

深信服AI基础设施解决方案

深信服云BG

www.sangfor.com.cn

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

AI基础设施发展趋势及挑战



AI转型的本质：重构企业生产方式

不是增加一个工具，而是从“人操作系统”走向“人设目标、Agent完成任务”

三层跃迁路径

搬到线上 → AI辅助 → 目标驱动自主执行

1 第一阶段｜数字化

把线下流程搬到线上

数字化｜人操作系统

流程从线下迁移到线上，形成标准化、可管理的数字流程。

核心能力

- 流程线上化
- 规则标准化
- 数据结构化
- 操作自动化
- 管理可视化

2 第二阶段｜AI化

在原有流程中增加AI能力

AI化｜AI辅助人操作系统

保留原有流程，由Copilot提供认知与分析能力。

核心能力

- 智能问答 / 内容生成
- 数据分析 / 知识检索
- 辅助判断
- Copilot辅助操作

3 第三阶段｜AI Native

围绕AI能力重新设计流程

AI Native｜Agent自主执行

人设定目标，Agent编排系统与工具完成任务。

核心能力

- 目标驱动 / Agent自主执行
- 动态任务编排
- 多Agent协作
- 系统自动调用
- 持续反馈与进化

AI化是在原有流程中增加AI；AI Native是以目标为起点，重新设计流程、组织与系统。

中大型组织的Agent开始转向规模化生产实践

业务目标、核心矛盾与技术底座同步演进

AI Native企业

01 前期｜创新探索期

点状创新

核心目标 激活使用与创新

核心矛盾 创新不足 vs. 创新失控

管理抓手 AI文化 · 工作坊 · 快速Demo

Agent形态｜自定义配置 Agent

02 中期｜规模发展期

平台规模化

核心目标 统一基建，复制高价值场景

核心矛盾 需求快速增长 vs. 平台治理滞后

管理抓手 网关 · Agent平台 · 数据知识 · 权限审计

Agent形态｜定制 + 第三方 + 平台内 Agent

03 后期｜深度变革期

企业生产方式重构

核心目标 优化成本，重构流程、岗位与组织

核心矛盾 技术已经可行 vs. 组织机制阻力

管理抓手 语义路由 · 性能优化 · 岗位融合

Agent形态｜多 Agent 协同 / 长时复杂工作闭环

Agent技术架构

能力由左向右累积，
不是简单替换

Prompt	Context	Harness	Loop
提示驱动	上下文驱动	编排治理驱动	多Agent协同
单点智能	语境智能	系统智能	复杂任务自主闭环

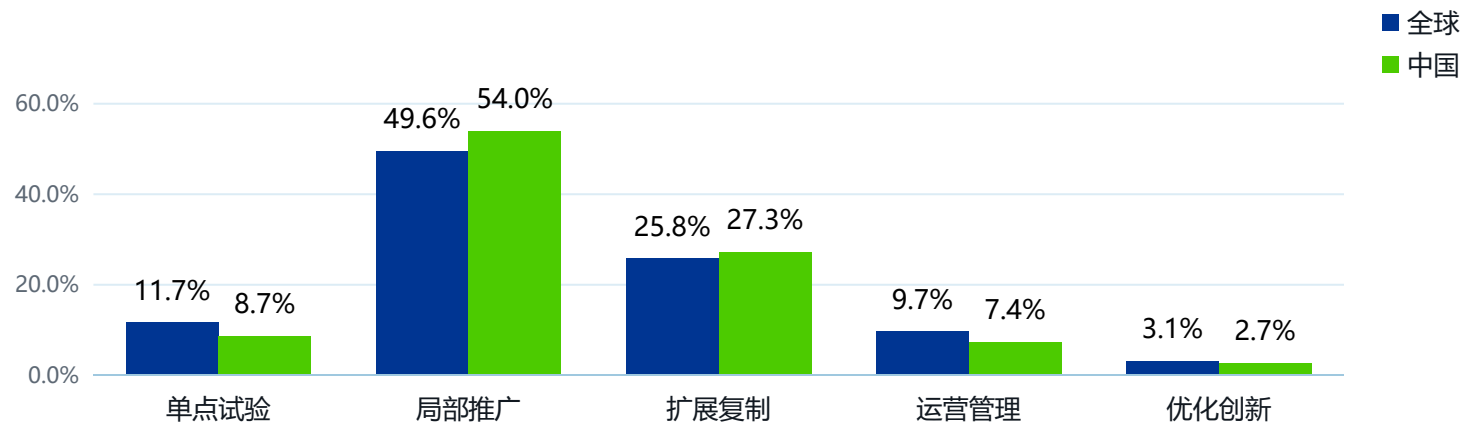
演进逻辑：先让AI可用 → 再让Agent可复制、可治理 → 最终让复杂业务自主闭环

目标｜AI Native企业

AI应用加速进入生产，成熟度仍处爬坡期

全球与中国的应用重心集中在局部推广和扩展复制，优化创新仍为少数

全球与中国AI应用成熟度分布（2026）



全球样本：IDC 1,900家组织 / 20个市场；中国数据据用户提供的IDC材料整理

关键判断

规模复制成为主战场

81.3% 中国集中于局部推广 / 扩展复制

61.3% 全球仍位于前两个成熟度阶段

优化创新仅占 3.1%（全球）/ 2.7%（中国）

AI正在重塑产业与组织的八个维度

从效率工具延伸至业务、组织与商业模式

01 生产工具重塑
视觉质检提速3倍

02 业务流程重塑
贷款审批缩短至10分钟

03 用户体验重塑
推荐转化率提升40%

04 产品服务重塑
肺结节识别准确率>95%

05 生态伙伴重塑
供应链伙伴协同生产

06 决策模式重塑
预测驱动库存与促销

07 员工组织重塑
数字员工与人机协同KPI

08 商业模式重塑
AI即服务、按结果收费

资料来源：IDC MaturityScape Benchmark: AI-Fueled Organization Worldwide, 2026；中国数据及案例据用户提供的IDC材料整理

AI规模化进生产的挑战：运行条件发生根本变化

试点Demo只是冰山一角，真正的挑战隐藏在动态、复杂、持续运行的生产环境中

试点Demo

可见的成功



规模化生产

水下的系统性挑战

为什么Demo容易成功?

- ✓ 数据由人工准备
- ✓ 用户容错度高
- ✓ 调用量较小
- ✓ 场景边界清晰
- ✓ 权限相对简单
- ✓ 效果靠专家调试

为什么生产运行挑战大?

