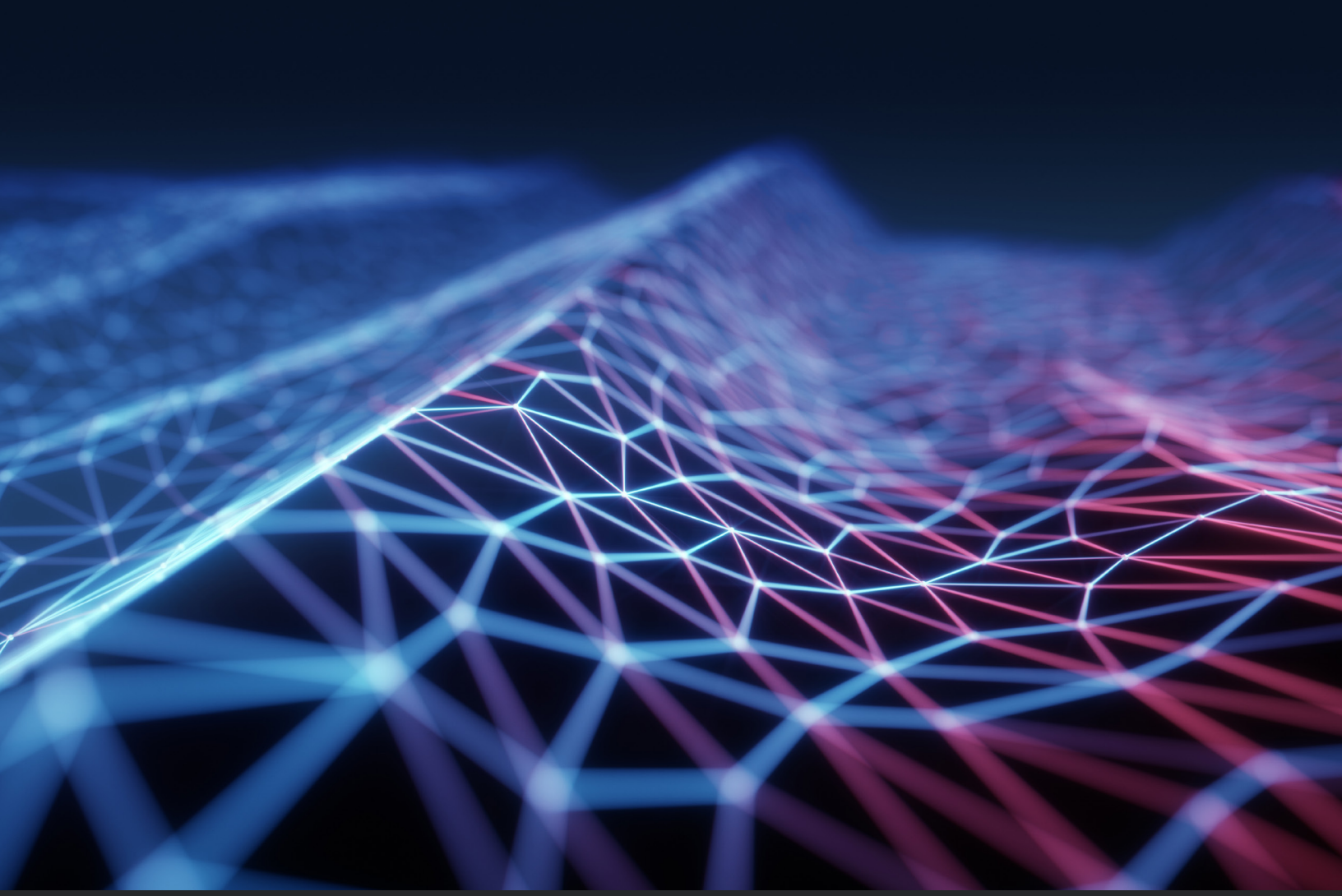


InsightHPC

Increase HPC speed and accuracy with the power of AI

PENGUIN COMPUTING

Contents

Solution-at-a-Glance

Features

- A comprehensive, proven compute architecture for AI-empowered HPC.
- High-performance DL training environment for large-scale, multi-user development teams.
- Workload-optimized compute servers.
- Rapid image-based provisioning of bare metal servers.
- High throughput, low latency networking.
- Flexible infrastructure to support heterogeneous compute.
- Cloud-native workload portability tools.
- Intuitive and centralized cluster and cloud management, reporting, and diagnostics.
- High-performance Virtual Desktop Interface.

Benefits

- Jump-start initiatives with a ready-to-run, AI-empowered, cloud-enabled enterprise HPC cluster.
- Discover more insights and take action faster.
- Use AI to narrow and optimize search and design spaces of HPC workloads.
- Dramatically reduce run times and increase accuracy.
- Engage HPC and AI platforms separately or as a unified data pipeline
- Improve the performance, adaptability, and accessibility of your data platforms.
- Achieve cloud-native workload portability.
- Build true hybrid clouds that scale across all compute assets.

AI-Empowered HPC

Artificial Intelligence has become an increasingly pervasive component of many aspects of modern technology. It also serves as a complementary technology to HPC. AI architectures share many architectural elements with those of its HPC counterparts. Both require balance between the computing, networking, and data systems to perform optimally. Both are connected by the fact that they are ultimately data processing and production platforms but perform complementary tasks in that effort.

Deploying HPC and AI platforms in a common infrastructure enables users to engage them for workload-specific tasks or combine them for a data-driven HPC pipeline. Adding AI and ML capabilities to a traditional HPC pipeline leverages data analysis to narrow and optimize the search and design spaces of HPC workloads. This approach can dramatically reduce run times and increase the accuracy of traditional HPC techniques.

These advanced pipelines are creating a convergence of HPC and AI that is being driven by large and growing amounts of data. When HPC and AI platforms are deployed in the same architecture, researchers can collaborate by leveraging a unified data infrastructure to operate on the same data for faster time to insight.

“Data is driving the convergence of HPC and AI.”

Designing the right AI-empowered HPC platform for your workloads is a complex task. Ensuring that the compute, storage, and networking subsystems are well designed individually and function in a balanced manner together is critical. With each new technological advance comes more choices. A poor design choice at any point in the process can negatively impact performance and reliability and can significantly reduce the value of your HPC investment.

