AI PLAYBOOK

Quali | NVIDIA

Why You Don't Need to Keep the Lights On:

Smarter AI Infrastructure Management with Quali Torque and NVIDIA

AI generated Image

Smarter AI Infrastructure Management with Quali Torque and NVIDIA

AI is becoming an integral part of business operations, and its increasing reliance on **GPU-based workloads** is creating significant financial and operational challenges for IT organizations. The demand for AI isn't just about technical feasibility anymore, it's about **cost efficiency, sustainability, and business alignment**.

Many believe the high costs associated with AI workloads will stabilize over time, assuming economies of scale will eventually lower prices. But the reality is, **the challenge goes deeper**. GPUs run significantly hotter than CPUs, leading to greater cooling requirements, higher power consumption, and ultimately, skyrocketing operational costs. Beyond the financial strain, there's a growing environmental impact to consider, with regulatory pressures mounting as businesses are increasingly held accountable for their carbon footprints.

The High Cost of Running AI Workloads 24/7

For years, organizations have been leaving workloads running continuously, largely out of fear that stopping and restarting infrastructure could lead to downtime, lost productivity, or unexpected failures. This practice, which may have been acceptable with CPU-based workloads, is becoming unsustainable with GPUs.

Running a high-performance AI workload 24/7 can cost over **$23,000 per month**, compared to just **$500 for CPU-based workloads**. GPUs also consume over **20 times** the power of their CPU counterparts, adding another layer of operational complexity and expense.

Source	Hourly Cost	Monthly Cost (24×7)	Power Consumption
High-End CPU	$0.68/hr.	~$500	~150 watts
NVIDIA GPU	$32.77/hr.	~$23,670	~3,200 watts

This massive cost difference isn't just a budgeting headache; it's an urgent signal that businesses need **smarter AI infrastructure strategies** to avoid waste and inefficiency.

AI Workloads Consume More Than Just Power

To put things into perspective, training a single large AI model, such as **GPT-3**, can consume as much energy as **120 U.S. homes in a year**. The problem isn't simply about keeping the lights on, it's about understanding **when and how** to turn them on and off.

The flexibility of cloud computing has made it easy to let AI-optimized instances run indefinitely. Many organizations leave resources running 'just in case' they're needed, leading to ballooning operational costs and a growing carbon footprint. In today's world, where financial sustainability is just as important as environmental sustainability, this approach is no longer viable.

Smarter AI Workload Management with Quali Torque and NVIDIA

Running AI workloads 24/7 isn't the only way to operate. In fact, it's one of the **worst** ways to do so. Implementing smarter scheduling and automation can **reduce GPU costs by over 50%**, ensuring resources are only used when they truly provide value.

For example, running workloads for **10 hours per day, Monday through Friday**, instead of continuously, can bring monthly costs down from **$23,670 to $9,830**, a savings of more than $13,800.

Source	Hourly Cost	Monthly Cost -10 hrs./day, weekdays	Savings
High-End CPU	$0.68/hr.	~$215	57%
NVIDIA GPU	$32.77/hr.	~$9,830	58%

Recognizing the cost implications is just the first step. The real challenge is **controlling infrastructure consumption without inhibiting innovation or disrupting performance**, and that's where **Quali Torque**, in collaboration with **NVIDIA**, provide the most advanced and game-changing solution.

Torque offers an intelligent, automated approach to AI workload management, **aligning resource usage with real business needs, enforcing governance, and providing deep visibility** into consumption patterns.

Torque's advanced automation capabilities ensure workloads run **only when necessary**, leveraging smart scheduling, real-time monitoring, and policy-driven governance to:

- **Reduce cloud spend** by dynamically allocating resources based on demand.
- **Eliminate waste** through automatic shutdown and restart schedules.
- **Maintain optimal performance** without the risk of disruption.
- **Achieve sustainability goals** by reducing power consumption and environmental impact.

