



主權AI不是選擇,是逼出來的:從 架構抉擇到治理實踐

Joey.Cheng

SUSE Solution Architect

joey.cheng@suse.com





88%

的全球企業
至少在一個業務功能中使用了 GenAI

資料來源：McKinsey, State of AI Global Survey, 2025



選擇了雲...然後...

Downdetector解析 受到影響

大當機災情，超過60個國家與3,500家公司

在10月19日與20日，公有雲服務 的雲端區域US-EAST-1發生部分服務故障的狀況，後續導致數千家公司受到影響，網路測速與網路品質分析服務公司Ookla在22日發布報告，具體呈現災情的樣貌。他們根據旗下Downdetector收到1,700萬份通報進行分析，總計影響超過60個國家、3,500家公司

Uber Burns Its 2026 AI Budget In Four Months On

By [Janakiram MSV](#), Senior Contributor. ⓘ I cover emerging technologies with ...

Follow Aut

Published May 17, 2026, 11:18pm EDT, Updated May 27, 2026, 12:17pm EDT

Incident Report

Summary

To our customers whose businesses were impacted during this outage, we sincerely apologize. This is not the level of quality and reliability we strive to offer, and we are taking immediate steps to improve the platform's performance and resilience.

Water Leak and Fire Damage

On Tuesday, 25 April 2023 at 16:46 US/Pacific, a cooling system water pipe leak occurred in one of the data centers in the europe-west9 region. The leak originated in a non-Google portion of the facility, entered an associated uninterruptible power supply (UPS) room, and led to a fire. The fire required evacuation of the facility, engagement from the local fire department, and a power shutdown of the entire data center building for several hours. The fire was successfully controlled on 26 April 2023 at 04:11 US/Pacific.

Regional Impact

Europe-west9 contains three buildings with independent cooling, power, and networking. Regional Spanner supports the regional control plane for most Google Cloud services in the region. In europe-west9, regional Spanner's replicas were not correctly distributed across the three buildings in the region to ensure quorum was available after the impacted building was powered down. As a result, the control plane services for many Google Cloud services were interrupted in europe-west9, causing unavailability and/or elevated error rates for these services in this region. Regional Spanner was restored on 26 April 2023 at 12:47 US/Pacific.

The Identity and Authentication Management (IAM) service in europe-west9 depends upon regional Spanner. After regional Spanner came back online, IAM recovered on 26 April 2023 at 14:42 US/Pacific. IAM recovery took longer than expected, driven by the time to refresh the policies for the region as a result of the outage. Some Google Cloud services that were impacted by this event began recovering soon after Spanner and IAM were back online.

Google Compute Engine (GCE) and Persistent Disk (PD) experienced regional impact in both the data plane (GCE Virtual Machines (VMs), PD Volumes) and control plane systems. Both systems had extended recovery periods after Spanner and IAM were restored. GCE VMs and PDs in europe-west9 zones were online by 19:00 US/Pacific, except for 58% of GCE VMs in zone europe-west9-a, which were directly impacted by the water leak. Soon after Spanner, IAM, GCE/PD were online, the europe-west9 region was primarily recovered and operational.



