

TandemSpec: Total Cost of Ownership Analysis

Appendix to the TandemSpec paper (October 2026)

PSALab
POSTECH
Pohang, Republic of Korea

A Total Cost of Ownership Analysis

In this appendix, we estimate the total cost of ownership (TCO) of TandemSpec and the baselines of §4. We report the cost per 1M output tokens, which divides the hourly cost of a server by the number of output tokens per hour. All throughput values are measured. We estimate the hourly cost from the price and the power of each component (Tables 2 and 3).

Cost Metric. Table 1 lists the abbreviations and symbols of this appendix.

Term	Meaning
T_{out}	Output token throughput of a system (tok/s)
C_{hour}	Hourly cost of the server that runs the system (\$/h)
N_{ref}	Output tokens of TandemSpec for a request; used for every system
CapEx	Capital expenditure: purchase price of GPUs, CPU, DDR5 DIMMs, and server chassis (\$)
f	Monthly amortization factor of CapEx (5 years, 13% per year): 0.022753
P	Average power of the server (kW)
PUE	Power usage effectiveness: facility power divided by server power
c_e	Electricity cost (\$/kWh)
c_{colo}	Colocation price for rack space, cooling, and power delivery (\$/kW per month)
730	Hours per month

Table 1. Abbreviations and symbols.

We denote by T_{out} the throughput of a system in output tokens per second and by C_{hour} the hourly cost of the server that runs the system, in dollars. Then

$$\text{tokens per dollar} = \frac{3600 T_{\text{out}}}{C_{\text{hour}}}, \quad (1)$$

$$\text{cost per 1M tokens} = \frac{10^6 C_{\text{hour}}}{3600 T_{\text{out}}}. \quad (2)$$

For a single request, $T_{\text{out}} = N_{\text{ref}} / (\text{TTFT} + N_{\text{ref}} \cdot \text{TPOT})$, where N_{ref} is the number of output tokens of TandemSpec for the request. We use the same N_{ref} for every system so that all systems generate the same number of output tokens. For batch serving, T_{out} is the number of output tokens divided by the wall-clock time of the run.

Hourly Cost. We follow the ownership model of Semi-Analysis [10]. The hourly cost is

$$C_{\text{hour}} = \frac{f \cdot \text{CapEx} + P (\text{PUE} \cdot 730 \cdot c_e + c_{\text{colo}})}{730}, \quad (3)$$

where $f = 0.022753$ amortizes the CapEx over five years with a rate of 13% per year, P is the average power of the server in kW, c_e is the electricity cost per kWh, and c_{colo} is the price per kW per month for colocation. We set the maintenance cost to zero and the utilization to 100% for every configuration.

Every configuration is one server with one CPU socket (Table 4). TandemSpec uses one GPU and stores the full KV cache in host memory. When the KV cache exceeds the memory of one GPU, vLLM uses tensor parallelism (TP), so a TP= n configuration has n GPUs in the same server. The instance of §4 exposes 48 of the 64 cores and 858 GiB of DRAM. We price the full socket with 512 GB of DDR5, which holds the model weights and the full KV cache of the longest request, at most 178 GiB for Granite-4.2-30B. For the CPU power, we use the TDP for TandemSpec and NEO, which perform attention on the CPU, and 30% of the TDP for the other systems (Table 3).

Component	Price	Source
NVIDIA L40S 48 GB	\$13,879 per GPU	[5]
NVIDIA RTX PRO 6000 96 GB	\$17,850 per GPU	[4]
Intel Xeon Gold 6530, 32 cores	\$2,128 per socket	[12]
Intel Xeon 6776P, 64 cores	\$7,764 per socket	[3]
DDR5 RDIMM 64 GB, Q4 2025	\$450 per DIMM	[2]
Server chassis (4U GPU barebone)	\$9,527 per server	[7]

Table 2. Component prices. GPU, CPU, and chassis prices are listed on 2026-10-21.

