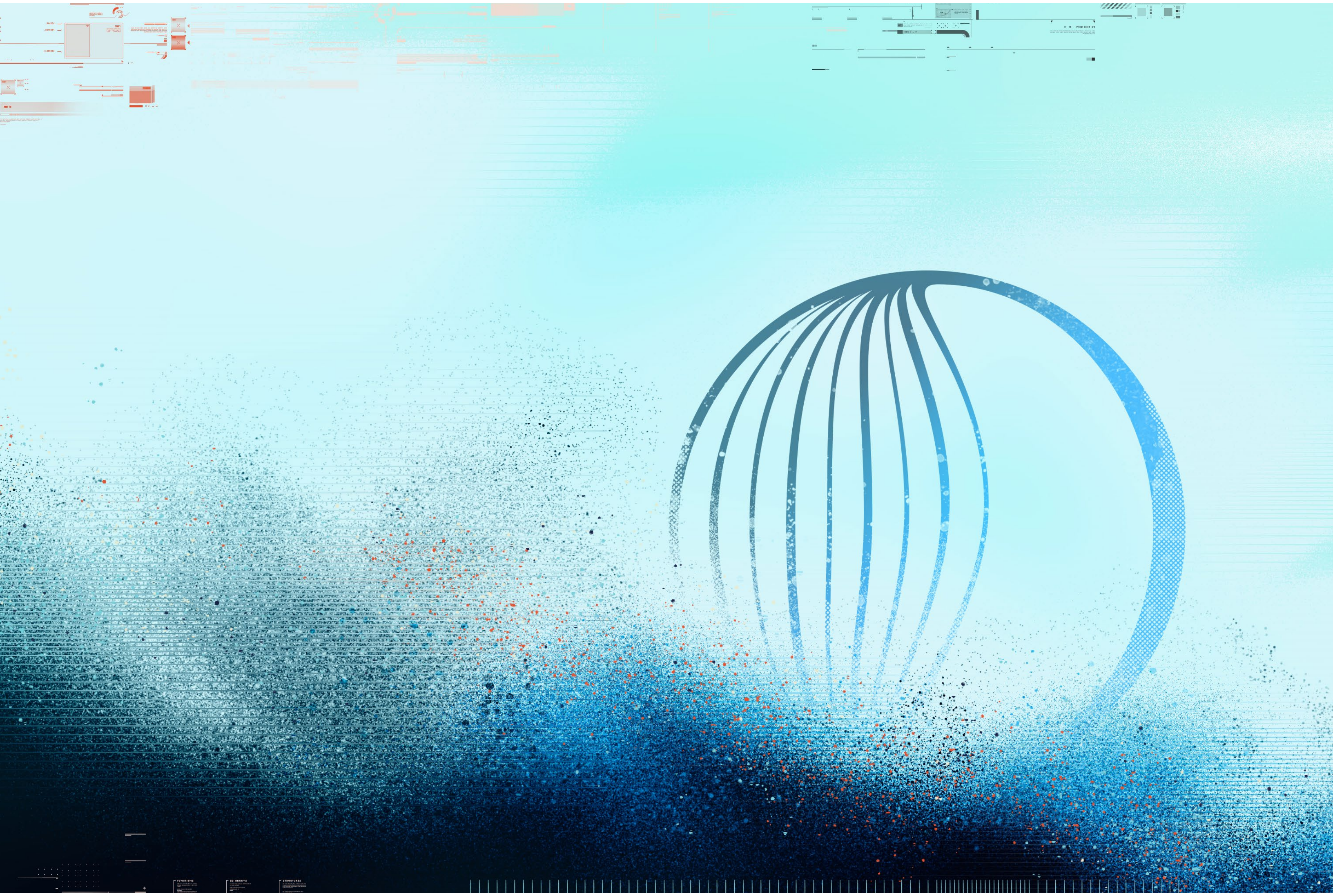
signal65



AI On-Premises: A Look at OpenAI GPT-OSS-120B

AUTHOR

Brian Martin
AI and Data Center Lead | Signal65

IN PARTNERSHIP WITH

DELL Technologies

SEPTEMBER 2025

Introduction

gpt-oss-120b

OpenAI's release of gpt-oss-120b marks a pivotal moment in the democratization of large language model deployment. As organizations increasingly seek to maintain data sovereignty, the availability of a 120-billion parameter model optimized for on-premises deployment addresses a critical gap in the enterprise AI landscape. This model represents not just a technical achievement, but a strategic shift toward enabling organizations to leverage powerful AI capabilities within their own infrastructure boundaries with affordable hardware solutions.

