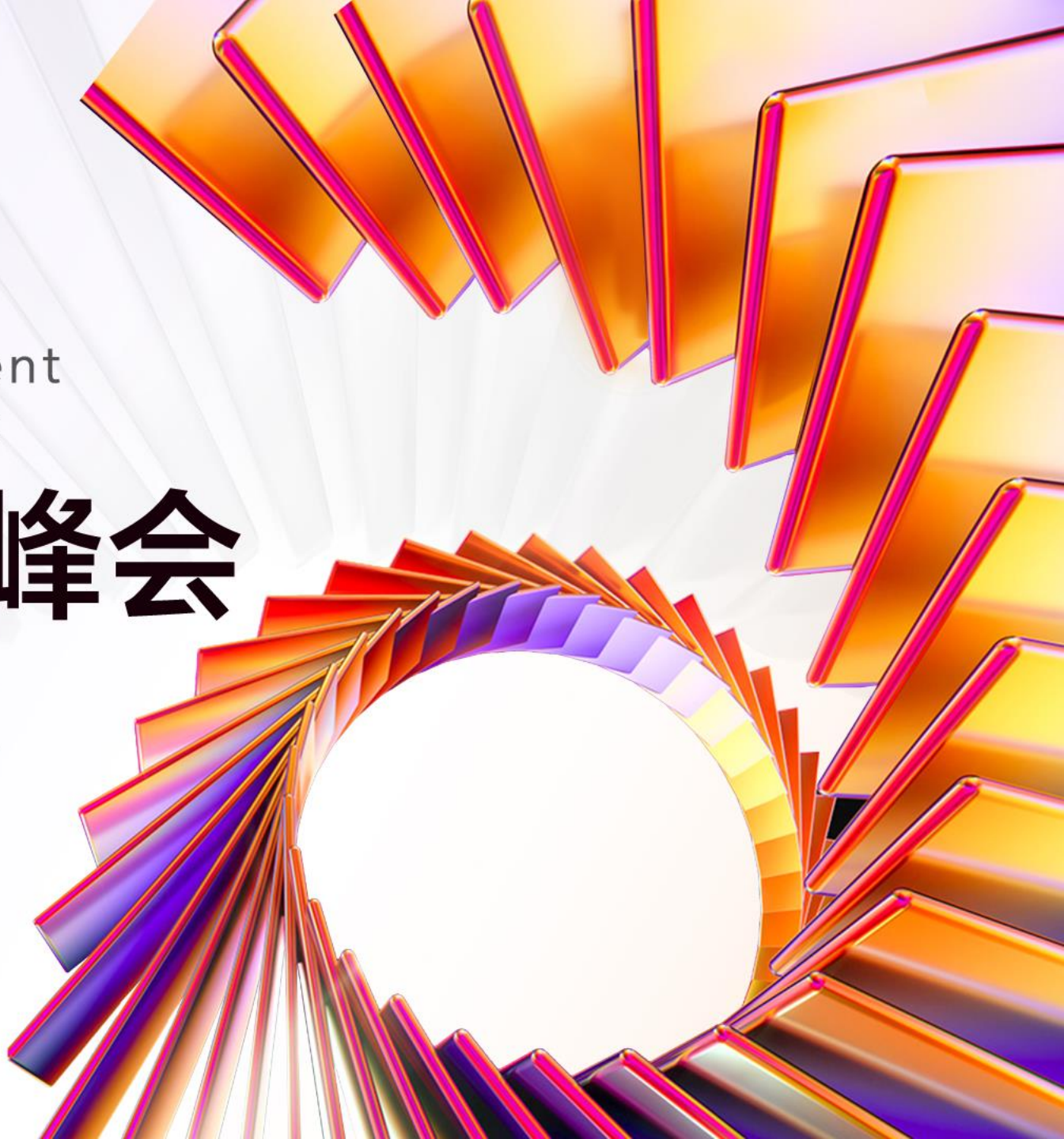




第8届 AI+ Development Digital Summit AI+研发数字峰会

拥抱AI 重塑研发

11月14-15日 | 深圳





2026 AI+ PRODUCT INNOVATION SUMMIT

01.16-17 · ShangHai

AI+产品创新峰会



Track 1: AI 产品战略与创新设计

从0到1的AI原生产品构建

论坛1: AI时代的用户洞察与需求发现

论坛2: AI原生产品战略与商业模式重构

论坛3: Agentic AI产品创新与交互设计

2-hour Speech: 回归本质



用户洞察的第一性

——2小时思维与方法论工作坊

在数字爆炸、AI迅速发展的时代，
仍然考验“看见”的“同理心”



Track 2: AI 产品开发与工程实践

从1到10的工程化落地实践

论坛1: 面向Agent智能体的产品开发

论坛2: 具身智能与AI硬件产品

论坛3: AI产品出海与本地化开发

Panel 1: 出海前瞻



“出海避坑地图”圆桌对话

——不止于翻译:

AI时代的出海新范式



Track 3: AI 产品运营与智能演化

从10到100的AI产品运营

论坛1: AI赋能产品运营与增长黑客

论坛2: AI产品的数据飞轮与智能演化

论坛3: 行业爆款AI产品案例拆解

Panel 2: 失败复盘



为什么很多AI产品“叫好不叫座”?

——从伪需求到真价值:

AI产品商业化落地的关键挑战

智能重构产品 数据驱动增长
Reinventing Products with Intelligence, Driven by Data

80+

业界大咖

汇聚产品大咖的真知灼见

40+

创新案例

开拓全新AI+产品视角

10+

分论坛

涵盖产品到商业增长的全链路

800+

听众规模

共同探索产品的智能重构

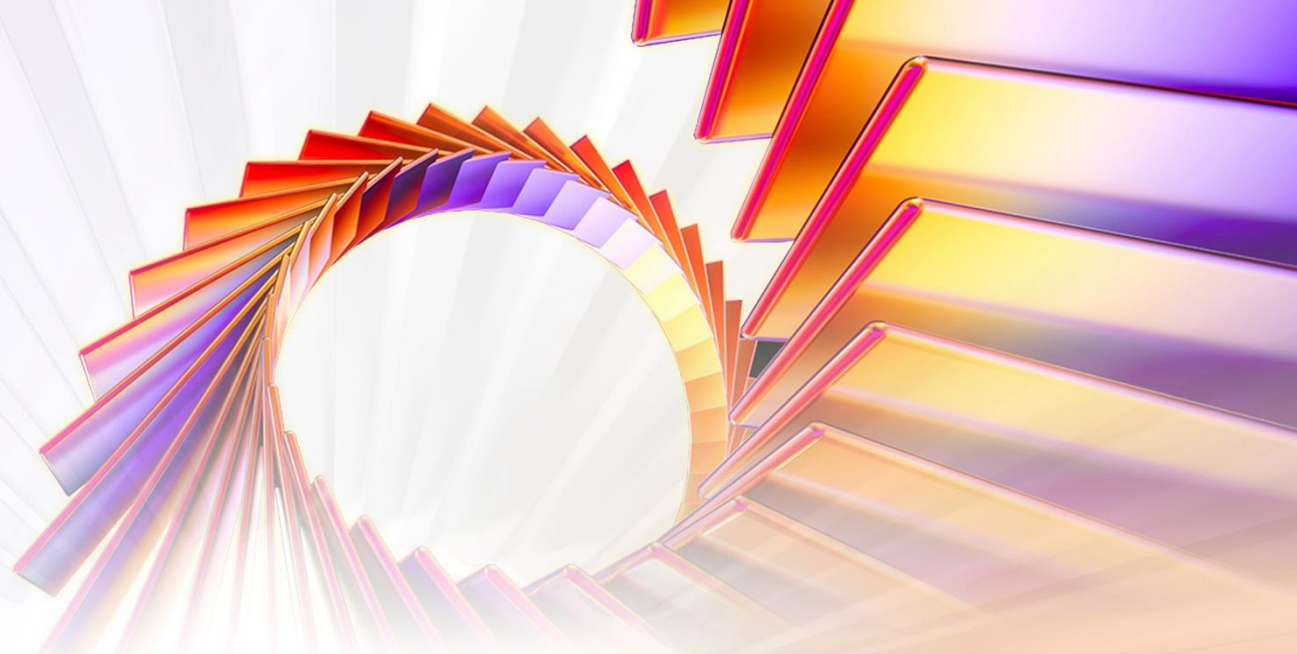


扫码查看会议详情

AI+DD 8th 2025 | 11月14-15日 | 深圳

第8届 AI+ Development Digital Summit AI+研发数字峰会

拥抱AI 重塑研发



基于大模型应用特征分析的 算力适配与应用优化

宋志方 | 北京并行科技股份有限公司



宋志方

并行科技 应用优化总监

深耕高性能计算与 AI 模型优化领域，兼具底层技术研发与产业落地经验，擅长通过软硬协同优化、并行架构设计及算法创新，解决大规模计算场景下的效率瓶颈与 AI 模型推理部署难题，参与优化过CFD、石油、电力等多款大型国产工业软件，主导了公司MaaS平台推理模型性能优化，大幅度提高了模型推理效率，为公司面向具身智能、AIGC、生物医药、工业仿真等领域的算力选型和7*24小时服务提供关键技术支撑。

目录

CONTENTS

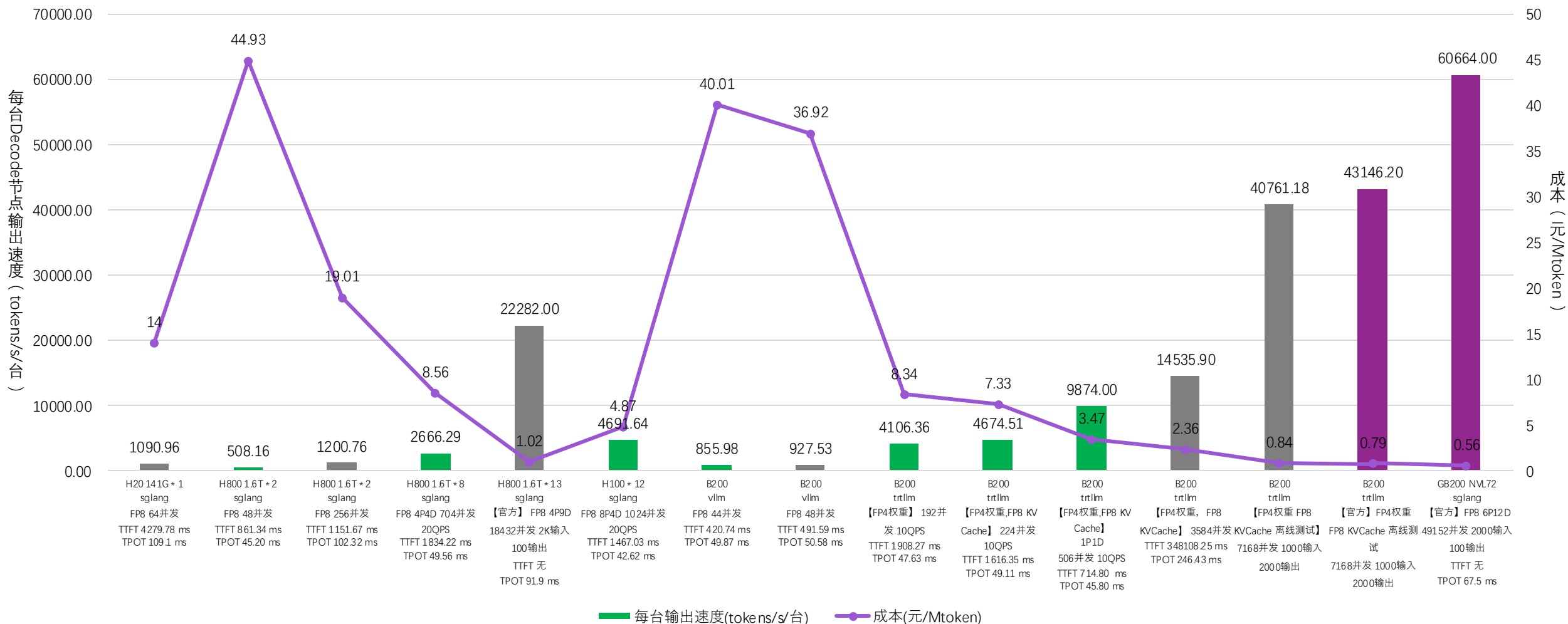
- I. DeepSeek应用运行特征分析
- II. 应用运行特征分析方法介绍
- III. 大模型性能优化案例
- IV. 总结与展望