Mitigating these risks, building optimal infrastructures, and planning for the future of your AI-empowered HPC environment is not simple. You need a secure HPC infrastructure that is optimized for your unique workloads and cloud needs and is engineered by a partner who knows HPC, AI, and the cloud.

Penguin Computing InsightHPC

Penguin Computing InsightHPC™ combines decades of HPC and AI design experience with the best in cloud-native technologies to provide proven, streamlined, cloud-enabled HPC architectures.

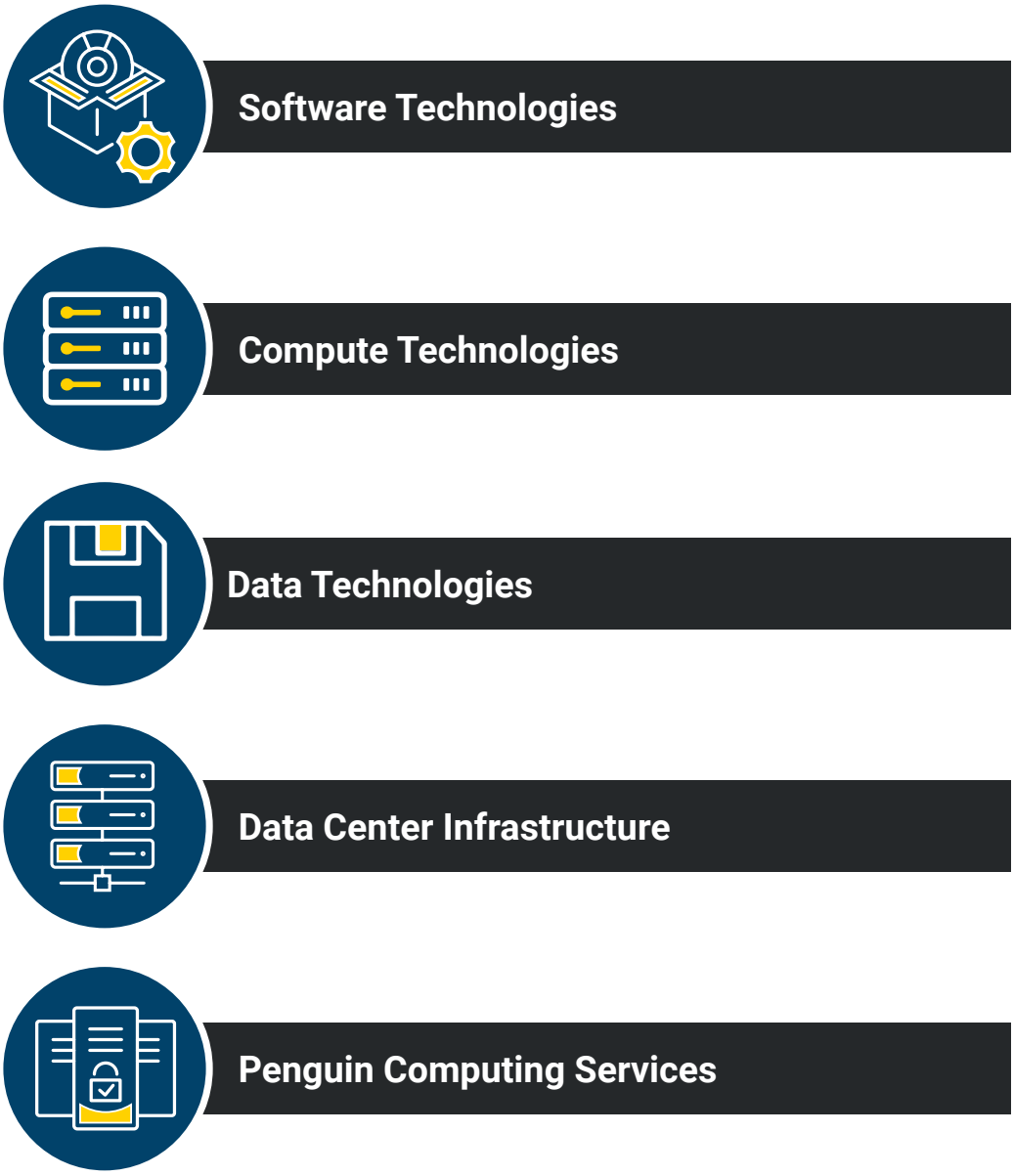
InsightHPC is built on HPC and AI-optimized server building blocks and cloud orchestration technologies to provide an on-premises, data-driven HPC cloud-native solution that allows you to leverage an AI pipeline to reduce time-to-insight for HPC workloads.

InsightHPC provides a complete software, hardware, and management platform built on our compute-optimized hardware, Red Hat Cloud Suite technologies, and Penguin Computing Scyld cloud and cluster orchestration and management software. This out-of-the-box, cloud-enabled HPC solution also leverages high-performance, low-latency networking and storage technologies to deliver optimized HPC for your workloads.

