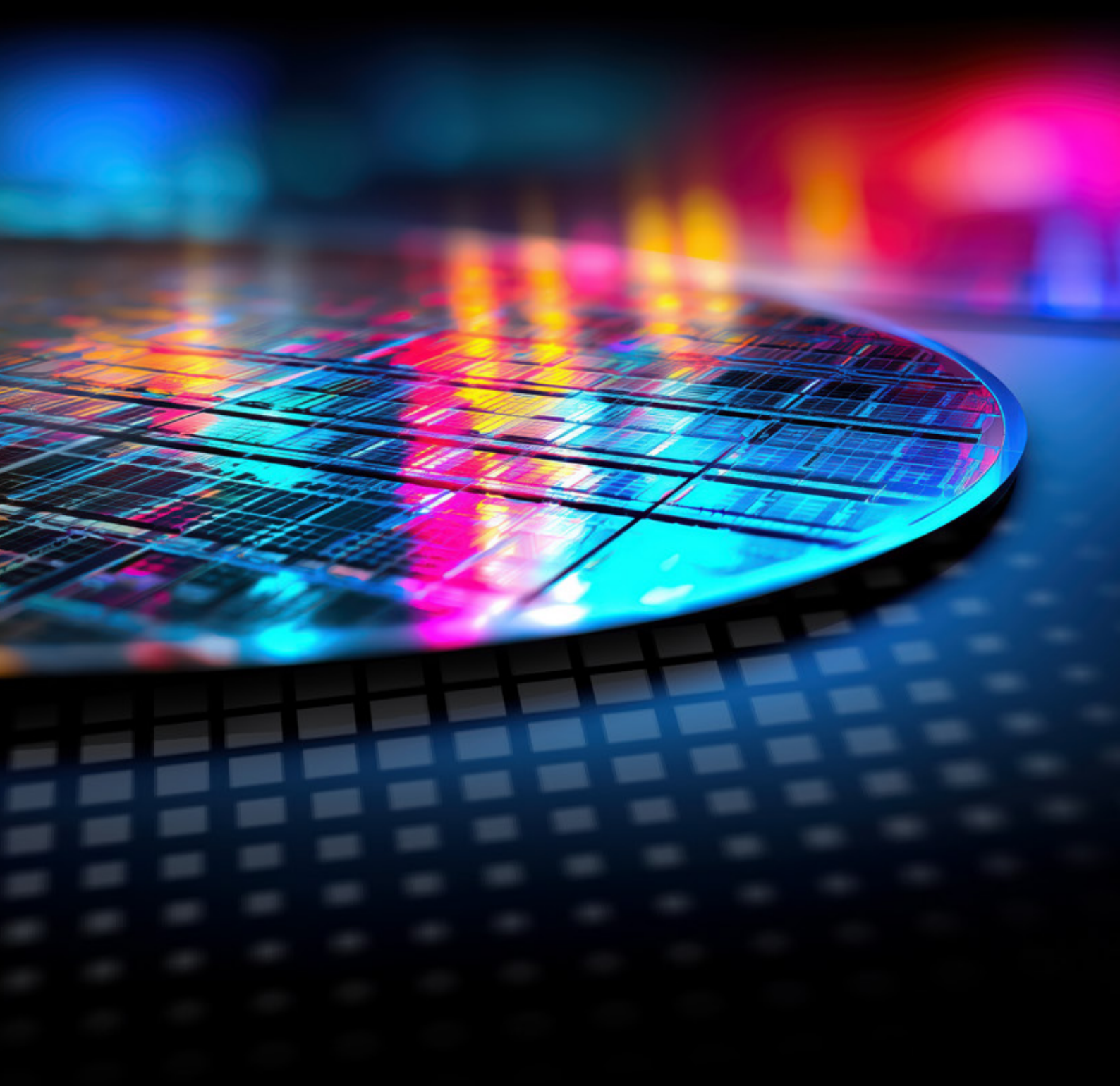
HCLTech | Supercharging Progress™

Unlocking the AI opportunity within existing high-performance computing environments

A practical modernization blueprint for enterprises and research organizations



Executive summary

The convergence of high-performance computing (HPC) and artificial intelligence (AI) represents one of the most consequential inflection points in enterprise technology today. Over the past three years, it has also become one of the most commercially compressed—often framed by oversimplified narratives and inflated expectations that obscure technical reality.

