希姆计算
STREAM COMPUTING

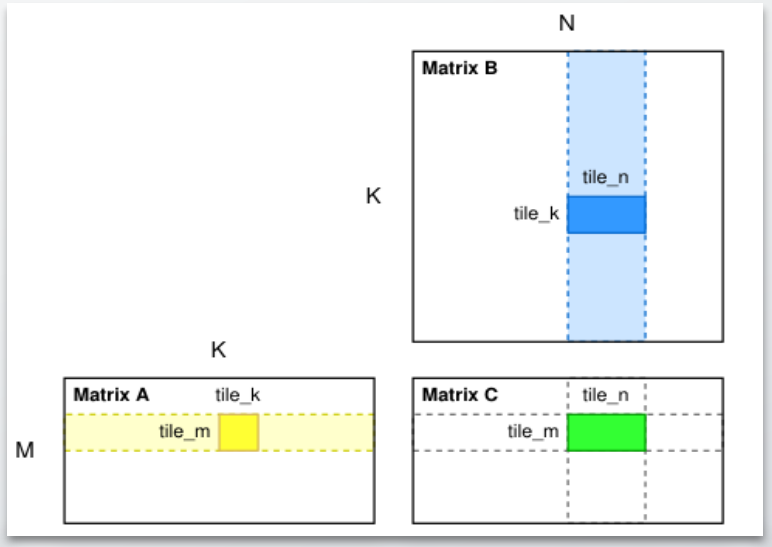
RISC-V加速AI创新

行业智能体的实践

主讲人：肖正宇

指令集标准化

- Tile-based Matrix Multiplication.
- RISC-style Instructions & GPR Architecture.
- Configurable Parameters for Implementations.
- Separate Tile Registers & Accumulation Registers.
  - 8 architectural tile registers, tr0 ~ tr7
  - 8 architectural accumulation registers, acc0 ~ acc7
- Decoupled with Implementations.
  - SIMD, WS Array, OS Array, MAC Cube, CIM, etc.



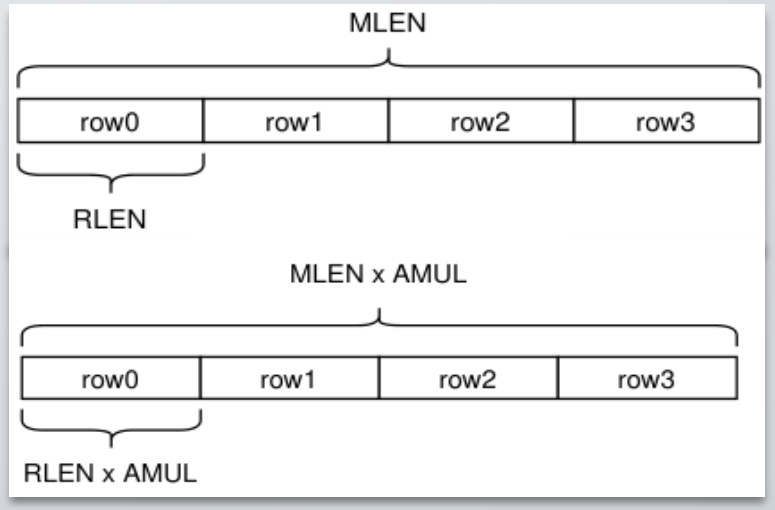

Stream Computing RISC-V Matrix Extension Open Source Project Upgrades to Version 0.5, Supporting Vector+Matrix Implementation

