

WHITE PAPER

Agent Native Compute

面向企业 Agent 时代的新一代 AI 计算基础设施

Powered by NovuWing™ Architecture

基于 NovuWing™ 技术架构

浙江星流计算科技有限公司

VERSION 1.0 | 2026.08

版权页

Agent Native Compute White Paper (Powered by NovuWing™ Architecture)

《Agent Native Compute：面向企业 Agent 时代的新一代 AI 计算基础设施》（基于 NovuWing™ 技术架构）

版权所有 © 2026 浙江星流计算科技有限公司 (NovuInfra)

本白皮书用于介绍 Agent Native Compute 技术范式、NovuWing™ Agent Native Compute Architecture、企业级 AI 基础设施体系以及玄矩™ AI 产品设计方向。

未经浙江星流计算科技有限公司书面许可，不得复制、修改、传播本白皮书的文字、图表、架构设计及版式表达，或用于商业用途；本白皮书不构成对任何商标、专利或其他知识产权的明示或默示许可。

知识产权声明

- “星流计算” “NovuInfra” 及星流计算图形标识为浙江星流计算科技有限公司及其全资子公司星流九天（上海）科技有限公司依法使用的公司商标或标识，相关权利依法受到保护。
- “Agent Native Compute” 用于描述面向 Agent 生命周期进行系统级优化的计算范式和技术理念；“NovuWing™” 是浙江星流计算科技有限公司及其全资子公司星流九天（上海）科技有限公司基于该范式构建并依法申请或使用的技术架构品牌及技术标识。
- “玄矩™” 及其图形标识、“SerpMind™” 及其图形标识为浙江星流计算科技有限公司及其全资子公司星流九天（上海）科技有限公司依法使用的产品商标或标识，相关权利依法受到保护。
- 核心专利（申请号：202611065591.X）：《基于 Agent 语义的三级分层 KV 缓存智能调度方法及系统》。
- 除非另有说明，本白皮书所含文字、图表、架构设计及版式表达的著作权及相关权益归浙江星流计算科技有限公司及其全资子公司星流九天（上海）科技有限公司所有。

版本：Version 1.0

发布日期：2026.08

星流计算 NovuInfra

Enterprise Full-Stack AI Infrastructure Company

目录

01 ~ 02	执行摘要
03	架构总览
04	AI 基础设施的转折点
05 ~ 06	第一部分 Agent 时代
07 ~ 09	第二部分 现有 AI 基础设施为何不足
10 ~ 14	第三部分 Agent 原生计算架构
15 ~ 17	第四部分 Agent 内存架构
18 ~ 21	第五部分 玄矩™ A1 架构
22 ~ 25	第六部分 AI 基础设施经济学
26 ~ 29	第七部分 企业级全栈 AI 基础设施
30 ~ 33	第八部分 行业实践与参考架构
34 ~ 36	第九部分 企业 AI 基础设施的未来
37 ~ 39	第十部分 玄矩™ A1 技术架构与规格概览
40 ~ 42	附录 核心架构图

执行摘要

EXECUTIVE SUMMARY

面向 Agent 时代重新定义企业 AI 基础设施

人工智能正在经历一次新的范式迁移。

过去几年，大语言模型（Large Language Model, LLM）的突破，使人工智能具备了前所未有的理解、推理和生成能力。

然而，当 AI 从实验环境进入企业生产环境，一个新的挑战正在出现：

企业真正需要的，不只是一个更强大的模型，而是一套能够支撑智能持续运行、规模化应用和不断演进的基础设施。

AI 正从 Model Era（模型时代）进入 Agent Era（智能体时代）。

在模型时代，AI 系统主要围绕模型训练和推理优化：

更大的参数规模、更强的计算能力、更高的推理性能。

而在 Agent 时代，企业 AI 的运行方式发生根本变化。

Agent 不再只是生成内容，而是成为能够理解目标、执行任务、调用工具、访问知识、连接业务系统，并持续完成工作的智能主体。

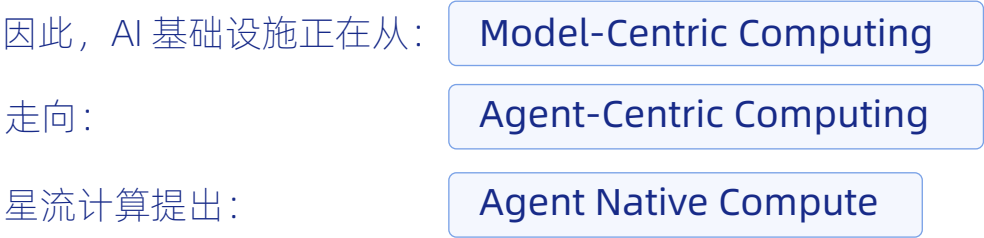
这对基础设施提出了新的要求：

- 持续运行（Continuous Operation）；
- 多 Agent 协同（Multi-Agent Collaboration）；
- 长期上下文管理（Long Context & Memory）；
- 企业级可靠性（Enterprise Reliability）；
- 规模化智能部署（Scalable Intelligence）。

传统 AI 基础设施主要面向模型计算设计。

其核心目标是提升单次推理性能。

但 Agent 时代需要解决的问题，是如何让越来越多的智能主体在企业环境中长期、高效、可靠运行。



面向企业 Agent 时代重新定义 AI 计算范式。基于这一范式，星流计算构建了 NovuWing™ Agent Native Compute Architecture。NovuWing™ 架构不只是提升计算能力，而是围绕 Agent 生命周期重新设计计算、内存与运行环境。

它关注：

- 如何管理 Agent 状态；
- 如何优化上下文访问；
- 如何提升资源利用效率；
- 如何支撑大规模企业智能运行。

基于这一理念，星流计算推出：**玄矩™ A1 —— Agent Native AI Server**

玄矩™ A1 不是传统 AI Server 的简单升级，而是面向 Agent 时代重新设计的新一代计算基础设施。

同时，星流计算构建：

Enterprise Full-Stack AI Infrastructure
企业级全栈 AI 基础设施。

通过：

NovuWing™ Agent Native Compute
提供智能运行所需的计算底座

SerpMind™ Enterprise Agent OS
提供 Agent 创建、运行、管理和治理能力

行业智能应用
连接真实业务场景。

最终帮助企业从：**使用 AI** 走向：**拥有 AI**

未来十年，AI 将成为企业新的基础设施。
就像过去企业拥有 ERP、CRM 和云平台一样，未来每一家企业都将拥有属于自己的 AI。

架构总览

ARCHITECTURE OVERVIEW

NovuInfra Agent 原生 AI 基础设施总览

面向企业 Agent 时代，NovuInfra 将业务应用、行业 Agent、企业级 Agent OS、Agent Run-time、Agent Native Compute 与玄矩™ A1 计算基础设施连接为一套完整的全栈体系。以下总览图用于在进入详细章节前说明各层之间的关系。



图1 从企业应用到底层基础设施的全栈架构

Figure 1. NovuInfra Agent Native AI Infrastructure — Full-Stack Architecture Overview

AI 基础设施的转折点

THE AI INFRASTRUCTURE INFLECTION POINT

为什么是现在：从模型扩展走向智能运行

模型时代的基础设施主要围绕参数规模、GPU 规模和模型性能持续优化。进入 Agent 时代后，基础设施需要同时承载更多 Agent、更长上下文和更持续的运行状态，资源压力随之从单一计算扩展到内存、调度和生命周期管理。优化目标因此从一次模型推理，转向企业智能的持续运行效率。

