Standardizing Compute for Futures Markets

Proposal for an Effective Compute Hour (ECH) Benchmark and Cash-Settled Futures

Prepared by: Asher Vandevort (252 Capital / 252.capital)

Version: November 17, 2025 (Draft 5)

Executive Summary

Compute is becoming the most volatile, fast-depreciating input in modern industry. The current infrastructure imbalance is defined by a massive **technological risk basis**, the risk that a purchased GPU is made obsolete by a newer, more efficient chip, and a **financial risk basis**, the volatility of spot pricing and scarcity. Top-tier GPUs age on 24–36 month cycles; every idle hour is value that never returns. Yet pricing remains frustratingly bilateral and opaque, forcing firms to over-commit capital or face crippling spot market instability. This paper proposes a practical, low-friction path to financialize compute responsibly: **standardize a performance-anchored unit (Effective Compute Hour, or ECH)**, publish a robust hub-based index, and list cash-settled futures. This approach echoes the critical step in energy and agricultural history where standardization unlocked massive capital pools. The ECH path mirrors these mature commodity and power markets, emphatically moving beyond nominal "GPU hours" to emphasize **performance, time, and location**; to create a unified, transparent language for hedging, financing, and policy engagement.

Bottom line: The market structure is sound if we transition pricing away from hardware inputs and toward **throughput** (e.g., tokens/sec on benchmark reference workloads) by **time block** and **location**. Futures must be **cash-settled** to an independently governed index, with clearly defined **Quality Grades** (e.g., SXM NVSwitch vs. PCIe Ethernet) to minimize basis risk and ensure fungibility of performance.

1) How Futures Markets Bring Buyers and Sellers Together

In grain and oil, standard contracts created shared language: size, grade, and delivery point. Compute today resembles a pre-benchmark era: fragmented SKUs, inconsistent performance, and no common unit.

A standardized **Effective Compute Hour (ECH)**, defined as **performance-normalized throughput over a time block**, can serve the role of the *bushel* or *barrel*.

- **ECH (Concept):** One ECH equals delivering a minimum verified throughput on an agreed reference workload during a one-hour block at a defined hub, with specified cluster/fabric characteristics.

$$ECH = \frac{Throughput_{Config}}{Throughput_{Base}} \cdot Hours$$

Why ECH Works:

1. **Performance-Anchored:** Hedges what buyers truly pay for (tokens/sec), not just the label-hour of a chip.
2. **Time-Boxed:** Treats compute like a non-storable commodity (akin to electricity), enabling peak/off-peak pricing.
3. **Hub-Based:** Normalizes to locational nodes where liquidity can concentrate.

2) Standardization, Hubs, and Quality Grades

Even cash-settled contracts reference a physical or logical **Hub**. For compute, the hub is a latency/interconnect cluster where many providers aggregate capacity. We launch with a **Trifecta** of logical hubs: **US-West, US-Central, and US-East**, to reflect critical latency zones.

Quality is defined by **Grades** to minimize quality basis risk:

Quality Taxonomy (Grade)	Description (Quality Ladder)	Example Hardware/Fabric
Grade A	High-density training cluster, emphasizing low-latency fabric.	≥8 H100/B-series SXM, NVSwitch/IB ≥3.2 Tb/s aggregate
Grade B	Cost-optimized training/large-scale inference cluster.	≥8 H100/B-series PCIe, IB/Ethernet with defined minima

Each other region/provider/hardware class then trades as a **basis** (premium/discount) to the anchor.

3) Spreads, Offsets, and Basis Trades in Compute

Once a benchmark exists, basis markets naturally arise, allowing market participants to hedge **relative** exposure:

- **Regional Spreads:** East vs. West (reflecting latency, power costs).
- **Hardware Spreads:** Grade A vs. Grade B (reflecting fabric and cluster guarantees).
- **Temporal Spreads:** Near-month scarcity vs. far-month abundance.

4) Price Discovery as Coordination

The forward **ECH curve** signals **where and when** scarcity bites and where investment should flow (model releases and new features). Futures transform bilateral opacity into public signals that aid budgeting, siting, and scheduling, a function critical to energy markets regulated by FERC.

5) Why “the Same H100” Prices Differ

The price difference observed between nodes using ostensibly similar GPU hardware (e.g., AWS p5 vs. Azure ND H100 v5) is not random noise, but a reflection of the non-fungible **bundle** of performance characteristics and services bundled with the chip that the ECH must normalize out. You are not simply buying a GPU-hour; you are buying three critical layers: the **Fabric/topology**, the **Node composition**, and the **Commercial scaffolding**.