01 AI基础设施就绪

硬件、模型、工具链贯穿建设与运维；
资源协同就绪与持续运维压力增加。

02 Agent开发质量与效果运营

场景复杂度上升，Agent效果易衰减；
“80分”难达生产要求，需持续评测优化。

03 数据就绪

数据复杂且实时变化，Agent需跨系统取数；
历史数据孤岛阻碍连接与有效使用。

04 业务连续性

模型、工具或系统异常可能中断业务；
需超时重试、降级切换与人工接管。

05 算力持续投入与成本管控

多算力、多场景消耗难穿透、难归因；
缺乏统一调度，难判断投入是否合理。

06 安全性保障

模型黑盒叠加跨系统调用，增加泄密
与对外服务合规风险，拦截、溯源更难。

同一个Agent：受控条件下容易成功，动态业务中必须面对稳定性、安全、成本和运营

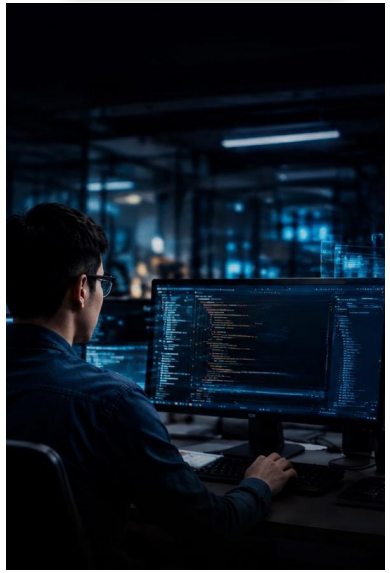
SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