希姆计算 STREAM COMPUTING


RISC-V

RISC-V Matrix Specification

Table of Contents

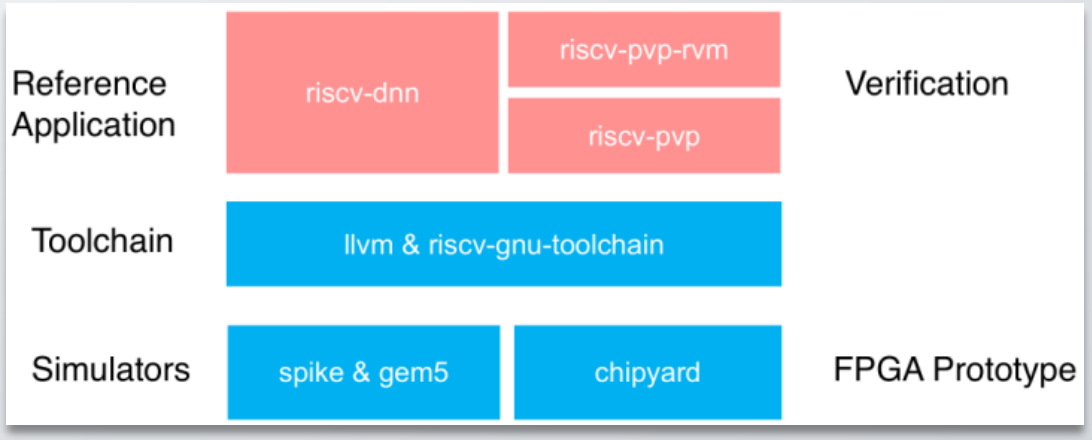


https://riscv.org/blog/2024/11/stream-computing-risc-v-matrix-extension-open-source-project-upgrades-to-version-0-5-supporting-vectormatrix-implementation/
https://github.com/riscv-stc/riscv-matrix-spec/tree/main

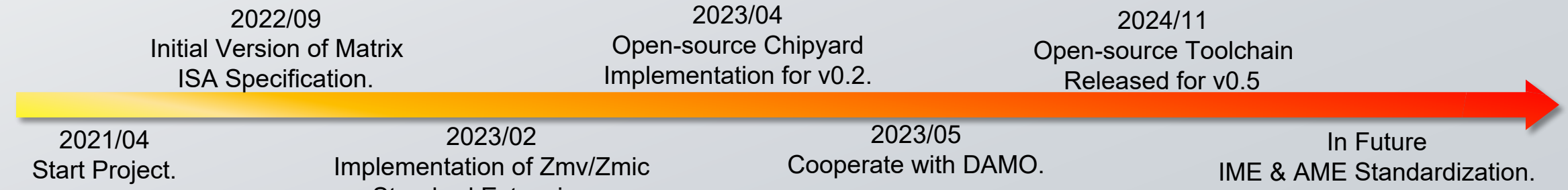
指令集标准化–开源生态项目

- riscv-matrix-spec (https://github.com/riscv-stc/riscv-matrix-spec)
  - The matrix extension proposal.
- llvm-project (https://github.com/riscv-stc/llvm-project)
  - LLVM toolchain to support matrix extension proposal.
- riscv-openocd-matrix (https://github.com/riscv-stc/riscv-openocd-matrix/tree/matrix)
  - GDB debug tool.
- riscv-isa-sim (https://github.com/riscv-stc/riscv-isa-sim)
  - Spike ISS to support matrix extension proposal.
- chipyard (https://github.com/riscv-stc/chipyard)
  - Chipyard project to support matrix extension proposal.
- riscv-pvp-matrix (https://github.com/riscv-stc/riscv-pvp-matrix)
  - RISC-V matrix extension ISA verification using RISC-V PVP.
- riscv-dnn (https://github.com/riscv-stc/riscv-dnn)
  - A small DNN library for RISC-V, using RISC-V vector and matrix extensions.