- **Fabric/Topology:** This is the most critical differentiator. High-speed, high-bandwidth interconnects like NVSwitch and Infiniband (IB) are fundamental to multi-GPU training speed, dictating collective communication overhead. A Grade A contract demands specific fabric minimums because a training job involving 100 billion parameters will stall catastrophically if the interconnect's latency and throughput are insufficient. Performance, and therefore the true $/throughput, diverges wildly even when the core GPU labels match. **The ECH unit is specifically designed to collapse this performance delta when priced to actual, verified throughput.**
- **Node Composition and Commercial Scaffolding:** Node composition includes memory (HBM), local storage (NVMe IOPS), and core-to-core latency. Commercial scaffolding involves pricing mechanisms that create implied forward contracts, such as long-term commitment programs (Enterprise Discount Program / Microsoft Azure Consumption Commitment) or reserved instance capacity. The ECH Index strips away these financial wrappers to reveal the true underlying value of the performance being delivered, creating a clean, exchange-tradable benchmark.

6) Stress-Testing Feasibility: Where a Compute Futures Market Can Fail

To succeed, the market must address the following critical risks:

- **Commodity Definition:** Solved by using the **Performance-anchored ECH** unit.

- **Delivery Complexity:** Solved by **Cash Settlement** to a published index.
- **Fast Tech Cycles:** Solved by defining **Grades** based on functional characteristics (e.g., fabric speed) rather than specific chip models, allowing for substitution.
- **Index Manipulation:** Addressed by **IOSCO-Grade Governance** and transparency (Section 8).

7) Efficiency, Energy, and Societal Utility — "Tokens per Dollar per Watt"

The ECH market framework provides the transparency necessary to incentivize energy efficiency and steer compute usage toward high-utility workloads.

7.1) Why this Metric

Our North-star metric is **Tokens per Dollar per Watt (TP$·W)**, the practical synthesis of cost, energy, and capability. Optimizing this aligns provider incentives (lower opex/capex), buyer incentives (lower $/token), and public goals (lower energy intensity per useful output). As Microsoft CEO Satya Nadella has stated, the future of AI infrastructure is optimizing for $TP\$ \cdot W$, anchoring this market focus in industry leadership and current supply constraints.

7.2) Definitions & Measurement (Refer to Appendix D)

We define the core ratios for legibility, which will be published alongside the ECH index price:

- **TPW (Tokens per Watt):** Measures energy efficiency; $TPW = \frac{T}{W} \cdot Hours$
- **TP$ (Tokens per Dollar):** Measures cost efficiency; $TP\$ = \frac{T}{\$ \cdot Hours}$ (all-in run cost).
- **TP$·W (Tokens per Dollar per Watt):** The ultimate efficiency frontier metric; $TP\$ \cdot W = \frac{T}{\$ \cdot W \cdot Hours}$.

Normalization includes PUE (Power Usage Effectiveness) assumptions, regional energy price inputs, and carbon-intensity tags (gCO_2e/kWh) published by Hub and Grade.

7.3) Scaling Laws and Compute Growth

Empirical scaling laws show that performance/loss follows power-law relationships with compute, model size, and data (Kaplan et al.; Chinchilla). This justifies the ECH's performance-anchored design, as every efficiency gain (a rise in $TP\$ \cdot W$) translates into tangible economic returns given the diminishing returns per marginal compute $log_{intelligence} \sim log_{compute}$.

7.4) Jevons Paradox and Market Design

Efficiency lowers the unit cost of compute, which, rather than reducing overall consumption, often unlocks massive new demand that can increase total aggregate energy consumption as delineated in the well-documented **Jevons Paradox**. In compute, this means that as $TP\$ \cdot W$

improves, the marginal cost of training new, ever-larger foundation models drops, leading to a surge in total tokens consumed. To address this inevitable surge, markets must be designed to channel the throughput toward socially valuable ends and cleaner energy sources.

- **Mitigations via Market Design:** This is not a problem for regulation alone. We propose market-based solutions, primarily through optional **Carbon-aware hubs/tenors** and **peak/off-peak blocks**. By aligning contract delivery periods to regional clean-energy windows (e.g., high solar/wind output), we reward clean supply without mandating technology choices. This creates a powerful financial incentive for demand and response style discounts and helps internalize the carbon externality, a core public goal. This market structure allows participants to trade the carbon footprint of their compute as a tradable characteristic via Value Overlays (Section 9).

7.5) Policy & Societal Utility

Higher $TP\$ \cdot W$ serves a critical social function by widening access (more tokens per school/dollar; more research per grant), ensuring the scarcity barrier does not solely restrict AI development to a few highly capitalized firms. The market structure provides essential, regulated coordination points vital for public confidence and orderly growth:

- **Regulatory Alignment:** Voluntary disclosures and structure should align with **FERC/RTO** constructs. This is necessary because the fundamental nature of compute hinders on locational, time sensitive, and non-storable parameters which parallels the physics and finance of electricity grid management. This alignment provides a blueprint for managing capacity surges and ensuring regional stability.
- **Index Governance:** Adoption of **IOSCO-style** benchmark governance is non-negotiable for the Compute Index. This framework ensures transparency, reliability, and non-manipulability, which are preconditions for widespread financial and regulatory adoption of the ECH as a primary economic indicator.

8) Index Methodology & Contract Spec (IOSCO-Style)

The index is cash-settled to the published **Hub ECH Index** (VWAP of validated contributors).