AI基础设施的转折点：为什么是现在

AI INFRASTRUCTURE INFLECTION POINT

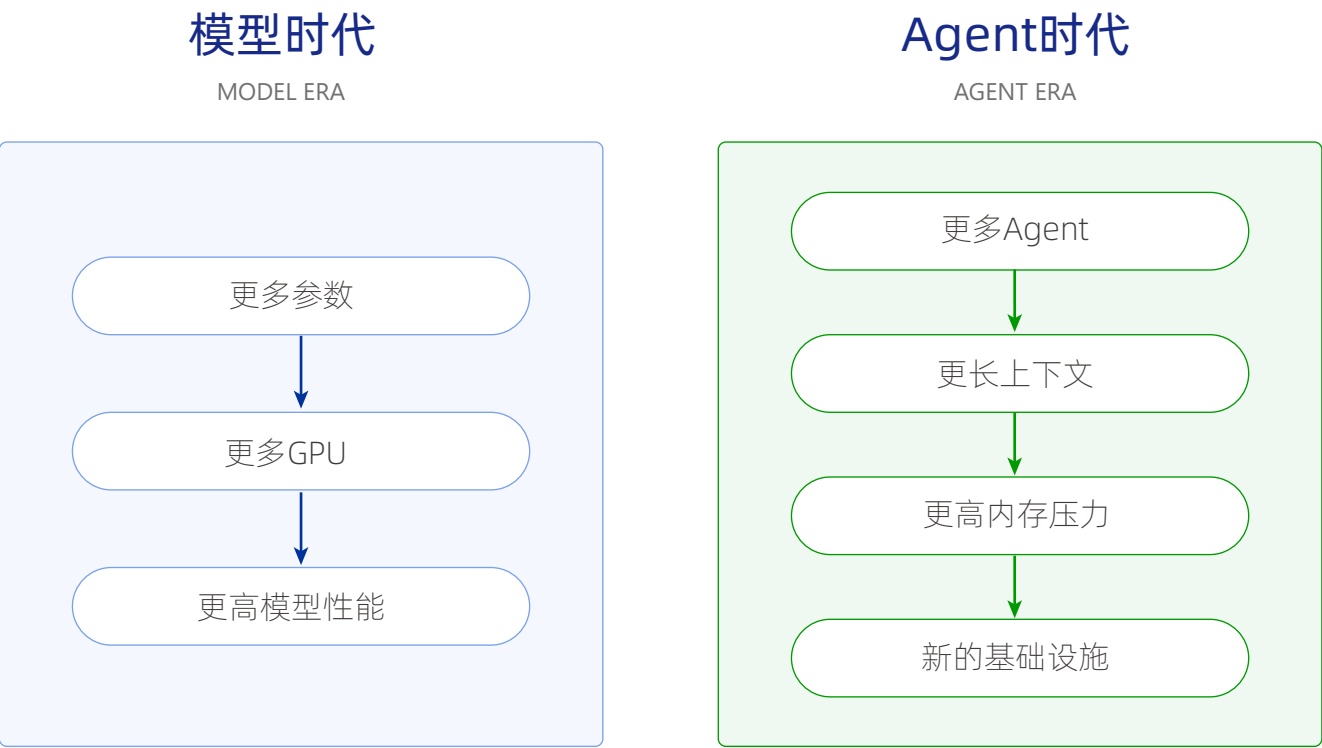


图2 优化目标从单次模型性能转向企业智能的持续运行效率

Figure 2. Model Era vs. Agent Era — The AI Infrastructure Inflection Point

第一部分 Agent 时代

PART 1 THE AGENT ERA

从模型智能到企业智能：AI 正进入 Agent 时代

1.1 AI 正在经历新的范式迁移

过去数十年，计算技术的发展始终伴随着软件形态的变化。

从个人计算机时代，到互联网时代，再到云计算时代，每一次计算平台的演进，都重新定义了软件的构建方式和企业运行模式。

过去几年，以大语言模型（Large Language Model, LLM）为代表的生成式 AI 技术取得突破，使机器首次具备了接近人类语言理解、知识推理和内容生成的能力。

模型能力的提升，解决了人工智能发展的一个核心问题：机器是否具备理解和推理能力。

然而，当 AI 从实验环境进入企业生产环境，一个新的问题逐渐显现：企业如何让智能持续、安全、稳定地运行？

模型能力的提升，是人工智能进入企业的起点，而不是终点。

未来企业需要的不只是能够回答问题的模型，而是能够理解目标、执行任务、调用工具、连接业务系统，并持续完成工作的智能主体。

这推动人工智能进入新的阶段：从 Model Era（模型时代），走向 Agent Era（智能体时代）。

1.2 从 Model 到 Agent：企业 AI 的运行方式正在改变

传统 AI 应用主要围绕模型推理展开。

其基本模式：

User Request → AI Model → Generated Response

企业需要的不是一次性的回答，而是能够长期工作的智能系统。

一个企业级 Agent，需要理解业务目标、制定执行计划、调用企业工具、访问知识和数据、管理上下文和长期记忆，并与其他 Agent 协同完成任务。

模型负责产生智能，Agent 负责执行智能。

AI 正从“生成内容”的工具，逐渐成为“完成工作”的生产力。

1.3 企业 Agent 与个人 AI 助手存在本质区别

消费级 AI 更多关注单用户体验、单次交互和快速响应。

而企业 Agent 面向的是复杂生产环境。企业需要管理的不再是一个 AI 助手，而可能是几十、几百甚至更多 Agent 的协同运行。

未来关注点将从如何训练更强大的模型，转向如何运行和治理一个由大量 Agent 构成的智能系统。

1.4 Agent 时代带来的新计算需求

Agent 的持续运行，对计算基础设施提出了新的要求。

传统 AI Server 主要围绕 GPU 算力、显存容量和单次推理性能优化。

但企业 Agent 需要持续运行、多 Agent 并发、长上下文与持续记忆管理。

随着上下文长度增加，KV Cache 成为新的资源瓶颈。

1.5 从模型智能到企业智能

大模型解决了智能能力的问题。

但企业真正需要的是可以运行、可以管理、可以持续进化的智能。

未来企业 AI 的完整体系将由 Enterprise Applications、Industry Agents、Enterprise Agent OS、Agent Runtime 和 AI Compute Infrastructure 共同组成。

1.6 Agent Native Compute：为 Agent 时代重新设计计算

随着企业 AI 从单次推理（Inference）走向 Agentic Application，计算架构必须重新设计。

传统 AI Compute 以 GPU 为中心，而 Agent 时代需要 CPU、GPU、Memory 与 Runtime 的系统级协同。

企业 AI Compute 正从 GPU-Centric 走向 System-Centric。

因此，星流计算提出 Agent Native Compute：一种面向企业 Agent 持续运行场景的新一代计算范式。

基于这一范式，星流计算进一步构建 NovuWing™ Agent Native Compute Architecture，通过 CPU、GPU、Memory 与 Runtime 的系统级协同，为企业 Agent 提供长期、稳定和高效运行的基础。

它关注的不只是让一次推理更快，而是让越来越多 Agent，在企业环境中长期、稳定、高效运行。

小结

人工智能正在进入新的阶段。

过去：模型定义智能。

未来：Agent 承载智能。

而企业真正需要的是支撑智能持续运行的基础设施。

Agent Era 的到来，不仅改变软件形态，也将重新定义 AI 计算架构。

第二部分 现有 AI 基础设施为何不足

PART 2 WHY EXISTING AI INFRASTRUCTURE IS NOT ENOUGH

为什么传统 AI 基础设施无法支撑 Agent 时代

2.1 从模型部署到智能运行：AI 基础设施正在面临新的挑战

过去几年，企业 AI 的基础建设主要围绕大模型训练和推理展开。计算资源的核心目标，是提升模型性能：更大的 GPU 算力、更高的显存容量、更快的推理速度。

这种以模型为中心（Model-Centric）的架构，有效支撑了大模型快速发展。

然而，当企业 AI 从模型调用进入 Agent 持续运行阶段，基础设施面临的问题正在发生变化。未来企业运行的不再只是一个模型，而是一组持续工作的智能 Agent。这些 Agent 需要长期保持上下文，需要调用工具，需要访问企业数据，需要与其他 Agent 协同完成复杂任务。

因此，企业 AI 面临的新问题已经不只是：“模型能不能运行”。

而是：

“大量 Agent 能否长期、稳定、经济地运行。”

2.2 传统 AI Server 的局限：为一次推理优化，而非持续智能优化

传统 AI Server 的设计目标，是优化一次推理过程。

典型架构关注：

- GPU 计算能力；
- HBM 显存带宽；
- 模型吞吐量；
- 单次响应延迟。

这种架构适合传统 AI 推理场景。

但 Agent 的运行方式完全不同。

一个企业 Agent 在生命周期中可能处于不同状态：

- Active：正在执行任务。
- Idle：等待下一步输入。
- Sleeping：暂时休眠。
- Recovering：恢复上下文并继续执行。

在这些状态之间，Agent 需要持续管理：

- 上下文状态；
- 可恢复运行状态；
- KV Cache；
- 工具调用状态。

如果基础设施仍然按照传统模型推理方式分配资源，就会导致大量计算和存储资源无法被高效利用。

2.3 新的瓶颈：从计算墙到内存墙

随着模型规模增长，AI 系统的瓶颈正在从单纯计算能力逐渐转向内存和数据访问效率。

在大模型和 Agent 场景中，KV Cache 成为影响系统效率的重要因素。

KV Cache 保存模型生成过程中已经计算过的上下文信息，可以避免重复计算，提高推理效率。但随着：

- 上下文长度增加；
- Agent 数量增加；
- 并发任务增加。

KV Cache 规模快速增长。传统方式通常依赖 HBM 保存大量运行状态。然而 HBM 具备极高性能，同时成本高、容量有限。如果所有 Agent 状态长期占用 HBM，将导致：

- 显存资源快速耗尽；
- GPU 扩展成本持续增加；
- 资源利用效率下降。

因此，企业 AI 的核心挑战正在从：

“拥有更多算力”

转向：

“如何更智能地管理计算与内存资源。”

