Run:ai on OCI: Cloud-native approach to maximizing GPU utilization and accelerating AI workloads

May 30, 2024

Kubernetes has become the standard platform for automating deployment, scaling, and management of modern containerized applications. Data scientists and machine learning (ML) engineers have also gravitated to Kubernetes since its infrastructure abstraction capabilities enable them to better focus on their workloads. An increasing number of data science platforms are today built on top of containers and plugged into Kubernetes.

GPU-based compute infrastructure that data science and ML workloads run on tend to be quite expensive, so there's strong motivation to maximize their utilization. The ideal scenario is one where available GPU resources are pooled together and managed by a scheduling system that provides multiple queues with different priorities, through which jobs are submitted. Users are assigned quotas to ensure that everyone gets a fair share. Also, when resources are idle,

jobs are allowed to use resources above their quota limits. When a user's job or higher priority request comes along, the scheduler preempts jobs as needed in accordance with a configurable policy. Gang scheduling is supported to ensure that all parts and pieces of a job get started together. When choosing GPU nodes to run a job, the scheduler considers the topology of interconnections between them to maximize performance. Detailed resource utilization easily provides accounting information for appropriate cost sharing.

Implementing such an environment is relatively straightforward using schedulers, such as Slurm and LSF, which are predominantly used in traditional CPU-based high-performance computing (HPC) setups. But while Kubernetes provides first class support for orchestrating containerized workloads, its scheduling mechanism lacks the functionality to implement such an environment to maximize GPU utilization.

The Run:ai solution

Run:ai is a cloud native AI orchestration platform that simplifies and accelerates AI and ML through dynamic resource allocation, comprehensive AI lifecycle support, and strategic resource management. Pooling GPU resources from both public and private cloud infrastructure and utilizing Run:ai significantly enhances GPU efficiency and workload capacity. Its policy engine, open architecture, and deep visibility into AI workloads foster strategic alignment with business objectives, enabling seamless integration with external tools and systems. This combination results in significant increases in GPU availability, workloads, and GPU utilization, all with zero manual resource intervention, accelerating innovation and providing a scalable, agile, and cost-effective solution for enterprises. Run:ai supports most popular versions of Kubernetes.

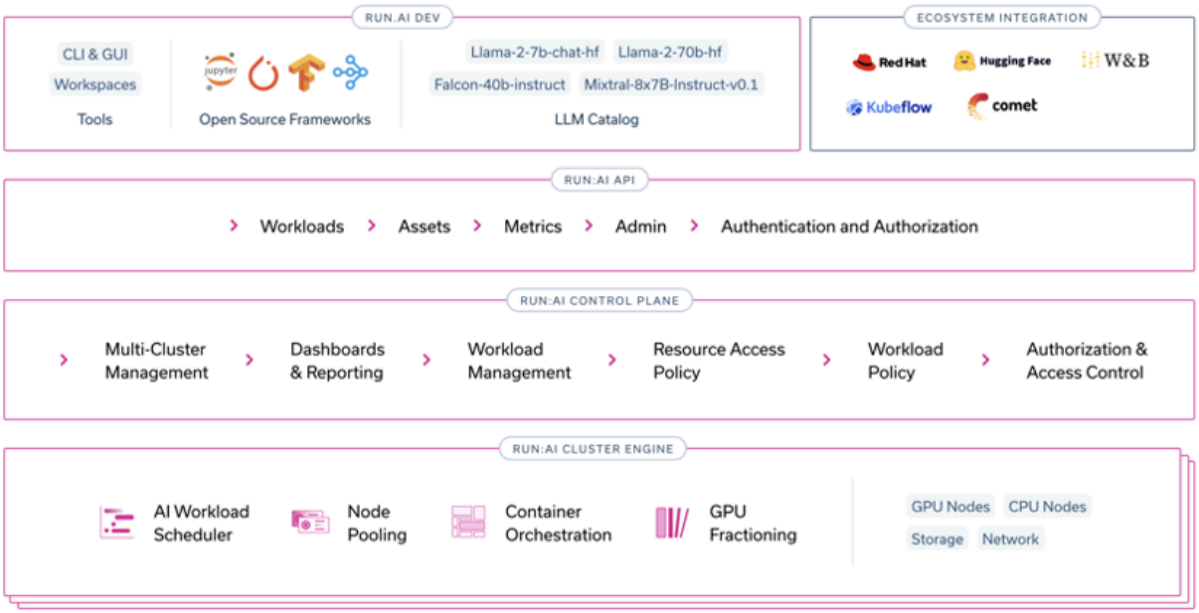

Figure 1: Run:ai Overview

In a joint exercise, Oracle and Run:ai engineers set up and tested the Run:ai solution in the Oracle Cloud Infrastructure (OCI) environment. This article highlights some of the key features tested during this activity.

