Scalable MLOps & AI Training Solution with cnvrg.io and Lenovo ThinkSystem AI-Ready Servers
Solution Brief

Machine Learning and the data science economy will drive $3.9 trillion in 2022. Organizations, departmental roles and project flows are being redefined. Programming-heavy tasks and mega-experimentation are blended with infrastructure, DevOps and production monitoring. All can be unified onto one pipeline to bring them all together. A new platform is needed for this shifted and rapidly growing market. Enter the MLOps control center - cnvrg.io

Despite extraordinary advancements in the field, Machine Learning and Deep Learning have seen slow success in the enterprise. It's reported that nearly 80% of enterprises fail to scale AI deployments across the organization. Data scientists are forced to spend 65% of their time on non-data science tasks and managing disconnected infrastructures. Deploying and maintaining ML solutions at scale demands a unified MLOps platform to operationalize the full ML lifecycle from research to production.

Benefits of this solution

- Optimize Server Utilization
- Maximize Data Science Productivity
- Quick Integration & Simple UI
- Accelerate Development & Deployments

- Collaboration & Project Visibility
- Decrease IT Technical Cost
- Enterprise-Ready Software Out-of-the-Box
- Advanced Resource Management & Visibility

Industry Challenges

Data Scientists Inefficiency: 65% of a data scientist's time is spent on non-data science tasks: DevOps, hardware and software configuration across diverse compute architectures, versioning, data management, tracking experiments, visualization, deployments, monitoring performance and more.

Resource Management: ML and data science projects are treated as a monolith, each project needs to run in one setup: either on public cloud or on-prem. Different components of a single pipeline cannot be deconstructed and run on different environments. As of 2019, 57% of data science projects are executed on-prem, while 43% are executed on public cloud.

IT Visibility: Data science projects infra costs are more than doubling each year, however IT teams lack the tools to manage, optimize and budget ML resources properly, on-prem or in the cloud.

ML Technical Debt: More than 80% of ML models don't make it to production due to engineering heavy tasks like data verification, monitoring, configuration, compute resource management, serving infrastructure and feature extraction.
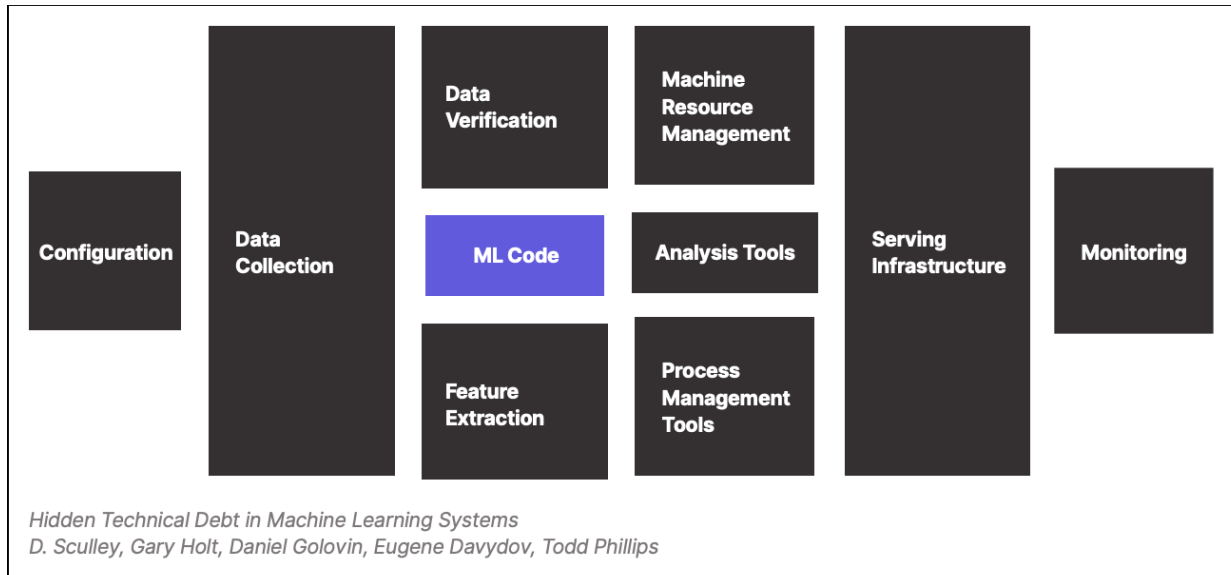
Figure 1. Hidden technical debt in ML systems (from Hidden Technical Debt in Machine Learning Systems by Sculley, Holt, Golovin, Davydov, and Phillips, presented at NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems, December 2015)

Key Features

MLOps Utilization Dashboard

Improve visibility and increase infrastructure utilization by up to 80% with advanced resource management and visibility across all ML runs. cnvrg.io gives data engineers and IT the tools to monitor utilization, properly size the compute components and visualize who is using what, with extensive visualization of cluster utilization.