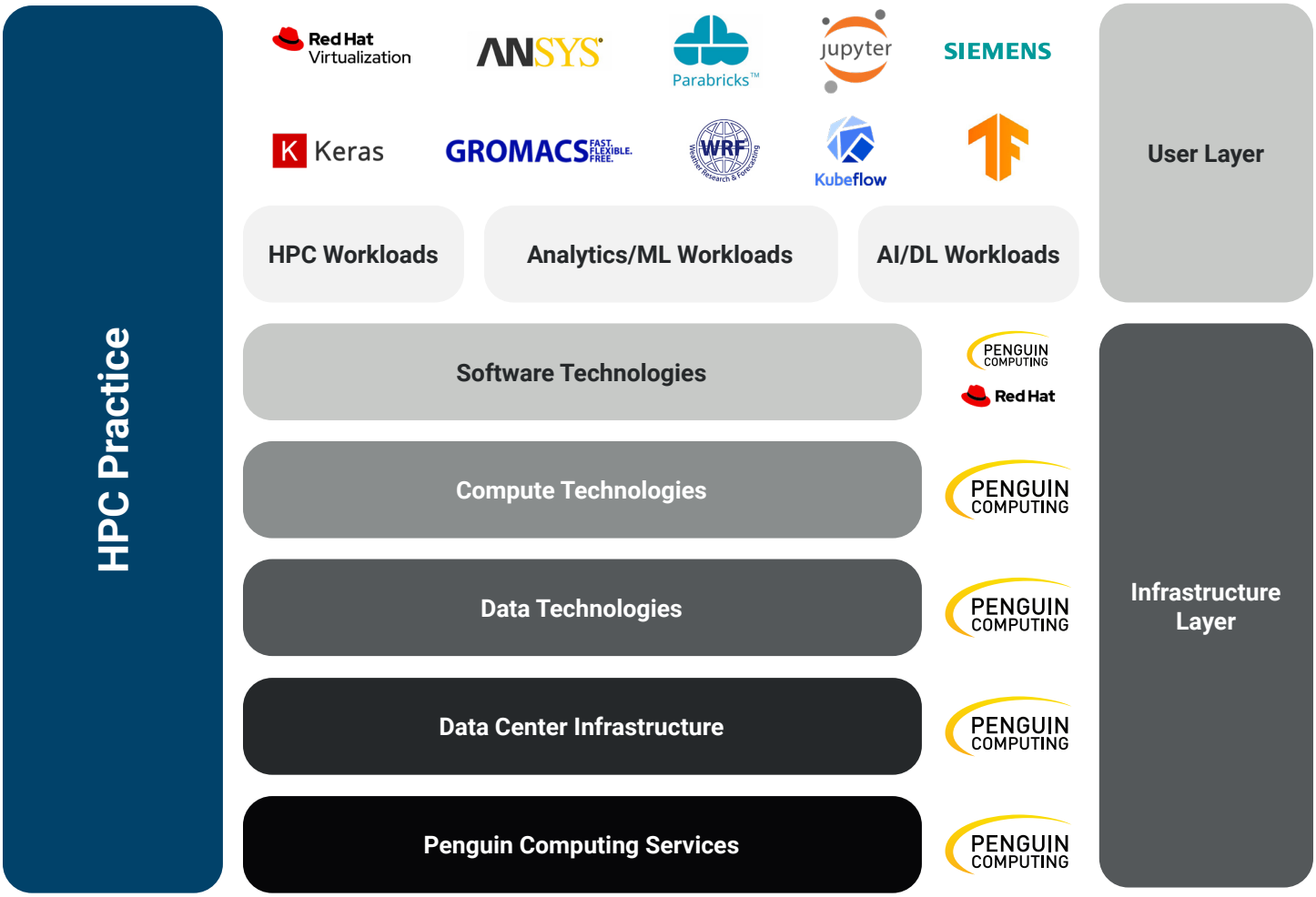
InsightHPC allows you to get the most performance out of your underlying HPC cluster from day one. With the powerful cloud and cluster management of Scyld ClusterWare and Scyld Cloud Manager and expert management from the Penguin Computing services teams, your end users can engage with these resources with the same ease and even better performance than they need to keep innovating.

You can combine InsightHPC with other Penguin Computing solutions for Data, Cloud, and AI/Analytics. When you combine the benefits from these different technologies, you can easily build complex, high-performance environments across many facets of your IT infrastructure.

InsightHPC includes:



Penguin Computing InsightHPC Components



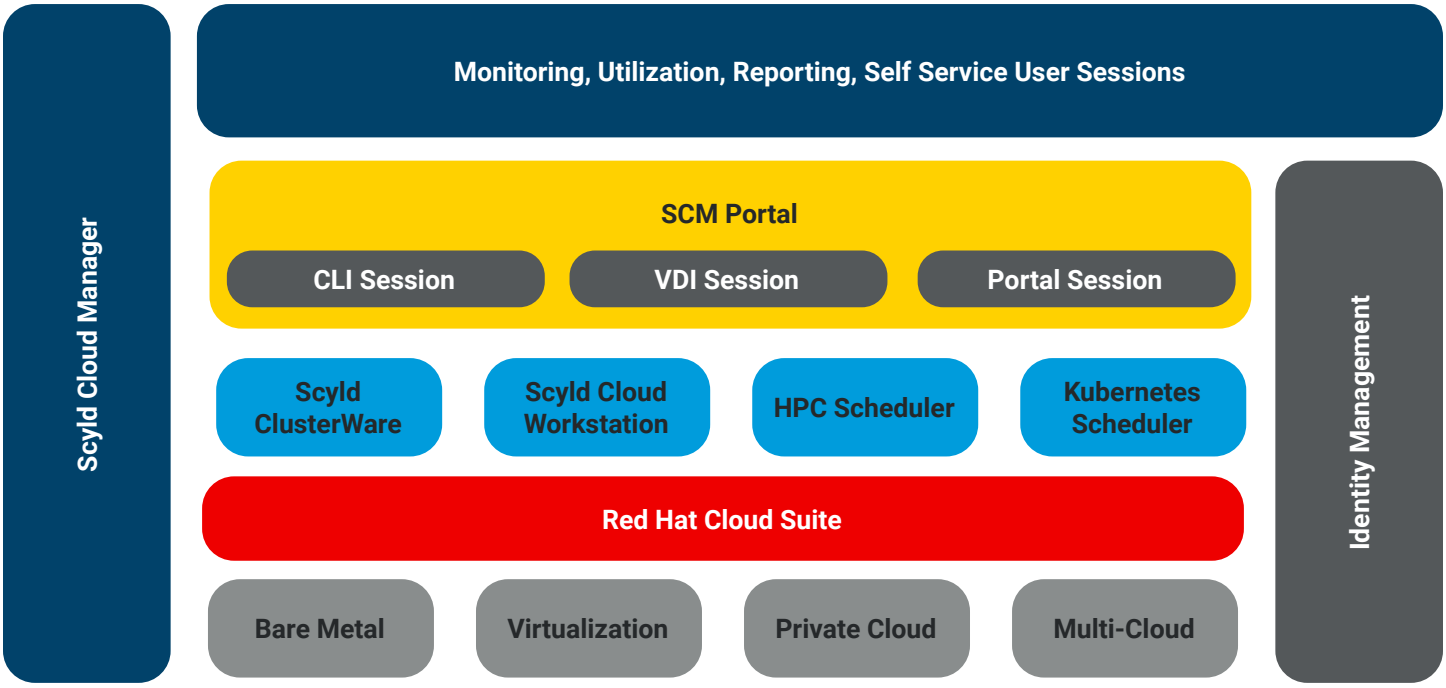
Software Technologies

Scyld Cloud Manager

HPC Cloud Portal

InsightHPC leverages Penguin Computing's Scyld Cloud Manager cloud software suite to enable a complete HPC cloud solution. Scyld Cloud Manager (SCM) provides administrators and users with a single pane of glass for registration, access management, system provisioning, and more. Managers can access and view cluster utilization metrics in a single location, giving them insight into what's happening on their cluster and who's using it, enabling them to plan environment upgrades and expansions for the next year or next three years. In addition

to powering the InsightHPC solution, Scyld Cloud Manager also powers Penguin Computing's public HPC cloud: Penguin On-Demand (POD).