KV Cache Capacity. Table 5 lists the maximum context length of vLLM with a batch size of 1. We derive the remaining values by dividing the memory for the KV cache on each GPU by the KV cache size per token. This memory is the remaining memory after subtracting the model weights and the activation memory from 90% of the GPU memory. We calibrate the activation memory with the measured TP=1 value. For Qwen3-30B-A3B, TP=8 provides little more capacity than TP=4 because the model has four KV heads, so two GPUs store each KV head.

Long Context on L40S. This experiment is not part of §4; it runs Qwen3-4B-Thinking-2507 on a separate L40S server.

Device	Power	Basis
GPU	TDP: L40S 350 W, RTX PRO 6000 600 W	NVIDIA specification
CPU, active	TDP: Gold 6530 270 W, 6776P 350 W	Attention runs on the CPU
CPU, idle	30% of TDP: 81 W, 105 W	ServeTheHome: idle package power is 22% of TDP [11]
DDR5	10 W per DIMM	Micron: PMIC limit 15.3 W [6]
Server base	100 W	NVIDIA: NIC 11 W [8]; Samsung: SSD 13.5 W [9]
Parameter	Value	Source
f	5 years, 13% per year: 0.022753	SemiAnalysis [10]
PUE	1.3	SemiAnalysis [10]
c_e	\$0.0977/kWh	EIA, U.S. industrial average, July 2026 [13]
c_{colo}	\$194.95/kW per month	CBRE, North American average, 2025 [1]

Table 3. Power of each device and cost parameters.

	TS	TP=1	TP=2	TP=4
<i>L40S server: Gold 6530, DDR5 64 GB × 4</i>				
GPUs	1	1	2	4
CapEx (\$)	27,334	27,334	41,213	68,971
$f \cdot \text{CapEx}$ (\$/month)	622	622	938	1,569
P (W)	760	571	921	1,621
Electricity (\$/month)	70	53	85	150
Colocation (\$/month)	148	111	180	316
C_{hour} (\$/h)	1.151	1.077	1.647	2.789
<i>RTX PRO 6000 server: 6776P, DDR5 64 GB × 8</i>				
GPUs	1	1	2	4
CapEx (\$)	38,741	38,741	56,591	92,291
$f \cdot \text{CapEx}$ (\$/month)	882	882	1,288	2,100
P (W)	1,130	885	1,485	2,685
Electricity (\$/month)	105	82	138	249
Colocation (\$/month)	220	172	290	523
C_{hour} (\$/h)	1.653	1.556	2.349	3.935

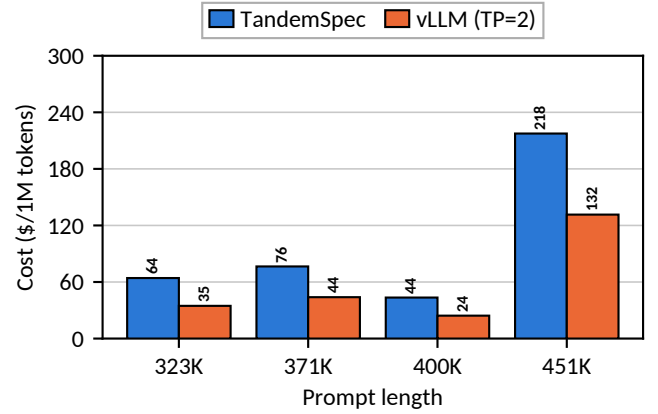
Table 4. Hourly cost C_{hour} of each configuration. TS is TandemSpec; TP= n is vLLM with n GPUs in the same server.

Model (GPU)	KiB/token	TP=1	TP=2	TP=4	TP=8
Qwen3-4B (L40S)	144	225*	515*	1,076	2,209
Qwen3-30B-A3B (RTX PRO 6000)	96	292*	1,206	3,034	3,345
Granite-4.2-30B (RTX PRO 6000)	256	112*	446	1,116	2,456

Table 5. Maximum context length of vLLM (K tokens, batch size 1). *Measured; others are derived.