深信服解决方案



建设理念：逐步构建完整的AI Infra体系，释放企业AI价值

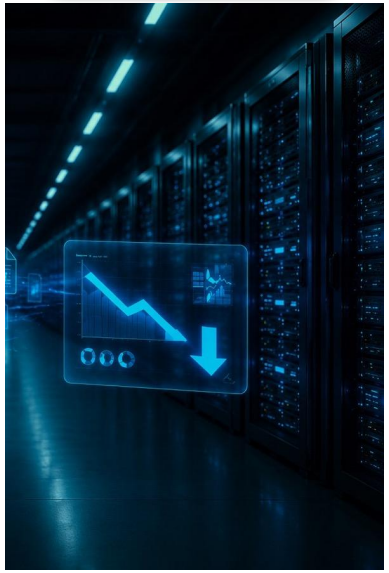
01
开发体系
简单快速 | 效果更好



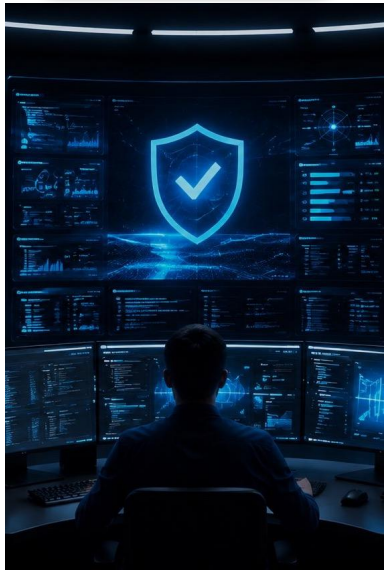
02
数据与知识底座
数据治理 | 知识增强



03
模型与成本治理
模型路由 | 成本优化



04
Agent治理
权限 | 版本 | 审计

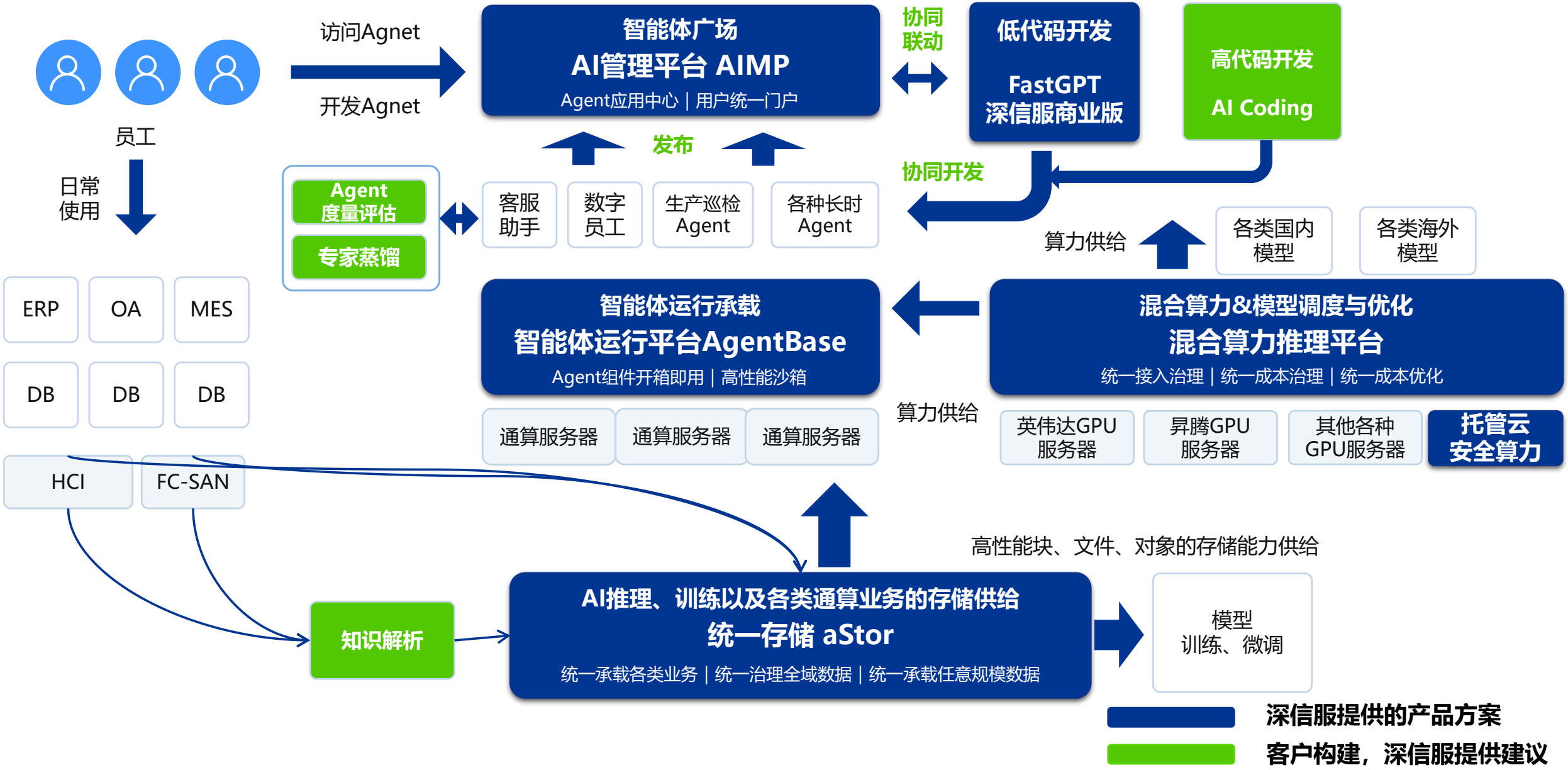


03
安全治理
数据 | 内容 | 合规



围绕AI建设旅程的五个核心阶段，体系化建设AI Infra，保障AI规模化进生产。

深信服AI解决方案整体架构



Agent开发与评估

01
开发体系
简单快速 | 效果更好

02
数据与知识底座
数据治理 | 知识增强

03
模型与成本治理
模型路由 | 成本优化

04
Agent治理
权限 | 版本 | 审计

03
安全治理
数据 | 内容 | 合规

Agent开发与运营，需要形成持续进化的价值闭环

开发做出效果，广场放大复用，评估验证价值，蒸馏沉淀组织能力

Agent全生命周期运营

评估反馈驱动持续迭代 ↻

01 → 02 → 03 → 04

Agent开发

效果优先，工具、数据与组织协同

效果是第一位

- 以任务效果为第一标准，用好工具与高质量数据共同保障
- IT与业务深度耦合，形成面向场景的产品型组织
- 从低代码快速起步，逐步演进至高低代码混合开发

智能体广场

统一发布、成果曝光、复用推广

统一分发入口

- 统一发布、编目、搜索和权限管理组织内所有Agent
- 让各部门成果被看见，以案例、榜单和激励促进创新
- 认证高价值Agent并全员推广，减少重复建设

Agent度量和评估

用效率、成功率与ROI验证价值

价值量化

- 评估任务完成效率与成功率，定位失败环节
- 核算单次成本、节省工时和业务收益，形成ROI
- 分析人工接管与收尾工作量，驱动版本持续优化

专家能力蒸馏

把专家经验沉淀为组织级数字能力

能力沉淀

- 通过工具平台与工作机制采集专家流程、规则和反馈
- 持续蒸馏为可复用的高价值SKILL与标准工作流
- 随业务变化迭代，长期构建匹配组织需要的数字员工

开发产生Agent，广场加速复用，评估验证价值，蒸馏沉淀能力——最终形成规模化运营飞轮

深信服与FastGPT强强联手，打造FastGPT深信服商业版

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

拥抱开源，围绕客户需求快速迭代，提供更好的使用体验



开源驱动，快速迭代 + 高准确率的对话 Agent

成熟的知识库能力 + 内置专家级最佳实践

灵活的应用构建能力 + 企业级平台能力

知识解析引擎&自调优引擎开发效果更好的Agent

SANGFOR 深信服科技 | 信服云 SANGFOR CLOUD

低代码
可视化
全栈Agent
开发

知识库管理		Agent应用构建			应用评测	平台管理	
多源数据导入	文件处理	简易对话Agent	工作流应用				
文件	智能分块	自动学习	简易应用	工作流编排	数据集生成/导入	团队管理	数据面板
web站点	PDF增强解析	数据纠错	应用配置	插件	数据集质量评估	用户管理	日志管理
数据库直连	知识增强	效果对比	MCP工具集	MCP服务	自动评测	权限管理	资源管理
第三方知识库	索引增强	兜底回复	应用调试	应用发布	评测维度管理	操作审计	系统配置
同义词管理	搜索测试	模板市场	多渠道对接	对话日志	评测总结	模型对接	模板/工具配置

