Cloud AI Thinks. Edge Autonomy Acts.

How the Next AI Platform Layer Embodies the

Architecture of Human and Biological Swarm Intelligence

Denis Garagić, CTO and co-founder, Palladyne AI

Ben Wolff, CEO and co-founder, Palladyne AI

Executive Summary

Cloud-based AI systems have created extraordinary value by excelling at language, reasoning, and digital workflows. These systems are optimized for thinking, and their valuations reflect that strength.

But intelligence that must act in the physical world operates under fundamentally different constraints. Machines that move, coordinate, and make decisions in real time cannot rely solely on centralized reasoning, constant connectivity, or energy intensive computation. They must operate locally, respond in milliseconds, collaborate with other machines, and do so within tight limits on power, size, and cost.

Nature created the perfect architecture for intelligence and swarming. Palladyne AI has modeled that architecture to enable machines not just to think, but to act in the real world.

We refer to this architectural class as Decentralized Embodied Collaborative Autonomy (DECA). DECA describes biologically inspired autonomy in which physically embodied agents perceive, decide, and act locally while collaborating through decentralized interaction rather than centralized control. Artificial intelligence is the enabling mechanism, but autonomy and execution occur entirely at the edge.

This paper is not an argument against cloud AI. It is an argument that machines which must act continuously and autonomously require a different intelligence architecture. We've seen this movie before. In the 1980s and 1990s, an analogous change in compute architecture ushered in a shift from centralized systems dominated by the likes of IBM to personal computers and operating systems from companies like Microsoft and Apple. Cloud systems remain powerful and complementary, but they cannot serve as the real-time execution layer for autonomous systems operating under tight physical constraints.

Nature solved this problem long ago. Human intelligence depends on massive local filtering and predictive processing, while conscious thought intervenes selectively. Likewise, ants and bees, which are arguably the most scalable biological systems, coordinate through decentralized swarms with no central controller, yet achieve speed, resilience, and efficiency that centrally managed systems cannot.

The next major AI platform layer is being built to act, not only think. That layer is edge-native and decentralized by necessity. It is designed for systems in which each machine operates autonomously while collaborating locally with others, without reliance on continuous centralized control or excessive compute.

Palladyne AI has already implemented early versions of a DECA architecture through its Palladyne IQ and SwarmOS platforms, proving the approach in real systems today while remaining at the early edge of a much larger and inevitable opportunity: becoming the intelligence layer for autonomous machines operating beyond the reach of the cloud.

I. Why Humans Act Locally and Think Selectively

Human intelligence works because most of it never reaches conscious thought.

At any moment, the human body generates on the order of 1 billion bits of sensory data per second from sight, sound, touch, taste, smell and proprioception[1]. Of that immense stream of information, roughly 10 bits per second enters conscious awareness. More than 99.9999% of sensory data is filtered, predicted, and acted upon automatically by fast, local systems.

You do not consciously calculate how to keep your balance while walking. You do not reason your way through catching a falling object. Those actions are handled by local intelligence operating far faster and more efficiently than conscious thought ever could.

Conscious reasoning is slow, metabolically expensive, and bandwidth limited. It intervenes selectively, primarily when prediction fails because something unexpected occurs.

This architecture is not accidental. It is the only way a biological system can act in real time while operating within strict limits on energy, bandwidth, and processing capacity.

The lesson is simple: intelligence that must act continuously cannot afford to think about everything.

II. What Cloud AI Is Exceptionally Good At

Modern cloud AI systems excel at exactly the kinds of problems conscious reasoning is good at.

They aggregate massive datasets, identify abstract patterns, generate language and plans, and support human decision-making. Latency is acceptable. Connectivity is assumed. Energy and compute are abundant relative to edge systems.

This is why cloud AI has created enormous value. It is ideally suited to digital workflows such as search, writing, analysis, design, and long-horizon planning.

These systems are optimized for thinking.

III. Acting Is a Different Class of Intelligence

When intelligence moves from the digital world into the physical one, the rules change.