PART 01

DeepSeek应用运行特征分析

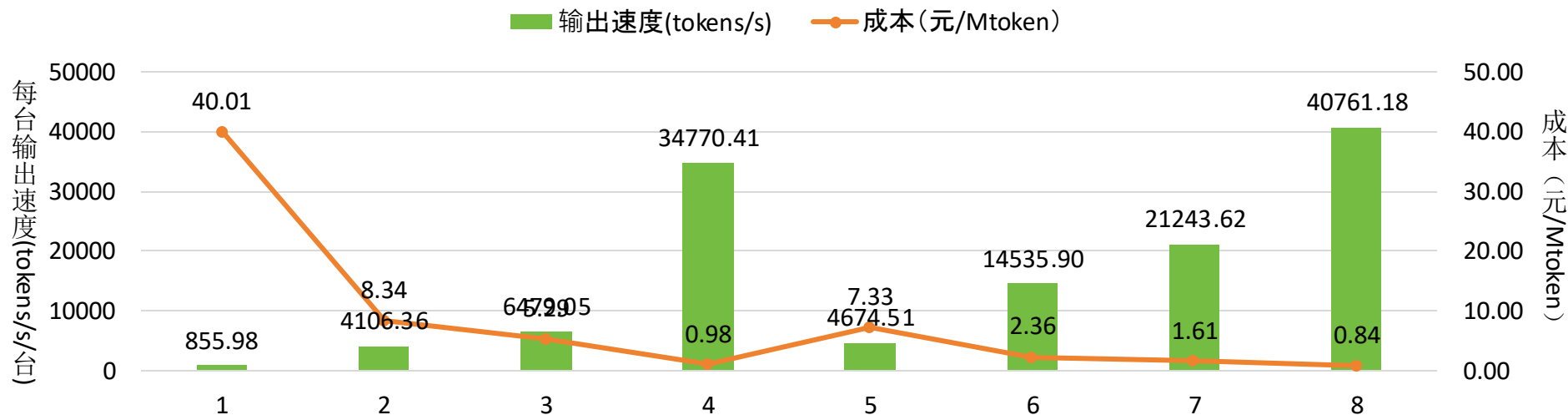
DeepSeek成本核算关键因素

DeepSeek-R1/V3不同环境输入3500输出1500下每台Decode节点输出速度和输出成本
(R1/V3/R1-0528/V3-0324成本相同)



▶ DeepSeek成本核算关键因素

DeepSeek-R1在B200上不同优化方法对于平均每台性能的影响



FP4 权重	X	●	●	●	●	●	●	●	
不管不顾(不限TTFT和TPOT)	X	X	●	●	X	●	●	●	
离线测试	X	X	●	●	X	X	X	●	否则为在线测试
缩小输入输出长度	X	X	X	●	X	X	●	●	
FP8 KVCache	X	X	X	X	●	●	●	●	
输入长度	3500	3500	3500	1000	3500	3500	1	1000	
输出长度	1500	1500	1500	2000	1500	1500	400	2000	
台数	1	1	1	1	1	1	1	1	
并发	44	192	256	3072	224	3584	7168	7168	
QPS	2	10	inf	inf	10	inf	inf	inf	
输出速度(tokens/s)	855.98	4106.36	6479.05	34770.41	4674.51	14535.90	21243.62	40761.18	



▶ DeepSeek的优化框架

DeepSeek的底层技术都是【并行计算+性能优化】！

Token间并行与缓存

PD分离

共享KV Cache

×1.5 性能提升

Token内并行

Tensor Parallel

Pipeline Parallel

Expert Parallel

×3 性能提升

Data Parallel

计算通信重叠

负载均衡

算法优化

MTP

×2 性能提升

算子优化

FlashMLA

DeepGEMM

DeepEP

×1.5 性能提升

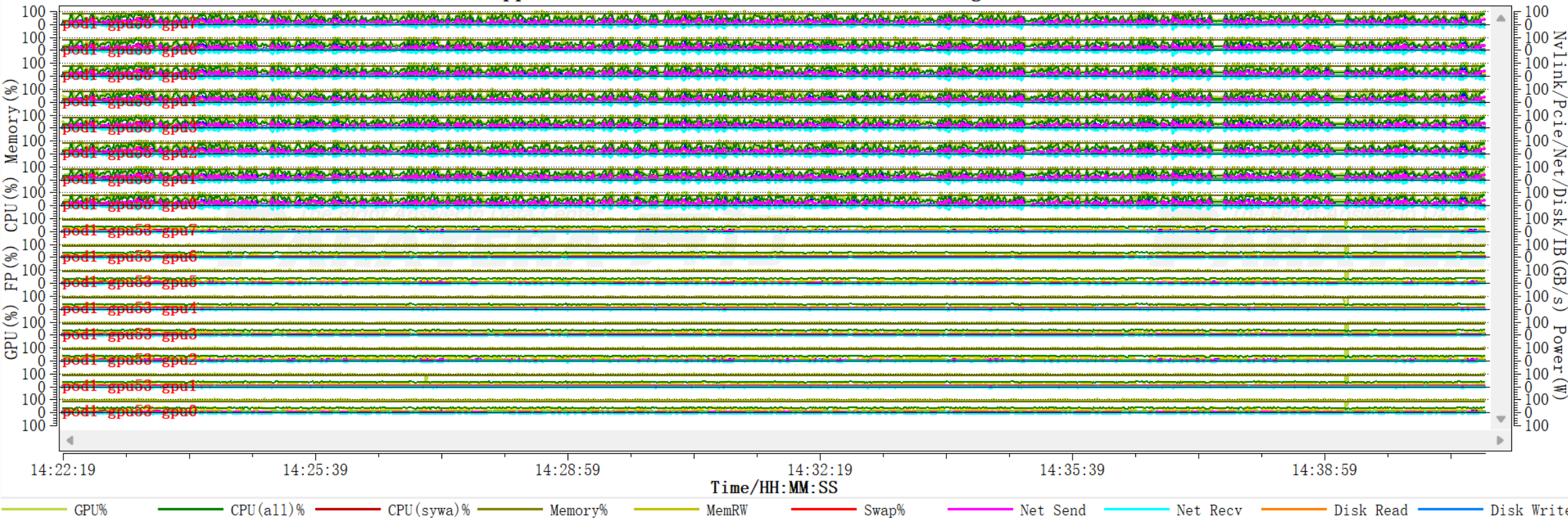




DeepSeek-V3.1 PD分离部署应用运行特征

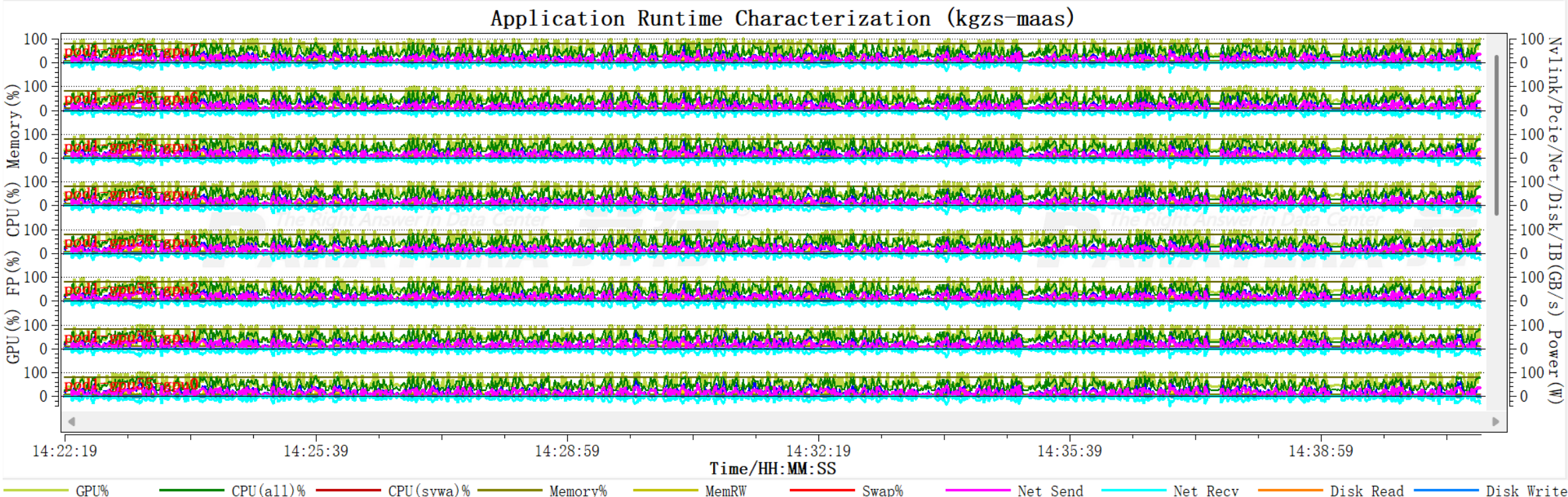
基本信息	平均值信息	峰值信息	总量信息	瞬时值信息	详细信息									
Gpu Index	GPU%	Tensorcore%	Fp64%	Fp32%	Fp16%	GPU Mem%	GPU MemBW%	Pcie TX(GB/s)	Pcie RX(GB/s)	NVLINK(GB/s)	GPU Power(W)	CPU(all)%	Memory%	IB Send(GB/s)
pod1-gpu55-gpu7	50.77	2.18	0.00	0.60	0.00	80.12	4.27	1.42	1.44	17.72	184.93	46.86	7.95	14.48
pod1-gpu55-gpu6	53.48	2.22	0.00	0.61	0.00	79.54	4.36	2.85	1.74	17.77	191.09	46.86	7.95	14.48
pod1-gpu55-gpu5	52.79	2.25	0.00	0.62	0.00	79.25	3.87	1.41	1.45	17.79	194.86	46.86	7.95	14.48
pod1-gpu55-gpu4	56.22	2.16	0.00	0.59	0.00	80.04	4.29	2.91	1.75	17.56	183.56	46.86	7.95	14.48
pod1-gpu55-gpu3	52.28	2.13	0.00	0.59	0.00	79.67	4.04	1.42	1.45	17.65	182.38	46.86	7.95	14.48
pod1-gpu55-gpu2	52.49	2.23	0.00	0.61	0.00	80.94	4.36	2.68	1.52	17.73	189.08	46.86	7.95	14.48

Application Runtime Characterization (kgzs-maas)