高质量的
知识解析引擎

SANGFOR

0专家
好效果

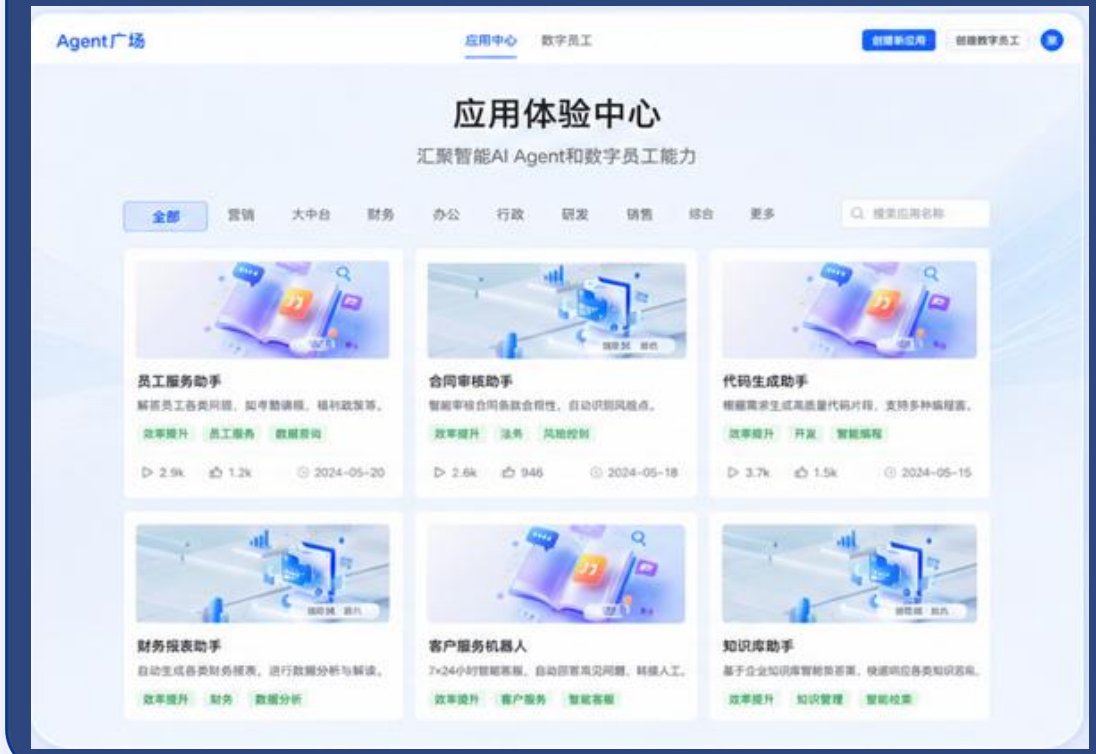
首次上线准确率 85%

效果持续进化的
自调优引擎

SANGFOR

统一发布运营，并持续挖掘高价值场景

统一的Agent和数字员工Portal



01 统一门户，方便业务人员使用、收藏、点赞评论

后台管理实现应用审核、上下架、排名统计



02 后台管理，实现应用上下架和运营，促进Agent推广

深信服统一存储 aStor

01
开发体系
简单快速 | 效果更好

02
数据与知识底座
数据治理 | 知识增强

03
模型与成本治理
模型路由 | 成本优化

04
Agent治理
权限 | 版本 | 审计

03
安全治理
数据 | 内容 | 合规

AI等创新业务推动着软件定义存储进入SDS 3.0 统一存储 承载AI却不止AI

业务类型决定存储形态

海量数据
备份归档场景

阶段一：SDS 1.0


AI质检


芯片设计


影视制作

海量非结构化数据
生产存储场景

阶段二：SDS 2.0


AI、块、文件
高性能
核心生产场景

阶段三：SDS 3.0

统一存储

Start Small　　Scale Big



2U盘控一体形态：
2节点起步，节点直连，无需交换机；
容量和性能**线性扩展**；
适用于数据库、企业级云、EDA、测绘等数据密集型场景
AI、HPC等新型应用

