AI PLAYBOOK

Quali | NVIDIA

Simplifying AI Workloads

How Quali and NVIDIA Drive Scalable Innovation

Streamlining AI Operations with Intelligent Automation, Unified Management, and Scalable Innovation

AI generated Image

Does Your AI Infrastructure Strategy Unlock NVIDIA's Full Potential?

The rapid adoption of AI in sectors like finance, healthcare, and advanced manufacturing has elevated the demand for high-performance infrastructure. However, many organizations face significant challenges in optimizing their AI workloads, particularly with NVIDIA GPUs, which are the cornerstone of modern AI innovation.

Key Challenges

AI Workload Complexity

Traditional infrastructure tools struggle to support the dynamic nature of AI/ML workloads, often leading to underutilization of NVIDIA GPUs and wasted resources.

Fragmented Management

Siloed tools and workflows make it difficult to orchestrate GPU-accelerated environments across multi-cloud and hybrid deployments.

Rising Costs

Without real-time optimization, organizations often face runaway expenses in GPU utilization, reducing the ROI of their AI initiatives

Quali's Approach

Quali Torque transforms these challenges into opportunities. With Torque, organizations can:

- Optimize GPU performance with intelligent orchestration and dynamic resource scaling, ensuring that every GPU cycle is maximized.
- Unify AI workload management across on-premises, cloud, and edge environments for seamless operations.
- Align infrastructure utilization with business priorities, embedding governance and cost optimization directly into AI workflows.

By providing a single platform that integrates NVIDIA GPU infrastructure into a cohesive management strategy, Quali helps organizations unlock the full potential of their AI investments, driving scalability, efficiency, and innovation.

This new paradigm enables businesses to achieve unparalleled performance and cost control, ensuring their NVIDIA GPU deployments contribute directly to transformative outcomes.

Gartner

Global IT spending is projected to surpass **$7 trillion** by 2028, with generative AI (GenAI) accounting for over **$1 trillion** of that expenditure.

Source: Gartner IT Spending Forecast

How to Manage AI Infrastructure Challenges with Quali Torque

Leveraging Quali Torque simplifies AI infrastructure management, accelerates innovation, and optimizes resources by addressing the complexities of AI workload orchestration.

Innovate with Confidence

Simplify AI workload orchestration across hybrid and multi-cloud environments.

Enabled: Torque provides turnkey, ready-to-run blueprints for AI environments, seamlessly integrating on-premises and cloud resources under a unified governance model.

Value Delivered: Secure, compliant, and consistent infrastructure that supports AI model development without compromising on innovation.

AI Model Performance

Deliver GPU-optimized environments tailored to your workload demands.

Enabled: Torque ensures AI workloads are placed in GPU environments that maximize resource utilization, offering real-time orchestration and high-speed connectivity for demanding AI tasks.

Value Delivered: Reduced latency, enhanced GPU performance, and faster AI model iteration cycles.

Teams Focus on Innovation

Automate infrastructure management, reducing operational overhead.

Enabled: Torque automates provisioning, policy enforcement, and lifecycle management, enabling teams to focus on high-value tasks rather than manual setup and monitoring.

Value Delivered: Improved efficiency, less reliance on specialized skills, and greater bandwidth for strategic AI projects.

Optimize Costs & Reduce Waste

Control AI infrastructure spending with real-time visibility and governance.

Enabled: Torque's dashboards provide actionable insights into resource utilization and costs, while proactive waste management prevents unused infrastructure from inflating expenses.

Value Delivered: Predictable, optimized infrastructure costs aligned with business priorities

Maximizing GPU Investments with Quali Torque:
Turning Challenges into Opportunities

AI and GPU workloads are transforming IT infrastructure, driving innovation but posing challenges. GPUs are powerful and high-value resources requiring effective management to prevent inefficiencies and overspending. Torque provides the control and visibility needed to maximize GPU investments on-premises or in the cloud.

The GPU Waste Challenge: A Crisis in Utilization

Organizations waste up to 30% of their cloud budgets due to inefficiencies, with overprovisioning of AI resources being a common cause of cost overruns. As the GPU market is projected to grow to $1.4 trillion by 2034, managing these high-value resources is increasingly critical. (1)

What's Needed? Real-Time Utilization Tracking:

Dashboards monitor GPU consumption to identify underutilized resources.