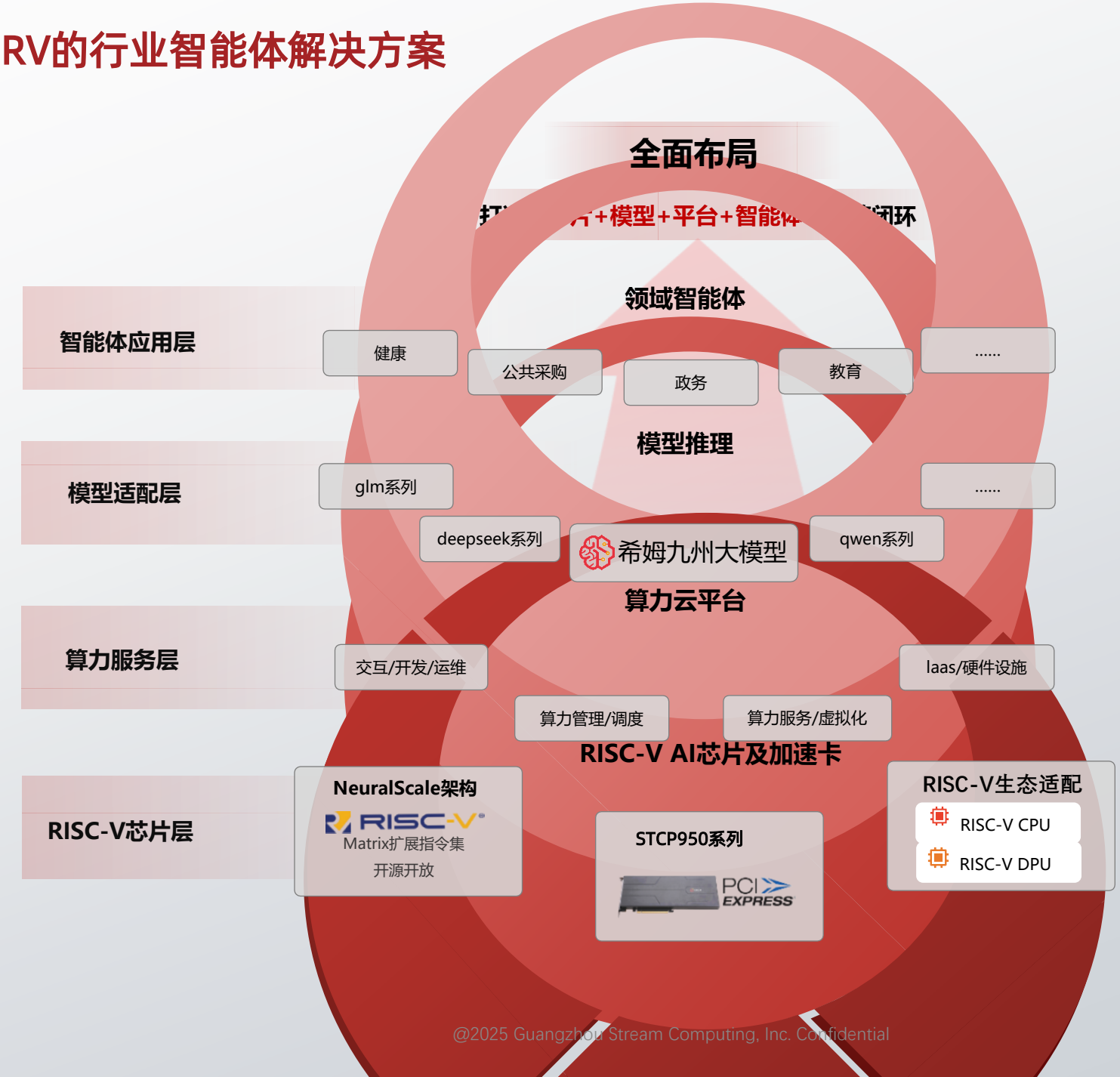
RISC-V Matrix Project Hierarchy.