SDS 3.0

容量归档存储



独立服务器形态
3节点+交换机起步，容量和性能线性扩展；
配置大容量的HDD介质，少量缓存盘
性能与可靠性不足，仅适用于归档场景

SDS1.0

性能提升
可靠性增强

企业生产存储



独立服务器形态
3节点+交换机起步，容量和性能线性扩展；
配置大容量的HDD介质，少量缓存盘
最适合热数据空间需求较小的非结构化生产场景

SDS 2.0

创新业务加速，效率推动

既能承载现在，又能承载AI未来

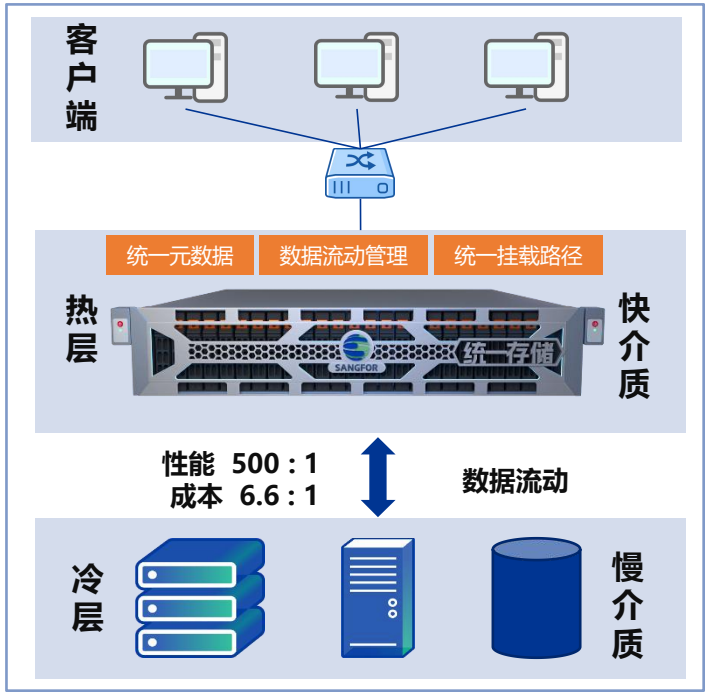
深信服统一存储解决方案的核心优势

产品创新释放数据价值

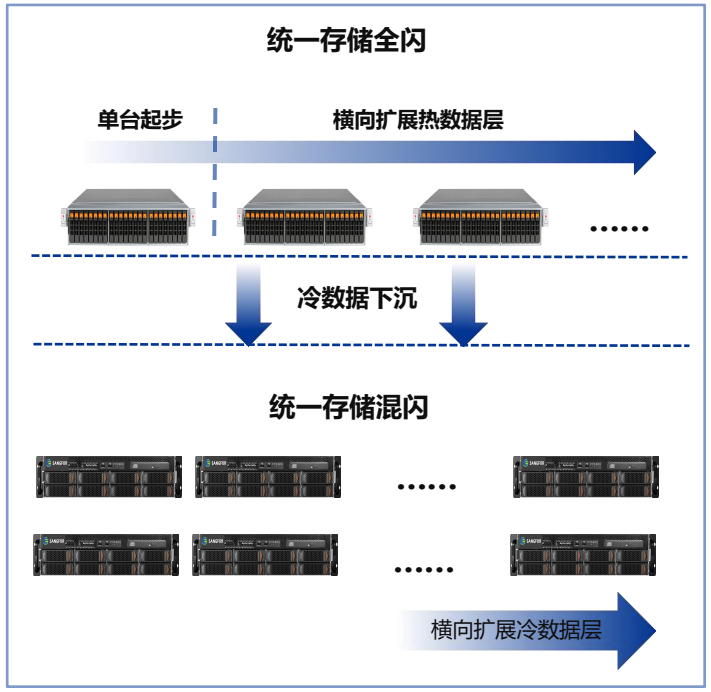
统一承载各类业务
传统业务稳得住，AI创新跑得快



统一治理全域数据
全域数据充分利用



统一储存任意规模数据
数据海量增长成本长期可控



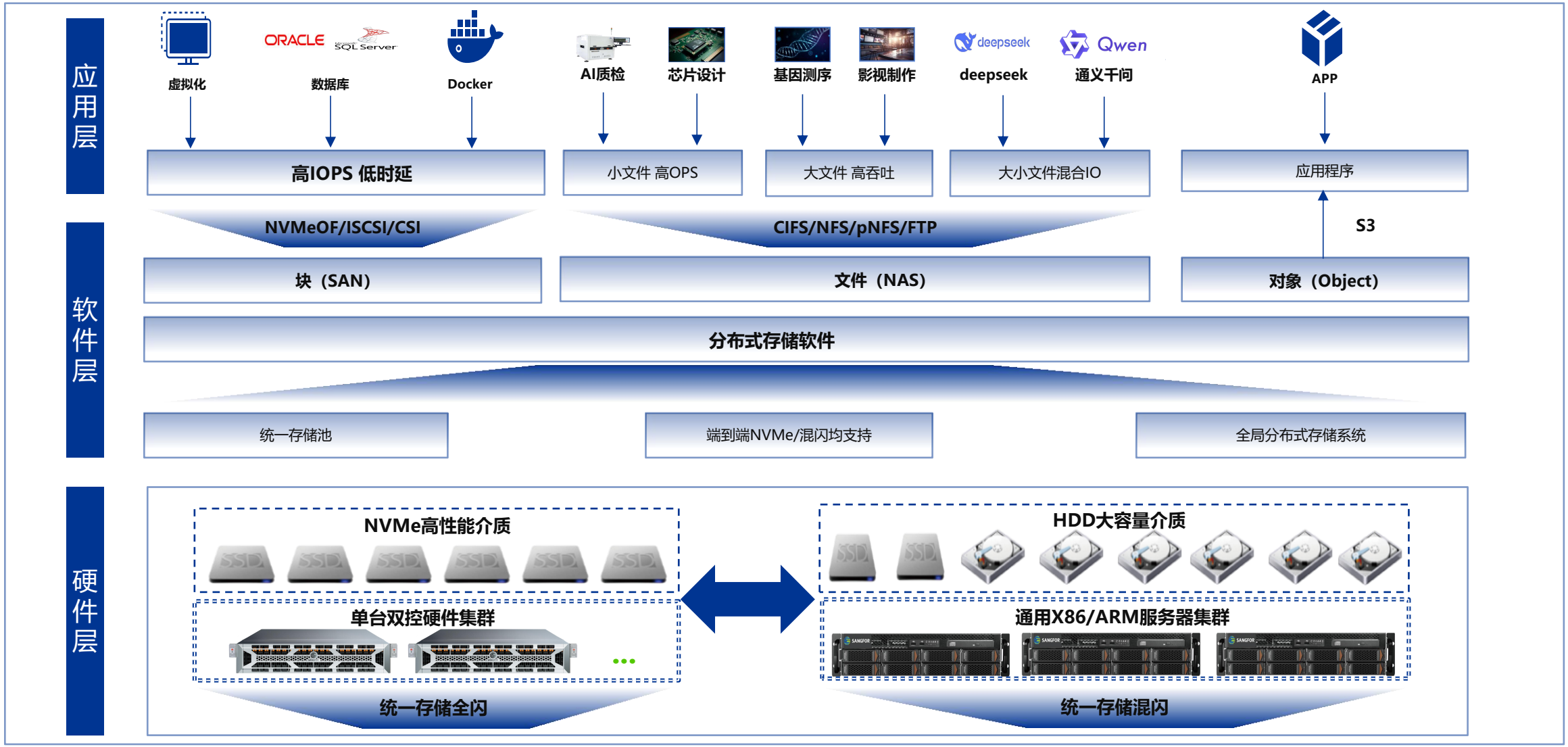
极致性能
稳定可靠

数据流动
统一管理

灵活扩展
高效经济

一套存储软件、两种硬件形态，满足各类业务工作负载

产品创新释放数据价值



统一治理全域数据：让数据统一可见、按需流动，该快的快、该省的省

产品创新释放数据价值



数据不搬，全域可见

异构纳管NAS、对象及云存储构建统一视图