We use the four LongBench-v2 prompts of 323K–451K tokens that do not fit in one L40S under vLLM, which correspond to the Ultra-long bucket of \$4. TandemSpec handles these requests on one GPU, whereas vLLM requires TP=2, and

SparseSpec does not support TP=2 for this model. The cost of TandemSpec is 1.7–1.9× that of vLLM with TP=2 (Table 6 and Figure 1).

**Figure 1.** Cost per 1M tokens on L40S (Qwen3-4B-Thinking-2507, Ultra-long prompts of 323K–451K tokens, batch size 1).

Prompt	System	TTFT (s)	TPOT (ms)	N_{ref}	T_{out} (tok/s)	\$/1M tokens
323K	TandemSpec	182.3	124.5	2,391	4.98	64.2
	vLLM, TP=2	85.6	39.9	2,391	13.20	34.7
371K	TandemSpec	236.7	131.6	2,194	4.18	76.6
	vLLM, TP=2	112.7	44.6	2,194	10.41	43.9
400K [†]	TandemSpec	272.4	132.8	83,230	7.35	43.5
	vLLM, TP=2	130.1	51.5	83,230	18.84	24.3
451K	TandemSpec	347.3	179.8	694	1.47	217.5
	vLLM, TP=2	162.8	52.5	694	3.48	131.4

Table 6. Ultra-long-context cost on L40S (Qwen3-4B-Thinking-2507, 323K–451K tokens, batch size 1). C_{hour} is \$1.151/h for TandemSpec (1 GPU) and \$1.647/h for vLLM (TP=2). [†]Both systems generate until the output limit.

Long Context on RTX PRO 6000. Table 7 and Figure 2 compare the five systems on one GPU. In the Long buckets, TandemSpec routes all requests to the GPU-resident path and achieves the same TPOT as vLLM, so the two systems differ in cost only by the CPU power. In the Ultra buckets, vLLM and SparseSpec run out of GPU memory. Compared to NEO, TandemSpec reduces the cost by 3.1× and 1.9× for Qwen3-30B-A3B and Granite-4.2-30B, respectively. Compared to HeadInfer, it reduces the cost by 10.4× and 8.4×, respectively.

Short Context, Batched Serving. On AIME25 with concurrency limits of 8 and 16, the cost of TandemSpec is 7.1% and 4.7% higher than that of vLLM, respectively (Table 8 and Figure 3).

Limitations. For the power, we use the TDP as an upper bound instead of measured power. We estimate the DDR5 power and the base power of the server. The GPU and CPU prices are listed on October 21, 2026. For DDR5, we use