Proactive Waste Removal:

Automatically decommissions idle GPU instances, ensuring efficient resource allocation.

Outcome:

Organizations minimize waste, ensuring every GPU dollar directly supports innovation and value creation.

Energy and Sustainability: The Hidden Cost of GPUs

A single AI model training run, such as GPT-3, can consume 1,287 MWh—the equivalent of powering 120 U.S. homes for a year. Power and cooling costs for on-premises GPU deployments are becoming a critical concern. (2)

What's Needed? Environment-Aware Deployment:

Optimizes workload placement to balance GPU utilization and power consumption.

Cooling and Power Efficiency:

Integrates cooling capacity and energy considerations into resource allocation.

Outcome:

Reduced energy costs and environmental impact without compromising performance.

Cloud Costs: Managing the $187 Billion Problem

Global cloud spending has reached $187 billion annually, with GPU costs adding complexity due to their intensive consumption patterns and high price tag (3)

What's Needed? Cost Governance Tools:

Aligns GPU usage with business objectives for predictable and justifiable costs.

Workload Prioritization:

Dynamically adjusts AI task alignment to fit budgets while maintaining performance

Outcome:

A balanced approach to cost and performance, ensuring GPU investments drive measurable business value.

(1) Source: McKinsey ; (2) University of Michigan ; (3) Trendforce & FinOps Foundation

Revolutionizing GPU Workload Management: Quali Torque's Transformative Approach

Total Control for AI and GPU Workloads: Unlocking the Full Potential of NVIDIA GPUs with Torque.

As AI and GPU workloads reshape IT infrastructure, the need for comprehensive management and optimization has never been greater. NVIDIA GPUs are at the forefront of this revolution, powering advanced AI models, high-performance computing, and data-driven innovation.
However, without the right tools, organizations risk inefficiencies, spiraling costs, and underutilized resources. Quali Torque redefines how businesses manage infrastructure and GPU workloads, offering total visibility, control, and alignment with business objectives to ensure every GPU cycle delivers value.

Maximizing GPU Utilization

Quali Torque integrates seamlessly with NVIDIA GPUs to provide real-time insights into workload performance and resource allocation. By continuously monitoring GPU usage across environments, Torque identifies underutilized or idle resources and optimizes their deployment. Automation ensures that GPUs are allocated where they are needed most, without manual intervention.

Organizations achieve higher efficiency and resource utilization, ensuring every GPU investment contributes to innovation and operational performance. This reduces waste and improves ROI, particularly for large-scale AI workloads.

Controlling Costs and Optimizing Expenditure

Torque embeds cost management into every layer of infrastructure, aligning GPU usage with strategic business priorities. Its cost governance tools provide visibility into spending and allow dynamic workload prioritization, ensuring critical AI tasks take precedence without exceeding budget constraints.

Predictable, optimized infrastructure costs empower organizations to scale AI initiatives confidently. By eliminating unnecessary expenses and aligning spending with value, businesses can fully capitalize on their NVIDIA GPU investments.

Unifying GPU Workload Mgmt.

Torque simplifies operations by providing a unified control plane for managing hybrid, on-premises, and multi-cloud GPU environments. With built-in governance policies and automation, Torque ensures consistent operations and compliance across all deployment models.

Organizations benefit from streamlined operations and scalable infrastructure mgmt. Whether deploying AI workloads on-premises, in the cloud, or at the edge, Torque ensures NVIDIA GPUs are fully utilized and easily managed across environments

Optimizing AI Workloads with Quali Torque: Unified Management, Cost Control, and Efficiency

The following sections explain how Torque empowers organizations to maximize NVIDIA GPU value, control costs, and unify AI workload management across diverse environments

1

Maximizing GPU Utilization:

Harnessing the Full Power of NVIDIA GPUs

2

Controlling Costs:

Aligning GPU Investments with Business Priorities

3

Unified Management:

Simplifying GPU Workloads Across Hybrid and Multi-Cloud Environments

1 Maximizing GPU Utilization

Harnessing the Full Power of NVIDIA GPUs

Why Optimize GPU Utilization?