Projects

Run:ai uses projects to streamline resource allocation and prioritize work. An administrator creates a project in the user-interface and assigns researchers to it. They also specify the number of GPUs, CPUs, CPU memory allowed for use. Policies regarding allowing “over quota” usage and scheduling rules are also configured in a project.

Internally, Run:ai projects are implemented as Kubernetes namespaces. A researcher submitting a job must associate a project name to the request for the Run:ai scheduler to determine eligibility and allocate resources.

Workspaces

A workspace is a simplified tool for researchers to conveniently conduct experiments, build AI models, access standard MLOps tools and collaborate with their peers. A workspace consists of all the setup and configuration needed for research in a single space, including container images, data sets, resource requests, and all required tools for the research. With facilitating

research, workspaces ensure control and efficiency while supporting various needs. An active workspace is a Run:ai interactive workload.

The following figures show connecting to an active workspace to invoke a Jupyter session.

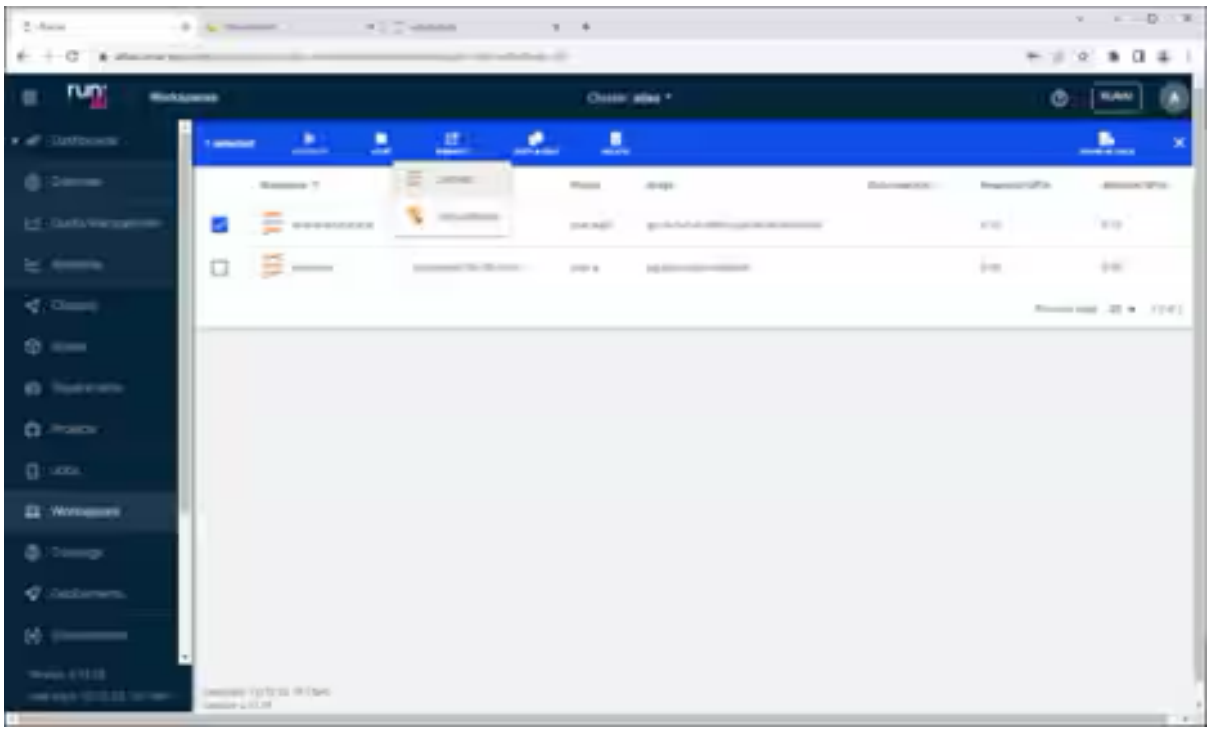

Figure 2: Running a Jupyter notebook in a workspace

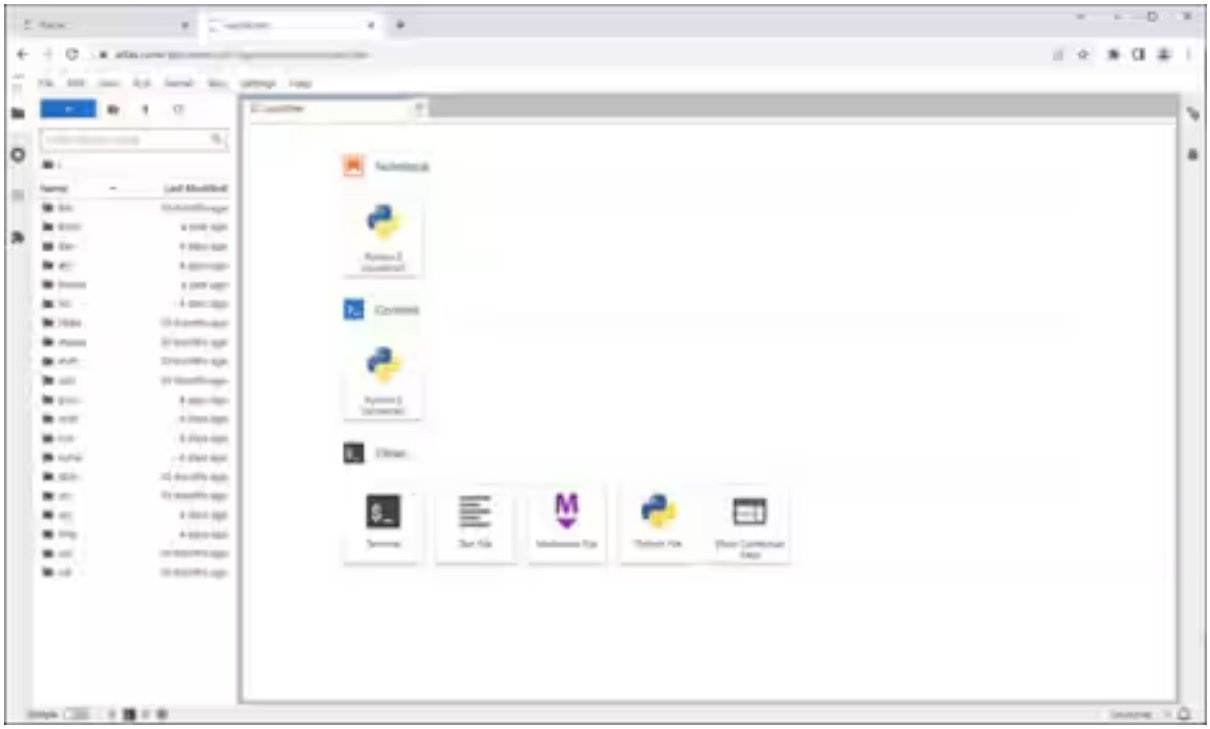

Figure 3: Running a Jupyter notebook with options in a workspace

Launching workloads

The engineers submitted various kinds of workloads through the CLI and tested them successfully, including unattended training, interactive build (including exposing internal ports to user), hyperparameter optimization (HPO) and inference workloads. They also tested overquota allocation, bin packing, and preemption behaviors.

GPU fractions

Run:ai provides the capability to allocate a container with a specific amount of GPU RAM specified at the time of submitting the job. Attempting to reach beyond the allotted RAM results in an out-of-memory exception. This configuration enables multiple containers to simultaneously run on a single GPU, each utilizing its own specified fraction of the total GPU memory. All workloads running on a single GPU share the compute capabilities in parallel and on average get an even share of the compute. Specifically, the GPU fractions feature does not require a GPU that supports NVIDIA multi instance GPU (MIG) technology.

