中国电信 CHINA TELECOM | 研究院 CTRI

基于RISC-V架构的高性能AI大模型工作站

杨玉模
中国电信研究院 技术专家
2025年7月

求真求实 向新向上

中国电信 CHINA TELECOM | 研究院 CTRI

1 云计算领域发展趋势与RISC-V现状

2 北海2.0 RISC-V智算云探索

3 RISC-V AI领域发展思考

求真求实 向新向上

数据中心计算架构发展趋势

IaaS云计算依然是数据中心主要形态

- 近年来PaaS和SaaS快速发展，但IaaS仍然是数据中心的核心形态，尤其在大型企业和政府机构中，IaaS服务为业务的数字化转型提供了基础支撑。
- 从裸机到虚拟机到容器再到函数计算，尽管计算的封装粒度和形态在不断演讲，但虚拟机和容器在未来很长一段时间仍然是主要计算形态。

AI驱动智算爆发式增长

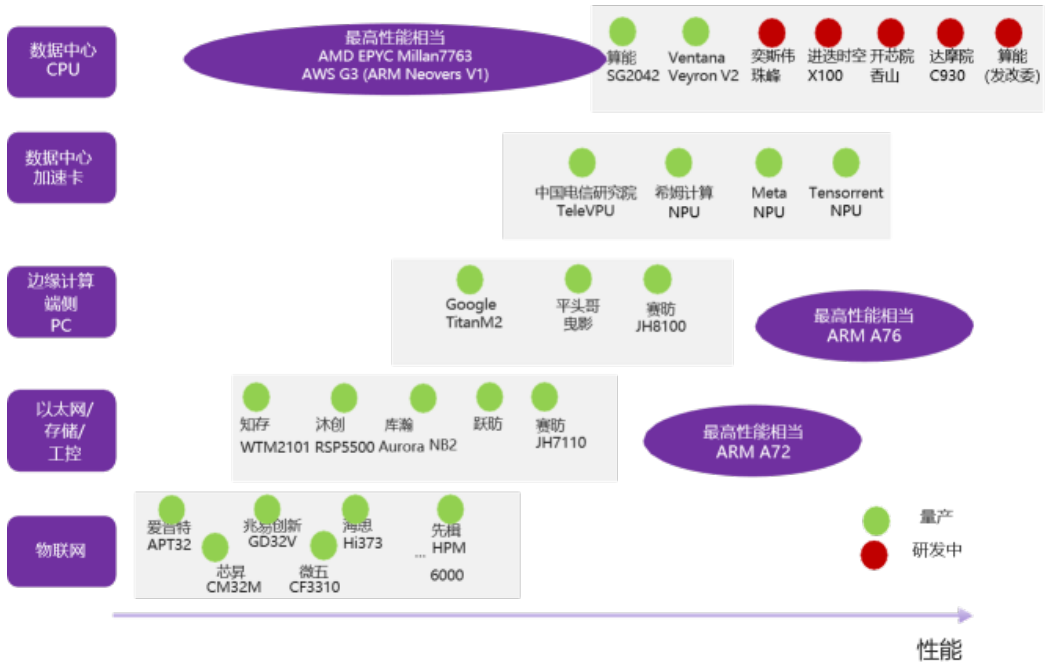
- 人工智能的快速发展推动智算需求爆发式增长，数据中心从传统的通用计算转向以AI为核心的智算。
- AI模型训练所需算力每3-4个月翻一倍，单次训练成果或高达数千万美元，新型计算架构芯片、超节点、集群成为提升算力规模的关键突破点。

领域专用架构DSA成为新趋势

- 领域专用架构（DSA）可以针对特定领域任务进行计算架构优化，与传统通用计算相比，在性能和能效方面具有显著优势。
- 随着AI、视频处理等领域对算力需求的快速增长，DSA成为数据中心发展的新趋势，其可定制性和高效性满足了特定领域的极端需求。

