SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

Sangfor AICP

輕鬆釋放GPU算力, 軟件定義AI政務創新

Steven Lo | 方案顧問

Steven.lo@sangfor.com | 深信服科技

DeepSeek趨勢下AI進入普惠階段，但依然成本挑戰巨大

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

專業人才成本高

- AI能力要求高
- 供需失衡
- 行業競爭激烈
- 培養成本高

基礎設施成本高

- 硬體資源成本
- 供電成本高
- 平臺工具成本
- 開發運維成本

運維成本提高

- 資源分享管理難
- 版本管理難
- 視覺化監控難
- 故障定位難

從擁抱AI的一開始到大範圍使用都要考慮ROI

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

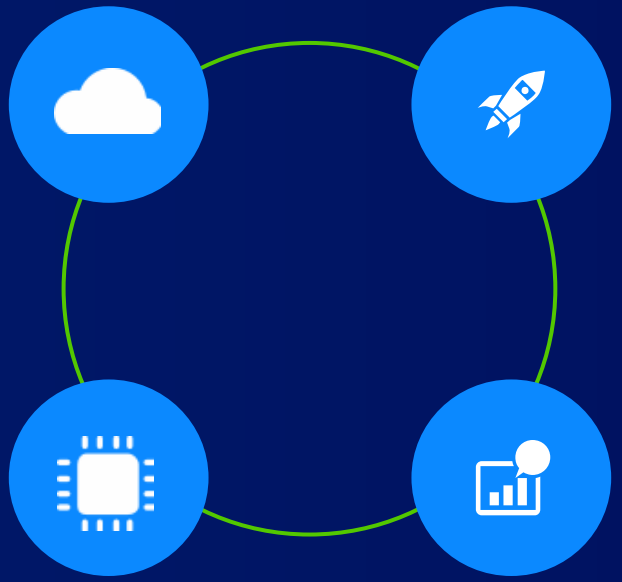
基於場景選擇模型，基於模型選擇顯卡，同時關注平臺性能優化用確定的顯卡發揮更大的效能

平臺性能優化

算力平臺不僅僅是一個資源管理器，更應該通過平臺自身的性能優化提升客戶推理性能

AI優化新技術

比如異構調度優化、昇騰優化、語義Cache等新技術的發展，大幅降低客戶的推理成本



基於訓練/推理以及信創要求選擇顯卡

結合自身的建設需求以及需要的模型大小，在4090D、H20、沐曦、昇騰、天數、等顯卡中進行選擇

根據場景選擇模型大小

以Deepseek為例，要基於使用場景選擇671B、70B、32B、14B、7B、1.5B

深信服長期佈局AI產業，團隊和技術能力積累扎實

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

深信服AI升級的解決方案：AICP 及 SF-FastGPT

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

AICP

私有雲獲取AI能力

RAG加速能力 | Deepseek模型服务

AI模型 AI模型 AI模型 AI模型

- 豐富的模型與應用
- 降低本地投入成本
- 平滑獲取 AI 能力

新一代超融合解決方案

傳統應用 | AI 應用

llama | qwen | deepseek

深信服 HCI + AICP GPU 節點

- 異構算力管理
- 性價比提升4倍
- 相容各類硬體和最新模型

承載場景

SF-FastGPT

承載AI模型 → 輔助應用創新

數據清洗 | 内容提取 | 概要生成 | QA對

知識問答 | 辦公助手 | 智能客服 | .....

應用發佈

- 應用構建極致簡單
- 應用效果評估和輔助建議，降低調優門檻
- 智慧化資料運營

應用構建場景

深信服AI平台用户

深信服新一代AICP平臺架構

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

AI應用場景

文案生成助手	企業知識問答助手	商機分析助手	AI教學品質評估	AI電子卷宗	醫護copilot
對話式應用改造	智能客服助手	經營分析助手	辦公智慧一體化助手	宣發助手	AI智能導診

常用模型

Deepseek　QWQ/Qwen　LLAMA　各類用途的小模型

向上承載各類大小模型，服務好客戶的業務應用

AICP 算力平台

全場景全模型低成本起步

AI起步階段：671B低成本起步
GPU顯卡異構納管統一演進
自訂模型服務、最佳實踐一鍵發佈

性能卓越

大參數量MOE性能優化
稠密模型性能優化
昇騰優化、語義cache

GPU資源利用率提升

vGPU、多實例智慧負載2.0
異構算力調度優化
模型服務API多Key精細運營

立體安全能力

大模型加密
AI內容安全
AI閘道負載

向下開放相容，廣泛適配並調優、調度各硬體廠商硬體基礎設施

主流硬件

NVIDIA	沐曦 MetaX	天数智芯	Ascend	intel	AMD	HYGON	Mellanox
國際廠商GPU	國內廠商NPU	國內廠商GPU	國內廠商NPU	CPU	CPU	國內廠商CPU	Network

