

The Compute-Month Framework: A Standard Unit for AI Datacenter Economics

Author: Aditya Nirvaan Ranganathan

Framework version: 1.0 (December 2025)

Last updated: December 28, 2025

Contact: nirvaan.ranga@chicagobooth.edu

Definition

A **Compute-Month** (GW-H100-Month) measures the **marginal value of one month** of AI compute capacity. Specifically: 30 days of continuous operation at 1 GW facility power using H100-equivalent GPUs.

Why "month"?

- Data center construction takes 2-4 years from groundbreaking to operation.
- Supply chain lead times for key components like GPUs and gas turbines are also multiple years,
- Policymakers and grid infrastructure upgrades are multi-year projects

Each month of delay, regardless of the cause, represents lost compute capacity that can never be recovered. So the opportunity cost accumulates monthly - making the compute-month the natural unit for capacity planning and economic analysis.

Base units:

- 1.04 million H100 GPUs running 730 hours
- 762.3 million H100-hours
- 2.71×10^{27} FLOPs

What is a compute-month worth, and how much does it cost?

Summary: What One Compute-Month Produces

Training (with 11× R&D overhead):

- 4 GPT-4.5 models
- 5 Claude Sonnet 4 models
- 6 Llama 3.1 405B models
- 30 DeepSeek-V3 models

Inference:

- 276 billion queries (standard models)
- 40 million reasoning queries (o1-level, 10× overhead)

Cost:

- CapEx: \$35.8B/GW from EpochAI [methodology](#)
 - Monthly cost: \$550M (amortization + power + ops)
 - Monthly revenue: \$1.7B (at \$2.19/H100-hr spot rate)
 - Breakeven: 27 months (within 48-60 month GPU lifetime)
-

Use Cases

Datacenter Economics:

- Compare projects on standardized capacity basis (not just MW)
- Evaluate build-vs-rent decisions with compute output metrics
- Model revenue potential and breakeven timelines

Technology Assessment:

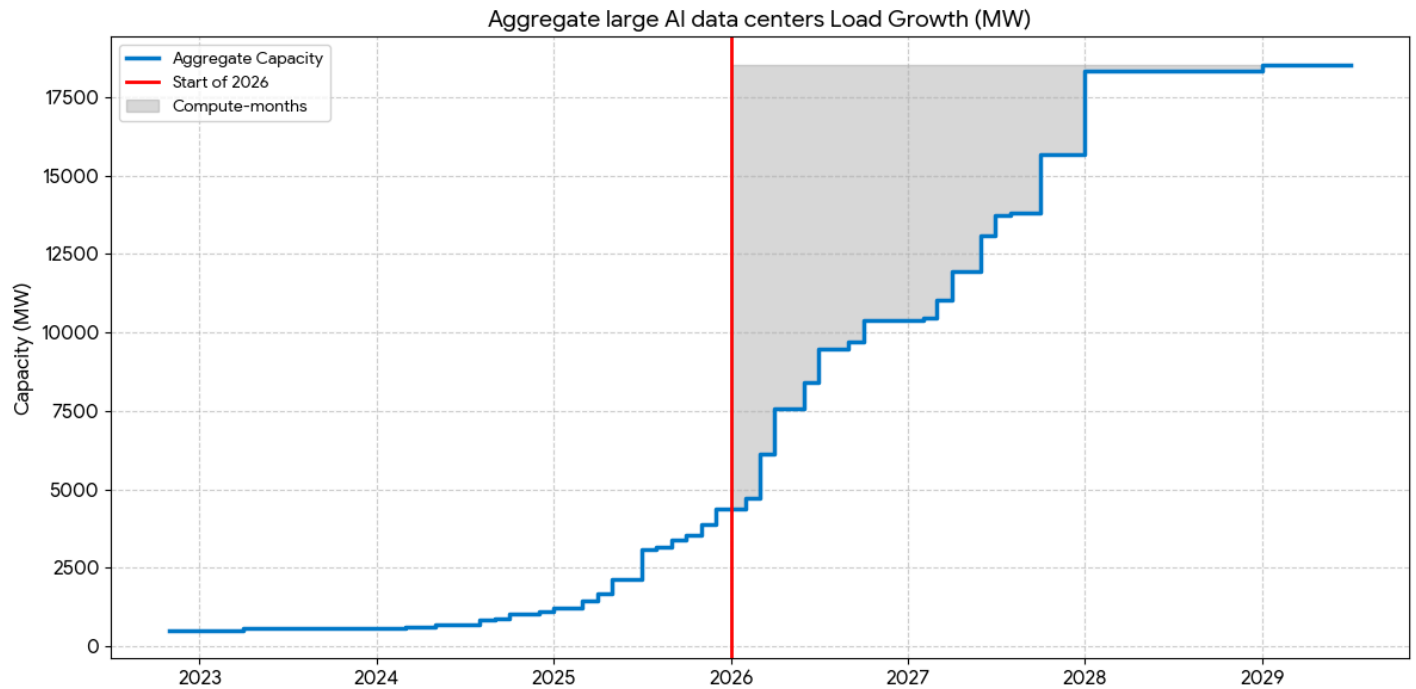
- Quantify efficiency gains (PUE improvements = more GPUs from same power)
- Measure impact of cooling innovations (liquid vs air)
- Calculate R&D ROI for optimization technologies