Heterogeneous Compute Pipelines

Launch and manage end-to-end heterogeneous ML pipelines where each component or stage (in a single pipeline) can run on a different compute architecture that is optimized for the specific use-case: preprocessing and/or inferencing on CPU, deep learning training on GPUs and inference in the edge

Open Platform

cnvrg.io was designed to be an open and flexible platform. Users can develop their own models, run experiments of their choice and modify behavior and code with Jupyter notebooks (or RStudio, VSCode). New tools and utilities can be easily integrated into the platform. cnvrg.io's container-based platform offers flexibility and control to use any image or tool.

High-performance Infrastructure

cnvrg.io provides better utilization of your infrastructure and Lenovo bring the highest performance for both your AI training and Inferencing workloads. The Lenovo ThinkSystem line-up of AI-Ready servers are award winning AI products with demonstrated performance leadership for AI workloads. Our unique modular design ensures that we support the architecture of your choice in a standard and scalable platform and when you're ready to move your AI workloads into production, our high performance AI-Ready edge & inference servers are designed to support the most demanding performance requirements in the data center and at the edge.

- Unique modular design
- Edge to Core

Learn more at: https://www.lenovo.com/AI

Integrated Solution Benefits

Advance ML infrastructure Solution

Lenovo is a premier AI infrastructure supplier, with a range of CPU only servers, CPU+GPU rich servers, edge servers, gateways and support infrastructure to span the wide variety of AI use-cases and customer requirements. cnvrg.io bundled with Lenovo ThinkSystem AI-Ready servers creates a managed, coordinated and easy solution to consume ML infrastructure.

cnvrg.io serves the extreme needs of DevOps and IT Ops by providing:

- Sophisticated meta-scheduling that accomplish high infrastructure utilization (>90%)
- Infrastructure utilization dashboards, with different views (per data scientist, project, server, etc). Ability to segment available compute infrastructure per team or as one pool.
- Quick and easy onboarding, all based on containers and packaged for simple installation and stand-up (helm, Kubernetes, etc.).
- Accelerated workloads, with ability to launch any ML or DL framework in one click with direct integration to optimized containers from both Intel and NVIDIA optimized to increase performance of Lenovo servers featuring 2nd gen Intel Xeon Processors and NVIDIA GPUs for AI workloads.
- Compatible with existing infrastructure, hence no silos need to be created. Growing your Lenovo AI servers with cnvrg.io is simple. With one click, you can add more installed hardware to the managed

infrastructure

Accelerate Development, and Deploy ML to Production at Scale

The cnvrg.io platform, running on Lenovo ThinkSystem AI-Ready servers provides data scientists with accelerated project execution from research, to training, to production. cnvrg.io enables enterprises to easily manage and scale ML in any environment. Data scientists can run ML pipelines on diverse workloads to achieve maximum performance and accelerate time to production.

Realize AI-Driven Results Fast with Out-of-the-Box Productivity
In one unified solution, data scientists can easily manage, experiment, track, version and deploy models in one click. It's designed to be agnostic, portable, and solves key MLOps challenges to help data scientists deliver more models to production quickly.