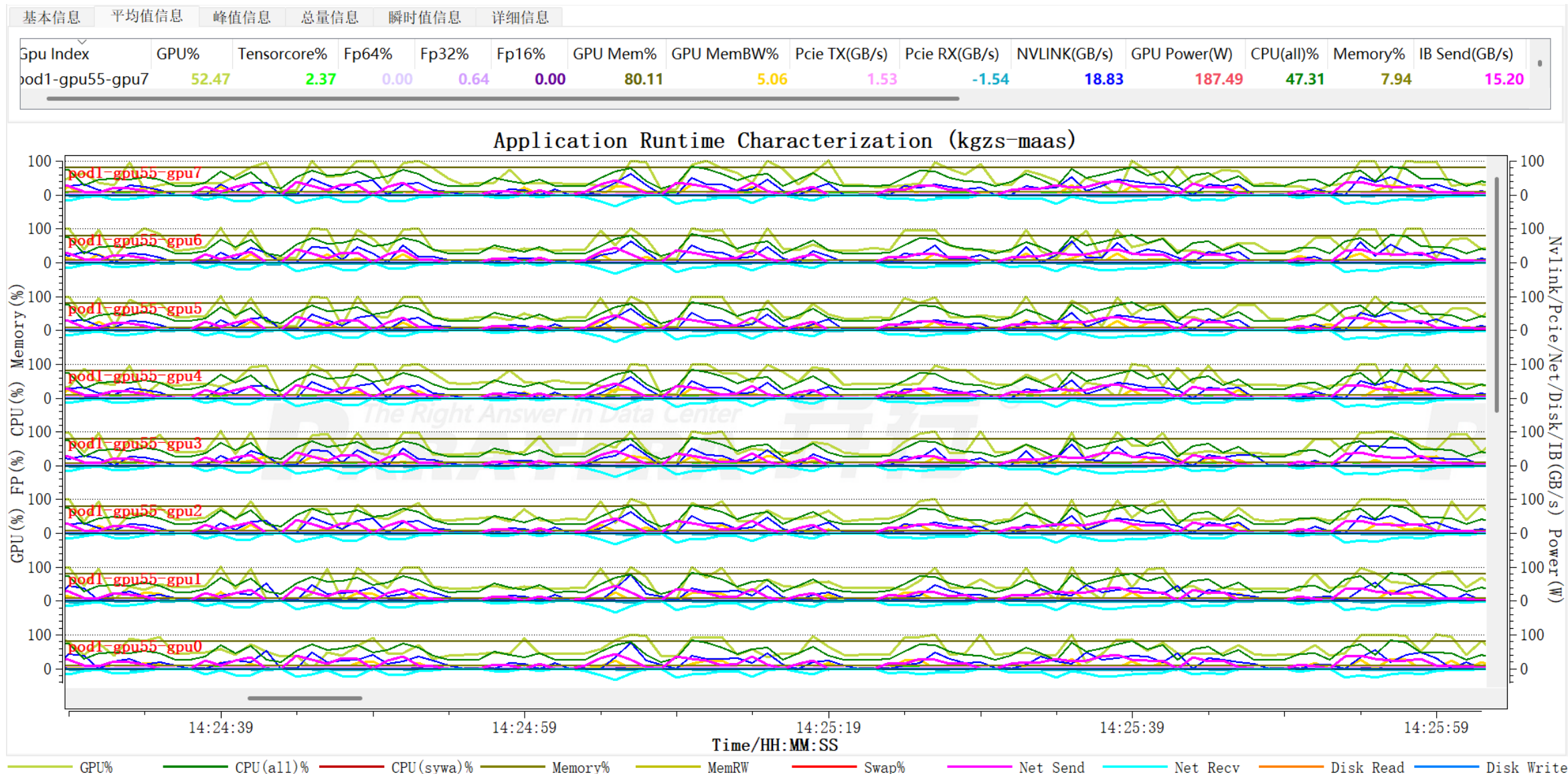


DeepSeek-V3.1 P阶段应用运行特征

基本信息	平均值信息	峰值信息	总量信息	瞬时值信息	详细信息										
Gpu Index	GPU%	Tensorcore%	Fp64%	Fp32%	Fp16%	GPU Mem%	GPU MemBW%	Pcie TX(GB/s)	Pcie RX(GB/s)	NVLINK(GB/s)	GPU Power(W)	CPU(all)%	Memory%	IB Send(GB/s)	
pod1-gpu55-gpu7	50.77	2.18	0.00	0.60	0.00	80.12	4.27	1.42	1.44	17.72	184.93	46.86	7.95	14.48	
pod1-gpu55-gpu6	53.48	2.22	0.00	0.61	0.00	79.54	4.36	2.85	1.74	17.77	191.09	46.86	7.95	14.48	
pod1-gpu55-gpu5	52.79	2.25	0.00	0.62	0.00	79.25	3.87	1.41	1.45	17.79	194.86	46.86	7.95	14.48	
pod1-gpu55-gpu4	56.22	2.16	0.00	0.59	0.00	80.04	4.29	2.91	1.75	17.56	183.56	46.86	7.95	14.48	
pod1-gpu55-gpu3	52.28	2.13	0.00	0.59	0.00	79.67	4.04	1.42	1.45	17.65	182.38	46.86	7.95	14.48	
pod1-gpu55-gpu2	52.49	2.23	0.00	0.61	0.00	80.94	4.36	2.68	1.52	17.73	189.08	46.86	7.95	14.48	
pod1-gpu55-gpu1	51.43	2.19	0.00	0.60	0.00	80.60	4.42	1.42	1.46	17.75	184.29	46.86	7.95	14.48	
pod1-gpu55-gpu0	57.85	2.21	0.00	0.61	0.00	79.72	4.47	2.87	1.75	17.70	187.15	46.86	7.95	14.48	



DeepSeek-V3.1 P阶段应用运行特征



▶ DeepSeek-V3.1 P阶段应用运行特征-平均值

基本信息 平均值信息 峰值信息 总量信息 瞬时值信息 详细信息

Gpu Index	GPU%	Tensorcore%	Fp64%	Fp32%	Fp16%	GPU Mem%	GPU MemBW%	Pcie TX(GB/s)	Pcie RX(GB/s)	NVLINK(GB/s)	GPU Power(W)	CPU(all)%	Memory%	IB Send(GB/s)
pod1-gpu55-gpu7	50.77	2.18	0.00	0.60	0.00	80.12	4.27	1.42	-1.44	17.72	184.93	46.86	7.95	14.48
pod1-gpu55-gpu6	53.48	2.22	0.00	0.61	0.00	79.54	4.36	2.85	-1.74	17.77	191.09	46.86	7.95	14.48
pod1-gpu55-gpu5	52.79	2.25	0.00	0.62	0.00	79.25	3.87	1.41	-1.45	17.79	194.86	46.86	7.95	14.48
pod1-gpu55-gpu4	56.22	2.16	0.00	0.59	0.00	80.04	4.29	2.91	-1.75	17.56	183.56	46.86	7.95	14.48
pod1-gpu55-gpu3	52.28	2.13	0.00	0.59	0.00	79.67	4.04	1.42	-1.45	17.65	182.38	46.86	7.95	14.48
pod1-gpu55-gpu2	52.49	2.23	0.00	0.61	0.00	80.94	4.36	2.68	-1.52	17.73	189.08	46.86	7.95	14.48
pod1-gpu55-gpu1	51.43	2.19	0.00	0.60	0.00	80.60	4.42	1.42	-1.46	17.75	184.29	46.86	7.95	14.48
pod1-gpu55-gpu0	57.85	2.21	0.00	0.61	0.00	79.72	4.47	2.87	-1.75	17.70	187.15	46.86	7.95	14.48

