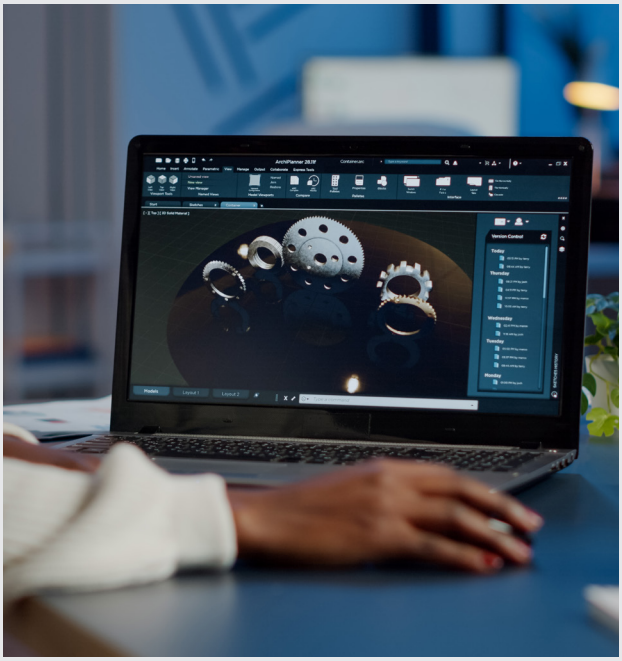
sonet.io

White Paper

Solving the GPU Bottleneck: Why Infrastructure Isn't the Answer

The Growing Demand for GPU Access



The rise of hybrid work, global teams, and application-heavy workflows has made GPU access a critical need, not just for engineers, but for operations teams across industries. From CAD reviews to 3D visualization to quality assurance, GPU-enabled applications are no longer the exception. They're part of day-to-day work.

Yet most organizations still treat GPU delivery like it's 2015: shipping expensive laptops to remote users, provisioning dedicated cloud desktops, or maintaining complex VDI infrastructure. These approaches are slow, rigid, and expensive. They create bottlenecks where there should be flexibility.

According to a 2024 survey reported by insideHPC, the majority of enterprise GPUs remain underutilized even during peak periods, and just 42% of companies have the ability to manage GPU partitioning to improve utilization. While 40% plan to implement orchestration and scheduling technologies, most organizations still lack the agility to align GPU compute with real-world usage patterns (insideHPC, March 2024).

Most users don't need full-time access to high-performance hardware.They need secure, reliable GPU access when the work calls for it. This paper explores why traditional infrastructure-heavy delivery models no longer fit, and how a browser-based approach enables IT teams to deliver the performance users need without the cost and complexity of overprovisioned hardware.

The Infrastructure Bottleneck

The traditional way to deliver GPU access is hardware-heavy and resource-intensive. Organizations ship pre-configured laptops

with dedicated GPUs, stand up virtual desktop infrastructure (VDI), or spin up full-blown cloud desktop environments. These methods assume every user is a full-time power user and that IT has unlimited time and budget to support them.

The result? Massive overprovisioning and painful delays.

Shipping and procurement cycles slow down onboarding, especially for global teams and contractors. Imaging, configuring, and securing these machines takes days or weeks. In many cases, users don't receive their GPU-equipped device until after the work they needed it for has already started.

Legacy delivery models also introduce significant overhead. VDI and cloud desktops require constant management, security patches, licensing, and infrastructure investment. They're rigid, expensive, and often still fail to deliver the performance users need.

And for every fully utilized machine, there are five that sit idle waiting for the next CAD review or design sprint.

According to insideHPC, even at peak times, most enterprise GPUs remain underutilized, and organizations are still struggling to implement the orchestration tools needed to align compute availability with actual user demand.

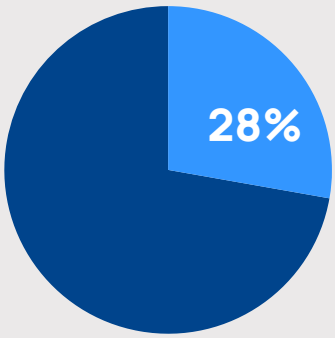
These infrastructure-based approaches aren't just inefficient, they're misaligned with how modern teams work. Users need access that matches the pace and scale of their projects, not a $5,000 workstation collecting dust between bursts of activity.

What the Industry Data Tells Us

The challenges around GPU delivery aren't hypothetical—they're playing out daily inside IT teams. From underutilization to allocation conflicts to security concerns, organizations are under growing pressure to rethink how they support GPU-enabled applications.

