



NUTANIX

全面解放 Agentic AI 以 Nutanix 打造現代化 AI 企業營運

KubeSummit 2026

Stanley Yeh | Senior Systems Engineer

Agentic AI 的下一個瓶頸不是模型，而是營運

Agent 從回答問題進化為替企業採取行動。

當每一次行動都會讀資料、呼叫工具、使用模型，平台就必須同時治理能力、成本與風險。



多輪 reasoning × token 消耗 × 寫入動作

DATA SOVEREIGNTY

資料該留在哪裡？誰能取用？

COST

哪一個模型，值得每一次呼叫？

RISK

工具與寫入權限如何收斂？

OPERATIONS

大量的 Agent，誰在觀測與維運？

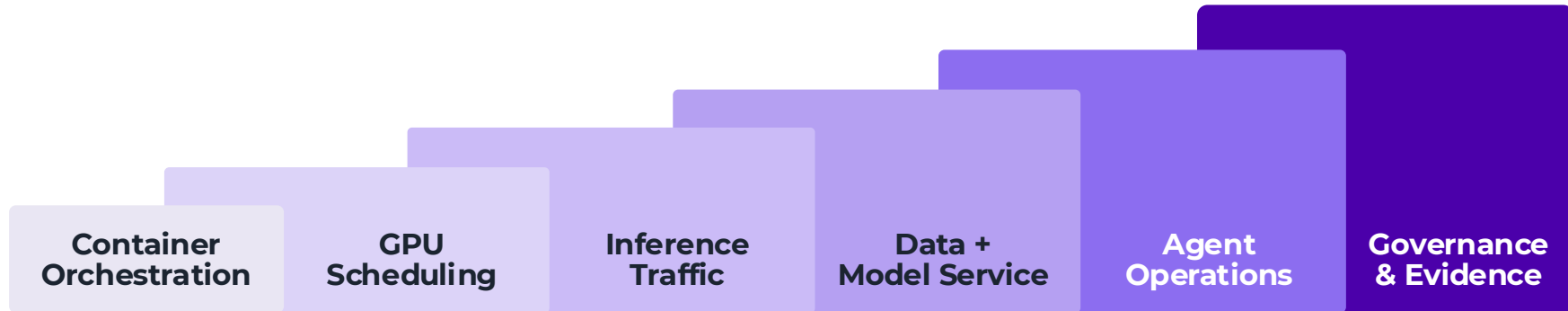
全面解放 Agent，不代表全面放任 Agent!

| Kubernetes 正從容器底座升級為 AI-Native 營運層

Cloud Native 管理部署與運行；AI-Native 還要管理模型、GPU、資料與 Agent 行為

從「把 **workload** 跑起來」到「讓 **AI workload** 長期可營運」

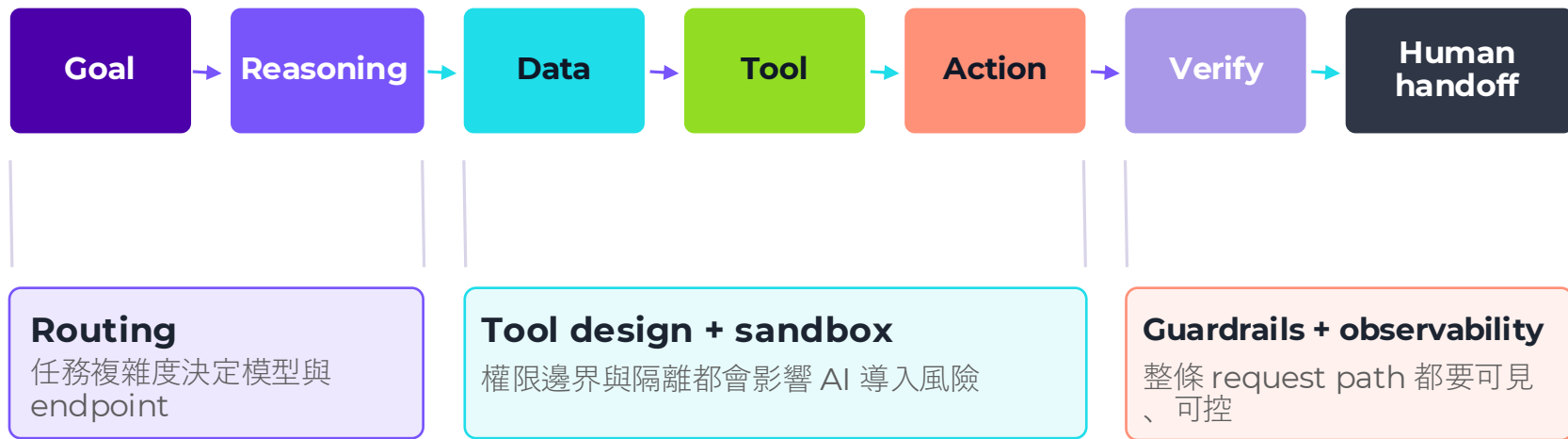
- Cloud Native : application deployment + operations
- AI-Native : models、GPUs、inference traffic、data、Agent behavior
- Kubernetes : 逐步成為 AI compute 與 model-service 的 operating layer