性能特征平均值：模型推理周期内，各指标采集性能均值

重点参考数值及相对硬件配置密集程度如下：

gpu利用率 53.41%，低

TensorCore利用率： 2.19%，低

FP32单元利用率： 0.60%，低

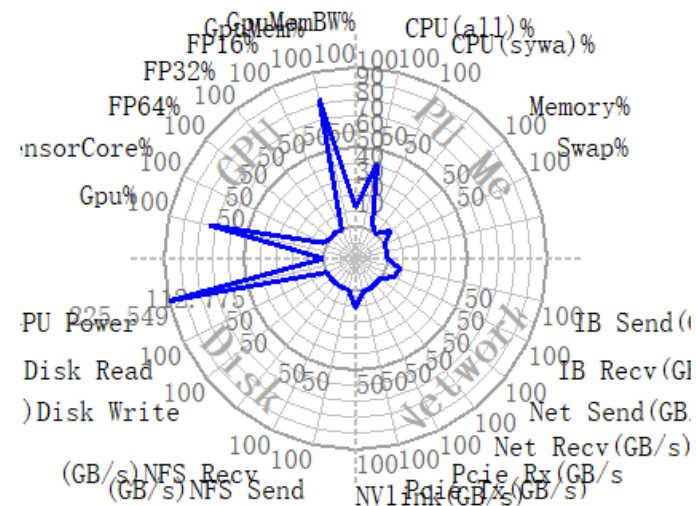
FP16单元利用率： 0.00%，低

显存利用率 79.98%，高

显存带宽利用率： 4.26%，低

NVLINK带宽： 17.72GB/s，低

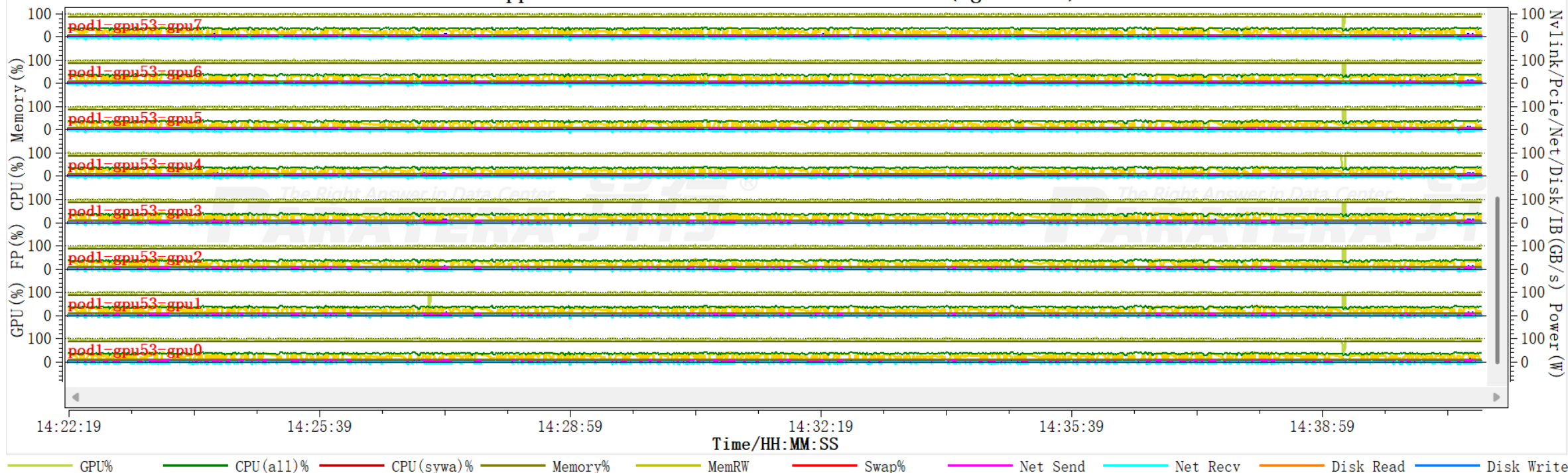
CPU利用率： 46.86%，高



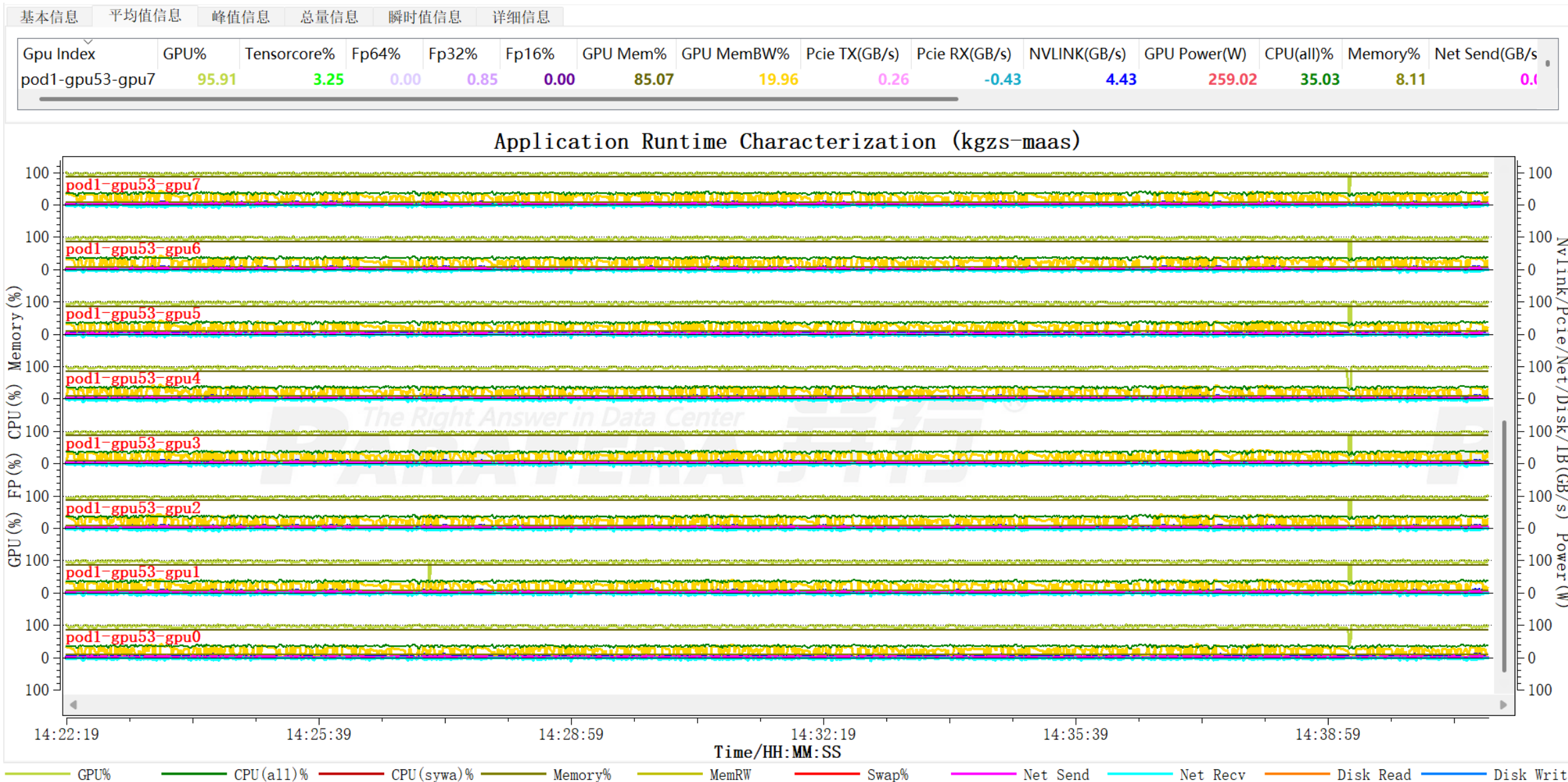
DeepSeek-V3.1 D阶段应用运行特征

基本信息		平均值信息		峰值信息		总量信息		瞬时值信息		详细信息				
Gpu Index	GPU%	Tensorcore%	Fp64%	Fp32%	Fp16%	GPU Mem%	GPU MemBW%	Pcie TX(GB/s)	Pcie RX(GB/s)	NVLINK(GB/s)	GPU Power(W)	CPU(all)%	Memory%	IB Send(GB/s)
pod1-gpu53-gpu7	95.91	3.25	0.00	0.85	0.00	85.07	19.96	0.26	-0.43	4.43	259.02	35.03	8.11	2.17
pod1-gpu53-gpu6	95.93	3.24	0.00	0.84	0.00	85.96	19.83	0.25	-0.47	4.45	269.61	35.03	8.11	2.17
pod1-gpu53-gpu5	95.91	3.20	0.00	0.84	0.00	85.96	19.57	0.28	-0.49	4.43	280.04	35.03	8.11	2.17
pod1-gpu53-gpu4	95.62	3.26	0.00	0.85	0.00	85.96	20.02	0.32	-0.55	4.51	257.87	35.03	8.11	2.17
pod1-gpu53-gpu3	95.59	3.27	0.00	0.86	0.00	85.96	20.31	0.35	-0.63	4.55	266.95	35.03	8.11	2.17
pod1-gpu53-gpu2	95.72	3.35	0.00	0.87	0.00	85.96	20.33	0.36	-0.63	4.59	264.53	35.03	8.11	2.17
pod1-gpu53-gpu1	95.64	3.15	0.00	0.83	0.00	85.96	19.46	0.36	-0.64	4.59	250.75	35.03	8.11	2.17
pod1-gpu53-gpu0	95.79	3.10	0.00	0.81	0.00	85.78	19.18	0.36	-0.71	4.61	262.68	35.03	8.11	2.17

Application Runtime Characterization (kgzs-maas)



DeepSeek-V3.1 D阶段应用运行特征



▶ DeepSeek-V3.1 D阶段应用运行特征-平均值

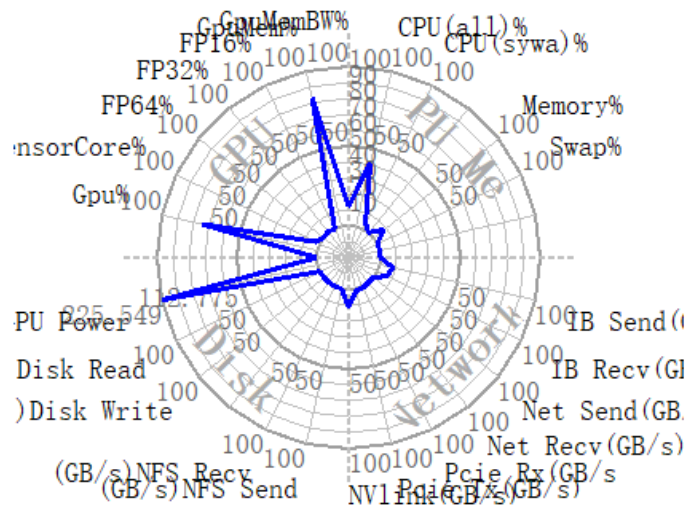
基本信息 平均值信息 峰值信息 总量信息 瞬时值信息 详细信息

Gpu Index	GPU%	Tensorcore%	Fp64%	Fp32%	Fp16%	GPU Mem%	GPU MemBW%	Pcie TX(GB/s)	Pcie RX(GB/s)	NVLINK(GB/s)	GPU Power(W)	CPU(all)%	Memory%	IB Send(GB/s)
pod1-gpu53-gpu7	95.91	3.25	0.00	0.85	0.00	85.07	19.96	0.26	-0.43	4.43	259.02	35.03	8.11	2.17
pod1-gpu53-gpu6	95.93	3.24	0.00	0.84	0.00	85.96	19.83	0.25	-0.47	4.45	269.61	35.03	8.11	2.17
pod1-gpu53-gpu5	95.91	3.20	0.00	0.84	0.00	85.96	19.57	0.28	-0.49	4.43	280.04	35.03	8.11	2.17
pod1-gpu53-gpu4	95.62	3.26	0.00	0.85	0.00	85.96	20.02	0.32	-0.55	4.51	257.87	35.03	8.11	2.17
pod1-gpu53-gpu3	95.59	3.27	0.00	0.86	0.00	85.96	20.31	0.35	-0.63	4.55	266.95	35.03	8.11	2.17
pod1-gpu53-gpu2	95.72	3.35	0.00	0.87	0.00	85.96	20.33	0.36	-0.63	4.59	264.53	35.03	8.11	2.17
pod1-gpu53-gpu1	95.64	3.15	0.00	0.83	0.00	85.96	19.46	0.36	-0.64	4.59	250.75	35.03	8.11	2.17
pod1-gpu53-gpu0	95.79	3.10	0.00	0.81	0.00	85.78	19.18	0.36	-0.71	4.61	262.68	35.03	8.11	2.17

性能特征平均值：模型推理周期内，各指标采集性能均值

重点参考数值及相对硬件配置密集程度如下：

- gpu利用率 95.91%，高
- TensorCore利用率： 3.26%，低
- FP32单元利用率： 0.85%，低
- FP16单元利用率： 0.00%，低
- 显存利用率 85.96%，高
- 显存带宽利用率： 20.02%，中
- NVLINK带宽： 4.43GB/s，低
- CPU利用率： 35.03%，中



分析：GPU利用率高，主要利用Tensorcore计算，FP32很低，显存带宽利用率中等，NVLink带宽利用率低，显存利用率高，CPU利用率中等。



不同硬件平台的性能比值

类别	卡型号	Model Name	FP16 算力 Tflop s	显存 带宽 TB/s	卡间 通信 GB/s	计算访存比 Flop/Byte	访存计算比 Byte/MFlop	计算卡间通 信比 双向, Flop/Byte	卡间通信计 算比 双向, Byte/Mflo p
Ada Lovelace系列	5090	RTX 5090-32G	419	1.79	128	234.08	4,272.08	3,273.44	305.49
Blackwell系列	B200	B200-180G-SXM	2200	7.7	1800	285.71	3,500.00	1,222.22	818.18
Blackwell系列	GB200	GB200-186G-SXM	2500	8	1800	312.50	3,200.00	1,388.89	720.00
Hopper 系列	H800	H800-80G-SXM	989	3.35	400	295.22	3,387.26	2,472.50	404.45
Hopper 系列	H200	H200-141GB-SXM	989	4.8	900	206.04	4,853.39	1,098.89	910.01
Hopper 系列	H100	H100-80GB-SXM	989	3.35	900	295.22	3,387.26	1,098.89	910.01
Hopper 系列	H20	H20-96G-SXM	148	4	900	37.00	27,027.03	164.44	6,081.08
Ascend	910C	Ascend 910C-128G	800	3.2	392	250.00	4,000.00	2,040.82	490.00
Ascend	910B	Ascend 910B3-64G	313	1.6	392	195.63	5,111.82	798.47	1,252.40



► B200和H200主要的性能参数对比

特性	B200	H200
FP64	37 TFLOPS	34 TFLOPS
FP32	75 TFLOPS	67 TFLOPS
FP64 Tensor	37 TFLOPS	67 TFLOPS
TP32 Tensor	1.1 PFLOPS	494 TFLOPS
BF16 Tensor	2.2 PFLOPS	989 TFLOPS
FP16 Tensor	2.2 PFLOPS	989 TFLOPS
FP8 Tensor	4.5 PFLOPS	1,979 TFLOPS
FP6 Tensor	9 PFLOPS	N/A
FP4 Tensor	4.5 PFLOPS	N/A
INT8 Tensor	4.5 POPS	1,979 TOPS
GPU Memory	180GB HBM3e	141GB HBM3e
GPU Memory Bandwidth	7.7 TB/s	4.8TB/s
Interconnect	NVLink: 1.8TB/s PCIe Gen5: 128GB/s	NVLink: 900GB/s PCIe Gen5: 128GB/s



► P阶段性能预测数据分析和结论

预测公式为：

$$Time_{target} = Time_{H100} / \left[k_1 * \left(\frac{FP16_{H100}}{FP16_{target}} \right) + k_2 * \left(\frac{GMemBW_{H100}}{GMemBW_{target}} \right) + k_3 * \left(\frac{NVLink_{H100}}{NVLink_{target}} \right) + k_4 * \left(\frac{IB_{H100}}{IB_{target}} \right) \right]$$

平台	FP16 Tensor TFLOPS	GPU Mem GB	GPU MemBW GB/s	NVLink/PCIe GB/s	预测耗时 (s)	性价比
H100	989	80	3350	900	1128 (1台)	1
H200	989	141	4800	900	1128 (1台)	0.85
B200	2252.8	180	7884.8	1843.2	495 (1台)	1.13



► D阶段性能预测数据分析和结论

预测公式为：

$$Time_{target} = Time_{H100} / \left[k_1 * \left(\frac{FP16_{H100}}{FP16_{target}} \right) + k_2 * \left(\frac{GMemBW_{H100}}{GMemBW_{target}} \right) + k_3 * \left(\frac{NVLink_{H100}}{NVLink_{target}} \right) + k_4 * \left(\frac{IB_{H100}}{IB_{target}} \right) \right]$$