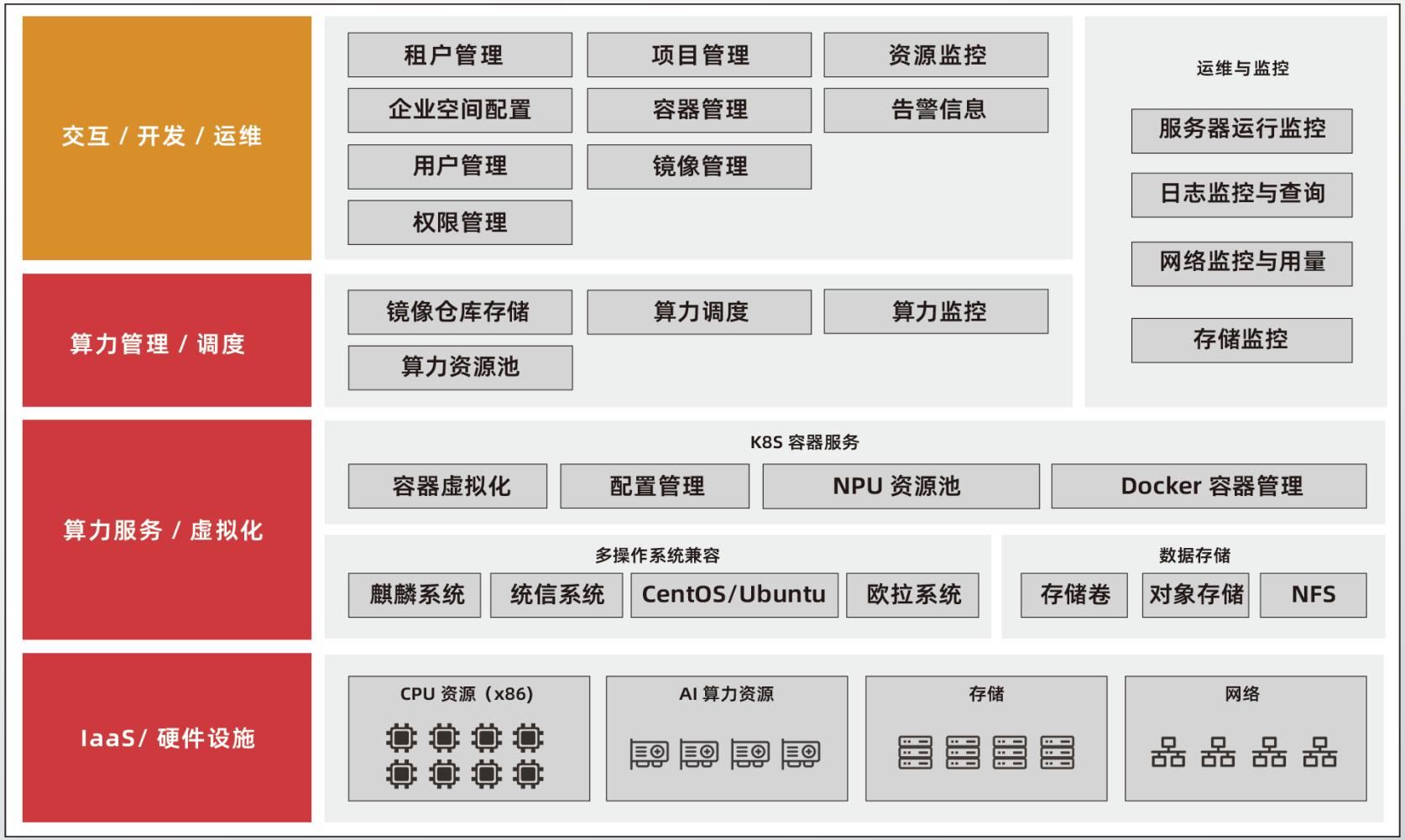
Milestones:



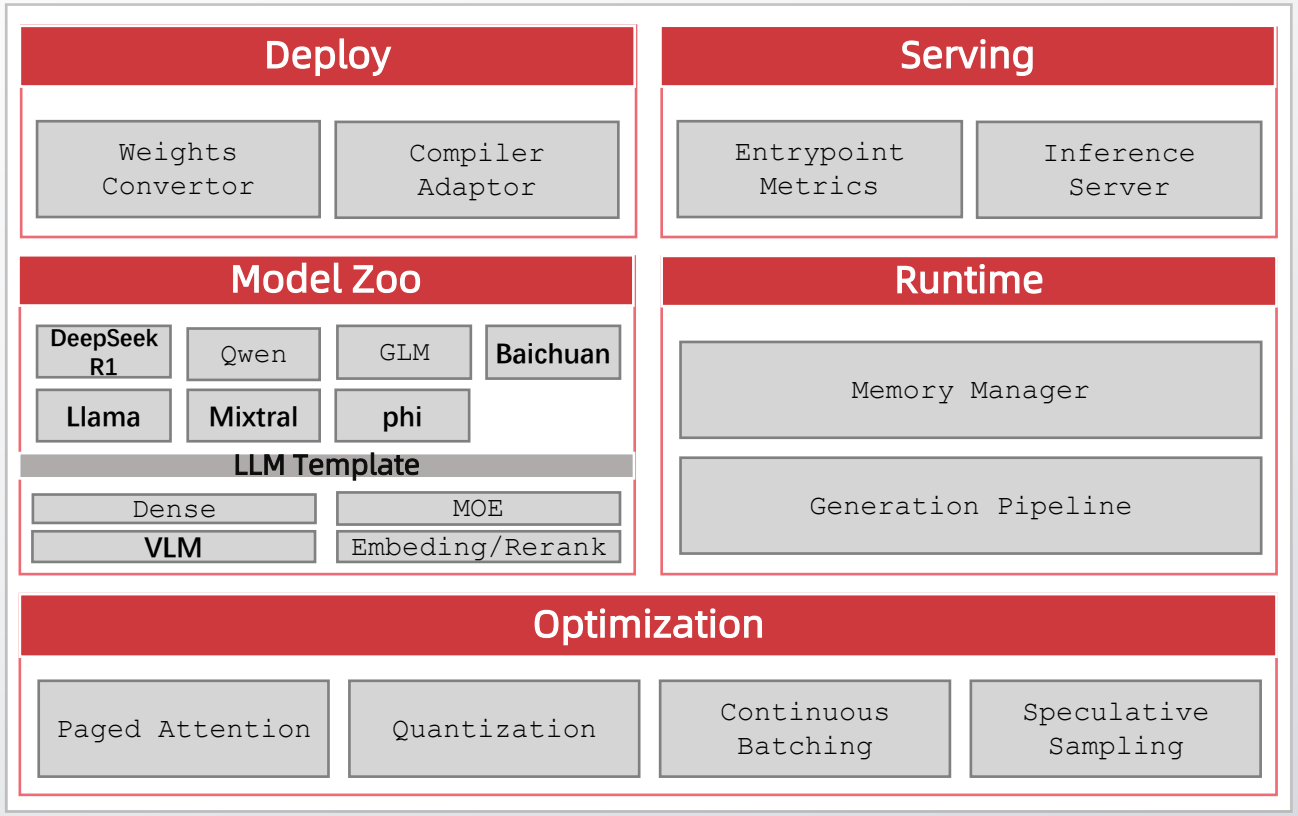
希姆计算-基于RV的行业智能体解决方案



希姆计算-算力云平台



希姆计算-模型推理框架



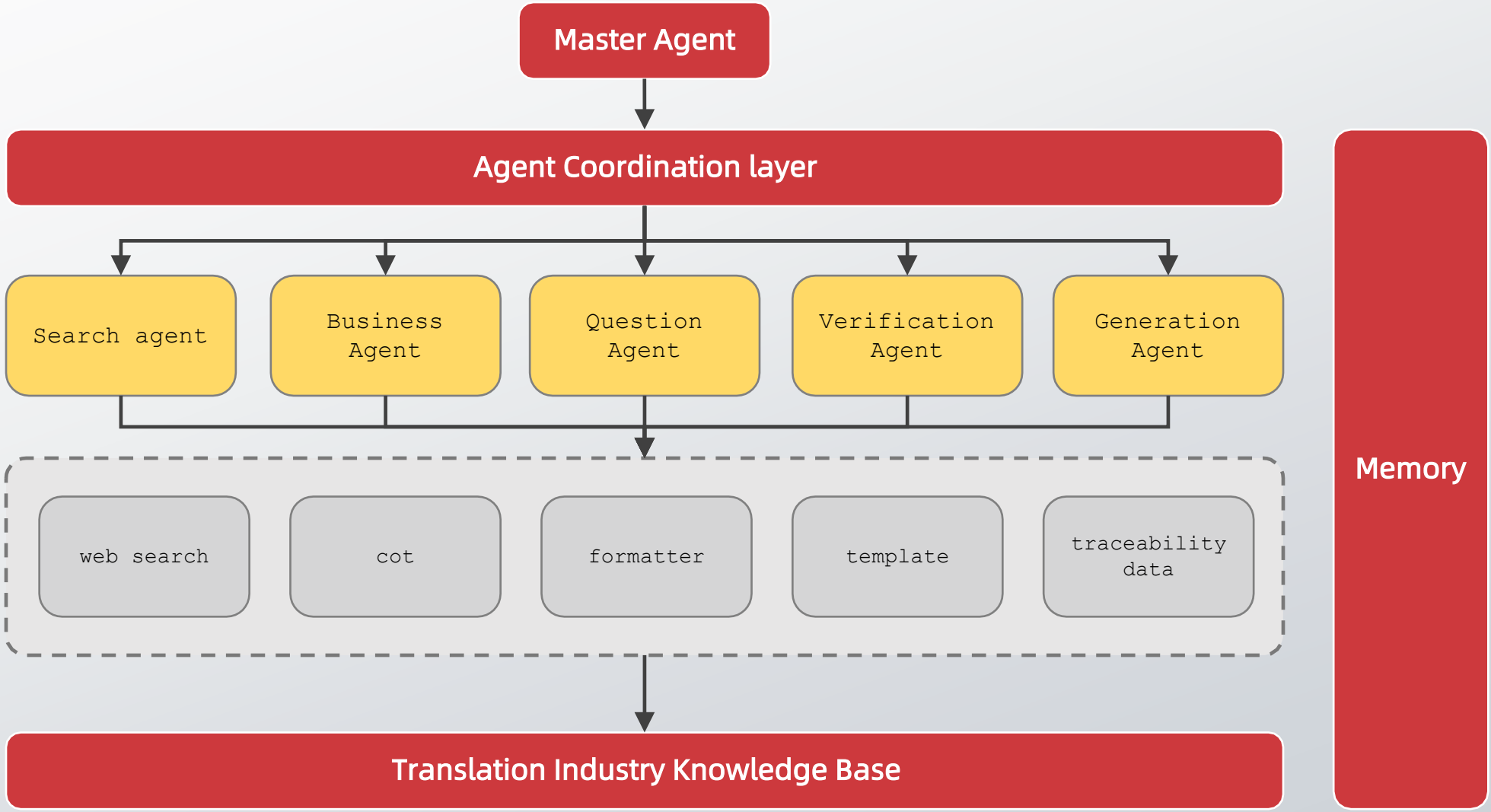
需要解决的问题

- 国内外各种大模型及网络结构层出不穷，如何快速跟进
- 模型参数量大，如何在既有硬件条件下达到最佳性能
- 开源优化方案大多基于 GPU进行，如何取长补短，快速迁移适配
- 灵活性 vs 兼容性

框架特点

- 适配主流的大模型算子
- 通过模板化、参数化管理各类网络结构
- 根据既有的硬件及指令集特点，开发针对性的优化策略
- 提供标准的 Entrypoint，适配主流大模型开发框架，例如 LangChain，llama-index，Dify 等
- 支持云原生技术部署推理集群，确保生产环境的稳定性与扩展性

希姆计算-多智能体Muti-Agent System架构



希姆计算-行业智能体案例

彩翼公共采购智能体平台

广州黄埔、广州开发区政务智能体

希姆计算-MAS多智能体的特点

希姆计算
STREAM COMPUTING

多模型协同