因此，近期「Sovereign AI」被提到的頻率越來越高

企業想要AI能力，資料、模型、運算環境的掌控權留在自己手上——不管是法規要求，還是IT治理的既定原則。

- 資料自主：企業的資料，企業控管
- AI自主：AI在企業環境做了什麼事情，企業要了解
- 成本自主：企業會在AI的花費是什麼，企業自主控管



66%

的全球企業

是透過 **K8S 架構**

運行 GenAI Inference 應用

CNCF Annual Cloud Native Survey, 2025



應用生態系

- GenAI應用微服務結構
- 主流AI推論工具鏈綁定



高可用性

- 應用面
- 叢集節點面



擴展彈性

- 應用效能擴展
- 叢集高可用性



資源使用與成本效益

- GPU資源分配
- 初期投入成本可控

85%

的全球企業

GenAI應用的POC無法成功落地

Gartner 2025 AI deployment survey



K8S管理複雜

- Workload部署
- 叢集更新
- Day2維運



合規性

- 開源工具漏洞問題



安全性

- 容器安全性
- Hugging Face被駭案例



可視化

- K8S AI應用運作情況
- Token使用量不可見
- GPU使用率不知道



SUSE AI Factory with NVIDIA

SUSE AI Factory with NVIDIA

企業主權AI的統一解決方案



強大且友善的 K8S叢集管理

建立在可信賴的 K8S管理工具 SUSE Rancher Prime之上

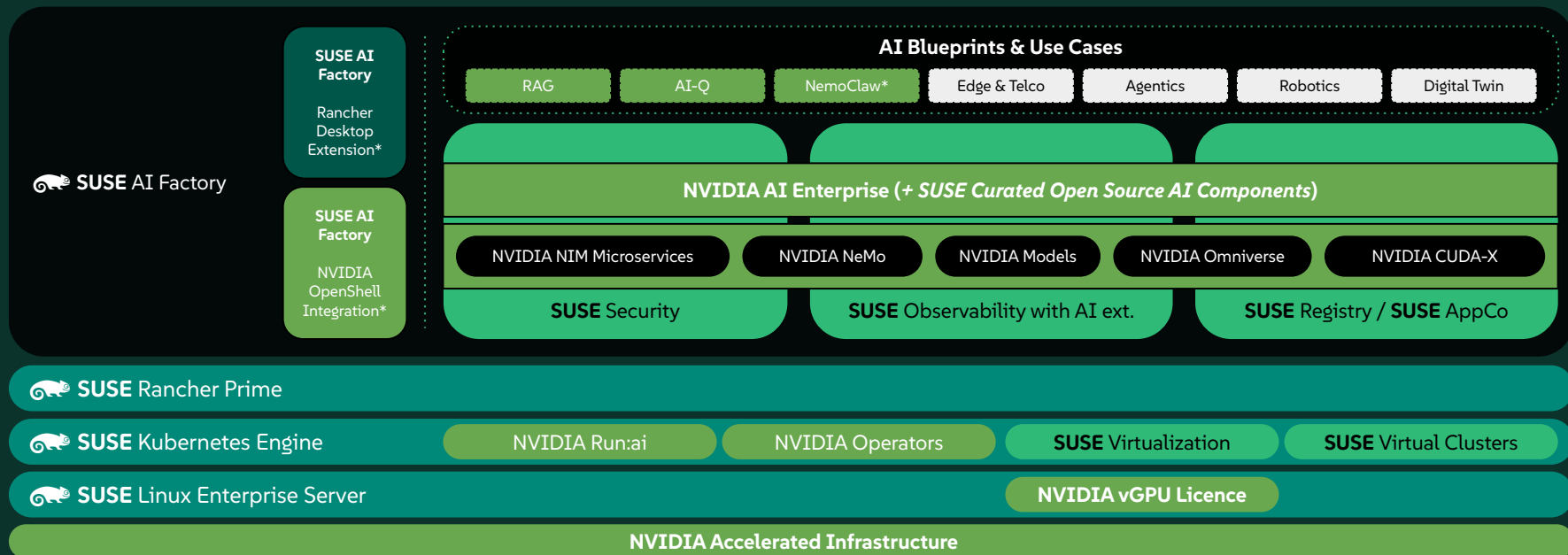
- 多叢集管理與大規模應用部署
- 容器應用合規性
- 監控與安全性
- GenAI智能助手，全面且方便地進行維運管理

NVIDIA AI生態系深度整合

將NVIDIA AI Enterprise生態系工具與應用整合至 SUSE AI Factory治理框架

- 最佳化 Nvidia GenAI應用的推論效能
- 快速部署 Nvidia相關應用
- 企業AI成本可視化
- 技術支援鏈整合

SUSE AI Factory with NVIDIA 結構一覽



SUSE AI Factory - K8S管理面優勢

跨叢集方便管理

- 將跨機房、多K8S叢集納管至單一 UI 管理介面
- 零遷移導入管理

大規模部署一致性

- 以Github Repo內的helm, yaml, Kustomize部署形式, 進行應用的自動化一致性部署
- 提供安全、受控且可在地端管轄的容器Image管理環境

