

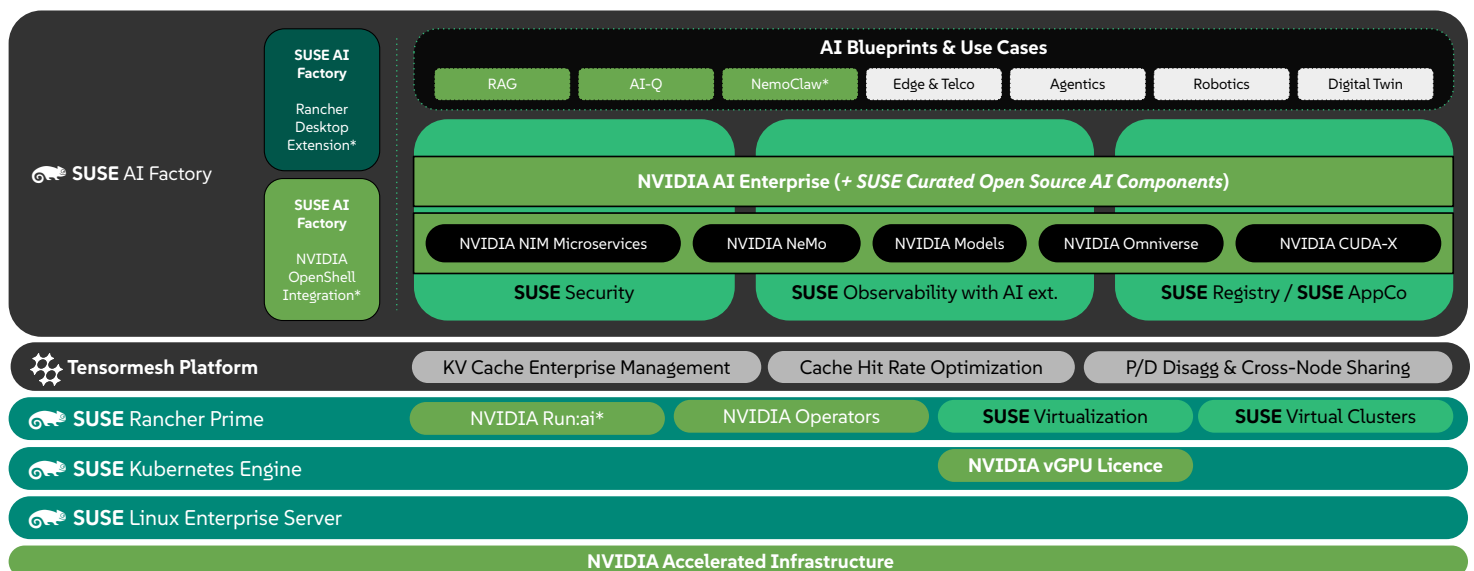
SUSE AI Factory with NVIDIA and Tensormesh

Maximizing Token Throughput with Intelligent KV Caching

SUSE AI Factory with NVIDIA is a Private Enterprise AI solution that natively embeds NVIDIA AI Enterprise directly into SUSE's sovereign, hardened infrastructure.

Built on top of the popular LMCache project, **Tensormesh Platform** enables and manages an enterprise-ready KV caching layer, which is the most effective and efficient way of optimizing inference costs and latency.

Together, they address two of enterprise AI's toughest challenges: securing data and models to meet compliance requirements, and keeping inference costs predictable at scale, especially in regulated industries and in RAG and agentic workloads.



Cost Reduction

Tensormesh Platform cuts AI inference compute costs by managing your KV cache intelligently across your entire infrastructure — GPU memory, CPU memory, and external storage. Whether you rely on prefix or non-prefix prompt caching, or need advanced configurations such as prefill/decode disaggregation and KV cache sharing across GPU nodes, the platform adapts to your infrastructure and workload to deliver best-in-class cache hit rates and token throughput.

Ease of Deployment and Enterprise Support

Deploying and operating the platform is just as efficient. Helm chart-based installation gets you running quickly, and a web UI paired with a command-line interface gives you the tools you need for both Day 1 setup and Day 2 support of your AI inference stack.

Multi Vendor Support

The joint solution is built to run on the infrastructure you have, not the one a vendor prescribes. It integrates across the stack at every layer: accelerators, inference engines, and any storage provider. It's a single data management layer for AI inference, tightly integrated with SUSE, with no vendor lock-in.

