Zero Latency (0.lat)

Distributed AI inference, and the VR network that shares its name

Explainer · June 1, 2026

A Charlottesville-based distributed AI inference startup that routes inference workloads to edge compute by latency and data-gravity constraints — and the unrelated free-roam VR network of the same name

COMPANY	Zero Latency (0.lat), formerly Hyphastructure	HEADQUARTERS	Charlottesville, Virginia, USA
OWNERSHIP	Private	TOTAL RAISED	Not publicly disclosed
CEO	Michael Huerta (co-founder/CEO)		

ABSTRACT

"Zero Latency" is a name that points at two unrelated organizations. The first is Zero Latency (0.lat), a Charlottesville, Virginia distributed AI inference startup, formerly named Hyphastructure, that operates "Zerogrid," a network designed to route inference workloads to edge capacity by latency, data-gravity, and burst constraints. Its conceptual borrow is from the energy sector: it models itself on distributed virtual power plants (VPPs). The second is Zero Latency VR, a Melbourne-founded free-roam virtual-reality entertainment business that lets up to eight untethered players roam a warehouse-scale arena, sold to operators on a B2B licensing model. There is no known corporate connection between the two beyond the shared phrase. The honest framing across both is that "zero latency" functions as branding rather than a measured engineering spec: neither company's public materials quantify the motion-to-photon or millisecond figures the name implies. This explainer disambiguates the two and characterizes each on its own terms.

Keywords: Zero Latency; 0.lat; distributed AI inference; edge compute; virtual power plant; free-roam VR; Zerogrid; physical AI

1. Snapshot

Two distinct entities answer to "Zero Latency," and conflating them is the single most common error.

Zero Latency (0.lat) is a distributed AI inference company based in Charlottesville, Virginia, formerly named Hyphastructure, with Michael Huerta as co-founder and CEO. The Charlottesville geography comes from trade press rather than first-party filings, so treat it as secondary-grade. It launched a closed beta of its "Zerogrid" product in May 2026.

Zero Latency VR was founded in Melbourne, Australia, in 2013 by Tim Ruse, Scott Vandonkelaar, and Kyel Smith, opening its first public venue in August 2015. It was reported acquired by private-equity firm Advent Partners in 2021, though the evidence does not pin down whether that was a full buyout or a majority/growth stake.

Key unknowns for both are substantial: 0.lat's total capital raised, true GPU scale, revenue, and headcount are undisclosed, and Zero Latency VR's revenue and current venue count are unverified.

2. Thesis: Why This Disambiguation, Why Now

The reason these two collide is timing and search collision, not strategy. The AI-infrastructure wave has minted a new 0.lat that uses the same phrase a decade-old VR brand built equity in, and the two have no stated relationship. The honest stance is "no known connection," not "definitively separate," because neither company has publicly addressed the shared name.

For 0.lat, the bet is that inference, not training, is where demand decentralizes. As physical-AI workloads (robots, drones) proliferate, latency and data-gravity push compute toward the edge, and 0.lat's pitch is that a routing layer over distributed capacity serves these better than centralized cloud. That "poorly served by centralized cloud" comparison is the company's own positioning, not an independently benchmarked result.

For Zero Latency VR, the thesis is more mature: location-based, warehouse-scale VR as out-of-home entertainment, sold to arena operators. The market is real and operating, with the company reporting more than five million player experiences over its history.

3. The Core Idea in Plain English

0.lat's core idea borrows directly from the power grid. A virtual power plant stitches together many small, distributed energy sources and dispatches them on demand instead of relying on one central plant. 0.lat applies the same shape to AI inference: rather than sending every request to a hyperscale data center, you define the service constraints a job needs, and Zerogrid routes it to whichever edge node can satisfy them right now, treating inference as a "first-class routing primitive."

Zero Latency VR's idea is physical, not infrastructural. Old world: VR meant one person, tethered, standing in place. New world: up to eight people walk freely through a shared physical space the size of a small warehouse, seeing and hearing each other inside the same game.

4. The Technical Space

These two occupy entirely different technical categories, and "good" means different things in each.

Distributed inference. The problem 0.lat addresses is matching AI inference requests to compute under competing constraints. Standard practice routes inference to centralized cloud regions chosen mostly for cost and availability. The dimensions that matter here are latency (round-trip time to the model), data gravity (the tendency for computation to be most efficient near where data originates or must remain), and burst handling (absorbing spiky demand without overprovisioning). For latency-sensitive physical-AI workloads, a request that must traverse the public internet to a distant region can be disqualifyingly slow, which is the gap edge-routing aims to close.

Free-roam VR. The problem here is sustaining immersion for multiple untethered players in a shared space. What matters is precise position tracking, low motion-to-photon delay (the lag between a head movement and the display updating), reliable wireless streaming of high-resolution frames, and keeping multiple players spatially synchronized without collisions. The historical constraint was that high-fidelity VR required either a tether or a backpack PC, both of which limit how many players can roam and how freely.

