verda

MLSys 2026

AI-native cloud
Built for AI builders

Find us at booth 5

Arturs Polis

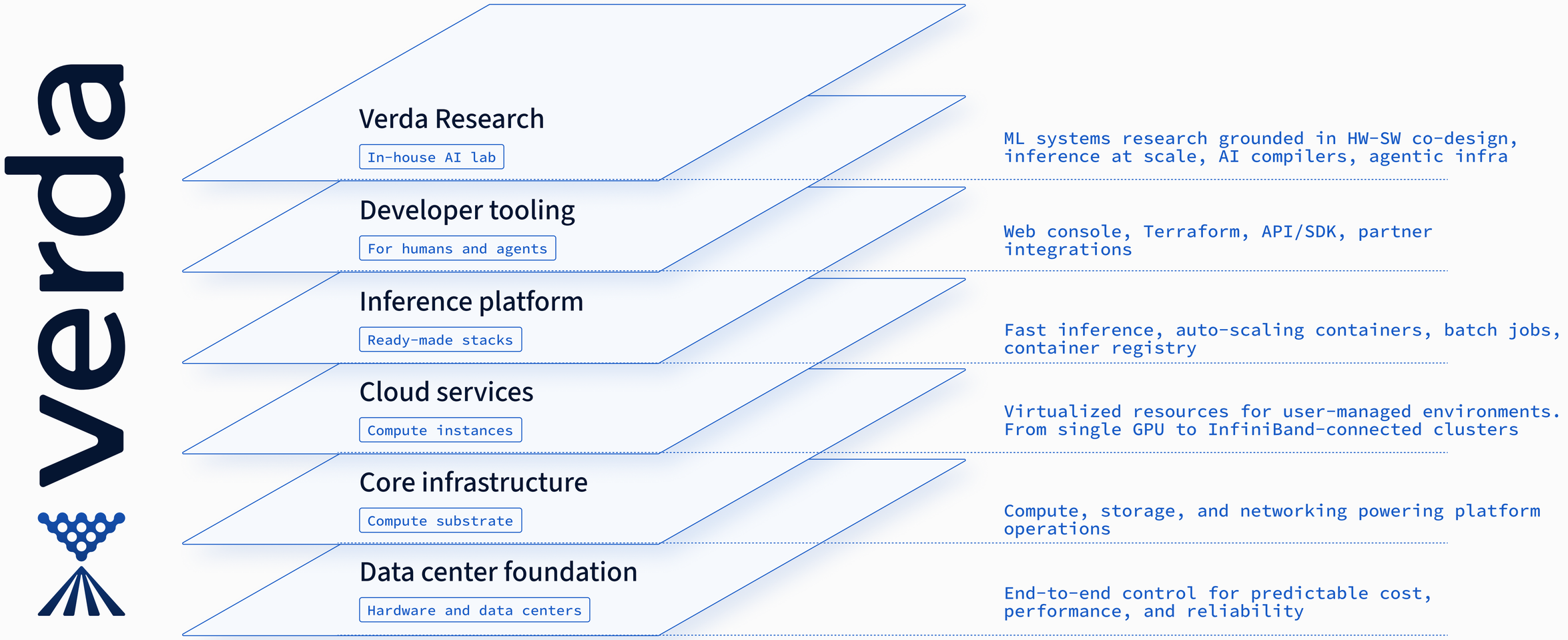
Chief Technology Officer / arturs@verda.com

Overview on Verda

verda
MLSys 2026

We build the whole stack

Hardware to applied research, all in-house. Pick your layer, or move between them.



MLSys

verda
MLSys 2026

Is AI infrastructure solved yet?

Training, inference, and agents need infrastructure built for them

Should training infra be hard? 1

The cadence mismatch

Research moves in hours. Procurement moves in weeks

Integrated stacks exist – for some teams, for some workloads

The harder question is what integration looks like as workloads diversify

Resilience gap

At scale, hardware will misbehave. Strength comes from how you recover

Can inference be more efficient? 2

The techniques are open

Composing them optimally is still bespoke

New shapes, new bottlenecks

MoE shifts the bottleneck to expert routing. Long context shifts it to KV-cache

Inference is an infrastructure problem

Efficiency gains compound at the boundary between serving and the substrate it runs on

Is the tooling around agents ready? 3

Real sandbox platforms exist

Pushing density, speed, and isolation together, not picking two out of three, is still open

The harness wants to live on the substrate

When rollouts are parallel and runs are long, network round-trips and orchestration overhead become the bottleneck

Infra and software – together

Fast cold starts, fast weight loading, RL environments, sandbox lifecycle - the substrate has to deliver a lot of pieces together

MLSys

verda
MLSys 2026

What we’re building

Optimizing across the full stack, end to end

Autonomous infrastructure

1

Cells over global coordination

Each site is a self-sufficient cell. The global plane handles only what must be globally consistent

Declarative reconciliation, agents on the tail

Configuration declares the target state. Agents reconcile reality to it.

Typed tools, not ad-hoc scripts

Every action that mutates state is a versioned, audited tool call with declared blast radius

Inference built for modern workloads

2

Modern serving stack

Prefill-decode disaggregation, wide expert parallelism, KV-cache reuse, full parallelism control

Closing the gap to peak throughput

Kernel-level work and tuned parallelism layouts. Pushing MFU closer to what the hardware can actually deliver

Tuned for MoE and long context

Wide EP, context parallelism, KV management built for these workloads from the ground up

Agent-native cloud

3

Sandboxes as first-class primitives

CPU and GPU sandboxes with proper isolation and fast cold starts. Same infrastructure for agent tool calling and RL rollouts

Agents as customers

Bursty, programmatic, short-lived, and orders of magnitude more sessions than human users

Purpose-built

GPU sandboxes, CPU sandboxes, RL rollouts, and agent runtimes, built for these workloads, not retrofitted

verda

MLSys 2026

Thank You!

Find us at booth 5

Arturs Polis

Chief Technology Officer / arturs@verda.com

We are hiring!
verda.com/jobs