The adoption of GPUs has revolutionized computing, particularly for AI and high-performance workloads. NVIDIA GPUs deliver industry-leading performance, enabling organizations to accelerate AI model training, execute complex data processing, and unlock real-time insights.

However, to fully harness their potential, GPUs must be utilized efficiently. Idle or underutilized GPUs represent wasted investments and missed opportunities to innovate.

NVIDIA's cutting-edge technology, like the A100 and H100 GPUs, offers unparalleled power, scalability, and efficiency.

These GPUs provide the computational horsepower to tackle the most demanding AI and HPC workloads, with features like multi-instance GPU (MIG) for workload isolation and elasticity. However, maximizing their value requires intelligent orchestration and real-time optimization. This is where Quali Torque excels.

Without proper optimization, organizations face:

- **Escalating Costs:** Underutilized GPUs drive up expenses without contributing to ROI.
- **Performance Bottlenecks:** Misaligned workloads result in delayed AI model training and inefficiencies.
- **Resource Fragmentation:** Managing GPUs across hybrid and multi-cloud environments can lead to operational silos.

1 Harnessing the Power of NVIDIA GPUs with Torque

Intelligent Management & Optimization

The Challenge:

Organizations struggle to manage GPU workloads effectively due to resource fragmentation, siloed tools, and the lack of real-time insights. This results in underutilized GPUs, inflated costs, and reduced innovation capacity.

Our Solution:

Torque integrates seamlessly with NVIDIA GPUs, leveraging real-time resource monitoring, dynamic workload orchestration, and proactive optimization to ensure every GPU cycle is maximized. By providing a unified control plane for hybrid and multi-cloud environments, Torque eliminates inefficiencies and enables organizations to align GPU utilization with business objectives.

