doitnow
HPC Services

White paper

HYBRID HPC, REAL ROI.

Maximizing **on-premises** utilization with governed cloud bursting.

WWW.DOITNOWGROUP.COM

Hybrid HPC, Real ROI.

Maximizing on-premises utilization with governed cloud bursting and efficient, consistent, intelligent environments.

EXECUTIVE OVERVIEW

High-performance computing (HPC) and AI infrastructures have reached a critical inflection point. Organizations are under pressure to justify large capital investments in on-premises clusters while maintaining the agility to respond to surges in computational demand. Both a purely on-prem and a cloud-only approach rarely deliver optimal ROI.

This white paper outlines how a hybrid strategy, anchored in **performance efficiency, software environment consistency, and intelligent workload governance**, enables organizations to maximize utilization, contain costs, and accelerate results. State-of-the-art open-source technologies, such as Spack, EasyBuild and EESSI, are instrumental to achieve this objective. Drawing from real client data and operational experience, this paper demonstrates how to transform technical infrastructure into a quantifiable business advantage.

01.

Understanding the Economics of Compute Models

ON-PREMISES CLUSTERS

An on-premises cluster provides full control over infrastructure, data, and cost structure. For stable, predictable workloads, it remains the most cost-efficient model. However, capital expenses, limited scalability, and inevitable underutilization during low-demand periods often erode the perceived ROI.

The three-to-five-year hardware lifecycle further compounds the issue; budget allocations are locked upfront, even though workload intensity fluctuates dramatically over time.

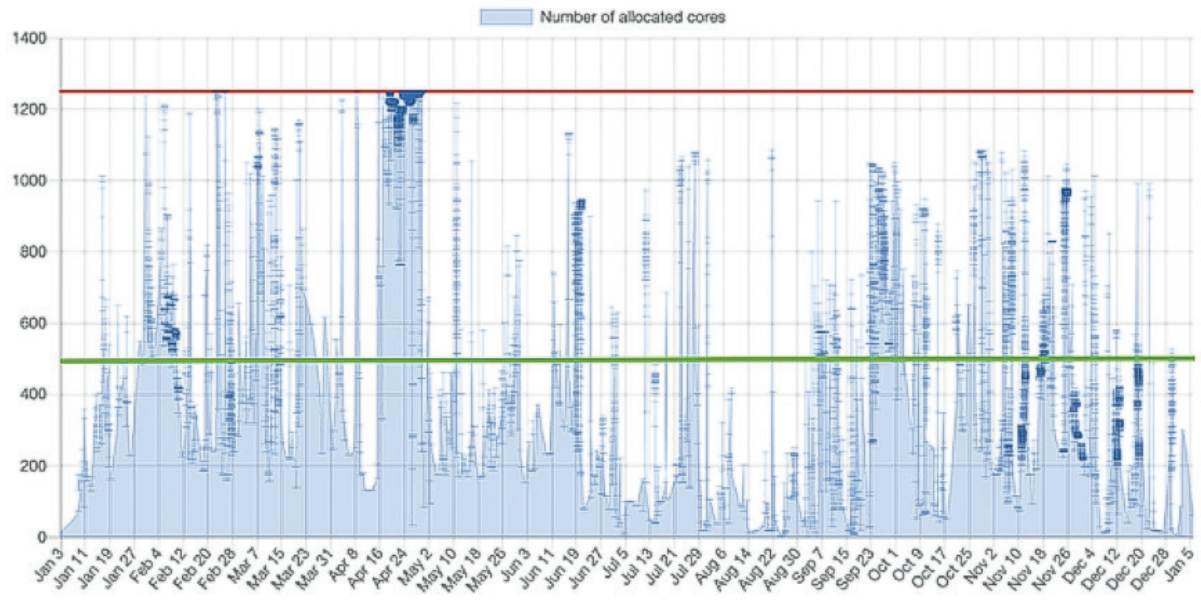


Figure 1. Year-long allocation of cores on an on-premises cluster. Static capacity (red) contrasts with fluctuating demand, with long periods of underutilization below the efficiency threshold (green).

COMMERCIAL CLOUD

Cloud computing introduces flexibility and scalability. Organizations can scale to thousands of cores in minutes and pay only for what they consume. Yet, the benefit of elasticity comes at a premium. The unit price per CPU/GPU hour is significantly higher than in-house resources, and without governance, operational expenses can spiral quickly.

A balanced model is needed: one that leverages the predictability of on-prem assets while tapping into the cloud for elasticity and resilience.

02.
Cloud Bursting: The Hybrid Model that Delivers

Cloud bursting bridges the gap between static capacity and dynamic demand. It enables HPC centers to extend compute capacity to the cloud only when needed, maintaining cost control without compromising agility.

Key advantages:

- Immediate capacity growth to meet peak demand.
- Reduced Total Cost of Ownership (TCO) through optimized hybrid utilization.
- Natural Disaster Recovery (DR) capability in case of local outages.
- Simplified migration during hardware refresh or software transition.
- Decreased capital expenditure (CapEx) with predictable operational cost scaling.

