CYANIS AI:

Pilot (Alpha) Technical Brief

PUBLIC TECHNICAL BRIEF. Approved for distribution. July 2026.

Table of Contents

> Compute where the power already is - A measured case for distributed AI inference in Europe.
>
> CYANIS AI, Alpha Technical Brief, July 2026.

Executive Summary

In July 2026, **CYANIS AI completed its Pilot (Alpha)**, validating distributed orchestration across multiple GPUs, two independent cloud providers and mixed GPU classes, serving models from 0.8B to 675B parameters. The figures in this brief are result of that work.

Europe does not have an AI-compute problem so much as an AI-compute siting problem. Demand is rising quickly, but the constraint is rarely silicon: it is grid connections, permits, land and time. Meanwhile, the industry's default answer, ever-larger centralised facilities in a handful of regions, adds a latency cost to every interaction and concentrates dependency in places that European buyers increasingly cannot use for regulated workloads.

This brief sets out a different approach: deploy modular compute where power, grid connection and fibre already exist, and connect it through a single control layer into one coherent service. It argues three things and supports each with measurement rather than assertion.

1. **Distributed deployment scales where centralised cannot**. Capacity is added at powered sites as a repeatable unit rather than as a construction project, and the fleet fills as one pool. The Alpha validated the preconditions: near-linear scale-out across servers and providers, and zero errors under deliberate saturation.
2. **Distributed placement is measurably faster.** Routing traffic to a closer region reduced median time to first token by 150 milliseconds, or 28.3%. In multi-turn and agentic workloads, that cost is incurred on every turn.
3. **Not every request needs the largest model tier.** On knowledge-type workloads, roughly 80% of requests were served by a model with a ten-times smaller memory footprint. The quality cost is real, measurable and moves with the offload rate, which is why both numbers are published together.

A fourth point runs underneath all three: distributed infrastructure only works commercially if every node is treated as operating in an untrusted environment from the outset. That is a design constraint, not a hardening exercise.

1. The Constraint Is Siting, Not Silicon

European AI-compute capacity is expanding, but slowly relative to demand, and the bottleneck sits upstream of the hardware. New data centres in Germany can wait years for a grid connection, and in some markets greenfield connection timelines run to the better part of a decade. Land, permitting, local acceptance and construction add more.

The demand side is not waiting. European data-centre electricity demand is projected to grow substantially by 2030, while enterprise adoption of AI across the European Union continues to rise. Public funding programmes have also committed significant capital to closing the infrastructure gap.

The result is a structural mismatch:

Capacity is needed in quarters, while conventional greenfield capacity arrives in years.

There is, however, a large installed base of sites that already have what a greenfield project spends years acquiring: a grid connection, transformer capacity, physical security and, in many cases, fibre along existing corridors.

Deploying modular compute into that existing footprint can decrease delivery times from years to months.

Hyperscale sites remain necessary, particularly for training. It is an argument that they cannot be the only answer to Europe's near-term inference requirement, and that the fastest available capacity is often capacity that does not require a new grid connection.

2. The Distributed Approach Enables Scaling No Single Site Can Match

A centralised facility scales until its site does: one grid connection, one permitting process, one construction programme. Every additional megawatt competes for the same constrained resources at the same location, which is precisely where European timelines are longest.

A distributed fleet scales along a different axis. Capacity is added wherever power, grid connection and fibre already exist, from energy and infrastructure sites through commercial rooftops to household PV-and-battery environments, and each addition is a repeatable unit rather than a construction project. The addressable footprint is the installed base of powered sites across Europe, which is orders of magnitude larger than the pipeline of new data-centre connections.

Scaling that way places one hard requirement on the software: adding capacity must not degrade the service, and the fleet must fill as one pool rather than as many islands. Utilisation is an orchestration property, not a site property. A single site can serve only the demand that reaches it; a fleet behind one control plane places workloads where capacity is free and absorbs overflow across sites, so no individual location depends on attracting its own customers.

The Alpha tested exactly these preconditions. Heterogeneous GPUs across two independent cloud providers were aggregated behind a single API and control layer. Doubling capacity retained 96.2 to 98.7 % of theoretical linear throughput depending on configuration, and the system processed 19,468 requests under deliberate full saturation with zero errors. Those are the properties scale-out depends on: capacity that can be added without losing efficiency, and a control layer that remains dependable when the fleet is full.

3. Latency Is Paid on Every Turn

Network distance adds a fixed cost to every remote inference call. In a single question-and-answer exchange, this may be a minor inconvenience. In the workloads the market is moving toward, it compounds.

