SENIOR ENGINEER'S BRIEFING

The Token Arbitrage: Inference Market 2025

Navigating the Speed War, the Price War, and the hidden trends shaping the future of LLM infrastructure.

The Economic Shift

99%

Price Reduction

From $20/1M tokens (OpenAI 2024) to $0.20/1M tokens (Llama-3 2025).

?

The Catch

Reliability is the new currency. We benchmark the top 5 serverless providers to find the balance.

Market Bifurcation: Two Wars

The Speed War

Groq vs. Cerebras

The battle for the lowest latency and highest throughput. Critical for voice agents and real-time applications.

> Metric: Time To First Token (TTFT)

The Price War

DeepInfra vs. Hyperbolic

The race to the bottom on cost. Providers use consumer GPUs and spot instances to offer commodity pricing.

> Metric: Cost per 1M Tokens

Round 1: Speed Demons

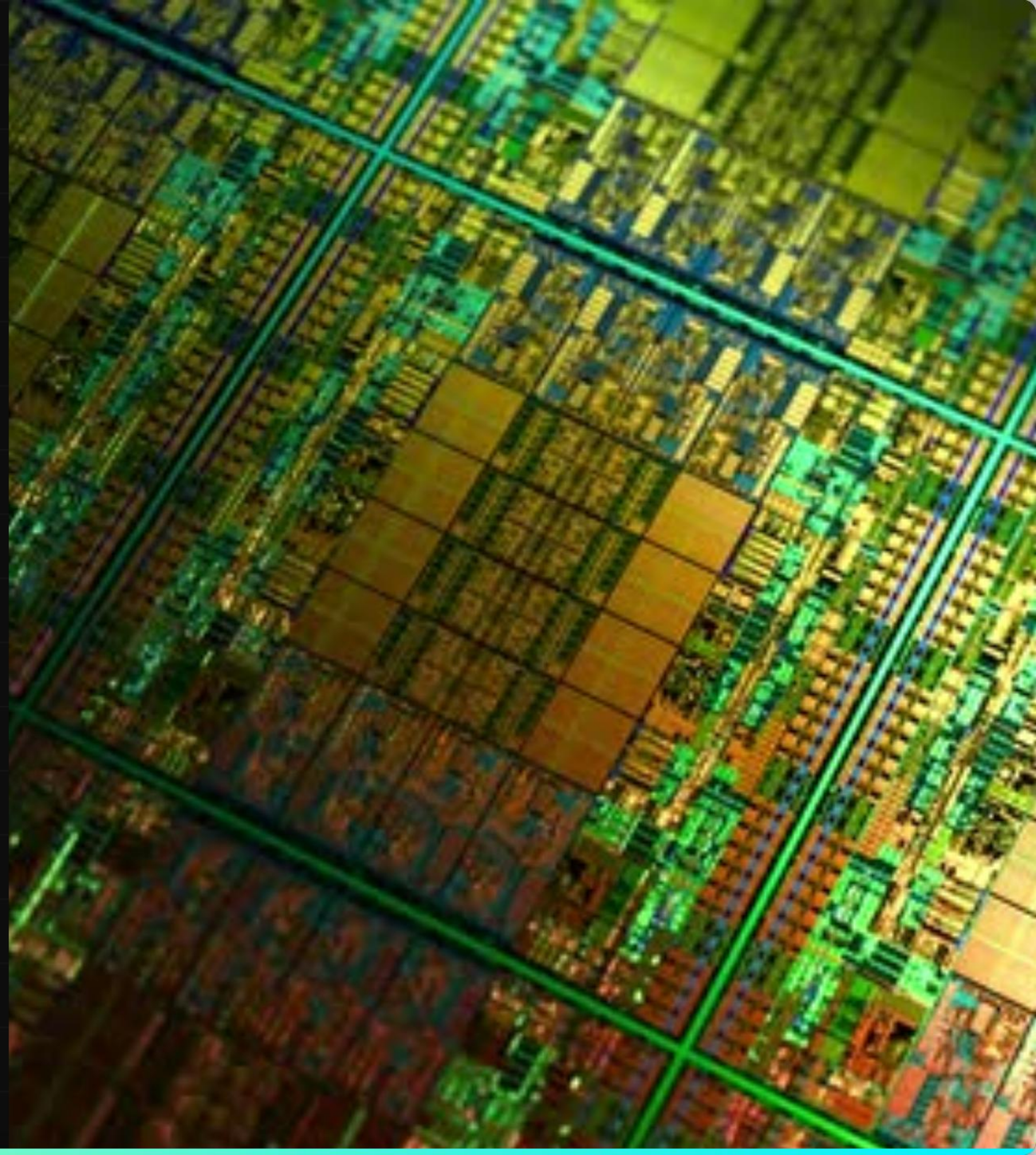
Groq (LPU)