A critical truth is frequently missed: for many enterprises, government laboratories, financial institutions and research universities, the most cost-effective path to AI-ready infrastructure is not a greenfield build. It runs through the HPC environments they already operate.

At the same time, the notion that any legacy HPC environment can be seamlessly repurposed for AI is both simplistic and risky. Organizations that move forward without rigorous technical assessment routinely encounter costly surprises related to power density, network bandwidth, storage performance and operational readiness.

Equally flawed, however, is the belief that existing HPC investments are obsolete in the AI era. According to Hyperion Research, 78% of global HPC sites already run AI workloads alongside traditional simulation[1]. Convergence is no longer theoretical—it is occurring in production environments today.

Organizations that clearly understand which assets can be reused, which must be upgraded, and which should be replaced are best positioned to achieve superior economic and operational outcomes. This paper examines the market forces driving HPC–AI convergence, the technical constraints that shape it, the structured methodology required for success and how HCLTech enables organizations to navigate this transformation with confidence.

The forces driving convergence between HPC and AI

HPC–AI convergence is a structural market shift

The global high performance computing and technical computing market is projected to exceed USD 100 billion by 2028[2], driven primarily by the convergence of simulation based computing and machine learning across industries such as pharmaceutical research, climate modeling, semiconductor design, automotive engineering and financial risk analytics. This trajectory reflects more than market expansion. It signals a fundamental shift in how advanced computation is applied to complex, data intensive problems.

For many organizations, this shift does not begin with a net new investment. Enterprises, government laboratories, financial institutions and research universities with large installed HPC environments already possess substantial assets that can be repositioned to support AI workloads.

When approached with technical rigor and architectural discipline, these environments represent a practical and economically advantaged starting point for AI ready infrastructure.

The opportunity, however, is not automatic. Capturing value depends on understanding the technical realities of existing systems and applying a deliberate transformation strategy rather than assuming that growth alone will deliver results.

[1]https://www.hpcwire.com/off-the-wire/hyperion-research-announces-hpc-ai-market-grew-by-23-5-in-2024/
[2]https://www.hpcwire.com/aiwire/2025/04/08/hyperion-research-announces-hpc-ai-market-grew-by-23-5-in-2024/

HPC–AI convergence is a production reality

Traditionally, HPC workloads—such as computational fluid dynamics, molecular dynamics, finite element analysis and seismic processing—were dominated by tightly coupled MPI based parallel jobs with deterministic communication patterns, high memory bandwidth requirements and moderate GPU utilization. AI training workloads, by contrast, emphasize massively parallel matrix operations, all reduce collective communication, GPU utilization targets in the 75–90% range and bursty, metadata-intensive storage access. These represent fundamentally different infrastructure demands.

Since 2022, however, this separation has eroded in practice. AI enhanced simulation techniques—such as neural operators that accelerate physics solvers and surrogate models replacing costly Monte Carlo methods—now require infrastructure that supports both workload types simultaneously. Climate modeling organizations are applying transformer based architectures to emulate dynamical cores, while drug discovery pipelines increasingly combine molecular dynamics with graph neural networks within a single workflow.

In these scenarios, organizations cannot choose between 'HPC infrastructure' and 'AI infrastructure'. They require both, operating on a shared fabric. This integration is the technical foundation of HPC–AI convergence and organizations unable to support it face a growing and compounding competitive disadvantage as AI augmented workflows become the production norm.

Cloud economics break down at sustained AI scale

For organizations with intermittent or highly variable AI workloads, cloud based infrastructure can be a sensible choice. However, for established HPC users operating at sustained scale, cloud economics quickly become unfavorable.

Training large AI models on proprietary datasets is extremely data intensive. A 10,000 GPU training environment, for example, can require sustained read bandwidth on the order of five terabytes per second simply to keep accelerators fully utilized.

Moving petabyte scale datasets into the cloud introduces significant data transfer costs, increased latency and elevated security and compliance exposure—constraints that are unacceptable for many regulated or IP sensitive industries. As organizations move from proof of concept AI to production deployment, these factors increasingly drive workloads back on premises, where cost, performance and control are more predictable.

