Etched's Inference Stack, El Segundo's Hardtech Fund, and AI Cost Compression

VC Tech Radar

2026-07-01

Etched's Inference Stack, El Segundo's Hardtech Fund, and AI Cost Compression

By VC Tech Radar • July 1, 2026

This brief highlights a capital-intensive AI hardware thesis, several early-stage teams showing real technical depth or traction, and fresh evidence that inference costs are falling quickly. It also tracks local AI, hiring data from heavy adopters, and where platform risk is still blocking applied AI products.

1) Funding & Deals

- **A new $30M hardtech fund targets earliest-stage industrial tech.** Jakob Diepen announced an oversubscribed $30M fund for the earliest-stage hardtech founders building for critical industries. The launch messaging positions El Segundo as a destination for hardtech builders, and Mark Suster described Diepen as part of LA's early-stage hardtech push. [1, 2]
- **Etched's early-2024 Series A came together around a full-stack inference thesis.** The company says it raised about $100M after circulating a 30-page technical memo, even though every major Valley investor initially passed. The thesis was that modern AI inference required the full system — chip, boards, interconnects, cooling, and rack — not just a chip. Founders Rob and Gavin had dropped out of Harvard; Rob traces the urgency to GPT-4V flagging a tumor in an old photo, and Gavin brought kernel experience from Xnor and Octo. The round later compounded into follow-on support from existing backers. [3]

2) Emerging Teams

- **Screenpipe shows notable open-source traction in local agent perception.** A market map posted to r/SideProject puts Screenpipe

(YC S26) at 19,566 GitHub stars and actively shipping, versus OpenRecall at 2,874, with Microsoft's Recall bundled into Windows and OpenAI reportedly building something similar. The same post argues the stronger wedge is developer-facing agent infrastructure, not privacy-first consumer positioning. [4]

- **Lumbox is building infrastructure for the parts of agent workflows that usually still require hands.** A solo founder says the product combines a real inbox for OTPs and verification links, a credential vault where plaintext never reaches the model, TOTP generation, and an MCP server with 100+ tools. The goal is to let an agent sign up, verify, store credentials, and log back in behind 2FA without human intervention. [5]
- **DVForge is an early traction signal in vertical hardware education.** The founder, a recent electronics engineering graduate, built a LeetCode-style platform for chip design verification and reports 120+ signups in 24 hours, the highest-voted post in the target subreddit, and inbound from a venture scout. The UI is AI-assisted, but the founder says the practice problems come from textbooks and are verified by experienced engineers. [6]
- **Applied AI for SMB lead response is already showing narrow-workflow ROI.** One founder says an AI agent that responds within 60 seconds, qualifies buyers, and books directly onto calendars helped a real-estate brokerage move from 2.3% to 6.1% conversion after five weeks of work on making the handoff to humans sound natural. [7]

3) AI & Tech Breakthroughs

- **Etched is making one of the strongest architectural claims in inference hardware.** The company says its first-generation low-voltage inference technology runs at under half the voltage of other AI chips. It also says its custom interconnect cuts point-to-point latency by more than 5x relative to the Blackwell figure cited in the interview, making "cluster-scale memory" more usable for decode workloads. Etched is building the chip, rack, boards, cold plates, interconnects, and parts of production in-house. [3]
- **Software optimization still looks like the fastest path to more AI capacity.** At RAAIS, ElevenLabs' Angelos Peri said batching, fp8, speculative decoding, and kv-cache compression let the company serve 70x more users on the same GPUs. [8]
- **Local-model usability is improving at the tooling layer.** Hugging Face now lets users filter public models by the hardware they already own; Clement Delangue says that still leaves 800k+ public models that fit on his M5 24GB via llama.cpp. [9]

- **Fusion crossed a public demo milestone.** Vinod Khosla said RealtaFusion powered a lightbulb using electricity harvested directly from WHAM via direct electricity conversion, which he described as the first such public demonstration by a private company. [10]