Cloud bursting is not a technology, but rather a strategy for **business continuity, risk mitigation, and performance assurance**.

03.

Foundational Challenges and How to Address Them

3.1 THE EFFICIENCY IMPERATIVE

The financial success of hybrid HPC depends on efficiency. Every wasted CPU cycle or misconfigured job translates directly into cost. Before scaling to the cloud, organizations must ensure their local clusters are fully optimized.

In a 30-day monitoring sample, one customer experienced a **50% job failure rate**, predominantly from user errors and configuration issues. Had these workloads been executed in the cloud, the equivalent cost would have exceeded **USD 400 000**. Identifying inefficiencies early converts lost compute time into measurable savings.

The **Do IT Now Monitoring Stack** exposes a comprehensive real-time dashboard that streamlines the detection of workload inefficiencies, user mistakes, and misbehaving jobs and the **prompt identification of measures to improve efficiency.**

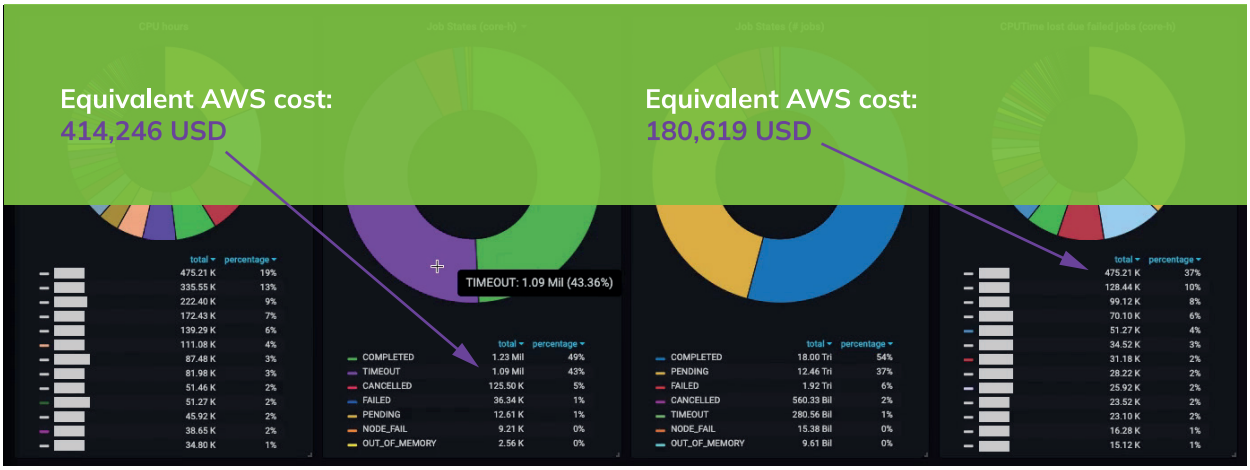


Figure 2. Real-time efficiency monitoring and equivalent AWS cost of failed jobs

3.2 CONSISTENCY ACROSS ENVIRONMENTS

To make workloads portable and reproducible, on-premises and cloud software environments must mirror each other, with identical compilers, libraries, dependencies, and runtime behavior. Without this parity, jobs risk producing inconsistent results or failing after migration.

The **Spack, EasyBuild** and **EESSI** (European Environment for Scientific Software Installations) frameworks standardize and automate this alignment, ensuring that applications run seamlessly across HPC & AI clusters, Kubernetes environments, and public clouds. This reduces context switching, accelerates onboarding, and guarantees identically optimal performance characteristics wherever the job executes.

3.3 DEFINING BUSINESS RULES

Effective hybrid governance depends on clear, enforceable rules:

- **Runtime thresholds** for cloud eligibility.
- **Data sensitivity flags** for compliance.
- **User- or group-based ACLs** for offload rights.
- **Remediation logic** to halt inefficient jobs automatically.

The **Do IT Now Slurm Plugin** integrates these business rules into the workload scheduler itself, enabling **cost-aware decisions** and automated job routing without human intervention.

04.
Enabling Technologies for Hybrid ROI

4.1 REAL-TIME EFFICIENCY MONITORING

Comprehensive observability, covering CPU, GPU, memory, and I/O efficiency, provides the foundation for informed decisions. Real-time dashboards expose underused nodes, highlight user errors, and reveal structural inefficiencies in workflows.

- Identify users or groups needing training.
- Detect and remediate bottlenecks in I/O, memory, or parallelization.
- Quantify the cost of inefficient jobs in equivalent cloud expenditure.
- Provide actionable insight before expanding to hybrid models.

