

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

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

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理形态，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进，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。

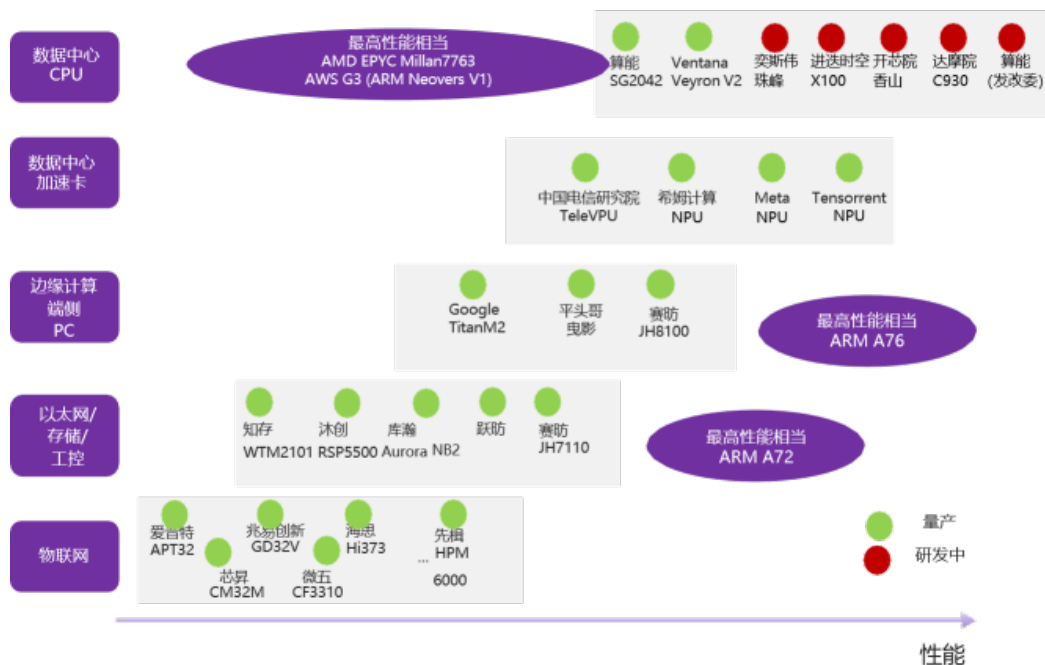
AI驱动智算爆发式增长

- 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

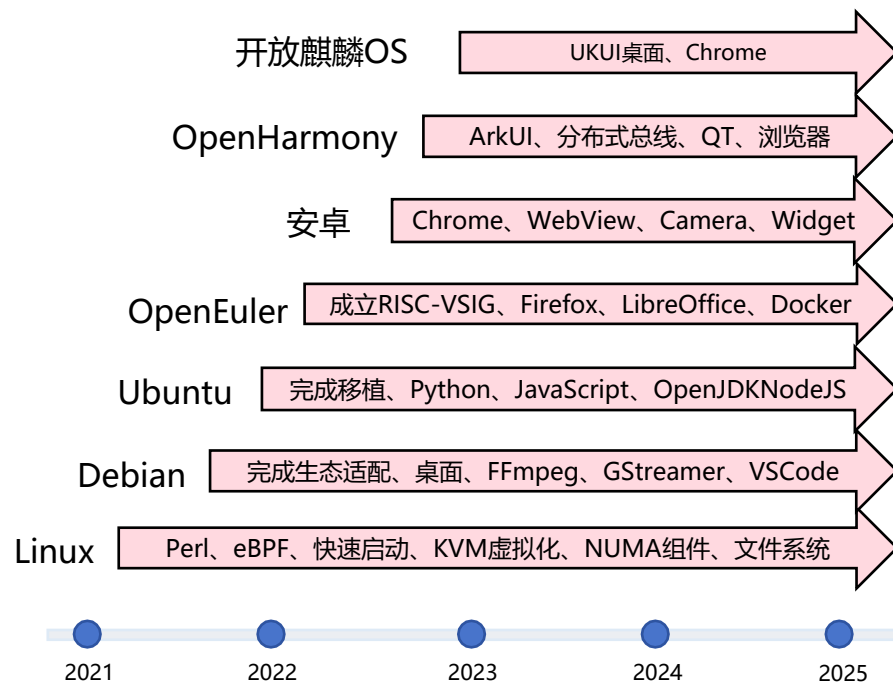
- 领域专用架构（DSA）可以针对特定领域任务进行计算架构优化，与传统通用计算相比，在性能和能效方面具有显著优势。
- 随着AI、视频处理等领域对算力需求的快速增长，DSA成为数据中心发展的新趋势，其可定制性和高效性满足了特定领域的极端需求。