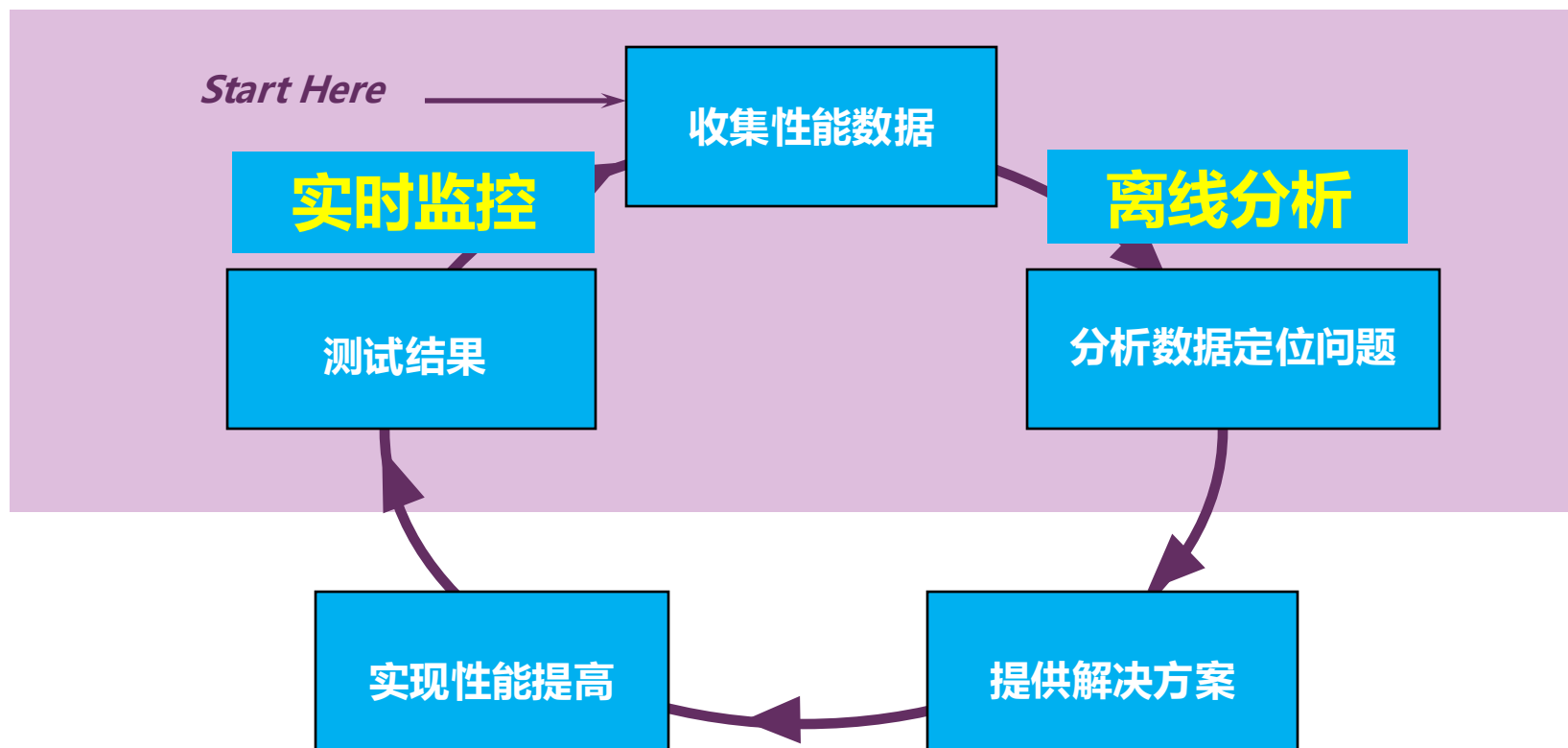
平台	FP16 Tensor TFLOPS	GPU Mem GB	GPU MemBW GB/s	NVLink/PCIe GB/s	预测耗时 (s)	性价比
H100	989	80	3430.4	900	1128 (1台)	1
H200	989	141	4915.2	900	815 (1台)	1.18
B200	2252.8	180	7884.8	1843.2	495 (1台)	1.13



PART 02

应用运行特征分析方法

可查看、可记录、可分析、可交流



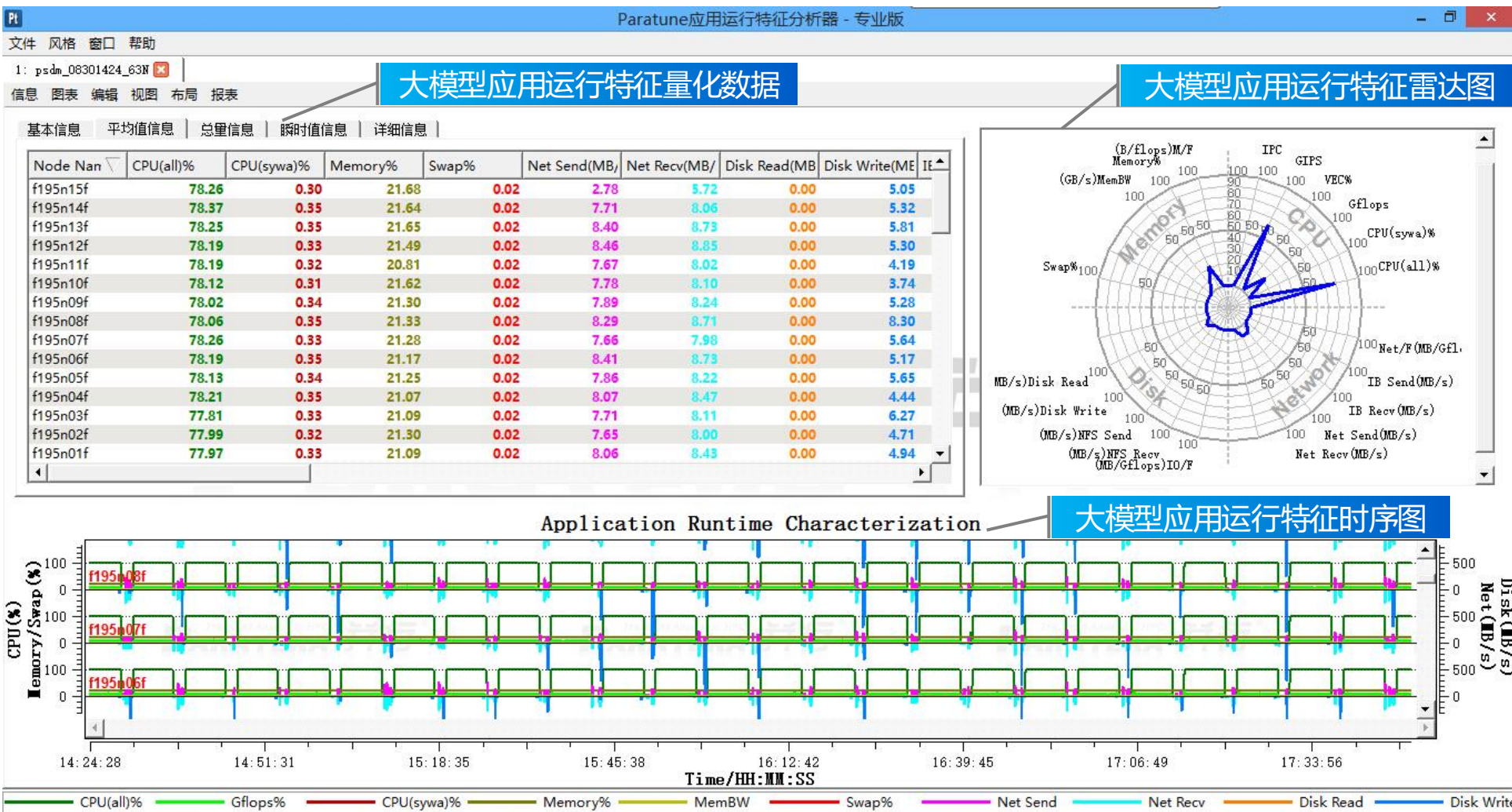
大模型应用运行特征分析方法论

采集应用运行特征

- GPU利用率
- Tensor Core利用率
- FP32利用率
- FP16利用率
- 显存利用率
- 显存带宽利用率
- PCIe利用率
- NVLink利用率
- 硬盘读写速率
- IB读写速率

分析应用瓶颈

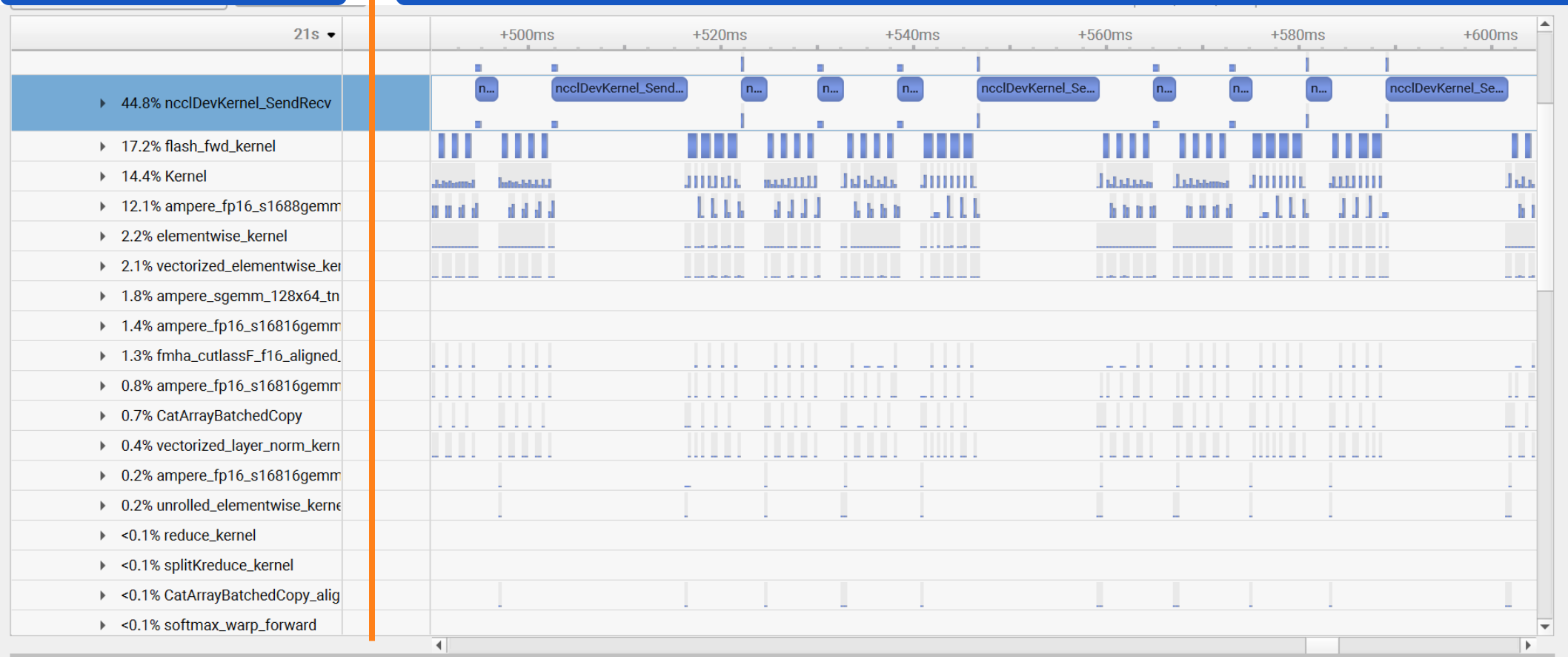
- 计算密集型
- 访存密集型
- 通信密集型



大模型应用运行特征分析方法论

应用函数热点

应用函数执行特征



- 1.应用运行特征：计算占比55.2%，通信占比44.8%，计算与通信交替执行。
- 2.最大的函数热点为ncclDevKernel_SendRecv，耗时占比44.8%。
- 3.最大的计算函数热点为flash_fwd_kernel,耗时占比为17.2%。

H100 3.2T RoCE vs H200 3.2T RoCE (Decode 单并发)

