Hardware & Networking

Data Center Watch: Token Spend Accelerates While GPU Rental and Memory Spot Pricing Track More Mixed in September

IT Hardware/ Telecom & Networking Equipment

Joseph Cardoso AC
(1-212) 622-9036
joseph.cardoso@jpmchase.com

Manmohanpreet Singh
(1-212) 622-4527
manmohanpreet.singh@jpmchase.com

Marc Vitenzon
(1-212) 622-3342
marc.vitenzon@jpmchase.com
J.P. Morgan Securities LLC

In the Data Center Watch report, we outline key data points relevant to AI infrastructure demand, including LLM token volumes and prices, GPU rental prices, and memory costs. Taken together, the latest data on LLM token spend growth suggest AI infrastructure demand remains strong, but pricing signals from GPU rentals and memory spot prices are more mixed. LLM token spend accelerated, driven by substantial m/m volume growth helped by volume-weighted pricing compression and heavier adoption of cheaper high-volume models. Meanwhile, GPU rental prices declined for A100s and H100s, while B200 pricing rebounded modestly. With regards to memory, DRAM spot prices declined modestly after five straight months of expansion, while NAND was effectively flat m/m. *Key highlights from Data Center Watch report:*

- **Token spend growth accelerated on OpenRouter in September.** LLM growth trends on OpenRouter in September were strong: token volume tracked to +71% m/m and 30x y/y, while pricing tracked to +4% m/m and +20% y/y. Net-net, overall spending was robust, with growth tracking to +28% m/m and 14x y/y. Volume-weighted pricing decreased -25% m/m and -55% y/y driven by: (1) mix shift toward newer, cheaper, high-volume models, with DeepSeek V4.1 Flash and GLM 5.3 Flash alone adding ~41 tn tokens (*30% of total volume*) with pricing well below the prior open-weight average; and (2) same-model price compression across Kimi K3, GLM 5.3, and GPT-5.6 Sol.
- **A100 and H100 GPU rental prices declined in September, while B200 prices rebounded modestly.** Sequential growth trends for GPU rental prices were mixed in September for Nvidia GPUs with regards to non-hyperscaler capacity. A100s and H100s both declined sequentially, while B200s rebounded after August's first decline in over six months. Pricing for B200s was ~2.2x that of H100s, and H100 pricing was ~1.7x that of A100s, with both ratios increasing sequentially.
- **Memory spot price growth stalled in September.** DRAM spot prices declined modestly m/m after five straight monthly gains, though DDR5 16Gb remains up more than 6x y/y, while NAND spot prices were effectively flat m/m and remain up nearly 5x y/y.

LLM Token Tracker

Our LLM Token Tracker captures a single week of usage and pricing data each month, providing a monthly snapshot or sampling of trends across ~300 LLMs listed on OpenRouter, an aggregation platform that routes API requests to models through a single endpoint. Keep in mind, OpenRouter skews towards developer, startup, and agentic-coding traffic and excludes first-party API volume (direct OpenAI/Anthropic, hyperscaler-hosted endpoints).

Key takeaways:

- **Token volume growth accelerated in the month of September, following strength in August.** Token volume growth accelerated on a sequential-basis

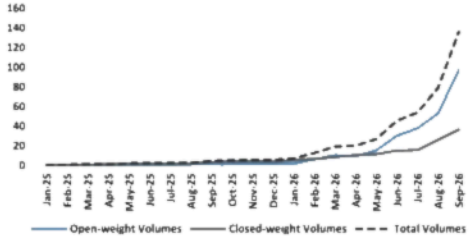
See page 5 for analyst certification and important disclosures.

in September, expanding +71% m/m (vs. +47% in August and −18% in July), and 30x y/y (vs. 28x in August and 21x in July). Volumes for closed-weight models (OpenAI, Anthropic, etc.), which represented 27% of the total (vs. 33% in August and 29% in July), saw growth moderate in September, expanding +39% m/m (vs. +66% in August and +3% in July) and 10x y/y (vs. 14x in August and 8x in July). Volumes for open-weight models (DeepSeek, Moonshot AI, etc.) saw growth accelerate sharply in September, expanding +83% m/m (vs. +39% in August and +26% in July) and 101x y/y (vs. 56x in August and 63x in July).