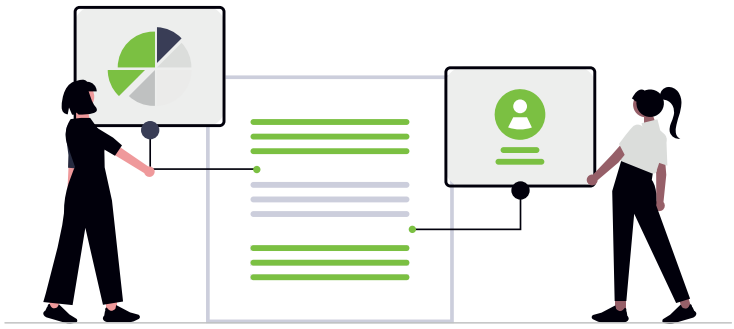
Continuous efficiency monitoring ensures the cluster operates at **peak ROI**, not just peak performance.

4.2 CONSISTENT SOFTWARE ENVIRONMENTS

The **EESSI** distribution layer and **Spack/EasyBuild** automations create a consistent and performance-tuned software stack that spans the entire computational continuum, from laptops to supercomputers to cloud instances.

Strategic benefits:

- **Simplified user workflows:** same commands, same modules, everywhere.
- **Reproducibility:** identical results across architectures.
- **Rapid debugging:** replicate issues locally before scaling.
- **Lower operational overhead:** centralize support and maintenance.
- **Optimized resource use:** shift workloads intelligently based on efficiency, not habit.



4.3 INTELLIGENT SCHEDULING AND GOVERNANCE

Embedding logic into the workload scheduler elevates job orchestration from reactive queue management to **proactive business control**.

For example:

- Short, low-impact jobs automatically backfill on-prem resources.
- Long, expensive jobs are retained locally to avoid high cloud costs.
- Medium-sized jobs are dynamically allocated between on-prem and cloud based on queue congestion and budget.
- Special-hardware workloads (GPU, FPGA) burst selectively to cloud instances.

Governance extends to automatic **react-prevent-improve** cycles (Table 1).

Action	Purpose	Example
React	Detect and notify inefficiency.	Flag jobs with poor scaling efficiency.
Prevent	Enforce thresholds and hold wasteful jobs.	Block repetitive misconfigured workloads.
Improve	Apply dynamic limits or rerouting.	Rebalance utilization by user or queue.

*Table 1. React–prevent–improve governance cycle. This approach embeds **ROI control directly into the operational layer.***

05.
Performance Optimization: Unlocking Latent Value

Performance isn't just a technical metric; instead, it's a cost function. Poorly optimized binaries, legacy builds, or architecture-agnostic software can **reduce performance by factors of four to eight**, directly inflating compute costs.

In particular, **microarchitecture awareness**, namely compiling software for the exact CPU or GPU family, is essential to maximize throughput and minimize energy consumption. Every instruction set upgrade (e.g., AVX-512, SVE) increases FLOPs per cycle. Ignoring these features leaves new hardware underutilized and delays ROI on infrastructure investments.

Also featured in the **EESSI** distribution, software compiled with microarchitecture awareness unlocks performance optimization that converts sunk cost into active productivity.

Microarchitecture	Year	Instruction Set	DP FLOPs/Cycle	Speedup vs 2008
Next Gen	?	AVX-1024	64	16
Diamond Rapids	2026	AVX-10.2	32	8
Skylake	2015	AVX-512	32	8
Haswell	2013	AVX2	16	4
Sandybridge	2011	AVX	8	2
Nehalem	2008	SSE	4	-

Table 2. Evolution of x86 microarchitectures and double-precision FLOPs per cycle.

06.
Quantifying the Real ROI

By unifying on-prem optimization, cloud bursting, and intelligent governance, organizations can:

- Improve infrastructure utilization by **30–50%.**
- Cut job failure-related waste by **up to 40%.**
- Decrease cloud over-spend by enforcing runtime and efficiency policies.
- Accelerate research and development outcomes with reduced queue delays.
- Enhance sustainability through lower energy consumption per result.

ROI in HPC is not theoretical, instead it is measurable in optimal infrastructure utilization, cost avoidance, and faster time-to-results.

07.
A Practical Roadmap

Each stage compounds ROI through iterative improvement.

1. **Assess:** Run a 30-day monitoring campaign to identify inefficiencies and bottlenecks.
2. **Align:** Standardize environments using Spack/EasyBuild/EESSI for consistency and reproducibility.
3. **Automate:** Deploy Slurm plugins for policy-driven scheduling.
4. **Adopt:** Implement a controlled cloud bursting strategy with clear cost governance.
5. **Accelerate:** Continuously tune performance to leverage hardware and software advances.

08.
Conclusion

HPC and AI investment yields real value only when efficiency, scalability, and governance converge. By blending **on-premises stability** with **cloud agility**, and enforcing intelligent rules to keep workloads efficient and portable, organizations achieve a self-optimizing infrastructure where every CPU hour contributes measurable business return.

Do IT Now helps enterprises and research centers move from infrastructure management to strategic value creation, translating performance into profitability, and compute time into competitive advantage.

DRIVING HPC INNOVATION FORWARD

www.doitnowgroup.com

Discuss your HPC needs today

Contact us at info@doitnowgroup.com

doitnow

WE FOCUS ON YOUR NEEDS SO
YOU CAN FOCUS ON YOUR BUSINESS