Agent 把 Inference API Request 變成需要治理的工作流程

動態選擇 reasoning steps 與 tools ; 每一步都可能改變成本、風險與結果

一個 goal, 可以展開成多次模型呼叫、多次 MCP tool call、資料讀取與寫入動作



參考：Anthropic public guidance on effective and trustworthy agents

Nutanix 的平台 DNA, 正好對應 AI-Native 的複雜度

把基礎設施的複雜度收斂成一致的 control plane

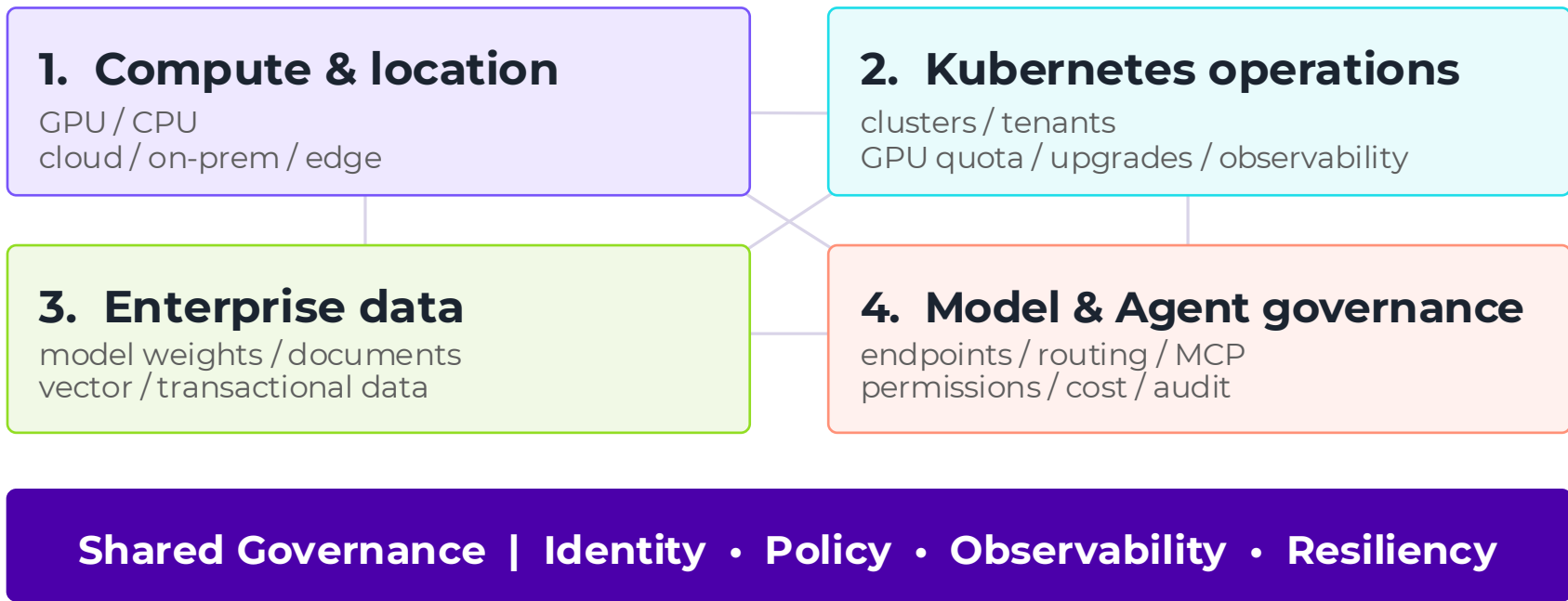
讓企業安全、彈性、可預期地使用模型的平台化基礎設施



平台價值：減少基礎設施複雜度，讓團隊把時間花在 Agent outcomes。

| 企業 Agentic AI 需要四個彼此連動的控制面