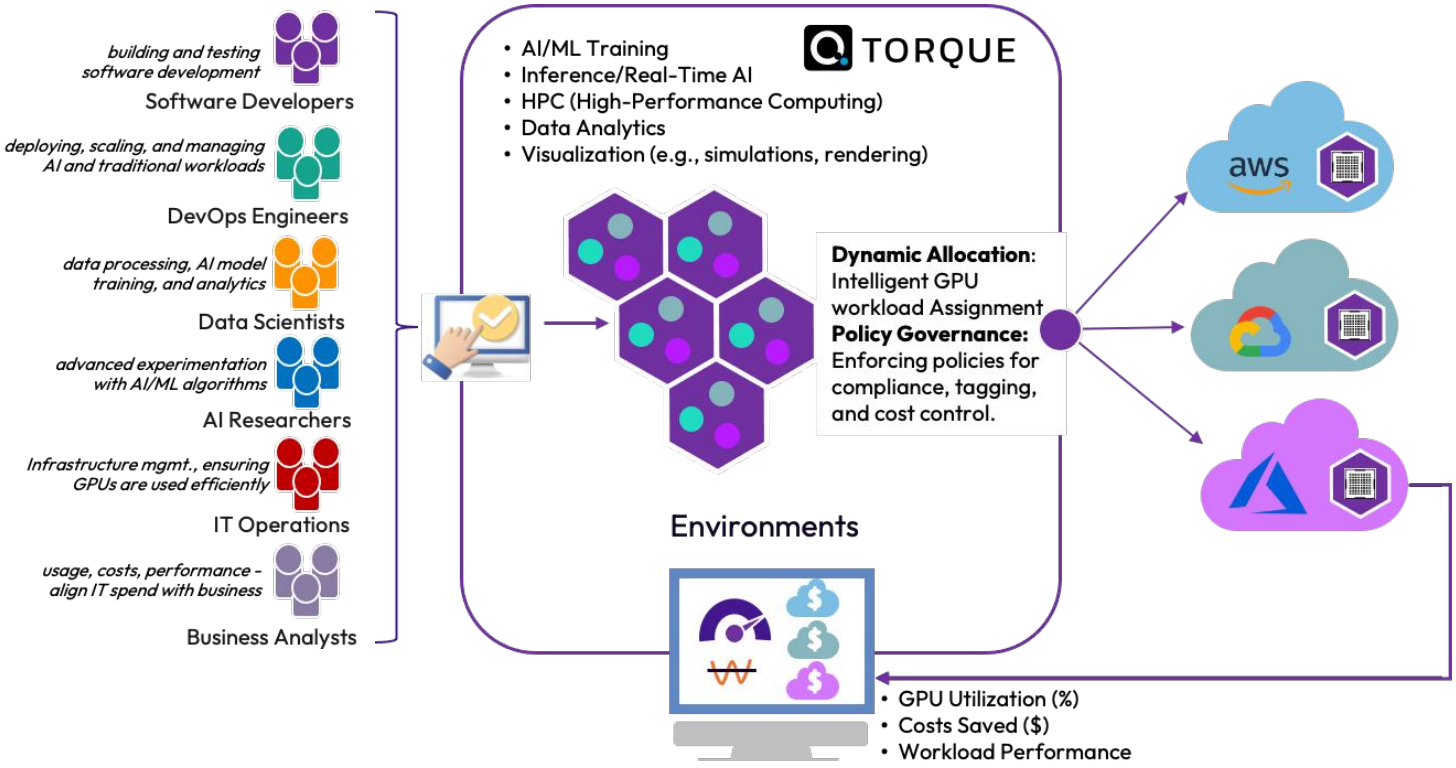
How Torque Maximizes GPU Resources

Torque takes AI to manage AI, leveraging its agentic AI-driven workflows to optimize GPU utilization across environments. By combining intelligence with automation, Torque ensures that every GPU cycle delivers measurable value. Here's how Torque achieves this in five key steps:

1. **Real-Time Resource Discovery:** Torque automatically detects and inventories GPU resources across on-premises, cloud, and hybrid environments, providing a unified view of all assets.
2. **Dynamic Workload Allocation:** Using AI-driven decision-making, Torque allocates GPU resources to workloads based on performance requirements and business priorities, ensuring optimal utilization.
3. **Proactive Waste Identification and Removal:** Torque identifies idle or redundant GPU instances in real time and decommissions them automatically, preventing resource wastage.
4. **Scalable Resource Orchestration:** Torque supports NVIDIA's MIG technology, enabling organizations to partition GPU resources for multiple workloads, maximizing flexibility and efficiency.
5. **Continuous Monitoring and Optimization:** Torque provides real-time dashboards and predictive analytics, enabling teams to monitor GPU performance and fine-tune workloads dynamically.

By embedding AI into its infrastructure management, Torque ensures that AI workloads are intelligently managed, delivering on the promise of NVIDIA's advanced GPU capabilities

1 Unlocking GPU Potential with Intelligent Optimization



1. Users initiate requests via the Self-Service Catalog for environments tailored to specific NVIDIA GPU workloads.

2. Torque dynamically orchestrates workloads across environments, for optimal GPU usage through policy enforcement and real-time insights.

3. NVIDIA GPUs are partitioned using MIG, allowing multiple workloads to run efficiently on the same hardware.

4. Policies ensure compliance, cost alignment, and workload prioritization while real-time dashboards provide visibility into operations.

5. Optimized environments enable higher utilization and significant cost savings, with measurable outcomes highlighted in the dashboard.

Organizations using Quali Torque for NVIDIA GPU workloads achieve up to **30% greater utilization efficiency** and **reduce operational costs by up to 40%** while improving workload performance across AI and traditional tasks.

2 Controlling Costs

Aligning GPU Investments with Business Priorities

Why Control GPU Costs?

GPUs are essential for driving AI innovation, offering unprecedented computational power and efficiency through technologies like NVIDIA's A100 and H100 Tensor Core GPUs.

These GPUs enable complex AI training, real-time inference, and high-performance computing workloads. However, this power comes at a premium, making cost control a critical factor for maximizing ROI.

To address these challenges, cost governance must be built into infrastructure management workflows.

Quali Torque ensures that GPU investments are aligned with business priorities, delivering real-time visibility, proactive controls, and dynamic optimization to eliminate waste and align expenditures with organizational goals.

Without effective cost management, organizations risk:

- **Runaway Expenses:** GPU workloads often involve unpredictable scaling, leading to budget overruns.
- **Underutilized Investments:** Poor visibility into usage can result in overprovisioned resources that go idle.
- **Misaligned Spending:** Resources may be consumed by non-critical tasks, reducing their value to the business.