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。

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

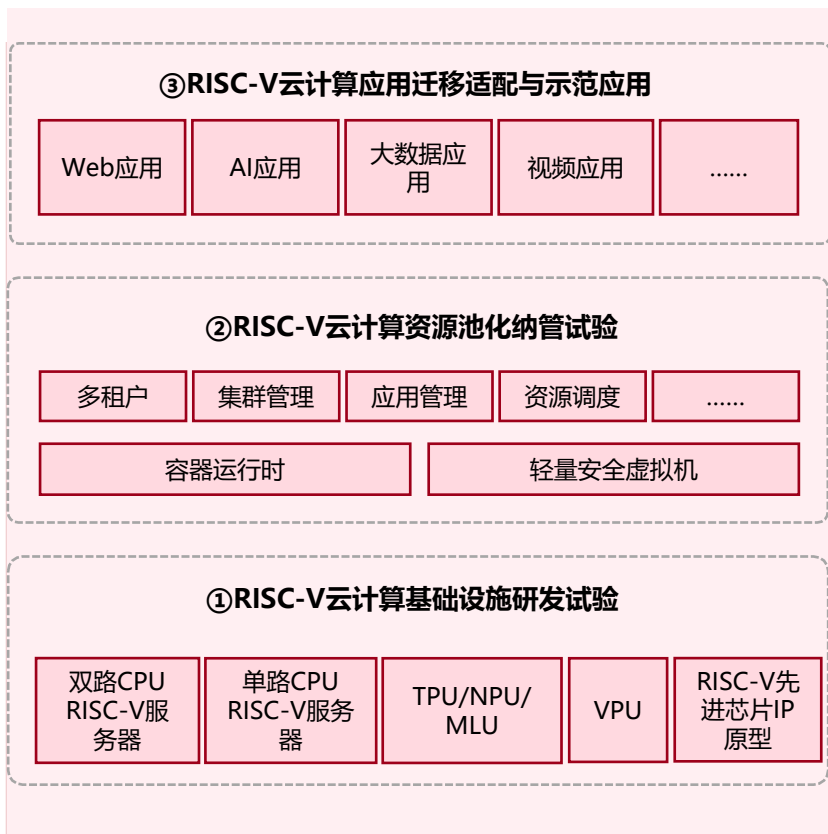
3

RISC-V AI领域发展思考

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



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



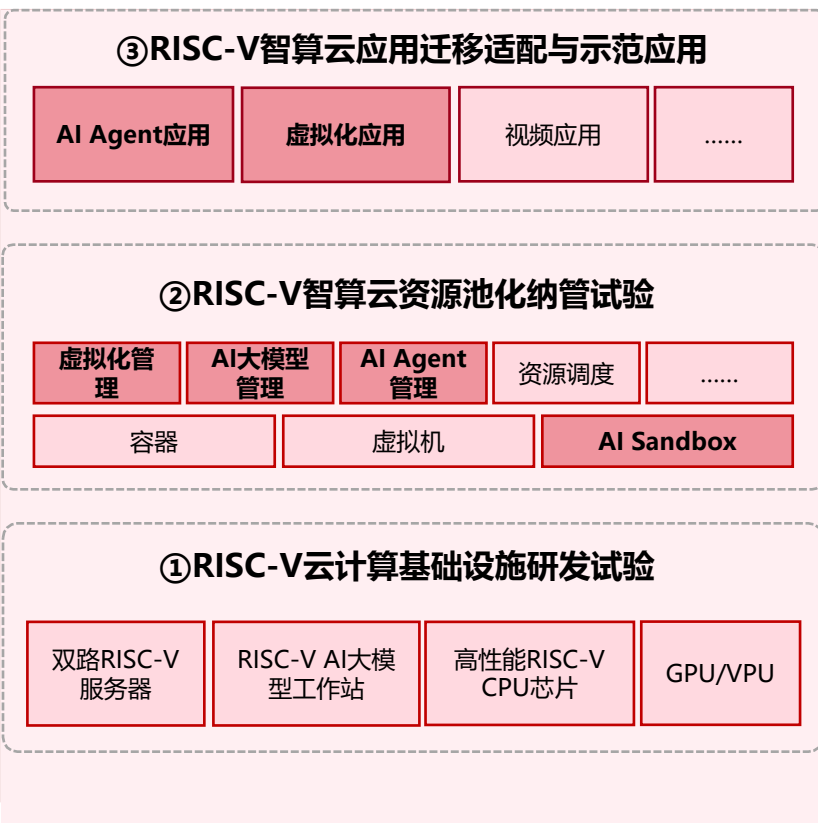
2024 RISC-V中国峰会发布



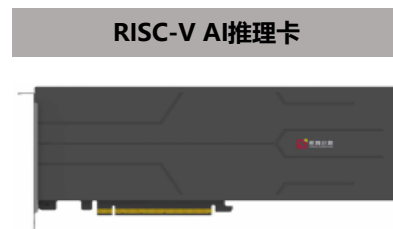
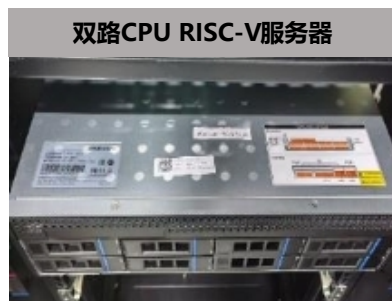
相关成果得到媒体报道