四個面向缺一不可；把資源、平台、資料與行為納入一個共同的治理框架

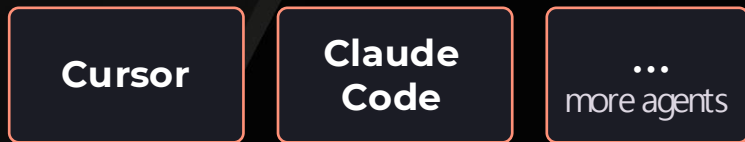


Nutanix 內部 coding-agent platform 案例：先解自己的問題

用平台解決方案回應 token cost explosion 與 provider dependency

開發團隊擴大使用 AI 後，token cost 與 provider 依賴變成營運問題

BEFORE



cost ↑ provider dependency ↑

coding-agent bills 太高，因此開始調整 provider 與 model mix

DIRECTION



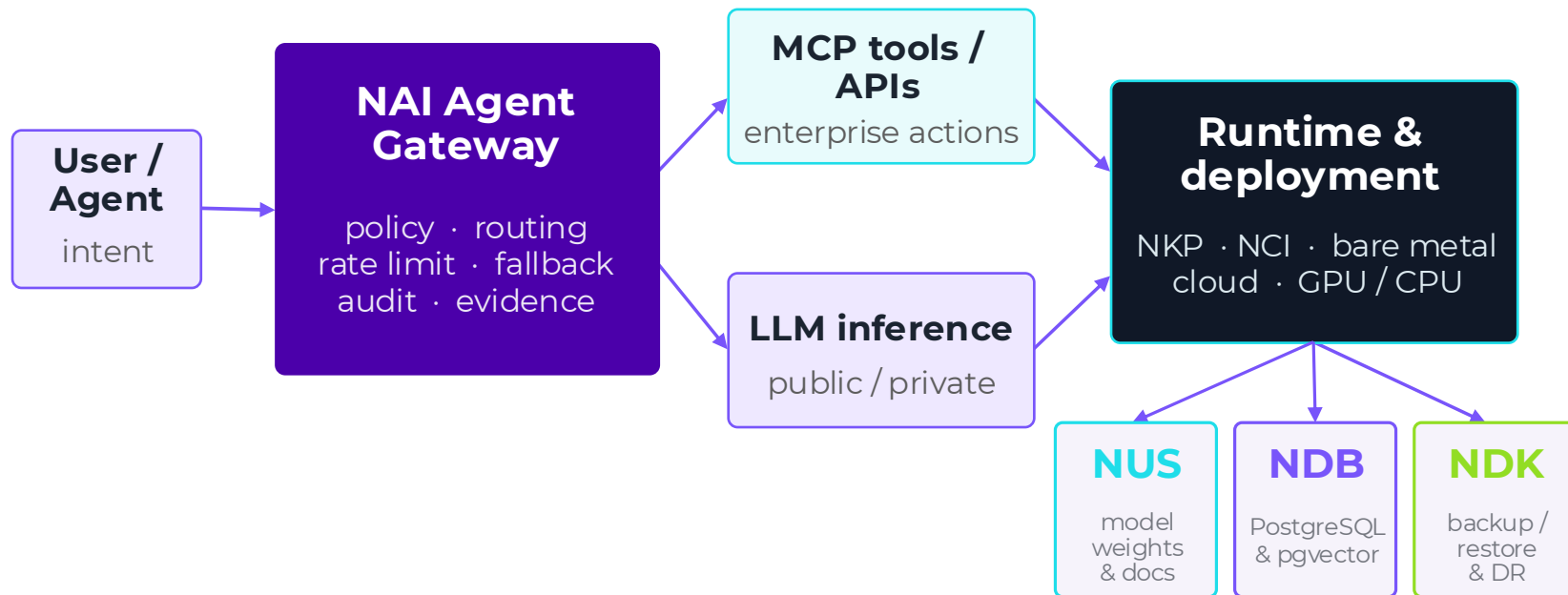
**model routing by
task complexity + cost**