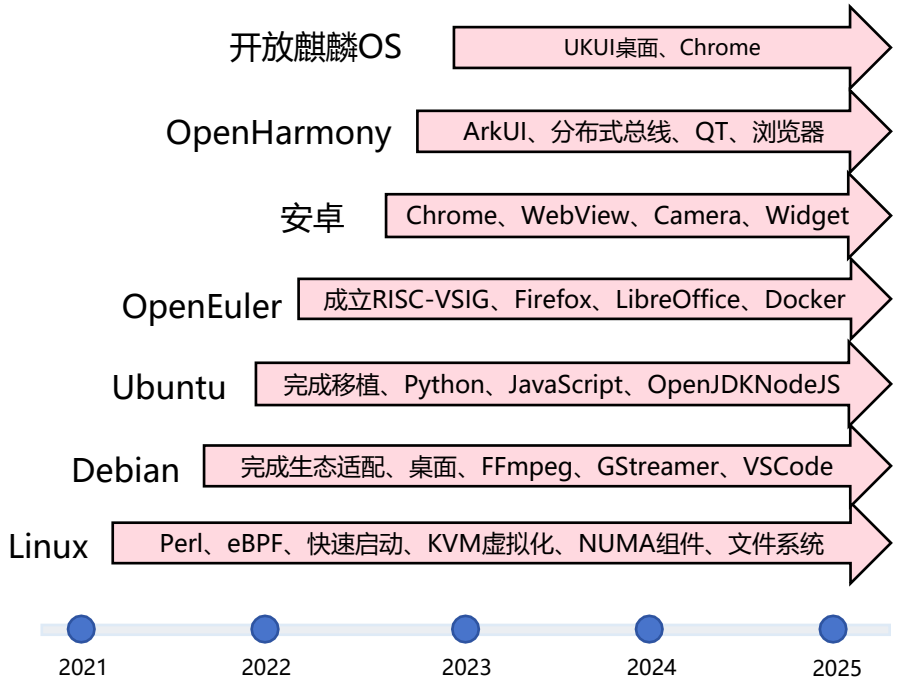
RISC-V在数据中心崭露头角

RISC-V硬件已覆盖主要场景，并向高性能场景演进



RISC-V已经形成较为丰富的产品矩阵，但相距数据中心标杆应用所需性能仍有差距，能用的场景很多，有优势的场景有待发掘。

RISC-V基础生态已具备，应用生态不断完善



RISC-V软件生态快速发展，基础软件基本完成版本适配，但应用软件由于依赖库多样、版本多样迁移依然困难。

中国电信 CHINA TELECOM | 研究院 CTRI

1 云计算领域发展趋势与RISC-V现状

2 北海2.0 RISC-V智算云探索

3 RISC-V AI领域发展思考

求真求实 向新向上

初步探索：“北海”云计算开放试验平台

针对问题：RISC-V云计算软硬件生态不成熟、缺乏规模化的应用验证。
目标：以中国电信丰富应用场景牵引，推动RISC-V架构在云计算行业的成熟与规模应用。

“北海” RISC-V云计算试验平台架构

③RISC-V云计算应用迁移适配与示范应用

Web应用	AI应用	大数据应用	视频应用	……

②RISC-V云计算资源池化纳管试验

多租户	集群管理	应用管理	资源调度	……

容器运行时	轻量安全虚拟机

①RISC-V云计算基础设施研发试验

双路CPU RISC-V服务器	单路CPU RISC-V服务器	TPU/NPU/MLU	VPU	RISC-V先进芯片IP原型

依托中国电信研究院RISC-V大科创装置

双路CPU RISC-V服务器

单路CPU RISC-V服务器

已构建运营商首个超千核RISC-V云计算集群

2024 RISC-V中国峰会发布

人民邮电报>正文

中国电信“北海”RISC-V云计算试验平台发布

发稿时间：2024-08-26 08:32 来源：人民邮电报 作者：翼研 A A 分享

8月21日，在第四届RISC-V中国峰会上，中国电信研究院、天翼云科技公司联合中国科学院软件所、达摩院等合作伙伴发布业界首个RISCV云计算试验平台“北海”，RVEI战略指导委员会主任、中国工程院院士倪光南共同参与发布。RISC-V指令集以其开源开放、精简高效、模块化可定制等技术优势，逐步从物联网、终端向高性能计算领域迈进。

中国电信研究院副院长傅志仁在发布会主题演讲中指出，依托中国电信云网融合大科创装置已完成“北海”平台部署，该平台以RISC-V CPU为核心，同时支持TPU、NPU、VPU

相关成果得到媒体报道

拓展研究：“北海2.0”智算云试验平台

北海2.0 RISC-V智算云增加了云原生虚拟化、AI大模型和AI Agent管理功能，推出RISC-V高性能AI大模型工作站，为虚拟化和AI Agent应用提供基础设施和适配验证方案，推动智算行业的成熟与规模应用。