An agentic workflow that makes eight sequential model calls incurs the network cost eight times. Voice interfaces are bounded by human conversational tolerance, where a few hundred milliseconds can make the difference between natural and stilted interaction. Retrieval-augmented generation and multi-step reasoning patterns multiply network round trips by design.

The industry is already re-architecting around this behaviour. NVIDIA's July 2026 architecture disclosure for the Vera CPU describes agentic AI workflows in which agents execute code, call tools and retrieve context between model interactions. As these loops run concurrently, they shape both per-agent responsiveness and overall throughput.

Every millisecond in that loop may be paid repeatedly, and the portion attributable to network distance cannot be removed by a faster processor alone.

Measured Regional Placement Result

Serving identical infrastructure and an identical model, the Alpha compared requests from a client in Romania against two European endpoints:

Route	Median time to first token
Romania to Paris	0.53 s
Romania to Warsaw	0.38 s
Improvement	**150 ms lower, 28.3 %**

Routing Romanian-origin traffic to Warsaw rather than Paris reduced median time to first token by **150 milliseconds**, or **28.3%**.

A separate Paris-to-Paris measurement returned a median time to first token of **0.29 seconds**. Because that measurement used a different request origin, it is not directly comparable with the Romania-origin measurements. It nevertheless provides a directional indication of the response times available when compute is placed closer to the user.

Country- and city-level routing will be tested in a later phase.

Two qualifications matter:

1. The controlled comparison validates **regional placement** and potential for in-country / intra-city routing.
2. The latency advantage is the property of network geography, not of any one software provider. What differentiates a platform is whether it collects that advantage automatically, placing each request on the right site across a heterogeneous fleet without the customer managing sites individually. The Alpha demonstrated this under verification: each benchmarked request confirmed which backend served it.

4. Not Every Request Needs the Largest Model Tier

A large share of production inference traffic is not exceptionally difficult. Classification, extraction, summarisation, routine question answering and short generations do not always require the largest available model.

In many deployments, however, every request is sent to the same model because only one model tier has been integrated.

Serving a request with an appropriately sized model is straightforward in principle and consequential in practice. A smaller model with approximately one-tenth of the memory footprint can require materially less compute for suitable requests because a significant component of token-generation cost is the repeated movement of model weights through memory.

The trade-off must nevertheless be shown honestly. The Alpha measured two public benchmarks:

Workload	Share served by the smaller model	Benchmark retention
Mathematical reasoning, GSM8K	11 %	94.8 %
General knowledge, MMLU	approximately 83 %	79.5 %

Two conclusions follow.

Offload Rate Is a Property of the Traffic

On mathematical reasoning, most requests remained on the larger model because they required it. On general knowledge, most did not. A single universal offload percentage therefore says more about the benchmark mix than it does about the system itself.

Compute Reduction and Quality Retention Move Together

At the high-offload operating point, the modelled compute reduction reaches approximately **72%**, derived from the measured routing distribution and the relative memory footprint of the two models. At the same operating point, benchmark retention is approximately **79.5%**.

These are not independent claims. They are two ends of one quality-versus-compute curve and should always be quoted together.

For a buyer, the relevant question is not simply: "How much can the system save?"

It is: **"What does my traffic look like, and what quality threshold am I prepared to set?"**

That is measurable, and assessing it against real customer traffic should be one of the first steps in a deployment.

5. Trust Boundaries in Distributed Infrastructure

Distributing compute raises a fair question: infrastructure outside a conventional data-centre environment may sit across a wider range of physical locations, so what protects the workload?

The **CYANIS™ target architecture** treats every node as operating in an untrusted environment by default, including nodes located at sites controlled by the group. The premises operator, site host and connecting network remain outside the core trust boundary.

Workload isolation, key custody and integrity verification must therefore be properties of the platform rather than assumptions based on the building in which the hardware is located.

This is a design constraint from the outset. Retrofitting a coherent trust model onto a large, deployed fleet would be materially more difficult.

The wider industry direction supports this approach. Confidential-computing capabilities are moving into silicon. NVIDIA's July 2026 architecture disclosure describes per-VM encryption keys, confidential VM isolation, secure device assignment and authenticated encryption across coherent links, establishing a trust domain from server to rack scale.

Hardware-level attestation and memory encryption make it materially more practical to run regulated workloads on distributed infrastructure than it was a hardware generation ago.

For buyers in regulated sectors, the relevant questions remain:

1. What is inside the trust boundary?
2. Who holds the encryption keys?
3. Which components and states are attested?
4. How is attestation performed?
5. What is technically verifiable rather than merely contractually asserted?

6. What We Measured

The Alpha was a software validation of distributed GPU infrastructure connected behind a single control layer and running real inference workloads.