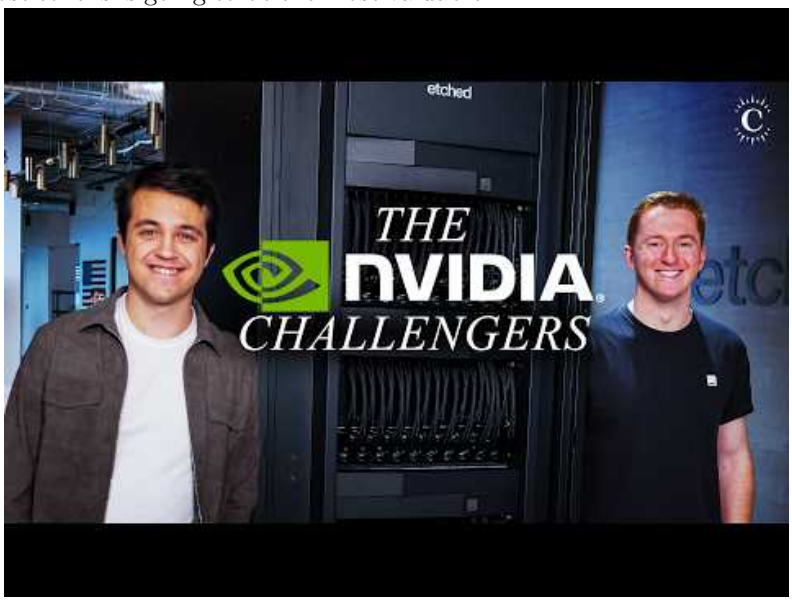
4) Market Signals

- **Inference efficiency is improving quickly at the software and systems layer.** Harry Stebbings says five founders — from 10-person startups to a $200B public company — reported cutting inference spend by 75% or more with little effort, no performance loss, and better latency. Nathan Benaich amplified a similar point from ElevenLabs, arguing that GPU scarcity is increasingly an engineering problem. [11, 8]
- **Local AI is shifting from ideology to cost-control.** Delangue cites a Stanford result that 71.3% of ChatGPT queries could be answered accurately by a local model and argues that a large share of enterprise AI work could run locally relative to frontier API costs, while also reducing dependence on rented models. [9]
- **The labor read-through from AI adoption remains more expansionary than contractionary.** Ramp and Revelio Labs, analyzing more than 21,000 U.S. firms, found that heavy AI adopters grow headcount 10% and entry-level headcount 12% over the two years after adoption, with growth showing up after roughly 6-12 months. The caveat is that these adopters are more technical, higher-paying, and more likely to be venture-backed, and average AI spend is still only about $33.67 per employee per month. [12]
- **Harmonic's Hot 25 remains a useful demand-side watchlist.** ResolveAI reclaimed the top spot, Salient jumped to #2, aaruHQ ranked #3, Simile debuted at #6, and BrainCo_AI was the biggest mover at +12 places. [13]
- **Platform approvals are still a major choke point for applied AI.** One restaurant voice-agent team says its production-ready system was rejected by Clover because the company lacks a policy for third-party AI applications, and separate attempts with Toast and Deliverect also failed to unlock distribution. In this case, integration approvals appear to be the gating issue rather than voice capability. [14]

5) Worth Your Time

- **Invest Like The Best: Etched founders** — the best single source in this set for the inference-first hardware thesis: why they think inference is the largest market, how they recruit "legends," and why they chose chip-plus-rack vertical integration. [3]

"Whoever produces the most tokens is going to be the most valuable company in the world." [3]



The Two Harvard Dropouts Who raised $800M to take on NVIDIA (10:05)

- **Harmonic's Hot 25 Report** — the full ranking behind ResolveAI at #1, Salient at #2, and aaruHQ at #3. [13]
- **Big Technology on AI adoption and hiring** — a concise write-up of the Ramp/Revelio dataset covering 21,000+ firms, 10% headcount growth for heavy adopters, and the 6-12 month lag before benefits show up. [12]
- **Clement Delangue on local AI** — useful if you're tracking how much AI work can move from rented APIs to owned local models, and how Hugging Face is reducing discovery friction for that shift. [9]

Sources

1. X post by @jakobdiepen
2. X post by @msuster
3. The Two Harvard Dropouts Who raised $800M to take on NVIDIA
4. r/SideProject post by u/Terrible-Painter5422
5. r/SideProject post by u/kumard3
6. r/SideProject post by u/Dry-Letterhead7890
7. r/EntrepreneurRideAlong post by u/Warm-Reaction-456
8. X post by @nathanbenaich
9. X post by @ClementDelangue
10. X post by @vkhosla

11. X post by @HarryStebbings
12. Heavy AI Adoption Linked To More Hiring, Not Layoffs, New Data Shows
13. X post by @HarryStebbings
14. r/SaaS post by u/Capital_Act8480