Powerful Model and Orchestration Designed for Next Generation AI Systems

Solution Brief

Highlights

Easily manage, run, and scale your AI & ML workloads on Lenovo HPC with WEKA Data Platform and UbiOps

- The fastest route to production-grade, highly scalable, ML / AI systems. Deploy AI models and run ML training up to 10x faster. Manage countless AI workloads simultaneously from a single control plane.
- Lower your product TCO and increase productivity. Reach faster convergence in your model dev cycle and a shorter time to market with AI solutions.
- Leverage a highly scalable, distributed, data and compute infrastructure for even the most demanding workloads.
- Get on-demand access to the entire Lenovo HPC stack on-prem and in hybrid cloud. Lenovo TruScale HPC gives you the freedom to focus on your core competencies and not worry about HW and SW stack management.

Building and running AI systems is a challenging task

AI & ML workloads require organizations to manage both code and massive data sets. For many teams, building and running AI systems is a challenging task: a long time to train models, data management, workload management and orchestration in a complex cloud landscape, cost overruns, and dealing with large data pipelines among others. All these challenges take resources and money to manage operations.

To overcome these challenges, Lenovo HPC, WEKA, and UbiOps provide HPC and AI teams with a powerful, production-ready solution for deployment, training, and management of machine learning and deep learning models and workflows.

Whether you are at the start of your HPC/AI journey or already have AI/ML integrated into your organization, you can benefit from this solution and start running real-time, data-intensive, applications today:

- Train, deploy, and manage all your data science and machine learning models in a turn-key production environment with a scalable compute infrastructure.
- Scale your AI workloads dynamically with usage and grow easily with autoscaling capability.
- Leverage state-of-the-art high-performance GPU for accelerating deep learning.
- Create and orchestrate workflows – reuse and share modular pipeline steps and data.
- Train models using data in the petabyte scale on-prem and/or in the hybrid cloud.
- Gain on-demand access to powerful hardware with serverless workload distribution. Deploy models on your own infrastructure or private cloud or scale out to hybrid and multi-cloud environments to optimize costs, compliance, and performance.

The challenges

Data and AI teams in both large and small organizations often encounter the same challenges when it comes to developing and deploying AI & ML systems at scale:

- **Infrastructure:** Getting the infrastructure right to run ML models, scale-out and handle large amounts of data in a production-grade setting is often a huge investment of time and DevOps/IT resources.
- **Data Quality**: Maintaining the quality of data used to train and test models is critical for AI & ML.
- **Applications**: it can be challenging to ensure that data is consistent, accurate, and up to date at scale.
- **Model Versioning:** Keeping track of different models and their respective performances can be difficult, especially as the number of models deployed increases.
- **Collaboration:** Collaboration between data scientists and operations teams can get complicated and inefficient.
- **Monitoring and Debugging:** Monitoring the performance of machine learning models in production environments and debugging issues that arise can be very time-consuming and complex.
- **Security and Compliance:** Ensuring that ML models are deployed and managed in a secure and compliant way is crucial, but it can be challenging to achieve at scale.
- **Lack of Automation:** Data science and machine learning often still requires manual processes, which can be time-consuming and lead to errors. Automation is essential for scaling MLOps processes.
- **Limited Resources**: Organizations may need more resources to devote to MLOps, making scaling difficult.
- **Hybrid-cloud & on-prem deployments:** Dealing with multiple environments where your data and compute resources reside can be a very challenging task for many teams.

The latest AI computing solution for HPC AI & ML applications

Lenovo, WEKA, and UbiOps have joined forces to offer the best-in-class solution to address these challenges. This solution dynamically allocates resources for the entire workflow; therefore, manages both the infrastructure for your development, as well as your operational needs.