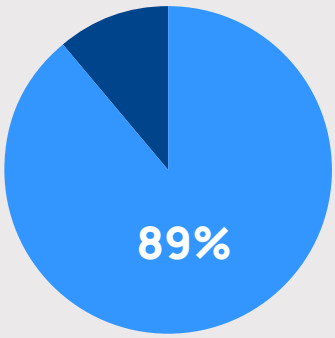
According to Run:AI's 2023 State of AI Infrastructure Survey

Only 28% of organizations have on-demand access to GPU compute.



Most still rely on manual ticketing systems to allocate resources, making flexible access nearly impossible.

89% of organizations face allocation issues regularly, with 40% encountering GPU conflicts weekly and 13% daily.



These aren't rare pain points—they're persistent blockers that delay work, create internal friction, and lead to shadow IT practices.

The problem isn't just about access. It's about timing and alignment. Most teams don't need GPUs full-time. They need access for periodic tasks: QA checks, CAD updates, rendering passes, or visualizations. But with existing infrastructure, IT has no way to provide that kind of burst-ready compute without overprovisioning or introducing delays.

Compounding the issue, nearly half of surveyed IT leaders expect reduced GPU access budgets in 2025, even as demand continues to grow. This financial pressure underscores the need for more efficient, right-sized GPU delivery strategies.

Security and compliance are also key concerns. As TierPoint reports, GPU cloud workloads often involve sensitive data making robust access controls, encryption, and regulatory compliance critical. Especially for teams working across borders or bringing in external collaborators, the traditional methods of delivering GPU access can increase risk, not reduce it.

The takeaway is clear: today's GPU provisioning models are slow, rigid, and costly. What teams need instead is a secure, flexible, usage-aligned delivery model, one that can scale without hardware, delays, or overbuilding.

Provisioned for the Past: Why IT Needs a Smarter GPU Model

Most GPU delivery strategies were built around a different era of work: centralized offices, full-time power users, and predictable workflows. But today's teams are distributed, dynamic, and driven by short bursts of high-performance needs.

Despite that shift, IT teams are still forced to overprovision expensive infrastructure "just in case" someone might need it. Whether it's a $5,000 GPU laptop shipped to a contractor, a persistent VDI instance running 24/7, or pre-allocated cloud desktops for every user, these models assume full-time usage that rarely materializes.

Overbuilding Creates a Triple Bind

Wasted Spend:
Resources sit idle while IT budgets get squeezed.

Delayed Onboarding:
Time-to-productivity suffers when access depends on hardware procurement or VDI provisioning.

Security and Compliance Risk:
Broad access to powerful resources increases the attack surface, especially when physical devices are involved.

And perhaps most critically, these models make it hard to scale when it matters most. When demand spikes, say during a design sprint, R&D push, or QA cycle, IT can't move fast enough. On the flip side, when work slows down, those same resources continue incurring cost.

The consequences aren't minor. A recent Techstrong.ai survey of 1,367 AI professionals found that 85% of organizations have delayed projects due to lack of GPU access, with 39% experiencing delays of three to six months. These aren't edge cases, they're daily realities caused by rigid infrastructure and global chip scarcity.

It's time for a smarter GPU model: one that treats compute as a flexible, secure, on-demand resource delivered when needed, and only for as long as it's needed.

The Sonet.io Approach

Sonet.io delivers GPU access that aligns with the way teams actually work—short bursts of high compute need, distributed contributors, and variable usage patterns that don't justify full-time infrastructure.

Instead of provisioning physical laptops or overbuilding static environments, Sonet.io offers secure, browser-based access to GPU-enabled applications with no agents, no shipping, and no long lead times. It gives IT the control to scale up or down based on project needs while giving users the performance they expect, right when they need it.

Whether it's a product designer updating a CAD model once a week or an R&D contractor validating a simulation monthly, Sonet.io ensures GPU power is available without overprovisioning, IT intervention, or idle hardware. GPU access becomes just another cloud-delivered service, controlled by policy, metered by usage, and secured by default.

With Sonet.io, IT teams can:

- **Add GPU-enabled users in minutes, not weeks**
- **Avoid CapEx-heavy hardware refresh cycles**
- **Eliminate data exposure from device shipping or unmanaged endpoints**
- **Support bursty or periodic workflows without performance trade-offs**

The result? Lower costs, faster onboarding, stronger security and finally, a GPU access model that scales with your business, not against it.

Use Cases and Deployment Models

GPU-powered workflows aren't limited to AI researchers or full-time 3D artists. Across industries, teams increasingly rely on high-performance compute to review models, render visualizations, validate simulations, and analyze data.