In the following tests, two jobs are run simultaneously on a single GPU each taking 50% and 20% of available GPU memory.

```
ubuntu@master-instance-20231011-1622:~$ runai submit frac05-3 -i gcr.io/run-ai-demo/quickstart
-g 0.5 --interactive

Job frac05-3 submitted successfully.

You can check the status of the job by running:

        runai describe job frac05-3 -p user-agm

ubuntu@master-instance-20231011-1622:~$ runai bash frac05-3

root@frac05-3-0-0:/workload#

root@frac05-3-0-0:/workload# nvidia-smi

Wed Dec 13 18:22:43 2023

+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 545.23.06              Driver Version: 545.23.06    CUDA Version: 12.3     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  NVIDIA A10                     On  | 00000000:00:05.0 Off |                    0 |
|  0%   39C    P0             86W / 150W |  10960MiB / 11514MiB |     14%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
+---------------------------------------------------------------------------------------+

root@frac05-3-0-0:/workload#
```

```
ubuntu@master-instance-20231011-1622:~$ runai submit frac03-3 -i gcr.io/run-ai-demo/quickstart
-g 0.3

Job frac03-3 submitted successfully.

You can check the status of the job by running:

        runai describe job frac03-3 -p user-agm

ubuntu@master-instance-20231011-1622:~$

ubuntu@master-instance-20231011-1622:~$

ubuntu@master-instance-20231011-1622:~$ runai bash frac03-3

root@frac03-3-0-0:/workload#

root@frac03-3-0-0:/workload#

root@frac03-3-0-0:/workload# nvidia-smi

Wed Dec 13 18:25:10 2023

+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 545.23.06              Driver Version: 545.23.06    CUDA Version: 12.3     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  NVIDIA A10                     On  | 00000000:00:05.0 Off |                    0 |
|  0%   49C    P0              88W / 150W |   6676MiB /  6908MiB |     10%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
+---------------------------------------------------------------------------------------+

root@frac03-3-0-0:/workload#
```

```
ubuntu@master-instance-20231011-1622:~$ runai list jobs

Showing jobs for project user-agm

NAME             STATUS     AGE  NODE                            IMAGE
TYPE         PROJECT   USER                    GPUs Allocated (Requested)  PODs Running
(Pending)  SERVICE URL(S)

anand-workspace  Suspended  4d                                   gcr.io/run-ai-demo/jupyter-
tensorboard  Interactive  user-agm  anand.manian@oracle.com  0.00 (0.10)                0 (0)

frac03           Failed     12h  -                               gcr.io/run-ai-demo/quickstart
Train        user-agm  ubuntu                  - (0.30)                   0 (0)

frac03-3         Running    2m   worker-instance-20231012-1723  gcr.io/run-ai-demo/quickstart
Train        user-agm  ubuntu                  0.30 (0.30)                1 (0)

frac05           Failed     12h  -                               gcr.io/run-ai-demo/quickstart
Interactive  user-agm  ubuntu                  - (0.50)                   0 (0)

frac05-3         Running    4m   worker-instance-20231012-1723  gcr.io/run-ai-demo/quickstart
Interactive  user-agm  ubuntu                  0.50 (0.50)                1 (0)

ubuntu@master-instance-20231011-1622:~$

ubuntu@master-instance-20231011-1622:~$ runai delete job frac03-3

Job frac03-3 deleted successfully.

ubuntu@master-instance-20231011-1622:~$ runai delete job frac05-3

Job frac05-3 deleted successfully.

ubuntu@master-instance-20231011-1622:~$ runai delete job frac03

Job frac03 deleted successfully.

ubuntu@master-instance-20231011-1622:~$ runai delete job frac05

Job frac05 deleted successfully.

ubuntu@master-instance-20231011-1622:~$


ubuntu@master-instance-20231011-1622:~$ runai list jobs

Showing jobs for project user-agm

NAME             STATUS     AGE  NODE  IMAGE                                    TYPE
PROJECT   USER                    GPUs Allocated (Requested)  PODs Running (Pending)  SERVICE
URL(S)

anand-workspace  Suspended  4d        gcr.io/run-ai-demo/jupyter-tensorboard  Interactive
user-agm  anand.manian@oracle.com  0.00 (0.10)               0 (0)

ubuntu@master-instance-20231011-1622:~$
```

Dynamic MIG

Run:ai also provides a way to dynamically create a MIG partition NVIDIA GPUs that support the technology, such as NVIDIA A100. This process happens spontaneously, without needing to drain workloads or involve an IT administrator. The partitions are automatically deallocated when the workload finishes. The partition isn't removed until the scheduler determines that it's needed elsewhere.

