Real-Time AI Inference at Enterprise Scale

Solution Brief

Silk is the only software-defined SAN built to deliver live context data to AI inference workloads — at enterprise scale, without compromising production.

Your AI pilot was promising. Now, as it moves toward production, powering real-time risk recalculation, clinical decisioning, logistics optimization, or embedded AI in your SaaS platform, it's breaking down. Latency spikes unpredictably. Infrastructure costs climb. Your engineering team patches around the problem with replicas, ETL pipelines, and isolated environments that introduce data lag. Model accuracy suffers. Business outcomes slip.

The model isn't the problem. Your data infrastructure is.

AI inference generates short, intense I/O bursts, extreme read concurrency, and latency sensitivity that overwhelms storage stacks designed for predictable, human-scale workloads. Traditional fixes, overbuilding infrastructure, tolerating stale replicas, isolating AI workloads, drive costs up while sacrificing the live data grounding your models need to perform.

Cost Avoidance: What's at Risk

Every month this goes unresolved, your AI investment underperforms, your engineering team carries unnecessary operational risk, and competitors who've solved the infrastructure layer move faster.

Why Everything You've Already Tried Has a Ceiling

Hyperscaler-native storage, traditional SANs, and database replicas share a common flaw: they were built for predictable workloads. When AI inference, analytics, and OLTP (Online Transaction Processing) compete for the same data layer, something always breaks — performance, data freshness, or cost discipline.

Most teams don't discover this is an infrastructure problem until they've already overspent on compute, restructured their database architecture, or quietly accepted that their AI systems will always run on slightly stale data.

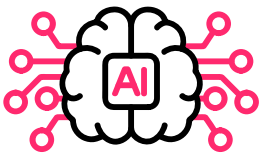
Built for the Problem Everyone Else Designed Around

Silk is a software-defined SAN and cloud acceleration layer built specifically for real-time AI inference on live operational data. It sits beneath your existing databases and applications — no code changes, no query rewriting — and dynamically governs I/O behavior so AI inference, analytics, and OLTP workloads safely coexist.

Three capabilities separate Silk from every alternative:

Silk Live (performance decoupled from capacity)	**Silk Flow** (patented adaptive I/O)	**Silk Echo** (zero-footprint clones)
Inference demand spikes are absorbed without resizing infrastructure, delivering stable response times and measurably lower cost per inference at scale. No more overprovisioning to stay safe at peak load.	Autonomously detects I/O sizes per workload and adjusts block sizing and throughput in real time. Your OLTP systems stay fast. Your inference workloads get the throughput they need. No noisy neighbors, no manual tuning.	Instant, point-in-time database clones let AI systems consume live context without duplicating data or touching production systems. Replace ETL pipelines and delayed replicas with immediate, accurate data access.

What Changes for Every Leader in the Room



For AI and Infrastructure leaders:

Deterministic performance under extreme concurrency. Safe coexistence of AI and mission-critical systems. Elimination of the operational risk that comes with mixed workload contention.

For CTOs and Technology executives:

Faster pilot-to-production timelines. Infrastructure that scales with AI demand without linear cost increases. Reduced engineering overhead from manual tuning and workload isolation.

For Business and Finance leaders:

AI initiatives that deliver on their ROI promise. Models grounded in live data produce more accurate outputs — better decisions, fewer errors, competitive advantage that compounds.

ptc

What It Looks Like When the Cloud Works for AI

PTC embeds AI directly into Windchill+, powering intelligent part reuse, visual search, and real-time engineering assistance for enterprise customers worldwide. Running on Silk, PTC delivers AI inference directly against live Oracle databases — nearly 1PB under management — with sub-millisecond data retrieval, zero ETL, and no performance risk to production systems.

Result: SaaS-scale AI performance, faster feature delivery, and mission-critical reliability — simultaneously.

50% reduction in infrastructure costs	**70%** reduction in manual effort to complete DB operations	**90%** reduction in time needed to complete DBOperations

Control at the Data Layer Is Now a Competitive Advantage

The organizations that solve it first will move faster, spend less, and build more accurate AI systems than those still working around it. Silk is purpose-built to make that transition possible without rearchitecting what you've already built.

Learn more at www.silk.us/artifical-intelligence

About Silk

Silk is an adaptive, software-defined SAN that serves as the cloud acceleration layer for mission-critical workloads, making cloud performance predictable at scale. It delivers real-time data access, consistent application performance, and instant recovery - without performance-driven overprovisioning or application refactoring.

Trusted by enterprises across finance, healthcare, SaaS, and more, Silk accelerates databases, analytics, and AI workloads with built-in resiliency. By enabling independent scaling of performance and capacity, Silk helps organizations unlock the full value of their data.