智慧融合架構，降低複雜性，服務化的提供基礎設施能力

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

應用1：研發部門
Constrict
（日常使用）

應用2：IT部門
智能客服問答
（小規模試點）

應用3：業務部門A
報告解讀業務應用
（開發測試）

...

應用N：業務部門N
XXXXX應用
（生產業務）

新AI建設範式

AI閘道

智能融合架構（SFA）

模型倉庫 | API多KEY | 用量觀測 | Open API | vGPU | 大模型/自訂模型服務

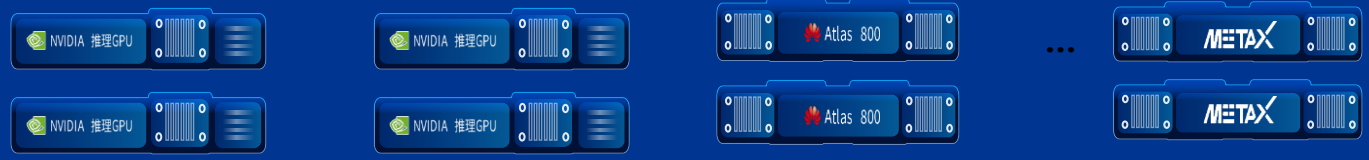
多模型快速上線

自我調整硬體遮罩層 Smart HAS

監控告警 | Device-Plugin調度 | **資源調度** | 集群管理 | vGPU切分

顯卡異構統一調度管理

Nvidia高端顯卡 | Nvidia中低端顯卡 | 國產卡A | 國產卡N



資源精細化運營與保障

算力基礎設施服務化，多模型、多卡相容不綁定硬體

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

deepseek Qwen-VL Embedding Rerank OCR ……

緊跟趨勢，不斷相容適配更多卡商、各類開源模型、多模態模型與微調模型

不斷適配調優更多型號

NVIDIA			METAX 沐曦	天数智芯	HYGON	Ascend
Nvidia A100	Nvidia 4090D	Nvidia L40S	N260	智鎧MR V100	K100-AI	910B3
Nvidia A800	Nvidia 5090D	Nvidia A6000	C500	天垓BI 150	……	910B4
Nvidia 4090	Nvidia L40	Nvidia H100	C550	……		300I DUO PRO
Nvidia H20	Nvidia L20	……	……			……

深信服 AICP 算力平台

- 英偉達顯卡 **快速相容**
- 新品牌型號顯卡 **1個月完成適配**
- 擁抱開源 **最新推理框架即時支持**
- 異構算力及模型資源 **統一管理**
- 異構算力資源 **統一調度與性能優化**
- OS、AICP軟體、容器平臺 **層層解耦**

