RIGHT-SIZING HYBRID MULTI-CLOUD INFRASTRUCTURE

Unlocking Agility and Cost Efficiency with Parallel Works

Produced by TCI Media Custom Publishing in conjunction with:

HPCwire

Parallel Works

Introduction: The ROI of Right-Sizing HPC Systems

In an era of rapidly evolving HPC and AI workloads, traditional on-premises systems often fall short in delivering the agility, scalability, and cost-efficiency required to stay competitive. By adopting a flexible, "right-sized" infrastructure approach — integrating on-premises resources, single or multi-cloud deployments, and hybrid multi-cloud frameworks — organizations can more effectively navigate the complexities of modern computing demands. Right-sizing an organization's infrastructure to include hybrid cloud is key to balancing performance with cost and offers flexibility that traditional environments cannot provide. The goal of right-sizing infrastructure is to optimize both capacity and capability to limit idle resources. These shifts are happening because businesses are searching for better ROI from their HPC systems and looking for tools to operationalize this. This paper describes some of the issues organizations face in effectively using infrastructure capacity and how using the Parallel Works ACTIVATE platform provides a "right-sizing" solution needed to determine the most cost-effective use of a hybrid multi-cloud infrastructure and achieve a clear ROI.

The Hybrid Multi-Cloud Environment: Drivers and Dynamics

Organizations that run HPC and AI workloads using on-premises infrastructure face a number of challenges as described by the AIMultiple Research study.

- **Capital expense:** Setting up an on-premises HPC infrastructure requires significant upfront investment in specialized hardware, facilities, and cooling systems.
- **Maintenance and upgrades:** On-premises HPC clusters necessitate ongoing maintenance and periodic upgrades, adding to the total cost of ownership.
- **Scalability:** Scaling on-premises HPC resources to meet fluctuating demand can be challenging and expensive.
- **Expertise:** Maintaining an on-premises HPC setup requires access to skilled personnel with expertise in high-performance computing, which can be a scarce resource.

Reasons for Implementing a Hybrid Multi-Cloud Environment

The challenges described above are alleviated by implementing a hybrid multi-cloud environment.

Organizations need to consider the rapidly evolving technology of cloud providers who maintain state-of-the-art infrastructure including CPUS, GPUS, FPGA and storage solutions capable of efficiently storing the massive amounts of data created by HPC, AI and simulation workloads. Cloud providers are also looking at future technology such as quantum computing and will have staff trained on its use.

In addition to issues with hardware capabilities, using outdated on-premises infrastructure can have a negative effect on user flexibility. On-premises systems are often monolithic and siloed, which causes challenges in collaboration and availability.

Hybrid computing isn't just a technological shift; it's a necessity for organizations aiming to remain competitive, agile and cost-effective.

Economics of Hybrid Multi-Cloud Infrastructure: Capacity and Capability

The main areas to consider when using hybrid HPC environments to right-size infrastructure include the driving factors of workloads, organization skills sets, and economics. The two main economic benefits focus on capacity and capability.

- Capacity optimization using hybrid systems allows organizations to recover resources that would otherwise be idle in an on-premises environment.
- Capability optimization requires matching workloads to the ideal heterogenous resources, whether that is on-premises, in the cloud, or utilizing specialized resources such as GPUs or quantum processors. This flexibility leads to significant cost reductions by ensuring that workloads don't consume more resources than they need.

Organizations need a tool to help them determine costs being spent across hybrid, cloud and on-premises infrastructure to optimize their HPC systems. The Parallel Works ACTIVATE platform provides this ability.

Making the Economics Work: Real-World Use Cases

Organizations need tooling to help them determine costs being spent across hybrid, cloud and on-premises infrastructure to optimize their HPC systems. The Parallel Works ACTIVATE platform provides this ability. See how ACTIVATE provides cost information to aid in right-sizing infrastructure capacity and capability costs in the examples below. For more information about Parallel Works ACTIVATE, see the How Parallel Works ACTIVATE Solves Hybrid Multi-Cloud Complexity section.