2.4 Agent 时代需要新的内存架构

Agent 与传统模型推理最大的区别，在于它具有生命周期。

不同阶段的数据价值不同。

正在执行任务的 Agent，需要最高性能资源。

暂时等待的 Agent，需要保留状态，但不需要占用最昂贵资源。

低活跃上下文和可恢复运行状态，则需要更大容量、更具成本效率的资源承载。

因此，未来 AI 基础设施需要根据 Agent 状态进行智能资源分配。

这要求计算系统具备：

- 多层内存协同能力；
- 上下文生命周期管理能力；
- Agent 状态感知能力；
- 自动化资源调度能力。

这也是星流计算提出三级内存架构（3-Tier Memory）的重要背景。

2.5 从 GPU 中心走向系统中心

传统 AI 基础设施以 GPU 为中心。

GPU 提供计算能力，HBM 提供高速数据访问，系统围绕模型运行进行优化。

但 Agent 时代需要的是系统级协同。

未来 AI 基础设施需要让：

- CPU 负责任务调度；
- GPU 负责高效计算；
- Memory 负责上下文管理；
- Runtime 负责 Agent 生命周期管理。

计算资源不再只是被动执行任务，而需要理解智能运行过程。

因此，AI 基础设施正在从：

GPU-Centric Computing

转向：

System-Centric Computing

这不是简单增加硬件规模，而是重新设计计算系统，使其适应 Agent 持续运行的需求。

2.6 企业 AI 的下一代基础设施需求

未来企业 AI 基础设施，需要同时满足：

- 持续运行能力（Continuous Operation）：支持 Agent 长时间稳定运行。
- 多 Agent 协同能力（Multi-Agent Collaboration）：支持大量 Agent 并发工作。
- 智能资源管理能力（Intelligent Resource Management）：根据 Agent 生命周期动态分配计算和内存资源。
- 企业级可靠性（Enterprise Reliability）：满足安全、治理、审计和业务连续性要求。
- 可持续扩展能力（Scalable Intelligence）：随着企业 Agent 数量增长，基础设施能够持续演进。

因此，企业 AI 的未来，不只是需要更强大的模型，也需要重新定义承载智能运行的基础设施。

2.7 小结

AI 正在从模型时代进入 Agent 时代。

传统 AI Server 解决的是：

而 Agent 时代需要解决的是：

如何让模型运行得更快。

如何让越来越多的智能体长期、稳定、高效地运行。

这要求计算架构从模型中心走向 Agent 原生。

这也是 Agent Native Compute 被提出的根本原因。

第三部分 Agent 原生计算架构

PART 3 AGENT NATIVE COMPUTE ARCHITECTURE

面向 Agent 时代重新定义 AI 计算架构

3.1 从基础设施演进到计算范式重构

前两部分已经说明，Agent 工作负载正在将企业 AI 基础设施的优化目标从单次模型执行，推向持续智能运行与系统级资源协同。

这一变化要求计算系统围绕 Agent 生命周期、上下文状态、资源调度和企业级运行要求进行整体设计。基于此，星流计算提出 Agent Native Compute。

因此，未来 AI 基础设施的核心问题不再只是：

如何让模型计算更快。

而是：

如何让智能系统长期、高效、可靠地运行。

基于这一变化，星流计算提出：

NovuWing™ Agent Native Compute

面向企业 Agent 时代的新一代 AI 计算架构。

3.2 什么是 Agent Native Compute

Agent Native Compute 是一种以 Agent 生命周期为核心设计对象，围绕持续运行、上下文管理、内存编排、资源调度和企业级治理进行系统级优化的新一代计算范式。

星流计算基于这一范式，构建了 NovuWing™ Agent Native Compute Architecture，将上述设计原则落实为可部署的计算、内存、Runtime 与系统协同架构。

传统架构主要围绕以下指标进行优化：

- 模型参数规模；
- 推理吞吐；
- GPU 利用率；
- 显存带宽。

而 NovuWing™ Agent Native Compute 关注：

- Agent 持续运行；
- 动态资源调度；
- 状态保存与恢复；
- 多 Agent 协同；
- 上下文管理；
- 企业级可靠运行。

其核心理念是： 计算系统不再只是执行模型，而是理解并服务智能运行过程。

换句话说：

传统 AI 基础设施以模型计算为主要优化目标。
Agent Native Compute 以智能工作负载的持续运行为主要优化目标。
NovuWing™ 将这一目标落实为系统级计算架构。

正式定义 / FORMAL DEFINITION

Agent Native Compute

一种围绕企业 Agent 生命周期管理、上下文与内存编排、计算资源调度以及持续智能运行效率进行系统级优化的新一代计算范式。

其设计对象不是单次模型调用，而是在企业环境中长期运行、可恢复、可协同且可治理的智能工作负载。

NovuWing™ 是星流计算对 Agent Native Compute 范式的技术架构实现。

3.3 从 GPU 中心计算走向系统中心计算

过去的 AI 计算架构主要以 GPU 为中心。

GPU 提供核心计算能力，HBM 提供高速数据访问，整个系统围绕模型推理流程进行优化。这种架构在模型时代非常有效。

但 Agent 时代要求更高层次的系统协同。

未来 AI 基础设施需要同时协调：

- CPU：负责任务管理、系统调度和业务逻辑处理。
- GPU：负责大规模并行计算和模型推理。
- Memory：负责上下文、状态和知识访问。
- Runtime：负责 Agent 生命周期管理和资源调度。

因此，AI 计算架构正在从：

GPU-Centric Computing

而是

System-Centric Computing。

这意味着 AI Server 不再只是计算设备，而将成为企业智能运行的基础平台。

3.4 NovuWing™ 架构的核心设计原则

为了支撑企业 Agent 的规模化运行，NovuWing™ Agent Native Compute 遵循四项核心设计原则。

一、Continuous Intelligence

持续智能

企业 Agent 需要长期运行，而不是一次性响应。

计算基础设施需要支持持续状态保存、持续上下文管理和持续任务执行。

二、Context-Aware Computing

上下文感知计算

Agent 的价值来自持续积累的上下文。

系统需要理解不同上下文的重要程度，并进行智能资源分配。

三、Multi-Agent Optimization

多 Agent 优化

未来企业将同时运行大量不同角色的 Agent。

基础设施需要支持 Agent 之间的协同、隔离和调度。

四、Enterprise-Scale Efficiency

企业级效率

企业 AI 的核心指标不只是单次性能，而是：

- 每个 Agent 的运行成本；
- 整体系统效率；
- 长期运营能力。

3.5 Agent Runtime：连接智能与计算的核心层

在 Agent 时代，Runtime 将成为 AI 基础设施的重要组成部分。

传统软件时代，操作系统负责管理计算资源。

云计算时代，云平台负责管理基础设施资源。

Agent 时代，需要新的 Runtime 管理智能资源。

Agent Runtime 负责：

- Agent 生命周期管理；
- 知识访问；
- 任务调度；
- 状态保存；
- 上下文管理；
- 安全治理。
- 工具调用；

它连接上层企业应用和底层计算基础设施。

因此，未来 AI 系统的核心架构将从：

Application → Model → Hardware

演进为：

Application → Agent Runtime → Compute Infrastructure

3.6 面向 Agent 生命周期的计算架构

Agent 的运行具有明显生命周期。

一个 Agent 可能经历：

- 创建（Create）；
- 配置（Configure）；
- 运行（Execute）；
- 协同（Collaborate）；
- 休眠（Suspend）；
- 恢复（Resume）。

不同阶段对计算资源的需求不同。

因此，未来计算系统需要根据 Agent 状态动态调度资源。

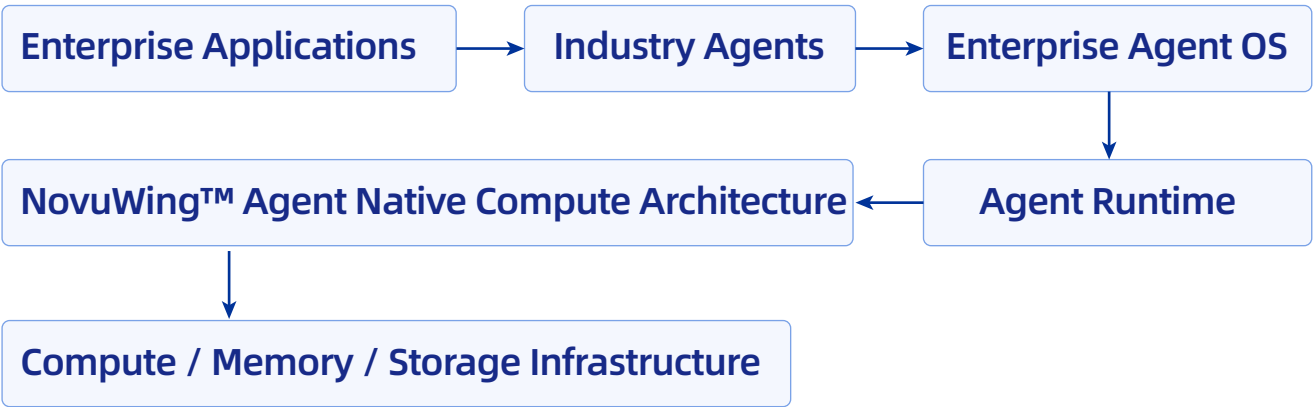
例如：

- 正在执行任务的 Agent，需要高性能计算资源；
- 暂时等待的 Agent，需要保存上下文，但无需持续占用高价值计算资源；
- 长期知识和历史状态，需要更大容量、更高成本效率的存储层。