Optimize Server Utilization with Advanced Resource Management

To scale AI, you need to integrate data science workflow within an IT/DevOps approach. The MLOps Dashboard improves visibility of workloads so you can maximize Lenovo server utilization. cnvrg.io MLOps helps streamline AI application delivery, so data science teams and IT can more effectively manage users, workloads, models, datasets, experiments, and more, while speeding continuous application delivery.
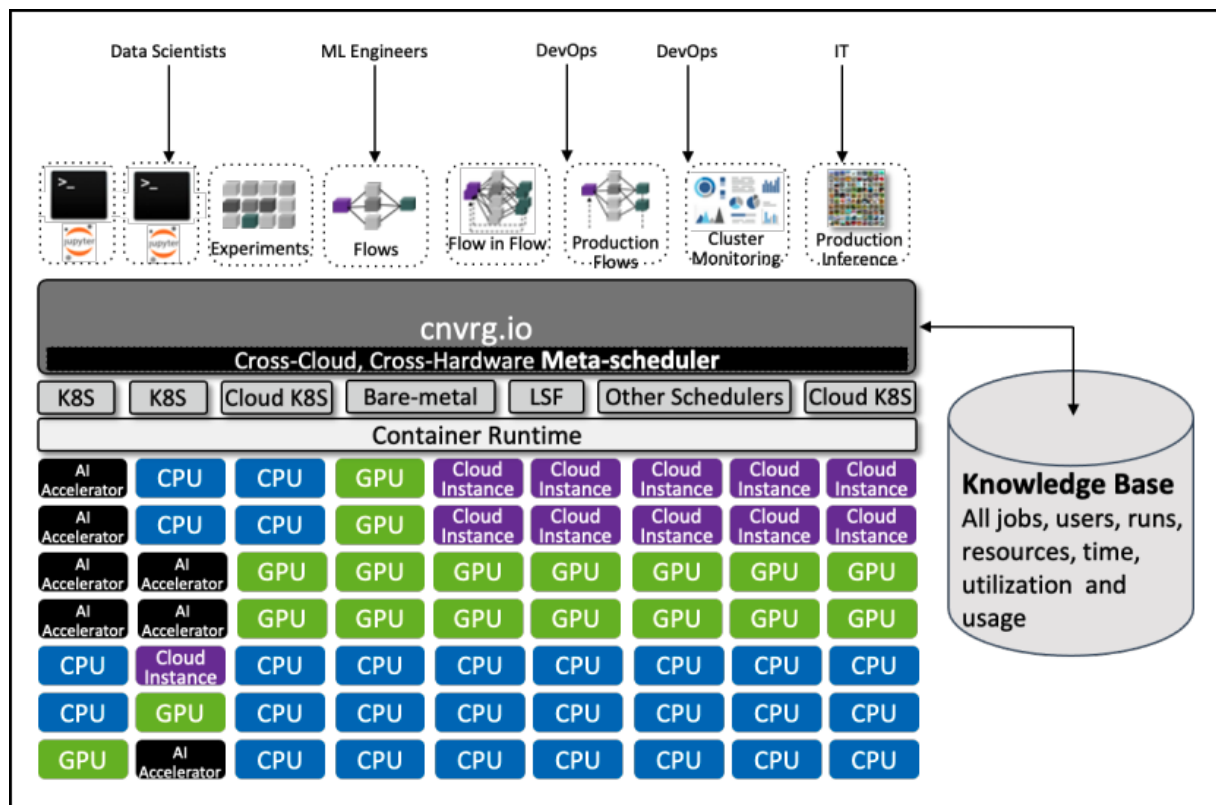

Figure 2. Cnvrg.io

Lenovo and cnvrg.io - Leading Enterprise AI

A focus on Productivity and User Experience

The cnvrg.io platform, running on Lenovo ThinkSystem AI-Ready servers provides data scientists with a unified environment to build, deploy and manage machine learning workloads, develop and innovate new models, or deploy and monitor models in on top of the Lenovo AI purpose built solutions and products. cnvrg.io supports data scientists at every stage of the AI lifecycle, delivering solutions to enhance research production.

cnvrg.io and Lenovo integrated solution also provides a single place for IT engineers, DevOps engineers, ML engineers, data scientists and researchers to collaborate, share and achieve AI-driven results. With cnvrg.io and Lenovo, data scientists can build dynamic end-to-end ML/DL solutions on heterogeneous compute by running different tasks on maximized CPUs and GPUs.

- cnvrg.io offers data scientists execution of a complete pipeline on Lenovo ThinkSystem AI-Ready servers without any need to understand the infrastructure, DevOps semantics or any resource conflicts or dependencies.

- cnvrg.io was designed by data scientists for data scientists. All pipeline stages from datasets versioning and management, through research, experimentation and deployment are offered as either Python SDK, CLI, REST APIs or GUI. The use cases range from ML CI/CD integration, to a single organization cockpit for ML infrastructure and individual data scientist development environment.