“北海2.0” RISC-V智算云试验平台架构

③RISC-V智算云应用迁移适配与示范应用

AI Agent应用	虚拟化应用	视频应用	……

②RISC-V智算云资源池化纳管试验

虚拟化管理	AI大模型管理	AI Agent管理	资源调度	……
容器	虚拟机	AI Sandbox		

①RISC-V云计算基础设施研发试验

双路RISC-V服务器	RISC-V AI大模型工作站	高性能RISC-V CPU芯片	GPU/VPU

依托中国电信研究院RISC-V大科创装置

双路CPU RISC-V服务器

RISC-V AI推理卡

RISC-V高性能桌面机

RISC-V高性能CPU

已构建RISC-V AI大模型工作站

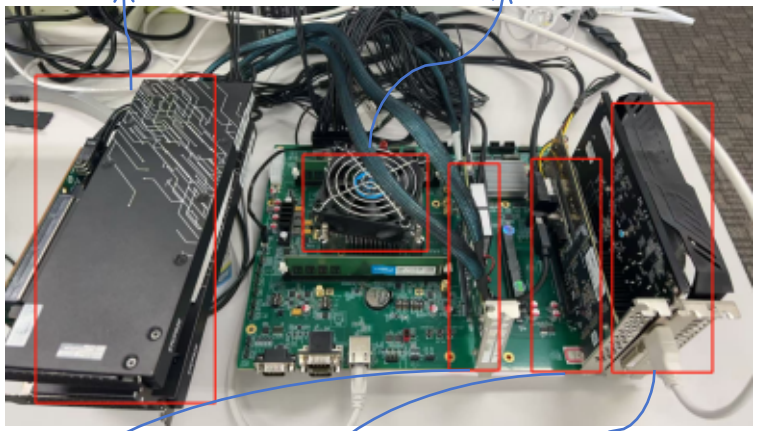
基础设施：RISC-V高性能AI大模型工作站

基于国产RISC-V CPU芯片和国产AI推理卡，构建云原生虚拟化平台及AI Agent应用平台。通过软硬协同定制优化，提供高性能、智能化、可扩展的AI计算环境。

RISC-V高性能AI大模型开发环境

国产AI推理卡或国外AI推理卡

国产RISC-V CPU芯片



PCIe扩展

多功能扩展卡：USB、SSD

普通显卡：HDMI、桌面图形显示

RISC-V高性能AI大模型桌面机

硬件：基于RISC-V的CPU和AI加速卡

- 8核高性能RISC-V CPU芯片，单核SPECInt2006性能达到10.4/GHz。
- PCIe x16可扩展2个x8或4个x8，确保4张AI推理卡稳定运行。
- 搭载国产AI推理卡，单卡16GB显存，可完成标量、向量和矩阵的异构计算，支持FP16和INT8运算。

软件：适配RISC-V云计算和AI软件生态

- 基于GCC和LLVM的私有编译器，优化AI大模型的性能。
- 丰富RISC-V软件生态，支持Ubuntu、OpenEuler、K8s、KubeVirt、AI Agent等。
- 适配1.5B-32B的DeepSeek和Qwen大模型。