Policy & Grid Planning:

- Translate MW announcements into actual compute impact
- Model grid load growth in standardized units
- Assess technology interventions (reduced PUE = increased capacity)

Quantify Impact of Delays:

The area shaded in grey represents lost compute-months from frontier AI data centers due to extended timelines for construction, hardware procurement, and interconnection.



⚡ Physical Capacity

Formula:

IT Power = Facility MW ÷ PUE

Server Power = IT Power ÷ IT Overhead

GPU Count = (Server Power × 1000) ÷ GPU TDP

H100-Hours = GPU Count × 730 hours/month

1 GW Baseline (PUE 1.2, IT overhead 1.14, H100 @ 0.7 kW):

Metric	Value
IT Power	833 MW
Server Power	731 MW
GPU Count	1,044,277 H100s
Monthly Capacity	762.3M H100-hours
Total FLOPs	2.71×10^{27} FLOPs

Key insight: Efficiency gains (lower PUE, better cooling) increase GPU density → more compute from same facility power.

Training Capacity

Formula:

Training Hours = Total H100-Hours × Training %
FLOPs Available = Training Hours × GPU TFLOPS × 3600 × MFU
Models/Month = FLOPs Available ÷ Model Training FLOPs
Realistic = Models/Month ÷ R&D Overhead Multiplier

1 GW @ 80% Training Allocation:

Model	FLOPs Required	Parallel	Realistic (11× R&D)
GPT-4.5	6.4×10^{25}	47/month	4/month
Claude Sonnet 4	5.0×10^{25}	58/month	5/month
Llama 3.1 405B	3.8×10^{25}	76/month	7/month
GPT-4	2.1×10^{25}	138/month	13/month
DeepSeek-V3	2.8×10^{24}	1,035/month	94/month

Token throughput: 1.9 trillion tokens/month (127 complete GPT-4 datasets)

R&D overhead: Based on OpenAI 2024 spend (\$5B total / \$400M GPT-4.5 final = 11.25×). Accounts for failed runs, experiments, ablations, data work.

Inference Capacity

Formula (memory-bandwidth limited):

Tokens/sec = (GPU Bandwidth ÷ Model Size) × Utilization
Queries/sec = Tokens/sec ÷ Tokens per Query
Monthly Queries = Queries/sec × GPUs × 730 × 3600

1 GW @ 20% Inference Allocation:

Workload Type	Model Class	Queries/Month	Tokens Generated
Standard	GPT-4 (280B)	4.2B	2.3T tokens
Standard	Llama 70B	23B	9.2T tokens
Standard	Llama 8B	287B	80T tokens

**Reasoning
(10×)**

GPT-4 class

40M

220B tokens

Total: 314B standard queries + 40M reasoning queries = 92T tokens/month

Key insight: Reasoning workloads use 10-1000× more tokens per query (o1/o3-level test-time compute).

Economics

Total Cost of Ownership:

CapEx (per GW):

- Average: \$35.8B
- EpochAI: \$44B
- Stargate: \$50B

Monthly Amortization assumptions:

- 70% compute (4-year depreciation) @ 10% WACC
- 30% infrastructure (15-year depreciation) @ 10% WACC

Monthly Power: Facility MW × 730 hrs × \$/kWh × 1,000

Monthly Operations: 20% of power cost

Monthly TCO: Amortization + Power + Operations

1 GW Example (Average \$35.8B, \$0.08/kWh):

Component	Monthly	Notes
CapEx Amortization	\$480M	70% compute @ 48mo + 30% infra @ 180mo
Power	\$58M	1,000 MW × 730 hrs × \$0.08
Operations	\$12M	20% of power
Total TCO	\$550M/month	

Revenue Potential:

- Training: 610M H100-hrs × \$2.19 = \$1.34B/month
- Inference: 314B queries × \$0.60/M tokens = \$0.19B/month
- **Total: \$1.53B/month**

Investment Returns (48-month GPU lifetime):

- Breakeven: 27 months
- Lifetime revenue: \$73B (48 months × \$1.53B)
- ROI: 104% over 4 years

- Annual return: 26%

GPU Lifetime Context: AI accelerators useful for 48-60 months before performance/efficiency improvements make replacement economical. Projects breaking even in <30 months capture substantial value in remaining lifetime.