The importance of this model extends beyond mere availability. For industries handling sensitive data like healthcare, finance, or government, the ability to run sophisticated language models locally means maintaining complete control over proprietary information while still benefiting from state-of-the-art natural language processing capabilities. This model bridges the gap between the desire for advanced AI capabilities and the necessity of data governance and regulatory compliance.

Additionally, as a Mixture-of-Experts (MoE) model, it performs very well on single GPUs with 80GB or more memory, so it scales linearly as GPUs are added. This allows organizations to grow as they need, in affordable, incremental steps. The models offers reasoning with Chain-of-Thought (CoT) and reduces overthinking common in many other current open models.

Signal65 tested gpt-oss-120b on NVIDIA H200, AMD MI300X, and NVIDIA RTX Pro 6000 GPUs. The model runs well across all three accelerators, generating impressive token rates up to 64 or 128 (H200) simultaneous requests on a single GPU. Batch sizes were increased until per thread TPS dropped below 20.

H200 Performance was measured across two workload shapes: 2048 input tokens and 128 output tokens ("summarize this post") and 2048 input tokens and 2048 output tokens ("update this document"). A single H200 consistently returned over 21,000 tokens/sec for "summarize this post" and over 7,100 tokens/sec for "update this document".

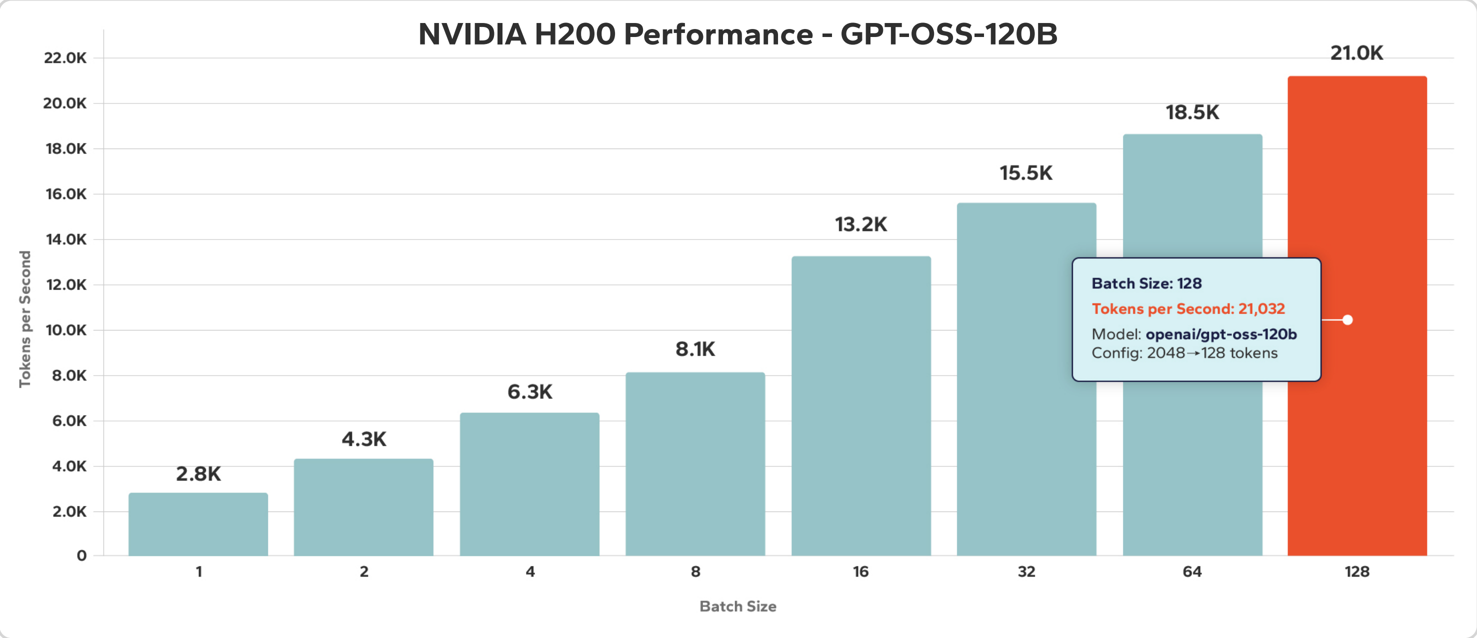


Figure 1: *tokens/sec for "summarize this post" workload*

Batch	Response TPS	TTFT	Total Output TPS	Total (In/Out) TPS
1	190	0.08s	190	2,764
2	154	0.12s	308	4,280
4	120	0.20s	480	6,325
8	95	0.40s	760	8,076
16	73	0.52s	1168	13,152
32	51	0.93s	1632	15,537
64	35	1.71s	2240	18,542
128	21	3.16s	2668	21,032

3.95
Output TPS/Watt

30.9
Total TPS/Watt

680W GPU power at batch size 128

Drawing 680W average across both workloads, the H200 delivers approximately 4-6 output tokens per second across workloads. Somewhat surprisingly, returning more response tokens proves to be more power efficient that fewer response tokens.

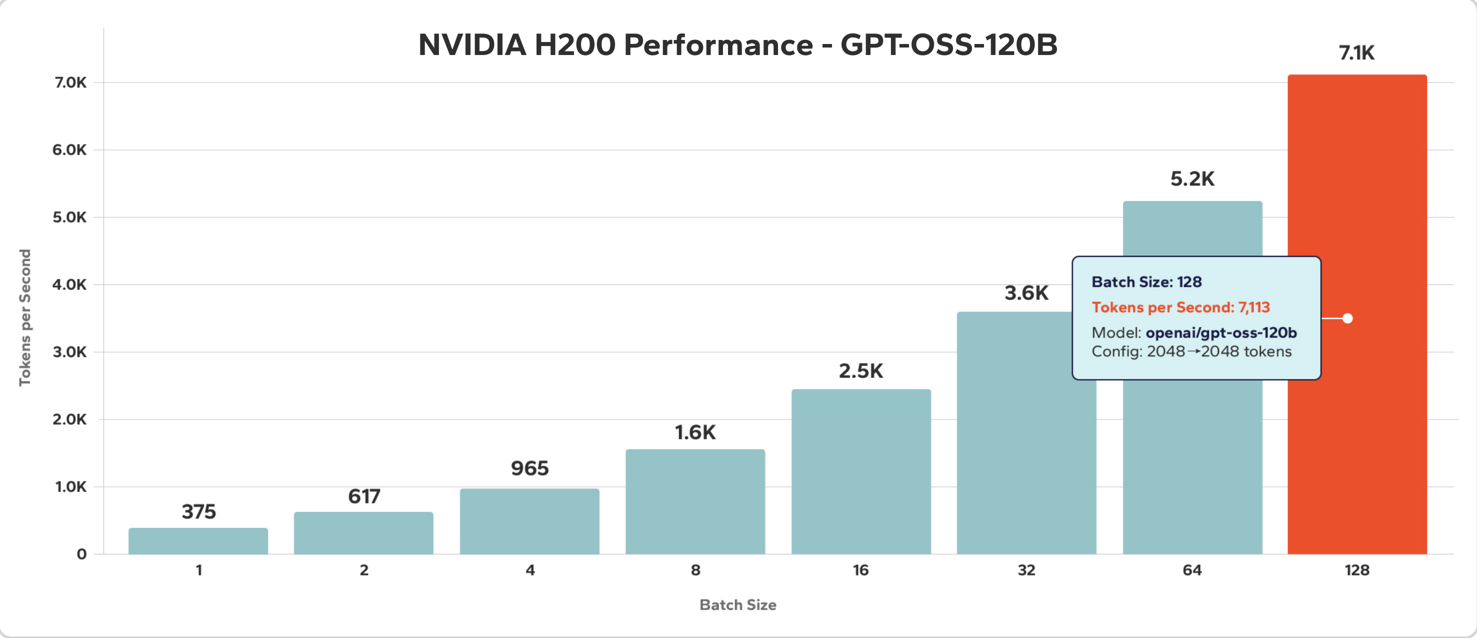


Figure 2: *tokens/sec for "update this document" workload*

Batch	Response TPS	TTFT	Total Output TPS	Total (In/Out) TPS
1	188	0.10s	188	375
2	155	0.13s	310	617
4	122	0.22s	488	965
8	99	0.33s	792	1550
16	79	0.52s	1264	2452
32	59	0.89s	1888	3604
64	44	1.75s	2816	5244
128	31	3.48s	3968	7113