这种以 Agent 生命周期驱动资源管理的方式，是 NovuWing™ Agent Native Compute 与传统 AI Server 的根本区别。

3.7 Agent 原生计算总体架构

面向企业 Agent 时代，新的 AI 基础设施将形成完整分层架构：



其中：

- Enterprise Applications 定义业务需求；
- Agents 执行业务任务；
- Agent OS 提供企业级管理能力；
- Agent Runtime 管理智能运行过程；
- NovuWing™ Agent Native Compute 提供持续运行所需的计算基础。

3.8 小结

Agent 时代正在改变 AI 基础设施的设计目标。

过去：

更快的模型推理。

未来：

更高效的智能运行。

NovuWing™ Agent Native Compute 不是简单提升服务器性能，而是重新定义 AI 计算系统。它让计算基础设施从支持模型，走向支持企业智能。

第四部分 Agent 内存架构

PART 4 AGENT MEMORY ARCHITECTURE

面向 Agent 时代重新定义 AI 内存架构

4.1 从计算能力到内存能力：AI 基础设施的新挑战

在模型时代，AI 系统的核心竞争力主要体现在计算能力。
更大的模型规模、更高的 GPU 算力、更快的推理速度，推动了人工智能能力的快速提升。
然而，当 AI 进入 Agent 时代，系统瓶颈正在发生变化。
Agent 不只是执行一次推理，而是在企业环境中持续运行。
它需要长期维护当前任务上下文、历史交互信息、工具调用状态、企业知识访问状态以及长期记忆。
未来 AI 系统的核心竞争，不仅是计算能力竞争，也是内存架构竞争。

技术范围说明 / TECHNICAL SCOPE

本文所述 Agent Memory Architecture 包含两个相互关联但不完全相同的层面：一是面向推理运行态数据的内存与缓存分层，包括上下文状态、KV Cache 及 Agent 可恢复运行状态；二是面向长期记忆、企业知识与业务数据的数据和知识基础设施。三级智能内存架构与 Intelligent KV Cache Scheduling 主要聚焦第一类运行态数据。长期记忆、知识库和业务数据可以使用底层存储与检索能力，但其数据模型、访问方式和生命周期不等同于 KV Cache。

4.2 KV Cache：Agent 时代的新型关键资源

在大语言模型推理过程中，KV Cache 用于保存已经计算完成的 Key 和 Value 信息，通过缓存历史上下文避免重复计算。
在企业 Agent 场景中，一个 Agent 可能持续处理长周期任务、多轮业务流程、大量企业知识和多系统交互记录。
随着上下文长度和 Agent 数量增长，KV Cache 规模快速扩大。
KV Cache 不再只是模型优化技术，而成为企业 AI 基础设施中的关键资源。

4.3 HBM 的价值与限制

HBM 为现代 AI 加速器提供极高的数据访问带宽，是大模型训练和推理的重要基础。
但在 Agent 时代，如果所有 Agent 状态长期驻留 HBM，将面临容量不足、成本过高和扩展困难。
因此，未来 AI 基础设施不能简单依赖更大的 HBM，而需要建立更加智能的内存体系。

4.4 面向 Agent 的三级智能内存架构

星流计算提出 Three-Tier Intelligent Memory Architecture。
该架构主要面向 Agent 推理运行过程中的上下文状态、KV Cache 及可恢复运行状态，根据数据访问频率、实时性要求、恢复成本和 Agent 生命周期状态进行动态分层管理。
三级智能内存架构不等同于企业全部数据与知识存储体系。Agent 长期记忆、企业知识库和业务数据属于更广义的数据与知识基础设施，可使用底层容量资源，但不属于 Intelligent KV Cache Scheduling 的直接调度对象。

L0: HBM 高性能层 (Hot Tier)

用于保存模型推理所需的活跃 KV Cache、当前高频上下文状态及其他对时延和带宽高度敏感的运行态数据。

L1: DDR 温层 (Warm Tier)

用于保存暂时非活跃但可能在较短时间内重新使用的 KV Cache、阶段性上下文和可快速恢复的 Agent 运行状态。

L1.5: 扩展温层 (Extended Warm Tier)

由 CXL.mem / NVMe 等容量型资源提供，用于承载较低活跃度、较长时间不访问但仍需保留的 KV Cache、上下文快照和暂停 Agent 的可恢复运行状态。CXL.mem / NVMe 在访问语义和性能特征上不同：CXL.mem 更接近扩展内存，NVMe 更适合作为 KV offload、快照与恢复后端。

通过三级内存架构，不同状态的数据可以在不同层级之间智能流动。
需要特别说明的是，企业知识库、向量索引、长期业务数据以及 Agent 的持久化长期记忆，并不等同于 KV Cache。它们可以部署在 L1.5 所依托的容量资源或其他数据基础设施之上，但由知识管理、检索增强、持久化存储和数据治理机制负责管理，而不是由 Intelligent KV Cache Scheduling 直接调度。

4.5 Intelligent KV Cache Scheduling

三级内存架构的核心，不只是增加存储层级，而是实现智能调度。
系统需要理解：

- 什么数据正在被使用；
- 什么数据可以迁移；
- 什么数据即将被使用；
- 什么数据需要长期保存。

Intelligent KV Cache Scheduling 根据 Agent 生命周期和上下文价值，对 KV Cache 进行动态管理。

核心专利

申请号：202611065591.X

《基于 Agent 语义的三级分层 KV 缓存智能调度方法及系统》

这使 AI 基础设施从静态资源管理，走向面向智能生命周期的动态资源管理。

4.6 从 Memory Hierarchy 到 Intelligence Hierarchy

传统计算系统中的内存层级主要依据访问速度和成本划分。

但 Agent 时代需要进一步理解数据价值。

- 当前任务上下文承载实时运行状态；
- 历史上下文与可恢复状态具有不同的访问价值；
- 企业知识与长期记忆则构成更广义的组织智能基础。

因此，面向 Agent 的基础设施，需要从单纯的 Memory Hierarchy，进一步走向对智能状态价值进行分层管理的 Intelligence Hierarchy。基础设施不仅存储数据，更需要管理智能状态。

这里所说的 Intelligence Hierarchy，并不是将所有企业知识和业务数据都纳入同一套缓存调度机制，而是强调基础设施在管理运行态数据时，需要同时考虑数据的访问频率、恢复成本、所属 Agent、任务阶段和业务优先级。运行态 KV Cache 的分层调度，与长期记忆和企业知识资产的持久化管理，应由相互协同但职责清晰的机制分别承担。

4.7 Agent Memory Architecture 的价值

面向企业 AI，Agent Memory Architecture 带来的价值包括：

- 提高资源利用效率；
- 提升 Agent 扩展能力；
- 降低企业 AI 总体成本；
- 支持持续智能运行。

这使 AI 基础设施从单纯计算平台，演进为智能运行平台。

4.8 小结

Agent 时代改变了 AI 基础设施的核心挑战。

过去，计算能力决定 AI 性能。

未来，计算能力与内存架构共同决定智能规模。

KV Cache、上下文管理和 Agent 生命周期，正在成为企业 AI 基础设施的新核心。

通过 Agent Memory Architecture，星流计算希望解决：

如何让越来越多 Agent 在企业环境中持续、高效、可靠地运行。

第五部分 玄矩™ A1 架构

PART 5 XUANJU A1 ARCHITECTURE

玄矩™ A1：第一代 Agent Native AI Server

5.1 从计算理念到产品实现

Agent Native Compute 的提出，并不是对传统 AI Server 的简单升级，而是针对企业 Agent 时代重新定义计算基础设施。

传统 AI Server 的设计目标，是提升模型训练和推理效率。

而玄矩™ A1 的设计目标，是支撑企业级 Agent 长期、稳定、高效运行。

它将 Agent 生命周期、上下文管理、计算资源调度和企业级部署需求纳入整体架构设计。

玄矩™ A1 是星流计算基于 NovuWing™ Agent Native Compute Architecture 打造的第一代 Agent Native AI Server。