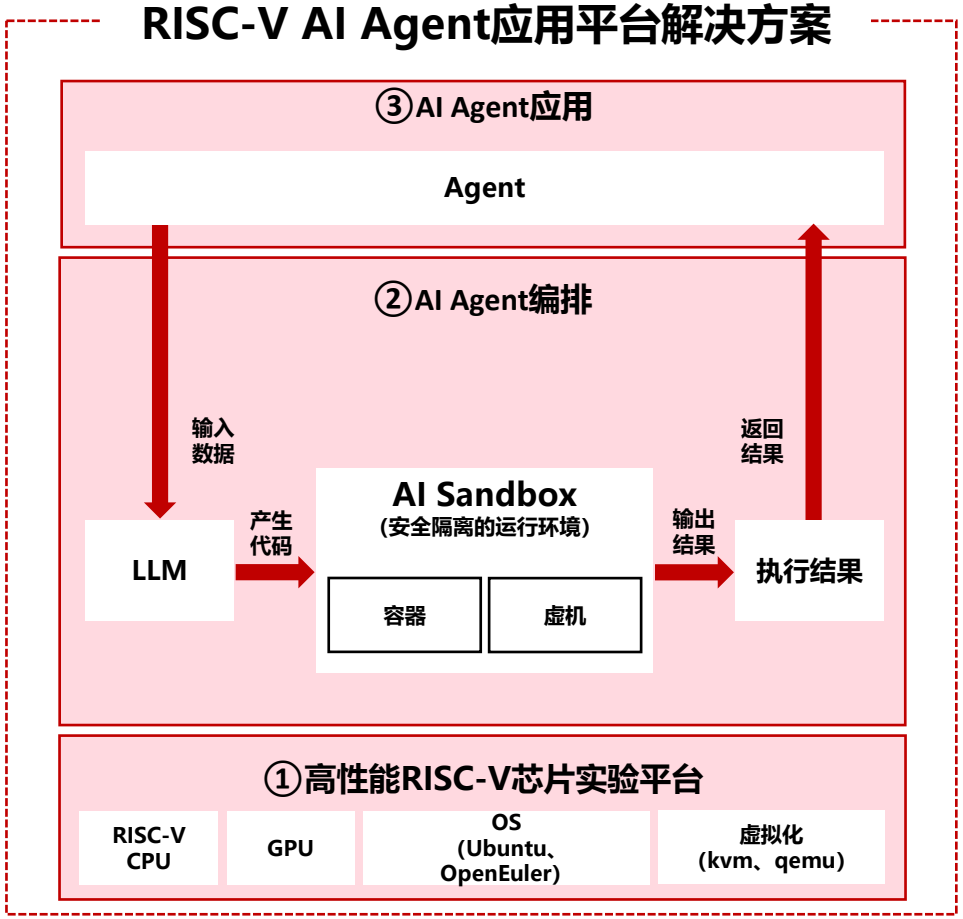
原创突破：填补了RISC-V云原生虚拟化管理平台的空白

RISC-V云原生虚拟化方案：基于高性能RISC-V CPU和硬件虚拟化技术构建业界首个云原生虚拟化管理平台，实现以集群模式统一管理虚拟机和容器，推动了RISC-V虚拟化的商业规模应用。

原创突破：填补了RISC-V AI Agent管理平台的空白

RISC-V AI Agent应用平台方案：针对AI Agent在执行LLM动态生成代码时面临的安全性、隔离性和环境一致性问题，构建安全隔离的沙盒环境，覆盖Agent从开发到部署的全流程，推动AI Agent的适配验证和商业规模应用。