算子	H100 3.2T RoCE * 2		H200 3.2T RoCE * 1	
	耗时(ms)	占比(%)	耗时(ms)	占比(%)
attention core	2.15	7.53%	2.17	8.06%
moe	7.87	27.63%	8.80	32.74%
gemm	0.63	2.22%	0.43	1.59%
nccl reducescatter	11.47	40.28%	11.10	41.28%
nccl allgather	1.91	6.69%	0.98	3.66%
nccl allreduce	0.12	0.41%	0.00	0.00%
deepen dispatch	0.88	3.08%	0.72	2.69%
deepen combine	3.46	12.16%	2.68	9.97%
overall	28.47	100.00%	26.88	100.00%



大模型应用运行特征分析方法论

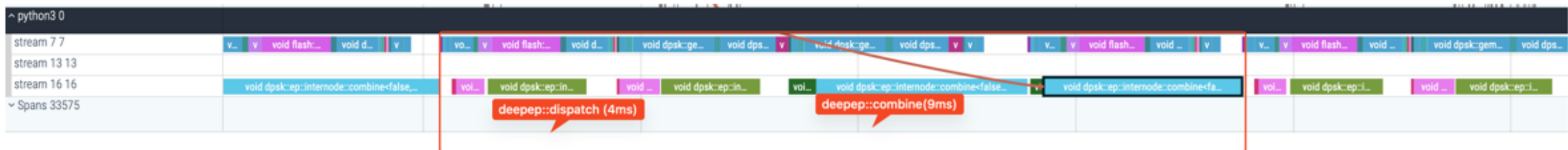
通信函数热点分析，程序通信结构及负载均衡性分析

Flat Profile						Load Balance						Call Tree						Call Graph					
Children of All_Processes						Show Pies						Children of All_Processes						Show Pies					
Name	TSelf	TSelf	TTotal	#Calls	TSelf /Call	Name	TSelf	TSelf	TTotal	#Calls	TSelf /Call												
Group Application	182.224e+3 s		496.09e+3 s	0	n.a.	Process 242	43.8442 s		43.8442 s	368611	118.944e-6 s												
MPI_Waitany	131.207e+3 s		131.207e+3 s	59196544	2.21646e-3 s	Process 243	34.6129 s		34.6129 s	361116	95.8499e-6 s												
MPI_Wait	72.6074e+3 s		72.6074e+3 s	307458010	236.154e-6 s	Process 244	31.5401 s		31.5401 s	363597	86.7446e-6 s												
MPI_Waitall	41.1964e+3 s		41.1964e+3 s	257475522	160.001e-6 s	Process 245	36.5823 s		36.5823 s	348589	104.944e-6 s												
MPI_Bcast	27.4708e+3 s		27.4708e+3 s	35201600	780.386e-6 s	Process 246	42.7672 s		42.7672 s	336635	127.043e-6 s												
MPI_Allreduce	20.6491e+3 s		20.6491e+3 s	67338608	306.646e-6 s	Process 247	37.0474 s		37.0474 s	319441	115.976e-6 s												
MPI_Alltoallv	7.35054e+3 s		7.35054e+3 s	621312	11.8307e-3 s	Process 248	42.6861 s		42.6861 s	327557	130.317e-6 s												
MPI_Recv	5.6033e+3 s		5.6033e+3 s	3609928	1.55219e-3 s	Process 249	46.2897 s		46.2897 s	318475	145.348e-6 s												
MPI_Barrier	3.83101e+3 s		3.83101e+3 s	1605376	2.38636e-3 s	Process 250	46.319 s		46.319 s	317293	145.982e-6 s												
MPI_Isend	1.67779e+3 s		1.67779e+3 s	1011992018	1.65791e-6 s	Process 251	42.8874 s		42.8874 s	282868	151.616e-6 s												
MPI_Rsend	825.112 s		825.112 s	107160261	7.6998e-6 s	Process 252	48.0882 s		48.0882 s	250865	191.689e-6 s												
MPI_Send	691.892 s		691.892 s	110470855	6.26312e-6 s	Process 253	42.6505 s		42.6505 s	269961	157.988e-6 s												
MPI_Irecv	417.92 s		417.92 s	1226013288	340.877e-9 s	Process 254	50.4058 s		50.4058 s	202701	248.671e-6 s												
MPI_Wtime	233.849 s		233.849 s	1990666949	117.473e-9 s	Process 255	48.3329 s		48.3329 s	180609	267.61e-6 s												
MPI_Alltoall	29.956 s		29.956 s	18176	1.64811e-3 s	Process 256	475.296 s		475.296 s	3264	145.618e-3 s												
MPI_Comm_rank	28.9046 s		28.9046 s	183924062	157.155e-9 s	Process 257	471.571 s		471.571 s	3264	144.476e-3 s												
MPI_Comm_dup	12.5292 s		12.5292 s	8705	1.43931e-3 s	Process 258	471.566 s		471.566 s	3264	144.475e-3 s												
MPI_Comm_size	11.5708 s		11.5708 s	62519860	185.074e-9 s	Process 259	471.562 s		471.562 s	3264	144.474e-3 s												
MPI_Comm_split	8.9435 s		8.9435 s	2560	3.49356e-3 s	Process 260	471.561 s		471.561 s	3264	144.473e-3 s												
MPI_Comm_create	3.13722 s		3.13722 s	3072	1.02123e-3 s	Process 261	471.557 s		471.557 s	3264	144.472e-3 s												
MPI_Reduce	1.95772 s		1.95772 s	40320	48.5545e-6 s	Process 262	471.552 s		471.552 s	3264	144.471e-3 s												
MPI_Scatterv	1.72674 s		1.72674 s	7424	232.589e-6 s	Process 263	471.547 s		471.547 s	3264	144.469e-3 s												
MPI_Gatherv	1.00005 s		1.00005 s	512	1.95322e-3 s	Process 264	471.54 s		471.54 s	3264	144.467e-3 s												
MPI_Info_create	959.992e-3 s		959.992e-3 s	2048	468.746e-6 s	Process 265	471.533 s		471.533 s	3264	144.465e-3 s												
MPI_Get_processor_name	652.998e-3 s		652.998e-3 s	512	1.27539e-3 s	Process 266	471.523 s		471.523 s	3264	144.462e-3 s												
MPI_Allgather	589.489e-3 s		589.489e-3 s	512	1.15135e-3 s	Process 267	471.513 s		471.513 s	3264	144.459e-3 s												
						Process 268	471.501 s		471.501 s	3264	144.455e-3 s												
						Process 269	471.489 s		471.489 s	3264	144.451e-3 s												
						Process 270	471.475 s		471.475 s	3264	144.447e-3 s												
						Process 271	471.456 s		471.456 s	3264	144.441e-3 s												
						Process 272	471.543 s		471.543 s	3264	144.468e-3 s												

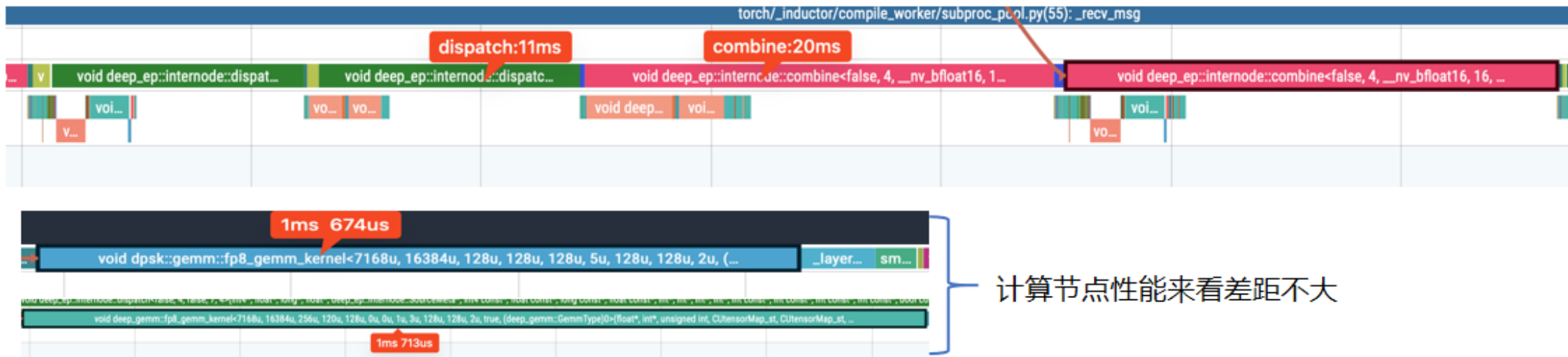


大模型应用运行特征分析方法论

DeepSeek官方Profile:



可以看出，整个循环是由两次`deepep::dispatch` 和两次 `deepep::combine`组成，DeepSeek官方单次的时间分别为4ms和9ms。



