SOLUTION BRIEF

Speed data science initiatives with NetApp AI and Run:AI

Simplify orchestration of AI workloads to run more experiments in less time

NetApp | run:ai

Faster AI experimentation with full GPU utilization
Speed up AI by running many experiments in parallel, with fast access to data, utilizing limitless compute resources. Run:AI enables full GPU utilization by automating resource allocation, and NetApp® ONTAP® AI proven architecture allows every experiment to run at maximum speed by eliminating data pipeline bottlenecks. Together, companies scaling AI with NetApp and Run:AI technology see a double benefit: faster experiments on top of full resource utilization.

Challenge

Speed is critically important in AI; fast experimentation and successful business outcomes of AI are directly correlated. And yet AI projects are rife with inefficient processes. The combination of data processing time and outdated storage solutions creates bottlenecks. In addition, workload orchestration issues and static allocation of GPU compute resources limit the number of experiments that researchers can run.

Key features

AI data center cluster management
Automated job queuing, orchestration, and resource allocations optimize GPU sharing and maximize cluster utilization.

Access to centralized data for many workloads running in parallel
Experiments run in parallel on different compute nodes with fast access to datasets that are on shared storage.

Control and visibility
Out-of-the-box tools help you set up policies, priorities, and quotas around resource allocations. You get a multicluster holistic view of how resources are consumed and jobs are orchestrated, and you can monitor metrics such as GPU utilization, workload run time, and wait times.

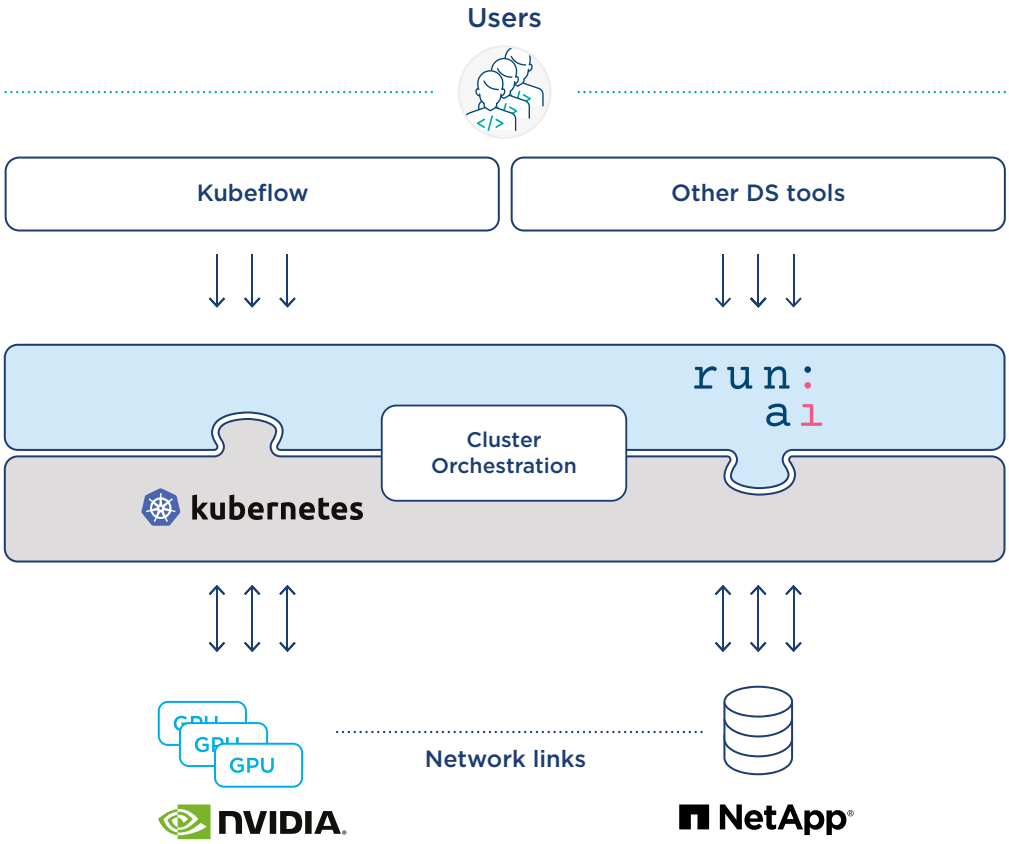
Solution

NetApp and Run:AI have partnered to simplify the orchestration of AI workloads, streamlining the process of both data pipelines and machine scheduling for deep learning (DL). You can fully realize the promise of AI and DL by simplifying, accelerating, and integrating your data pipeline with the NetApp ONTAP AI proven architecture. Run:AI's orchestration of AI workloads adds a proprietary Kubernetes-based scheduling and resource utilization platform to help researchers manage and optimize GPU utilization. Together, the products enable numerous experiments to run in parallel on different compute nodes, with fast access to many datasets on centralized storage.

Benefits of the solution include:

- **Faster time to innovation.** By using Run:AI's centralized resource pooling, queueing, and prioritization mechanisms together with the NetApp storage system, researchers are removed from infrastructure management hassles and can focus exclusively on data science. With Run:AI and NetApp technology, you can increase productivity by running as many workloads as you need without compute or data pipeline bottlenecks.
- **Increased productivity.** Run:AI's fairness algorithms guarantee that all users and teams get their fair share of resources. For example, you can preset policies for prioritization. You can also allocate resources dynamically from one user or team to another so that everyone gets timely access to coveted GPU resources.
- **Improved GPU utilization.** With the Run:AI Scheduler and virtualization technology, you can easily use fractional GPUs, integer GPUs, and multiple nodes of GPUs for distributed training on Kubernetes. In this way, AI workloads run based on need, not capacity. Data science teams can run more AI experiments on the same infrastructure.

If you're looking to implement these solutions, reach out to both NetApp and Run:AI for more information.



NetApp and Run:AI data science solution.

About Run:AI

Run:AI has built the world's first orchestration and virtualization platform for AI infrastructure. By abstracting workloads from underlying hardware, Run:AI creates a shared pool of GPU resources that can be dynamically provisioned, enabling efficient orchestration of AI workloads and optimized utilization of GPUs. Learn more at www.run.ai.

About NetApp

In a world full of generalists, NetApp is a specialist. We're focused on one thing, helping your business get the most out of your data. NetApp brings the enterprise-grade data services you rely on into the cloud, and the simple flexibility of cloud into the data center. Our industry-leading solutions work across diverse customer environments and the world's biggest public clouds.

As a cloud-led, data-centric software company, only NetApp can help build your unique data fabric, simplify and connect your cloud, and securely deliver the right data, services and applications to the right people—anytime, anywhere.

NetApp