約 80% 的 coding prompts 可以透過 high-quality open-weight models 處理

AI Sovereignty and Cost Control

| 在 Nutanix AI Platform 上放心處理每個 Agent Request

從 Agent 到 model、tool、data 的全方面治理



資料、模型與系統韌性，決定 Agent 能否被長期信任！

| AI Factory 成果、內部實驗、專案目標與 benchmark

ACTION / 已採取方向

provider + model mix

coding-agent bills 太高，因此調整供應商與模型組合

EXPERIMENT / 內部實驗

1.4–2.4x

透過 token-saving 的技術導入所實現的成本節省，best case 近 3x

TARGET / 專案目標

≥ 80%

多數推論在自己的基礎設施內完成；減少第三方 frontier-model 的支出

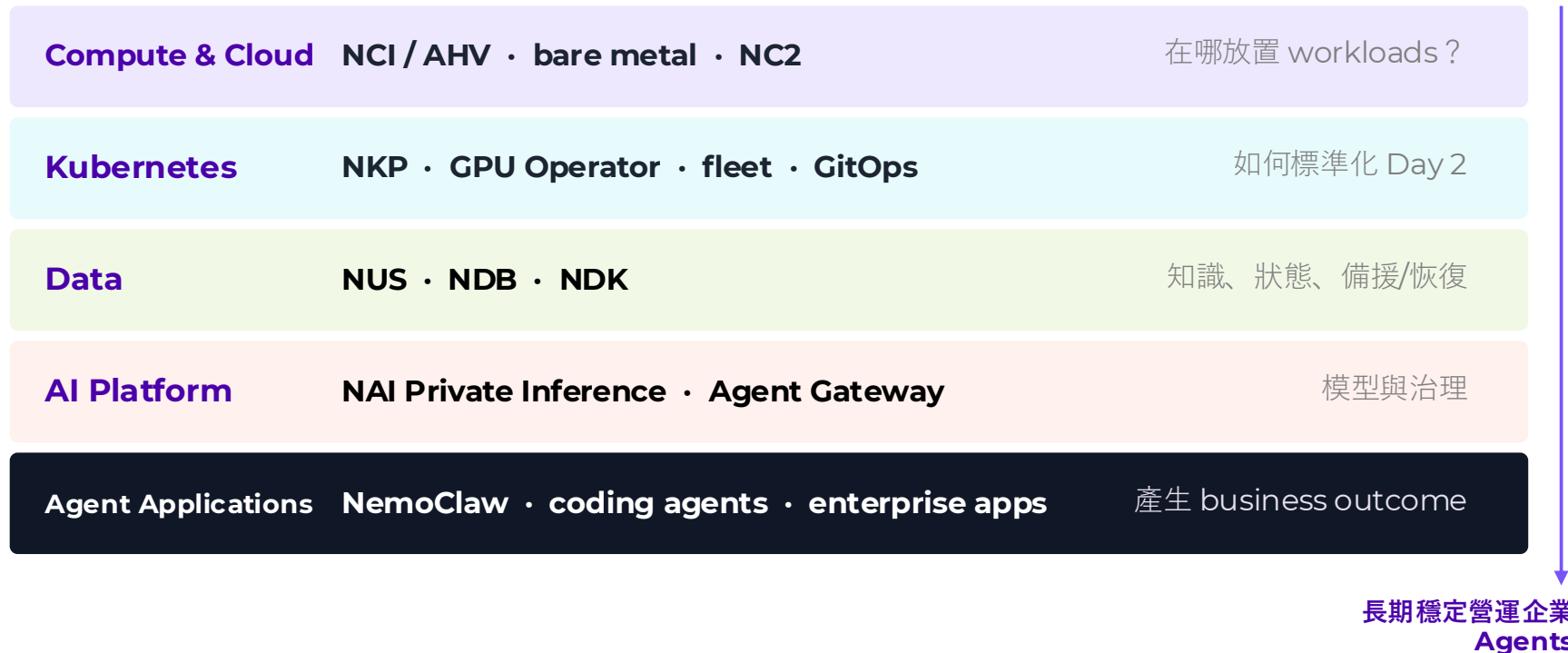
EXTERNAL TOKEN-SAVING BENCHMARK

36%+ / 25%

AI Gateway + caching 可達約 36%+ 的 token-saving；另一銀行案例約 25%

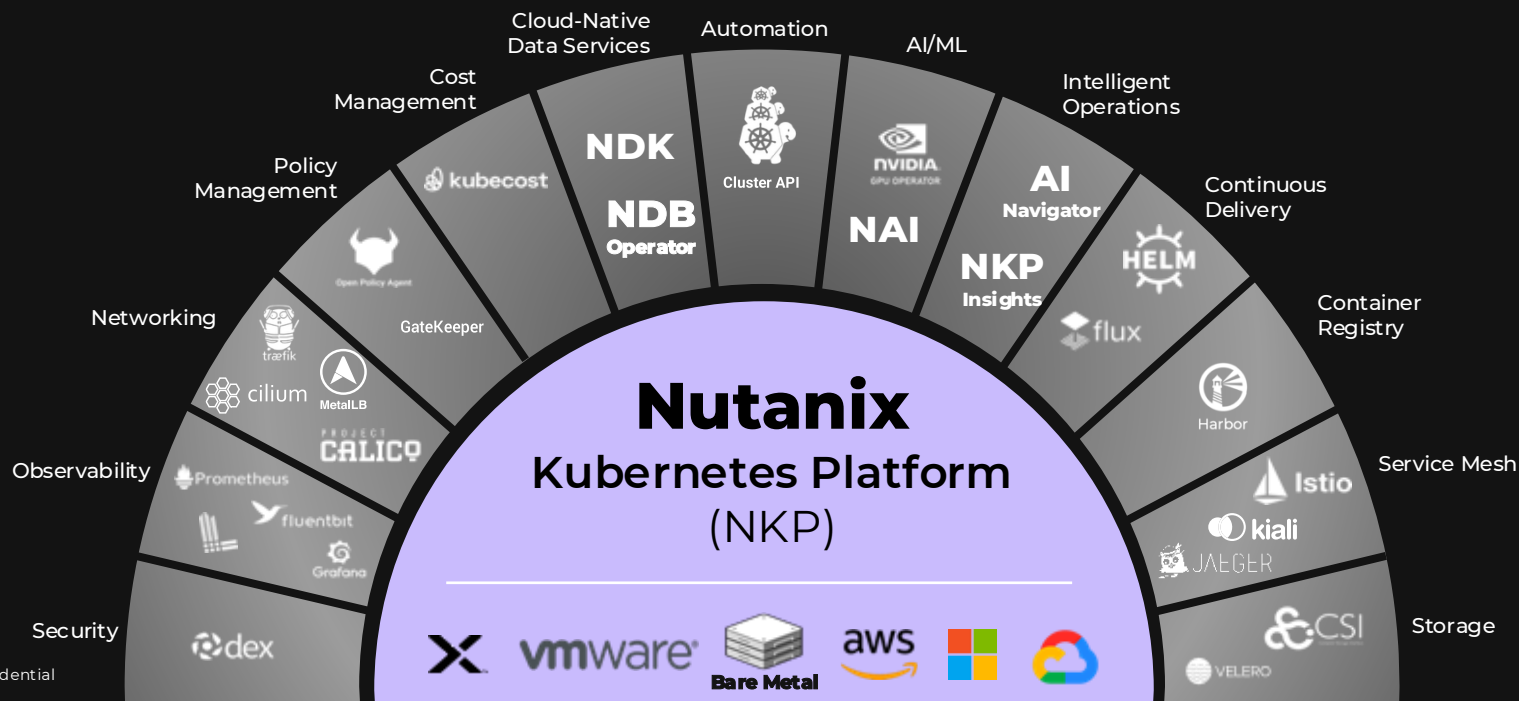
