

Fast Isn't Enough.

Why latency still breaks streaming systems.

Kranti Parisa

Founder & CEO, LaserData · Apache Iggy PPMC

REAL-TIME AI, WHEN IT WORKS

A sales call with live transcription.

Every word appears as it is spoken. Action items surface automatically. When it works,
the AI is invisible.



END-TO-END < 100 MS · ZERO DROPPED WORDS

THEN THIS HAPPENS

One delay. The conversation breaks.

AUDIO IN
still running

**MESSAGE
STREAM**

ASR
waiting...

REP SCREEN
words missing

GC PAUSE · 320 MS

WORDS DISAPPEAR

A sentence drops. The rep loses context mid-call.

ACTION ITEMS WRONG

AI captures the wrong task from an incomplete sentence.

TRUST COLLAPSES

The rep stops reading the transcript. The AI becomes noise.

WHERE SYSTEMS FAIL

The tail is where experience breaks.

MEDIAN

P50

Unnoticed.

95TH

P95

Noticeable.

99TH

P99

Painful.

99.9TH

P99.9

System failure.

Your worst users — the ones most likely to churn — **live at the tail.**

THE UNCOMFORTABLE TRUTH

Everyone's building workarounds.

Capacity.

Scale E2E infra artificially.

Cache.

To mask broker latency.

Bypass.

Custom RPCs around the broker.

Patch.

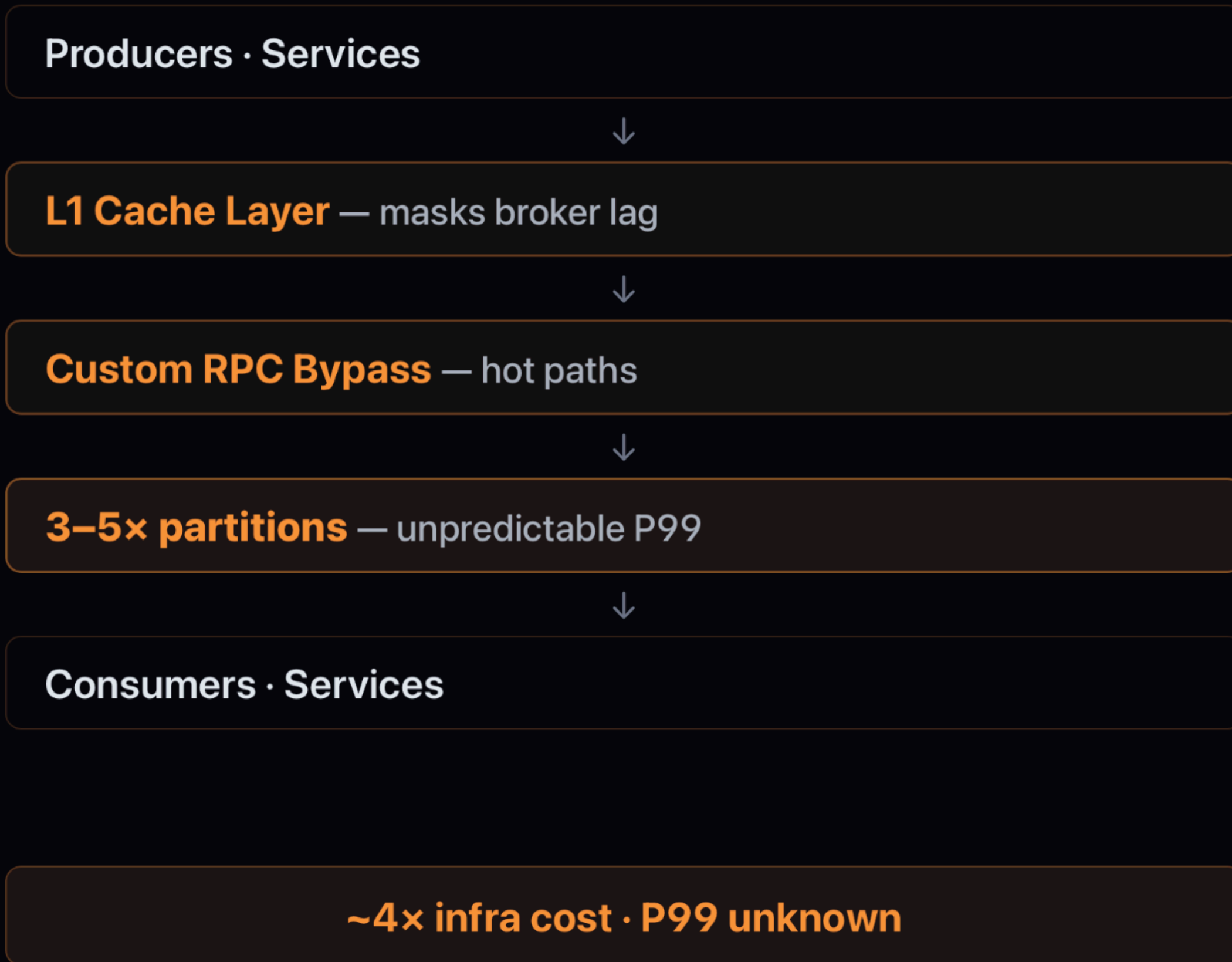
Hand-rolled pub-sub for hot paths.