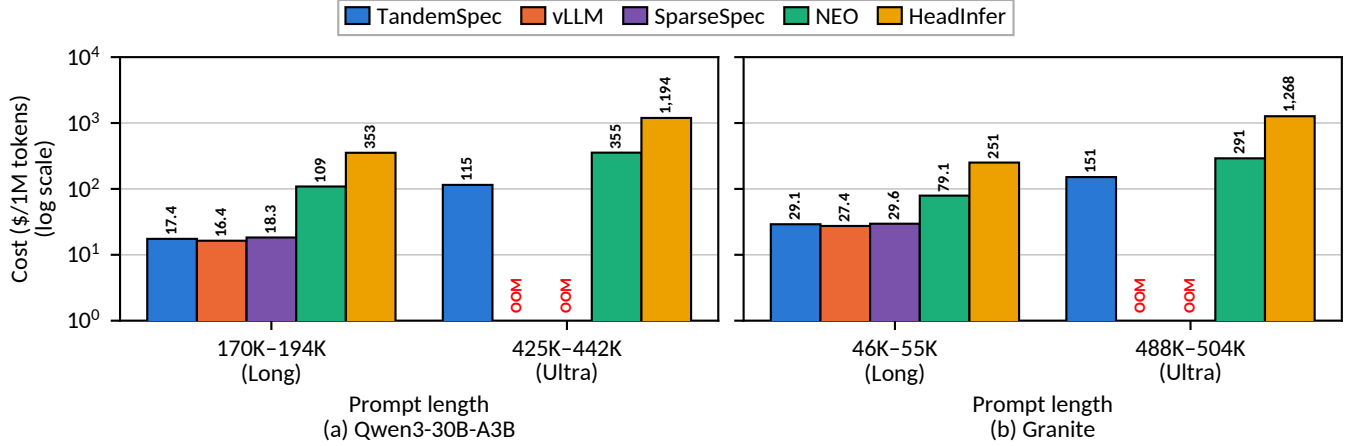


Figure 2. Cost per 1M tokens on RTX PRO 6000 (LongBench-v2 Long and Ultra buckets, batch size 1).

System	TTFT (s)	TPOT (ms)	E2E (s)	T_{out} (tok/s)	\$/1M tokens
<i>Qwen3-30B-A3B, Long, 170K–194K, $N_{ref}=3,950$</i>					
TandemSpec	41.5	18.7	150.1	26.32	17.4
vLLM	41.0	18.7	149.4	26.44	16.3
SparseSpec	83.2	14.7	167.2	23.62	18.3
NEO	47.1	155.2	935.9	4.22	108.8
HeadInfer	60.2	549.6	3,228.6	1.22	353.3
<i>Qwen3-30B-A3B, Ultra, 425K–442K, $N_{ref}=3,763$</i>					
TandemSpec	242.0	76.2	942.7	3.99	115.0
vLLM			Out of memory		
SparseSpec			Out of memory		
NEO	238.7	273.9	2,911.1	1.29	355.2
HeadInfer	279.1	1,043.9	10,394.6	0.36	1,193.9
<i>Granite-4.2-30B, Long, 46K–55K, $N_{ref}=2,914$</i>					
TandemSpec	11.9	49.7	184.6	15.79	29.1
vLLM	11.9	49.7	184.6	15.79	27.4
SparseSpec	14.1	53.0	199.8	14.59	29.6
NEO	14.0	140.8	501.8	5.81	79.1
HeadInfer	16.9	483.4	1,692.4	1.72	251.0
<i>Granite-4.2-30B, Ultra, 488K–504K, $N_{ref}=8,190$</i>					
TandemSpec	673.5	247.4	2,696.4	3.04	151.2
vLLM			Out of memory		
SparseSpec			Out of memory		
NEO	589.1	563.1	5,194.3	1.58	291.2
HeadInfer	681.0	2,852.4	24,008.1	0.34	1,267.0

Table 7. Long-context cost on RTX PRO 6000 (Qwen3-30B-A3B-Thinking-2507 and Granite-4.2-30B, LongBench-v2 Long and Ultra buckets, batch size 1, median over 20 Long and 5 Ultra documents). C_{hour} is \$1.653/h for TandemSpec and NEO and \$1.556/h for the others.

the price of Q4 2025, before the surge of 2026; the cheapest listing on October 21, 2026 is $5.0\times$ higher. The results for long contexts are measured with a batch size of 1, and we leave the evaluation of larger batch sizes for long contexts to future work.

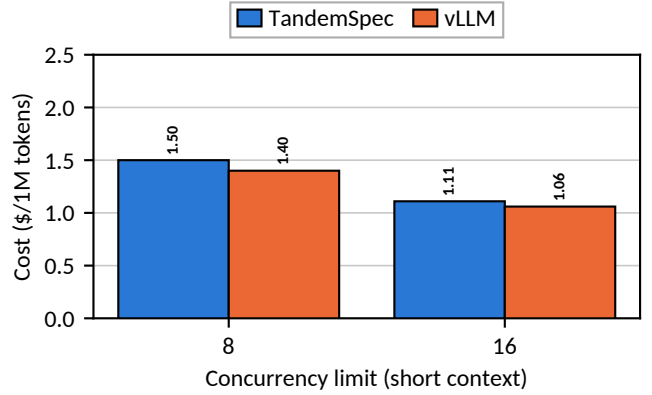


Figure 3. Cost per 1M tokens on RTX PRO 6000 (AIME25, Qwen3-30B-A3B-Thinking-2507, short context).

