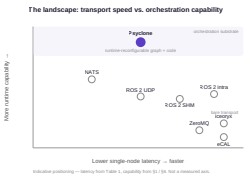
Psyclone AIOS — the runtime-reconfigurable orchestration substrate that is also latency-competitive

CMLabs Technical Report · v2.2, July 2026 · one-page summary

Psyclone moves messages competitively fast and lets an LLM rewire the dataflow at sub-millisecond speed and reprogram the system's code while running — the capability that matters when the shape of an AI system has to change as fast as its inputs. It pays a small, predictable latency premium over a bare transport, and in exchange gives you an orchestration substrate: crash-isolated processes, dynamic scaling, in-runtime simulation of new functionality, and a dataflow graph that is a runtime object.



Indicative landscape. Dedicated transports (iceoryx, eCAL, ZeroMQ) win on raw latency but supply only a message path; NATS trades latency for broker reach; tuned ROS 2 sits between. Psyclone alone occupies the orchestration-substrate zone.

What makes it different — five capabilities

1. **Sub-millisecond global rerouting.** One context switch re-routes every affected stream at once (≈0.11 ms), graph-wide.
2. **Trigger batching & multi-input joins.** Many named triggers feed one queue → one activation; fusion without a synchroniser node.
3. **Ephemeral compute.** Spin a module into the live graph on demand (≈0.30 ms) and tear it down after — per-request compute inside the fabric.
4. **Runtime subscription updates.** A live module's inputs rewired by the module itself, an LLM supervisor, or a human — no redeploy.
5. **Live reprogramming — including code.** A running module's logic replaced in place. Human-driven today; LLM-driven by release.

The systems we compare

NATS — distributed broker; horizontal fan-out and durability, two socket hops per message on-node.

ROS 2 (Fast-DDS) — robotics middleware; measured in three regimes: default UDP, tuned SHM, and single-process intra-process.

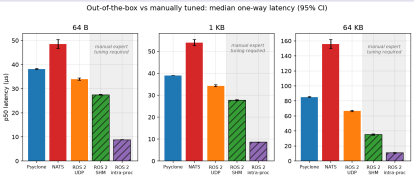
Eclipse iceoryx — true zero-copy shared-memory transport; the fastest cross-process path measured.

eCAL — SHM pub/sub middleware; single-digit-µs class at small payloads.

ZeroMQ `inproc` — in-process library channel; no process isolation.

The last three are bare transports — no broker, no orchestration, no runtime-reconfigurable graph. They set the raw-latency floor Psyclone is read against.

The comparison that matters: day-one, zero-tuning vs expert-tuned



Median one-way per-hop latency: default installs vs manually tuned configurations (hatched). The two configs that beat Psyclone both require expert tuning or surrender process isolation. Full data: Table 1 of the paper.

Headline numbers (one-way per-hop latency, single node)

System (config)	64 B p50	64 KB p50	64 KB p99.9
iceoryx (zero-copy transport)	6.2 µs	12.3 µs	16.3 µs
ROS 2 intra-process (tuned)	8.8 µs	10.9 µs	16.1 µs
ROS 2 Fast-DDS UDP (default)	33.9 µs	66.6 µs	87.7 µs
Psyclone (default install)	**38.1 µs**	**85.1 µs**	**124.0 µs**
NATS (default)	48.5 µs	155.7 µs	761.4 µs

10 trials × N=100,000, mean ± 95% CI, CPU-pinned; validated on bare metal (z1d.metal). Psyclone's p99.9 stays within ≈1.8× of its median at every payload — the tightest, most predictable tails measured.

What to take away

- **Comparably fast, not fastest.** Dedicated transports and tuned ROS 2 post lower medians — those numbers are reported plainly. Psyclone's premium over a bare transport is small and predictable.
- **Best tail predictability.** Bounded p99.9 (≤124 µs everywhere), smallest CIs — confirmed on bare metal.
- **The only runtime-reconfigurable system measured.** Routing, subscriptions, and module code are all runtime objects — the property GenAI/agentic systems need.
- **Open source (LGPL v3).** Every system compared, Psyclone included, is buildable from published source and independently reproducible.

Read the full technical report →

Methodology, all tables, bare-metal replication, cross-node estimates, and the differentiator measurements are in the full paper. AWS z1d.2xlarge & z1d.metal, kernel 7.0.0-1006-aws · NATS 2.14.3 · ROS 2 lyrical / Fast-DDS 9.4.8 · Psyclone AIOS 2.1.0 (LGPL v3).