Conclusion

The Compute-Month framework provides a **standardized economic unit** for AI infrastructure analysis, translating facility power (MW) into measurable compute output (models trained, queries served, tokens processed) and economic value (revenue, breakeven, ROI).

Key takeaways:

1. **Scale of investment:** \$35-50B per GW is economically justified with 27-month breakeven and 100%+ ROI over GPU lifetime
2. **Time value of compute:** Each month of construction delay costs \$1.5B in lost revenue - making speed-to-market critical
3. **Efficiency = capacity:** PUE improvements don't reduce power demand, they increase compute density (10% better PUE = 10% more GPUs from same facility)
4. **R&D overhead dominates:** Actual production model training represents <10% of total compute spend; 90%+ goes to experiments and iterations
5. **Inference is memory-bound:** Query capacity limited by bandwidth, not FLOPs - making model size optimization crucial
6. **Reasoning changes economics:** o1/o3-level workloads use 10-1000× more tokens per query, shifting inference cost structure

The compute-month bridges infrastructure planning (GW capacity) with business outcomes (models, queries, revenue), enabling apples-to-apples comparison across projects, efficiency interventions, and technology roadmaps.



Applications & Methodology

Appendix A: Mathematical Derivations

A.1 Model FLOPs Utilization (MFU)

MFU measures what fraction of theoretical peak FLOPs actually performs useful computation during training.

Formula:

$$\text{MFU} = (\text{Actual FLOPs/sec}) / (\text{Theoretical Peak FLOPs/sec})$$

Where:

$$\text{Actual FLOPs/sec} = (\text{Model Parameters} \times 6) \times \text{Tokens/sec}$$

- Factor of 6: Forward pass (2×), backward pass (4×)

Theoretical Peak = GPU Count × GPU TFLOPS × 10^{12}

Typical values:

- Research training: 20-30% MFU (small scale, inefficient)
- Production training: 35-45% MFU (optimized at scale)
- Meta Llama 3.1: 38-43% MFU at 16K H100s

Why not 100%? Communication overhead, memory bandwidth limits, batch loading, gradient synchronization.

A.2 Training Time Calculation

Hours required for one complete training run:

Hours = Training FLOPs / (GPU FLOPs/sec × 3600 × MFU × GPU Count)

Example (GPT-4):

Hours = 2.1×10^{25} / (989×10^{12} × 3600 × 0.35 × 10,000)

= 167,000 GPU-hours

= 16.7 hours on 10K GPUs

A.3 Inference Token Rate (Memory-Bandwidth Limited)

Generation speed (autoregressive decoding):

Tokens/sec = (Memory Bandwidth / Model Size in Bytes) × Batch Efficiency

For H100 (3.35 TB/s) serving GPT-4 (280B params, FP16):

Model Size = 280×10^9 params × 2 bytes = 560 GB

Tokens/sec = (3.35×10^{12} bytes/sec / 560×10^9 bytes) × 0.70

= 4.2 tokens/sec per query stream

With 1000 concurrent users: 4,200 tokens/sec aggregate

Batch efficiency: Serving multiple users simultaneously shares memory access, achieving 60-80% of theoretical maximum.