包含多个大模型（CoT）、OCR、Embedding、Reranker等模型

高并发需求

多智能体并行执行，需要高效的任务调度

复杂工作流

一个任务可能涉及多次模型调用，形成链式推理

低延迟要求

实时交互场景（如对话、决策）需要快速响应

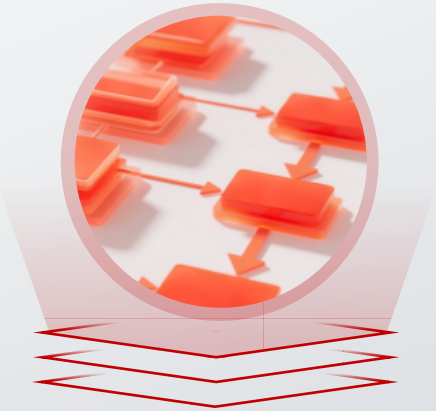
基于RV的MAS Infra的核心技术

高效的资源分配

算力云平台动态调整计算资源,根据实际负载情况自动扩展或缩减资源使用

宏KernelTemplate编译

构建LLM Kernel Template,将各种结构的LLM进行参数化及配置化。
将全模型计算图编译为单一持续内核,消除层间内核启动开销。



硬件指令集融合

直接调用NPU原生ISA接口,绕过传统内核调度栈,实现算子级指令流水线优化

优化深度学习编译器

通过分层扩展的硬件系统和高性能深度学习编译器,通过整图调度、动态切分策略、算子融合自动软流水技术,释放硬件极致性能。

异构通信抽象

构建统一P2P传输层(Register/DMA/PCIe),实现跨卡-跨Cluster零拷贝通信,隐藏延迟

THANKS

感谢大家

在数据中心高端处理器领域占据一席之地