Nutanix 提供一條完整的 Agent production path

每一層的技术實踐都讓下一層的營運更容易



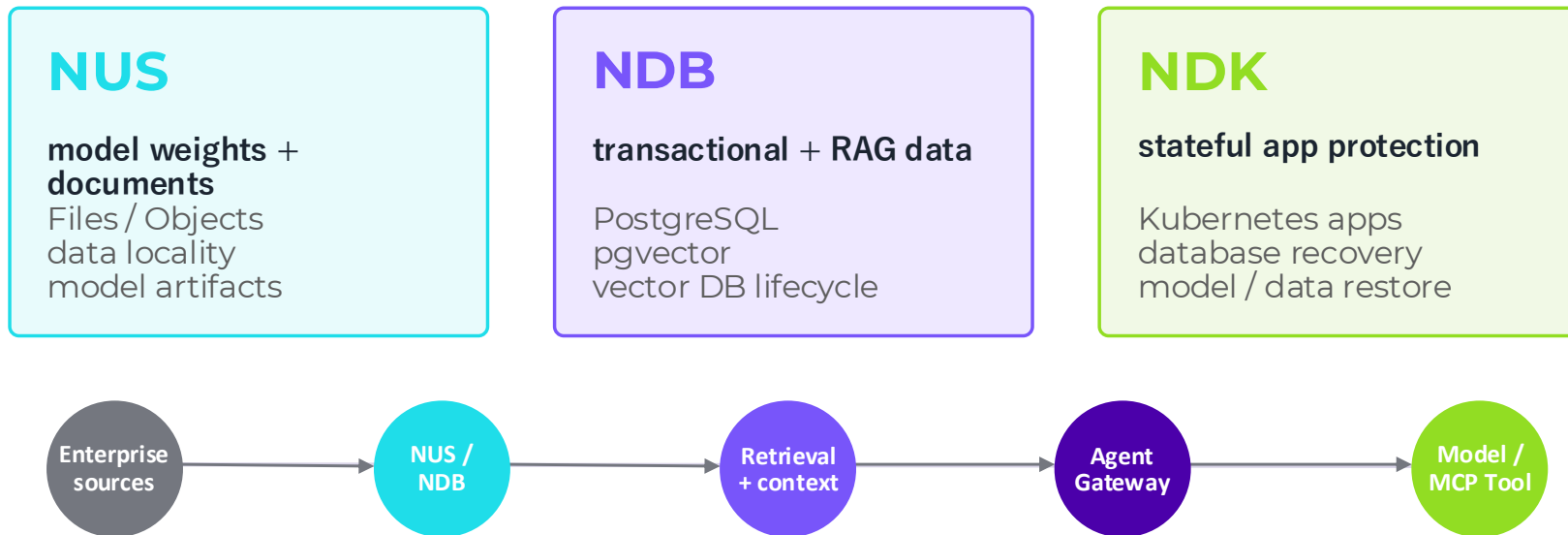
整合並掌握您的雲原生作業

部署、安全、管理與升級：
可規模化的企業級雲原生堆疊



資料控制面決定 Agent 能否可靠地使用企業知識

Nutanix 提供多樣資料解決方案協助維護資料品質



NAI = Agent Gateway + Private Inference

Start with Gateway → add Private Inference as sovereignty, cost, and usage grow.