What's changed isn't the need for power, it's the frequency and distribution of that need. Sonet.io supports a range of use cases that don't justify full-time infrastructure, but still require high-performance access.

Engineering & Design Reviews

Mechanical engineers, architects, and product designers often need to open large CAD files, make minor updates, or run simulations, tasks that don't justify dedicated GPU workstations but can't run effectively on a laptop.

Simulation & Validation Workloads

From QA teams to R&D units, simulation cycles often happen in bursts, especially late in development. Sonet.io gives these users access only when needed, without incurring idle cost.

Remote & Field-Based Contributors

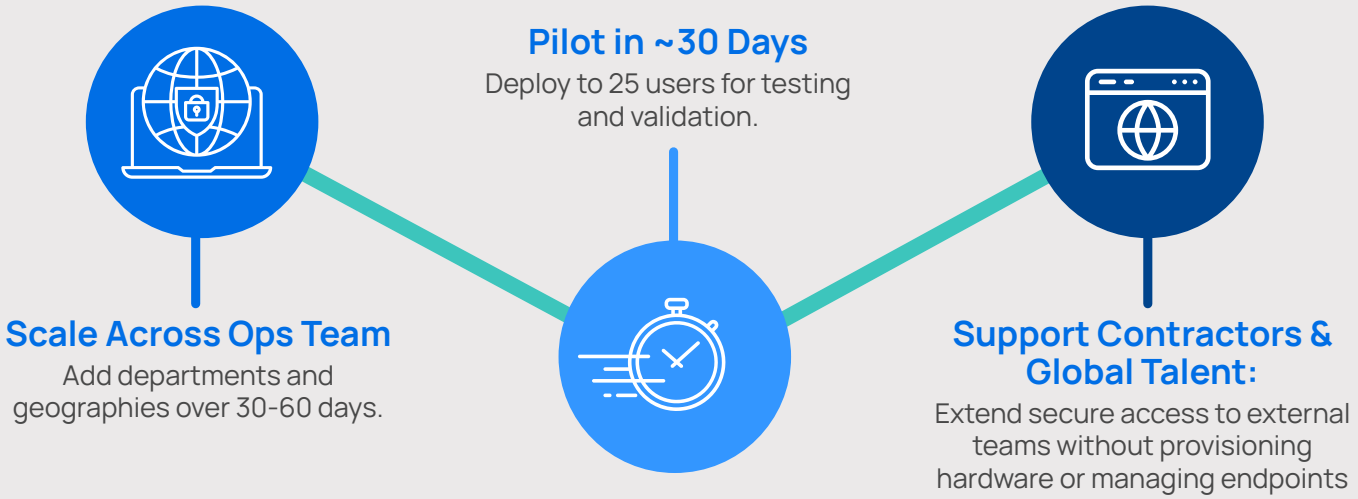
Field engineers or global contractors often need secure access to GPU-powered applications for troubleshooting, inspection, or updates. With Sonet.io, they can access those tools without shipping devices or punching holes in the network.

Visualization & Rendering

Creative and technical users need high-performance environments for rendering 3D models, visualizations, or animation. Sonet.io delivers this power through the browser: scalable, secure, and globally accessible.

Deployment Models That Fit

Sonet.io is designed to integrate with your environment and roll out on your terms



No infrastructure, no re-architecting, and no learning curve for users. Just GPU power when and where you need it.

Next Steps

The GPU bottleneck isn't a hardware problem, it's an access problem. While demand for GPU-accelerated applications continues to grow, the traditional models for delivering that compute are too slow, too rigid, and too expensive to keep up.

Sonet.io offers a better path: right-sized GPU access, delivered securely via browser, without the delays and cost of legacy infrastructure. Whether you're supporting 25 power users or enabling 2,500 global contractors, Sonet.io scales with your needs giving users the performance they require, and IT the control it demands.

Why Customers Choose Sonet.io

- Cut hardware and provisioning costs by up to 80%
- Deploy GPU access in minutes, not weeks
- Reduce risk with built-in Zero Trust security
- Support hybrid, global, and contractor workforces without compromise

Ready to Modernize Your GPU Strategy?

- Start with a pilot: Validate Sonet.io with a 25-user deployment
- Scale at your pace: Expand to full teams in 30-60 days
- Eliminate the need for device shipping, CapEx, and overprovisioning

See how fast, secure GPU access should work

Book a Demo Today

Contact Us :

sales@sonet.io

sonet.io

Work Securely From Any Device

sales@sonet.io

Find us online at **sonet.io**

3031 Tisch Way, 110 Plaza West
San Jose, CA 95128