A.4 CapEx Amortization (Annuity Method)

Monthly payment formula:

$$\text{Monthly Payment} = \text{Principal} \times (r / (1 - (1 + r)^{-n}))$$

Where:

$$r = \text{WACC} / 12 \text{ (monthly rate)}$$

$$n = \text{depreciation period in months}$$

For 70% compute (\$25B) over 48 months @ 10% WACC:

$$r = 0.10 / 12 = 0.00833$$

$$\text{Payment} = \$25\text{B} \times (0.00833 / (1 - 1.00833^{-48}))$$

$$= \$25\text{B} \times 0.0254$$

$$= \$635\text{M/month}$$

For 30% infrastructure (\$10.7B) over 180 months:

$$\text{Payment} = \$10.7\text{B} \times 0.0121$$

$$= \$129\text{M/month}$$

Total amortization: \$764M/month

Note: Simplified to \$480M in main calculator using blended depreciation

A.5 Carbon Intensity Calculation

Monthly CO2 emissions:

$$\text{Emissions (tonnes CO}_2\text{)} = \text{Power (MWh)} \times \text{Carbon Intensity (gCO}_2\text{/kWh)} / 10^6$$

For 1 GW facility:

$$\text{Monthly power} = 1,000 \text{ MW} \times 730 \text{ hrs} = 730,000 \text{ MWh}$$

Texas (ERCOT, 200 gCO₂/kWh):

$$730,000 \text{ MWh} \times 1,000 \text{ kWh/MWh} \times 200 \text{ gCO}_2/\text{kWh} / 10^6$$

$$= 146,000 \text{ tonnes CO}_2/\text{month}$$

California (CAISO, 80 gCO₂/kWh):

$$= 58,400 \text{ tonnes CO}_2/\text{month (60\% less)}$$

Appendix B: Data Sources & Citations

GPU Specifications

- **NVIDIA H100:** [NVIDIA H100 Datasheet](#)
 - 989 TFLOPS (FP16 Tensor Core)
 - 700W TDP
 - 3.35 TB/s memory bandwidth
- **EpochAI ml_hardware.csv:** GPU specs across vendors (AMD, Google TPU, Amazon Trainium)

Training FLOPs by Model

- **GPT-4:** 2.1×10^{25} FLOPs - [EpochAI Parameter Database](#)
- **Llama 3.1 405B:** 3.8×10^{25} FLOPs - [Meta Llama 3.1 Report](#)
- **Claude 3.5 Sonnet:** 3.6×10^{25} FLOPs - EpochAI estimate
- **DeepSeek-V3:** 2.8×10^{24} FLOPs - [DeepSeek Technical Report](#)
- **GPT-4.5, Claude Sonnet 4, Gemini 2.0 Pro:** Estimates based on company announcements and compute trends

Model FLOPs Utilization (MFU)

- **Meta Llama 3.1:** 38-43% MFU - [Meta Llama 3.1 Report, Fig. 5](#)
- **Default MFU (0.35):** Conservative estimate for frontier model training at scale

Inference Throughput

- **vLLM benchmarks:** [vLLM Performance Data](#)
- **Token generation rates:** Memory-bandwidth calculations validated against NVIDIA TensorRT-LLM benchmarks

Reasoning Overhead

- **OpenAI o1:** "Uses chain-of-thought reasoning that can consume 10× or more tokens" - [OpenAI o1 Blog](#)
- **OpenAI o3:** High-complexity tasks showing 100-1000× token usage - [OpenAI o3 Announcement](#)

Economic Data

- **CapEx per GW:**
 - Validated (\$35.8B): Bottom-up analysis from component costs
 - EpochAI (\$44B): [EpochAI Datacenter Cost Model](#)
 - Stargate (\$50B): Derived from \$500B / 10 GW announcement
- **Ornn Spot Pricing:** \$2.19/H100-hour - [Ornn Index](#) (December 2025)
- **API Pricing:** OpenAI, Anthropic, Google published rates (blended to \$0.60/M tokens)

Regional Power & Carbon

- **Power costs:** State utility commission filings, EIA commercial rates (2024-2025)
- **Carbon intensity:** Grid CO2 Intensity Details
 - ERCOT, PJM, MISO, CAISO: Grid operator sustainability reports (2024)
 - National average (384 gCO2/kWh): EPA eGrid 2024

Infrastructure Efficiency

- **PUE benchmarks:**
 - Google: 1.10 (2024 datacenter average)
 - Meta: 1.09 (2024 fleet average)
 - Microsoft: 1.18 (2024 Azure regions)
 - Industry standard: 1.2 (Uptime Institute 2024 survey)
- **IT overhead (1.14):** NVIDIA DGX GB200 NVL72 specifications (networking, storage, management)

Water Usage

- **1.5 L/kWh:** Modern closed-loop liquid cooling systems
 - Crusoe datacenter specifications
 - FluidStack immersion cooling data
 - Industry average (traditional): 3-5 L/kWh

GPU Lifetime

- **48-60 months:** Industry standard for AI accelerator replacement cycles
 - Economic obsolescence vs. physical failure
 - Validated through hyperscaler procurement cycles