Accounting & Chargeback

Many large organizations depend on internal chargeback and tracking models to ensure resources are being consumed by the users that help generate revenue or budget for the environment. The InsightHPC solution provides organizations with the same billing system that Penguin Computing uses to run Penguin On-Demand - extending the same billing capabilities into their own environments. Organizations can track job completion time, job resource utilization, and who ran the job - giving them the ability to bill by user, group, or project.

Virtualization & Bare Metal Computing

InsightHPC provides the best features of a cloud environment with the best features of an on-premises HPC cluster. Head nodes, login nodes, and other administrative nodes are virtualized to enable process redundancy and flexibility across the infrastructure. Compute nodes are provisioned as bare metal servers to ensure HPC jobs run optimally and consistently in the cluster. Users can expect predictable job performance from the cluster, and administrators can expect flexible and rapid provisioning across the environment. The InsightHPC solution is designed to integrate with workload optimized computing solutions, such as those available in Penguin Computing's TrueHPC™ solution.

Integration & Security

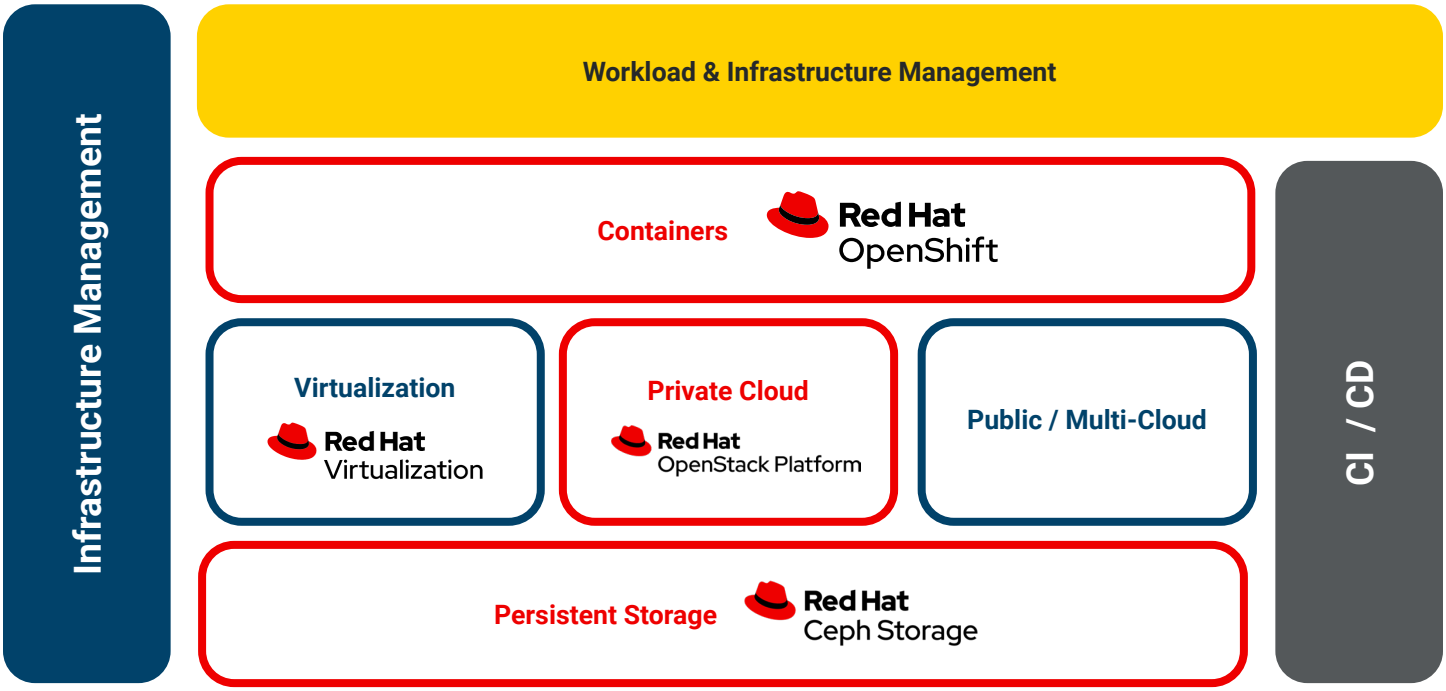
InsightHPC is tested and delivered as an integrated solution with security and support as paramount components. The InsightHPC solution can integrate into existing customer environments - providing the flexibility needed to integrate with many different networks, identity managers, authentication policies, and user workflow models. The InsightHPC architecture provides a reliable cloud-enabled foundation with endpoints that are integrated into the private HPC networks and services as well as enterprise networks and services. The architecture is built on a solid cloud-enabled framework providing a secure platform to integrate with and strengthen security compliance.

Scyld Cloud Manager Features

- Fully integrated, HPC cloud solution for administrators and end-users
- Guided self-registration Web portal to on-board users onto HPC resources
- Web portal interface allows access rights management and system monitoring
- Live and historical usage reporting for billing and chargeback management
- Web service application programming interfaces (APIs) for reporting, resource management, and access control
- Flexible VM configurations that support remote 3D visualization desktops

Red Hat Cloud Suite

Penguin Computing CloudBase™ with Red Hat® Cloud Suite provides a container-based application development platform, built on massively scalable cloud-native infrastructure, all managed through a common management framework. Customers can move existing workloads to scale-out cloud infrastructure and accelerate new cloud-based services for private cloud and application development. With CloudBase, an operations team can deliver public cloud-like services on premises to developers and business units while maintaining control and visibility. CloudBase enables a unified management framework to truly enable workload portability edge-to-core strategies to deliver the same user experiences on-premises or in public cloud environments.



