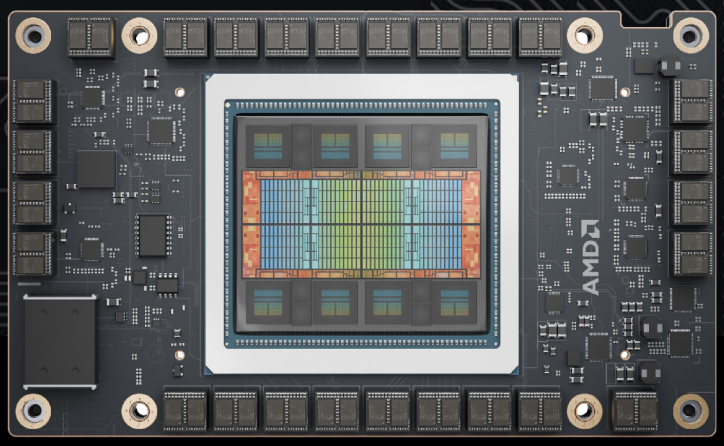


AMD Instinct MI355X Kimi K3 Can Generate Up to $32B of Revenue per GigaWatt per Year

Executive Summary - Agentic Inference Serving Economics Analysis

Published by SemiAnalysis InferenceX Team - September 4, 2026

AMD Instinct MI355X Kimi K3 Can Generate Up to $32B of Revenue per GigaWatt per Year

Executive Summary - Agentic Inference Serving Economics Analysis · Published by SemiAnalysis InferenceX Team · September 4, 2026

Summary

InferenceX AgentX measures how many agentic-coding tokens a production serving stack delivers at a fixed interactivity target. Priced at Kimi K3's public OpenRouter rates and scaled to one gigawatt of provisioned power, the MI355X vLLM result is **$32.6 billion of revenue per year** at P90 45 tok/s/user and 60% utilization.

Owned at hyperscaler volume ($1.50 per chip-hour), the fleet keeps $16.5 billion of profit at a 50.7% margin. Rented on a 3-year contract at August 2026 pricing ($3.00 per chip-hour), it keeps $10.3 billion at 31.5%.

$32.6B
Revenue per GW-year
Kimi K3 at OpenRouter list price, 60% utilization

$16.5B
Profit per GW-year
Owned at hyperscaler volume, $1.50 per chip-hour · **$3.95** per chip-hour

50.7%
Profit margin
After compute expense and 30% model license fee

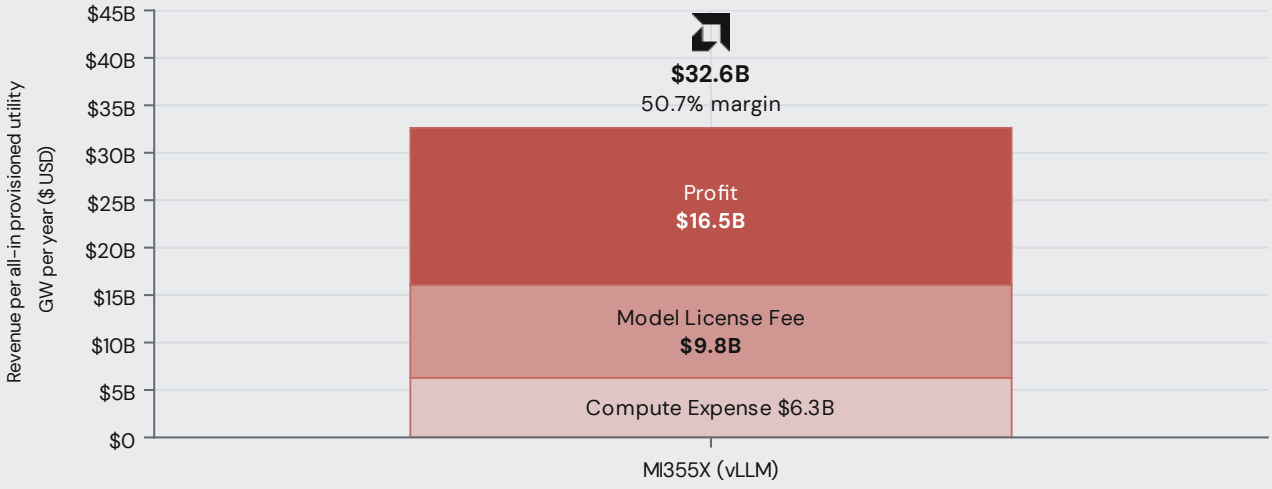
Figure 1

Kimi K3 2.8T Agentic Revenue & Profit Estimates per GigaWatt Per Year at P90 45 tok/s/user Interactivity

Cost Tier: Owning Hyperscaler **Utilization:** 60% **Model License Fee Assumption:** 30% **Updated:** 2026-09-04 **Source:** SemiAnalysis InferenceX™

TCO $/chip/hr: MI355X: 1.5 Source: SemiAnalysis Market August 2026 Pricing Surveys & AI Cloud TCO Model ↗

Selling Price per Million Tokens: Input $3 · Cached Input $0.3 · Output $15 (OpenRouter)



Owned at hyperscaler volume. Compute is charged on every provisioned GPU-hour at $1.50; the model lab takes 30% of revenue as a license fee; the operator keeps $16.5B, or $3.95 per chip-hour.