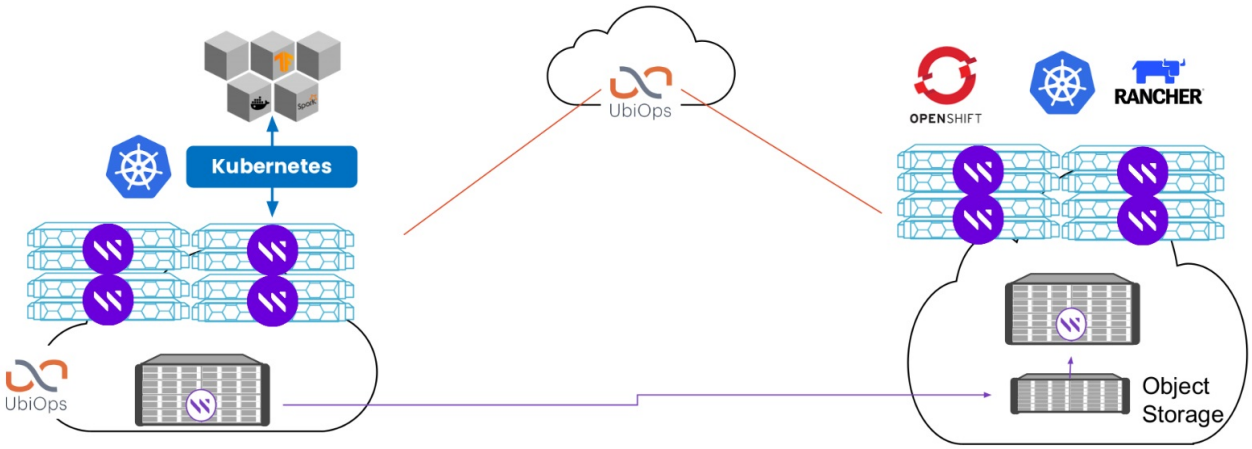

Figure 1. AI computing solution for HPC AI & ML applications

Advantages provided by UbiOps

UbiOps gives your team powerful AI model serving and orchestration capabilities with unmatched simplicity, speed, and scale. It enables the management, training, and running of countless AI/ML jobs simultaneously, from a single control plane.

UbiOps lets data scientists easily collaborate on models, data, and workflows. It comes with a realm of built-in MLOps features like version control, simple rollback, monitoring, and logging for deployed models and jobs. UbiOps as an integrated MLOps solution increases your team's productivity and cuts your time-to-market for AI solutions and products.

Moreover, UbiOps can be used as SaaS or installed across hybrid-cloud environments. Gain on-demand access to powerful CPU & GPU hardware with serverless, multi-cloud workload distribution. Optimize compute nodes to match your models and build modular, optimized pipelines with our data workflow management tool.

Overview of Lenovo and WEKA advantages

WEKA has built a software-defined Data Platform that leverages Lenovo's cutting-edge servers and storage, and fast networking technologies to unleash the value of your data.

WEKA Data Platform running on Lenovo servers delivers consistent lightning-fast access to data at terabytes to exabytes scale when needed across the AI workflow. This fast access is supported by the award-winning hardware platform from Lenovo TruScale and the patented architecture supports both small and extremely large configurations without compromising performance. WEKA Data Platform eliminates the need for multiple storage options and data copies across MLOps workflows, reducing operational complexity, enhancing pipeline efficiency, and increasing GPU utilization.

With WEKA, a single data platform supports all popular data access methods, including the POSIX-compliant file system, NFS, SMB, S3, CSI for Kubernetes, and GPU Direct Storage (for direct data movement between GPUs and storage). In addition, because of SDDP, it enables both on-prem, private cloud, and hybrid cloud deployment. WEKA also makes moving data across different cloud configurations.