Red Hat OpenStack and OpenShift

At its infrastructure foundation, InsightHPC leverages Red Hat Cloud Suite to build a private cloud based either on Red Hat OpenStack Platform with public cloud-like scalability, or Red Hat Enterprise Virtualization, based on high-performance virtualization. Both underlying choices provide secure, scalable foundations for hosting the OpenShift® Container Platform. OpenShift automates the development and administration of container-based applications.

Containerized, Cloud-Native Workload Portability

As cloud-native technologies like containers and Kubernetes mature rapidly, they are quickly becoming the preferred way to build new software experiences and modernize existing applications, workloads and workflows at scale and across on premises to public clouds, and multi-cloud. Creating value depends on the ability to deliver applications and workloads faster. This is being driven by the explosion of data-driven workloads in AI/ML, Analytics, IoT and other emerging technologies. Cloud-native technologies are driving this innovation culture. Enterprise customers now seek container development platforms that accelerate and simplify the development and operations (DevOps) of cloud-native apps wherever and however firms build and deploy them. InsightHPC not only provides comprehensive container infrastructure lifecycle operations from the data center to the cloud to the edge; it also helps developers modernize apps and innovate workloads with integrated service catalogs and microservices, service mesh, and serverless features.

InsightHPC powered by Red Hat OpenShift Container Platform delivers a balanced blend of development and operations features that:

- **Simplify cloud-native app development with rich development services** — The leading solutions draw developers in and jump-start both development and app modernization with microservices frameworks, serverless support, continuous integration and delivery (CI/CD) integrations, dependency management, and app lifecycle management features like code quality checks and vulnerability scanning. InsightHPC helps developers focus on business logic with comprehensive service catalogs and prebuilt DevOps automations and integration.
- **Enable distributed infrastructure operations from data center to cloud to edge** — Data-driven workloads are increasingly distributed and hybrid. InsightHPC offers model driven configuration, monitoring, security, and cluster lifecycle features for unified multi cloud cluster operations. InsightHPC extends operational control to the edge and supports thousands of clusters.

Scyld Cloud Workstation

Scyld Cloud Workstation is a remote desktop solution designed to deliver real-time interactive enterprise class visualization through a standard browser without plugins. Users simply connect to the remote environment from virtually any device running Firefox, Internet Explorer, Edge, Chrome, or Safari.

3D Accelerated Engineering VDI

Scyld Cloud Workstation provides engineering class visualization for HPC and AI/ML Engineers running graphical applications, such as CFD, CAE, or FAE codes. Scientists in weather and chemistry are another class of users that commonly need to visualize data. Traditional HPC and AI/ML environments require users to download large data files to on-premises workstations for pre/post processing, model development, and data analysis offline from the computing resource and centralized storage. This is a time-consuming process that makes it hard to create an efficient workflow with predictable time to results.

Scyld Cloud Workstation offers significant time savings by moving pre- and post-processing to a workstation with direct access to a cluster's data storage — eliminating the need to download large data files. Users can use the same GUI tools as on their local workstations, ensuring continued productivity.

Remote Collaboration on Shared Desktops

Scyld Cloud Workstation enables multi-user collaboration and remote desktop access for up to ten temporary or permanently authorized users. Desktop control can be passed from one user to another while our QoS algorithm intelligently adjusts the frame rate per client to ensure an optimal experience. Through Scyld Cloud Workstation, customers can deploy large scale remote desktop environments using open source or commercial virtual machine platforms for provisioning.

Expansive OS Support

Scyld Cloud Workstation can run in any bare-metal or virtualized environment, enabling you to couple advanced VDI solutions into an existing environment. Whether you are enabling high-end 3D accelerated desktops, deploying GPU enabled workstations for AI/ML, enabling remote access to software suites for content creators, or building a platform to enable thousands of users with intuitive secure access to a familiar remote desktop, Scyld Cloud Workstation is able to meet your needs.

With support for Linux, Microsoft Windows and Apple macOS, integration with existing workflows allows for frictionless enablement of your user base into a familiar environment.

Secure Remote Workforce

Scyld Cloud Workstation supplies secure access through HTTPS, requiring no additional ports through the firewall. This unique architecture saves bandwidth, simplifies implementation for IT departments, improves image quality, and ensures near-universal accessibility for users. IT departments can use custom SSL certificates and couple authentication into centralized identity

Scyld Cloud Workstation Features

- Browser-based, HTML5 remote desktops
- No client installation or plug-in needed
- Supported on macOS®, Linux, and Windows®
- Enables 3D accelerated interactive workflows
- Operates in a wide variety of network conditions
- Secure HTTPS authentication from anywhere
- Collaborative, secure multi-user sessions
- Supports multiple monitors
- Intelligent QoS for minimizing network usage
- Dual channel stereo audio support
- Up to 4K, 3840x2160 resolutions at 30fps
- Outperforms traditional VDI

managers through Scyld Cloud Workstation's ability to pass authentication onto the operating system.

Lossless Remote Desktops

For users that require pixel accuracy, Penguin Computing delivers an optional client that delivers two classes of high quality video outputs at up to 4K UHD:

- Visually Lossless provides high fidelity video at near lossless compression to optimize your network.
- Alternatively, Lossless Video delivers uncompressed video streams.

Users can easily toggle between visually lossless and completely lossless interactive sessions.