By adopting a strategic approach powered by **Quali Torque and NVIDIA**, businesses can confidently scale their AI initiatives while maintaining full control over operational costs and performance. The future of AI infrastructure management isn't about running workloads indefinitely, it's about running them **intelligently and purposefully**.

Quali Torque: Smart AI Infrastructure Management

To overcome the challenges of runaway GPU costs and energy inefficiency, businesses need **automated governance, intelligent consumption policies, and real-time workload optimization. Quali Torque**, in partnership with **NVIDIA**, empowers organizations with the tools needed to optimize AI infrastructure **with precision and purpose**, ensuring that workloads run only when and where they provide value.

Torque provides a set of capabilities that help businesses optimize AI operations while maintaining agility and control:

These blueprints collectively provide a comprehensive framework that addresses the critical aspects of building AI agent applications:

Automated Start-Stop Schedules	• Workloads are automatically **shut down during off-hours** and restarted when needed, reducing costs without affecting operations. • Environments restart seamlessly with preserved configurations, so teams don't need to worry about disruptions.
Smart Auto-Scaling with Advanced Intelligence	• GPU resources are **dynamically provisioned** based on demand, ensuring only the required capacity is used. • AI-driven predictive scaling matches infrastructure needs with workload patterns, eliminating manual adjustments. • Underutilized resources are deallocated in real time, preventing costly idle consumption.
Budget and Usage Controls	• Consumption caps can be set to **avoid unexpected cost overruns**, ensuring financial predictability. • Real-time monitoring provides alerts when usage exceeds limits, keeping costs in check.
Cost Visibility and Governance	• Detailed dashboards provide a **clear picture of AI spending**, helping teams make data-driven decisions. • Governance policies ensure infrastructure usage is aligned with business priorities and compliance requirements.
Purpose-Driven Workload Alignment	• AI workloads are evaluated to ensure they run **only when they provide tangible value**, eliminating waste. • Compliance policies ensure that GPU usage aligns with sustainability and business objectives.

By embracing these capabilities, businesses can move from an inefficient **‘always-on’** approach to a **‘right-time’** strategy, where infrastructure runs when it matters most.

With these capabilities, businesses can **eliminate the inefficient 24/7 model** and instead embrace a ‘right-time’ approach that optimizes both costs and performance while ensuring workloads align with business purpose.

Quali Torque and NVIDIA are driving the industry forward by providing enterprises with a controlled, managed, and optimized approach to GPU infrastructure, ensuring that AI innovations deliver maximum value without the burden of runaway costs.

Overcoming the Fear of Shutting Down AI Workloads

One of the biggest challenges organizations face is **the fear of shutting down workloads**. Concerns about losing data, configuration drift, or delayed restart times often lead teams to leave infrastructure running unnecessarily.

Quali Torque eliminates these concerns by ensuring environments can be safely decommissioned and restarted without complexity. Torque preserves the state of AI workloads, automates environment restoration, and prevents operational hiccups, making it easier to adopt a cost-efficient infrastructure strategy.

NVIDIA and Quali Torque: Pushing AI Efficiency Forward

Together, **NVIDIA and Quali Torque** are driving AI infrastructure optimization forward by offering a **controlled, managed, and cost-effective approach**. This collaboration enables businesses to take full advantage of GPU capabilities without the fear of runaway costs.

By leveraging Torque’s intelligent automation and NVIDIA’s powerful GPUs, organizations can confidently scale AI workloads with:

- **Seamless integration into existing cloud environments.**
- **Automated enforcement of usage policies to prevent waste.**
- **Self-service portals that provide teams with instant access to pre-configured, cost-optimized AI environments.**

A Smarter Path Forward for AI Infrastructure

The days of leaving AI workloads running out of habit are over. With **Quali Torque and NVIDIA**, businesses can achieve the best of both worlds, **the power of AI without the burden of excessive costs**.

By making AI infrastructure **smarter, more sustainable, and more cost-effective**, organizations can focus on innovation without worrying about cloud waste or unpredictable bills.