Nutanix
Enterprise AI

Nutanix Agent Gateway

集中的控制層，用於治理、路由並確保 AI 代理 (Agents)、大型語言模型 (LLMs) 與企業工具之間的通訊安全。

Nutanix Private Inferencing

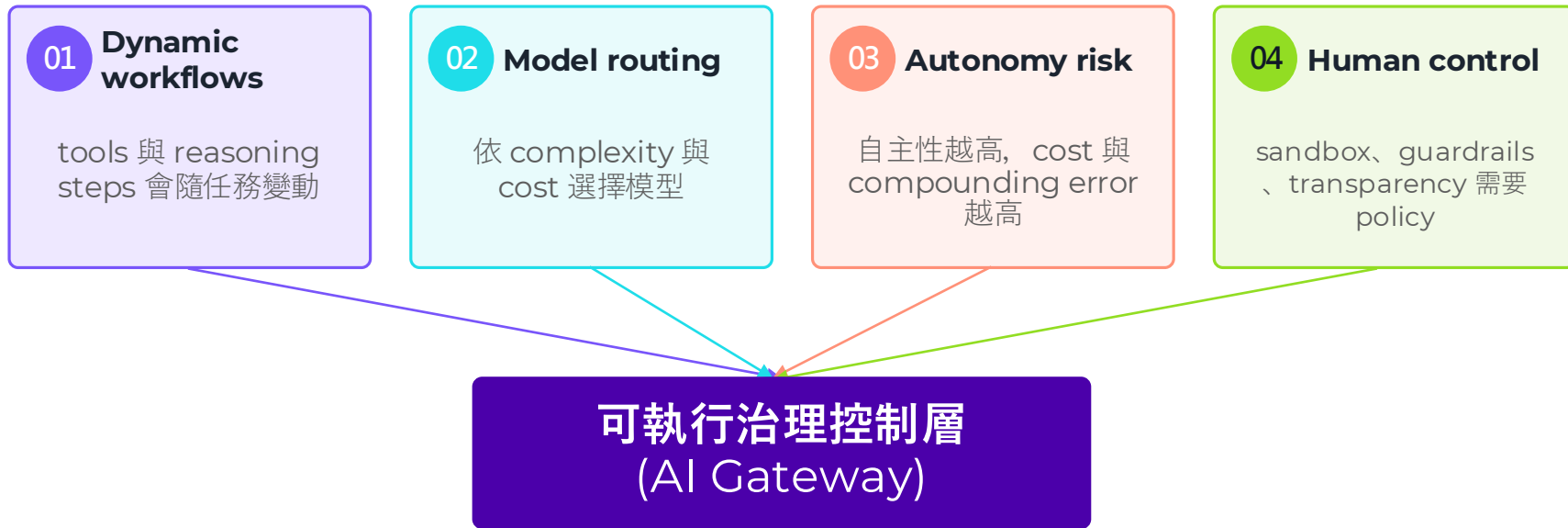
自託管模型的管理與高效推論。

Any CNCF Certified Kubernetes Distribution (Virtual or Bare Metal)

Run Anywhere - Edge, Core, Cloud, Neocloud

| 為什麼需要 Agent Gateway?

從 Anthropic 的 Agent 原則推導 Gateway 的必要性



NAI Agent Gateway 落實每個推論 request 的治理



一個統一的入口管理 agents、模型推論、MCP 工具及企業資料的使用

台灣工業電腦大廠 NemoClaw 案例，從 Agent 應用反推平台需求

跨 hardware、cloud、on-prem 的 AI platform

BUSINESS REQUIREMENTS

Fast Agent delivery

快速建立 enterprise Agents、RAG、multi-model workflows

Flexible execution

不同品牌的伺服器 / 彈性 GPU 算力分配、consistent operation

Cloud speed + control

如同 public cloud 的易用性；data sovereignty 與 cost control



PLATFORM REQUIREMENTS

Unified container platform

quota、multi-tenancy、可觀測的 shared compute

AI control plane

相同的 Agent Gateway / Private Inference story

Hybrid-ready operations

hardware、cloud、on-prem 的一致 Day 2 維運

NemoClaw 賦能員工加速使用 AI，放大了 platform operations 的重要性!