觀測性

- 全方位K8S效能監控與稽核日誌
- 提供容器相依拓樸圖與容器實時健康狀況
- 清楚了解容器關聯性, 加速維運除錯時間

安全性

- 容器執行期行為監控, 即時偵測橫向移動與提權
- 實時偵測並阻斷 DDoS、DNS 隧道(Tunneling)、SQL 注入及反向Shell 攻擊。

合規性

- 部署合規的第三方GenAI應用, 防堵應用漏洞

AI維運助手

- 智能應用部署、維運與除錯
- 依使用者本身RBAC權限運作, 不逾越權限、不在未經核准下執行動作
- 可外接企業自有模型, 資料主權由企業自主掌控

智能K8S管理助手 - Liz



SUSE AI Factory - Nvidia生態系整合優勢



Nvidia GPU資源效益最大化

- **Run:ai**: GPU動態排程與跨團隊資源共享
- **CUDA-X**: 提升AI工作負載效能



預驗證的Nvidia AI應用與藍圖

- Nemo/NIM: 進行AI模型微調、對齊、Guardrail, 並將模型封裝成推論容器, 正式上線接客
- NemoClaw: 企業級Agent secure runtime
- RAG, AI-Q: 現成藍圖, 直接部署無需重新兜整條技術棧



成本可視化與預測

- 追蹤 GenAI 應用的 token 消耗與 GPU 使用率
- 依模型計費倍率換算成本, 幫企業預測 AI 支出、防止預算超支



技術支援鏈整合

- SUSE負責L1/L2, 問題升級由SUSE主動對接NVIDIA L3
- 企業全程僅需對接單一窗口, 免去廠商間來回舉證。



SUSE



NVIDIA

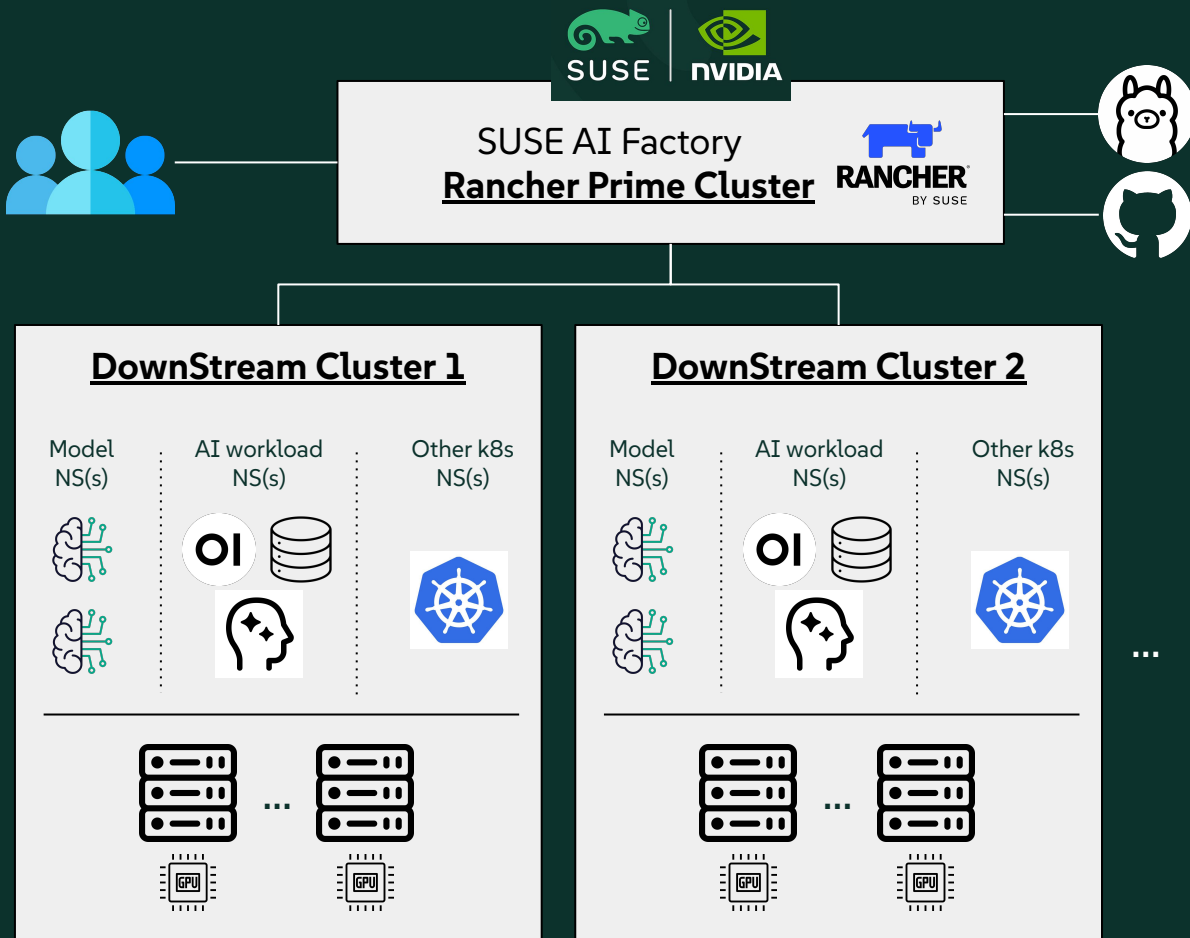
SUSE AI Factory

