### Tooling Compatibility

Works seamlessly with existing container tools like Docker and Buildah.

### Supply-Chain Security

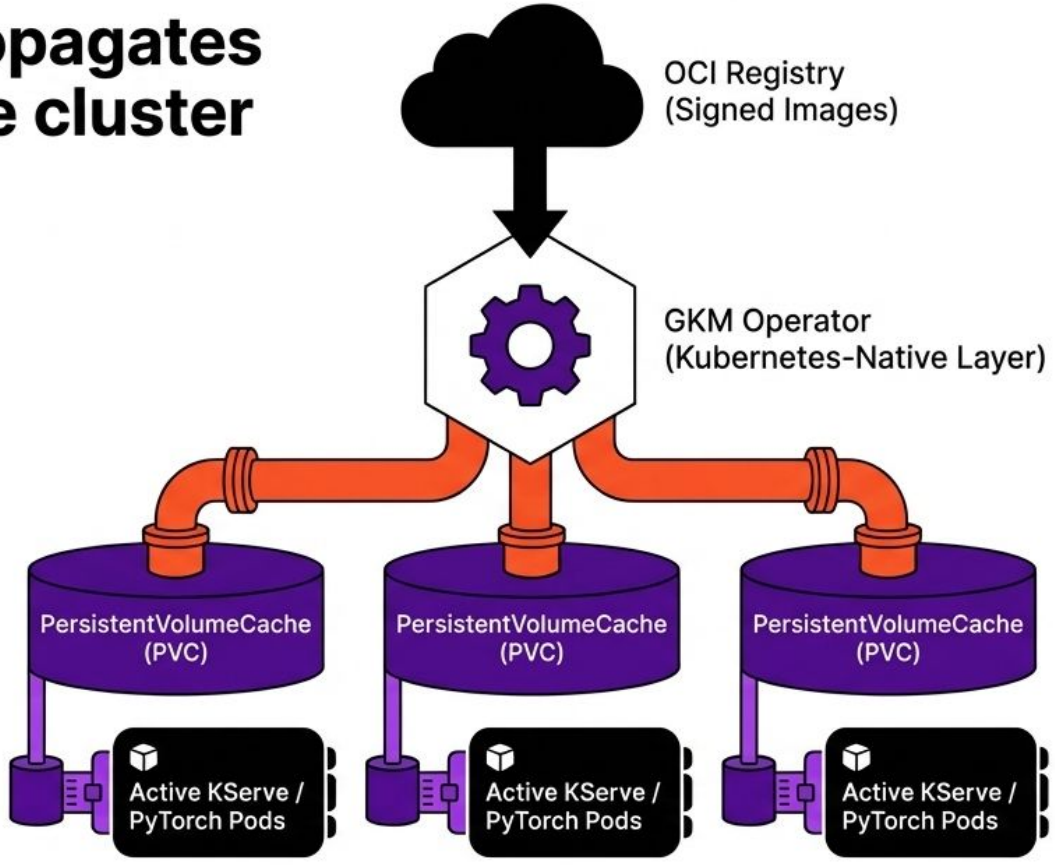
Automatically signs and verifies cache artifacts using Sigstore Cosign.

# GKM automatically propagates GPU caches across the cluster

**1. Automated Extraction**: Pulls signed OCI Images and extracts them into PVCs.

**2. Cluster Propagation**: Operator watches and distributes across your entire environment.

**3. Direct Mounting**: Mounts the verified cache directly into workloads.



# The automated logistics pipeline for instant execution


**1. JIT Compile**

Generate baseline cache on matching GPU.

**2. MCV Package & Sign**

Wrap directory into secure OCI Image.

**3. OCI Registry**

Store the signed image centrally.

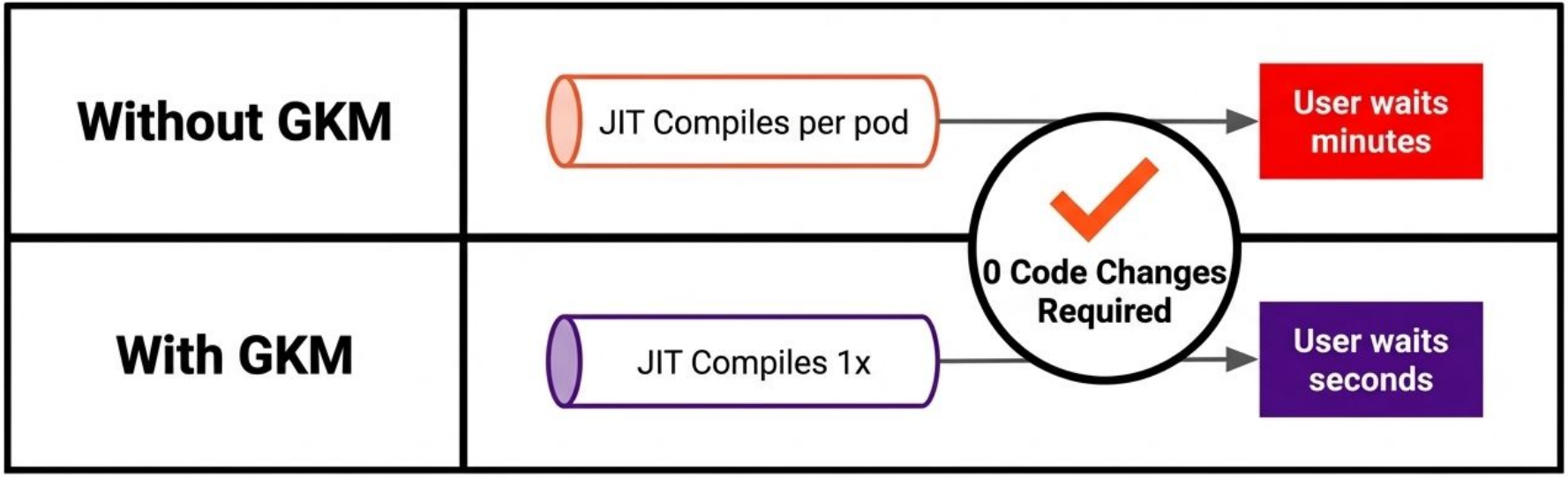
**4. GKM/KServe Operator:**

Extract image to PVC on target nodes.

**5. Warm Pod**

Mount cache and execute instantly.

# Cut startup latency in half with zero code modifications




| | |
|---|---|
| **Massive Latency Reduction:** Pod startup time decreased by up to 50%. | **Fully Transparent:** The workload is completely unaware the cache was pre-distributed. |

# The production blueprint for a secure ML supply chain

**Stop rebuilding caches** and **start pre-distributing them.**

---

**Secure your ML supply chain** with signed OCI images.

---

**Get Involved:** GKM is currently transitioning to the KServe community!

github.com/redhat-et/GKM

github.com/kserve/kserve

# Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.