vGPU大幅提升顯卡資源利用率，承載更多AI Agent應用

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

RAG、智能客服等AI應用建設
都需要大量小模型

vGPU

支持算力1%，
256MB級顯存資源切分

運行Embedding+Rerank+OCR
小模型場景

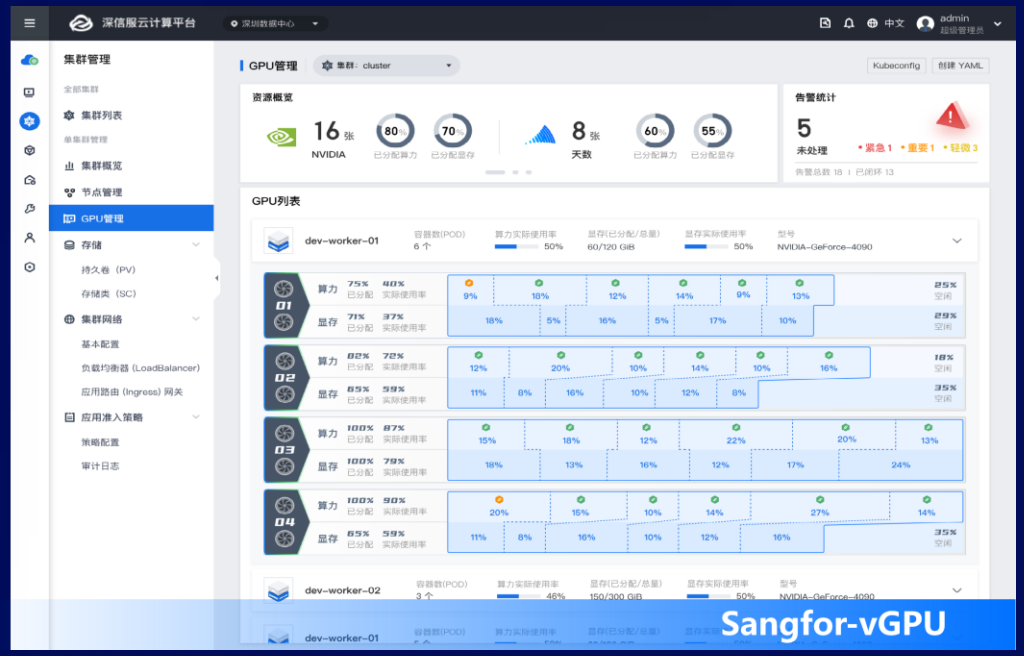
TCO分析

顯卡消耗

顯卡越高端

模型使用越多

越省錢



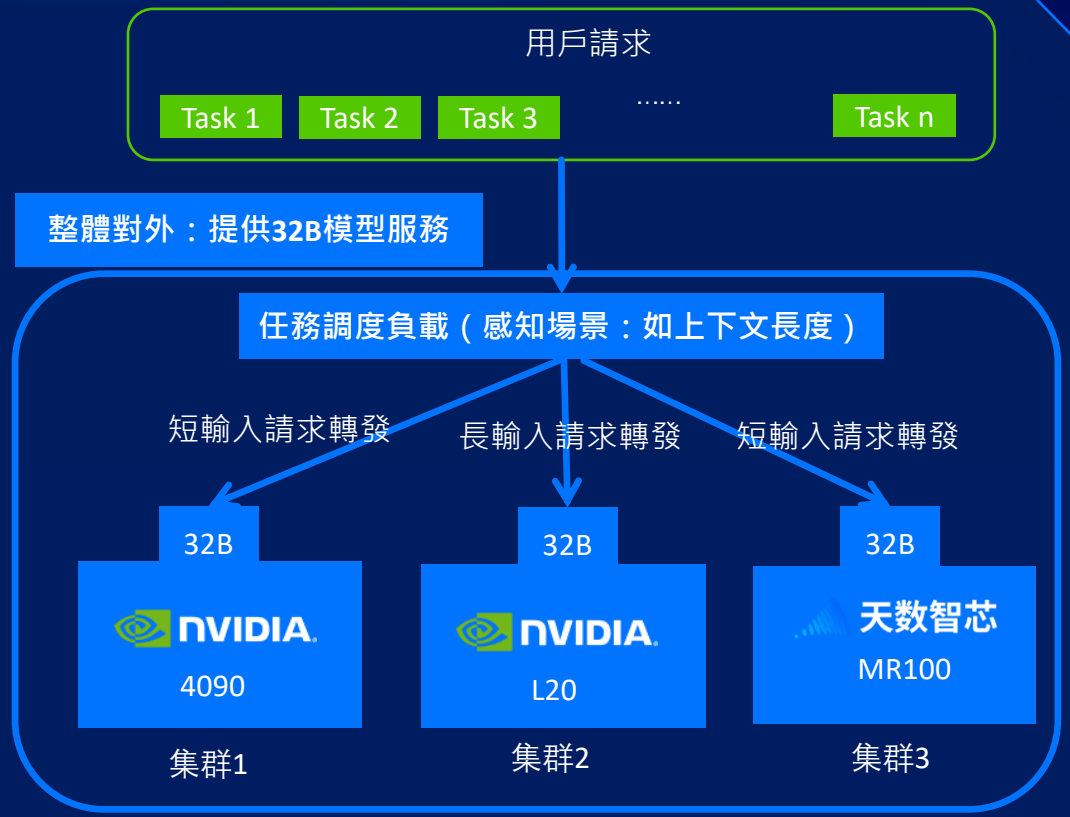
	未切分	vGPU切分
4090D	3張卡	1～2張卡
H20	3張卡	1～2張卡
L20	3張卡	1～2張卡