2 Controlling Costs with Torque

Aligning GPU Investments with Business Priorities

The Challenge:

GPU workloads are resource-intensive and inherently dynamic, leading to unpredictable costs and inefficiencies. Many organizations lack the tools to manage these expenses effectively, resulting in waste and misaligned budgets.

Our Solution:

Torque embeds cost control into GPU management, providing real-time dashboards, automated budget enforcement, and AI-driven workload prioritization. These capabilities ensure that GPU resources are allocated based on business priorities, delivering predictable costs and measurable ROI.

How Torque Aligns GPU Costs with Business Priorities

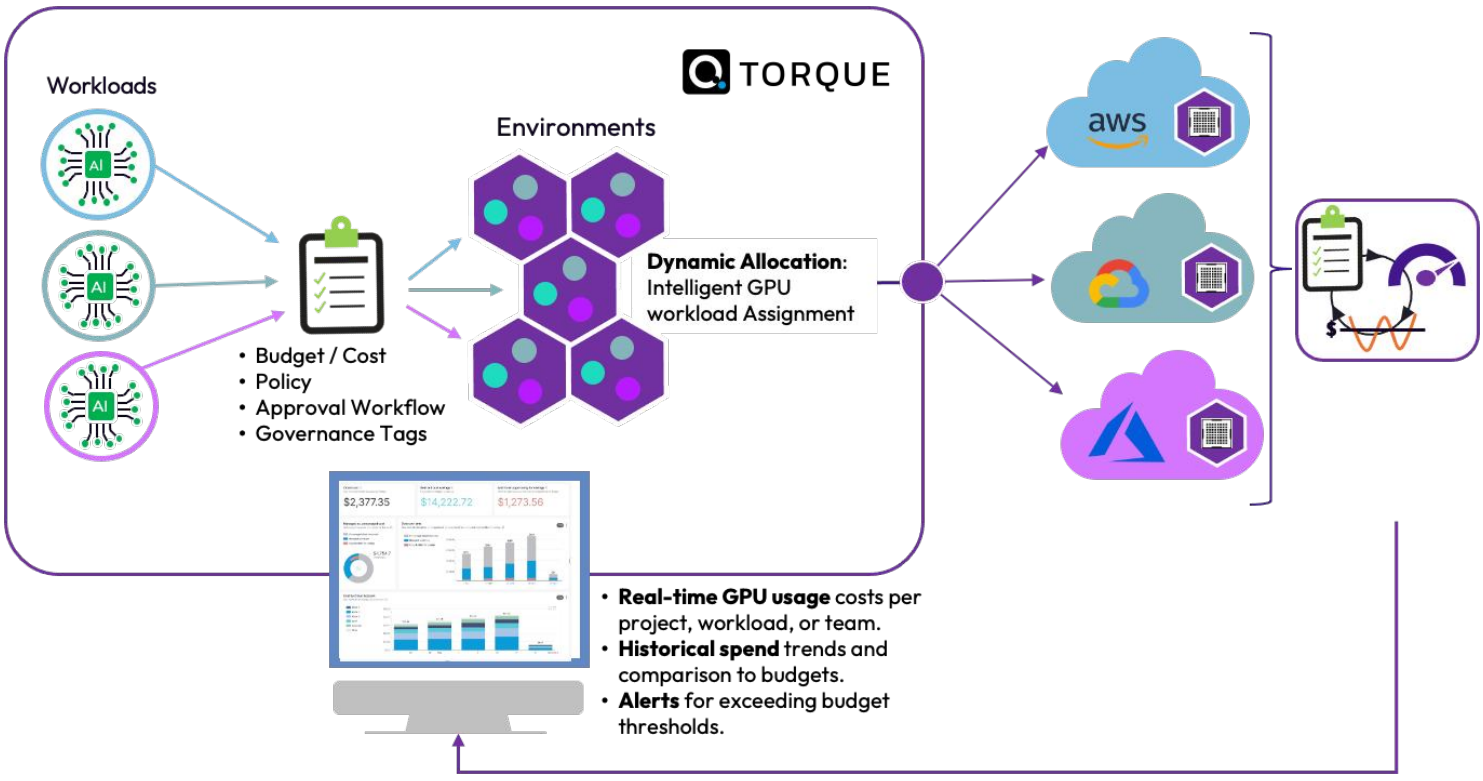
Torque embeds cost governance into every aspect of GPU workload management, combining visibility, automation, and intelligence to prevent overspending.