5.84
Output TPS/Watt

10.5
Total TPS/Watt

680W GPU power at batch size 128

MI300X performance was also measured across two workload shapes: 2048 input tokens and 128 output tokens (“summarize this post”) and 2048 input tokens and 2048 output tokens (“update this document”). A single MI300X returned approximately 12,500 tokens/sec for “summarize this post” and over 3,100 tokens/sec for “update this document” at a batch size of 64.

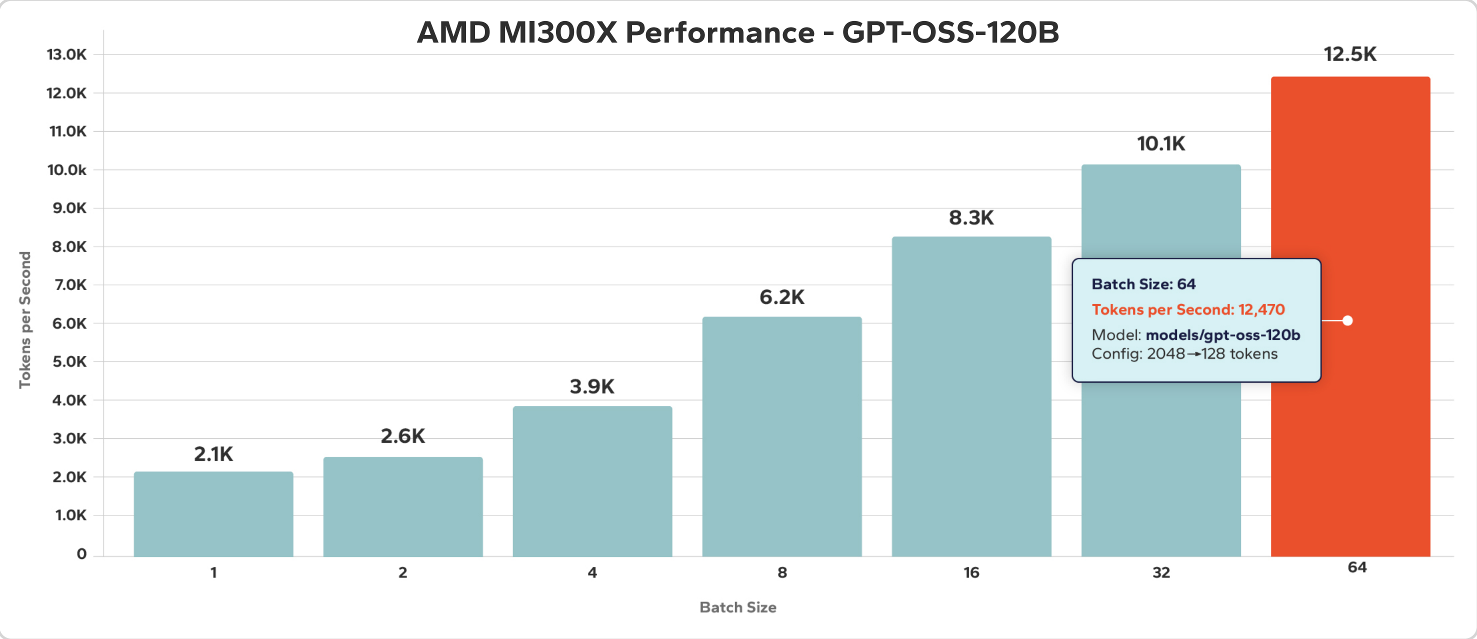


Figure 3: *tokens/sec for “summarize this post” workload*

Batch	Response TPS	TTFT	Total Output TPS	Total (In/Out) TPS
1	146	0.12s	146	2,128
2	89	0.18s	178	2,565
4	70	0.29s	280	3,857
8	61	0.47s	488	6,196
16	43	0.79s	688	8,255
32	30	1.32s	960	10,135
64	21	2.51s	1344	12,470

1.87
Output TPS/Watt

17.3
Total TPS/Watt

720W GPU power at batch size 64

At 720 Watts average across both workloads, the MI300X draws the highest power of the group, eking out slightly more at higher response tokens as well.