$32.6B
Revenue per GW-year
Kimi K3 at OpenRouter list price, 60% utilization

$10.3B
Profit per GW-year
Rented at $3.00 per chip-hour · **$2.45** per chip-hour

31.5%
Profit margin
After compute expense and 30% model license fee

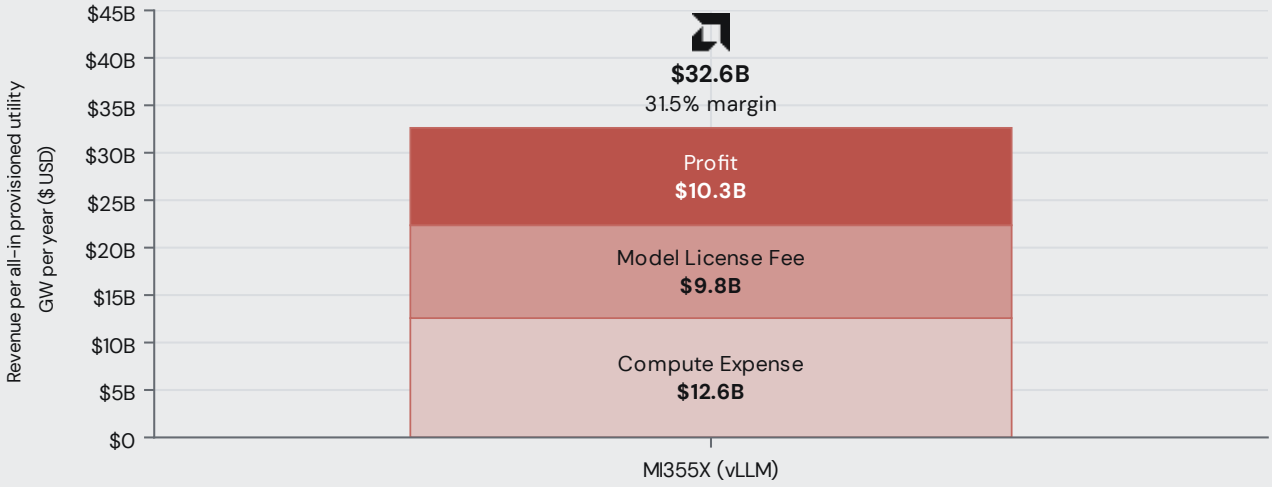
Figure 2

Kimi K3 2.8T Agentic Revenue & Profit Estimates per GigaWatt Per Year at P90 45 tok/s/user Interactivity

Cost Tier: Custom $/GPU/hr (August 2026 3-Year Rental Pricing) **Utilization:** 60% **Model License Fee Assumption:** 30% **Updated:** 2026-09-04 **Source:** SemiAnalysis InferenceX™

TCO $/chip/hr: MI355X: 3 Source: SemiAnalysis Market August 2026 Pricing Surveys & AI Cloud TCO Model ↗

Selling Price per Million Tokens: Input $3 · Cached Input $0.3 · Output $15 (OpenRouter)



Rented at $3.00 per chip-hour. Revenue and license fee are unchanged; compute expense doubles to $12.6B and profit falls to $10.3B, or $2.45 per chip-hour. The rented fleet breaks even at 33% utilization; the owned fleet at 16.5%.

How the estimate is built

- **6,115 tok/s per GPU** at P90 45 tok/s/user, interpolated on the measured MI355X vLLM frontier (FP4, MTP, TP8).
- **478,469 GPUs per gigawatt** at 2.09 kW all-in per chip; 60% of benchmarked throughput is sold.
- **$0.59 per million tokens** blended: 99.2% of tokens are input and 93.7% of those hit the prefix cache.
- **Profit** = revenue − compute (every GPU-hour, busy or idle) − 30% model license fee.

Sources: inferencex.semianalysis.com/profit-estimator-per-gigawatt · inferencex.semianalysis.com/agentx · openrouter.ai/moonshotai/kimi-k3 · semianalysis.com/ai-cloud-tco-model. Figures are estimates from benchmark throughput and public list prices, not forecasts of any operator's results.