Wins on **Single-User Latency**. The specialized LPU architecture delivers ~150ms TTFT.

Cerebras (Wafer Scale)

Wins on **Massive Throughput**. While Groq is fast for one, Cerebras is fast for 10,000 concurrent users.

> Verdict: Use Groq for Voice. Use Cerebras for Batch.



Round 2: The Price Floor

DeepInfra

Winning by running on older consumer GPUs (3090/4090 clusters). Effective, but "noisy neighbor" issues can occur.

Hyperbolic

Decentralized compute. Low cost ($0.28/1M), but reliability varies heavily based on node availability.

⚠️ Risk: Silent Rate Limiting

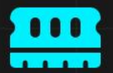
Q4 2025 Benchmarks (Llama-3-70B)

Provider	Superpower	Cost / 1M	Latency (TTFT)	Verdict
Groq	Low Latency (LPU)	$0.60	~150ms	**Real-Time Voice**
Cerebras	Throughput	$0.40	~180ms	**Bulk Processing**
DeepInfra	Price Floor	$0.25	~300ms	**Background Jobs**
Fireworks	Developer Exp	$0.50	~200ms	**JSON / Function Calling**

The "Hidden" Trends

Speculative Decoding

Providers are competing on predictive algorithms to draft tokens faster than the main model can verify them.



Context Caching

Drastic cost reductions for RAG. Only pay to upload the system prompt and documents once.

Consumer Hardware

The successful deployment of 3090/4090 clusters proving that enterprise H100s aren't always necessary.

