Tribunal Session Summary

Session: conclave-20260724-204633-which-specific-inputs-to-0ee3

Depth: T6

Advocates: 6 | Judges: 3 | Cost: $11.9388 | Time: 40m 11s

THE PROMPT: Which specific inputs to AI compute are supply-constrained right now, and what observable evidence separates a real shortage from allocation lag or vendor messaging?

THE RULING: The primary supply-constrained inputs to AI compute right now are **High Bandwidth Memory (HBM 3E/4)**, **power infrastructure components (specifically large transformers, switchgear, and grid interconnection)**, and **advanced packaging capacity (TSMC CoWoS-L)**. Real shortages are distinguished from allocation lag or vendor messaging by the presence of **persistent multi-year lead times that fail to compress despite capital expansion**, **price premiums impacting "unwilling payers" (spillover costs to unrelated sectors)**, and **verified physical capacity gaps** that cannot be resolved by simple queue management, rather than purely temporal allocation queues.

The Prompt

Which specific inputs to AI compute are supply-constrained right now, and what observable evidence separates a real shortage from allocation lag or vendor messaging?

The Ruling

The primary supply-constrained inputs to AI compute right now are **High Bandwidth Memory (HBM 3E/4)**, **power infrastructure components (specifically large transformers, switchgear, and grid interconnection)**, and **advanced packaging capacity (TSMC CoWoS-L)**. Real shortages are distinguished from allocation lag or vendor messaging by the presence of **persistent multi-year lead times that fail to compress despite capital expansion**, **price premiums impacting "unwilling payers" (spillover costs to unrelated sectors)**, and **verified physical capacity gaps** that cannot be resolved by simple queue management, rather than purely temporal allocation queues.

The strongest evidence points to a multi-layered bottleneck: Power infrastructure is the most durable and geographically specific constraint gating deployment; HBM is the most structurally tight semiconductor constraint due to physical wafer intensity and confirmed sold-out status through 2027; and CoWoS packaging remains a physical constraint during 2026 that is transitioning toward an allocation queue. Finished leading-edge logic (GPUs) is co-tight but not the primary binding constraint, as older generations (A100/H100) have largely normalized.

Core Analysis

The Three Binding Constraints

1. Power Infrastructure: The Primary Deployment Gate The most durable bottleneck limiting AI compute is not the silicon itself, but the electrical infrastructure required to energize it.

- **Specific Inputs:** Large Power Transformers (LPTs), High-Voltage Switchgear, and Transmission Interconnection capacity.
- **The Evidence of Real Shortage:**
  - **Physical Capacity Gap:** GE Vernova filings indicate a firm backlog of ~44 GW of heavy-duty gas turbines against an annual manufacturing capacity of only ~10 GW/year. The math dictates a multi-year wait regardless of price.
  - **Lead Times:** U.S. Department of Energy and NREL assessments document lead times for transformers extending from months to 1–3 years, with some large power units taking longer. Bloomberg and Reuters reporting corroborates 2–5 year lead times for critical grid equipment.
  - **Revealed Preference:** Developers have announced approximately 56–101 GW of behind-the-meter (on-site) natural gas generation to bypass grid bottlenecks. Building independent power plants is economically irrational compared to grid power and serves as a "cost you pay to escape a constraint" that

cannot be faked.

- **Market Signals:** The PJM 2025/26 capacity auction cleared at $269.92/MW-day versus $28.92/MW-day previously—roughly a ninefold increase—indicating scarcity pricing for reliable capacity, even if influenced by rule changes.

2. High Bandwidth Memory (HBM 3E/4): The Structural Semiconductor Constraint Memory bandwidth is the defining performance characteristic for current AI accelerators, making HBM the tightest semiconductor input. This is a structural shortage driven by physics, not just allocation.

- **Specific Inputs:** HBM3E and HBM4 memory stacks, and the underlying DRAM wafer capacity.
- **The Evidence of Real Shortage:**
  - **Supplier Disclosures:** Primary sources (SK Hynix, Micron earnings calls) confirm "sold-out" status for 2024 and 2025, with forward visibility into 2027. Pricing agreements are locked in well in advance.
  - **Spillover ("Unwilling Payer" Test):** The shortage has broken containment. Because HBM requires 3–5 times the wafer area of equivalent commodity DRAM due to vertical stacking (TSV) and thinning, its production cannibalizes general DRAM supply. This has caused price inflation of 50–125% in commodity DRAM markets (PC, automotive, handset). Buyers in these unrelated sectors—who have no stake in AI and every incentive to deny a shortage—are bearing the cost. This contagion is an unfakeable signature of physical scarcity.
  - **Process Physics:** Manufacturing adds ~19 extra process steps vs. standard DRAM (TSVs, stacking), limiting ramp speed and imposing a hard ceiling on how fast new capacity can come online.

3. Advanced Packaging (CoWoS-L): A Transitioning Physical Constraint The assembly of logic and memory into a functional accelerator is currently constrained, though capital investment is rapidly expanding capacity.