RISC-V AI Agent应用平台解决方案



RISC-V AI Agent应用平台界面

AI Agent平台总览

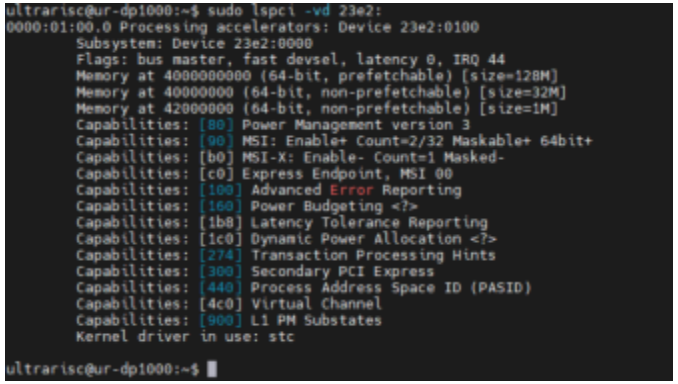
创建AI Agent

资源配置

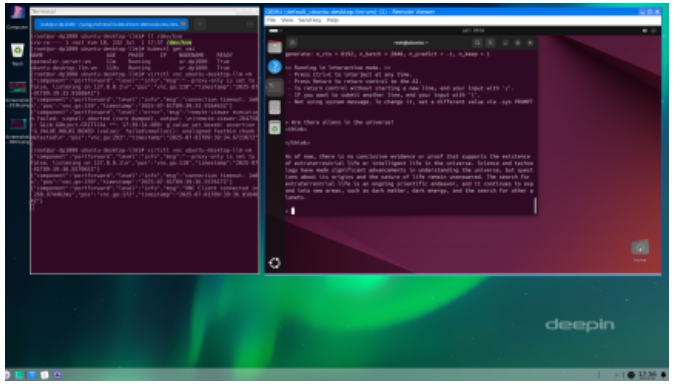
资源监控

适配验证：基于国产RISC-V AI卡的大模型适配验证

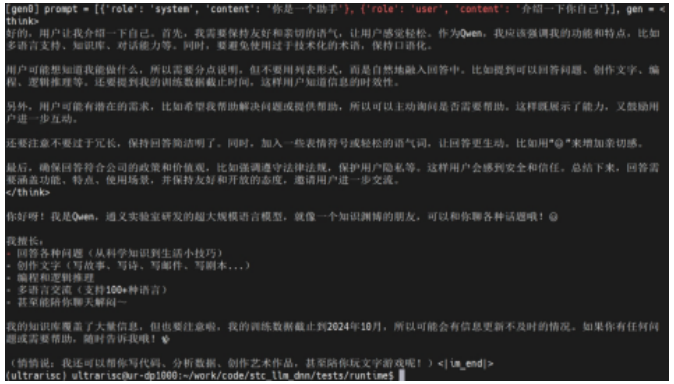
RISC-V AI大模型适配验证方案：通过软硬协同优化，构建高性能、智能化、可扩展的桌面级AI计算环境，搭配RISC-V国产AI加速卡，验证Qwen、DeepSeek等多种大模型。



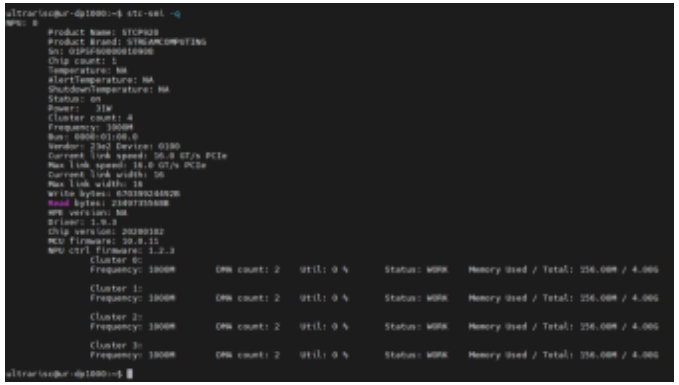
PCI查询RISC-V AI卡信息



RISC-V桌面虚拟机运行DeepSeek 1.5B



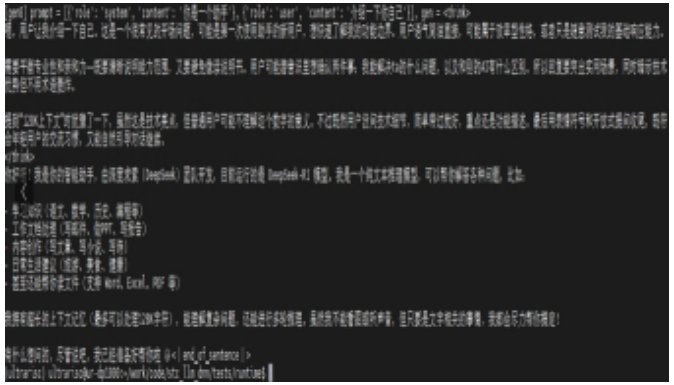
RISC-V运行QWEN3 8B



RISC-V AI板卡信息

Model	Batch Size	TTFT(ms)	提示词处理速度(tps)	产生token速度(tps)
DeepSeek-R1-Distill-Qwen-14B-Q4_K_M	1	386	674.14	50.24
DeepSeek-R1-Distill-Qwen-14B-Q4_K_M	8	3023	656.41	124.32
DeepSeek-R1-Distill-Qwen-32B-Q4_K_M	1	814	301.21	25.87
DeepSeek-R1-Distill-Qwen-32B-Q4_K_M	8	6528	301.12	71.53