All tests ran on cloud GPU infrastructure. Physical deployment and energy integration are separate programmes and are not covered by these results.

Distributed Scale-Out

Capacity was doubled at three levels, and the percentage of theoretical linear throughput retained was measured in each case:

Configuration	Scaling efficiency
Two GPUs within one server	98.7 %
Two servers, one GPU each	98.0 %
Two independent servers of eight H100 GPUs each, serving replicas of a 675B-parameter inference service	96.2 %

Scaling efficiency is the proportion of theoretical throughput retained when capacity is doubled.

For the 675B-parameter service, tensor parallelism operated within each eight-GPU server. Requests were then distributed across two independent server replicas.

The distributed service retained **96.2% of theoretical linear throughput** when the second server was added.

Robustness

Under deliberate full saturation, the system processed:

1. **19,468 requests**
2. **Zero errors**

Median time to first token remained between **0.448 seconds and 0.479 seconds**, while the 95th percentile widened with queueing, as expected under saturation.

Heterogeneity

The tested fleet combined:

1. NVIDIA L4 24 GB GPUs;
2. NVIDIA H100 80 GB GPUs;
3. Models ranging from 0.8B to 675B parameters;
4. Two independent cloud providers
5. One API and control layer.

7. How We Measured It

Benchmarks are easy to get wrong in ways that flatter the result. We document the methodology because we believe it is a stronger credibility signal than the headline figures alone.

Time to First Token

Naive timing returned values four to ten times lower on the same requests because it was measuring protocol activity rather than generated output. Time to first token was measured at the first non-empty content delta in the streamed response.

Same-Origin, Same-Run Comparisons

Path comparisons were conducted from the same origin and within the same run.

Comparing one measurement taken from a user device with another taken from inside the data centre can conflate hundreds of milliseconds of network distance with the infrastructure characteristic being tested.

Routing Verification

Routing was verified rather than assumed. Each benchmarked request confirmed which backend served it rather than relying only on the configured route.

Regional Test Method

Regional latency was measured from client instances physically located in the origin region, so each figure reflects the network path an actual user there would traverse. Within a run, only the destination endpoint changed; origin, model, payload and load level were held constant. The comparison therefore isolates placement, which is the variable being tested.

Known Limitations

The figures are point estimates and ranges from repeated runs, not statistically powered results with confidence intervals.

Comparisons against third-party frontier-model leaderboards have not been the focus of our testing as we are benchmarking the inference orchestration, not the performance of the models themselves.

8. Efficiency, Sustainability and the Regulatory Floor

German and European rules impose efficiency obligations on data centres, including requirements relating to power usage effectiveness, renewable-electricity coverage on an accounting basis, waste heat and reporting.

The regulatory regime was still evolving as of July 2026. An amendment to the German Energy Efficiency Act had cleared the Federal Cabinet and proposed changes to applicability thresholds, deadlines and waste-heat obligations, but had not completed the parliamentary process. The existing law therefore continued to apply at the time of the Alpha.

CYANIS AI takes a conservative design position:

Design for the requirements currently in force rather than relying on anticipated regulatory relief.

Capacity built to a high efficiency standard should require less remediation later, regardless of how the amendment process develops.

9. What Comes Next

The Alpha validated the orchestration software.

Work on the Beta has begun: the same control layer moves onto commercial, energy-linked infrastructure, and placement extends from regional toward country and sub-regional granularity.

In parallel, a staged 10 MW commercial rollout has kicked off, covering engineering and site readiness, hardware procurement and integration, and commercial preparation. The first tranche validates the deployment unit; every stage after it repeats that unit at scale.

About CYANIS AI

CYANIS AI is building distributed AI-inference infrastructure across Europe.

The company plans to deploy modular GPU capacity at sites where power, grid connection and fibre already exist and connect that capacity through a proprietary orchestration layer into one enterprise service under European data-placement rules.

CYANIS AI GmbH is registered in Hamburg, Germany. The group holds a pending German patent application relating to the orchestration of energy-linked and distributed AI infrastructure.

CYANIS™ is a registered trademark.

cyanis.ai · Partnership and capacity enquiries welcome · Further technical detail is available under NDA

Notes

All performance figures were measured during the Alpha in July 2026 on NVIDIA L4 24 GB and H100 80 GB GPUs running open-weight models, hosted on European cloud infrastructure, with orchestration validated across two independent cloud providers.

The compute-reduction figure is benchmark-derived from the measured routing distribution and relative memory footprint of the models involved. It is not a direct measurement of GPU time and is quoted with its corresponding benchmark retention.

Market and regulatory statements reflect the information contained in the supplied July 2026 draft and should not be relied upon as legal or investment advice.