Capacity Example for Midsize Life Science Organization: Recover Resources That Would be Idle in On-Premises Environment with Hybrid Infrastructure

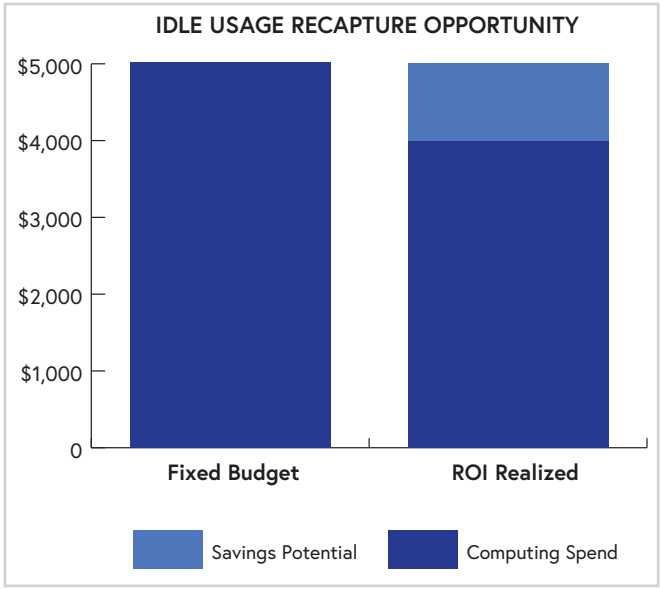
At one midsize life science organization, project managers in each department forecast their annual need for fixed on-premises computing resources. To avoid potential shortfalls, these teams often reserve more capacity than required, ensuring they have enough buffer to meet project demands. However, this conservative approach results in surplus computing resources sitting idle across the organization's on-premises data centers. By moving from an annual fixed compute budget to an allocated operational spend and integrating cloud-based capacity when on-premises infrastructure reaches its limits, organizations can shift from a fixed annual compute budget to a more dynamic, right-sized spending model. This hybrid approach—using both on-premises and cloud resources—allows for the efficient recapture of idle on-premises capacity, enhancing system utilization and driving a better return on investment (ROI). Adopting this strategy empowers teams to right-size their computing resources as needed, optimizing allocations without compromising on flexibility or performance. The Parallel Works ACTIVATE platform provides this ability through near-real-time cost control tools.

In the capacity example, the life sciences organization has an annual budget of $5 million to spend on computing. Project managers at the midsize life science organization typically allocate additional compute budgeting for running workloads in on-premises data center infrastructure. In reality, they are only using about 70% of their budget which results in 30% of the infrastructure sitting idle. The result of this annual surplus usage behavior causes high idle capacity across the organization. In this example, recovering the lost compute capacity has a potential savings on lost compute capacity of $1.5 million.

Spend on Computing	$5,000
Current Utilization	70%
Extra Capacity	30%
Savings Potential / Lost Capacity	$1,500,000
Target Utilization	90%
Saved Capacity	20%
ROI Realized	**$1,000,000**

The ACTIVATE platform contains cost control features that allow managers to capture this idle capacity by allocating computing spend/units

to teams. Project teams are responsible for "right-sizing" their computing infrastructure to their needs via the ACTIVATE uniform scheduler interface. In this case, managers set their target utilization at 90% on ACTIVATE which includes right-sizing their jobs by running workloads on both on-premises and cloud infrastructure. This change results in a 20% saved capacity with a $1 million dollar ROI savings. ACTIVATE also provides cost estimates and how much is being spent on workloads run on cloud infrastructure. Right-sized cloud usage is charged back to project teams on a ~3 min resolution, enabling tight capacity utilization. Operationally, computing allocations can be clawed back if teams don't sustain expected burn rates or can request additional allocations.