架構亮點

強大且友善的 K8S叢集管理

SUSE Rancher Prime

- 多叢集管理
- 大規模應用部署
- 開源容器應用合規性
- 監控與安全性
- GenAI智能助手



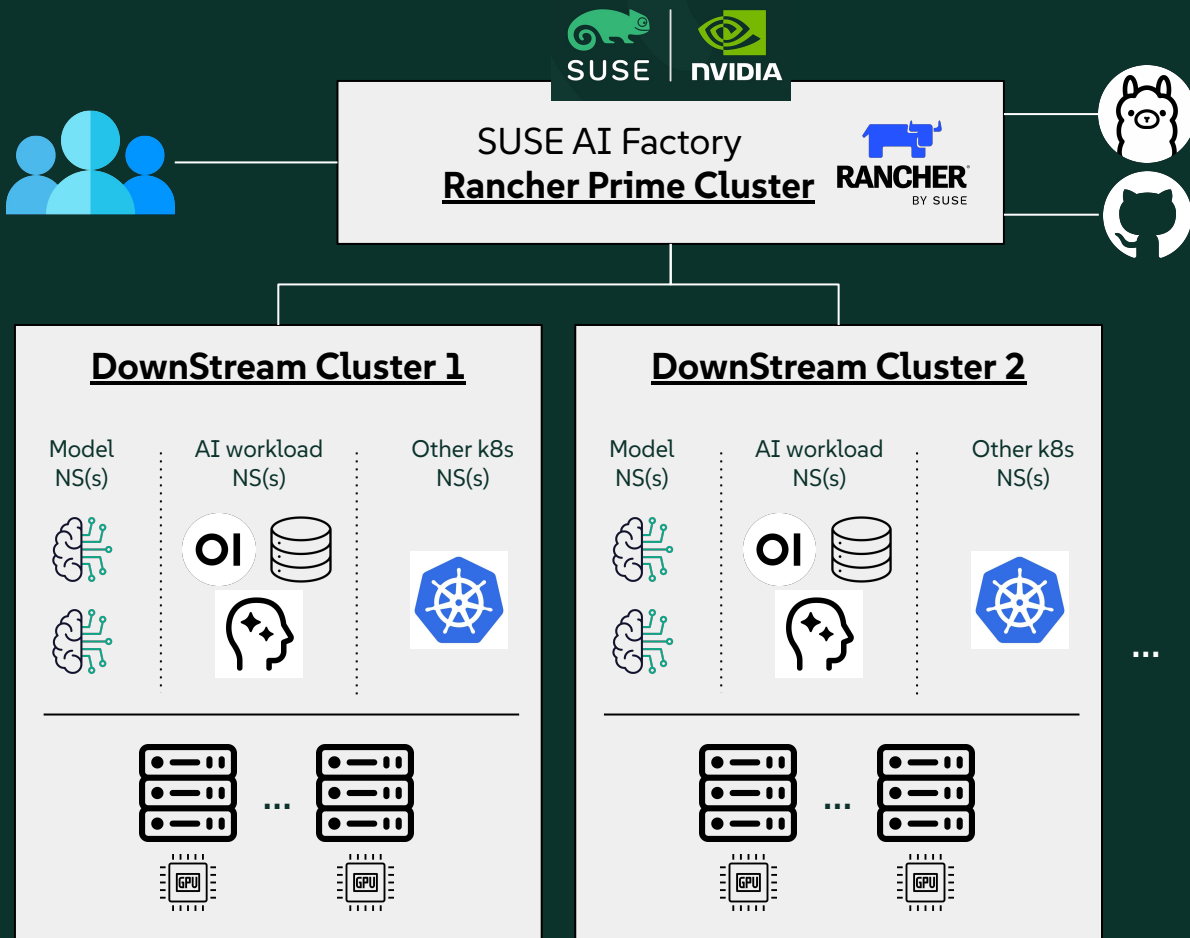
SUSE AI Factory

架構亮點

NVIDIA AI生態系深度整合

NVIDIA AI Enterprise生態系整合

- 最佳化 Nvidia GenAI應用的推論效能
- 快速部署 Nvidia相關應用
- 企業AI成本可視化
- 技術支援鏈整合



SUSE AI Factory with NVIDIA

大幅加速 GenAI 應用落地

數個月

傳統建置與測試



數天

SUSE AI Factory with NVIDIA

將 NVIDIA AI Enterprise 生態系工具與應用整合至 SUSE AI Factory 治理框架，
企業拿到的就是「已經驗證過、可以上生產環境」的 GenAI Inference 應用與平台。

Key Takeaway

- 為何企業會走向主權AI
 - 雲端不是絕對可靠 - 平台穩定、成本控管以及模型可控性
 - 掌控權(資料、模型、成本)本該留在自己手上,不是委外給第三方
- 企業主權AI的痛點
 - **K8s管理複雜**: Workload部署、叢集更新、Day2維運皆耗費大量人力與維運知識
 - **成本效益不透明**: GPU使用率如何? 模型產生了多少Token? 企業花了錢卻看不見花在哪
 - **合規性**: 採用開源工具的漏洞問題如何處理?
 - **安全性**: Agent提權導致企業內部系統被入侵, 如何預防?
- Why SUSE AI factory
