J.P.Morgan

Hardware & Networking

Data Center Watch: Token Spend Climbs, GPU Pricing Mixed, & Memory Prices March Higher

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

Akanksh Chauhan
(1-212) 622-0045
akanksh.chauhan@jpmorgan.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 suggest positive trends in the AI infrastructure demand environment, including continued expansion in LLM token spend, with substantial m/m volume growth helped by pricing discounts and open-weight model adoption. Meanwhile, GPU rental price for B200s modestly declined vs. rebounded for H100s. With regards to memory, spot prices for NAND returned to growth while DRAM continued on its upward trajectory. *Key highlights from Data Center Watch report*:

- **Token spend growth continues on OpenRouter in August, driven by a substantial uptick in volumes.** LLM growth trends on OpenRouter in August were mixed following a strong July: token volume tracked to *+47% m/m and 28x y/y*, while pricing tracked to *-5% m/m and +5% y/y*. Net-net, overall spending was robust, with growth tracking to *+7% m/m and 12x y/y*. Volume-weighted pricing *decreased -28% m/m and -56% y/y* driven by: (1) substantial list price discounts (*50% off on GPT-5.6 Sol and 80% off on GPT-5.6 Luna, for example*); and (2) considerable adoption of low-cost models (*Tencent's Hy3, DeepSeek v4 Flash-0731, and GPT-5.6 Luna accounted for ~99% of the m/m token volume increase*).
- **A100 and H100 GPU rental price growth rates were stable, while B200 price growth turned negative.** GPU rental price sequential growth trends were mixed in August for Nvidia GPUs with regards to non-hyperscaler capacity. B200s showed a first sequential decline in over six months, while A100s were flat sequentially and H100s rebounded modestly. Pricing for B200s was ~2.1x that of H100s, and H100 pricing was ~1.7x that of A100s, with the former ratio falling sequentially and the latter ratio increasing for the first time in four months.
- **Memory spot price growth continues for DRAM and rebounds for NAND.** DRAM spot prices climbed sequentially again in August, a fifth straight monthly gain that puts it up more than 8x y/y, while NAND spot prices strongly rebounded after four months of declines, supporting y/y growth of over 4x.

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 re-accelerated in the month of August, following a moderation in July.** Token volume growth re-accelerated on a sequential-basis in August, expanding +47% m/m (vs. +18% in July and +70% in June), and 28x y/y (vs. 21x in July and 20x in June). Volumes for closed-weight models (OpenAI, Anthropic, etc.), which represented 33% of the total (vs. 29%

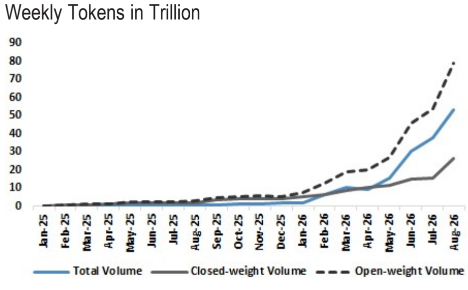
See page 5 for analyst certification and important disclosures.

in July and 33% in June), saw growth re-accelerate in August, expanding +66% m/m (vs. +3% in July and +34% in June) and 14x y/y (vs. 8x in July and 8x in June). Volumes for open-weight models (DeepSeek, Moonshot AI, etc.) also saw growth re-accelerate in August, expanding +39% m/m (vs. +26% in July and +97% in June) and 56x y/y (vs. 63x in July and 74x in June).