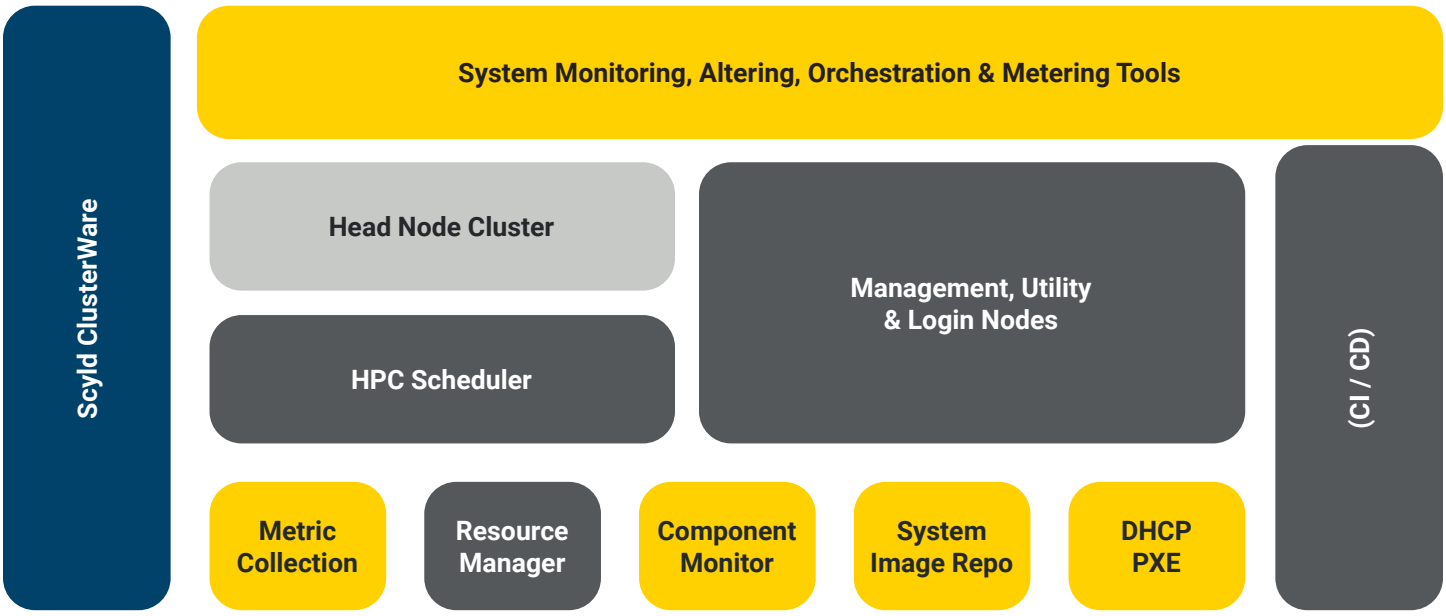
Scyld ClusterWare

HPC Cluster Orchestration

InsightHPC leverages Penguin Computing's Scyld ClusterWare cluster orchestration software. Scyld ClusterWare provides a complete HPC environment that supports Slurm, OpenPBS, and TORQUE to handle the scheduling and queueing of HPC jobs. Scyld ClusterWare also provisions the hardware to operate as a single, unified cluster by booting compute nodes using PXE, establishing IP address using DHCP, monitoring node health, and collecting metric data across the cluster.

Alerting and Monitoring

Scyld ClusterWare supports alerting features that can integrate into enterprise communication tools, such as email, Slack, PagerDuty, and more to send out important alerts to different groups within an organization regarding the current status of an InsightHPC implementation. This information is also centralized to the head node, providing system administrators with a single pane of glass displaying the status of every node in the cluster.

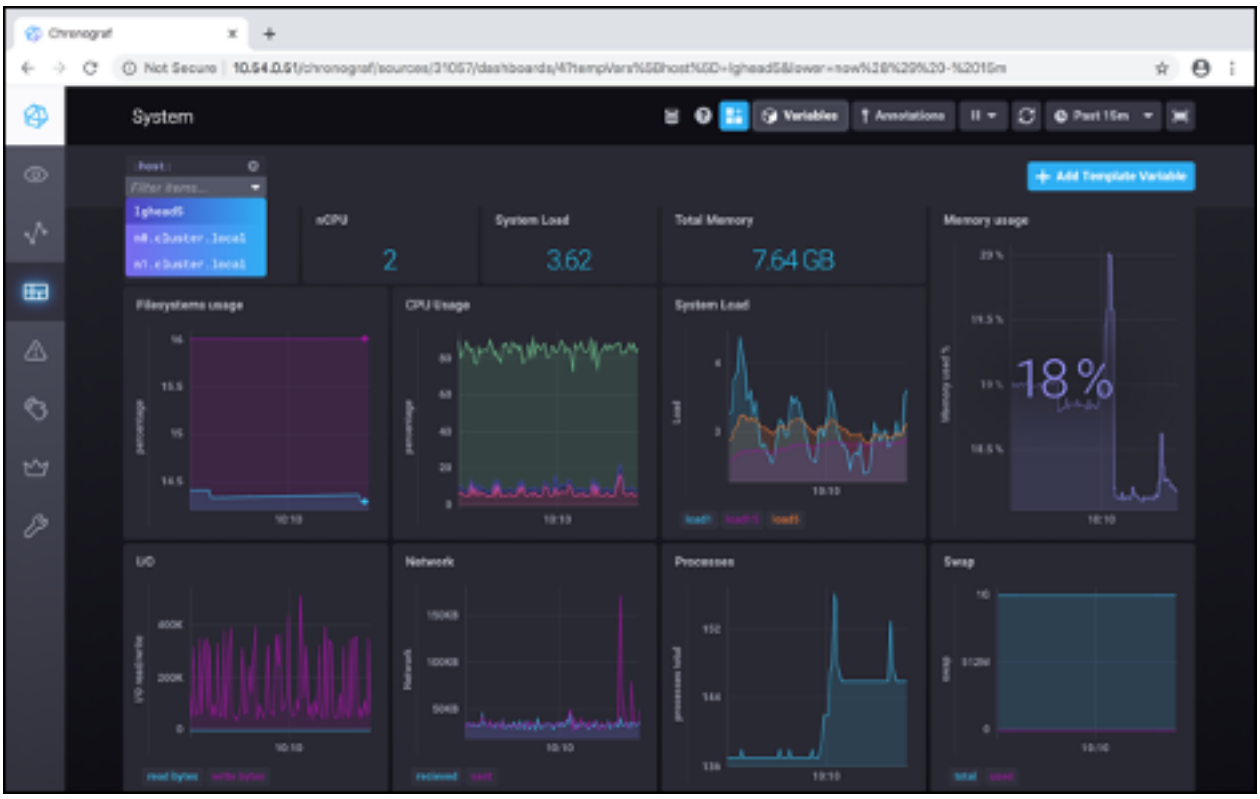


System Image by Application

Scyld ClusterWare also supports custom system image deployment. Users can save a compute node image into a repository managed by the Scyld ClusterWare head node. System images can be completely different from the operating system that the head node uses. The head node could be running RHEL 8, while the compute nodes are running RHEL 6, RHEL 7, RHEL 8, Ubuntu 16.04, Ubuntu 18.04, or some combination of operating systems and system versions across the cluster.