RISC-V运行DeepSeek对比实验



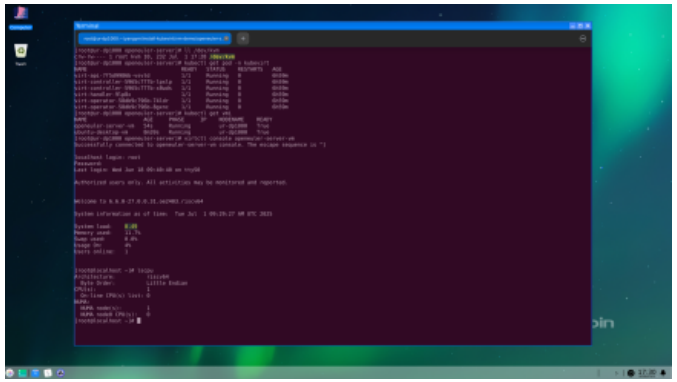
RISC-V运行DeepSeek 7B

适配验证：基于国产RISC-V CPU的云原生虚拟化适配验证

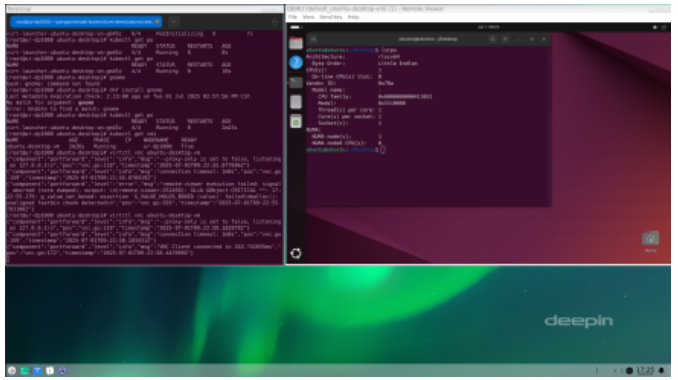
RISC-V云原生虚拟化适配验证方案：基于RISC-V高性能CPU芯片和硬件虚拟化，创建云原生虚拟化管理系统，验证了硬件虚拟化的性能损失约6%，性能是软件虚拟化的6倍，能够正常启动KubeVirt服务器版和桌面版虚拟机。

```
[root@ur-dp1000 vm-demo]# kubectl get po -n kubevirt
NAME                               READY   STATUS    RESTARTS   AGE
virt-api-7f5d9986b-prjtd           1/1     Running   0          63m
virt-controller-5965c777b-7fvfj    1/1     Running   0          63m
virt-controller-5965c777b-rjjzr    1/1     Running   0          63m
virt-handler-zrblz                 1/1     Running   0          63m
virt-operator-58db9c796b-hbmxg     1/1     Running   0          63m
virt-operator-58db9c796b-kpqdn     1/1     Running   0          63m
[root@ur-dp1000 vm-demo]# kubectl get vmi
NAME           AGE   PHASE     IP    NODENAME    READY
openeuler-vm   62m   Running         ur-dp1000   True
[root@ur-dp1000 vm-demo]# kubectl get po
NAME                               READY   STATUS    RESTARTS   AGE
virt-launcher-openeuler-vm-l9sqk   4/4     Running   0          62m
```