- **Token VWAP continued to compress.** Volume-weighted pricing decreased -25% m/m and -55% y/y, driven by: (1) mix shift toward newer, cheaper, high-volume models, with DeepSeek V4.1 Flash and GLM 5.3 Flash alone adding 41 tn tokens at an average price well below the prior open-weight average; and (2) same-model price compression, including Kimi K3, GLM 5.3, and GPT-5.6 Sol. Volume-weighted pricing for open-weight models declined -11% m/m (vs. -17% in August and −36% in July) and rose +30% y/y (vs. +60% in August and +74% in July) vs. -15% m/m in September for closed-weight models (vs. -37% in August and -5% in July) and ~flat y/y (vs. -29% in August and +18% in July). At the same time, average token pricing increased +4% m/m in September (vs. -5% in August and -3% in July), while year-over-year growth also improved, with pricing expanding +20% y/y (vs. −5% in August and +2% in July) given the addition of more expensive image models to the catalog.
- **Token spending growth accelerated in September.** Net-net, token spending growth strengthened in September, expanding +28% m/m (vs. +7% in August and +7% in July) and 14x y/y (vs. 12x in August and 11x in July). Spending for closed-weight models, which represented 72% of the total (vs. 78% in August and 80% in July), saw growth re-accelerate in September, increasing +18% m/m (vs. +4% in August and -2% in July) and 10x y/y (vs. 10x in August and 9x in July). By comparison, spending for open-weight models saw growth re-accelerate sharply in September, increasing +64% m/m (vs. +16% in August and +71% in July) and 132x y/y (vs. 89x in August and 109x in July).
- **One model overlapped in the top five by spend versus usage.** The top five models in September based on token usage were: (1) DeepSeek V4.1 Flash, (2) GLM 5.3 Flash, (3) Tencent's Hy4 preview, (4) GPT-5.6 Luna, and (5) DeepSeek V4 Flash, accounting for 53%+ of total token volume, up from 50% + in August. However, on a spending basis, the top five models in September were: (1) GPT-6 Astra, (2) Tencent's Hy4 preview, (3) Claude Fable 5.1, (4) Claude Opus 5, and (5) GPT-5.6 Sol, accounting for 50% of total token spend, up from 48% in August.

Figure 1: Token Volumes

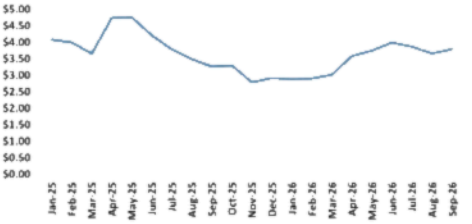
Weekly Tokens in Trillion



Source: OpenRouter, J.P. Morgan Estimates

Figure 2: Average Token Pricing

$ per Million Tokens



Source: OpenRouter, J.P. Morgan Estimates
* Includes models without observable volumes on OpenRouter