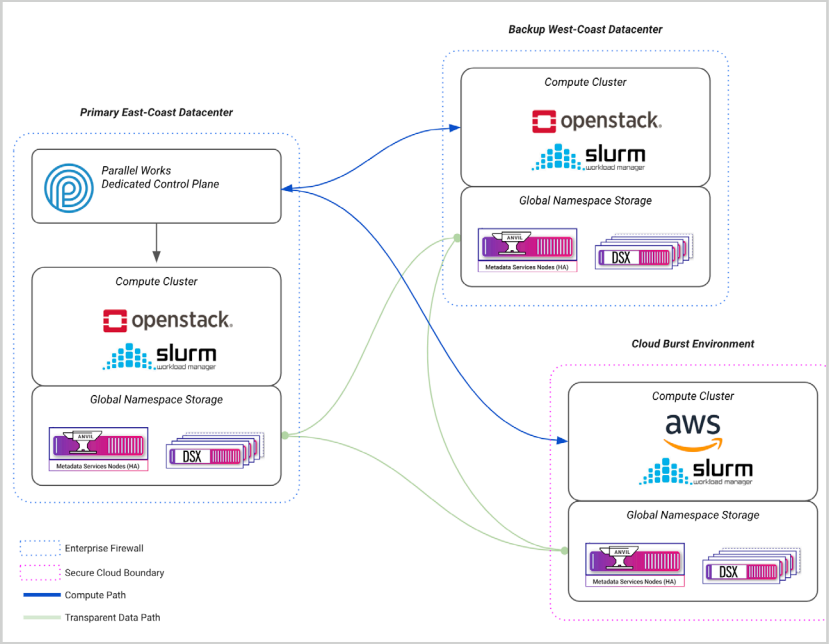
Capability Example for a Large Manufacturing Organization Fitting Workloads to the Right Resource

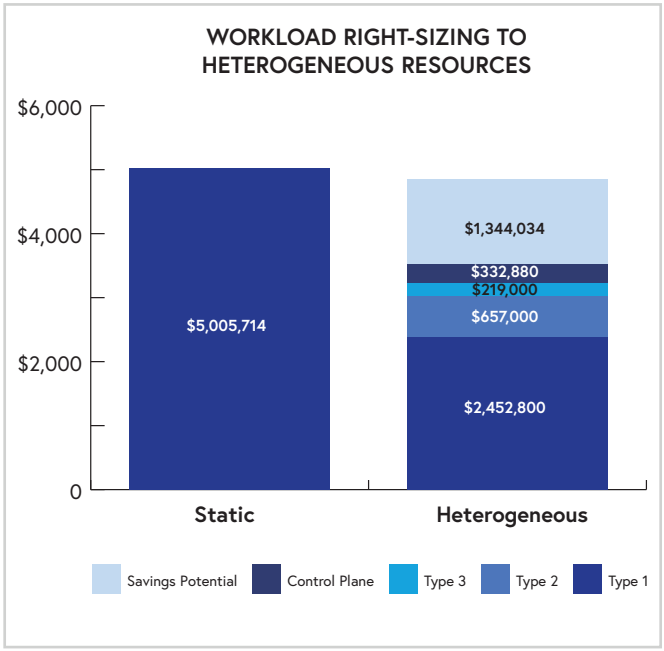
A large manufacturing organization began using the Parallel Works ACTIVATE platform to help manage costs and make it easy for staff to run workloads on the right resource by right-sizing workloads using hybrid multi-cloud heterogeneous compute resources. This workload right-sizing shifts the responsibility of resource optimization to the end users, making it easy for them to run workloads in various on-premises data centers or in the cloud.

In the capabilities example, the manufacturing organization has a total annual compute budget of $5 million using 10,000 compute cores. If workloads are run only in on-premises data centers (type 1) including the primary east-coast and backup west-coast datacenters, workloads use all 10,000 compute cores with 500,571 annual node hours for an annual compute cost of $5,005,714.