單卡承載模型數量翻8倍+

現有資源利用率同比3倍+

異構算力調度優化與精細化運營大幅提升顯卡資源利用率

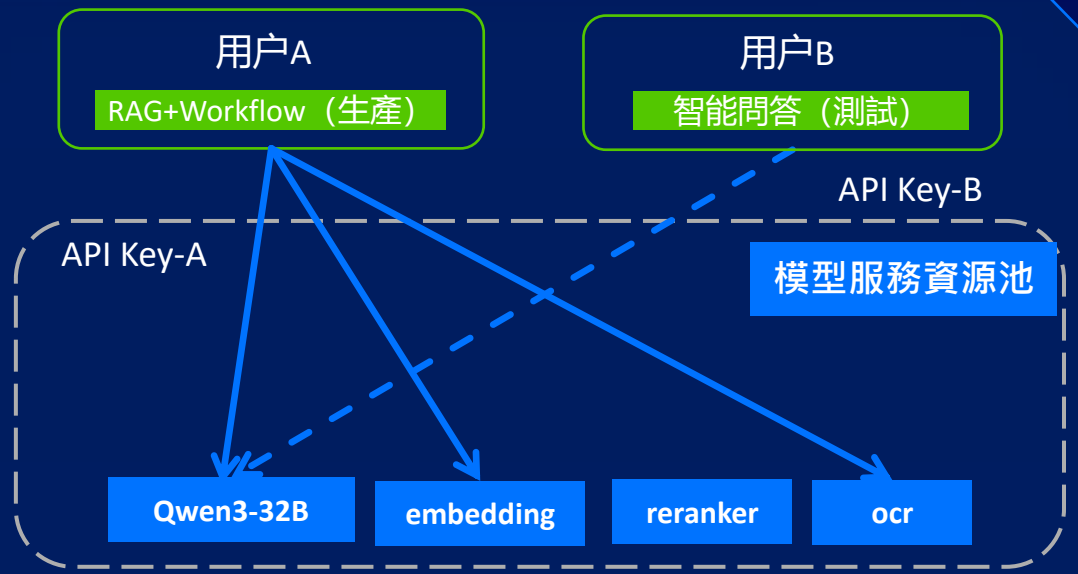
01 自動化的異構推理集群調度



感知應用場景和負載情況，自動選擇最優節點，提升整體資源利用率。

開發中，2026年Q2發佈

02 多API KEY高效利用顯卡資源，保障關鍵業務體驗



- 通過API Key為不同使用者分配不同模型服務。
- 優先保障關鍵用戶的業務穩定運行，非關鍵用戶限流（如併發數、吞吐、tokens等）
- 監控使用者資源使用情況如基於tokens、併發等，優化API Key分配規則

通過底層核心技術創新，發揮2~5倍智能算力設施效能

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

重載應用場景

創新優化啟動智慧算力效能2~5倍起步

H20	4090D-48G	910B3
單卡平均承載用戶數 2x倍提升	p99 TTFT 6.4x倍提升	總承載用戶數 2.25x倍提升
ROI： 2x倍提升	Decode速度 2.0x 倍提升	ROI： 2.0x倍提升
	p99 E2E Latency 2.0x倍提升	

模型： GLM4.6（H20）、GLM4.6-Air（4090D-48G、910B3）；

SLO標準： p99 TTFT<10s，p99端到端時延 <= 60秒；

推理框架： 相比sglang最新社區版本

通過關鍵技術創新，不斷提升端到端ROI

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

AI應用

AI 智能體　AI 智能客服　衛甯健康　衡石　派兹　鼎捷軟體　華東醫為　…

自我調整架構層Smart Arc

自我調整模組

- 自我調整多版本推理引擎
- 自我調整多類型推理引擎
- 性能瓶頸分析
- 自我調整調整
- 回流數據觀測
- 一鍵最佳實踐配置
- 指標採集
- 開源推理框架/自研推理框架：SIF（Sangfor Inference Framework）--加密、可靠優化、性能優化

自我調整原子能力

- AI 閘道 智能負載調度、限流、計量
- 基於檢索的投機解碼優化
- 算力資源自我調整 MOE運算元優化
- 基於場景化的分塊調度優化
- GPU間、節點間資料傳輸優化
- 基於場景語義感知的Cache智慧負載
- 長輸入性能優化
- 昇騰優化
- ……