它代表了一种新的计算范式：

- 从支持模型计算，
- 走向支持智能运行。

5.2 玄矩™ A1 的设计目标

面向企业 Agent 场景，玄矩™ A1 围绕四个核心目标进行设计：

- 支持持续智能运行（Continuous Intelligence）：企业 Agent 不再是一次性的请求响应，而需要长期运行。
- 玄矩™ A1 面向持续任务、长期上下文和复杂业务流程进行优化。
- 提升多 Agent 承载能力（Multi-Agent Scalability）：未来企业将运行大量不同角色的 Agent。
- 计算基础设施需要支持多 Agent 并发、隔离和资源调度。
- 优化 AI 基础设施效率（Infrastructure Efficiency）：企业关注的不仅是峰值性能，更关注长期运行成本。
- 玄矩™ A1 通过系统级资源协同，提高整体计算效率。
- 满足企业级可靠性要求（Enterprise Reliability）：面向生产环境，AI 基础设施需要具备稳定、安全和可持续运营能力。

5.3 Agent 原生服务器总体架构

玄矩™ A1 采用面向 Agent 的系统级架构设计。

整体架构由以下核心层组成：

Application Layer	企业业务应用和行业智能场景。
Agent Layer	负责具体业务任务执行的企业 Agent。
Agent Runtime Layer	负责 Agent 生命周期管理、任务调度、上下文管理和工具调用。
Compute Layer	提供 AI 推理和计算能力。
Memory Layer	提供面向 Agent 的智能内存体系。
Storage Layer	提供长期数据和知识资产存储。

通过各层协同，玄矩™ A1 将计算、内存和运行环境统一设计，使 AI 基础设施能够真正服务企业智能运行。

5.4 面向 Agent 的计算架构

传统 AI Server 主要优化单模型推理性能。

玄矩™ A1 更关注 Agent 工作负载。

Agent 工作负载具有：

因此，玄矩™ A1 需要在计算层面支持：

- 长上下文；
- 多步骤任务；
- 多轮交互；
- 动态资源需求；
- 持续运行状态。
- 高效推理计算；
- 动态资源分配；
- 多任务并行处理；
- Agent 状态恢复。

其核心思想是：

计算资源不再只是服务模型，而是服务智能生命周期。

5.5 面向 Agent 的内存体系

在 Agent 时代，内存成为影响系统效率的关键因素。玄矩™ A1 基于 Agent Memory Architecture 理念，构建面向推理运行态数据和可恢复 Agent 状态的分层内存体系。系统根据数据实时性、访问频率、恢复成本和生命周期状态，对 KV Cache、上下文状态及相关运行态数据进行分层管理。企业知识库、长期业务数据和 Agent 持久化记忆由上层数据与知识体系负责管理，可使用玄矩™ A1 提供的存储和扩展能力，但不与 KV Cache 使用完全相同的调度机制。

高频实时状态：	由高性能内存承担。
活跃上下文状态：	由容量和性能平衡的内存承担。
长期历史状态：	由大容量存储承担。

这种设计使系统能够在性能、容量和成本之间实现动态平衡。

5.6 Agent 缓存调度

玄矩™ A1 的核心设计之一，是面向 Agent 的缓存调度能力。传统缓存管理主要关注数据访问频率。而 Agent 场景需要进一步理解：

- 数据属于哪个 Agent 或 Session；
- Agent 当前处于 Active、Idle、Deep-Idle 还是 Recovering 状态；
- 当前任务处于什么阶段，数据再次被访问的概率有多高；
- 数据迁移、重新加载或重新计算的恢复成本是多少；
- 数据是否属于可调度的推理运行态数据，而不是长期知识或业务数据。

因此，系统需要根据 Agent 生命周期动态管理缓存资源。通过智能调度，使有限的高性能资源优先服务最需要的智能任务。

5.7 面向企业部署的系统设计

企业 AI 基础设施不仅需要性能，也需要稳定运行能力。玄矩™ A1 面向企业生产环境设计，关注：

- 长时间稳定运行；
- 企业级安全要求；
- 资源利用效率；
- 未来扩展能力。
- 运维管理能力；

它不是一个实验环境中的 AI 加速设备，而是企业智能时代的新型基础设施节点。

5.8 玄矩™ A1的技术价值

玄矩™ A1的价值，不只是提供更强的 AI 计算能力。
更重要的是，它重新定义了 AI Server 的角色。

过去：

AI Server 是模型计算设备。

未来：

AI Server 将成为企业智能运行平台。

玄矩™ A1通过 NovuWing™ Agent Native Compute 架构，让基础设施能够理解 Agent、支持 Agent，并服务 Agent 的持续演进。

5.9 小结

Agent 时代需要新的计算基础设施。
玄矩™ A1不是传统 AI Server 的延伸，而是面向企业智能时代重新设计的计算平台。
它将计算、内存和 Agent Runtime 协同起来，为企业构建持续运行的 AI 能力提供基础。

第六部分 AI 基础设施经济学

PART 6 AI INFRASTRUCTURE ECONOMICS

重新定义企业 AI 时代的成本模型

6.1 从算力规模到智能经济：AI 基础设施评价体系正在改变

在模型时代，AI 基础设施的核心评价指标通常集中于计算能力。

更大的 GPU 规模、更高的 FLOPS、更快的推理速度，成为衡量 AI 系统能力的重要标准。

然而，当企业 AI 进入 Agent 时代，评价体系正在发生变化。

企业不再只是关注：

模型是否能够运行。

而更加关注：

- 多少 Agent 可以稳定运行；
- 每一个智能任务的成本是多少；
- 整体系统是否具备长期运营价值。

因此，AI 基础设施的竞争正在从单纯的算力竞争，走向智能经济（AI Infrastructure Economics）竞争。未来衡量 AI 基础设施价值的核心，不只是峰值性能，而是单位智能产出的综合效率。

6.2 为什么更强算力并不等于更高 AI 价值

过去几年，企业 AI 投资大量集中在 GPU 资源。

GPU 提供强大的并行计算能力，是大模型发展的核心基础。

但在企业实际部署中，仅增加 GPU 数量并不能线性提升 AI 价值。

原因包括：

- GPU 利用率受到工作负载影响：企业 Agent 通常具有复杂生命周期，不会始终处于满负载计算状态。
- 内存和上下文成为新的瓶颈：大量 Agent 状态和 KV Cache 会占用宝贵资源。
- 系统运营成本持续增加。

包括：

- 硬件成本；
- 能源成本；
- 运维成本；
- 扩展成本。

因此，企业需要的不只是更多算力，而是更高效率的智能基础设施。

6.3 评测方法

本白皮书中的性能、效率与 TCO 相关表述，应在统一的测试边界和可复现的工作负载条件下进行评估。本方法论旨在建立统一、透明、可复现的评估与比较口径。

本白皮书 Version 1.0 聚焦 Agent Native Compute 的架构理念、系统设计与评估方法，不以产品级性能测试结果作为主要内容。为建立面向 Agent 工作负载的可重复评估体系，本节定义 Benchmark Framework、测试边界及核心评价指标。

玄矩™ A1 的产品级 Benchmark 数据将在完成标准化、可重复验证的测试后，以独立技术报告形式发布。

评测框架 / BENCHMARK FRAMEWORK

评估范围 / Evaluation Scope: 对比系统需采用等效模型、精度、服务等级与部署边界。

工作负载 / Workload Profile: 记录模型版本、上下文长度、并发 Agent 数量、任务步骤、工具调用模式及运行时长。

系统配置 / System Configuration: 记录 CPU、GPU、HBM、DDR、CXL.mem / NVMe、网络、软件栈、Runtime 与调度策略。

核心指标 / Primary Metrics: Agent Throughput、Task Completion Rate、P50/P95 Latency、Context/KV Cache Efficiency、Resource Utilization、Cost per Agent、Cost per Task 与 TCO。

测量流程 / Measurement Protocol: 预热后进行多轮重复测试，报告测试窗口、样本量、均值及尾延迟，并说明异常值处理方式。

可复现性 / Reproducibility: 保留配置、日志、监控数据与成本假设；所有对比结果应标注测试日期和适用条件。

注：本版本仅定义 Benchmark 方法、测试边界与报告规范。具体产品性能、资源利用率、Agent 承载规模及 TCO 测试结果，将在完成标准化、可重复验证的测试后另行发布。

6.4 从 FLOPS 到单位智能成本

传统 AI 系统主要关注 FLOPS（每秒浮点运算次数）。这一指标能够衡量计算能力，但无法完整反映企业 AI 的真实价值。

在 Agent 时代，更重要的问题是：

单位成本能够支撑多少智能任务？

因此，需要新的评价方式：

Cost per Agent 每一个 Agent 的运行成本。

Cost per Task 每一个业务任务的完成成本。

Intelligence Efficiency 单位资源产生的智能效率。

这种评价体系，将推动 AI 基础设施从计算性能导向，转向业务价值导向。

6.5 Agent 时代的主要成本挑战

企业部署 Agent 系统，成本主要来自多个方面：

- 计算成本：包括 GPU、CPU 以及相关计算资源。
- 内存成本：随着上下文长度增加，KV Cache 和状态管理带来巨大资源压力。
- 存储成本：企业 Agent 需要长期保存知识、历史状态和业务数据。
- 运营成本：包括系统管理、安全治理和持续优化。