Here's how Torque manages costs in five key steps:

1. **Cost Visibility Dashboards:** Torque provides real-time insights into GPU usage and spending across environments, enabling organizations to track consumption patterns and identify inefficiencies.
2. **Dynamic Budget Allocation**: Using AI-driven workflows, Torque aligns GPU usage with budgets, ensuring that critical projects receive priority while less important tasks are deprioritized.
3. **Proactive Budget Enforcement:** Torque automates the enforcement of spending limits, preventing workloads from exceeding predefined budgets without authorization.
4. **Predictive Analytics:** Leveraging historical data and AI, Torque predicts future GPU spending trends, allowing organizations to adjust plans and prevent surprises.
5. **Policy-Based Cost Management:** Torque integrates governance policies that automate tagging, allocation, and approval workflows to ensure costs align with strategic objectives.

By introducing these capabilities, Torque ensures that every dollar spent on NVIDIA GPUs contributes directly to business goals.

2 Cost Governance for GPU Workloads



1 **Workloads** (AI/ML training, HPC, etc.) are initiated by users or teams and routed to available GPU's.

2 The **Cost Dashboard** tracks real-time expenses and provides historical and forecasted insights into spending trends.

3 **Policy Enforcement** ensures all workloads adhere to governance rules, automating approvals, tagging, and budget caps.

4 **Predictive Analytics** forecasts future usage and offers recommendations for optimized GPU allocation.

5 **Outcomes**, such as cost reductions and predictable spending, are achieved through this closed-loop system of visibility, control, and enforcement.

Leveraging Quali Torque for NVIDIA GPU workloads **can achieve up to a 40% reduction** in operational costs by embedding real-time cost governance, proactive budget enforcement, and AI-driven workload prioritization — ensuring every GPU dollar is maximized for business impact.

3 Unified Management with Torque Simplifying GPU Workloads Across Hybrid Environments

Why Unify GPU Workload Management?
As organizations adopt hybrid and multi-cloud strategies, managing GPU workloads across diverse environments has become increasingly complex.

NVIDIA GPUs, with their unmatched computational power, enable AI-driven innovation across cloud, edge, and on-premises deployments. However, fragmented tools, inconsistent policies, and operational silos often create inefficiencies and governance challenges.

Quali Torque solves these challenges by providing a unified control plane for managing NVIDIA GPUs and AI workloads across all environments.

By automating workflows and ensuring consistency, Torque simplifies operations and enhances scalability.

Without unified management, organizations face:

- **Operational Inefficiencies:** Siloed infrastructure complicates workload orchestration and resource scaling.
- **Governance Gaps:** Inconsistent policies increase the risk of non-compliance and security vulnerabilities.
- **Limited Scalability:** Disconnected environments hinder the ability to adapt to growing workload demands.

3 Unified Management
Simplifying GPU Workloads Across Hybrid & Multi-Cloud Environments

The Challenge:

Managing GPU workloads across hybrid and multi-cloud environments leads to operational silos, governance challenges, and limited scalability. Disconnected tools and inconsistent policies further exacerbate these issues.

Our Solution:

Torque integrates orchestration, governance, and visibility into a single control plane, simplifying GPU workload management across environments. Its dynamic orchestration and policy enforcement ensure consistent, scalable, and secure operations.