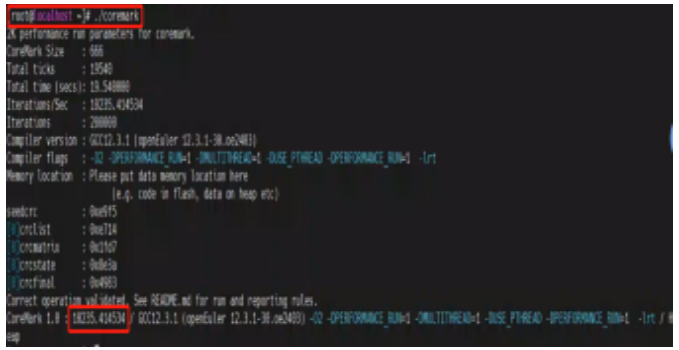
RISC-V KubeVirt运行状态



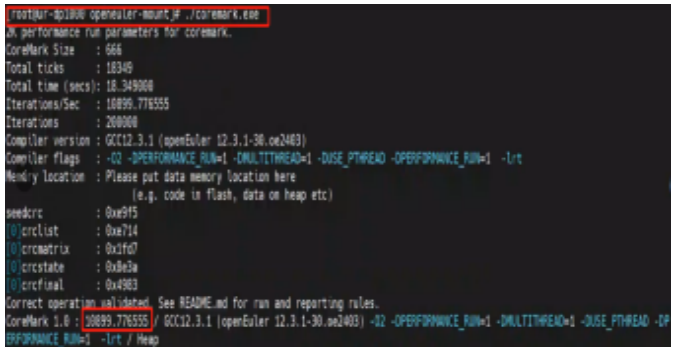
通过console进入KubeVirt OpenEuler虚拟机



通过vnc进入RISC-V KubeVirt桌面版虚拟机



RISC-V硬件虚拟化单核CPU跑分实验

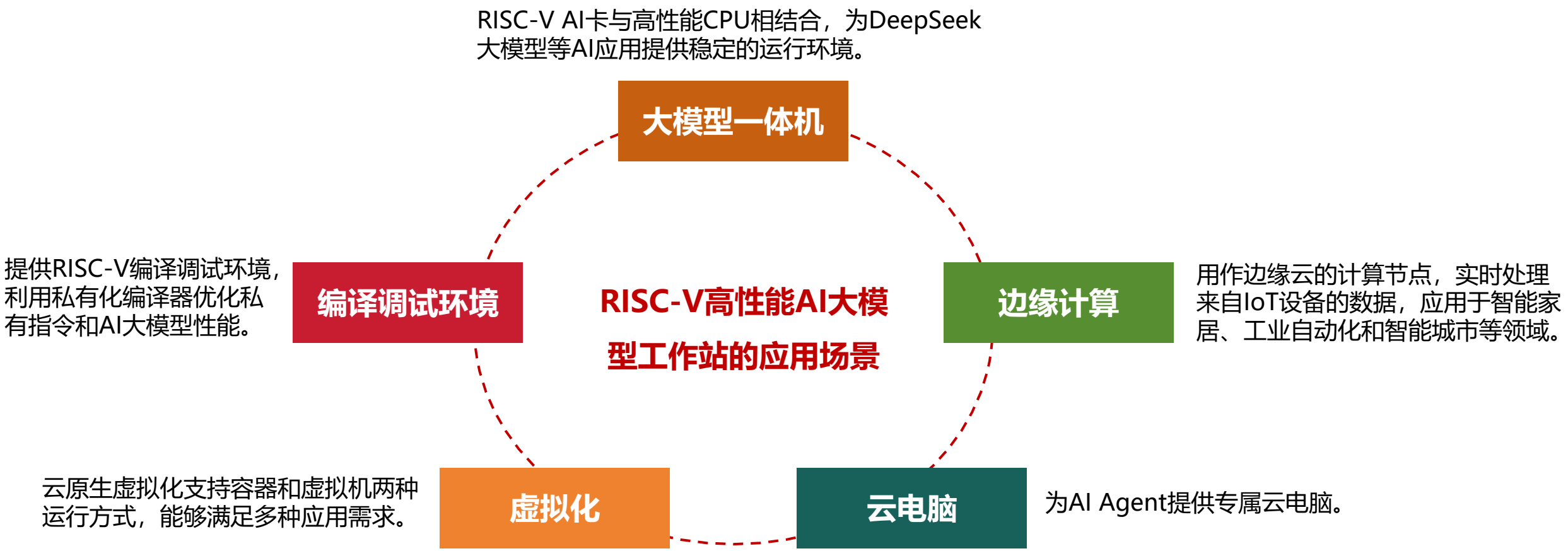


RISC-V宿主机单核CPU跑分

RISC-V软件虚拟化单核CPU跑分实验

应用场景

RISC-V高性能AI大模型工作站具有高性能、智能化、可扩展的特征，使其相比其他PC架构在技术、成本、市场以及发展潜力等方面表现出独特的优势，这为RISC-V 高性能AI大模型工作中多样化场景的应用提供了基础保障。



中国电信 CHINA TELECOM | 研究院 CTRI

1 云计算领域发展趋势与RISC-V现状

2 北海2.0 RISC-V智算云探索

3 RISC-V AI领域发展思考

求真求实 向新向上

运营商视角的RISC-V AI领域发展思考

找好定位

- 做AI基础设施平台和集成商
  - 以IaaS为基础做AI算力平台
  - 以PaaS为基础做AI应用平台

大模型如何为客户创造价值

- 立足现有业务，拓展新型业务
  - 大模型重新定义传统业务
  - 大模型探索新型业务场景

如何将理论研究转化为商业应用

- 与上下游厂商携手，推动商业落地
  - 健全RISC-V AI领域的开源生态
  - 从开源向产品化转变

中国电信 CHINA TELECOM | 研究院 CTRI

感谢聆听！

求真求实 向新向上