 - 跨叢集統一管理、Fleet大規模部署一致性與 AI助手降低維運門檻 —— 解決K8s管理複雜
 - GPU/Token用量可視化, —— 解決成本不透明
 - 開源元件強化, 大幅降低CVE漏洞數 —— 解決合規性
 - 偵測提權與橫向移動, 執行期即時攔截 —— 解決安全性
 - NVIDIA深度整合: GPU資源優化, 預驗證藍圖加速落地, 單一支援窗口

SUSE 工作坊

AI 寫得出程式，誰讓它上線？動手走一遍 agentic 交付流程，從應用部署到 AI 執行環境

時間：9/12 PM 15:30 - 17:00

地點：605+606 會議室 6F



主講者：Sam Chen

擁有 14 年以上開源技術實務經驗，現任 SUSE 解決方案架構師。專精於大型雲端環境、分散式系統及雲原生解決方案。擅長將複雜的開源架構轉化為企業級穩定方案，曾協助多家企業完成容器化管理與架構升級。



THANK YOU



GenAI典型企業應用

通用問答



1. 解決Shadow AI問題, 避免企業隱私外洩
2. 寫程式顧問(Coding Advisor)
3. 調整System Prompt/Model參數做特殊應用: 法務文件判別、製造業零件的資料表判讀

RAG - 知識庫



1. 技術客服: 操作手冊回覆並附上出處
2. 新人訓練: 加速新人上手效率
3. 業務/銷售支援: 該客戶適用哪個公司產品、有無對應相關成功案例?

Agentic AI



1. 企業內部流程自動化: 費用報銷審核、工單自動分派、員工請假
2. 客戶服務端到端處理: 客戶反映訂單問題, **主動查訂單狀態、判斷是否符合退換貨政策**
3. 外部服務AI客服: 根據客戶情況是查閱RAG亦或是開立Support Ticket