Physical systems must:

- Respond in milliseconds
- Operate continuously and safely
- Function with limited onboard power and compute
- Collaborate with other machines in real time
- Fail gracefully, not catastrophically

In this environment, centralized or compute-heavy reasoning alone becomes a liability. Bandwidth is constrained. Latency matters. Energy is scarce. The cost of error is physical.

This is why autonomy cannot simply be “cloud AI applied to machines.” The architectures that succeed at large-scale reasoning do not translate directly to systems that must act continuously in the real world.

Acting intelligence must live where arbitrary action occurs and it must do so efficiently.

IV. Why Raw Compute Is the Wrong Answer

It is tempting to assume that the path to machine autonomy is simply more compute.

That assumption collapses under basic physical constraints.

Consider the world’s fastest supercomputer, El Capitan, operated by Lawrence Livermore National Laboratory. El Capitan delivers roughly 2.7 exaflops of peak performance[2], only a few times greater than the estimated upper bound of human brain processing capabilities.

Achieving that performance requires extraordinary resources. El Capitan occupies approximately 6,800 square feet of dedicated data-center space. It consumes on the order of 30 megawatts of continuous electrical power which is roughly equivalent to the average electricity usage of about 25,000 U.S. homes. The system cost approximately $600 million, excluding long-term operational and infrastructure costs. This is what it takes for centralized machines to approach human-scale cognitive throughput through brute force.

A mobile machine cannot carry a 6,800-square-foot facility. It cannot draw tens of megawatts of power. It cannot justify hundreds of millions of dollars in compute infrastructure for each deployed system.

If we expect autonomous machines such as drones, vehicles and robots to act in real time, they must operate within severe constraints on size, weight, power, and cost. The question is not how to give machines more compute. The question is how to avoid needing it in the first place.

Nature answered this question long ago.

V. How Nature Solved Real-Time Intelligence at Low Cost

Biological intelligence does not scale by centralizing computation. It scales by filtering aggressively, predicting locally, and coordinating simply.

Humans do not process every sensory input with conscious thought. Ants and bees do not form or compute global plans; instead, each individual responds to local signals and interactions, and coordinated behavior emerges from those local rules.

This architecture enables:

- Real-time response with minimal energy
- Robust operation despite individual failures
- Scalable coordination without centralized control
- Intelligence embedded where action occurs

This is not a metaphor. It is the only architecture ever shown to support continuous, real-world intelligence under strict physical constraints.

VI. Applying Biological Architecture to Machines

Palladyne AI's development of DECA architecture reflects these same principles.

Rather than attempting to replicate human-level reasoning through brute-force computation, Palladyne AI structures intelligence the way biology does:

- Perception is filtered locally to what matters
- Decisions are made predictively, not reactively
- Each machine operates autonomously
- Collaboration emerges through decentralized interaction
- Cloud systems support learning and improvement, not real-time execution

This allows machines to act quickly, reliably, and efficiently even when power, bandwidth, and compute are limited.

These principles are already embodied in Palladyne AI's deployed platforms. What exists today represents the early stages of a much larger trajectory, as these systems scale in capability, autonomy, and scope.

The DECA architecture implemented by Palladyne AI is not theoretical. It has been implemented in operational systems and is patent protected, reflecting both technical novelty and real-world applicability.

VII. Why Acting Platforms Create Durable Value

Platforms that think are valuable. Platforms that act are harder to replace.

Once intelligence is embedded in machines that operate in the real world:

- Switching costs rise with every deployment
- Experience compounds defensibly
- Reliability and efficiency become barriers to entry
- Architecture matters more than raw compute
- Cost curves improve with scale rather than explode

These dynamics favor systems that are efficient by design, not dependent on ever-increasing compute. As with prior platform shifts, the architectures that align with physical reality and not simply theoretical maximum performance are the ones that persist.

Conclusion

Cloud AI has proven that intelligence optimized for thinking can create extraordinary value.

The next chapter of AI is being defined by systems optimized to act, not only think.

Nature created the perfect architecture for intelligence and swarming. Palladyne AI has leveraged that architecture to enable machines not just to think, but to act in the real world.

Notes

[1] Zheng, J., & Meister, M. (2024). "The unbearable slowness of being: Why do we live at 10 bits/s?"

[2] https://str.llnl.gov/str-december-2024/introducing-el-capitan