- **Token pricing declined sequentially in August.** Average token pricing fell -5% m/m in August (vs. -3% in July and +7% in June), while year-over-year growth modestly improved, with pricing expanding +5% y/y (vs. +2% in July and -5% in June). Volume-weighted pricing decreased -28% m/m and -56% y/y driven by: (1) substantial list price discounts (*50% off on GPT-5.6 Sol and 80% off on GPT-5.6 Luna, for example*); and (2) considerable adoption of low-cost models (*Tencent's Hy3, DeepSeek v4 Flash-0731, and GPT-5.6 Luna accounted for ~99% of the m/m token volume increase*). Volume-weighted pricing for open-weight models declined -17% m/m (vs. +36% in July and +12% in June) and rose +60% y/y (vs. +74% in July and +25% in June) vs. -37% m/m in August for closed-weight models (vs. -5% in July and +23% in June) and -29% y/y (vs.+18% in July and +76% in June).
- **Token spending continued to grow in August, following robust growth in July.** Net-net, token spending growth was firm in August, expanding +7% m/m (vs. +7% in July and +70% in June) and 12x y/y (vs. 11x in July and 16x in June). Spending for closed-weight models, which represented 78% of the total (vs. 80% in July and 88% in June), saw growth rebound in August, increasing +4% m/m (vs. -2% in July and +65% in June) and 10x y/y (vs. 9x in July and 14x in June). By comparison, spending for open-weight models saw growth decelerate in August, increasing +16% m/m (vs. +71% in July and +120% in June) and 89x y/y (vs. 109x in July and 92x in June).
- **No overlap in top five models by spend versus usage.** The top five models in August based on token usage were: (1) DeepSeek v4 Flash-0731, (2) Tencent's Hy3, (3) MiMo v2.5, (4) GPT-5.6 Luna, and (5) DeepSeek v4 Flash, accounting for 50%+ of total token volume. However, on a spending basis, the top five models in August were: (1) Claude Opus 5, (2) Kimi K3, (3) Claude Opus 4.7, (4) GPT-5.6 Sol, and (5) Claude Fable 5, accounting for 48% of total token spend.

Figure 1: Token Volumes

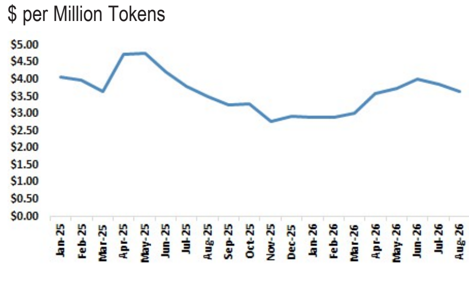
Weekly Tokens in Trillion



Source: OpenRouter, J.P. Morgan Estimates

Figure 2: Average Token Pricing

$ per Million Tokens

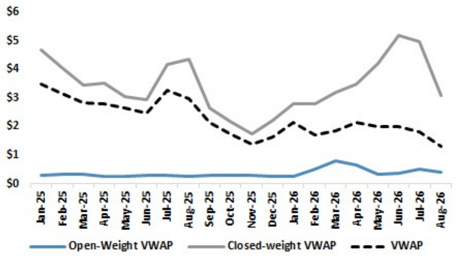


Source: OpenRouter, J.P. Morgan Estimates
* Includes models without observable volumes on OpenRouter