The manufacturing company began using hybrid multi-cloud infrastructure running the Parallel Works ACTIVATE platform to burst into the cloud environment with access to Amazon Web Services (AWS), Google Compute Platform (GCP), Microsoft Azure Cloud Computing Services (Azure), or Oracle Cloud instances using SLURM as the workload manager. In this example, the company reduced utilization of the on-premises infrastructure to 50% utilizing 5,000 cores by using hybrid multi-cloud instances to run their workloads. This example assumes that 50% of workloads run in on-premises data centers including the Parallel Works ACTIVATE Control Plane license. In addition, 30% run on one cloud instance (type 2), and 20% run on a different cloud instance (type 3 cloud). The total annual compute cost of running hybrid multi-cloud workloads

on-premises and on two different clouds results in a total annual cost of $3,328,800 along with a $332,880 cost for the ACTIVATE Control Plane license. The total annual compute cost of running hybrid multi-cloud workloads on-premises and on two different clouds results in a total annual cost of $3,328,800 along with a $332,880 cost for the ACTIVATE Control Plane license. This results in a total savings of 27% compared to running only in the on-premises data centers.





Static	Type 1 Hardware
	100%
On-Prem HPC Cores	10,000
Avg Cost / Node-HR	$10.00
Physical Cores / Node	175
Avg Core – HR Rate	$0.057
CSP Nodes	57.14285714
Annual Node-Hrs	500,571
Annual Compute Cost	$5,005,714

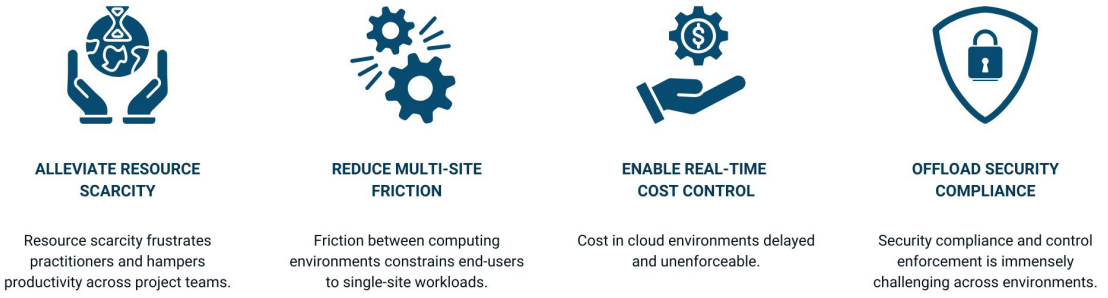
Heterogeneous	Type 1 Hardware	Type 2	Type 3	
	50%	30%	20%	100%
	5,000	3,000	2,000	10,000
	$10.00	$10	$10	
Physical Cores / Node	180	120	80	
Avg Core – HR Rate	$0.056	$0.025	$0.013	
CSP Nodes	28	25	25	
Annual Node-Hrs	245,280	219,000	219,000	
Annual Compute Cost	$2,452,800	$657,000	$219,000	$3,328,800
Control Plane License				$332,880
			Total Costs with Control Plane	$3,661,680
			Savings Compared to Static System	**27%**

Challenges to Implementing Hybrid Multi-Cloud Computing Environments

Organizations face challenges when implementing hybrid multi-cloud systems including:

- **Integration Complexity:** Integrating multiple cloud platforms and on-premises systems can be a technical and operational hurdle due to technical differences between on-premises infrastructure and various cloud providers.
- **Skill Gaps:** Implementing workloads to run on cloud instances requires a broad technical skill set, which is often lacking in organizations. In addition, there are technical differences across cloud vendors which can cause problems in managing diverse cloud environments.
- **Operational Complexities:** In a hybrid multi-cloud environment, managing multiple schedulers, monitoring systems, and budget controls can become overwhelming without proper tools.