[3]What are the optimal GPU utilization levels for different workloads? - Massed Compute
[4]https://www.vdura.com/2025/12/11/gpu-goliaths-are-devouring-supercomputing-and-legacy-storage-cant-feed-the-beast/

How a structured approach enables successful HPC-to-AI transformation

Avoiding failure in HPC-to-AI transformation requires a disciplined approach that balances ambition with operational realism. Successful organizations follow a structured, phased methodology that reduces risk while preserving the integrity of existing HPC operations.

Phase 1: Infrastructure assessment

Transformation must begin with a comprehensive infrastructure assessment conducted by practitioners with deep experience across both HPC and AI environments. This assessment should produce quantitative, defensible outputs across four critical areas:

Facility-level power availability and thermal headroom	Network topology, fabric characteristics and bisection bandwidth	Storage performance under AI-specific I/O and metadata access patterns	Scheduler, middleware and software stack readiness

Each component is then classified as reusable, upgradeable, or requiring replacement. This clarity forms the foundation for both the business case and the executable project plan.

Phase 2: Pilot deployment and validation

Before committing full capital investment, organizations should deploy a bounded pilot—typically 30 to 100 GPU nodes—and execute representative distributed training workloads. This phase validates performance assumptions and establishes critical operational practices, including container management, GPU-aware scheduling, monitoring and production runbooks.

Phase 3: Phased scale-out with rollback protection

Production scale-out should proceed incrementally, typically in 25–30% capacity expansions, with clearly defined validation gates between phases. Existing HPC workloads must remain fully operational throughout. Protecting production simulation and modeling workloads is a governance requirement, not simply a stakeholder preference.

Phase 4: Continuous optimization

Deploying hardware does not mark the end of the journey. GPU utilization below 70% represents significant capital inefficiency. Sustained performance requires continuous scheduling optimization, communication profiling, storage tuning and deliberate management of software updates. Co-scheduling AI and simulation workloads can materially improve overall cluster utilization and economic return.

What a well-executed upgrade delivers

A well-executed HPC-to-AI upgrade delivers measurable business value across three critical dimensions:

Faster time to production

While greenfield AI environments typically require 18–36 months to become operational, a correctly scoped upgrade can deliver AI-ready capability in as little as 6–12 months—accelerating value realization significantly.

Greater capital efficiency

By reusing existing facilities, interconnects and storage foundations, organizations avoid the most capital-intensive elements of new builds, materially reducing upfront investment.

Higher infrastructure utilization

Introducing AI workloads often increases average cluster utilization from 60–70% to 80–90%, unlocking substantial economic gains from assets already in place.

The analysis above leads to a clear conclusion: HPC to AI transformation does not start from a blank slate. It begins with a real, inherited environment—defined by existing fabric generations, storage architectures, scheduler versions and operational practices built up over years of production use.

The success of the transformation depends on how well that baseline is understood and how deliberately change is introduced. Organizations that apply a structured, phased approach—one that protects ongoing HPC operations while incrementally enabling AI capability—consistently achieve better outcomes. Those who bypass rigorous assessment or rely on assumptions almost inevitably discover, midway through the project, that unresolved constraints are the most costly and disruptive to address.

Critical constraints that influence HPC-to-AI transformation

Power density

The most immediate constraint in HPC to AI transitions is power density. Modern AI accelerator racks can draw up to 160 kilowatts per rack[5]—far exceeding what most legacy HPC facilities were designed to handle. Incremental improvements to air cooling are insufficient at these levels; the physical limits of air as a heat transfer medium impose a hard ceiling that cannot be overcome.

At these power densities, direct to chip liquid cooling becomes the only viable option. Implementing such solutions requires chilled water delivery to the rack, robust leak detection mechanisms, upgraded power distribution infrastructure and in many cases, significant mechanical and structural modifications to the facility.

While these upgrades are achievable, they are neither quick nor inexpensive. As a result, any AI upgrade strategy that does not begin with a detailed, facility level assessment of thermal and electrical capacity is fundamentally incomplete.

Storage

While computer hardware typically commands the most attention, storage is where AI infrastructure initiatives most frequently encounter unexpected performance limitations. AI training workloads place sustained demands on high bandwidth data reads while simultaneously generating intensive metadata activity through frequent file access and checkpointing.