未来企业 AI 基础设施需要在这些维度之间实现动态平衡。

高性能不是唯一目标。

高效率、可扩展和可持续运营，才是企业 AI 规模化发展的关键。

6.6 Agent 原生计算对 AI 经济模型的重新定义

Agent Native Compute 的核心价值，不只是提升单次计算性能。它通过系统级优化，重新定义 AI 基础设施经济模型。

从架构设计目标看，Agent Native Compute 通过生命周期感知、分层内存管理和软硬件协同，旨在改善企业 AI 基础设施的资源效率与单位智能成本。以下内容描述其主要优化方向，不代表本版本已经完成的产品级 Benchmark 结果：

- 提高资源利用率：通过 Agent 生命周期感知和智能资源调度，让计算资源服务于真正需要的位置。
- 降低上下文管理成本：通过智能内存架构和 KV Cache Scheduling，提高资源使用效率。
- 提高 Agent 承载规模：让企业能够在同等基础设施条件下运行更多智能任务。
- 降低长期运营成本：通过软硬件协同，提升整体 TCO 表现。

6.7 从服务器采购到智能基础设施投资

未来企业购买 AI 基础设施，不应只关注硬件规格。

传统服务器采购关注：

- CPU 数量；
- GPU 数量；
- 内存容量；
- 存储容量。

而企业 AI 时代，更应该关注：

- 能够支持多少业务 Agent；
- 能够承载多少智能任务；
- 能够持续运行多久；
- 每个智能任务的综合成本。

AI 基础设施正在从硬件采购，演变为企业智能能力建设。

服务器不再只是 IT 设备，而成为企业未来智能生产力的重要组成部分。

6.8 玄矩™ A1 的经济价值

作为第一代 Agent Native AI Server，玄矩™ A1 的设计目标不是简单提供更高峰值性能。

它关注的是企业 AI 长期运行过程中的综合效率。

通过：

- NovuWing™ Agent Native Compute 架构；
- 面向 Agent 的内存体系；
- 智能缓存调度；
- 系统级资源协同。

玄矩™ A1 希望帮助企业降低 AI 应用规模化部署的成本，并提升智能基础设施投资回报。

其价值不在于替代更多 GPU，而在于让每一份 AI 基础设施投入产生更大的智能价值。

6.9 小结

AI 基础设施的经济价值，应由可交付的业务结果衡量，而不能只由峰值算力衡量。企业需要同时观察 Agent 承载规模、任务完成质量、上下文资源效率、服务可靠性与单位任务成本，并在统一的服务等级目标下比较不同方案。

因此，基础设施优化的目标不是简单减少硬件投入，而是在满足安全、时延和可用性要求的前提下，提高单位资源能够稳定交付的智能任务数量。

第七部分 企业级全栈 AI 基础设施

PART 7 ENTERPRISE FULL-STACK AI INFRASTRUCTURE

从计算基础设施到企业智能基础设施

7.1 企业 AI 正进入全栈智能时代

人工智能正在从单一模型能力竞争，进入完整智能系统竞争。

过去，企业部署 AI 往往关注单一能力：

- 选择一个大模型；
- 建设一套推理环境；
- 开发若干 AI 应用。

这种方式可以解决局部问题，但难以支撑企业级 AI 的长期发展。

随着 Agent 成为新的工作主体，企业需要的不再是一个孤立模型，而是一套完整的智能基础设施。它需要连接：

- 企业应用；
- 业务流程；
- Agent；
- 知识体系；
- 计算资源；
- 运行环境；
- 安全治理。

7.2 企业级全栈 AI 基础设施

星流计算提出：

Enterprise Full-Stack AI Infrastructure

其核心理念是：

让企业从“使用 AI”，走向“拥有 AI”。

一套完整的企业 AI 基础设施，需要覆盖从底层计算到上层业务智能的完整链路：

应用层（Application Layer）

承载企业业务场景。

智能体层（Agent Layer）

负责业务任务执行。

运行层（Agent Runtime Layer）

负责 Agent 生命周期管理。

计算层（Compute Layer）

提供 AI 推理和计算能力。

基础设施层（Infrastructure Layer）

提供计算、内存、存储和网络能力。

通过全栈协同，企业 AI 才能够从实验走向生产。

7.3 玄矩™ A1 与 SerpMind™：双平台协同架构

面向企业 Agent 时代，星流计算构建两大基础平台：

玄矩™ A1

Agent Native AI Server

负责提供面向 Agent 的计算基础设施。

SerpMind™

Enterprise Agent OS

负责企业 Agent 的创建、运行、管理和治理。

两者共同形成：

Hardware + Runtime + Application

的完整智能基础。

玄矩™ A1 解决：

Agent 如何高效运行

SerpMind 解决：

Agent 如何被企业构建、管理和持续使用

二者协同，使企业 AI 从单点能力走向完整智能体系。

SerpMind™ Enterprise Agent OS 是企业 Agent 的管理与运行平台；Agent Runtime 是其中连接 Agent、模型、工具、知识与底层计算资源的核心运行时层。

7.4 Agent Runtime：企业智能运行的新中间层

在传统软件架构中，操作系统连接应用和硬件。

在云计算时代，云平台连接应用和基础设施。

在 Agent 时代，需要新的智能运行层。

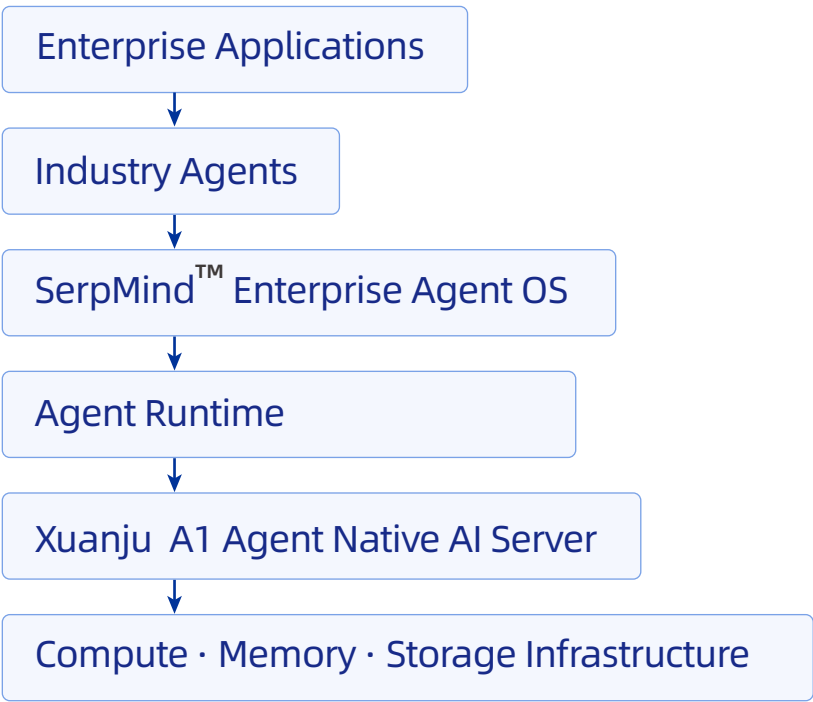
Agent Runtime 成为连接企业应用、Agent、模型、工具、知识与计算资源的核心中间层。

它负责：

- Agent 生命周期管理；
- 任务规划与执行；
- 上下文管理；
- 工具调用。
- 状态保存；
- 权限控制；
- 运行监控；

7.5 企业 AI 全栈架构

面向未来企业智能应用，完整架构如下：



其中：

- 应用定义业务目标；
- Runtime 管理智能运行；
- 玄矩™ A1 提供持续运行所需的计算基础。
- Agent 执行业务任务；
- Agent OS 提供企业级管理能力；

这是一套从业务到基础设施完整闭环的企业 AI 架构。

7.6 从工具到智能资产

传统软件时代，企业积累的是：

- 数据资产；
- 流程资产；
- 系统资产。

Agent 时代，企业将进一步积累智能资产。企业 Agent 会持续沉淀：

- 业务经验；
- 知识体系；
- 执行流程；
- 决策模式。

这些智能资产，将成为企业未来竞争力的重要组成部分。因此，企业 AI 的目标不是简单增加 AI 功能，而是建立属于企业自己的智能能力。

7.7 企业 AI 规模化落地的关键能力

企业 AI 从试验走向生产，需要同时具备：

- 计算能力：支撑大规模 Agent 持续运行。
- 运行能力：管理 Agent 生命周期和业务执行过程。
- 知识能力：连接企业数据和业务经验。
- 治理能力：保障安全、权限、合规和可控。
- 演进能力：支持企业智能持续优化。