Concurrency	System	Requests	T_{out} (tok/s)	C_{hour} (\$/h)	\$/1M tokens
8	TandemSpec	48	306.9	1.653	1.50
	vLLM	48	309.6	1.556	1.40
16	TandemSpec	96	415.3	1.653	1.11
	vLLM	96	409.3	1.556	1.06

Table 8. Short-context, batched cost on RTX PRO 6000 (AIME25, Qwen3-30B-A3B-Thinking-2507, concurrency limits of 8 and 16).

References

- [1] CBRE. 2026. North America Data Center Trends H2 2025, reported by MBA NewsLink: Data center market set records across indicators in 2025. <https://newslink.mba.org/?p=272101>. Accessed: 2026-10-22.
- [2] Counterpoint Research. 2026. Memory prices have surged 80–90% QoQ in Q1 2026: server-grade 64 GB RDIMM prices jumped from USD 450 in Q4 2025, reported by Newsquawk. <https://www.newsquawk.com/headlines/memory-prices-have-surged-80-90-q1-2026-driven-by-steep-gains-in-server-dram-and-rising-nand-and-hbm3e-costs-according-to-counterpoint-research>. Accessed: 2026-10-21.

- [3] Esaitech. 2026. Intel Xeon 6776P 64-core processor, product listing. <https://esaitech.com/products/intel-xeon-6776p-pk807200645540-0-64-core-2-3-ghz-processor>. Accessed: 2026-10-21.
- [4] Esaitech. 2026. NVIDIA RTX PRO 6000 Blackwell Server Edition 96 GB, product listing. <https://esaitech.com/products/nvidia-rtx-pro-6000-900-2g153-0000-000-96gb-graphic-card>. Accessed: 2026-10-21.
- [5] Ingram Micro. 2026. NVIDIA L40S 48 GB GPU computing processor, product listing. <https://global.channelonline.com/andrew/ingram/Products/overview/M025137492>. Accessed: 2026-10-21.
- [6] Micron Technology. [n. d.]. 288-pin DDR5 RDIMM core product description. https://www.mouser.com/datasheet/2/671/ddr5_rdimmm_core-3310292.pdf. Accessed: 2026-10-21.
- [7] Newegg. 2026. Supermicro SYS-421GE-TNRT 4U GPU server barebone, product listing. <https://www.newegg.com/p/N82E16816139439>. Accessed: 2026-10-21.
- [8] NVIDIA. [n. d.]. ConnectX-6 Dx Ethernet adapter cards user manual: specifications. <https://docs.nvidia.com/networking/display/ConnectX6DxEN/Specifications>. Accessed: 2026-10-21.
- [9] Samsung Electronics. [n. d.]. PM9A3 NVMe SSD data sheet. <https://www.content.shi.com/cms-content/accelerator/media/pdfs/samsung/samsung-091823-ssd-pm9a3-datasheet.pdf>. Accessed: 2026-10-21.
- [10] SemiAnalysis. [n. d.]. GPU Cloud Economics Explained. <https://newsletter.semianalysis.com/p/gpu-cloud-economics-explained-the>. Accessed: 2026-10-21.
- [11] ServeTheHome. [n. d.]. We Found Something Testing the 5th Gen Intel Xeon Scalable 1P CPU. <https://www.servethehome.com/we-found-something-testing-the-5th-gen-intel-xeon-scalable-1p-cpu-supermicro-kioxia-asus/3/>. Accessed: 2026-10-21.
- [12] TechPowerUp. 2026. Intel Xeon Gold 6530 specifications. <https://www.techpowerup.com/cpu-specs/xeon-gold-6530.c3425>. Accessed: 2026-10-21.
- [13] U.S. Energy Information Administration. 2026. Electric Power Monthly, Table 5.6.A: Average price of electricity to ultimate customers by end-use sector. https://www.eia.gov/electricity/monthly/epm_table_grap her.php?t=epmt_5_06_a. Accessed: 2026-10-22.