3000W文件跨存储找
从32分钟缩短到5秒

按需加速，冷热分层

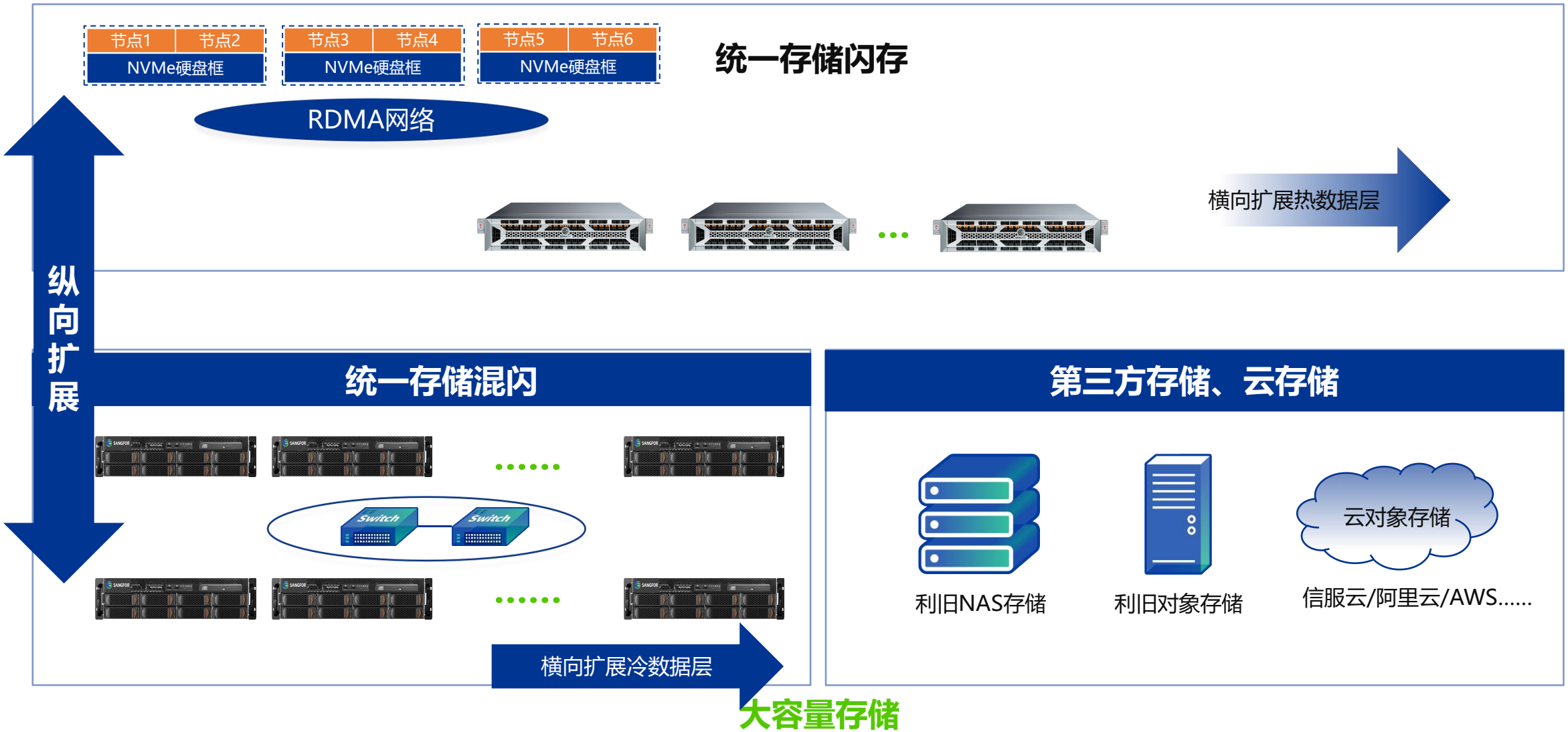
数据按冷热在全闪和混闪之间流动
兼顾业务性能与长期成本
TCO降低至少60%

多人协作，过程可控

通过版本、分支、发布和回滚管理数据
减少重复副本，让变更可追溯、结果

统一承载任意规模数据，全闪单节点起步，弹性灵活应对未来业务增长

产品创新释放数据价值



AI统一存储池：更高性能、更低成本、更灵活部署

数据更快可用

痛点：海量小文件拖慢链路

清洗、标注、入湖等待时间长

能力：40W IOPS

小文件并发读写与治理提速

训练更快产出

痛点：GPU 等数据

加载、写入、恢复拖慢训练

能力：120GB/s +GDS/RDMA

减少 CPU 中转和数据等待

成本持续下降

痛点：高成本、重复建设

高性能介质长期占用、容量浪费

能力：统一纳管流动 + QLC + 压缩

降低介质、空间和运维成本

资源按需增长

痛点：扩展不够灵活

性能容量绑定，旧资产难复用

能力：性能容量独立扩展

冷热分层，第三方/云统一接入

业务价值量化　从数据准备、训练执行、资源成本到规模化扩展，形成 AI 数据底座的完整经营收益闭环

20%+

版本迭代提速

从“等数据”转向“持续高效数据使用”