只有这些能力共同组成，企业 AI 才能够真正进入生产环境。

7.8 星流计算的企业级全栈 AI 架构

星流计算希望构建的不只是 AI 产品，而是一套面向企业智能时代的新基础设施。

通过：



提供 Agent Native Compute 能力。

通过：



提供企业 Agent 运行、治理能力以及核心Runtime层。

最终形成：计算基础设施 + Agent 平台 + 企业应用的完整体系。
让每一家企业，都能够拥有属于自己的 AI。

7.9 小结

企业 AI 的未来，不属于单一模型，也不属于单一应用。
它需要完整的基础设施体系。
从计算，到 Runtime，再到 Agent 和业务应用，企业需要重新构建面向智能时代的技术底座。
Enterprise Full-Stack AI Infrastructure
将成为企业进入 Agent 时代的重要基础。

第八部分 行业实践与参考架构

PART 8 INDUSTRY PRACTICES & REFERENCE ARCHITECTURES

从基础设施到行业智能：Agent Native Compute 的行业实践

8.1 从技术创新到产业应用

Agent Native Compute 面向的是企业生产环境中的持续智能运行需求。不同产业在业务流程、数据治理、实时性、安全性和可靠性方面具有不同约束，但其基础设施需求具有一定共性。

本章以医疗、政务和制造为代表场景，说明 Agent Native Compute 在典型行业中的适用方向、基础设施要求与参考架构。相关内容用于描述行业实践路径与架构设计，不构成特定客户项目的性能或业务效果声明。

8.2 行业 Agent 的核心特征

相比通用 AI 助手，行业 Agent 具有更高要求。

- 专业知识理解：行业 Agent 需要理解领域知识、业务规则和专业流程。
- 业务流程执行：行业 Agent 不只是回答问题，而需要参与实际业务。
- 多系统协同：行业任务通常需要连接多个业务系统和数据来源。
- 持续学习和优化：行业 Agent 需要随着企业业务发展持续演进。

因此，行业 Agent 的规模化应用，需要计算基础设施、Agent Runtime 和行业知识体系共同支撑。

8.3 医疗行业：企业智能化的重要实践场景

医疗行业是企业 AI 落地的重要场景。

医疗业务具有：

- 复杂流程；
- 专业知识密集；
- 高数据价值；
- 高安全要求。

传统医疗信息化主要解决业务流程数字化问题。

但未来医疗智能化，需要进一步实现：

从系统记录业务，

走向 Agent 辅助业务。

医疗 Agent 可以围绕：

- 临床辅助；
- 知识管理；
- 医院管理。
- 医疗运营；
- 患者服务；

等场景提供智能能力。

但医疗行业 AI 的规模化应用，需要满足：

- 数据安全；
- 业务可靠；
- 专业可信；
- 持续运行。

这正是企业级 AI 基础设施的重要价值。

8.4 政务与城市治理：从数字化到智能化

政务和城市治理场景具有大量复杂业务流程。

传统数字化系统解决：

- 信息管理；
- 流程审批；
- 数据共享。

而 Agent 时代，可以进一步实现：

- 智能咨询；
- 流程自动化；
- 辅助决策；
- 跨部门协同。

行业 Agent 能够连接政策知识、业务系统和公共数据，帮助政府机构提升服务效率。

但大规模部署需要稳定、安全、可治理的 AI 基础设施作为支撑。

8.5 制造与企业运营：智能生产的新阶段

制造行业正在从自动化走向智能化。

未来企业不仅需要自动化设备，也需要能够理解生产目标、分析运行状态和优化流程的智能 Agent。

制造 Agent 可以应用于：

- 生产优化；
- 设备运维；
- 供应链管理；
- 质量分析；
- 知识传承。

这些场景要求 AI 系统具备：

- 实时性；
- 可靠性；
- 持续运行能力。

NovuWingTM Agent Native Compute 为制造企业构建智能生产体系提供基础支撑。

8.6 从通用架构到行业智能

前文定义的企业 AI 全栈架构，为不同行业构建 Agent 应用提供统一的基础设施底座。

当这一架构进入医疗、政务、制造等具体行业时，其核心变化不在于底层技术栈本身，而在于 Agent、知识、工具与业务流程呈现出不同的行业特征。

- **医疗领域**，Agent 需要连接临床知识、医疗数据与诊疗流程；
- **政务领域**，Agent 需要连接政策知识、政务数据与跨部门流程；
- **制造领域**，Agent 需要连接设备、生产数据、工业知识与运营流程。

因此，行业智能并不意味着为每个行业重新构建一套 AI 基础设施，而是在统一的企业 AI 基础设施之上，形成面向不同产业的 Agent、知识、工具与业务应用体系。

统一基础设施，承载不同产业智能。

这些场景要求 AI 系统具备：

- 基础设施提供持续运行能力；
- Runtime 管理 Agent 生命周期；
- Agent 执行业务任务；
- 行业应用创造业务价值。

这是一套面向行业智能时代的新型技术架构。

8.7 从行业软件到行业智能平台

过去二十年，企业通过软件系统实现业务数字化。

未来十年，企业将通过 Agent 系统实现业务智能化。

软件时代：企业购买软件

AI 时代：企业将拥有智能能力

行业 Agent 将成为企业新的生产力载体。

而支撑这些 Agent 长期运行的基础设施，将成为企业智能时代的重要底座。

8.8 小结

人工智能正在进入产业深水区。真正有价值的企业AI，不只是模型能力展示，而是能够进入真实业务流程，并在生产环境中持续、可靠、可治理地运行。

医疗、政务、制造等行业对数据、流程、安全和可靠性具有不同要求，但都需要更加系统化的 Agent 运行基础设施。Agent Native Compute 通过计算、内存、Runtime 与企业级治理能力的协同，为这些行业构建 Agent 应用提供统一的基础设施参考。

随着行业实践不断深入，企业 AI 基础设施将进一步从“支持模型部署”走向“支撑智能持续运行”。

第九部分 企业 AI 基础设施的未来

PART 9 FUTURE OF ENTERPRISE AI INFRASTRUCTURE

企业智能时代的新基础设施

9.1 AI 基础设施正在进入新的发展阶段

企业 AI 基础设施正在从“承载模型”走向“承载智能系统”。随着 Agent 进入核心业务流程，未来基础设施的竞争重点将进一步转向持续运行能力、系统效率、治理能力与智能规模。