- Industry known tools and GUI are used for each stage of the pipeline, optimizing the data scientists' experience and eliminating the learning curve for new tools/GUI. Be it Jupyter Notebooks for experimentation, TensorBoard for visualization, Grafana for monitoring – cnvrg.io will have it running out-of-the-box on your Lenovo system. Need additional tools? cnvrg.io offers an open platform to integrate and launch any ML or DL tools**.**

- As AI teams scale their AI infrastructure, DevOps hand-holding and manual coordination of resources becomes impossible to manage. Moving to larger clusters requires the cnvrg.io platform to deliver large scale coordination and server consumption/scheduling at scale.
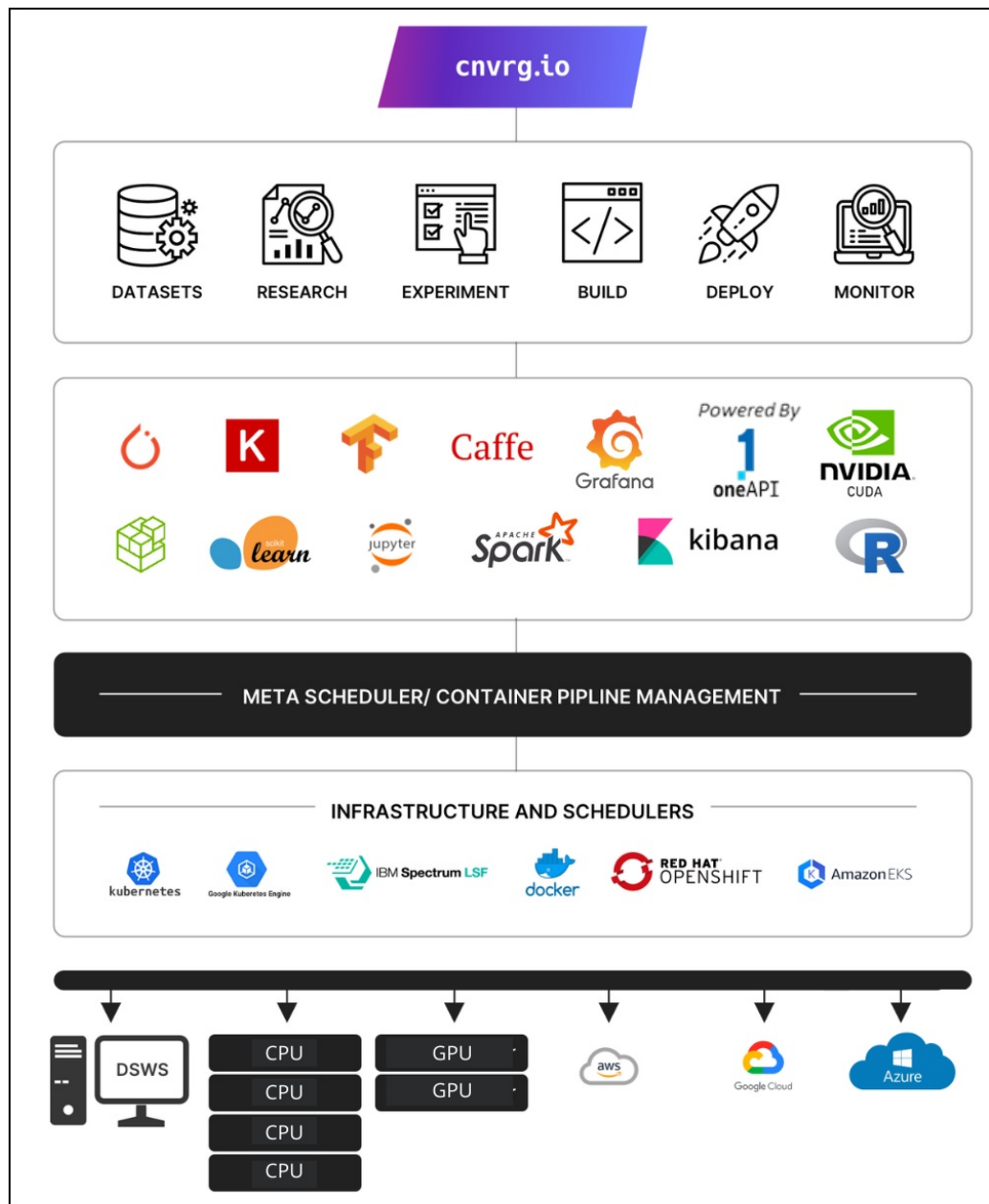
Figure 3. Architecture