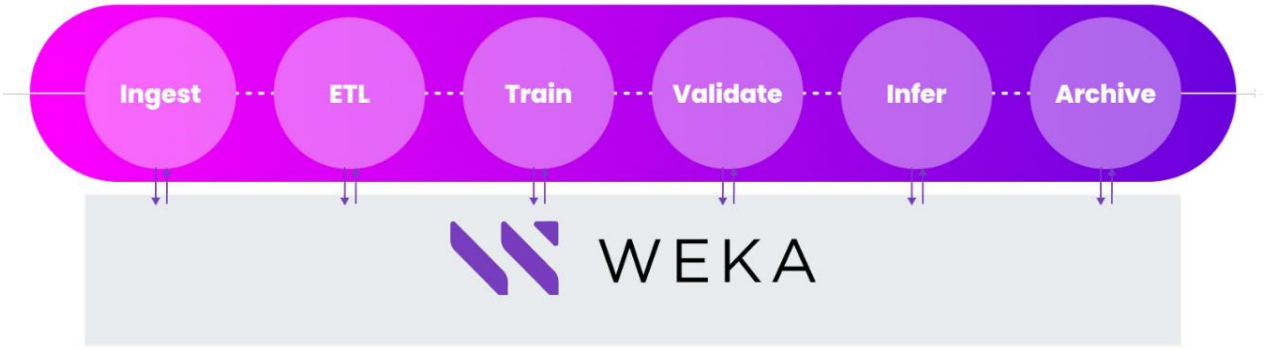

Figure 2. WEKA workflow

Additional benefits include:

- **Efficient move and backup of data.** Advanced data management capabilities enable data movement quickly and efficiently between different nodes and simplify backups to local or remote cloud regions.
- **Tier automatically.** WEKA can automatically tier cold data to low-cost object storage, on-prem, or cloud for better economics. All data remains in the namespace, and metadata stays on the flash tier for fast access.
- **Ensure security.** The WEKA data platform was architected to ensure the security of your data with advanced authentication, in-flight and at-rest encryption, and flexible key management.

Lenovo TruScale advantages

Lenovo TruScale for HPC provides simplified access to HPC technology through a flexible on-demand model combining hardware, software, and services into a single, configurable solution with a predictable and affordable regular fee.

Delivered as a service, TruScale HPC helps accelerate innovation by giving access to the latest industry-leading solutions through a no-risk, no-surprise, pay-as-you-grow model. In essence, it combines the best of on-prem and cloud in a single solution.

The benefits are clear. Firstly, there is no need for capital investment, and you can avoid long procurement cycles with our OPEX model. Additionally, TruScale simplifies data center management and the provisioning of IT, alleviating budget and staffing constraints. This gives you the time and resources to focus on accelerating business outcomes and achieving competitive advantages. TruScale HPC managed services take care of upgrades, maintenance, and growth plans for the customer. Therefore, you focus on your core competency while Lenovo provides the infrastructure along with our partners.

Some of the advantages of Lenovo HPC include:

- The #1 in server reliability for seven consecutive years
- Delivered products to six of the top 10 hyperscalers in the world.
- Certifications and coverage in 100+ markets.
- Central financing/invoicing/contracting/delivery model to support complex global deals.
- Well-established as-a-service programs in place.