计算节点性能来看差距不大

但是从我们平均每张卡输出的profile的log来看，dispatch和combine要慢2倍以上，计算不能掩盖通讯部分。



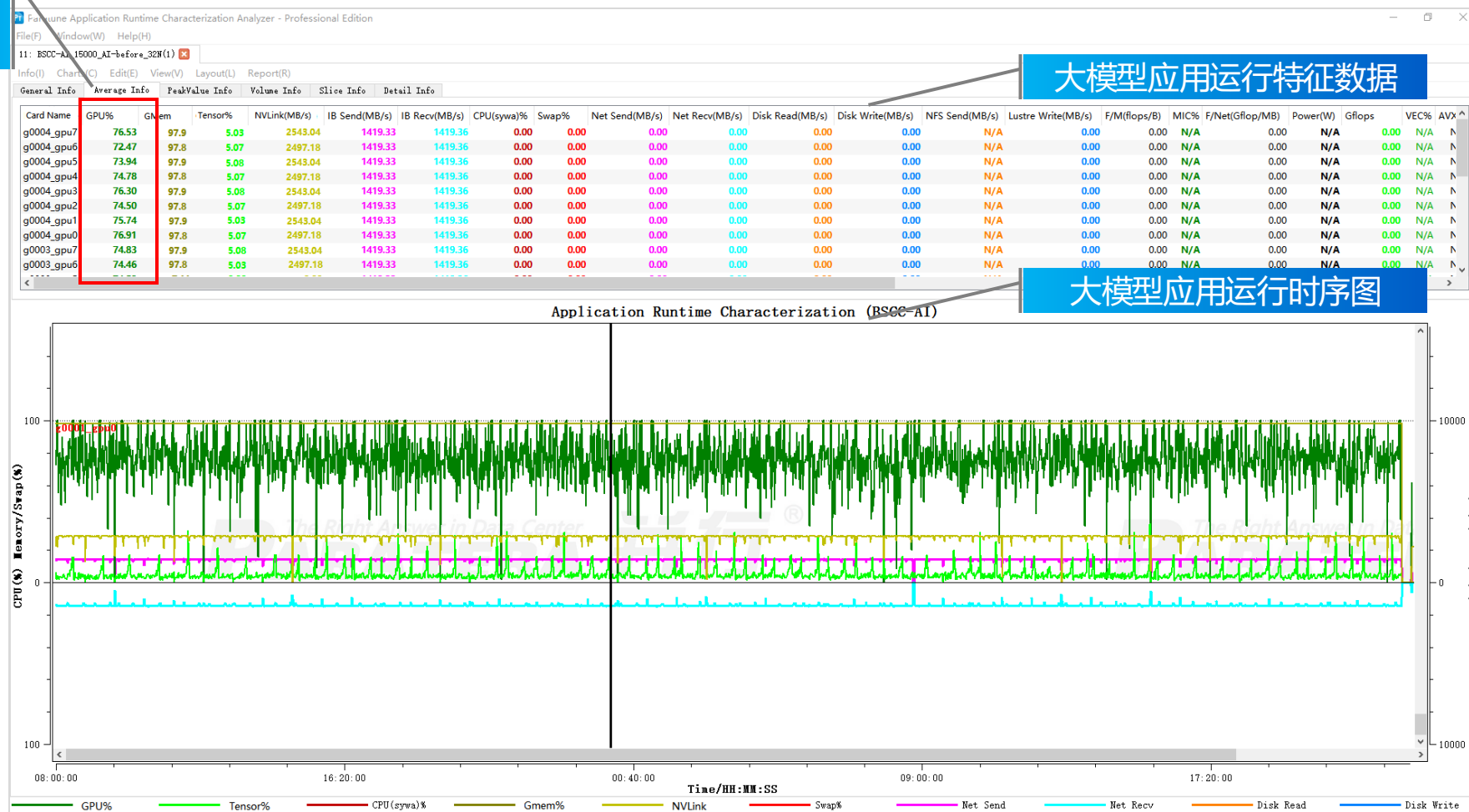
PART 03

大模型性能优化案例

某用户1300亿参数模型应用运行特征分析（优化前）

75%
GPU利用率

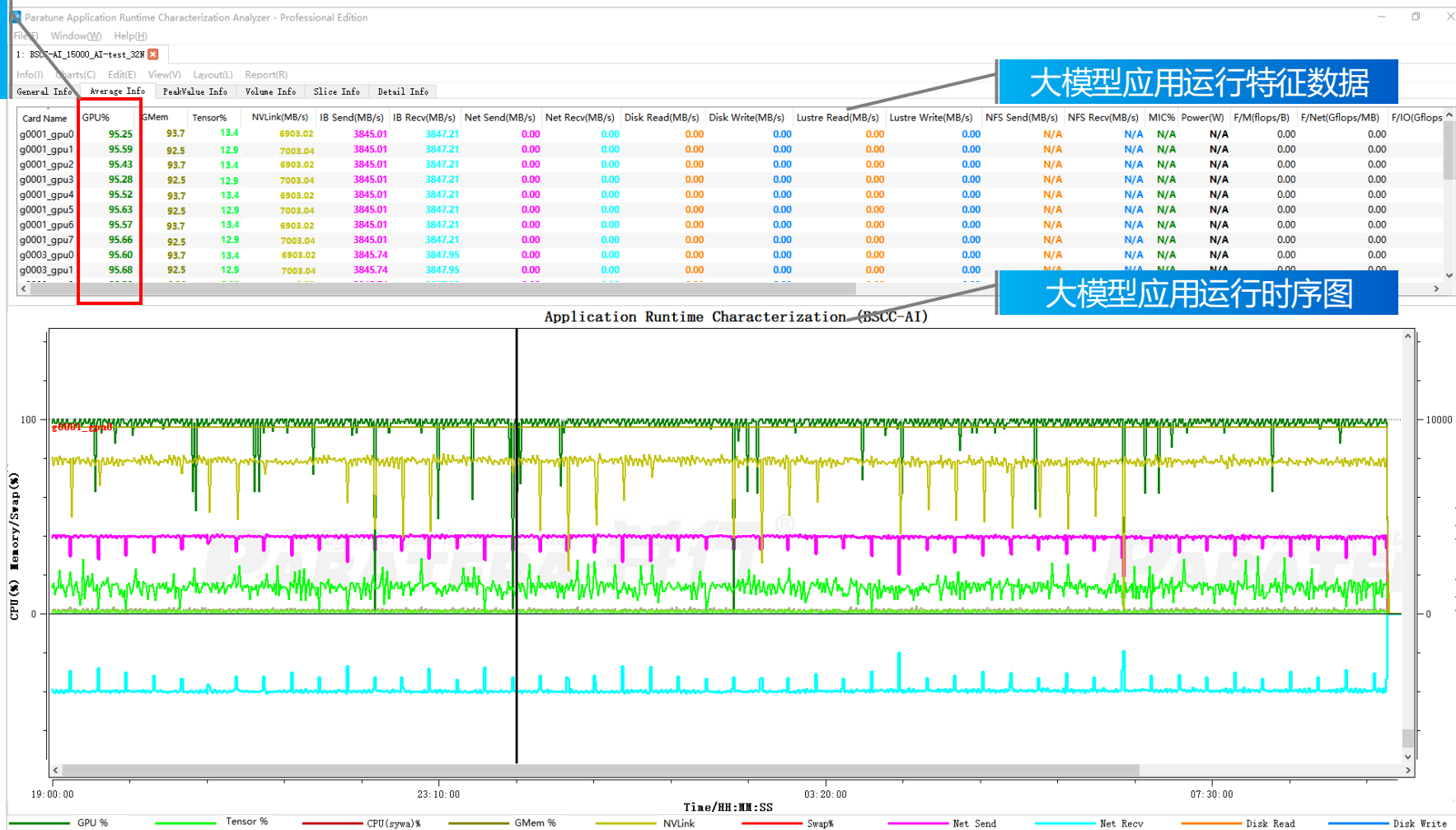
- 右图为上页某1300模型应用运行特征中所截取一段时间内的运行特征数据
- 分析：在该应用运行过程中GPU利用率75%左右，NVLink通信平均速度在2500MB/s，IB通信发送接收速度为1419MB/s，可见该应用最大瓶颈为GPU利用率有待提高
- 针对性优化措施：分析改进应用程序计算负载设计，充分利用GPU资源



某用户1300亿参数模型应用运行特征分析（优化后）

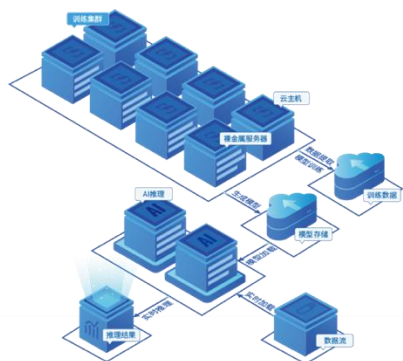
95%
GPU利用率

- 右图为上页某1300亿模型应用运行特征中所截取一段时间内的运行特征数据
- 分析：在该应用运行过程中GPU利用率95%左右，NVLink通信平均速度在7000MB/s，IB通信发送接收速度为3845MB/s
- 结论：随着GPU利用率的提升，NVLink和IB通信都得到了充分利用，整体效率提升了40%+



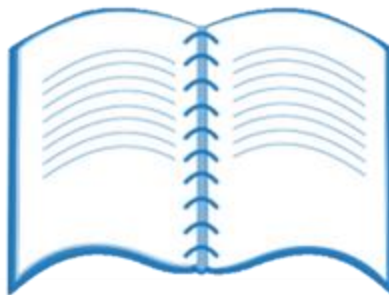
模型训练

- 对Llama3-70B / Gemma2-27B进行全参微调
- 使用LlamaFactory对微调模型进行DPO强化学习训练



小说生成

- 在昇腾平台部署Llama3 / Gemma2架构的模型推理服务
- 使用训练后的大模型生成小说

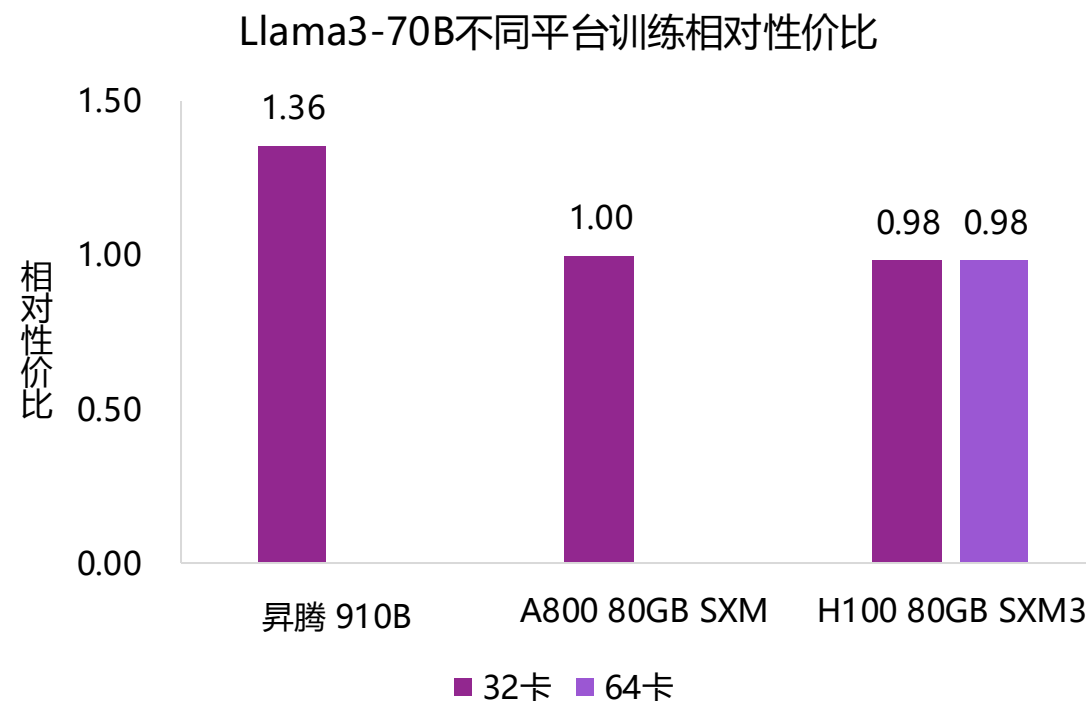
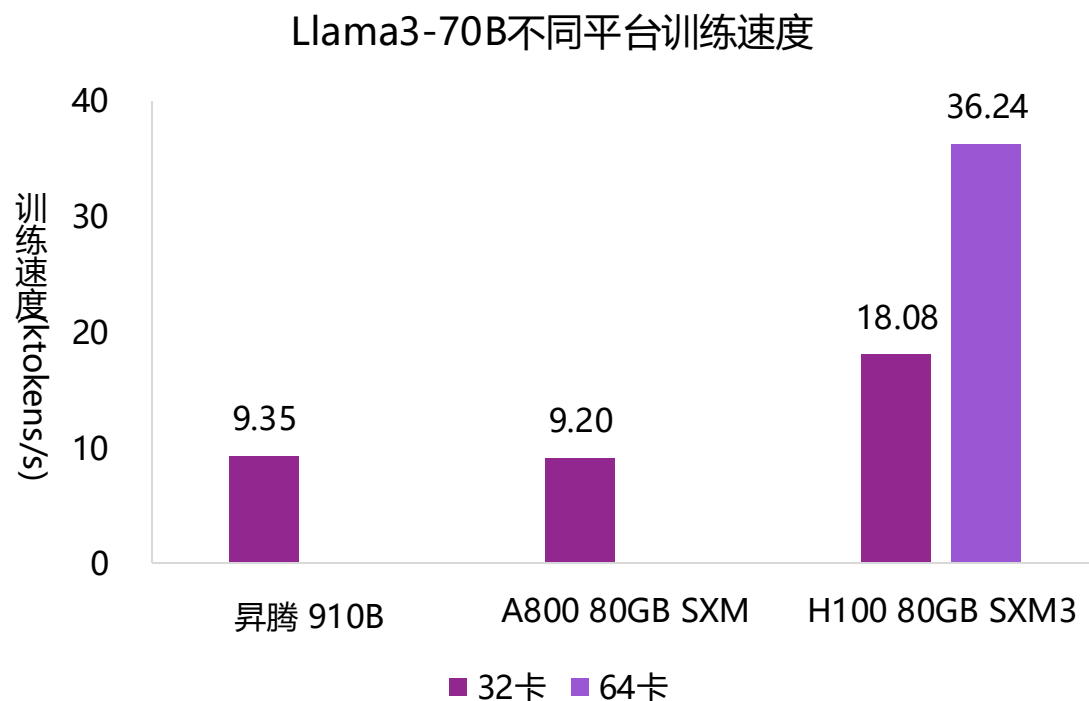