cnvrg.io Architecture for Lenovo systems

While cnvrg.io offers one-click development and deployment of ML pipelines to Lenovo AI-Ready Servers, it can also auto-detect lack of resources, and burst your workload to the cloud of your choice.

Unlike other platforms, each stage of the pipeline can be attached to a different resource, including the support of multi-cluster and heterogeneous compute pipelines. For example: Query and process large datasets in a CPU cluster, while running training experiments in various GPU based accelerator systems, deploy in the cloud and iterate, or any different combination/resource attached to each pipeline step.

cnvrg.io will support the environment of your choice, be it Kubernetes, bare-metal or any hypervisor (for example, KVM). All can co-exist with-in your Lenovo clusters.

Certified and Tested by the Lenovo AI Team

The cnvrg.io data science platform has been tested and optimized to run on Lenovo

ThinkSystem AI-Ready servers; both CPU based and GPU based platforms are supported. The integrated solution delivers a proven enterprise-grade ML lifecycle platform that increases data science productivity, accelerates AI workflows, and improves accessibility and utilization of AI infrastructure.

The Lenovo ThinkSystem SR650 AI-Ready server features 2 second-generation Intel Xeon Platinum processors and has been tested and certified with cnvrg.io featuring the Intel container repo integrated in cnvrg.io. This platform unlocks fullML/DL performance of Intel Xeon processor performance and provides a cost-effective solution for mixed workload environments. This is the ideal platform for Data Science teams that are working on traditional ML workloads, CPU based workloads, memory-bound workloads or mixed GPU+CPU workloads.

Figure 4. Lenovo ThinkSystem SR650

The Lenovo ThinkSystem SR670 AI-Ready server adds support for up to4 double-wide GPUs for extreme accelerated computing performance for your most demanding deep learning workloads or up to 8 single-wide NVIDIA GPUs for an extremely powerful DL Inference platform. With cnvrg.io's built in GPU optimized containers it is ideal for data parallel or deep learning dominant workloads where optimizing for max performance is the goal.

Figure 5. Lenovo ThinkSystem SR670

Start Today

cnvrg.io together with Lenovo GPU and/or CPU servers are available as a starter kit allowing a quick test drive for customers. It comes equipped with the hardware and software to jumpstart your AI, and example ML pipelines for customers to quickly start with NLP, computer vision and classical ML models. Easily unlock optimal performance of your servers for any ML or DL framework with Intel oneContainer integration and NVIDIA GPU Cloud containers.

Contact us at sales@lenovo.com or info@cnvrg.io.

About Lenovo Servers

Drive your most complex AI projects with ease thanks to the uncompromised performance, legendary reliability, and scalability of Lenovo from the core to the edge. Engineered to not just meet, but exceed the rigorous performance requirements of today's demanding AI, machine learning, and deep learning workflows, Lenovo ThinkSystem AI-Ready servers are equipped to suit your growing Artificial Intelligence needs.

About cnvrg.io

cnvrg.io is an AI OS, transforming the way enterprises manage, scale and accelerate AI and data science development from research to production. The code-first platform is built by data scientists, for data scientists and offers unrivaled flexibility to run on-premise or cloud. From advanced MLOps to continual learning, cnvrg.io brings top of the line technology to data science teams so they can spend less time on DevOps and focus on the real magic - algorithms. Since using cnvrg.io, teams across industries have gotten more models to production resulting in increased business value.

Related product families

Product families related to this document are the following:

- Artificial Intelligence
- ThinkSystem SR650 Server
- ThinkSystem SR670 Server

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, LP1455, was created or updated on March 20, 2021.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
https://lenovopress.lenovo.com/LP1455

- Send your comments in an e-mail to:
comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1455.

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®
ThinkSystem®

The following terms are trademarks of other companies:

Intel® and Xeon® are trademarks of Intel Corporation or its subsidiaries.

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