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



依托中国电信研究院RISC-V大科创装置



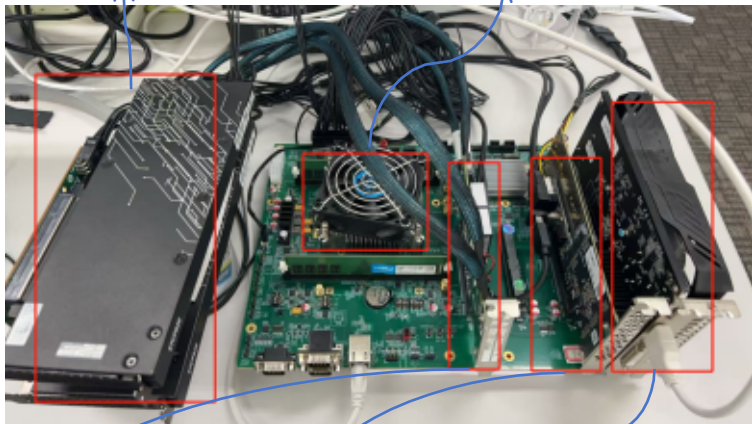
已构建RISC-V AI大模型工作站

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。

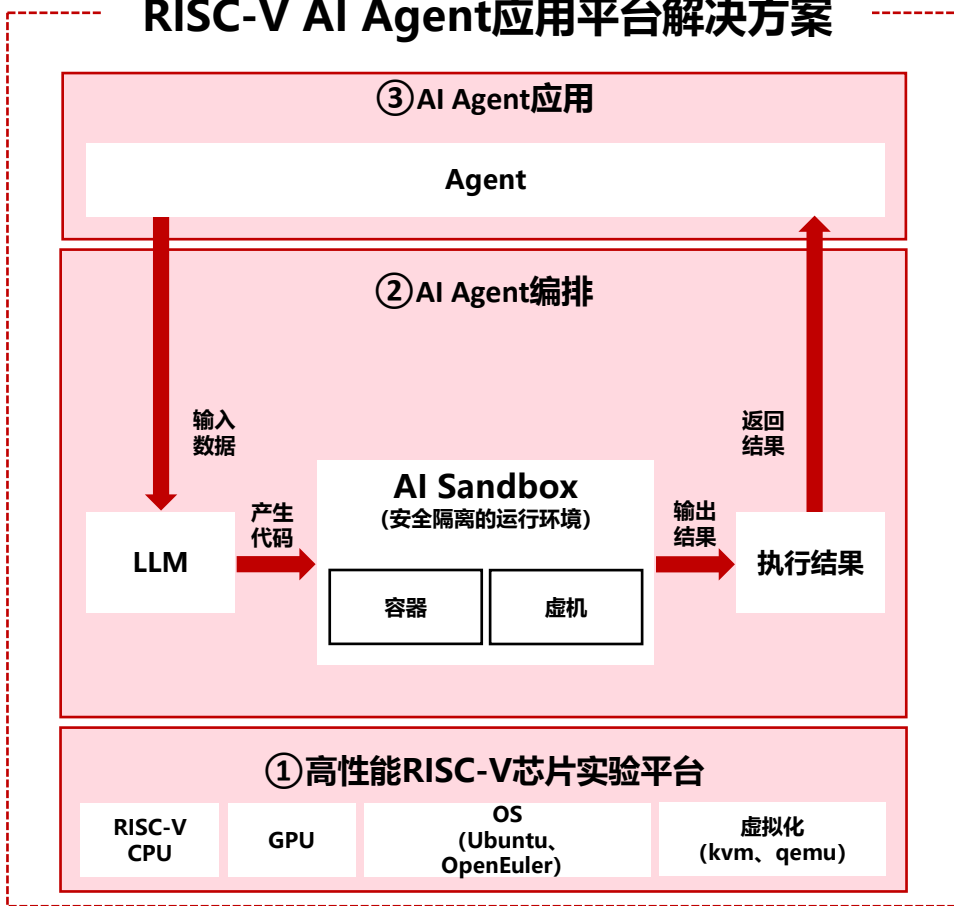
RISC-V云原生虚拟化适配验证方案



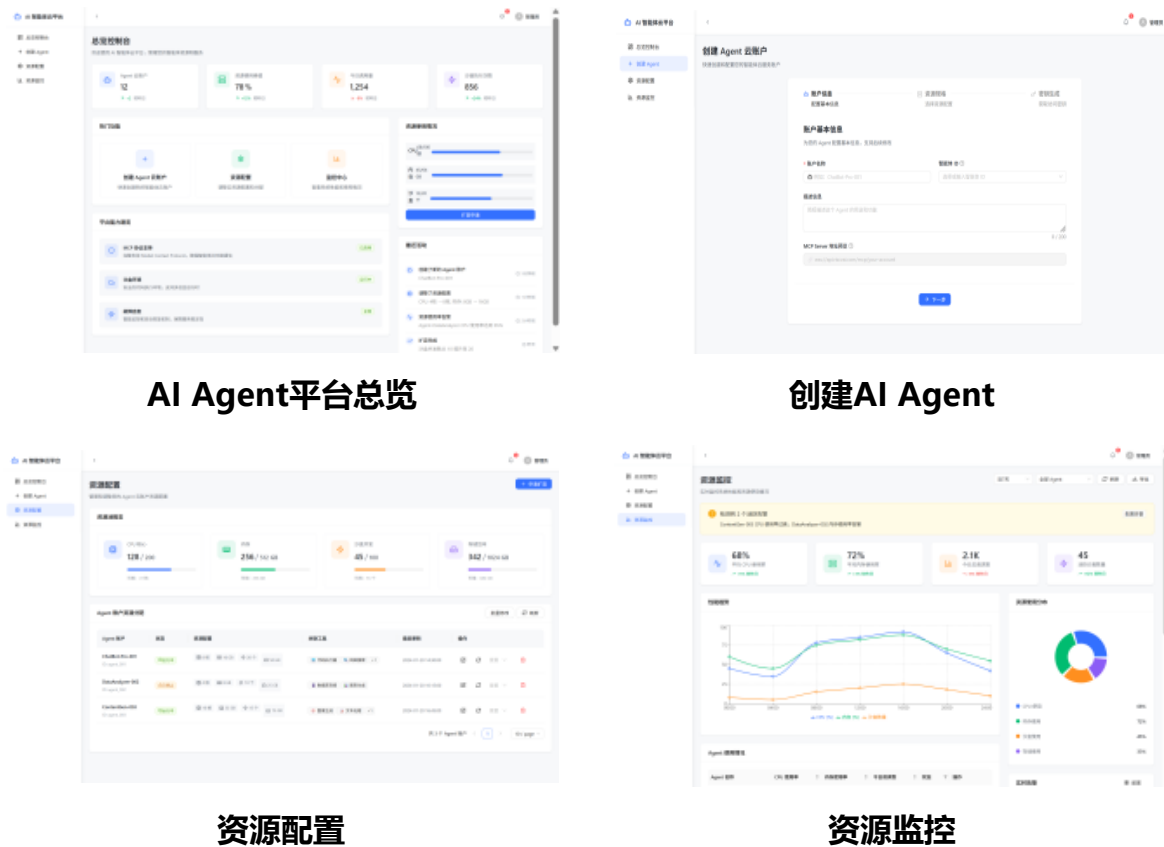
原创突破：填补了RISC-V AI Agent管理平台的空白

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

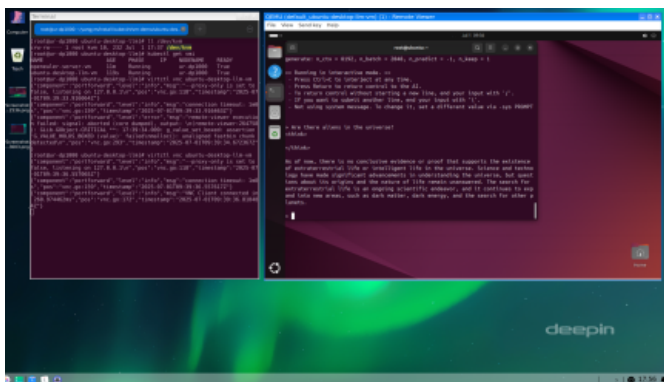


适配验证：基于国产RISC-V AI卡的大模型适配验证

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。

```
ultrarisc@ur-dp1000:~$ sudo lspci -vd 23e2:
0000:01:00.0 Processing accelerators: Device 23e2:0100
Subsystem: Device 23e2:0000
Flags: bus master, fast devsel, latency 0, IRQ 44
Memory at 4000000000 (64-bit, prefetchable) [size=128M]
Memory at 400000000 (64-bit, non-prefetchable) [size=32M]
Memory at 420000000 (64-bit, non-prefetchable) [size=1M]
Capabilities: [80] Power Management version 3
Capabilities: [90] MSI: Enable+ Count=2/32 Maskable+ 64bit+
Capabilities: [b0] MSI-X: Enable- Count=1 Masked-
Capabilities: [c0] Express Endpoint, MSI 00
Capabilities: [100] Advanced Error Reporting
Capabilities: [160] Power Budgeting <?>
Capabilities: [1b8] Latency Tolerance Reporting
Capabilities: [1c0] Dynamic Power Allocation <?>
Capabilities: [274] Transaction Processing Hints
Capabilities: [300] Secondary PCI Express
Capabilities: [440] Process Address Space ID (PASID)
Capabilities: [4c0] Virtual Channel
Capabilities: [900] L1 PM Substates
Kernel driver in use: stc
```