60%+

3年数据TCO优化

保护已有投资、全局流动减少全闪数据占用

灵活部署

业务先跑起来灵活演进

单台起步业务增长时按需扩展，灵活敏捷支撑业务

总结：AI 存储不只是基础设施升级，而是把数据准备、GPU 训练、存储成本和资源扩展变成可度量、可优化、可持续的业务生产力。

混合算力调度与优化平台



混合算力调度与优化平台

统一接入治理

- 一站式接入和管理混合算力，屏蔽异构差异，让AI能力像水电一样即开即用

统一成本治理

- 全局看清混合算力用量与成本，统一计量和分摊，让AI成本可见、可管、可控

统一成本优化

- 统筹效果与体验，持续优化模型服务与资源使用，降低单位有效Token成本



混合算力统一入口：全局算力真正看得清

云端
Token订阅MaaS服务

Overlay

深信服算力平台 | 第三方云/算力平台

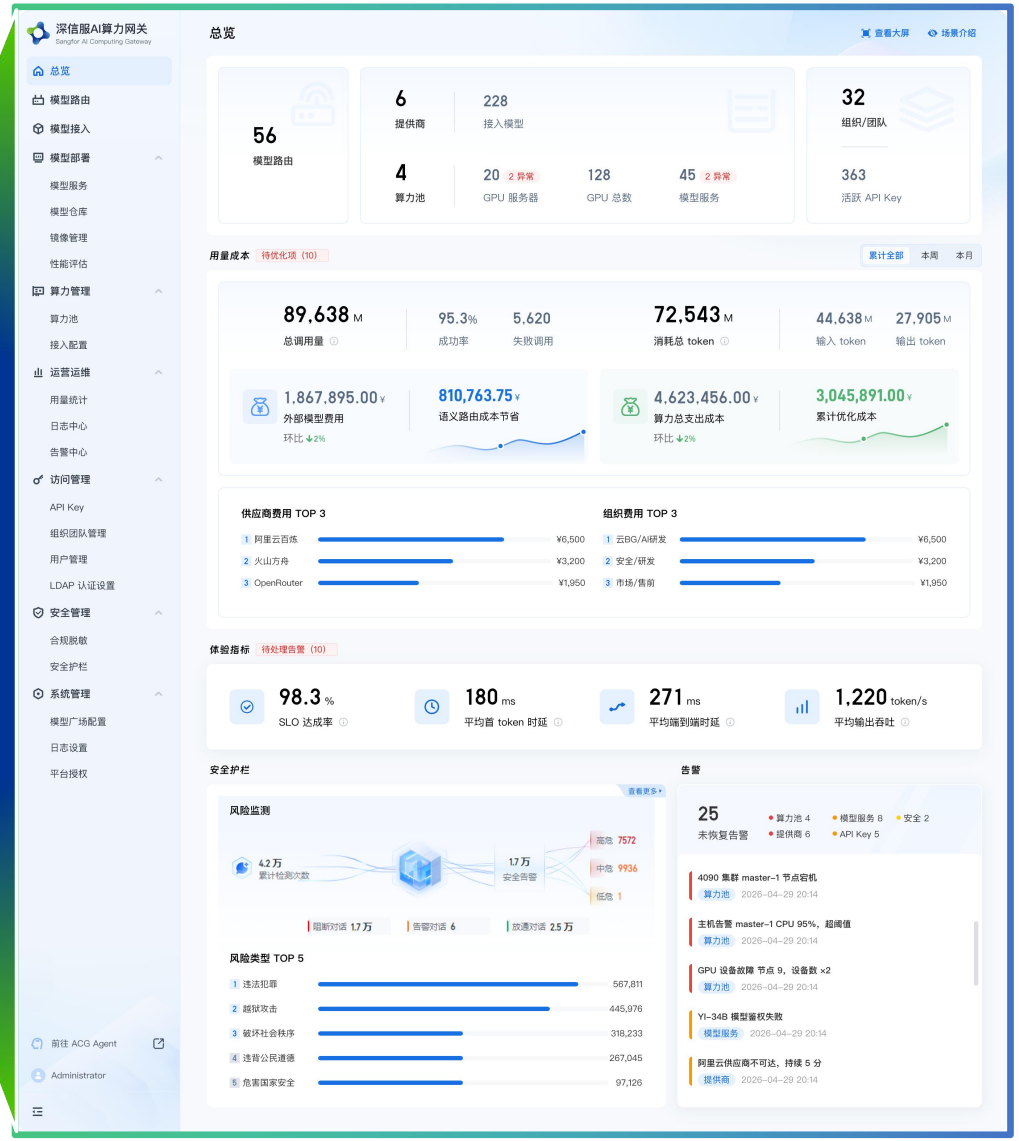
本地、各分支、部门
自建算力

Overlay

深信服算力平台 | 第三方云/算力平台

本地、各分支、部门
租用/托管/IDC算力

统一
入口



一人一 Key 一策略，更好的保护敏感数据

一个统一 API 地址 + 属于你的一把 Key

这把 Key 同时承载四件事，无需使用者理解背后的算力拓扑

身份 与 AD / LDAP / SSO 打通，Key 即人，谁在用一目了然

权限 能用哪些模型、哪些算力池、能否出域，由策略决定而非靠自觉

留痕 每一次调用可追溯到人、到 Agent、到具体业务场景

业务代码零改动

换模型、换卡、换供应商，对使用者完全无感知；算力侧的每一次优化，都不需要业务配合改一行代码。

接入成本从「天」到「分钟」

新人上手从申请多个平台账号，变成发一把 Key；部门自助开通，不再排队等运维。

用量自己先看得见

部门与个人拥有各自的用量看板，花了多少、花在哪先于账单被看见，超支不再是月底才知道。

对使用者越简单，对企业才越可控 —— 复杂度必须由平台吸收，而不是转嫁给每一个人

统一接入治理场景-业务价值总结

看得清

开放兼容与解耦，内部、云上、第三方模型服务统一接入、运营和扩展，不被算力平台和供应商绑定，应用 0 改造

全局/各部门/应用/个人Token用量可视、费用可视，推动各部门AI进度、发现AI明星员工、加速AI明星应用推广

提供AI算力网关Agent，MCP运维对接服务

管得住

调用行为、输入输出内容深度可视，让AI调用不再黑盒，行为管控有依据Token级智能限流、配额管理，每一分算力用有所值，算力使用不再失控

私有基础设施服务化：多卡多模型快速适配、API Key管理，N部门N模型显卡资源利用率倍增，一键限流关键业务有保障

用的稳

聚合路由+多来源容灾+限流熔断，业务体验更好，连续性更强

极低开销时延 + 多实例 + 生产级可靠底座 + 企业级产品可靠与服务

更安全

平滑对接大模型安全护栏、基于联邦隐私技术的托管AI安全算力

- **通过强大的可见性**

让AI创新在组织内的落地速度数倍提升

- **通过Token管控**

核心业务AI算力保障能力提升2-3倍，AI业务全局可靠性提升1个9

- **通过安全治理**

让核心资产不泄密，内容舆情更可控

*以上效能数据来自业界广泛Public AI和实践用户的能力落地后的成果统计

基于任务的智能调度引擎：让模型费用越用越省

基于缓存感知的任务调度算法



更省钱

解决缓存切换、脏数据等成本突增问题

基于模型和场景的知识图谱算法



更灵活、更精准

多档位模型协同，效果成本间综合选择

使用智能路由引擎对接ClaudeCode，成本大幅下降，不降智

复杂问题
“帮我重构这个xx代码模块，提升性能”

简单问题
“帮我写一篇护网通xx”

基于任务的智能调度引擎

语义与表示特征
句法与结构特征
文本与统计特征

决策因子

复杂 → CLaude4.8 Opus → “已经完成xx模块重构，性能提升了50%，你可以开始测试体验...”

简单 → DeepSeek v4 Pro → “护网通知编写完成，已保存到...”

CLAUDE CODE

节省¥100万

省

*使用基于任务的智能调度引擎后，1000人研发团队一个月调用外部模型费用

智能融合架构SFA2.0：
构筑持续的专属算力2～5倍ROI提升，成本优化百万起

通过工具、集成的监控等手段，辅助定位应用场景的性能瓶颈，再结合自适应的原子优化能力，实现应用端到端承载的ROI提升

创新自研自适应模块

定位性能瓶颈、采取性能优化配置

自适应原子能力

基于场景数据特征和性能瓶颈选择自适应原子能力优化

Agent应用

企业级 openclaw　卫宁健康　衡石　派兹　鼎捷软件　东华医为　...

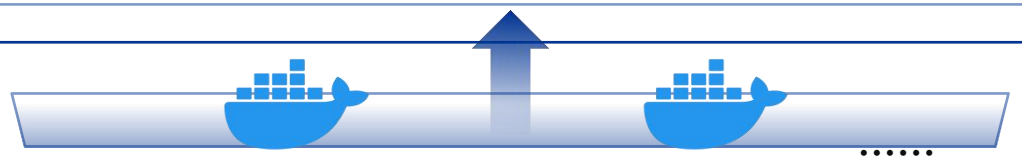
AI算力网关　**自适应架构层Smart Arc**

自适应模块	自适应多版本/类型推理引擎	Overlay指标采集	性能瓶颈分析	一键最佳实践配置	自适应调整

自适应原子能力	开源推理引擎与调参（保持更新）	深信服自研优化算法持续增强	
	单实例缓存优化 GPU数据传输优化 KVcahe共享 ……	全局智能调度	智能路由｜Cache感知|负载均衡｜多实例调度...
		推理引擎优化	自适应投机解码｜并行策略优化｜单实例性能优化...
		KV Cache协同	多级缓存｜缓存感知调度｜跨层高速传输...