Parallel file systems designed for traditional HPC workloads—often optimized for infrequent, large checkpoint writes and backed by spinning disks—struggle under these mixed and highly concurrent access patterns. Addressing this gap commonly requires introducing solid-state storage tiers, retuning file systems for smaller block I/O and in some cases, augmenting or replacing legacy storage platforms altogether.

Organizations that treat storage as a secondary consideration often find that GPU utilization falls well short of expectations, undermining both performance and economic efficiency.

Networking assets

Networking assets require careful, workload specific evaluation during HPC to AI transformation. Modern InfiniBand fabrics operating at 200–400 Gbps remain strong assets and align well with current GPU cluster reference architectures. In contrast, older InfiniBand generations in the 40–60 Gbps range frequently emerge as primary bottlenecks for distributed training workflows, particularly those dominated by all reduce collective communication.

Determining suitability goes beyond headline link speeds. Effective assessment must examine switch-level bisection bandwidth, fabric topology and collective communication latency to determine whether the existing network can efficiently support AI workloads at scale.

Only by evaluating these characteristics in combination can organizations accurately determine whether their current fabric represents a reusable asset.

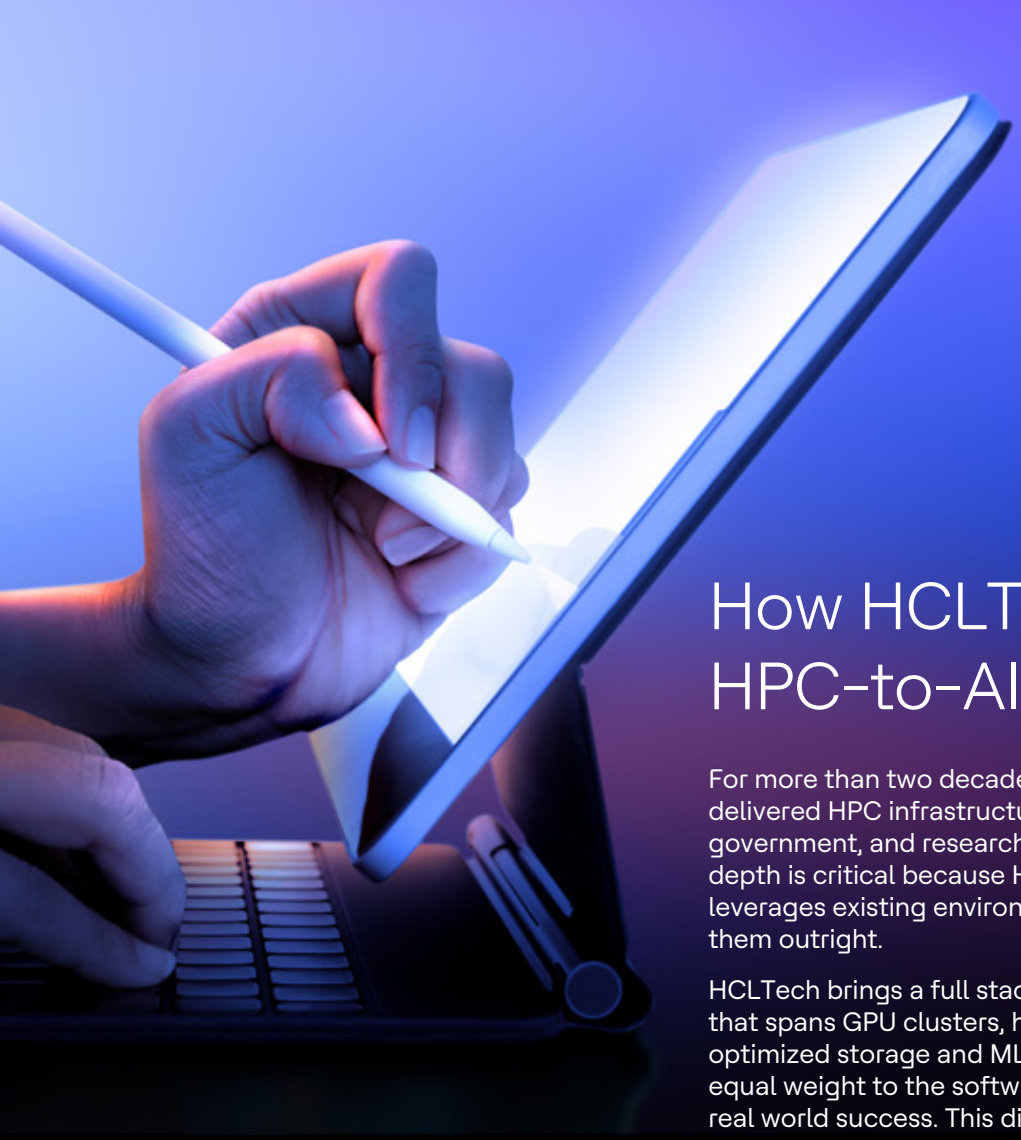
[5]Qualcomm Unveils AI200 and AI250—Redefining Rack-Scale Data Center Inference Performance for the AI Era | Qualcomm