Proof Points: Real-World Use Cases

1. AWS SageMaker HyperPod

LMCache integrates with Amazon SageMaker HyperPod's managed tiered storage as an L2 cache backend. In a joint benchmark, this delivered up to **1.67x** improvement in inter-token latency and **1.27x** higher throughput for multi-round chat workloads under high concurrency, compared to L1-only (GPU/CPU) configurations. *LMCache on Amazon SageMaker HyperPod: Accelerating LLM Inference with Managed Tiered KV Cache — blog.lmcache.ai*

2. PrimaLabs

As AI serving scales, multi-tier distributed caching becomes essential. PrimaLabs validates that Tensormesh Platform is tackling that hard problem well, and that the industry needs more of this kind of infrastructure.

3. SemiAnalysis AgentX

SemiAnalysis' AgentX benchmark put LMCache through sustained, high-concurrency agentic workloads on Kimi K3 and found real cost-efficiency gains: storage per token dropped by nearly **20x** for certain hybrid model architectures.

LMCache Agentic Optimizations — inference.semianalysis.com

4. Samsung

"As AI workloads grow, intelligent reuse of cached state has become one of the most powerful levers for performance and cost efficiency."Tensormesh's LMCache is built to take full advantage of next-generation storage, push[ing] the boundaries of what's possible across the AI-stack." — Leno Park, Vice President of NAND Product Planning, Samsung Electronics

A proven track record is what earned Tensormesh Platform certification and integration for **SUSE AI Factory with NVIDIA**. Enterprise customers can now bring the same high-performance caching capabilities into their own sovereign data centers, with automated Helm chart deployment and joint vendor support.

SUSE AI Factory with NVIDIA

Pre-Validated, Use-Case Aligned Blueprints

- Tightly packaged open-source AI components and 3rd-party ISV software into blueprints tailored for specific workloads (GenAI Inference, RAG, Computer Vision, MLOps)
- Initially aligned with proven SUSE domains (Edge, Telco, SAP) providing turnkey solutions as commercial beachheads

Scalable Deployments Anywhere

- Seamlessly provision AI workloads across any supported environment: tactical edge, core data center, public cloud
- Manage both decentralized, air-gapped footprints and large enterprise fleets through single pane of glass

Unified Governance & Observability

- Single common portal surfacing all critical telemetry
- Fine-grained, real-time visibility into GPU utilization, LLM token throughput, cost modeling, traffic patterns

Tensormesh Platform

State of Art Performance with KV Cache Enterprise Features and Support

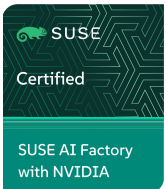
- Full visibility into what's cached and where, across your entire infrastructure
- Pin the prefixes that matter so they're never evicted
- Move data between tiers or clear it on demand, per workload, per tenant, or fleet-wide
- Declarative policies, versioned like the rest of your infrastructure

Prefix and Non Prefix Caching & External Storage offloading

- Extends KV cache beyond GPU memory onto CPU DRAM and shared storage
- Context computed once is reused everywhere, even persisting across restarts
- Faster time-to-first-token and more headroom on every GPU
- Cache scales with your storage, not your accelerator budget

P/D Disaggregation & KV Cache Sharing Across Nodes

- Prefill and decode can run on separate pools, sized independently to your traffic
- KV state moves GPU-to-GPU instead of being recomputed
- Any node can serve warm context to any other, over RDMA or plain TCP
- Predictable latency under load, with no GPU cycle wasted on duplicate work



SUSE AI Factory
Learn more about
NVIDIA solution
[Explore Details](#)

Tensormesh Platform
Certified for SUSE AI Factory
with NVIDIA
[link](#)



Tensormesh Platform Visit
official website
tensormesh.ai

SUSE is the leading sovereign open infrastructure company where hybrid AI runs. Backed by 30 years of community collaboration, we transform open source innovation into secure, enterprise-grade solutions for mission-critical environments. Built on choice and neutrality, SUSE is the only platform that manages any Linux and Kubernetes distribution. Through our open infrastructure platform and partner ecosystem, IT leaders gain control over their data, budgets, and sovereign AI strategies. Our solutions deploy across hybrid environments, from developer workstations and on-premises data centers to multi-cloud and the farthest edge, and are optimized and certified for SAP. With SUSE, choice happens.

Tensormesh is the leader in smart AI-native data. Founded by the team behind the open-source LMCACHE project, Tensormesh enables a multi-tier, scalable KV cache that turns the enterprise's AI data into a managed, optimized asset. Tensormesh is backed by Laude Ventures, AMD Ventures, CoreWeave, NVIDIA's NVenture, and more.

SUSE S.A.
11-13 Boulevard de la Foire
L-1528 Luxembourg
Grand Duchy of Luxembourg

benelux@suse.com
+31 23-8080010

[SUSE.COM](https://suse.com)

Copyright © 2026 SUSE LLC and/or its affiliates. All rights reserved. SUSE, the SUSE logo, and all SUSE marks are trademarks or registered trademarks of SUSE LLC or its subsidiaries in the United States and other countries. All third-party trademarks are the property of their respective owners.