全模型全場景低成本起步

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

671B低門檻起步：當前業界唯一商業化，4090D起步運行含671B的端到端AI應用場景

智慧問答應用 | RAG知識庫 | … | 安全GPT | ISV生態AI應用 | … | Embeding+Rerank+P2T等自訂大小模型0DAY適配

Deepseek 671B模型 + QWQ-32B模型 + X个小模型

混合部署：結合MOE模型特點+vGPU，將模型服務合理部署與顯存和記憶體

【方案1】全量8卡：8併發80tokens

【方案2】4+4：DeepSeek：4卡8併發80tokens；QWQ 32B：4卡100+併發1000+tokens

【方案3】4+2+X：DeepSeek：4卡8併發80tokens；QWQ 32B：2卡；X小模型2卡

其他價值優勢

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

廣泛ISV生態適配調優

中科闻歌 | 衡石 | 行云创新 CLOUD·TOGO | 鼎捷软件 DigiwinSoft

私有雲AI升級

- HCI加1台GPU節點即可輕鬆獲取AI能力
- AI能力與線上體驗一致

模型訓練一鍵微調

- 嚮導化配置一鍵微調

模型倉庫

- 支持DeepSeek等開源大模型統一運維管理上傳管理
- 支援從外部上傳自訂大小模型運維管理
- 通用場景、長輸入場景最佳實踐，一鍵發佈模型服務

構建立體AI安全能力

- 模型加密
- 數據內容安全

模型倉庫讓各類大小模型運維管理更簡單省心

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

開源模型

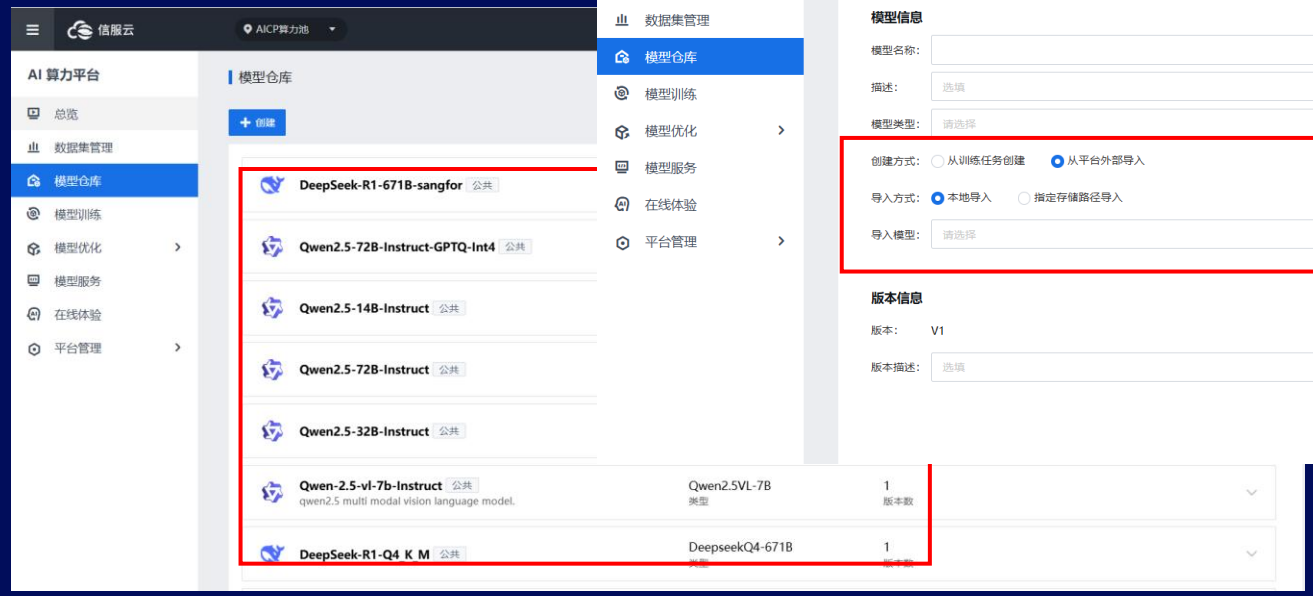