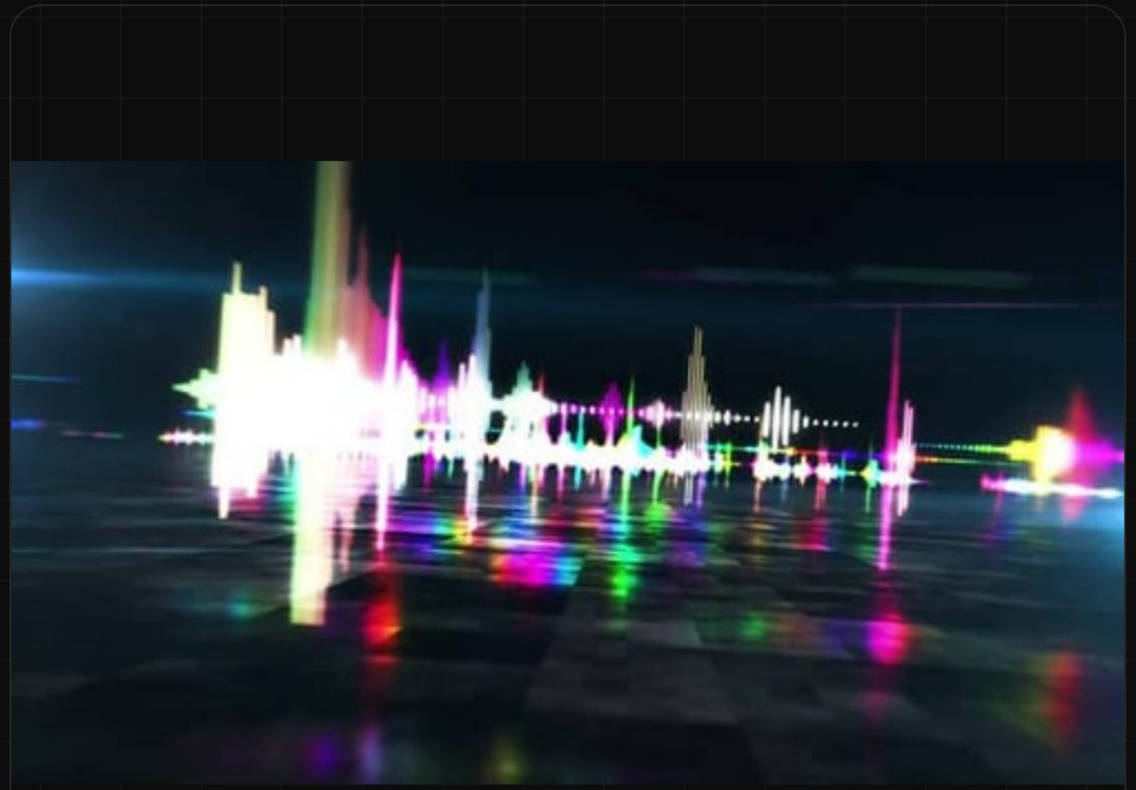
Use Case: Voice Agents

The Challenge

Human perception of latency is ~200ms. Any delay above this breaks the illusion of conversation.

The Winner: Groq

Pay the premium ($0.60). The LPU architecture guarantees the TTFT required for seamless interruption and response.



Use Case: RAG & Bulk

The Challenge

Summarizing 10,000 documents or processing massive vector databases. Latency matters less than cost and throughput.

The Winners

- **Cerebras:** If you need it done fast (high volume).
- **DeepInfra:** If you need it done cheap ($0.25).

Round 3: Developer Experience

The Problem

Generic providers often "break" on complex JSON schemas or function calling, returning malformed text instead of structured data.

The Winner: Fireworks AI

Optimized kernels specifically for structured output.

