STONETEK G5208 DS-R1-Cluster

DeepSeek 大模型推理性能白皮书

文档版本:	V1.1
发布时间:	2025/02/25
测试人员:	Yang Xin
报告撰写:	Yang Xin
报告审核:	Peter Cang
报告批准:	Peter Cang

文档说明

本文档作为 STONETEK G5208 DS-R1-Cluster 性能测试报告，旨在验证产品 DeepSeek-R1 系列大模型推理性能。

测试模型：

信息类别	详细信息
模型名称	DeepSeek-R1 系列
发布时间	2025.01.20
发布方	深度求索

目录

1. 推理参数

(1) 测试环境：vllm serve:0.7.3

(2) 推理参数:

A. 服务端:

a. tensor-parallel-size 8 / pipeline-parallel-size 6 (R1-671b-Premier)

b. tensor-parallel-size 8 / pipeline-parallel-size 4 (R1-671b-lite)

B. 客户端:

a. input:80 token / Max Output: 512 token

2. 性能参数

(1) 测试方法: 通过自定义脚本并发测试。

(2) 参数解释：

A. Conc. (Concurrency) 推理请求并发数；

B. RPS 每秒完成请求数(req/s)；

C. TPS(Tokens Per Second) 推理速度,每秒输出Token数(token/s)；

D. e2e_latency(End-to-End Request Latency)，推理总用时；

E. ITL(Inter-token Latency) 连续输出Token间隔时间；

F. TPOT(Time Per output Token) 每个Token输出的时间，等同于ITL

G. TTFT(Time to First Token) 首个token输出延迟；

3. DS-R1-Cluster 671b Lite 推理性能

(1) 硬件环境(采用4台STONETEK G5208，以下为单台配置):

部件大类	规格	数量
处理器	AMD EPYC 7542 32-Core Processor	2
内存	32GB DDR4 3200	16
系统盘	480G SATA SSD	2
数据盘	3.84T NVME SSD	1
GPU 卡	ASUS-TUF-4090-O24G	8
网卡	NVIDIA- MCX4111A-ACAT-25G Dual port-SFP28	1

(2) 软件环境：

软件	版本信息
操作系统	STONETEK AIOS DS Edition v1.0
NVIDIA 驱动版本	560.35.03
CUDA 版本	V12.4

(3) 推理性能：

Conc.	RPS	TPS	E2E	ITL/TPOT	TTFT（AVG / P50 / P99）
1	0.05	26.19	19.548s	0.038s	0.308s / 0.308s / 0.351s
2	0.09	48.78	21.560s	0.041s	0.494s / 0.498s / 0.758s
4	0.14	70.16	28.744s	0.057s	0.360s / 0.306s / 0.542s
8	0.22	115.92	34.472s	0.069s	0.655s / 0.602s / 1.402s
12	0.25	130.44	47.866s	0.092s	1.164s / 1.318s / 1.974s
16	0.31	172	47.738s	0.093s	1.672s / 1.728s / 2.770s
20	0.34	183.4	50.494s	0.109s	1.370s / 1.404s / 2.523s
24	0.35	192	62.969s	0.125s	1.530s / 1.287s / 2.971s
28	0.4	217.84	62.352s	0.129s	3.082s / 3.086s / 4.678s
32	0.43	231.68	69.040s	0.138s	2.425s / 2.235s / 4.557s

4. DS-R1-Cluster 671b Premier 推理性能

(1) 硬件环境(采用6台STONETEK G5208，以下为单台配置):

部件大类	规格	数量
处理器	AMD EPYC 7542 32-Core Processor	2
内存	32GB DDR4 3200	16
系统盘	480G SATA SSD	2
数据盘	3.84T NVME SSD	1
GPU 卡	ASUS-TUF-4090-O24G	8
网卡	NVIDIA- MCX4111A-ACAT-25G Dual port-SFP28	1

(2) 软件环境：

软件	版本信息
操作系统	STONETEK AIOS DS Edition v1.0
NVIDIA 驱动版本	560.35.03
CUDA 版本	V12.4

(3) 推理性能：

Conc.	RPS	TPS	E2E	ITL/TPOT	TTFT（AVG / P50 / P99）
1	0.03	16.53	31.352s	0.060s	0.438s / 0.438s / 0.454s
2	0.06	32.86	31.691s	0.061s	0.596s / 0.577s / 0.802s
4	0.12	62.88	33.594s	0.064s	1.076s / 0.720s / 2.357s
6	0.15	82.02	35.318s	0.075s	0.605s / 0.620s / 0.839s
8	0.18	93.84	42.539s	0.085s	0.605s / 0.603s / 0.832s
12	0.26	137.04	44.141s	0.088s	0.633s / 0.525s / 0.924s
14	0.3	152.6	42.567s	0.092s	0.655s / 0.672s / 0.940s
16	0.32	171.52	43.835s	0.093s	0.685s / 0.747s / 0.942s
18	0.37	191.34	45.734s	0.094s	0.666s / 0.720s / 0.956s
20	0.42	215.8	47.012s	0.093s	0.684s / 0.645s / 1.000s
24	0.49	252.72	45.430s	0.095s	0.769s / 0.819s / 1.046s
28	0.49	254.8	53.464s	0.110s	0.728s / 0.779s / 0.974s
32	0.51	267.52	57.653s	0.120s	0.737s / 0.653s / 1.059s

5. 推理质量对比

(1) 评测环境：

软件	版本信息
评测工具	EvalScope
数据集	gsm8k / humaneval / trivia_qa

(2) 评测对比：

Cluster	Dataset	Metric	Subset	Num	Score
Lite	gsm8k	AverageAccuracy	main	20	0.95
Lite	humaneval	Pass@1	openai_humaneval	20	0.7
Lite	trivia_qa	AverageAccuracy	default	20	0.85
Premier	gsm8k	AverageAccuracy	main	20	1
Premier	humaneval	Pass@1	openai_humaneval	20	0.8
Premier	trivia_qa	AverageAccuracy	default	20	0.85

注: 数据集分别为数学、编码、知识问答场景，得分越高表现越佳。

6. 结论

(1) DS-R1-Cluster 671b Lite：

搭载 DeepSeek-R1-AWQ 大模型，使用四个节点进行推理。在 1 个并发下 TPS 可达 26.19token/s，在 16 个并发时 TPS 依然有 10token/s。

(2) DS-R1-Cluster 671b Premier:

搭载 DeepSeek-R1 大模型满血未量化版本，使用六个节点进行推理。在 4 个并发内可达 16token/s，在 24 个并发时 TPS 依然有 10token/s。

(3) 场景推荐:

DS-R1-Cluster 系列可用于国家级科研机构进行前沿科学研究，大型企业进行复杂的商业决策、数据分析等。在准确性要求较高的场景，推荐使用 DS-R1-Cluster 671b Premier。