支持DeepSeek等開源大模型**統一運維管理**上傳管理

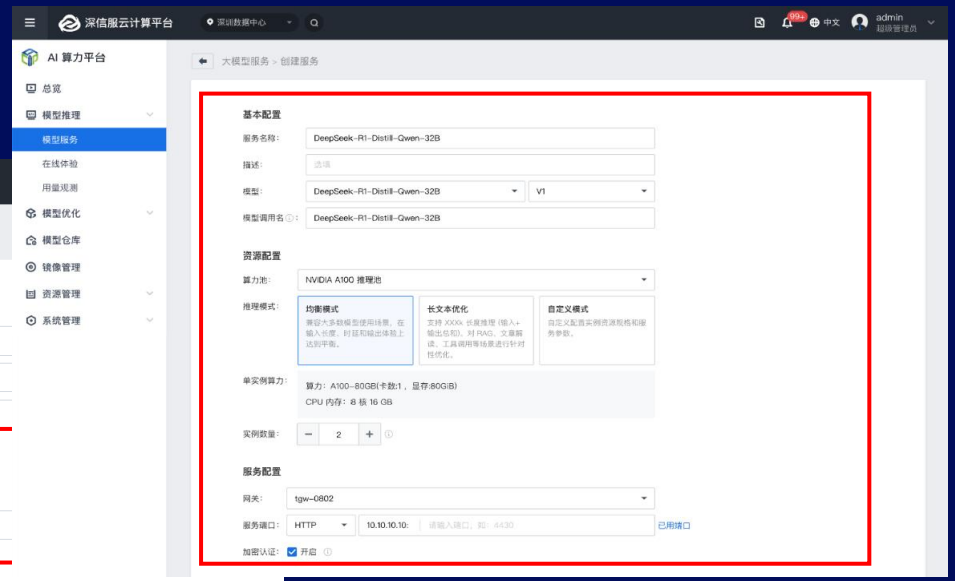
自訂大小模型

支持從外部上傳最新、自訂大小模型運維管理

一鍵發佈模型服務

無需配置算力，一鍵發佈均衡、長文本、自訂模型服務

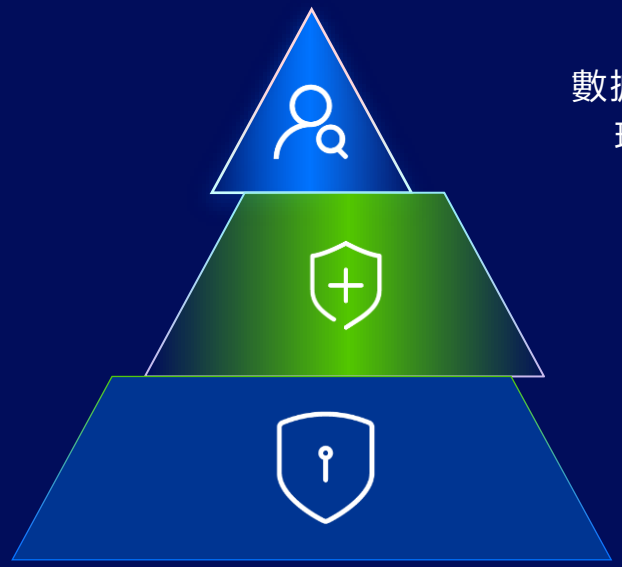




立體安全能力，構築AI時代新安全體系

大模型自身安全

業界首推模型動態加密，確保模型與私有訓練內容不被竊取使用



環境安全

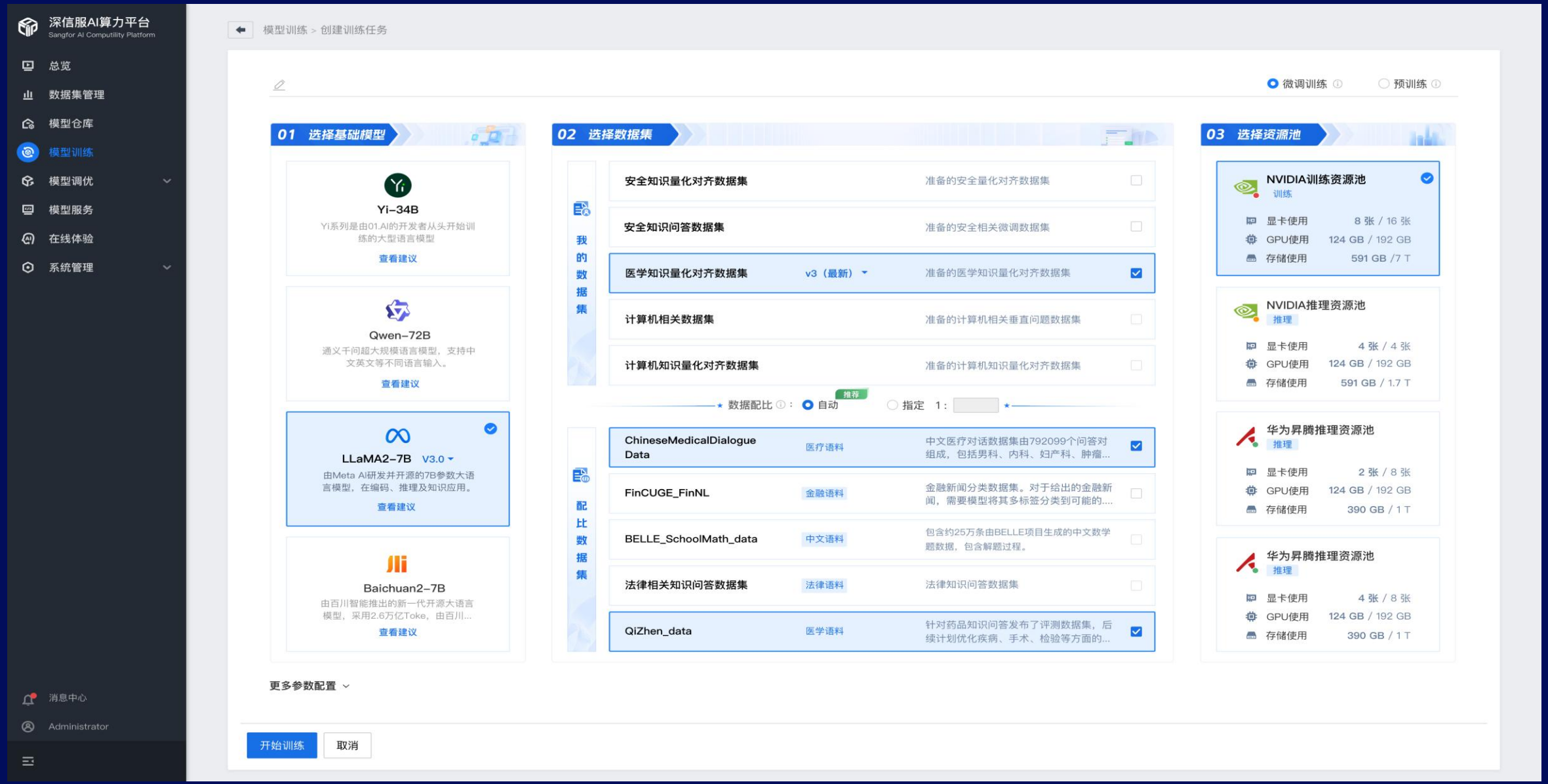
數據當地語系化，線上專屬雲AI環境，保障內網隱私資料安全

資料與參數內容安全

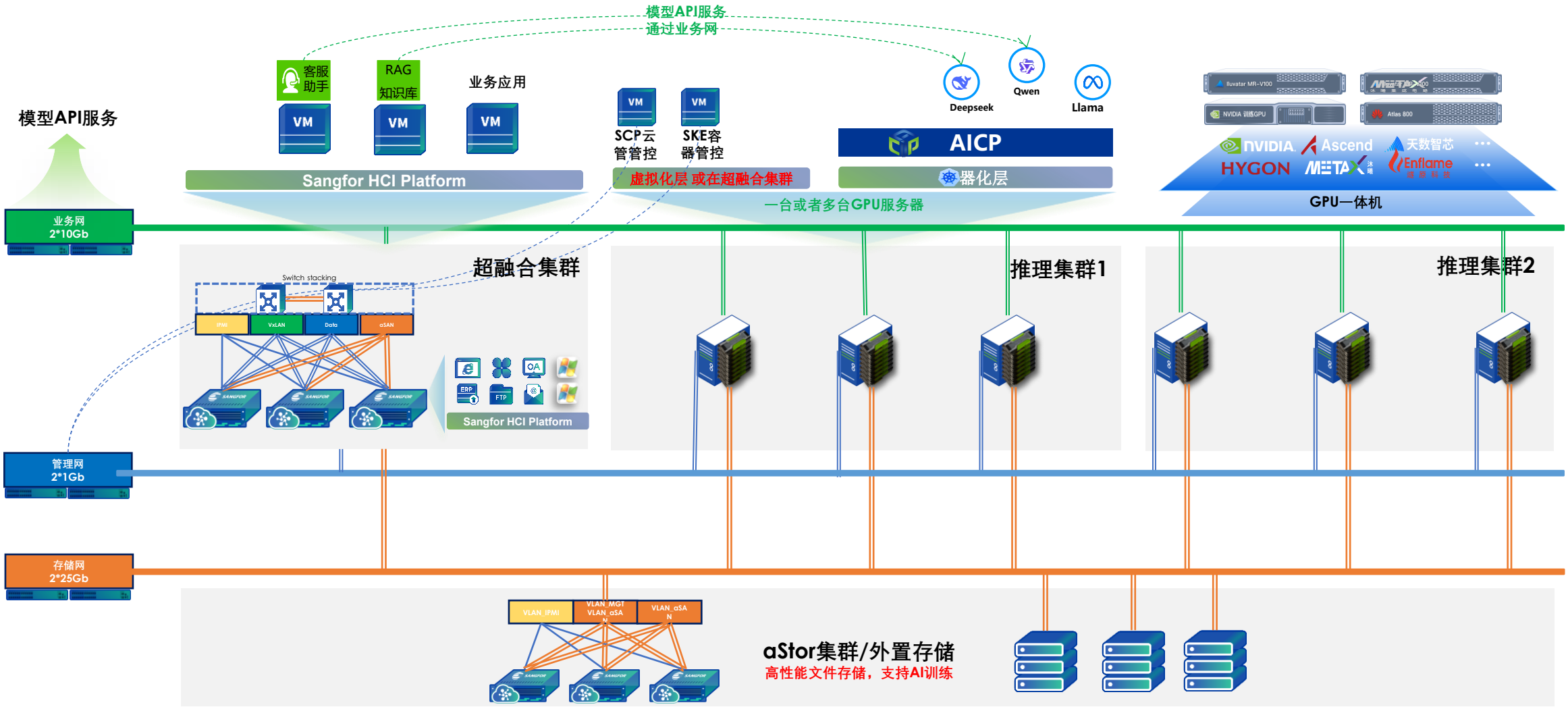
- 資料安全平臺：大模型應用安全護欄，實現提示注入攻擊防護以及有害或敏感資訊輸出的安全防護