Other challenges include resource scarcity, friction, security compliance, and cost as shown below:



Introducing Parallel Works

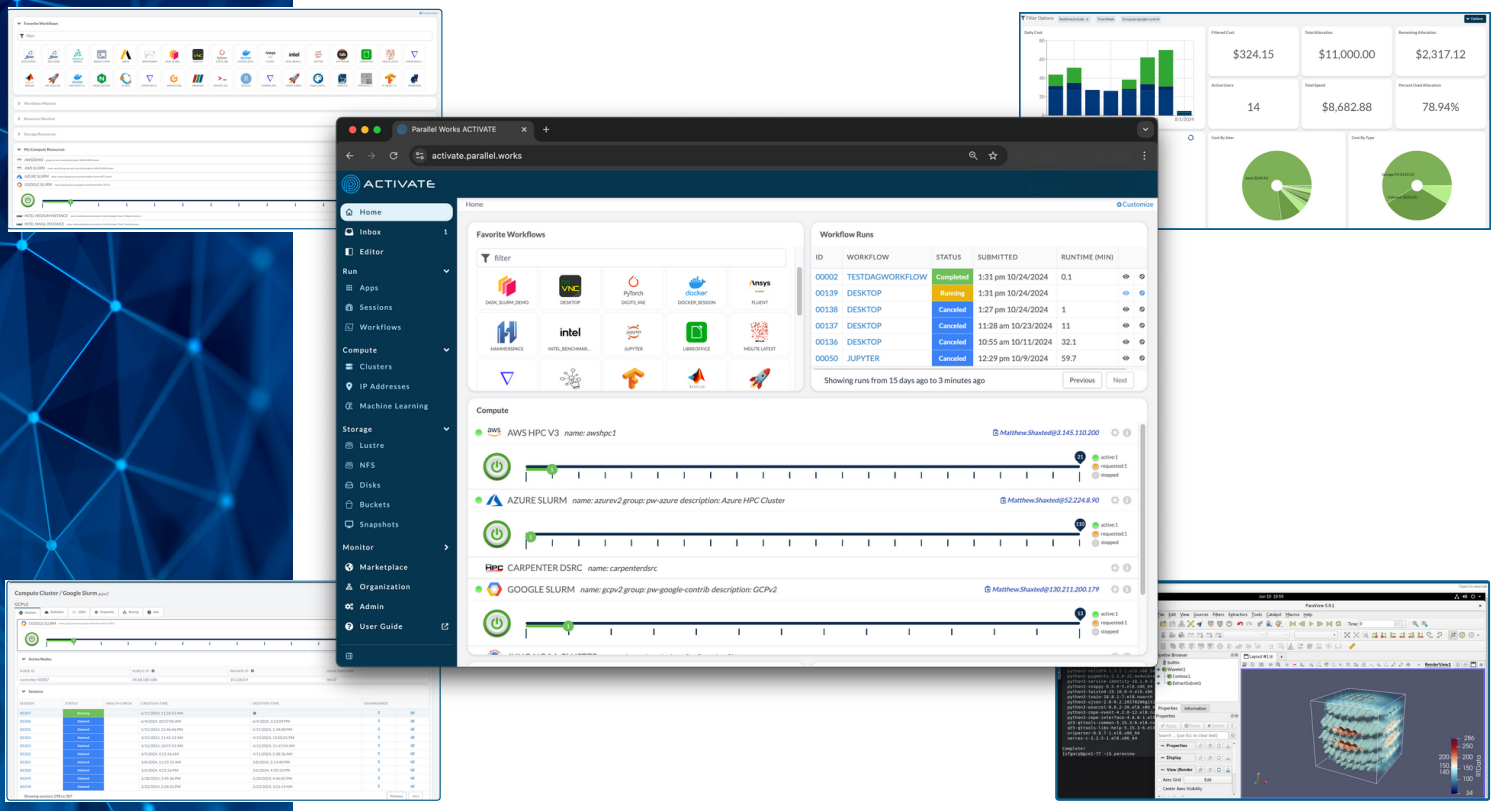
Parallel Works was founded in 2015 out of work done at Argonne National Laboratory. The Parallel Works goal is democratizing HPC & AI to reinvent the way R&D practitioners interact with diverse computing resources.