灵活部署，交付调优效果

支持对接用户已有的平台提供性能优化能力
（需要提供标准k8s接口）



DeepSeek满血版优化镜像　Qwen3-72B优化镜像　……

本地深信服算力平台GPU资源池　第三方算力平台GPU资源池

智能融合架构SFA2.0：

构筑持续的专属算力2～5倍ROI提升，成本优化百万起

深信服自研　开源生态

模型接入层	统一协议与模型接入	多协议统一转换（自研）	MaaS + 私有模型统一接入（自研）	OpenAI / Responses	Anthropic Messages	多模态模型	LLM

深信服技术深耕

服务路由层	跨模型与跨服务路由	请求级语义路由（自研）	缓存亲和路由（自研）	故障感知与迁移（自研）	上下文长度路由	权重 / 优先级	蓝绿 / Fallback	
推理调度层	多实例负载与SLO调度	自适应感知调度引擎（自研）	SLO感知调度（自研）	KV Cache感知调度（自研）	PD分离传输加速（自研）	请求抢占	PD分离	负载感知调度	队列感知调度
推理引擎层	引擎适配、解码与算子优化	业务感知推理引擎（自研）	自适应步长投机解码（自研）	vLLM / SGLang	vllm-ascend	Chunked prefill	Continuous Batching	TP/DP/EP并行
KV Cache层	缓存卸载与共享	Agent冷热数据 KV Cache分层管理（自研）	多级KV Cache卸载	跨实例KV Cache共享	分布式KV Cache存储			

硬件层	云原生高性能算力池	昇腾	NVIDIA	沐曦	海光	天数智芯	更多国产GPU

统一成本治理与优化场景-业务价值总结

创新智能路由引擎

语义路由、聚合路由缓存

创新智能融合架构SFA2.0

基于应用场景，打造自适应架构层，创新自研自适应模块、自适应原子能力

vGPU切分

vGPU支持GPU/NPU，显卡利用率翻倍

- 通过混合算力Token管控

让全局Token消耗降低50%

- 全局算力调度场景：费用越用越省，节省30%-50%，AI Coding场景：节省50%+，每年降本数百万~数千万

- 专属算力场景：让ROI提升5~10倍，成本节省数百万起步

让Private AI的ROI不断接近Public AI

深信服安全护栏

01
开发体系
简单快速 | 效果更好

02
数据与知识底座
数据治理 | 知识增强

03
模型与成本治理
模型路由 | 成本优化

04
Agent治理
权限 | 版本 | 审计

03
安全治理
数据 | 内容 | 合规

深信服大模型安全防护框架——强对抗 高可靠 可进化

大模型安全护栏

强对抗：攻防场景高检出，低误报

高可靠：对AI业务稳定0影响

可进化：伴随AI业务发展

应用层

- 风险总览
- 资产管理
- 策略管理
- 能力中心
- 日志检索
- 系统运维

逻辑层

- 威胁分类
- 响应策略
- 阻断提示
- 业务保护
- 通知提醒

GPT引擎

- 文本检测模型：文本检测&综合研判
- OCR模型：图片文字识别
- ASR模型：音频转文本
- 多模态模型：图片/音视频分片检测

规则引擎

- 内置PII库
- 自定义关键字
- 自定义正则
- 误报规则

数据层

- API
- SDK
- AI安全网关
- ...

AI风险评估

以”评”促”防”

内容安全评估

- 攻击语料自动泛化
- 多场景多轮对话诱导

Skills风险评估

- 模型分析
- 情报分析

运行环境扫描

- 30+主流框架指纹
- 400+AI框架漏洞

运行时防护

上线前检查

强对抗：红队数据训练垂域模型，提示词注入95%+高检出

罕见语种混淆
攻击者只用不常见语言翻译恶意问题

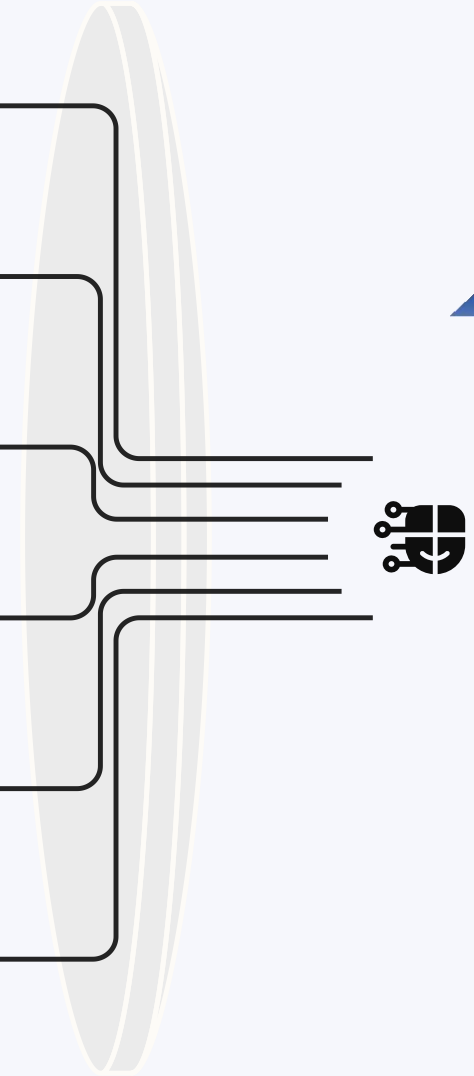
编码混淆
攻击者将恶意问题编码绕过模型输出有害内容

策略傀儡
攻击者将恶意指令伪装成策略文件输出恶意内容

对立响应
攻击者设定对立响应，诱导模型有害输出

反向诱导
攻击者伪装反对，实则诱导模型违背准则

威逼利诱
攻击者威逼利诱，诱导模型违背准则



95%+

护栏攻防手段高检出
提供安全领域专精的攻防手段检出能力，不止有害内容

海量红队数据门槛
20+年红队攻防、渗透测试、漏洞研究实战数据积累

垂域模型门槛
3000亿Token安全领域知识
保障训练数据国内首个安全垂域模型

实战测试：相较各类护栏，强对抗技术实力

以实战对比为例，本次参与评测的包括互联网厂商、海外厂商和国内安全厂商，对应实现方案分别是规则+小模型、规则为主和大模型方案。测试数据总共2000条，其中1500黑、500白，75%来自第三方数据源，25%来自攻防专家构造。

独立安全护栏的必要性

开源基础大模型具备一定的输入输出防护能力，但效果有限，随着模型应用复杂度增加，风险也会持续加剧

应用大模型技术构建安全护栏的必要性

基于规则的检测方案面对复杂的应用场景，泛化能力有限，检出能力弱，误报控制差

主要厂商	友商	检出率	检出精准率（黑白混合）
互联网厂商	T某	18%	38%
	A某	55%	88%
	W某	28%	48%
海外厂商	Meta Prompt Guard	26%	80%
安全厂商	A某	10%	10%（误报达90%）
	深信服	90%	95%


SANGFOR
深信服科技
信服云
SANGFOR CLOUD
谢谢聆听
让每个用户的数智化化更简单、更安全!