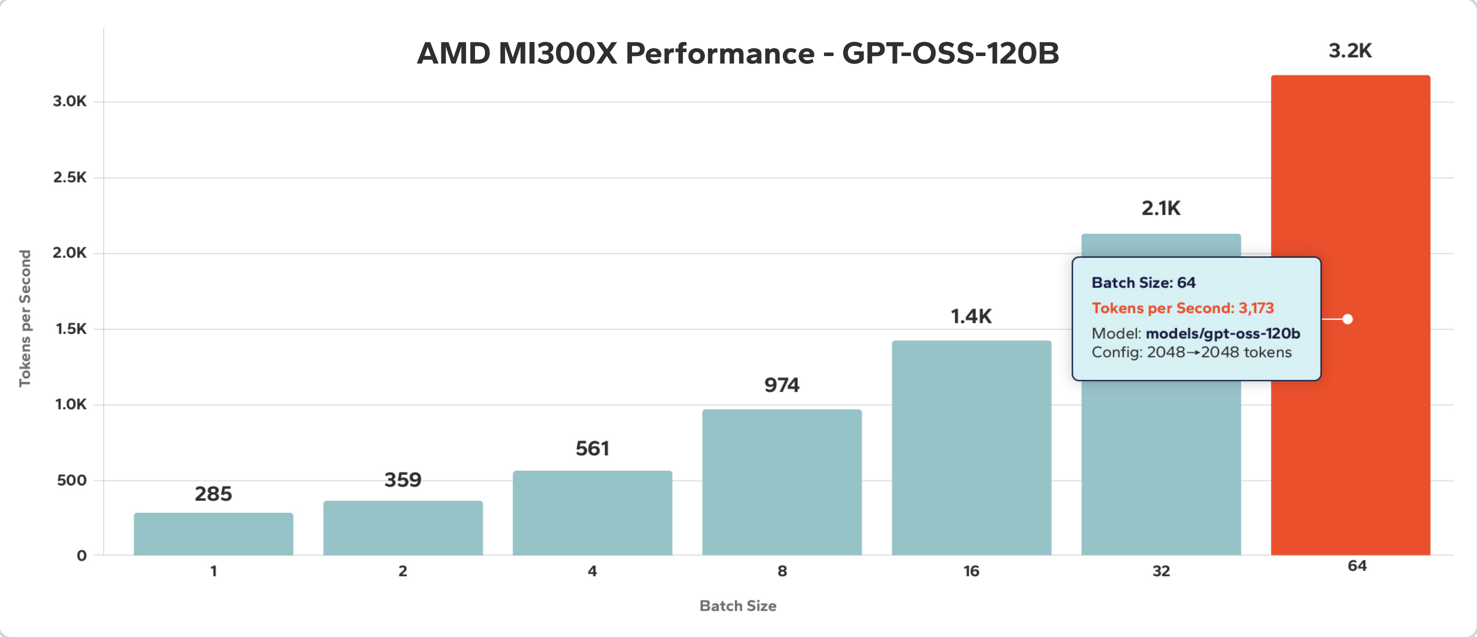


***Figure 4:** tokens/sec for "update this document" workload*

Batch	Response TPS	TTFT	Total Output TPS	Total (In/Out) TPS
1	145	0.12s	145	285
2	90	0.17s	180	359
4	71	0.30s	284	561
8	62	0.47s	492	974
16	45	0.78s	720	1421
32	35	1.35s	1120	2139
64	26	2.78s	1664	3173

2.31
Output TPS/Watt

4.4
Total TPS/Watt

720W GPU power at batch size 64

RTX Pro 6000 performance was similarly measured across the same “summarize this post” and “update this document” workload shapes. A single RTX Pro 6000 returned over 16,500 tokens/sec for “summarize this post” and over 4,900 tokens/sec for “update this document” at a batch size of 64.