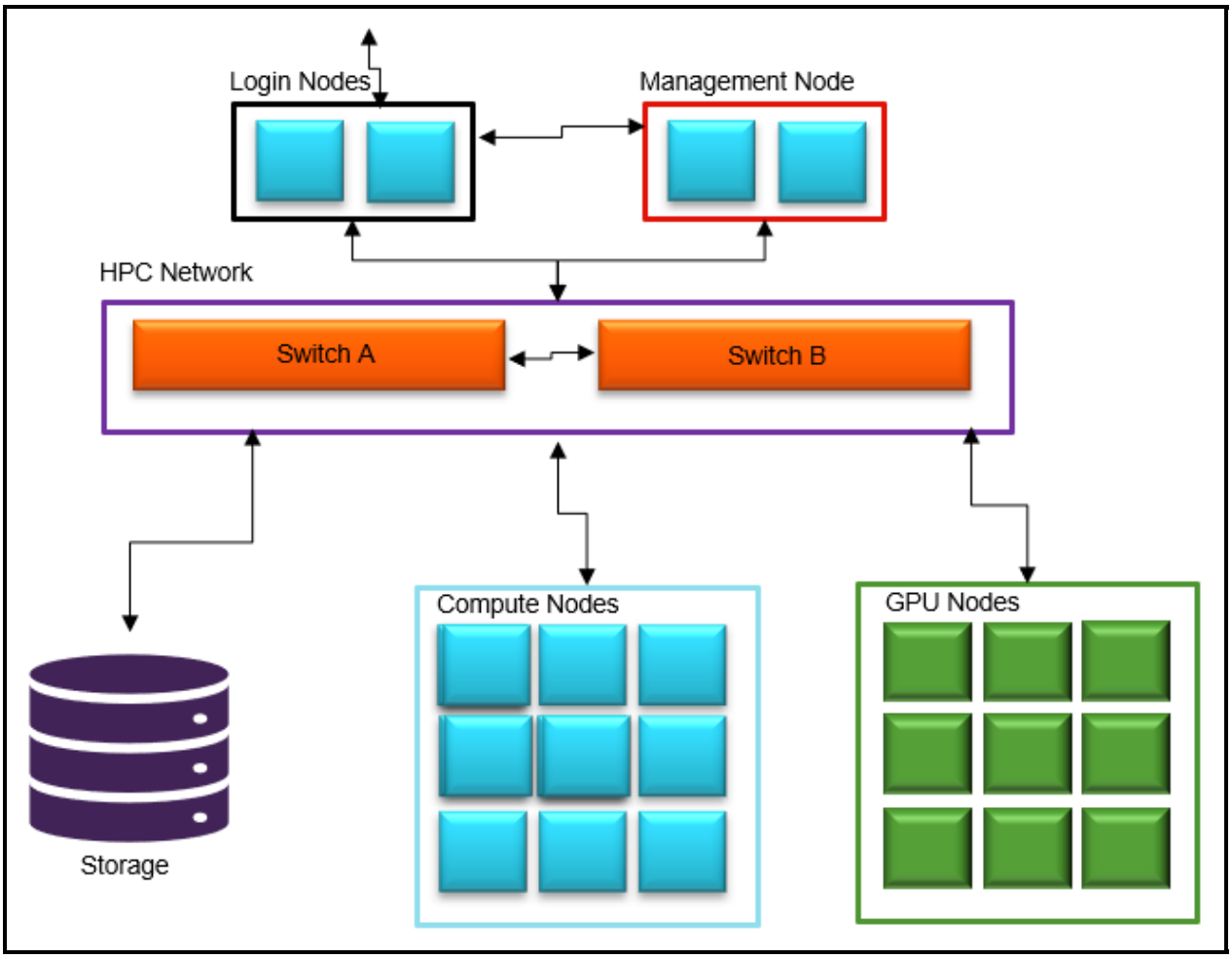

Figure 3. Sample TruScale HPC architecture

Conclusion

Lenovo, WEKA, and UbiOps joined forces to provide a complete solution to run, manage and orchestrate data intensive AI/ML workloads and provide MLOps capabilities to data science teams:

- Consolidates siloed stacks into a single governed data platform running on-prem and/or in a hybrid-cloud.
- Designed to deploy, run and manage state-of-the-art AI/ML solutions at scale.
- Provides your team with MLOps capabilities for capturing and tracking all project artifacts, including code, package versions, and parameters, to establish full visibility, repeatability, and reproducibility at any time for both model development and official release management.
- Features fully managed, best-in-class auto-scaling storage and compute nodes so that your team can have peace of mind and focus on their competencies instead of day-to-day infrastructure challenges.

Consumption-based billing makes this not only a powerful, but also very cost-efficient solution for any sized organizations to kick-start their AI/ML journey.

Get started today

If you are an organization investing in artificial intelligence, machine learning, and deep learning initiatives, this solution simplifies your journey with excellent ROI and expedites time to market. Email us at Truscale@lenovo.com, visit us at Lenovo.com, or contact your authorized representatives to learn more about this offer.

About the author

Mark Azadpour is a Sr. Strategic TruScale Product Manager at Lenovo, where his focus is on HPC, AI, virtualization (cloud) and software defined infrastructure. He has decades of enterprise software and hardware experience, and holds a PhD. In Computer Engineering and an MBA in strategic Marketing.

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP1788, was created or updated on August 2, 2023.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1788
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1788.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®

Other company, product, or service names may be trademarks or service marks of others.