How Torque Simplifies GPU Workload Management

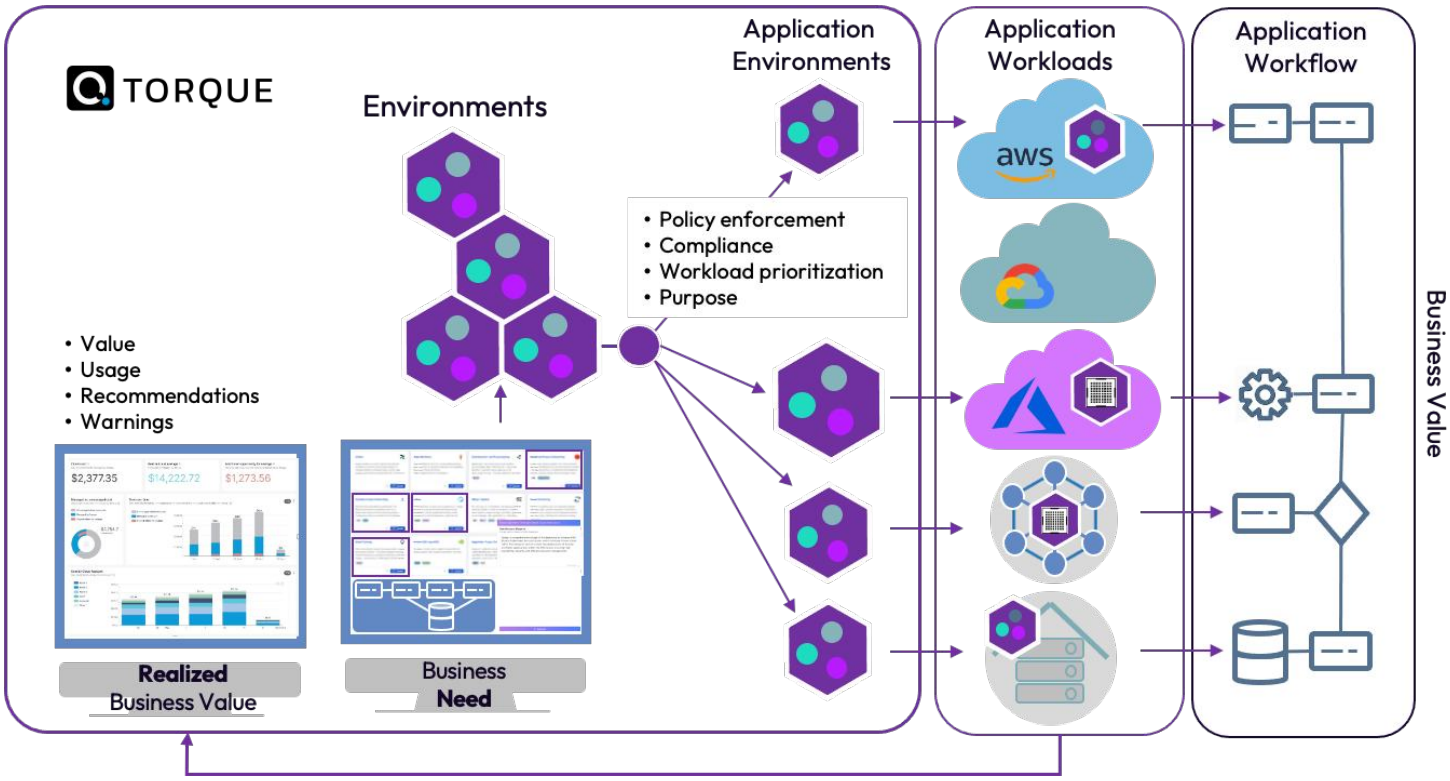
Torque unifies GPU management by integrating orchestration, governance, and automation into a single platform.

Here's how Torque achieves this in five key steps:

1. **Centralized Control Plane:** Torque provides a single management interface for hybrid, multi-cloud, and on-premises GPU environments, offering complete visibility and control.
2. **Automated Policy Enforcement**: Built-in governance ensures consistent security and compliance policies across all environments, minimizing risks and manual effort.
3. **Dynamic Orchestration:** Torque automates workload deployment and scaling, enabling seamless transitions between on-premises, cloud, and edge environments.
4. **Unified Visibility and Insights:** Real-time dashboards track GPU usage, performance, and compliance metrics, allowing organizations to monitor and optimize operations proactively.
5. **Scalable Infrastructure Management:** Torque's platform supports NVIDIA's GPU technologies, such as multi-instance GPU (MIG), to enable resource partitioning and workload elasticity across diverse deployments.

By unifying GPU management, Torque empowers organizations to streamline operations, maintain governance, and scale with confidence.

3 Unified GPU Workload Management



1 **Workloads** are initiated by teams, routed through Quali Torque, and allocated dynamically across environments (cloud, edge, on-premises).

2 Torque's **Policy Enforcement** ensures compliance, security, and workload prioritization at every stage.

3 **Dynamic Orchestration** enables seamless transitions of workloads between environments based on performance needs and resource availability.