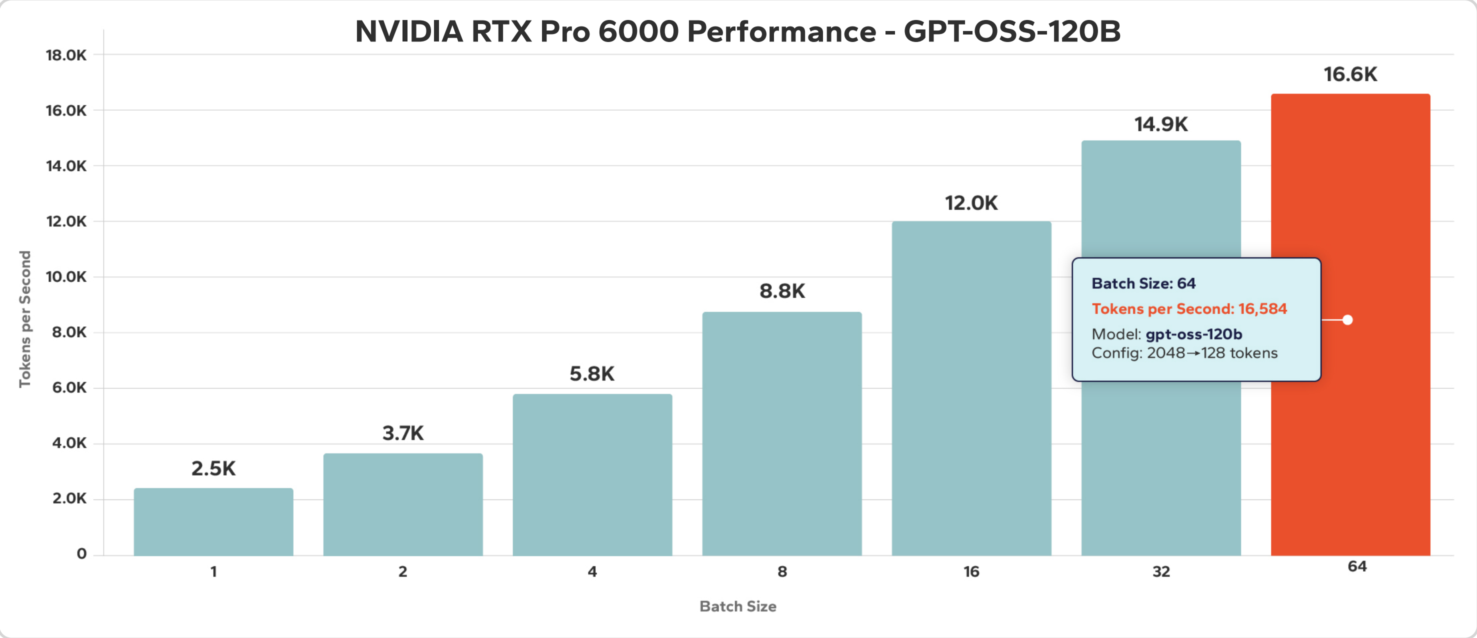


Figure 5: *tokens/sec for “summarize this post” workload*

Batch	Response TPS	TTFT	Total Output TPS	Total (In/Out) TPS
1	163	0.10s	163	2,454
2	129	0.18s	158	3,699
4	108	0.29s	432	5,807
8	92	0.47s	736	8,756
16	67	0.73s	1072	11,958
32	47	1.30s	1504	14,922
64	29	2.61s	1856	16,584

3.87
Output TPS/Watt

34.6
Total TPS/Watt

480W GPU power at batch size 64

The RTX Pro 6000 demonstrated very strong performance, consistently competitive with the H200 at 4-6 response tokens per second, and leading by 11% in total tokens per second per Watt in post summarization.