Scyld ClusterWare Features

- Rapid provisioning for technical computing environments
- Designed to manage optimized HPC clusters and their coupled enterprise services
- Single source for tested HPC middleware (MPI implementations, HPC schedulers)
- Image-based node management facilitates simplified change management
- Flexible provisioning options (for example, diskless, diskfull, network mounted)
- Robust high availability architecture prevents downtime when unexpected failures occur
- Supports SELinux in MLS mode and FIPS 140-2 implementations
- Monitoring GUI for visualizing system telemetry and building custom dashboards
- Notification and alerting integration with email, Slack, and PagerDuty



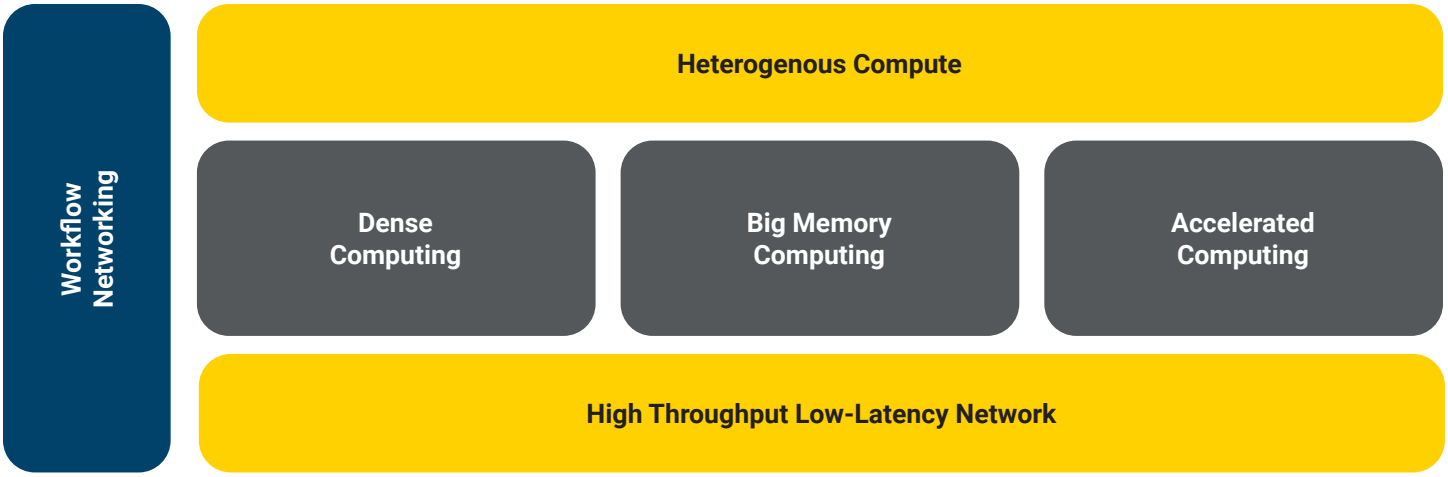
Compute Technologies

Heterogeneous Compute

InsightHPC leverages industry-leading technologies from Intel, AMD, NVIDIA and other technology providers to enable a complete technology ecosystem supporting many different workloads. HPC workloads often require high core count, high clock speed, high memory bandwidth, low latency communication, and/or accelerated computing using GPUs, FPGAs, and ASICs. InsightHPC supports heterogeneous computing environments within a single architecture using workload-optimized server building blocks for many types of high performance workloads.

Workload-Optimized Servers

HPC workloads typically require a large number of cores and high core clock speeds to achieve the best performance possible. These workloads also require high performance interconnects, because many HPC workloads span multiple servers, requiring constant node-to-node communication that benefits from high-throughput and low-latency network technologies. Optimized server building blocks for HPC workloads need to provide many cores with high clock speeds and low-latency, high-throughput interconnects to provide the best application performance possible.



Memory-centric workloads call for additional server memory resources to support applications that require extreme read and write performance and extremely low latency. Optimized server building blocks for big memory computing require as much memory bandwidth, capacity, and clock speed as possible to provide the best application performance.

Accelerated computing workloads require enterprise accelerators, such as GPUs, FPGAs, and ASICs, to drastically improve the performance of certain applications. Optimized server building blocks for computing accelerated require in-system, device-to-device communication optimizations to ensure that accelerators can communicate with CPUs, SSDs, NICs, and other accelerators without communication bottlenecks in order to provide the best application performance possible.

High Speed Low-Latency Interconnects

The InsightHPC solution supports the leading high-throughput, low-latency networking interconnects that help maximize the performance of an HPC cluster for certain workloads.

Data Technologies

HPC is moving toward data-driven workloads that consume and generate large amounts of data. This data growth drives the need for data solutions that can scale to exabyte capacities. HPC environments have data requirements that create data workflow and infrastructure challenges related to management and orchestration.

Data I/O requirements weigh heavily on the overall success of an HPC solution. I/O patterns and performance vary across different tiers of storage in the environment — from memory, to flash, to cold storage. InsightHPC integrates with the data solutions in Penguin Computing's Data Practice, which cover the entire spectrum of I/O — from memory, to flash, to cold storage — to support the entire data lifecycle.

Big Memory Computing

Some HPC workloads can require massive memory storage or high memory performance. By utilizing the LiveData™ solution, InsightHPC can support memory-centric workloads that require high memory capacity, persistent memory, and high memory tier performance.

High Performance Storage

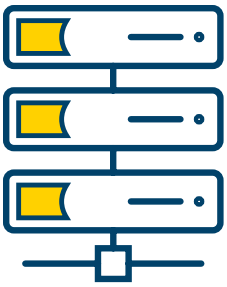
HPC workloads require high performance storage that can fulfill the ingress and egress demands of high performance workloads. InsightHPC can be paired with the ActiveData™ solution to provide high performance storage for data-heavy computing workloads. Just as with the InsightHPC solution, ActiveData can leverage the industry-leading technologies that are tuned for specific customer workloads.

High Capacity Storage

HPC workloads often ingest or output massive amounts of data that must be kept in a general purpose storage environment when not being used for computing. InsightHPC can be paired with the DeepData™ solution to provide scale out capacity optimized storage tier best suited for storing long-term data.

Multi-Site Data Fabric

Some HPC environments require connectivity to the cloud or another site. Some workloads might require cross organizational collaboration on datasets that span multiple locations. InsightHPC can access data sets across the world as if they were local using the DataNexus™ solution.