4 The **Real-Time Dashboard** provides visibility into all GPU workloads, allowing teams to monitor utilization, performance, and compliance metrics

5 **Realized business value** through reduced operational overhead and enhanced scalability, driving greater efficiency and agility in GPU workload management.

Quali Torque **reduces operational overhead by up to 50%** by unifying hybrid, multi-cloud, and on-premises GPU environments under a single control plane, ensuring seamless scalability and governance

Unlocking Business Value You Can Trust
Why Quali Torque Is the Only Choice for NVIDIA GPU Workloads

AI-Driven Infrastructure for AI Workloads

Quali Torque, the **only Platform solution** that uses **AI to manage AI**. With AI-driven orchestration, governance, and optimization, Torque dynamically adapts to the unique needs of NVIDIA GPU workloads, ensuring unparalleled performance and efficiency.

Unified Management for AI and Traditional Workloads

Unlike other solutions, Quali Torque is purpose-built to manage **both AI workloads** (e.g., training, inference) and **traditional workloads** (e.g., database operations, DevOps pipelines). This unified approach eliminates silos, enabling seamless operations across diverse environments.

Instant AI Enablement of Existing IaC Assets

Quali Torque uniquely turns **all existing Infrastructure as Code (IaC) assets**—Terraform, Ansible, Kubernetes YAML, and more—into GPU-ready, AI workload assets. This dramatically accelerates the adoption of NVIDIA GPUs without requiring rework or new skill sets.

Comprehensive Support for all NVIDIA GPU Workloads

Quali Torque is the only solution capable of optimizing the full spectrum of NVIDIA GPU workloads.

- **High-Performance Computing (HPC):** For scientific simulations, genomic analysis, and more.
- **Graphics Rendering:** Used in gaming, film production, and virtual reality.
- **Data Analytics:** Accelerated big data processing and real-time insights.
- **Blockchain and Cryptography:** Cryptocurrency mining and secure transaction verification.
- **Video Processing:** For high-resolution encoding, decoding, and streaming.

Democratized Access to GPU Instances

Quali Torque simplifies access to GPU cloud instances with **natural language-driven, AI-assisted self-service**. It empowers **data scientists, developers, IT operations teams, and non-technical users** to request, provision, and manage GPU resources effortlessly, removing barriers to entry for GPU-powered innovation.

NVIDIA® and Quali TORQUE

Empowering Organizations with Unified GPU Management and Control

Quali and NVIDIA: Unlocking the Full Potential of Accelerated GPU Workloads for AI Innovation

Quali Torque, in collaboration with **NVIDIA**, is revolutionizing GPU workload management by delivering unparalleled efficiency, scalability, and control. By unifying hybrid, multi-cloud, and on-premises environments under a single platform, Torque ensures that NVIDIA GPUs are utilized to their fullest potential while aligning costs and performance with business priorities. With its intelligent automation, real-time insights, and AI-driven workflows, Torque helps organizations maximize ROI on their GPU investments, reduce operational complexity, and unlock the true potential of AI and HPC workloads. Whether it's driving innovation, scaling operations, or ensuring cost control, Torque equips businesses with the tools they need to succeed in an increasingly AI-driven world.

Features

- **Unified Control Plane:** Seamlessly manage hybrid, multi-cloud, and on-premises GPU environments from a single platform.
- **Dynamic Workload Allocation:** AI-driven orchestration ensures GPU resources are allocated based on performance needs and priorities.
- **Real-Time Cost Governance:** Embedded tools monitor spending, enforce budgets, and provide predictive analytics for future planning.
- **Policy Automation:** Enforce security, compliance, and operational policies across environments automatically.
- **NVIDIA Integration:** Fully supports NVIDIA GPU technologies, including MIG for resource partitioning and workload isolation.

Benefits

- **Increased GPU Utilization:** Optimize GPU usage by up to 30%, ensuring resources are never wasted.
- **Cost Savings:** Achieve up to 40% reduction in infrastructure costs through real-time governance and dynamic allocation.
- **Streamlined Operations:** Reduce operational overhead by 50% with unified management and automated workflows.
- **Scalability:** Easily adapt to growing workload demands across cloud, edge, and on-premises environments.
- **Business Alignment:** Align GPU usage with strategic objectives, ensuring investments deliver measurable value.