5. How Their Technology Works (and What's Proprietary)

0.lat's stated architecture decomposes into a few components.

- **The routing layer.** Zerogrid is a distributed inference grid that routes workloads to edge capacity by latency, data-gravity, and burst constraints. The conceptual model is the distributed virtual power plant, treating inference as a routing primitive rather than a fixed destination. This framing is the most distinctive and explainer-worthy part of the company's story.

- **The compute stack.** 0.lat reports using NVIDIA Blackwell GPUs (NVIDIA's current-generation data-center GPU architecture) and Red Hat's Kubernetes-based AI platform (Kubernetes being the standard open-source container orchestration system), focusing on inference rather than training.
- **The physical footprint.** The company reports building its own edge facilities, with three online (in California and Florida) and six more planned, aimed at physical-AI use cases like robots and drones. Present this as a company report, not as independently verified live infrastructure. The "tens of thousands of GPUs" figure that appears in the company's own marketing copy has no independent corroboration in trade reporting, which describes only the three online facilities; treat it as an unverified marketing claim.

The hard question is defensibility. The orchestration logic, the constraint-aware routing that decides where each workload lands, is where genuine IP could sit, since the VPP analogy is conceptual rather than off-the-shelf. The stack underneath is not proprietary: Blackwell GPUs and Red Hat's platform are available to any well-funded competitor, and a hyperscaler could extend its own edge offering toward the same constraints. The edge facilities are real assets but, at three online, not yet a scale that competitors cannot match.

Zero Latency VR's "Gen3" system, developed with HTC, uses Wi-Fi 6E (the latest generation of Wi-Fi, operating in the 6 GHz band for higher throughput and lower interference) to stream 5K VR and eliminate the backpack PCs earlier systems required. These are reported press-release and venue-page specs, not first-party engineering documentation.

6. Business and Go-to-Market

The two run opposite commercial motions.

Zero Latency VR operates a B2B licensing and arena model: operators run the venues, with up to eight players roaming untethered across roughly 200 square meters. The company reports more than five million player experiences, which should be read as a self-reported KPI consistent with a ten-year network history, not an audited figure. First-party investor language affirms the licensed-arena structure. Venue counts disagree across sources (roughly 120-plus as of mid-2025, 150-plus per investor materials), with the discrepancy likely reflecting different definitions of "active" versus "launched." Revenue estimates from third-party aggregators vary widely and disagree even on headcount, so no single figure is defensible.

0.lat is far earlier. Its go-to-market centers on the Zerogrid closed beta launched in May 2026, targeting physical-AI workloads, and at this stage appears to be direct engagement with early design partners rather than a self-serve or channel model. There are no disclosed customers, revenue, or pricing in the available evidence. Unit economics are an open question by construction: building owned edge facilities is capital-intensive, and a distributed-inference business must clear the spread between what customers pay for low-latency routing and the cost of standing up and utilizing dispersed Blackwell-class capacity. Whether that spread is attractive at three facilities, versus at the scale the company aspires to, is unknown.

7. Competitive Landscape and Moats

The two face different fields and should be assessed separately.

For 0.lat, the competitive frame is the broader distributed-inference and edge-AI category, where it positions against centralized cloud. In practice it competes against "do nothing different" (centralized inference) and "roll your own edge" among sophisticated robotics teams. That "centralized cloud serves these workloads poorly" claim is the company's own and is not third-party benchmarked.

Routing IP is the only plausible technical moat, and it is unproven. If the constraint-aware orchestration genuinely outperforms generic edge routing, that is durable. But the VPP analogy is a concept, not a barrier, and a hyperscaler extending its edge platform toward latency and data-gravity routing is the obvious platform risk.

Owned edge facilities could become a moat but are not one yet. Three facilities online, with six planned, is a footprint a well-capitalized rival can match; durability would require scale and dense geographic coverage the company has not demonstrated.

For Zero Latency VR, the moat is more legible. **Installed base and operator relationships are the real asset.** A decade of deployed arenas, a reported five million-plus player experiences, and an HTC hardware partnership create switching costs for operators who have built venues around the platform. The threat is a competing free-roam VR licensor or a hardware shift, such as consumer headsets converging on the warehouse-scale experience, that resets the field, but the operating network is a tangible lead.

A coincidental aside: "0 lat" also evokes Null Island, the cartographic point at 0°N 0°E used as a GIS placeholder for bad or missing coordinates. It is unrelated to 0.lat.

8. Risks and Open Questions

The diligence questions diverge sharply by entity.