Making the HPC-to-AI transition work: A disciplined path forward

Upgrading existing HPC infrastructure for AI is not a default choice—it is a decision that must be earned through rigorous, technically grounded assessment.

Organizations that invest in honest, technically grounded evaluation consistently find that their existing HPC assets provide a materially stronger starting point than greenfield alternatives. Those who skip this step almost always discover, mid-project, that the constraints they failed to investigate are the ones that cost the most to resolve.

Getting the assessment right is what ultimately determines whether HPC-to-AI transformation delivers competitive advantage or becomes an expensive lesson.



How HCLTech enables HPC-to-AI transformation

For more than two decades, HCLTech has actively delivered HPC infrastructure services to enterprise, government, and research clients. This operational depth is critical because HPC to AI transformation leverages existing environments instead of replacing them outright.

HCLTech brings a full stack AI infrastructure practice that spans GPU clusters, high performance fabrics, AI optimized storage and MLOps platforms—while giving equal weight to the software foundations that decide real world success. This disciplined focus on containers, GPU aware schedulers, CUDA alignment and monitoring closes the execution gaps that undermine many infrastructure programs.

Every engagement start with assessment and Architecture Review that delivers a clear, quantitative gap analysis and a prioritized upgrade roadmap—without disrupting ongoing HPC operations or requiring any downtime.

References

1. https://www.hpcwire.com/off-the-wire/hyperion-research-announces-hpc-ai-market-grew-by-23-5-in-2024/
2. https://www.hpcwire.com/aiwire/2025/04/08/hyperion-research-announces-hpc-ai-market-grew-by-23-5-in-2024/
3. https://massedcompute.com/faq-answers/?question=What%20are%20the%20optimal%20GPU%20utilization%20levels%20for%20different%20workloads?#:~:text=Inference%20Workloads,50%25%20utilization%20to%20minimize%20latency
4. https://www.vdura.com/2025/12/11/gpu-goliaths-are-devouring-supercomputing-and-legacy-storage-cant-feed-the-beast/
5. https://www.qualcomm.com/news/releases/2025/10/qualcomm-unveils-ai200-and-ai250-redefining-rack-scale-data-cent#:~:text=This%20enables%20disaggregated%20AI%20inferencing,our%20optimized%20AI%20inference%20solutions.
6. https://www.nvidia.com/en-us/data-center/gb200-nvl72/
7. https://buy.hpe.com/us/en/compute/rack-scale-system/nvidia-nvl-system/nvidia-gb300-nvl72-by-hpe/p/1014890105
8. https://www.datacenterdynamics.com/en/news/microsoft-deploys-cluster-of-4600-nvidia-gb300-nvl72-systems-for-openai/
9. https://awesomeagents.ai/hardware/nvidia-gb300-nvl72/
10. https://www.hpcwire.com/2025/06/07/hyperion-hpc-ai-update-2024-boomed-loudly-will-2025-sputter/

Manpreet Singh

Product Manager

HCLTech

About the Author

Manpreet Singh is an experienced professional with a strong background in Hybrid Cloud and Networking, with a specialized focus on High-Performance Computing (HPC). With a proven track record in presales and GTM roles, he brings strategic insights that drive innovation and customer-centric solutions across these domains