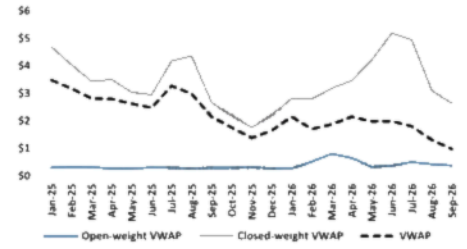
人在草木间

Joseph Cardoso [AC]
(1-212) 622-9036
joseph.cardoso@jpmchase.com

J.P.Morgan

Figure 3: Volume-Weighted Average Token Pricing

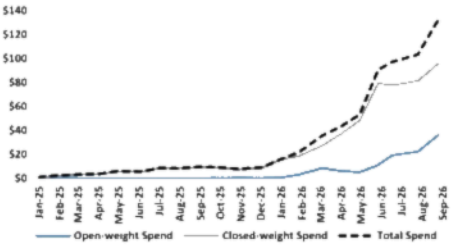
$ per Million Tokens



Source: OpenRouter, J.P. Morgan Estimates

Figure 4: Snapshot of Weekly Token Spend

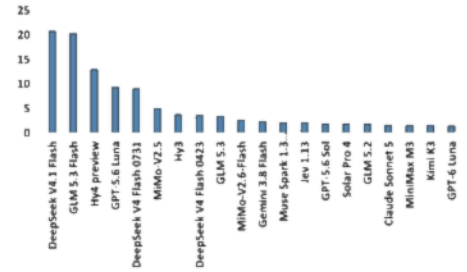
$ in Million



Source: OpenRouter, J.P. Morgan Estimates

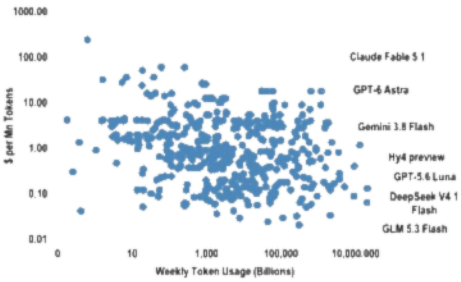
Figure 5: Top 20 Models Based on Volume

Weekly Tokens in Trillion



Source: OpenRouter, J.P. Morgan Estimates

Figure 6: Token Pricing vs. Volume



Source: OpenRouter, J.P. Morgan Estimates

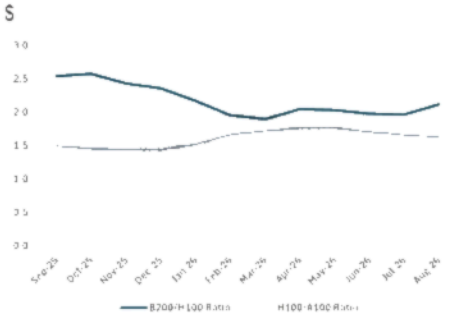
GPU Rental Price Tracker

Key takeaways based on Bloomberg indices:

- **A100 rental prices declined m/m in September.** A100 rental prices tracked to $1.59/GPU-hour on average in September, down -2.8% m/m (vs. -0.7% m/m in August and +0.9% m/m in July).
- **H100 rental prices declined m/m in September; H100-to-A100 ratio ticked higher.** H100 rental prices tracked to $2.64/GPU-hour on average during September, a -2.6% m/m decline (vs. +0.4% m/m in August and -0.6% m/m in July). The H100-to-A100 price ratio tracked to 1.66x on average in September (vs. 1.65x in August and 1.64x in July).
- **B200 rental prices rebounded modestly m/m in September; B200-to-H100 ratio moved higher.** B200 rental prices tracked to $5.70/GPU-hour on average in September, up +1.3% m/m (vs. -1.5% m/m in August and +7.2% m/m in July). The B200-to-H100 price ratio tracked to 2.16x on average in September (vs. 2.08x in August, 2.12x in July, 1.97x in June, and 2.58x around index launch in Sep-25).

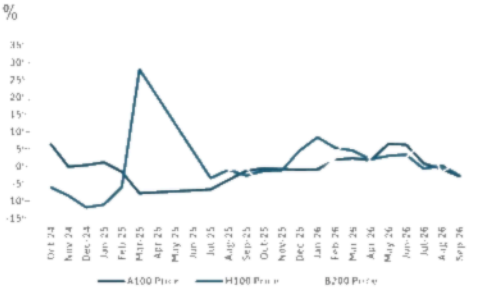
Joseph Cardoso [AC]
(1-212) 622-9036
joseph.cardoso@jpmchase.com

Figure 7: Avg. Monthly Rental Prices (Non-Hyperscaler)



Source: Bloomberg Finance L.P.

Figure 8: Avg. Monthly Rental Prices (M/M % Change)



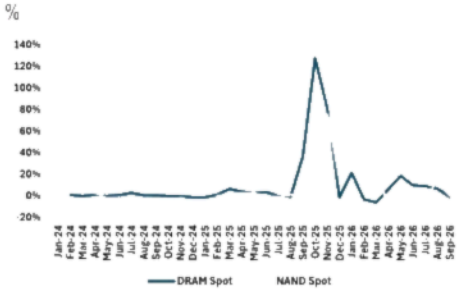
Source: Bloomberg Finance L.P

Memory Price Tracker

Key takeaways based on Bloomberg data:

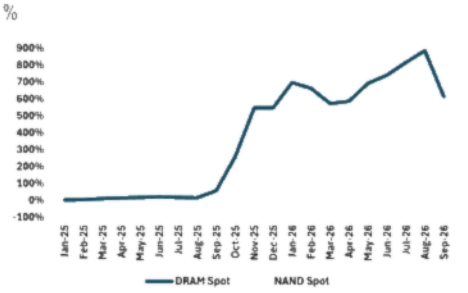
- **DRAM spot prices declined modestly in September, though DDR5 16Gb remains up more than 6x y/y.** DDR5 16Gb spot prices reached $49.70 in September, declining -1% m/m (vs. $50.20 in August and $47.00 in July) and increasing +614% y/y (vs. $6.96 in September 2025), marking the first m/m decline after five consecutive monthly increases.
- **NAND spot prices were effectively flat in September; NAND 1Tb remains up nearly 5x y/y.** NAND 1Tb spot prices tracked to $30.54 in September, an increase of +0.1% m/m (vs. $30.50 in August and $26.90 in July) and an increase of +470% y/y (vs. $5.36 in September 2025), marking stability following the first month of increase after four consecutive months of modest m/m declines.

Figure 9: Memory Prices (M/M % Change)



Source: Bloomberg Finance L.P.

Figure 10: Memory Prices (Y/Y % Change)



Source: Bloomberg Finance L.P.