開發中，2026年Q2發佈

模型訓練嚮導化配置，一鍵微調提升AI生產效率



AICP + HCI 架構圖


模型API服务
通过业务网
模型API服务
客服助手
RAG知识库
业务应用
VM
VM
VM
SCP云管管控
SKE容器管控
VM
VM
Deepseek
Qwen
Llama
AICP
Sangfor HCI Platform
虚拟化层 或在超融合集群
器化层
一台或者多台GPU服务器
NVIDIA
Ascend
天数智芯
HYGON
METAX
Enflame
GPU一体机
业务网
2*10Gb
超融合集群
Switch stacking
IPMI
VxLAN
Data
oSAN
ERP
FTP
Sangfor HCI Platform
推理集群1
推理集群2
管理网
2*1Gb
存储网
2*25Gb
VLAN_IPMI
VLAN_MGT
VLAN_oSAN
VLAN_oSAN
aStor集群/外置存储
高性能文件存储，支持AI训练

關鍵結論

AICP 能力要點

- 異構算力抽象與池化：多模型、多卡相容且不綁定硬體，持續適配 NVIDIA 與多家國產加速卡並統一調度與優化。
- 服務化接口與治理：每模型提供 API Gateway 與 API Key，支援按 Key 的限流與配額，保障多部門公平與穩定。
- 私有雲部署與全鏈路支持：主流/自有模型託管、資料版本化、單介面微調訓練、推理解耦優化（量化、蒸餾、KV cache）、模型加密。

成本與效能

- 藉由調度、混部與推理優化，在多款 GPU 上可實現約 2–5 倍效能提升，支撐高併發與長文本等重載場景，同時提升資源利用率與能效。
- 小模型（Embedding、Rerank、OCR 等）可使用 vGPU 資源分割能力，避免整卡佔用造成浪費。

完善生態協作

- 與資深 ISV 夥伴深度整合，提供從諮詢到運維的全流程支援

彈性起步，平滑擴展

- 最低僅需 1 個 HCI 節點 + 1 個 GPU 節點即可啟動
- 支援橫向擴展，達到公有雲級別的服務體驗

一站式開發平台

- 一鍵微調功能，快速適配業務場景
- 模型倉庫支援主流開源模型與自訂模型導入
- 覆蓋數據接入、訓練、管理、壓縮、加密到推理的完整生命週期

企業級安全合規

- 內建模型加密與數據安全機制
- 私有化部署，符合國際標準與行業合規要求

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

謝謝聆聽

讓每個用戶的數位化更簡單、更安全！

www.sangfor.com