NemoClaw 案例的成果與效益

APPLICATION

NemoClaw & Various Agents

負能員工進行 Agent 與 skills 開發，鼓勵員工使用 AI 加速 AI 應用落地

AI PLATFORM

NAI

提供 Private Inference 及 Agent Gateway，實現 Model Routing & Failover

OPERATIONS

NKP

提供 Kubernetes，建立 GPU 算力池、落實 multi-tenancy 並簡化 Day 2 維運

INFRASTRUCTURE

NCI / AHV / NC2

提供支持 hybrid cloud 的 AI 基礎設施，簡化資料中心的 Day 2 維運

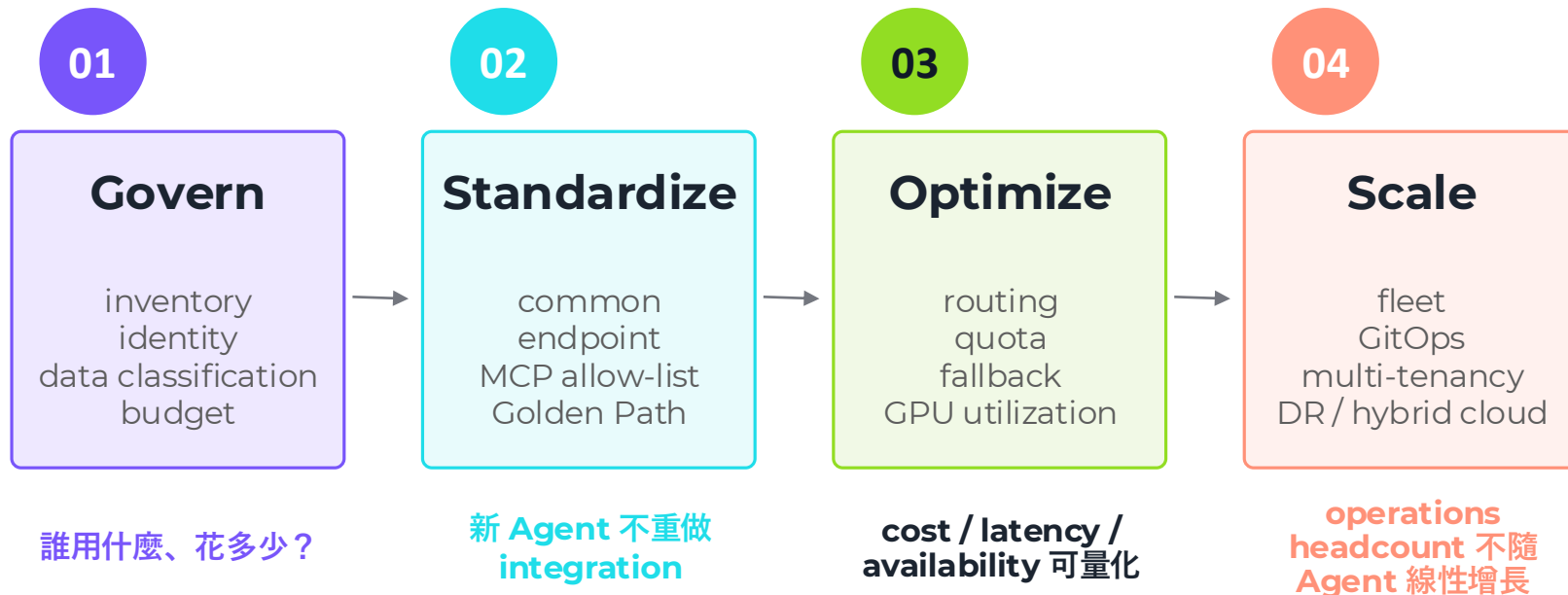
程式編寫流程
100 min → 30 min

地端 AI 同時在線使用人數
150+

實體伺服器部署與上線
months → one day

從單一 PoC 到 AI 企業營運，需要四階段導入

先建立可見性，再把共通路徑標準化；最後才是 fleet、GitOps、DR 與 hybrid scale



全面解放 Agent, 必須同時全面治理 Agent

從 model capability, 轉向 platform operations

- 1 Agentic AI 的瓶頸, 從 model capability 轉向 platform operations
- 2 企業需要 model-neutral、cloud-flexible、data-aware 的 AI control plane
- 3 Freedom 是讓 Agent 安全使用更多 models、data、tools；不是讓它無限存取一切

企業首先要建立的, 不是一個又一個的 AI agent, 而是一個能讓 **Agents** 安全行動、成本可控、資料留在正確位置, 並且可以長期營運的 AI 平台！



Ready for more?



Walk-through
of Nutanix
Enterprise AI



Test Drive
Nutanix
Enterprise
AI

The background of the image consists of a dark, textured surface with intricate, wavy, light-colored lines that resemble a topographical map or a complex network. In the center, there is a faint, large, four-pointed star or spark-like shape.

Thank you.

NUTANIX™