PCI查询RISC-V AI卡信息



RISC-V桌面虚拟机运行DeepSeek 1.5B

```
[qwen] prompt = [{"role": "system", "content": "你是一个助手"}, {"role": "user", "content": "介绍一下你自己"}], gen = <think>
好的，用户让我介绍一下自己。首先，我需要保持友好和亲切的语气，让用户感觉轻松。作为Qwen，我应该强调我的功能和特点，比如多语言支持、知识库、对话能力等。同时，要避免使用过于技术化的术语，保持口语化。
用户可能想知道我能做什么，所以需要重点说明，但不要罗列列表形式，而是自然地融入回答中。比如提到可以回答问题、创作文字、编程、逻辑推理等。还要提到我的训练数据截止时间，这样用户知道信息的时效性。
另外，用户可能有潜在的需求，比如希望我帮助解决问题或提供帮助，所以可以主动询问是否需要帮助，这样既展示了能力，又鼓励用户进一步互动。
还要注意不要过于冗长，保持回答简洁明了。同时，加入一些表情符号或轻松的语气词，让回答更生动。比如用“~”来增加亲切感。
最后，确保回答符合公司的政策和价值观，比如强调遵守法律法规，保护用户隐私等。这样用户会感到安全和信任。总结下来，回答需要满足功能、特点、使用场景，并保持友好和开放的态度，邀请用户进一步交流。
</think>
你好呀！我是Qwen，通义实验室研发的超大规模语言模型，就像一个知识渊博的朋友，可以和你聊各种话题哦！~
我擅长：
- 回答各种问题（从科学知识到生活小技巧）
- 创作文字（写故事、写诗、写邮件、写剧本...）
- 编程和逻辑推理
- 多语言交流（支持100+种语言）
- 甚至能陪你聊天解闷~
我的知识库涵盖了大量信息，但也要注意啦，我的训练数据截止到2024年10月，所以可能会有信息更新不及时的情况。如果你有任何问题或需要帮助，随时告诉我哦！~
（悄悄说：我还可以帮你写代码、分析数据、创作艺术作品，甚至陪你玩文字游戏哦！）</im_end>
ultrarisc@ur-dp1000:~/work/code/stc_llm_dnn/tests/runtime$
```

RISC-V运行QWEN3 8B

```
ultrarisc@ur-dp1000:~$ lspci -vv -s 01:00.0
01:00.0 Product Name: STC923
Product Brand: STC923AC00000000
ID: 0100:000000000000
Chip count: 1
Temperature: NA
AlertTemperature: NA
ShutdownTemperature: NA
Status: on
Power: 31w
Cluster count: 4
Frequency: 1000M
Board: 0000:01:00.0
Version: 23e2 Device: 0100
Current Link speed: 56.0 GT/s PCIe
Max Link speed: 16.0 GT/s PCIe
Current Link width: 16
Max Link width: 16
Writ System: 803030000000
Read bytes: 2389300000
MPS version: NA
B1000: 1.0.0
Chip version: 20280102
MCU Firmware: 20.0.11
MPS ctrl Firmware: 1.1.3
Cluster 0:
Frequency: 1000M DM count: 2 Util: 0.4 Status: WOK Memory used / Total: 156.00M / 4.00G
Cluster 1:
Frequency: 1000M DM count: 2 Util: 0.4 Status: WOK Memory used / Total: 156.00M / 4.00G
Cluster 2:
Frequency: 1000M DM count: 2 Util: 0.4 Status: WOK Memory used / Total: 156.00M / 4.00G
Cluster 3:
Frequency: 1000M DM count: 2 Util: 0.4 Status: WOK Memory used / Total: 156.00M / 4.00G
```

