EMPOWER INTELLIGENT COMPUTING

STONETEK G5208 DS-R1-AiO

DeepSeek 大模型推理性能白皮书

文档版本:	V1.2
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-AiO 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： SGLang serve:0.4.3

(2) 推理参数:

A. 服务端:

a. tensor-parallel-size 8 (R1-70b-premier)

b. tensor-parallel-size 4 (R1-70b-lite / R1-32b)

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-AiO 32b Pro

(1) 硬件环境:

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	4
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	0.09	42.59	11.385s	0.023s	0.195s / 0.193s / 0.247s
2	0.16	79.42	12.879s	0.025s	0.189s / 0.188s / 0.202s
4	0.31	138.85	11.569s	0.025s	0.180s / 0.161s / 0.260s
8	0.58	288.87	13.518s	0.027s	0.226s / 0.239s / 0.263s
16	1.06	527.4	14.789s	0.029s	0.234s / 0.235s / 0.274s
32	1.63	814.3	19.141s	0.038s	0.285s / 0.307s / 0.317s
64	2.88	1407.58	21.232s	0.043s	0.294s / 0.302s / 0.343s
72	3.1	1478.62	21.855s	0.045s	0.315s / 0.357s / 0.386s
80	3.32	1622.87	23.138s	0.047s	0.327s / 0.300s / 0.628s
85	3.16	1555.14	26.030s	0.052s	0.341s / 0.332s / 0.431s
90	3.24	1588.77	26.284s	0.053s	0.344s / 0.341s / 3.745s
96	3.25	1604.72	27.172s	0.054s	0.776s / 0.374s / 11.139s
128	3.49	1693.03	31.819s	0.059s	3.085s / 0.382s / 30.668s

注: DS-R1-AiO 32b Pro在80并发左右时体验较好；

4. DS-R1-AiO 32b Premier 推理性能

(1) 硬件环境:

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）
2	0.17	84.79	12.048s	0.023s	0.163s / 0.136s / 0.237s
4	0.31	146.47	11.944s	0.025s	0.177s / 0.179s / 0.210s
8	0.59	274.87	12.353s	0.026s	0.181s / 0.163s / 0.250s
16	1.06	525.2	14.249s	0.028s	0.207s / 0.231s / 0.275s
32	1.85	900.72	15.801s	0.032s	0.232s / 0.204s / 0.304s
64	3.07	1491.04	19.627s	0.040s	0.261s / 0.278s / 0.366s
96	4.27	2080.22	21.197s	0.043s	0.315s / 0.321s / 0.424s
128	5.24	2546.47	22.853s	0.046s	0.356s / 0.313s / 1.315s
144	5.56	2706.14	24.291s	0.049s	0.334s / 0.307s / 1.307s
160	5.84	2841.65	25.488s	0.052s	0.380s / 0.324s / 1.361s
180	5.88	2871.62	27.158s	0.055s	0.480s / 0.376s / 3.279s
192	6.41	3113.35	27.317s	0.055s	0.750s / 0.383s / 9.816s
256	6.27	3060.38	33.698s	0.062s	3.556s / 0.417s / 32.110s

注: DS-R1-AiO 32b Premier在160并发左右时体验较好；

5. DS-R1-AiO 70b Lite 推理性能

(1) 硬件环境:

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	4
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	0.07	35.29	14.338s	0.028s	0.214s / 0.190s / 0.306s
2	0.13	67.5	14.947s	0.029s	0.302s / 0.271s / 0.601s
4	0.26	130.38	15.503s	0.030s	0.309s / 0.344s / 0.421s
8	0.48	242.45	16.706s	0.032s	0.327s / 0.337s / 0.425s
16	0.83	420.19	19.251s	0.037s	0.367s / 0.351s / 0.462s
32	1.39	694.1	22.784s	0.045s	0.372s / 0.360s / 0.501s
48	1.87	936.93	25.492s	0.050s	0.478s / 0.379s / 0.976s
64	2.25	1134.13	28.365s	0.055s	0.435s / 0.417s / 0.615s

注: DS-R1-AiO 70b Lite在32并发左右时体验较好，70b Lite限制最大并发64；

6. DS-R1-AiO 70b Premier 推理性能

(1) 硬件环境:

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	0.08	38.15	13.260s	0.026s	0.217s / 0.217s / 0.270s
2	0.14	70.96	14.183s	0.028s	0.218s / 0.217s / 0.224s
4	0.27	137.71	14.633s	0.029s	0.259s / 0.262s / 0.318s
8	0.5	254.86	15.917s	0.031s	0.238s / 0.191s / 0.303s
16	0.83	413.01	18.934s	0.037s	0.323s / 0.321s / 0.350s
32	1.23	615.58	25.840s	0.051s	0.348s / 0.335s / 0.384s
64	2.1	1064.45	30.404s	0.059s	0.408s / 0.405s / 0.436s
96	2.27	1147.7	42.286s	0.083s	0.607s / 0.576s / 0.682s

注: DS-R1-AiO 70b Premier在32并发左右时体验较好；

7. 结论

(1) DS-R1-AiO 32b Pro：

搭载 DeepSeek-R1-Distill-Qwen-32B 大模型，在 80 个并发下 TPS 能达到 1623 Token/s，平均 TPOT 低至 0.047s，平均 TTFT 0.372s。

推荐场景：教学辅助、代码评审以及需要高并发问答的场景

(2) DS-R1-AiO 32b Premier:

搭载 DeepSeek-R1-Distill-Qwen-32B 大模型，在 160 个并发下 TPS 能达到 2842 Token/s，平均 TPOT 低至 0.052s，平均 TTFT 0.380s。

推荐场景：教学辅助、代码评审以及需要高并发问答的场景

(3) DS-R1-AiO 70b Lite:

搭载 DeepSeek-R1-Distill-Llama-70B-FP8-dynamic 大模型，实现在更低的硬件要求下运行 70b 大模型。在 32 个并发下 TPS 能达到 694 Token/s，平均 TPOT 低至 0.045s，平均 TTFT 0.372s。

推荐场景：法律、医疗、金融等专业领域，提供高质量的文本生成、知识检索和决策支持。

(4) DS-R1-AiO 70b Premier:

搭载 DeepSeek-R1-Distill-Llama-70B-大模型。在 32 个并发下 TPS 能达到 616 Token/s，平均 TPOT 低至 0.051s，平均 TTFT 0.348s。

推荐场景：法律、医疗、金融等专业领域，提供高质量的文本生成、知识检索和决策支持。

(5) 选择推荐:

A. 32b 系列适合多用户，高并发的场景;
B. 70b 系列适合对高并发需求较低，对专业度要求较高的场景;