这意味着，AI 基础设施将不再只是模型计算平台，而会逐步成为企业智能能力长期运行和持续演进的基础。

9.2 从模型智能到企业智能

模型赋予 AI 通用智能能力。

但企业真正需要的是能够融入业务流程的企业智能。

企业智能需要连接：

- 业务系统；
- 专业经验；
- 计算资源。
- 企业知识；
- 组织流程；

未来，每一家企业都将拥有属于自己的智能系统。

这些智能系统不是简单的软件功能，而是企业新的生产力基础。

因此，企业 AI 的竞争，将从单一模型能力竞争，走向完整智能基础设施竞争。

9.3 Agent Native Compute：下一代企业 AI 计算范式

随着 Agent 数量增长和业务复杂度提升，传统 AI 基础设施将难以满足企业长期运行需求。

未来计算架构需要：

- 支持持续智能运行；
- 支持多 Agent 协同；
- 支持企业级安全和治理；
- 支持长期上下文管理；
- 支持规模化部署。

NovuWing™ 是星流计算面向这一范式构建的技术架构，并将通过玄矩™ A1、SerpMind™ 及后续产品持续演进。

计算基础设施将从：支持模型计算 走向：支持智能运行

9.4 企业 AI 基础设施的演进方向

随着 Agent 逐步进入企业核心业务流程，企业 AI 基础设施将从以模型部署和推理为中心，进一步演进为面向智能持续运行的系统架构。
这一演进将呈现三个方向：

从模型计算走向智能运行

基础设施的优化对象将从单次模型调用，扩展到 Agent 生命周期、上下文状态与持续运行过程。

从单一资源优化走向系统级协同

CPU、GPU、Memory、Storage 与 Runtime 将从相对独立的资源与软件组件，走向更紧密的系统级协同。

从部署 AI 能力走向构建企业智能基础设施

企业将逐步形成能够承载 Agent、连接业务系统，并持续积累知识、流程与运行经验的智能基础设施能力。

计算基础设施将从“支持模型计算”，走向“支持智能持续运行”。

9.5 星流计算的使命

星流计算希望构建面向企业智能时代的新一代 AI 基础设施。

通过：

Agent Native Compute

重新定义 AI 计算架构。

通过：

Enterprise Agent OS

帮助企业构建、管理和运营智能 Agent。

通过：

Enterprise Full-Stack AI Infrastructure

连接计算、软件和行业应用。

让企业不只是使用 AI，而是真正拥有 AI。

9.6 企业 AI，从这里开始

未来十年，人工智能将成为企业新的基础设施。

就像过去企业拥有 ERP、CRM 和云平台一样，未来每一家企业都将拥有自己的 AI。

这不仅是技术演进，更是企业生产方式的改变。

- 从模型智能，到 Agent 智能；
- 从软件系统，到智能系统；
- 从数字化企业，到智能化企业。

新的时代，需要新的基础设施。

Agent Native Compute，将成为企业进入智能时代的重要计算范式。NovuWing™ 将作为星流计算面向这一时代持续演进的核心技术架构。

9.7 结语

人工智能的未来，不只是拥有更大的模型。

真正重要的是：

- 如何让智能持续运行；
- 如何让智能进入业务；
- 如何让智能成为企业自身能力。

星流计算相信：

未来属于那些能够构建、管理并拥有自己智能能力的企业。

而面向 Agent 时代的新一代 AI 基础设施，将成为这一未来的重要支撑。

第十部分 玄矩™ A1 技术架构与规格概览

PART 10 XUANJU A1 TECHNICAL ARCHITECTURE & SPECIFICATION OVERVIEW

产品架构、系统能力与规格体系

10.1 产品定位

玄矩™ A1是星流计算基于 NovuWing™ Agent Native Compute 理念打造的第一代 Agent Native AI Server。

不同于传统 AI Server 主要面向模型训练和推理优化，玄矩™ A1 面向企业 Agent 持续运行场景设计。其核心目标不是单纯提升峰值计算性能，而是通过计算、内存、运行环境和系统调度协同，为企业提供面向智能时代的新一代 AI 基础设施。

10.2 系统设计理念

玄矩™ A1 遵循 NovuWing™ Agent Native Compute 的核心设计原则：

- Continuous Intelligence：支持企业 Agent 长时间稳定运行。
- Context-Aware Computing：围绕上下文、状态和知识访问进行资源优化。
- Multi-Agent Optimization：支持多 Agent 并发运行和资源协同。
- Enterprise Reliability：满足企业级部署所需的稳定、安全和可管理能力。

玄矩™ A1 将 AI Server 从单纯计算设备，升级为企业智能运行节点。

10.3 系统架构

玄矩™ A1 采用面向 Agent 的系统级架构。
整体由以下层次组成：

Agent Runtime Layer	负责 Agent 生命周期管理、任务调度、上下文管理和资源协调。
Compute Layer	提供 AI 推理和计算能力。
Memory Layer	提供面向 Agent 的智能内存体系。
Storage Layer	提供长期数据、知识和状态存储。
Network Layer	提供高效的数据交换和集群互联能力。

10.4 计算子系统

玄矩™ A1面向企业 AI 推理和 Agent 工作负载设计。

计算子系统重点支持：

- 大模型推理；
- 企业 Agent 执行；
- 多任务并行计算；
- 行业智能应用。

相比传统 AI Server 关注单模型吞吐，玄矩™ A1更关注：

单位基础设施可承载的智能任务规模

计算资源根据 Agent 运行状态进行动态调度，提高整体资源利用效率。

10.5 智能内存系统

玄矩™ A1采用面向 Agent 的内存体系。

核心包括：

- L0：HBM 高性能层，用于活跃 KV Cache、实时推理上下文和高频运行状态。
- L1：DDR 温层，用于较短时间内可能重新激活的 KV Cache、阶段性上下文和可快速恢复状态。
- L1.5：CXL.mem / NVMe 扩展温层，用于较低活跃度 KV Cache、上下文快照及暂停 Agent 的可恢复运行状态；其中 CXL.mem 与 NVMe 分别提供扩展内存和持久化 offload / snapshot 能力。

通过 Intelligent KV Cache Scheduling，实现不同 Agent 状态在不同存储层之间动态流转。该调度主要针对 KV Cache、上下文和可恢复运行状态；企业知识库、业务数据和长期记忆由独立的数据与知识管理机制负责。

10.6 Agent Runtime 协同能力

玄矩™ A1与企业 Agent Runtime 深度协同。

系统能够感知：

- Agent 当前运行状态；
- 任务优先级；
- 上下文价值；
- 资源需求变化。

通过 Runtime 与硬件协同，实现：

- 计算资源智能分配；
- 上下文高效管理；
- Agent 生命周期优化。

这使基础设施从静态计算平台演进为智能运行平台。

10.7 企业级部署能力

面向企业生产环境，玄矩™ A1关注：

- 长时间稳定运行；
- 高可靠性设计；
- 安全部署能力；
- 运维管理能力；
- 横向扩展能力。

玄矩™ A1适用于：

- 企业私有 AI 基础设施；
- 行业智能平台；
- Agent 应用运行节点；
- 企业级 AI 推理环境。

10.8 技术规格说明

玄矩™ A1 已完成产品定型。本白皮书聚焦 Agent Native Compute 的技术架构、系统设计与核心能力，因此不展开具体硬件型号、容量、数量及接口配置等完整产品参数。

玄矩™ A1 的详细硬件配置、性能参数及可选配置，以正式产品规格书（Product Datasheet）及相应型号配置说明为准。

主要规格维度包括：

计算能力： CPU 架构与型号 GPU 加速能力 AI 推理性能指标	内存系统： HBM 容量 DDR 容量 CXL.mem / NVMe容量 Memory Hierarchy 支持能力。	扩展能力： PCIe 规格 网络接口 存储扩展能力	系统能力： Agent Runtime 支持 AI Framework 支持 模型部署支持 集群管理能力
---	--	------------------------------------	--

注：具体硬件配置、性能参数及可选配置，以玄矩™ A1 正式产品规格书及对应型号配置说明为准。

10.9 玄矩™ A1 的技术价值

玄矩™ A1 的核心价值，不在于成为又一款 AI Server。

它代表了一种新的基础设施方向：

从模型计算设备，
走向 Agent 智能运行平台。

通过 NovuWing™ Agent Native Compute 架构，玄矩™ A1为企业构建持续、高效、可靠的智能能力提供基础。

附录 核心架构图

APPENDIX CORE ARCHITECTURE DIAGRAMS

以下三张图分别展示 Agent 原生计算、三级智能内存与企业级全栈 AI 基础设施。三者共同构成从智能运行到底层资源的完整技术体系。

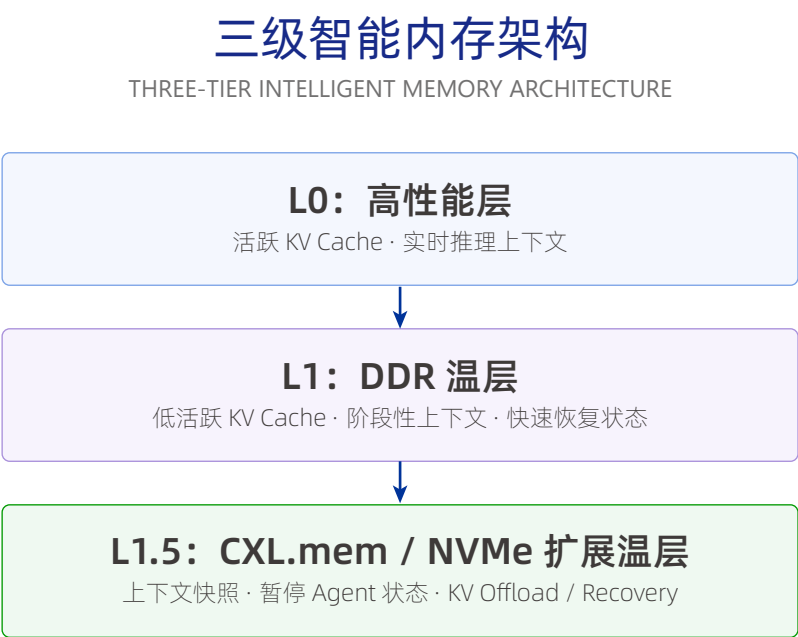


图3 Agent 原生计算架构

Figure 3. Agent Native Compute Architecture

附录 核心架构图

APPENDIX CORE ARCHITECTURE DIAGRAMS



注：三级智能内存架构主要管理推理运行态数据。企业知识库、长期记忆与业务数据属于更广义的数据与知识基础设施，不属于 Intelligent KV Cache Scheduling 的直接调度对象。

图4 三级智能内存架构

Figure 4. Three-Tier Intelligent Memory Architecture

附录 核心架构图

APPENDIX CORE ARCHITECTURE DIAGRAMS



图5 企业级全栈 AI 基础设施

Figure 5. Enterprise Full-Stack AI Infrastructure

架构设计原则

- AI 基础设施服务于 Agent 的持续运行，而不仅是单次模型推理；
- 内存与上下文管理是 Agent 时代的关键资源能力；
- 企业 AI 需要计算、运行时、治理与业务应用协同形成完整体系。



Agent Native Compute

面向企业 Agent 时代的新一代 AI 计算基础设施

浙江星流计算科技有限公司

www.novuinfra.cn