- **Specific Inputs:** TSMC CoWoS-L capacity and Outsourced Semiconductor Assembly and Test (OSAT) capacity.
- **The Evidence of Real Shortage:**
  - **The Math of the Ramp:** Capacity is ramping aggressively from ~35-40k wafers per month (wpm) in 2024 to ~65-80k wpm in 2025, targeting 120–140k wpm by end-2026. However, demand is estimated near 1 million wafers for 2026.
  - **Linear vs. Exit Rate Arithmetic:** A critical distinction made in the deliberation is that annualizing a *December exit rate* (140k wpm *12 = 1.68M) overstates supply. When calculating a* linear ramp average* from ~80k wpm to 130k wpm, the actual 2026 supply is closer to ~1.26M wafers. Against demand of ~1.0–1.38M and accounting for yield losses, the margin is razor-thin or negative for high-density flows (CoWoS-L) for much of the year.
  - **Customer-Level Proof:** Google's reported ~47% shortfall in TPU production—despite being a top-tier customer with effectively unlimited capital—proves the constraint is physical, not merely "monopolistic allocation." If capacity were sufficient and only hoarded by Nvidia, Google would have cleared the market by paying a premium. The capacity did not exist to buy.
  - **Status:** This is classified as a physical shortage during 2026 (especially for CoWoS-L) transitioning toward an allocation queue by late 2026/early 2027.

Distinguishing Real Shortages from Allocation Lag

The deliberation validated two primary frameworks to separate physical scarcity from market manipulation or temporary queues:

1. **The Unwilling-Payer Test (Claude Opus 5):** A constraint is physically real if a party with *no stake in the scarcity narrative* is measurably paying a cost. The DRAM price hikes hitting automotive and PC buyers,

and the transformer delays hitting industrial utilities, satisfy this test. A shortage that only exists in NVIDIA's earnings call or AI cloud pricing does not.

2. **The Binding Scarcity Test (GPT-5.6 Sol):** A shortage requires an **economic signal** (premiums, sold-out status) AND a **mandatory physical-causation gate** (tool throughput limits, yield constraints, factory build times). "Sold out" statements alone can mean the vendor is managing allocation conservatively; "sold out" combined with 3-year tooling lead times (HBM) or 5-year factory build times (Transformers) indicates a physical limit.

The Status of Leading-Edge Logic

While leading-edge logic wafers (3nm/2nm) are tight—CNAS analysis suggests demand exceeds capacity by 25–30%—they are **co-binding** rather than the primary bottleneck. The evidence shows that logic lead times for older generations (H100/A100) have collapsed, and finished accelerators often sit waiting for HBM or packaging, not the other way around. The serial nature of the supply chain means logic alone does not determine output.

How We Got Here

Scorecard

Judge	GPT-5.6 Sol	Grok 4.5	Gemini 3.1 Pro	Magistral Medium	DeepSeek V4 Pro	Claude Opus 5	Winner
DeepCogito v2.1 671B	ADOPT	ADOPT	REJECT	REJECT	ADOPT	ADOPT	Claude Opus 5

Convergence Assessment

Convergence was **epistemic** (evidence-driven). Position shifts were triggered by specific falsifications—such as the exposure of Magistral Medium's outdated data points, the math corrections on CoWoS run-rates, and the application of the "unwilling payer" test to transformer backlogs. Models did not shift positions due to social pressure or sycophancy; they shifted when their arithmetic or data vintage was demonstrably wrong.

Closing Statements

Claude Opus 5 pivoted from a confident ranking to a refined position emphasizing the serial link between memory and power. They noted that even if HBM mix shifts stop, conventional DRAM wafer-start growth (6–8%) is insufficient to restore historical bit growth (22%), proving a structural lag. They stressed that "sold out" is an order-book state, whereas physical throughput (wafer starts) is the unfakeable metric. They ended by proposing a tiebreaker: analyzing hyperscaler PP&E disclosures to see if Construction-in-Progress (CIP) is growing faster than assets placed in service (indicating power binding) or vice versa.

Gemini 3.1 Pro evolved from a binary taxonomy (Real vs. Artificial) to a multi-layered constraint matrix. They formally conceded their error on CoWoS, acknowledging that Google's inability to secure volume proved a physical aggregate shortfall, not just monopolistic capture. They concluded that the stack is gated by "Megawatt Queue Displacement"—a macroeconomic shift where compute deployments move out of prime geographies purely to secure power windows. They retained their "Cross-Market Cannibalization Index" as the strongest surviving evidence for HBM's physical reality.

GPT-5.6 Sol refined their hypothesis to acknowledge the "hybrid" nature of CoWoS (mix-specific scarcity + allocation) while maintaining high confidence in HBM and power. They emphasized that "capacity" is not fungible—installed wafers, qualified-good output, and finished shipments are different denominators. They concluded that the analytical obligation is to identify scarcity at the specific level of qualified product, location, and customer access, rejecting both "shortage maximalism" and reflexive skepticism.