RISC-V AI板卡信息

Model	Batch Size	TTFT(m s)	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

```
[qwen] prompt = [{"role": "system", "content": "你是一个助手"}, {"role": "user", "content": "介绍一下你自己"}], gen = <think>
好的，用户让我介绍一下自己。这是一个自我介绍的问题，我需要保持友好和亲切的语气，让用户感觉轻松。同时，也要避免使用过于技术化的术语，保持口语化。
用户可能想知道我能做什么，所以需要重点说明，但不要罗列列表形式，而是自然地融入回答中。比如提到可以回答问题、创作文字、编程、逻辑推理等。还要提到我的训练数据截止时间，这样用户知道信息的时效性。
另外，用户可能有潜在的需求，比如希望我帮助解决问题或提供帮助，所以可以主动询问是否需要帮助，这样既展示了能力，又鼓励用户进一步互动。
还要注意不要过于冗长，保持回答简洁明了。同时，加入一些表情符号或轻松的语气词，让回答更生动。比如用“~”来增加亲切感。
最后，确保回答符合公司的政策和价值观，比如强调遵守法律法规，保护用户隐私等。这样用户会感到安全和信任。总结下来，回答需要满足功能、特点、使用场景，并保持友好和开放的态度，邀请用户进一步交流。
</think>
你好呀！我是你的智能助手，由通义实验室（DeepSeek）研发开发，目前运行的是 DeepSeek-R1 模型，我是一个纯文本模型，可以帮助你解决各种问题，比如：
- 学习知识（语文、数学、历史、编程等）
- 工作文档处理（写邮件、写PPT、写剧本）
- 内容创作（写小说、写诗、写故事）
- 日常问答（天气、新闻、美食、健康）
- 甚至还能帮你写代码（Python、Java、C++、JS等）
如果你有学习上或工作上的问题，随时告诉我，我会尽力帮你解答。虽然我不能替你完成作业，但我可以帮你理清思路，提供思路和方法。
有什么问题的话，尽管告诉我吧！</im_end>
ultrarisc@ur-dp1000:~/work/code/stc_llm_dnn/tests/runtime$
```

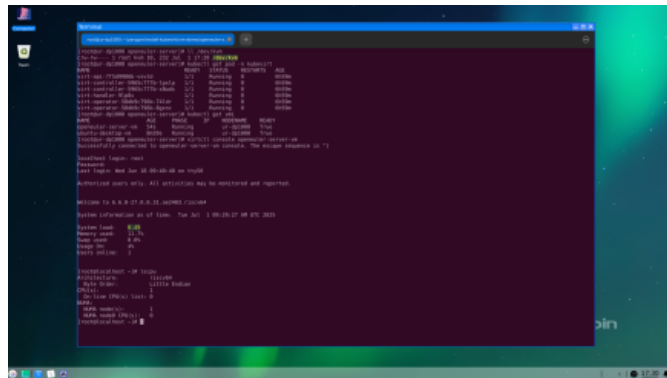
RISC-V运行DeepSeek 7B

适配验证：基于国产RISC-V CPU的云原生虚拟化适配验证

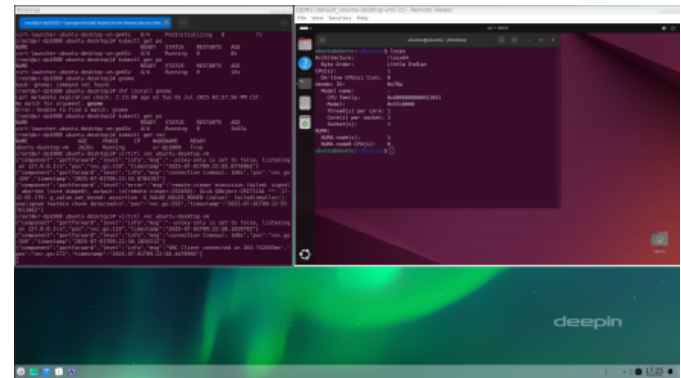
RISC-V云原生虚拟化适配验证方案：基于RISC-V高性能CPU芯片和硬件虚拟化，创建云原生虚拟化管理系统，验证了硬件虚拟化的性能损失约6%，性能是软件虚拟化的6倍，能够正常启动KubeVirt服务器版和桌面版虚拟机。

```
[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-kpoadn       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