小说配图

- 在昇腾平台部署ComfyUI文生图服务
- 使用ComfyUI为小说生成配图



Llama3-70B训练，昇腾910B性价比可达A800的1.36倍



01

平台精度对齐

用户对昇腾平台训练精度有疑问。通过比较训练loss曲线，证明昇腾精度可与N卡对齐。

02

框架精度对齐

用户对ModelLink训练精度有疑问。通过比较训练loss曲线，证明ModelLink精度可与torch原生对齐。

03

部署推理服务

对于华为已经官方支持的Llama模型，使用MindIE部署；对于华为未支持的Gemma2模型，使用transformers框架部署。

04

切换国产算力

用户正式切换到国产算力，帮助用户进行存储、网络搭建，基础软件环境搭建，数据迁移。

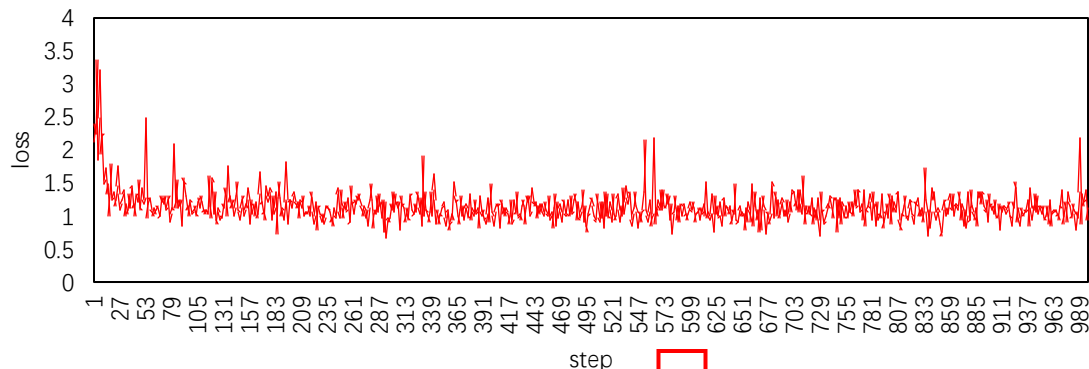




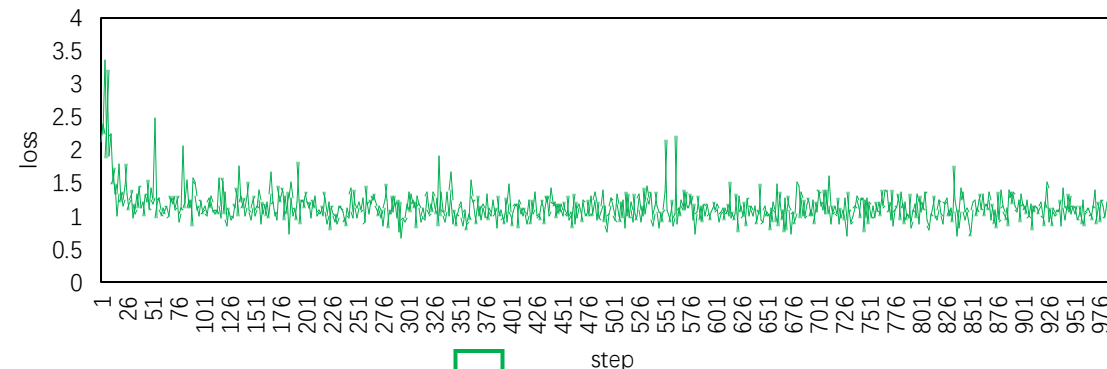
客户模型国产化移植内容

比较训练loss曲线，昇腾精度可与N卡对齐

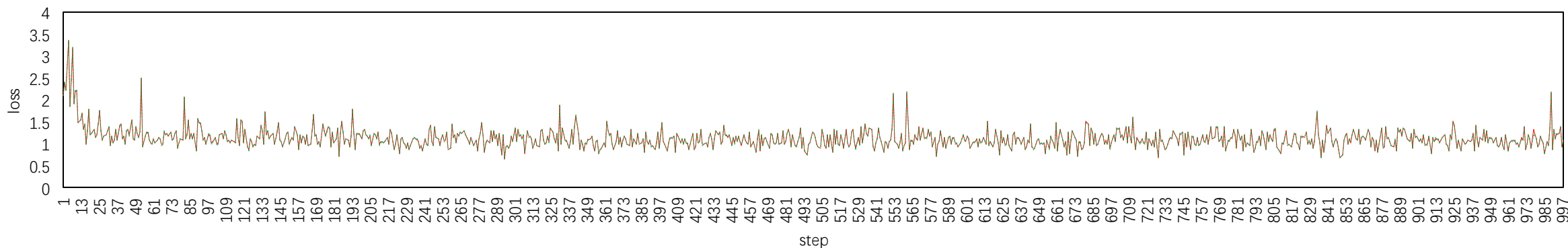
Llama3-8B在Stanford_Alpac数据集上使用LlamaFactory框架基于Zero3分布在910B上训练的loss曲线



Llama3-8B在Stanford_Alpac数据集上使用LlamaFactory框架基于Zero3分布在A100上训练的loss曲线



Llama3-8B在Stanford_Alpac数据集上使用LlamaFactory框架基于Zero3分布在910B和A100上训练的loss曲线



— 910B — A100

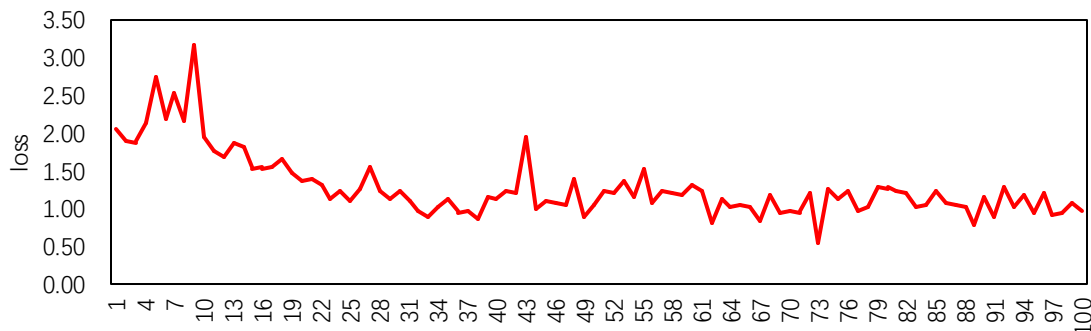




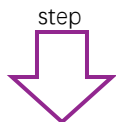
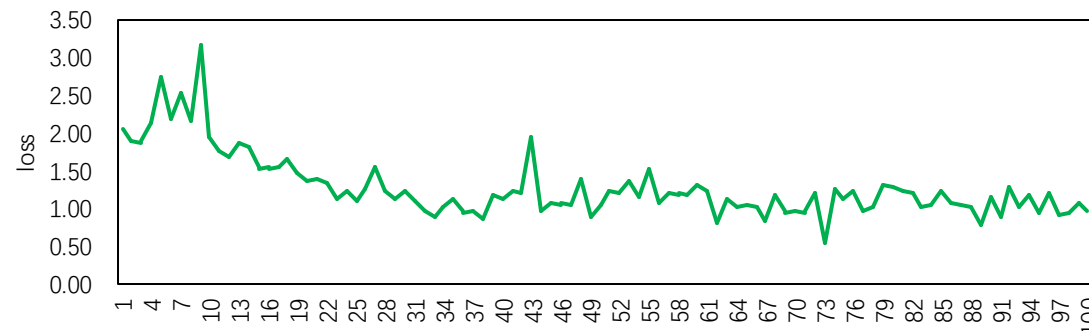
客户模型国产化移植内容

比较训练loss曲线，昇腾优化库可与主流库对齐

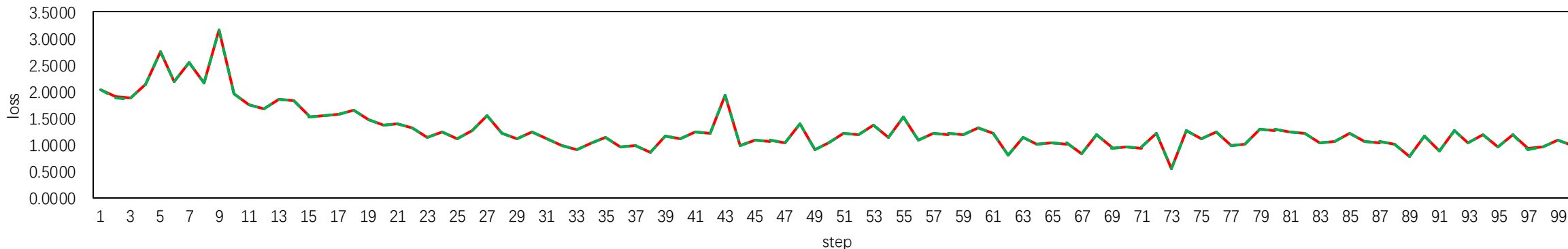
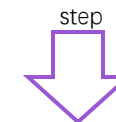
昇腾910B平台ModelLink
在staford_alpaca数据集上训练Llama3-70B的loss曲线



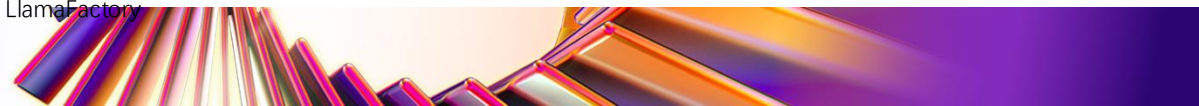
昇腾910B平台LlamaFactory
在staford_alpaca数据集上训练Llama3-70B的loss曲线



昇腾910B平台ModelLink和LlamaFactory在staford_alpaca数据集上训练Llama3-70B的loss曲线对比



— ModelLink — LlamaFactory



PART 04

总结与展望

■ 系统级分析

- ✓ 首先监控系统资源的使用情况，包括CPU、内存、网络、硬盘等，发现并排除系统资源使用瓶颈。

■ MPI通信级分析

- ✓ 分析MPI通信，观测应用程序所有MPI进程的通信函数，统计不同MPI进程间相同函数的比例变化，分析MPI进程负载均衡。

■ 应用级/函数级/微架构级

- ✓ 抽样采集MPI进程的函数信息，获取函数热点，从指令级分析程序指令执行效率、浮点运算效率。



科技生态圈峰会 + 深度研习

——1000+ 技术团队的选择



NiDD 峰会

NiDD 峰会 上海站

AI+研发数字峰会

时间: 2026.05.22-23

NiDD 峰会 北京站

AI+研发数字峰会

时间: 2026.08.21-22

NiDD 峰会 深圳站

AI+研发数字峰会

时间: 2026.11.20-21



AiDD峰会详情

NiDD × K AI+产品峰会

NiDD × K 上海站

AI+产品创新峰会

时间: 2026.01.16-17

NiDD × K 北京站

AI+产品创新峰会

时间: 2026.08.23-24



产品峰会详情



2026 AI+ PRODUCT INNOVATION SUMMIT

01.16-17 · ShangHai

AI+产品创新峰会



Track 1: AI 产品战略与创新设计

从0到1的AI原生产品构建

论坛1: AI时代的用户洞察与需求发现

论坛2: AI原生产品战略与商业模式重构

论坛3: Agentic AI产品创新与交互设计

2-hour Speech: 回归本质



用户洞察的第一性

——2小时思维与方法论工作坊

在数字爆炸、AI迅速发展的时代，
仍然考验“看见”的“同理心”



Track 2: AI 产品开发与工程实践

从1到10的工程化落地实践

论坛1: 面向Agent智能体的产品开发

论坛2: 具身智能与AI硬件产品

论坛3: AI产品出海与本地化开发

Panel 1: 出海前瞻



“出海避坑地图”圆桌对话

——不止于翻译:

AI时代的出海新范式



Track 3: AI 产品运营与智能演化

从10到100的AI产品运营

论坛1: AI赋能产品运营与增长黑客

论坛2: AI产品的数据飞轮与智能演化

论坛3: 行业爆款AI产品案例拆解

Panel 2: 失败复盘



为什么很多AI产品“叫好不叫座”?

——从伪需求到真价值:

AI产品商业化落地的关键挑战

智能重构产品 数据驱动增长
Reinventing Products with Intelligence, Driven by Data

80+

业界大咖

汇聚产品大咖的真知灼见

40+

创新案例

开拓全新AI+产品视角

10+

分论坛

涵盖产品到商业增长的全链路

800+

听众规模

共同探索产品的智能重构



扫码查看会议详情



第8届 AI+ 研发数字峰会
AI+ Development Digital Summit

感谢聆听!

扫码领取会议PPT资料