Data Center Infrastructure

InsightHPC can be built using both a traditional 19” rack platform and a modern 21” OCP (Open Compute Project) platform. Traditional 19” rack infrastructures are supported in almost every data center worldwide and in a variety of dimensions. Modern 21” OCP rack infrastructures require data centers that can support the most demanding physical and power densities. Penguin Computing has partnered with leading data center facility pioneers who can support the demanding characteristics of today’s HPC platforms.

Power

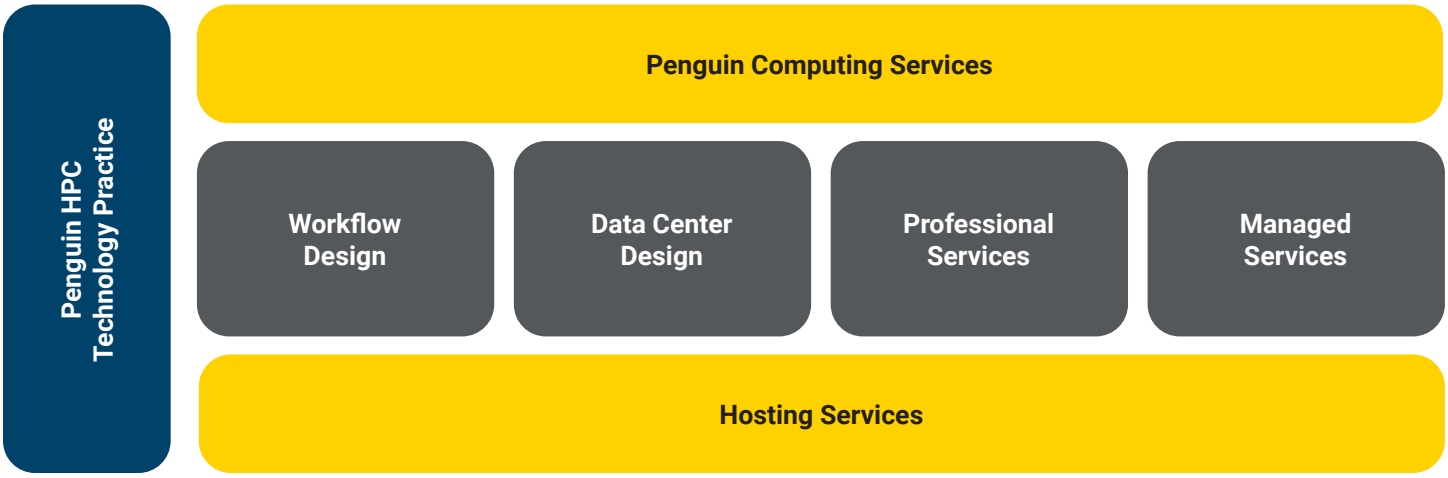
InsightHPC supports three-phase 50A or 60A, 208V, 277V, or 480V power options as well as A+B fully redundant power, or N+1 redundant power. 21” OCP also supports 12V or 48V power delivered directly to the servers, which enable much higher power density per rack.

Cooling

InsightHPC can be air cooled with traditional HVAC equipment. Penguin Computing recommends using a combination of air cooling and liquid cooling when deploying InsightHPC into a data center not designed for high-power equipment. Rear Door Heat Exchangers capture hot air exhaust at the rear of the rack, and can be deployed on most 19” and 21” rack infrastructures. InsightHPC is also designed to integrate Direct-To-Chip cooling options that capture heat directly from the CPU block. This cooling solution removes 85% of server heat before it’s transferred into the air, and can be used in select 21” infrastructures.

Penguin Computing Services

InsightHPC is a comprehensive, end-to-end solution that organizations can leverage to jump-start their HPC initiatives. In some cases, the solution will directly meet the needs of the organization, right out of the box. However, most often there will be additional design, deployment, integration, and hosting considerations that need to be addressed.



Penguin Computing provides services that consider rack and floor space, how to scale the environment, maximum rack power consumption, power phase balance, efficient heat removal, and the optimal networking topologies when using low-latency, high throughput interconnects.

InsightHPC is supported by Penguin Computing engineering services, including Design Services, Professional Services, Managed Services, and Hosting Services.

Data center hosting services are offered through Penguin Computing's strong partnerships with data center service providers. Our partners can provide the space, power, and cooling InsightHPC needs — as a service.

Design Services

Workflow Design

- Software Orchestration
- Compute Performance
- Multi-Node Communication
- Data Storage and Data Tiering
- Data Ingest and Egest
- Environment Sizing

Data Center Design

- Rack and Floor Space
- Environment Scalability
- Maximum Power Consumption
- Power Phase Balance
- Efficient Cooling and Heat Removal
- Optimal Networking Topologies

Professional Services

Stand Up and Initialization

- System Burn-In Testing
- Racking and Cabling
- Software Installation & Tuning
- On-Site Deployment and Integration

Hosting Services

Data Center Hosting

- Penguin Data Center
- Customer Data Center
- Power, Space, and Cooling Management
- Monthly or Annual Billing (As-A-Service)

Managed Services

System Administration:

- Complete Hands-Off Experience
- Augment Existing IT Capabilities
- Collaborate with Penguin Support
- Tens to Thousands of Servers
- Terabytes to Exabytes of Data
- Multi Data Center Support

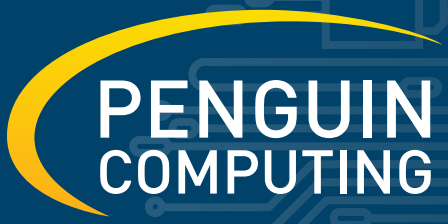
Conclusion

Penguin Computing InsightHPC™ is built on HPC- and AI-optimized server building blocks and cloud orchestration technologies to provide an on-premises, data-driven HPC cloud-native solution that allows you to leverage an AI pipeline to reduce time-to-insight for HPC workloads.

Penguin Computing can apply our decades of experience to create quality, integrated solutions for our clients. We offer a wide range of professional and managed services that can quickly bring your artificial intelligence, machine learning, and deep learning initiatives products to production.

Contact Us

Use this **form** or call Penguin Computing today at 1-888-736-4846 to find out how you can leverage AI pipelines to reduce time-to-insight for your HPC workloads.



Expanding the world's vision of what is possible