```
root@ur-dp1000 ~# ./coremark
A performance run parameters for coremark.
CoreMark Size : 666
Total ticks : 18349
Total time (secs): 18.349000
Iterations/Sec : 18335.41634
Iterations : 280000
Compiler version : GCC12.3.1 (openEuler 12.3.1-30.no403)
Compiler flags : -O2 -DPERFORMANCE_RUN=1 -DMULTITHREAD=1 -DUSE_PTHREAD -DPERFORMANCE_RUN=1 -lrt
Memory location : Please put data memory location here
(e.g. code in flash, data on heap etc)
seedcrc : 0xe9f5
| corList : 0xe714
| cronMatrix : 0xc1d7
| crcstate : 0x8e3a
| crcfinal : 0x4983
Correct operation validated. See README.md for run and reporting rules.
CoreMark 1.0 : 18335.41634 / GCC12.3.1 (openEuler 12.3.1-30.no403) -O2 -DPERFORMANCE_RUN=1 -DMULTITHREAD=1 -DUSE_PTHREAD -DPERFORMANCE_RUN=1 -lrt / Heap
```

RISC-V硬件虚拟化单核CPU跑分实验

```
root@ur-dp1000 ~# ./coremark
A performance run parameters for coremark.
CoreMark Size : 666
Total ticks : 18349
Total time (secs): 18.349000
Iterations/Sec : 18899.776555
Iterations : 280000
Compiler version : GCC12.3.1 (openEuler 12.3.1-30.no403)
Compiler flags : -O2 -DPERFORMANCE_RUN=1 -DMULTITHREAD=1 -DUSE_PTHREAD -DPERFORMANCE_RUN=1 -lrt
Memory location : Please put data memory location here
(e.g. code in flash, data on heap etc)
seedcrc : 0xe9f5
| corList : 0xe714
| cronMatrix : 0xc1d7
| crcstate : 0x8e3a
| crcfinal : 0x4983
Correct operation validated. See README.md for run and reporting rules.
CoreMark 1.0 : 18899.776555 / GCC12.3.1 (openEuler 12.3.1-30.no403) -O2 -DPERFORMANCE_RUN=1 -DMULTITHREAD=1 -DUSE_PTHREAD -DPERFORMANCE_RUN=1 -lrt / Heap
```

RISC-V宿主机单核CPU跑分

```
root@ur-dp1000 ~# ./coremark
A performance run parameters for coremark.
CoreMark Size : 666
Total ticks : 38333
Total time (secs): 38.333000
Iterations/Sec : 3836.763308
Iterations : 30000
Compiler version : GCC12.3.1 (openEuler 12.3.1-30.no403)
Compiler flags : -O2 -DPERFORMANCE_RUN=1 -DMULTITHREAD=1 -DUSE_PTHREAD -DPERFORMANCE_RUN=1 -lrt
Memory location : Please put data memory location here
(e.g. code in flash, data on heap etc)
seedcrc : 0xe9f5
| corList : 0xe754
| cronMatrix : 0x1f47
| crcstate : 0xb6c4
| crcfinal : 0x5275
Correct operation validated. See README.md for run and reporting rules.
CoreMark 1.0 : 3836.763308 / GCC12.3.1 (openEuler 12.3.1-30.no403) -O2 -DPERFORMANCE_RUN=1 -DMULTITHREAD=1 -DUSE_PTHREAD -DPERFORMANCE_RUN=1 -lrt / Heap
```

RISC-V软件虚拟化单核CPU跑分实验

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。

RISC-V AI卡与高性能CPU相结合，为DeepSeek大模型等AI应用提供稳定的运行环境。

大模型一体机

提供RISC-V编译调试环境，利用私有化编译器优化私有指令和AI大模型性能。

编译调试环境

RISC-V高性能AI大模型工作站的应用场景

用作边缘云的计算节点，实时处理来自IoT设备的数据，应用于智能家居、工业自动化和智能城市等领域。

边缘计算

云原生虚拟化支持容器和虚拟机两种运行方式，能够满足多种应用需求。

虚拟化

为AI Agent提供专属云电脑。

云电脑

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

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

感谢聆听!