TCI MEDIA
CUSTOM PUBLISHING

How Parallel Works ACTIVATE Solves Hybrid Multi-Cloud Complexity

The Parallel Works ACTIVATE platform is a unified control plane for HPC & AI resources that simplifies optimizing hybrid multi-cloud infrastructures. The ACTIVATE platform provides a unified interface to both on-premises and cloud resources, removing the need for manual integration. In addition, ACTIVATE helps recover idle resources, enabling organizations to right-size their infrastructure efficiently. Heterogeneous resource management is provided by allowing IT teams to easily provision and allocate resources to different workloads without worrying about underlying hardware complexity.

The Parallel Works ACTIVATE platform provides ready-to-use access to workflows. The ACTIVATE cluster module includes storage connectors. It is used to create and manage uniform elastic HPC & AI clusters across multiple cloud providers. Access module allows users to access clusters via notebooks, browsers, and virtual desktops to interactively run diverse applications. ACTIVATE provides advanced metering and cost management as well as user and group management and tracking. The Control module allows users to seamlessly administer and monitor all elements of the HPC & AI control plane elements using a unified user interface and API to efficiently interact with and share diverse computing resources.



Benefits of Using the ACTIVATE Platform

Using the ACTIVATE platform provides users with ease of use, cost management, and operational efficiency along with these benefits.

- **Seamless Orchestration:** The platform provides a unified interface to both on-premises and cloud resources, removing the need for manual integration.
- **Capacity Optimization:** Using the ACTIVATE cost control features helps organizations determine whether workloads will run in an on-premises data center or burst into the cloud to help recover idle resources, enabling organizations to right-size their infrastructure efficiently.
- **Capability Management:** ACTIVATE supports heterogeneous resource management, allowing IT teams to easily provision and allocate resources to different workloads without worrying about underlying hardware complexity.

User Testimonials

Users including scientists, researchers, developers, or those creating AI models need to be able to access computing resources without needing to understand hardware, batch schedulers, or command line interfaces. ACTIVATE provides organizations with the ability to orchestrate computing tasks across on-premises resources, legacy infrastructure, batch schedulers, virtualized environments, and move workloads into the cloud effectively.

Orion Space Solutions:

> "Parallel Works ACTIVATE lets us focus on research, not the digital plumbing that supports it," said Jeff Steward, Principal Scientist at Orion Space Solutions, an Arcfield company. "We are using it to run a digital twin model of the Earth's thermosphere and ionosphere, which lets us research and predict space weather more effectively and has significant ramifications for communications networks and satellites."

Albert Einstein College of Medicine:

> The ACTIVATE platform provides an interface which incorporates simplicity making it easy for users to run HPC workloads and compute without needing to understand all the complexity or be an expert. "Parallel Works ACTIVATE makes it easier for researchers to access their computing resources," said Shilesh Shenoy, Senior Staff Scientist and Assistant Dean for Einstein Information Technology at Albert Einstein College of Medicine. "ACTIVATE dramatically improves 'time to science' for bio-medically focused researchers in their efforts ranging from genomics and pathology to drug development."

Moving Forward with Hybrid Computing

Organizations have traditionally run large HPC and AI workloads on HPC systems in on-premises data centers. However, organizations need the ability to right-size their infrastructure solution to run workloads in diverse locations including on-premises, on a single cloud, using multiple cloud providers, or a hybrid multi-cloud approach using all of these resources. Running HPC and AI workloads on both the cloud and on-premises infrastructure can coexist to provide maximum right-sizing benefits including cost efficiency, and workload flexibility. Parallel Works ACTIVATE enables organizations to implement and scale hybrid infrastructures easily, making cloud resources feel like a natural extension of their existing HPC systems.

"Hybrid multi-cloud computing is no longer just an option—it's a strategic imperative for organizations aiming to remain competitive in today's rapidly evolving landscape. At Parallel Works, our mission is to empower teams to right-size their HPC and AI infrastructure with tools that deliver unmatched agility, cost-efficiency, and scalability," states Matthew Shaxted, Parallel Works CEO.

Contact Us for a Demo

We encourage readers to reach out for consultations, demos, or further discussions about how they can implement hybrid HPC in their organizations. Try the Parallel Works ACTIVATE platform by connecting to your desired cloud account and get up and running within 10 minutes. Click this link to request a demo.