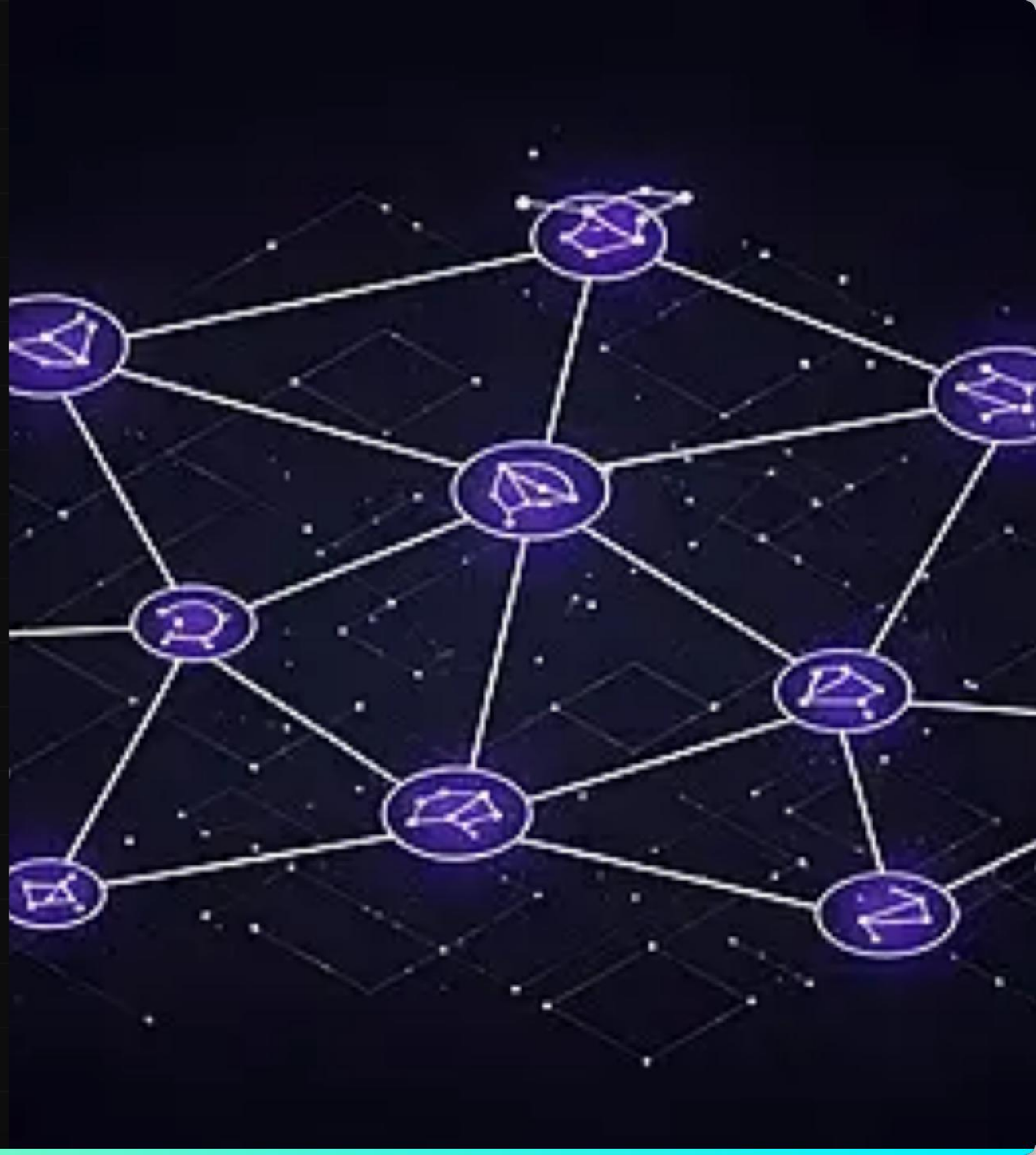
Verdict: **For Agent Logic, use Fireworks.** For Raw Text, use DeepInfra.

The Strategy: Inference Arbitrage

"Do not marry a provider. They are commodities."

```
// LiteLLM Router Configuration
if (task == "voice") return GROQ;
if (task == "background") return DEEPINFRA;
if (task == "agent") return FIREWORKS;
```

Use a proxy router to dynamically switch traffic based on the specific needs of the request.

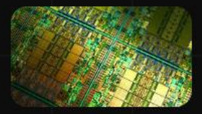


Questions?

Your Slide Deck on LLM Market 2025 is ready.

engineer@briefing.com

Image Sources



https://s3.envato.com/files/21a58144-724a-4a9a-8a65-641990aee6f4/inline_image_preview.jpg

Source: videohive.net

https://images.stockcake.com/public/e/7/1/e719fa8f-69c3-4f77-afbf-085e7933f5ce_large/futuristic-server-room-stockcake.jpg

Source: stockcake.com

https://elements-resized.envatousercontent.com/elements-video-cover-images/41d7c3e0-4fff-4aa4-a12e-b3de77efc616/video_preview/video_preview_0000.jpg?w=500&cf_fit=cover&q=85&format=auto&s=ca54fd7cc7797d1eb35c15ea4daa93c7dda8cf3ddd48a58ddeb98ef222a98399

Source: elements.envato.com

https://www.techtarget.com/rms/onlineimages/data_management-big_data_vs_mobile.png

Source: www.techtarget.com

https://png.pngtree.com/thumb_back/fh260/background/20251030/pngtree-abstract-network-of-glowing-nodes-and-connections-on-a-dark-background-image_20134343.webp

Source: pngtree.com