- **0.lat infrastructure reality.** The company markets a large always-on network, but trade coverage describes only three facilities online. How many GPUs does 0.lat actually own versus orchestrate from partners, and what is verifiable today?
- **0.lat performance proof.** The low-latency-edge advantage is asserted without published millisecond figures or SLAs. What measured latency and reliability does Zerogrid deliver against centralized cloud for a real physical-AI workload?
- **0.lat platform dependency.** With Blackwell GPUs and Red Hat underneath, what stops a hyperscaler from offering the same constraint-aware routing as a feature, and what is the mitigation if upstream platforms subsume routing?
- **0.lat capital and AI-capex cyclicality.** Edge facilities are capital-intensive; what is the runway, and if inference costs fall sharply through model-efficiency gains, does the addressable gap narrow faster than the buildout can return capital?
- **Zero Latency VR scale and structure.** Venue counts in company materials disagree, and the nature of the 2021 Advent Partners transaction (full acquisition versus majority or growth stake) is not pinned down.
- **The name itself.** Neither company has publicly addressed the shared "Zero Latency" branding, leaving any relationship, or trademark tension, an open question.

9. Bottom Line

The most important read is structural: "Zero Latency" is two unrelated stories, and treating them as one is the error to avoid. 0.lat's distinctive idea, inference routed like a virtual power plant, is genuinely interesting but rests on unproven IP and a three-facility footprint that marketing copy describes as far larger. Zero Latency VR is a real, operating, ten-year-old VR network whose moat is its installed base, not its branding. The thing to watch next is whether 0.lat can convert its VPP-for-inference concept into a benchmarked, scaled grid before a hyperscaler absorbs the idea.

10. For the Nerds

The deeper bet inside Zerogrid is that inference scheduling under data-gravity constraints is a fundamentally different optimization than cost-availability scheduling, and that a VPP-style dispatch model captures it. Formally, this is a variant of the online bin-packing and job-scheduling problem: assign each incoming job to a node given GPU memory, queue depth, and network distance, so as to minimize end-to-end latency subject to residency and burst caps. That problem is NP-hard in the general case but tractable with heuristics and predictive load models, which is why a workable design tends to blend a fast hot-path heuristic with slower periodic optimization plus admission control and regional circuit breakers.

The interesting open problem is statefulness. Inference is not perfectly stateless once you introduce KV-cache reuse, retrieval context, or model-affinity, so routing a request to the lowest-latency node can collide with routing it to the node that already holds the warm cache or the co-located data. Worse, model-weight cold starts (the cost of loading a large model into a node that has not recently served it) can dominate total latency at the edge. A genuine constraint-router has to price those tensions, and nothing in the public materials reveals how. There is also a utilization question: Blackwell-class GPUs are expensive to leave idle, and a dispersed footprint fragments demand against the batching that makes inference economical. The VPP analogy is seductive precisely because power dispatch tolerates idle capacity that AI economics punish.

WHERE SOURCES DISAGREE

Exact number of Zero Latency VR venues

One company-facing source (investor materials) cites "150+ active arenas," while a June 2025 press release cites "120+ venues," and earlier materials gave 100 (Sept 2024) or 103. Both higher figures are recent and company-originated, yet they disagree directly

Unresolved: the materials don't clarify whether counts reflect arenas "launched" versus currently operating, so only a timestamped range (e.g., "120+ as of mid-2025, 150+ per investor materials") is defensible

Zero Latency VR's annual revenue

One aggregator estimates roughly $49.8M (221 employees); another lists about $14.3M (106 employees); a third figure of ~$9.1M also appears. All are third-party modeled estimates that disagree on both revenue and headcount, and no audited figures or company disclosures are available, so no single number can be stated as fact.

Whether 0.lat owns its "tens of thousands of GPUs" or orchestrates partner capacity

The 0.lat homepage advertises "tens of thousands of GPUs" on an "always-on" network, while trade reporting describes only three edge facilities online with six more planned and enumerates no GPU count

Unresolved: the figure appears only in the company's own marketing copy with no independent infrastructure audit or third-party corroboration in the available evidence

The precise nature of the 2021 Advent Partners transaction

One account describes Zero Latency VR as having been "acquired" by the private-equity firm; the evidence does not distinguish a full acquisition from a majority or growth investment

Unresolved: the sourcing states the event without detailing the corporate-control structure

Whether the two "Zero Latency" entities have any relationship

One reading treats them as entirely separate companies; another notes that the evidence says only that any connection is "unclear," not that separation is confirmed

Unresolved: no statement from either company explicitly addresses the shared name, so the honest stance is "no known connection" rather than asserted independence

"Zero Latency" as a literal technical claim

In the VR context, immersion is described qualitatively ("tracked instantly") with no quantified motion-to-photon latency; in the AI context, low-latency edge routing is asserted without published millisecond figures or SLAs

Unresolved: neither company's materials quantify the latency their shared name implies, leaving it as aspirational branding rather than a measured spec