Not solutions. **Symptoms of unpredictability.**

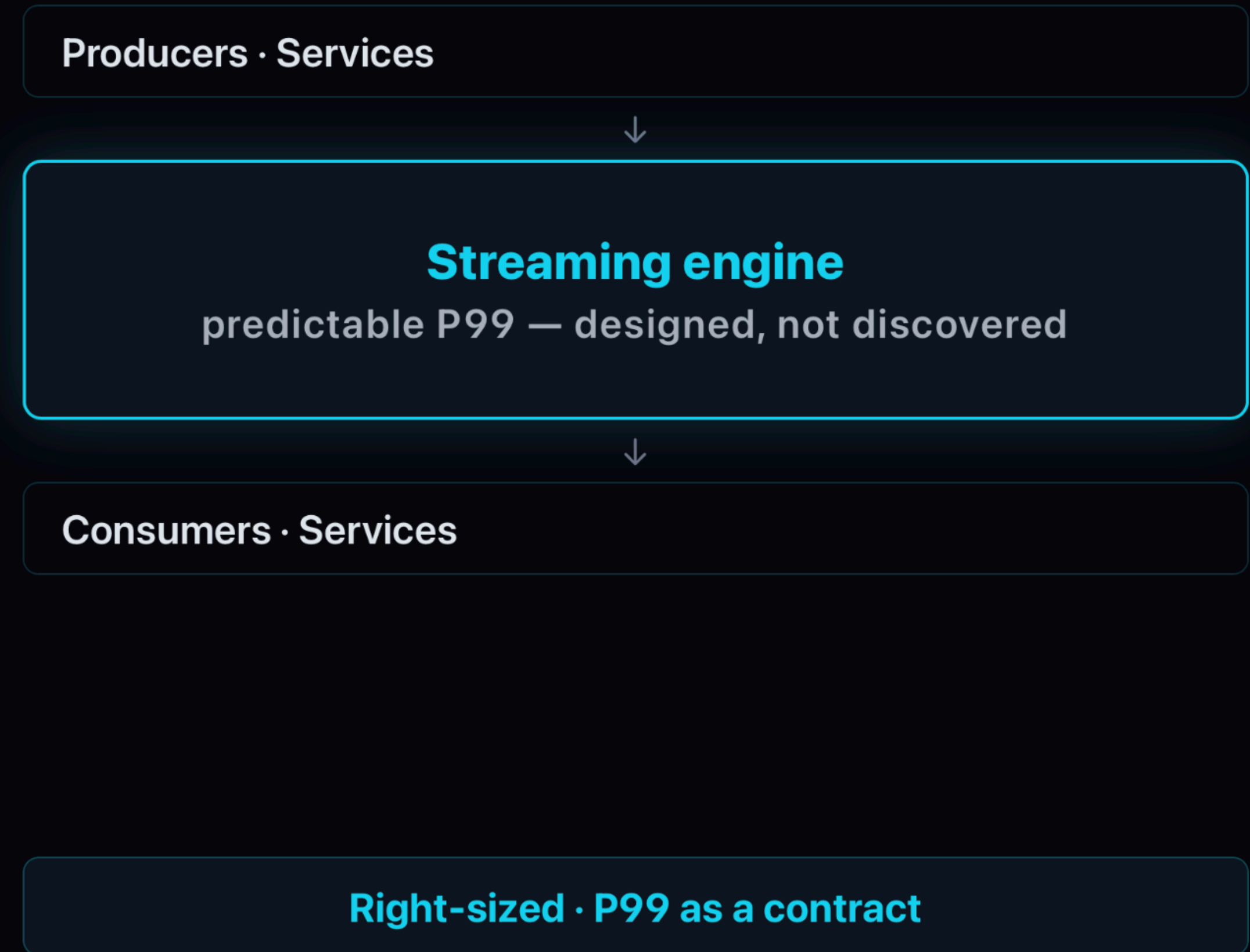
THE TAIL LATENCY TAX

Complexity as a coping mechanism.

TODAY



WITH A PREDICTABLE ENGINE



WHY NOW

Data is exploding — and going **real-time.**

394zB

GLOBAL DATA BY 2030 — 23% CAGR

~40%

NEEDS REAL-TIME
PROCESSING BY
2030

75 B

CONNECTED
DEVICES
STREAMING
EVENTS

Zettabytes



AI ERA

Not one prediction. Continuous reasoning.

TRADITIONAL ML

Move data to the model.

Batch

Nightly ETL · hours late is fine.

One hop

Request → prediction → done.

AGENTIC AI

Stream events between models.

Streams

Continuous, multi-source, always live.

Many hops

Agents, tools, memory — every turn.