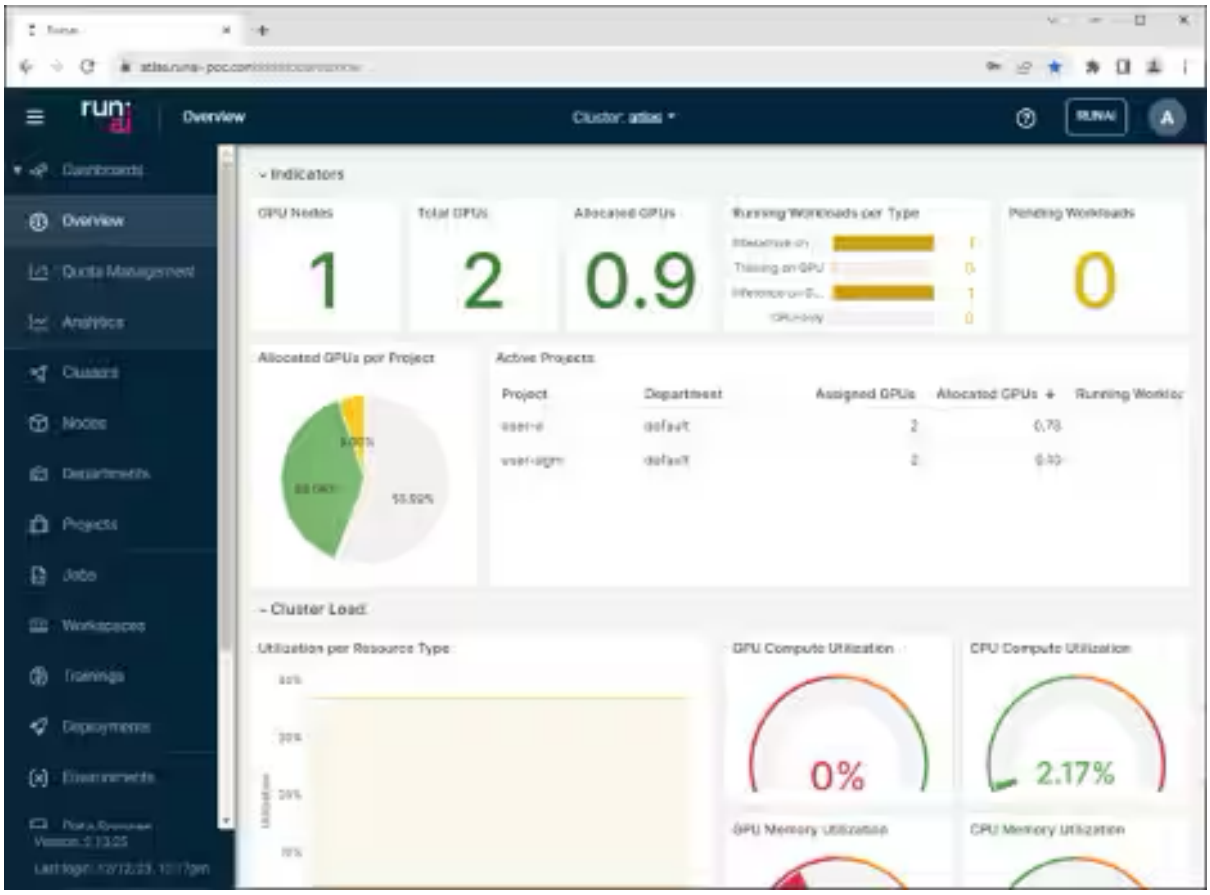

Figure 4: Run:ai user interface

Conclusion

The Run:ai solution offers more than just an improvement in GPU utilization. It's a transformative asset for enterprises aiming to lead in the competitive landscape of AI-driven industries. By optimizing resource allocation, Run:ai not only enhances the operational efficiency of Kubernetes environments, but also significantly accelerates the pace of AI innovation. This solution can directly translate into faster time-to-market for new AI solutions and improvements in cost efficiency. For organizations looking to scale their AI capabilities without corresponding increases in operational complexity or costs, Run:ai offers a compelling solution. We invite you to explore how Run:ai can tailor its capabilities to your specific business needs.

By integrating Run:ai's cloud native approach, enterprises can ensure that their AI and ML workloads are not just running, but running smarter, faster, and more cost-effectively.

Discover firsthand how Run:ai can revolutionize your AI projects by scheduling a personalized demo. For detailed tech briefs, case studies, and webinars that delve into the capabilities and successes of Run:ai solution and services, visit our resources page. If you're ready to discuss your organization's needs, visit our site today to get started on transforming your AI infrastructure.

For more information, see the following resources:

- KubeDL Gang Scheduling
- Slurm SchedMD documentation
- About IBM Spectrum LSF
- Run:ai Documentation Library Admin UI Setup
- Getting familiar with workspaces

About Cloudsway

Cloudsway is a subsidiary of Wangsu Science and Technology (stock code: 300017), established in March 2023. Wangsu Science and Technology is a global leading provider of information infrastructure platform services, with business spread across more than 70 countries and regions worldwide.

Cloudsway is one of the three innovation engines in Wangsu's “2+3” strategy, providing enterprises with integrated products and solutions, such as cloud strategy consulting, modernized application construction, generative AI, and enterprise-grade cloud hosting services. solutions based on AWS.

Cloudsway is committed to become a leading provider of hybrid cloud solutions,offering secure, efficient, and convenient cloud services to enterprises, helping them with digital and intelligent transformation, and boosting their operational efficiency.