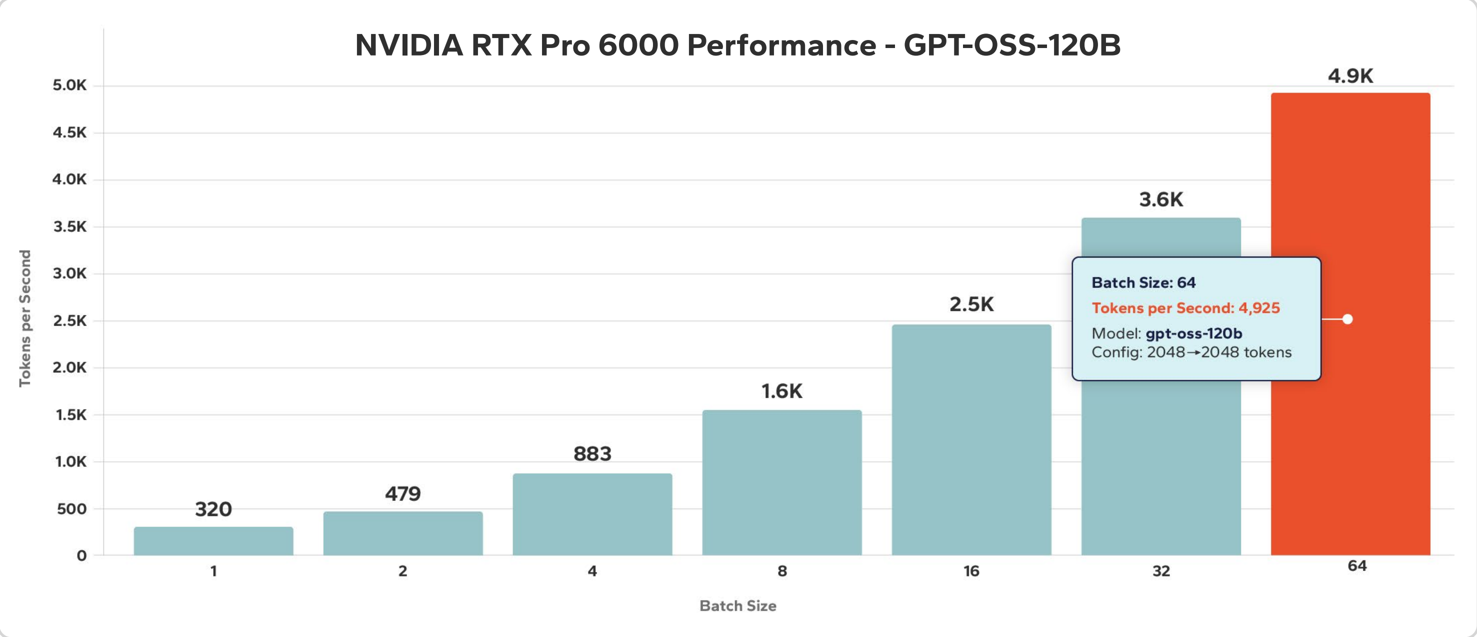


Figure 6: *tokens/sec for "update this document" workload*

Batch	Response TPS	TTFT	Total Output TPS	Total (In/Out) TPS
1	160	0.11s	160	320
2	137	0.15s	274	479
4	112	0.30s	448	883
8	99	0.46s	792	1552
16	80	0.73s	1280	2471
32	60	1.30s	1920	3605
64	42	2.59s	2688	4925

5.6
Output TPS/Watt

10.3
Total TPS/Watt

480W GPU power at batch size 64

Summary

Examining output and total tokens per second per Watt for the two workloads we see solid performance from the H200, especially at larger token sizes. The RTX Pro 6000 does exceptionally well summarizing posts, even at batch size 64 vs 128 on the H200. The MI300X performance was potentially limited by the software stack. Expect these numbers to get even better as MXFP8 support improves over time.

GPU	Workload	Output TPS/Watt	Total TPS/Watt
H200	Summarize Post	3.95	30.9
MI300X	Summarize Post	1.87	17.3
RTX Pro 6000	Summarize Post	3.87	34.6
H200	Update Document	5.84	10.5
MI300X	Update Document	2.32	4.4
RTX Pro 6000	Update Document	5.6	10.3

OpenAI's gpt-oss-120b represents more than just another open model, it is validation that on-premises AI is alive and well and growing. The performance metrics on Dell XE9680 servers with H200 or MI300X as well as Dell XE7745 with RTX Pro 6000 demonstrate organizations no longer need to choose between capability and control. With thoughtful optimization and appropriate hardware, enterprises can deploy sophisticated language models that rival cloud-based solutions while maintaining complete ownership of their data and infrastructure.

Looking ahead, success with gpt-oss-120b will encourage further innovation in open, locally deployable models. The path forward is clear: increasingly more accessible on-premises AI for organizations of all sizes. The era of democratized, high-performance language models is here.

Important Information About this Report

CONTRIBUTORS

Brian Martin

AI and Data Center Lead | Signal65

PUBLISHER

Ryan Shrout

President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

IN PARTNERSHIP WITH

DELL Technologies

ABOUT SIGNAL65

Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

signal65

signal65

CONTACT INFORMATION

Signal65 I signal65.com