Reality Checks

Claim	Challenge	Result
Magistral Medium: The most supply-constrained inputs are NVIDIA H100 and A100 GPUs.	**Claude Opus 5 / GPT-5.6 Sol:** A100 is 2020 silicon; H100 rental prices have fallen 64–75% from peaks; secondary market is saturated.	**Thesis Abandoned.** Magistral Medium conceded the entire position was based on 2023/24 data and acknowledged H100/A100 are no longer the binding constraint.
Gemini 3.1 Pro: CoWoS is constrained by "monopolistic allocation capture," not a total industrial shortfall.	**DeepSeek V4 Pro / DeepSeek V4 Pro:** Priority rationing is the signature of scarcity, not proof of abundance. If capacity existed, Google would have cleared the market. The math (linear ramp) shows supply < demand.	**Logic Refuted.** B conceded the economic logic was flawed and that CoWoS is a physical shortage, not just a queue.
Claude Opus 5: CoWoS is "allocation, resolving" because exit-rate arithmetic shows surplus.	**Gemini 3.1 Pro / Grok 4.5:** You cannot annualize a December exit rate; linear ramp math shows a razor-thin or negative margin. Outsourcing to Amkor is proof of physical capacity stress, not surplus.	**Argument Revised.** E accepted the math correction and reclassified CoWoS as a physical constraint for high-density flows.

Claim	Challenge	Result
DeepSeek V4 Pro: Logic dies are "allocation lag, not genuine shortage."	**GPT-5.6 Sol / CNAS Data:** N3/N2 utilization >120%; demand exceeds capacity 25–30%. Logic is co-tight with packaging/memory.	**Position Refuted.** A conceded logic is tight and co-binding, dropping the "allocation lag" label.

Dissenting Opinions

No dissenting opinions were filed in this session.

Glossary

Term	Definition
Advanced Packaging (CoWoS)	Chip-on-Wafer-on-Substrate, a 2.5D/3D packaging technology required to stack High Bandwidth Memory (HBM) beside a GPU logic die on a silicon interposer.
Allocation Lag	A temporary mismatch where orders exceed available supply due to logistics, scheduling, or qualification, but physical capacity exists or is imminent to resolve the queue.
Binding Scarcity Test	A framework holding that a real shortage requires both an economic signal (price premiums) and a mandatory physical gate (tool limits, factory lead times) preventing supply expansion.
DRAM (Dynamic Random-Access Memory)	Standard volatile memory used in PCs, servers, and mobile devices; distinct from HBM but competing for the same underlying wafer manufacturing capacity.
Grid Interconnection	The process of connecting a power generation source (or data center) to the electrical grid, involving studies, upgrades, and approvals that can take years.
HBM (High Bandwidth Memory)	High-performance RAM with a wide interface used in AI accelerators; requires ~3-5x the wafer area of standard DRAM to manufacture.
Interconnection Queue	A list of projects waiting for grid connection approval and capacity; often inflated by speculative requests ("optionality").
Leading-Edge Logic	The most advanced semiconductor manufacturing nodes (e.g., 3nm, 2nm) used to produce the central processing cores of AI accelerators.
Large Power Transformers (LPT)	Critical electrical equipment that steps down high-voltage transmission power to lower voltages suitable for data centers; lead times range 1–5 years.
TSV (Through-Silicon Via)	A vertical electrical connection passing completely through a silicon wafer or die; essential for 3D stacking in HBM and advanced packaging.
Unwilling-Payer Test	A framework positing that a shortage is physically real if it imposes costs on parties with no stake in the AI narrative (spillover effects).

How The Conclave Works

The Conclave is a structured multi-model deliberation system. Instead of asking one AI model a question and trusting its answer, the Conclave convenes a panel of independent models, forces them to argue, and subjects their conclusions to judicial review. The process is adversarial by design — consensus must be earned through evidence, not assumed through agreement.

Advocates are independent AI models (e.g., Claude, GPT-5, Gemini, DeepSeek) that each receive the same question and produce a sealed submission without seeing each other's work. Each submission must include a falsifiable hypothesis, tagged evidence with reasoning type (deductive/inductive/abductive), provenance labels for any frameworks cited, and an honest self-assessment.

Challenges follow submissions. Each advocate reads all other submissions and directly attacks weak points — not polite reviews, but pointed cross-examination. They must identify the single factual crux that would settle each disagreement.

Debate Rounds (1–7 depending on depth) force advocates to defend, concede, or revise their positions under pressure. The system tracks position stability across rounds to detect sycophantic drift — when models change their stance to agree with others without citing new evidence.

Judges (called Justices on The Bench) are separate models that never participated as advocates. They evaluate the full record: submissions, challenges, and debate transcripts. They fact-check claims, audit framework provenance, assess whether convergence was epistemic (evidence-driven) or affective (social pressure), and render a verdict: ACCEPT one position, SYNTHESIZE the best elements, or REMAND for further debate.

Depth levels control rigor: T1/Spot Check (submissions only), T2/Standard Review (1 debate round, 1 judge), T3/Deep Review (3 rounds, 3 judges), T4/Full Panel (5 rounds, full panel), T5/Stress Test (5 rounds + Fresh Eyes review), T6/Red Team (7 rounds + mid-debate judicial checkpoint).

The system is deterministic code — it cannot be sycophantic. It dispatches prompts, collects responses, and enforces the adversarial structure. The models argue; the code referees.