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

Figure 3: Volume-Weighted Average Token Pricing

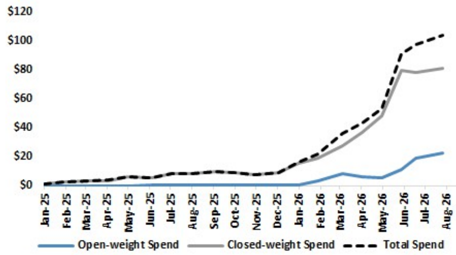
$ per Million Tokens



Source: OpenRouter, J.P. Morgan Estimates

Figure 4: Snapshot of Weekly Token Spend

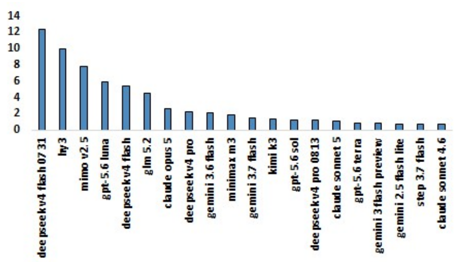
$ in Million



Source: OpenRouter, J.P. Morgan Estimates

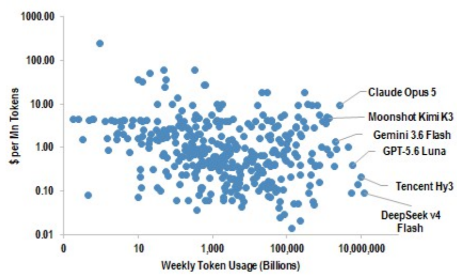
Figure 5: Top 20 Models Based on Volume

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 were flat m/m in August.** A100 rental prices tracked to $1.65/GPU-hour on average in August, flat m/m (vs. +0.9% m/m in July and +6.4% m/m in June).
- **H100 rental prices rebounded m/m in August; H100-to-A100 ratio ticked higher.** H100 rental prices tracked to $2.73/GPU-hour on average during August, a +1.2% m/m increase (vs. -0.6% m/m in July and +3.5% m/m in June). The H100-to-A100 price ratio tracked to 1.65x on average in August (vs. 1.64x in July and 1.66x in June).
- **B200 rental price declined modestly m/m in August; B200-to-H100 ratio falls.** B200 rental prices tracked to $5.63/GPU-hour on average in August, down -1.5% m/m (vs. +7.2% m/m in July and +2.7% m/m in June), marking the first sequential decline in over six months. The B200-to-H100 price ratio tracked to 2.06x on average in August (vs. 2.12x in July, 1.97x in June, and 2.58x around index launch in Sep-25).

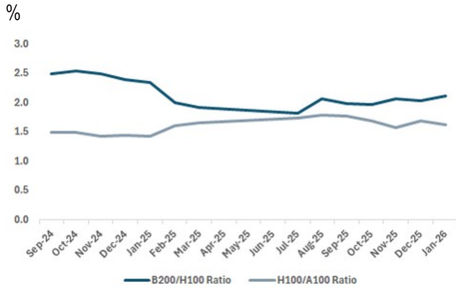
Figure 7: Avg. Monthly Rental Prices (Non-Hyperscaler)

$

Source: Bloomberg Finance L.P.

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

%



Source: Bloomberg Finance L.P.

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

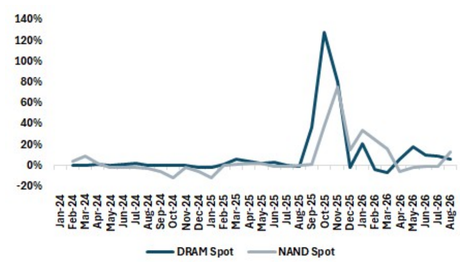
Memory Price Tracker

Key takeaways based on Bloomberg data:

- **DRAM spot prices continued to rise in August, with DDR5 16Gb up more than 8x y/y.** DDR5 16Gb spot prices reached $50.20 in August, rising +6% m/m (vs. $47.00 in July and $43.14 in June) and +880% y/y (vs. $5.09 in August 2025), marking a fifth consecutive m/m increase following modest sequential declines in February and March.
- **NAND spot prices rebounded in August; NAND 1Tb up more than 4x y/y.** NAND 1Tb spot prices tracked to $30.50 in August, an increase of +14% m/m (vs. $26.90 in July and $27.03 in June) and an increase of +470% y/y (vs. $5.29 in August 2025), marking the first month of increase following four consecutive months of modest m/m declines after the strong run-up through 1Q26.

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

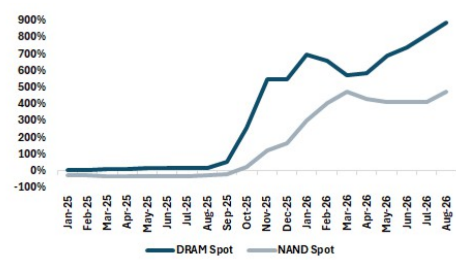
%



Source: Bloomberg Finance L.P.

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

%



Source: Bloomberg Finance L.P.