- **Contract Addendum:** TPW and TP$·W$ are reported as reference statistics for each contract month (non-binding for settlement but part of the official market data release).
- **Data Lineage:** Contributor data includes posted rates, audited fills, and cross-checks.
- **Governance:** Independent oversight, VWAP calculation (with trimming for outliers), and publicly available fallback methodologies.

9) Ancillary Services and Value Overlays

The core ECH price is a pure, performance-normalized commodity price. Optional wrappers and certifications provide additional value without distorting the core price signal.

- **Ancillary Services:** These are optional add-ons to the futures contract to manage specific risks, such as **Preemption Insurance** and **Checkpoint Credits** (for managing interruptible capacity).
- **Value Overlays:** Certifications like **GreenCAC** (certifying clean energy usage) or **OpenCAC** (certifying capacity for open-source models) avoid mandated change to the index but reward specific, socially valuable behaviors via voluntary premium payments.

10) Hedge Efficacy (Backtest Outline)

The backtest objective is to show that ECH futures reduce realized $/throughput variance compared to spot prices and quasi-forwards like Capacity Blocks. Basis decomposition will break down hedge error into **Quality basis**, **Locational basis**, **Calendar basis**, and **Energy-price basis**.

11) Acknowledgment of Other Participants & How This Proposal Differs

We recognize ongoing efforts by Ornn, OneChronos×Auctionomics, and Compute Exchange. This paper is differentiated/superior due to: Performance-based Commodity Definition (ECH), Quality Grades & Conversion Factors, IOSCO-Grade Index Governance, and a Regulatory Path to Listed Futures.

12) Regulatory Considerations

Plan for DCM/SEF (designated contract markets / swaps execution facilities) pathways, clearing arrangements, position limits, reporting, and market surveillance, aligning with CFTC-style oversight for cash-settled commodity-like contracts.

13) Roadmap

Publish ECH spec (Quality Ladder) and methodology. Collect contributor commitments. Run shadow indices. Trade bilateral NDFs/forwards against the index. Publish backtest efficacy. Migrate to listed, cleared futures.

Appendices

Appendix A — Draft Contract Spec (Illustrative)

Parameter	Detail
Contract	US-East ECH (Grade A), Monthly Strip
Unit of trading	100 ECH (100 verified one-hour blocks)
Settlement	Cash-settled to official **US-East Grade-A ECH Index**

Final settlement window	Last five business days of month

Appendix B — Quality Ladder (Illustrative)

Grade	Hardware	Fabric/Interconnect
Grade A	≥8 H100/B-series **SXM**	NVSwitch or IB ≥3.2 Tb/s aggregate
Grade B	≥8 H100/B-series **PCIe**	IB/Ethernet with stated minima

Appendix C — Reference Workloads (Illustrative)

- **Training:** LLM pretrain & finetune (e.g., MLPerf-aligned), vision transformer, diffusion.
- **Inference:** LLM tokens/sec across context lengths; batch-size grids.

Appendix D — Efficiency Metrics and Calibration

This appendix details the measurement methodology.

D.1: Efficiency Metrics (TP\$·W\$ Ratios)

The core ratios are derived from measured tokens (T), total all-in run cost in dollars (\\\$), and average power consumption in Watts (W).

1. **Tokens per Watt (TPW):** Measures raw energy efficiency.

$$TPW = \frac{T}{W} \cdot Hours$$

2. **Tokens per Dollar (TP\$):** Measures raw cost efficiency.

$$TP\$ = \frac{T}{\$ \cdot Hours}$$

3. **Tokens per Dollar per Watt (TP\$·W\$):** The synthesized efficiency frontier metric.

$$TP\$ \cdot W = \frac{T}{\$ \cdot W \cdot Hours}$$

D.2: Calibration and Normalization

- **PUE Treatment:** Power Input $W\$$ must be multiplied by a standardized PUE (Power Usage Effectiveness) assumption (e.g., PUE=1.2) to account for facility overhead.
- **Carbon Intensity Calculation:** Emissions per unit of output gCO_2e/T is calculated using regional marginal emissions data gCO_2e/kWh published alongside the index.

- **Worked Example (Normalization):** A dedicated section will normalize an AWS p5 node vs. an Azure ND H100 v5 node to $/ECH, demonstrating how the price delta collapses when normalized by throughput rather than nominal GPU-hours.

Appendix E — Service Parameters Matrix

Parameter	Classification (ECH)
Hardware Type	**Contract-Defining (Grade)**: Defines the Quality Ladder (A, B, C).
Node Fabric/Topology	**Contract-Defining (Grade)**: The single most important factor for performance, defined by the Grade minimums.
Location/Latency	**Contract-Defining (Hub)**: Defines the Hub (US-West, US-East).
Throughput (Tokens/sec)	**Contract-Defining (ECH Unit)**: The variable being standardized and priced.
Reliability/Uptime	**Basis/Optional Add-on (Ancillary Service)**: Managed via Checkpoint Credits/Insurance.
Job Preemption Risk	**Basis/Optional Add-on (Ancillary Service)**: Managed via Preemption Insurance.
Carbon Footprint	**Basis/Optional Add-on (Value Overlay)**: Managed via GreenCAC certification.