Same building blocks. **Completely different data movement.**

AT SCALE

Message streaming is the **nervous system**.



AI is the **brain**.

Streaming is the **nervous system**.

Compute evolved. Data movement didn't.

Streaming must evolve the way GPUs evolved for AI.

Not band-aids. A fundamental shift.

THE INSIGHT

Same fundamentals. **New engine.**

PREVIOUS GENERATION

Designed 2011 · JVM · Throughput-first

Streams	✓
Topics	✓
Partitions	✓
Consumer groups	✓

JVM · GC pauses

Complex, external consensus

OS page cache

Throughput-optimized

IGGY

Rust · Designed for predictability

Streams	✓
Topics	✓
Partitions	✓
Consumer groups	✓

Rust — no GC, zero-cost

Built-in consensus

Direct I/O — bypass page cache

Predictability-first

The next phase in data streaming.



Engineered for predictability, not throughput alone.

Written in Rust

Low protocol overhead

Minimal data movement

Predictability-first

4,100+

GITHUB STARS

Top 3

RUST PROJECT AT APACHE

PPMC

LEADING FROM THE FRONT

PROOF, NOT PROMISES

Benchmarked. Not estimated.

1M+ msg/s

Throughput
+1 GB/s

1.01 ms

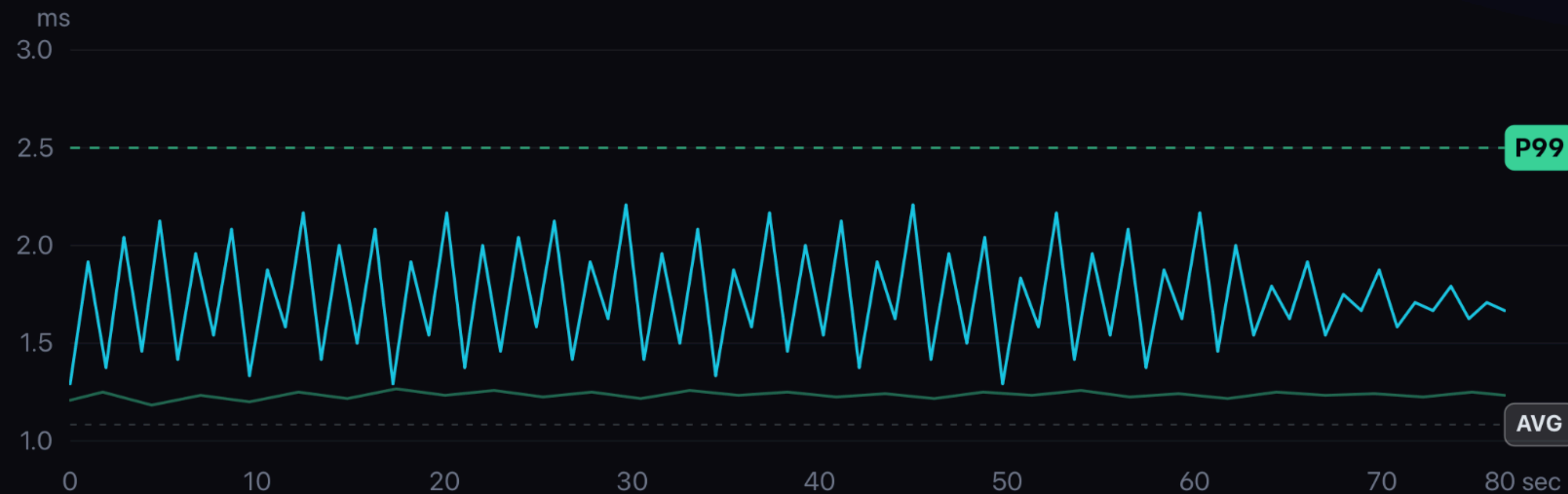
Avg latency
Producer-side, sustained

2.05 ms

P99 latency
The tail, measured

● Producer ● Consumer

40M messages · 80 s run



LASERDATA

lggy at internet scale.

Enterprise streaming for the AI era. Built on Apache Iggy.

✓ **Blazing Fast & Predictable**

1 ms P99 · 1 GB/s R+W · 1 M+ msgs/sec per node

✓ **Multi-Cloud**

AWS · GCP · more coming

✓ **Dedicated Infrastructure**

No noisy neighbors · Built-in observability

✓ **Open Source Core**

Creators of Apache Iggy · 4 protocols · 6 SDKs

✓ **Flexible Deployment**

Managed · BYOC · Private · On-prem

✓ **Zero-Trust Security**

Enterprise-grade · RBAC · No inbound, no vendor access

Fast is easy. Predictable is hard.

The compute layer got its GPU moment.
Data infrastructure is next.

If not the Rust community — who?

Obsessed about tail latencies?

Come build with us.

Register for the FREE TIER. Join the community.
Let's take this to the next level — together.

Kranti Parisa

Founder & CEO, LaserData · Apache Iggy PPMC

laserdata.com

iggy.apache.org



SCAN TO VISIT

Thank you.