Hyperscalers Run:ai Appliance

Introduction document

Friday, 12 August 2022

INTRODUCTION

Modern organisations expect rapid, efficient, and accurate results from their investments into AI technology. The ability to stand up appropriate models quickly can be paramount in terms of delivering critical insights responsively across key business and research dimensions.

Counterproductive to these efforts are long AI software development times and/or time lost with respect to AI infrastructure implementation, configuration and management.

Recent evolution in the world of AI algorithm commoditisation has seen the establishment of the Nvidia NGC Catalog. NGC is a library of commonly needed AI algorithms that have been implemented into a common container runtime format.

The Hyperscalers Run:ai Appliance offers a fully integrated hardware and software solution that supports deployment of common container-based AI applications designed for researchers, academics and business users.

AI infrastructure hardware is an expensive resource, yet most AI workload scheduling environments are not able to utilise all GPU resources optimally - meaning that unfortunately, AI hardware infrastructure must often be significantly over-provisioned.

The Hyperscalers Run:ai appliance stands out against this backdrop due to its ability to perform fine-grained sub-allocation of GPU resources between multiple simultaneous workloads. In addition, the Hyperscalers Run:ai Appliance can dynamically re-allocate GPU resources against any workload while it is being processed. This may mean actively releasing GPU resources if they are no longer required or adding GPU resources whenever they become needed.

The dynamic partitioning capability of Run:ai moves beyond traditional methods in which GPU profile size must be pre-configured to a pre-set fixed value in advance. This older method often fails to match the profile sizing requirements of workloads whenever they are eventually scheduled, further reducing the ability to fully utilise GPU resources.

Inclusion of fine-grained, dynamic sub- allocation capabilities within the Hyperscalers Run:ai Appliance is an industry leading approach that automatically optimises utlisation of your valuable AI infrastructure. Coupled additionally with the powerful deployment context of NGC containers (and/or Helm charts), the Hyperscalers Run:ai Appliance is truly a game-changer in the field of agile AI platform enablement.

The Hyperscalers Run:ai Appliance is a fully pre-integrated solution including hardware, software and support services that can be implemented on a small, medium or large basis as per your initial needs and scaled up later to whatever size may be required.

This kind of power puts an unprecedented level of AI agility and efficiency within the reach of any organisation.

The Hyperscalers Run:ai Appliance management console allows users without deep infrastructure expertise to schedule, monitor and manage their own workloads. This is truly disruptive in terms of placing NGC model execution support generally across your organisation atop the most efficient AI infrastructure environment possible.

Hyperscalers [1] in partnership with Run:ai is offering the HyperScalers Run:ai Appliance as a fully integrated and qualified solution to quickly, reliably and efficiently support the needs of AI users, using hardware supplied by Hyperscalers and software stack from Run:ai.

The following diagram illustrates the high-level relationship between NVIDIA NGC, 3rd Party MLOps and Model Serving tools, the Run:ai built-in tools and workflows and additionally its ability to operate within the context of numerous well-known Kubernetes implementations using qualified, reliable state of the art equipment by Hyperscalers:

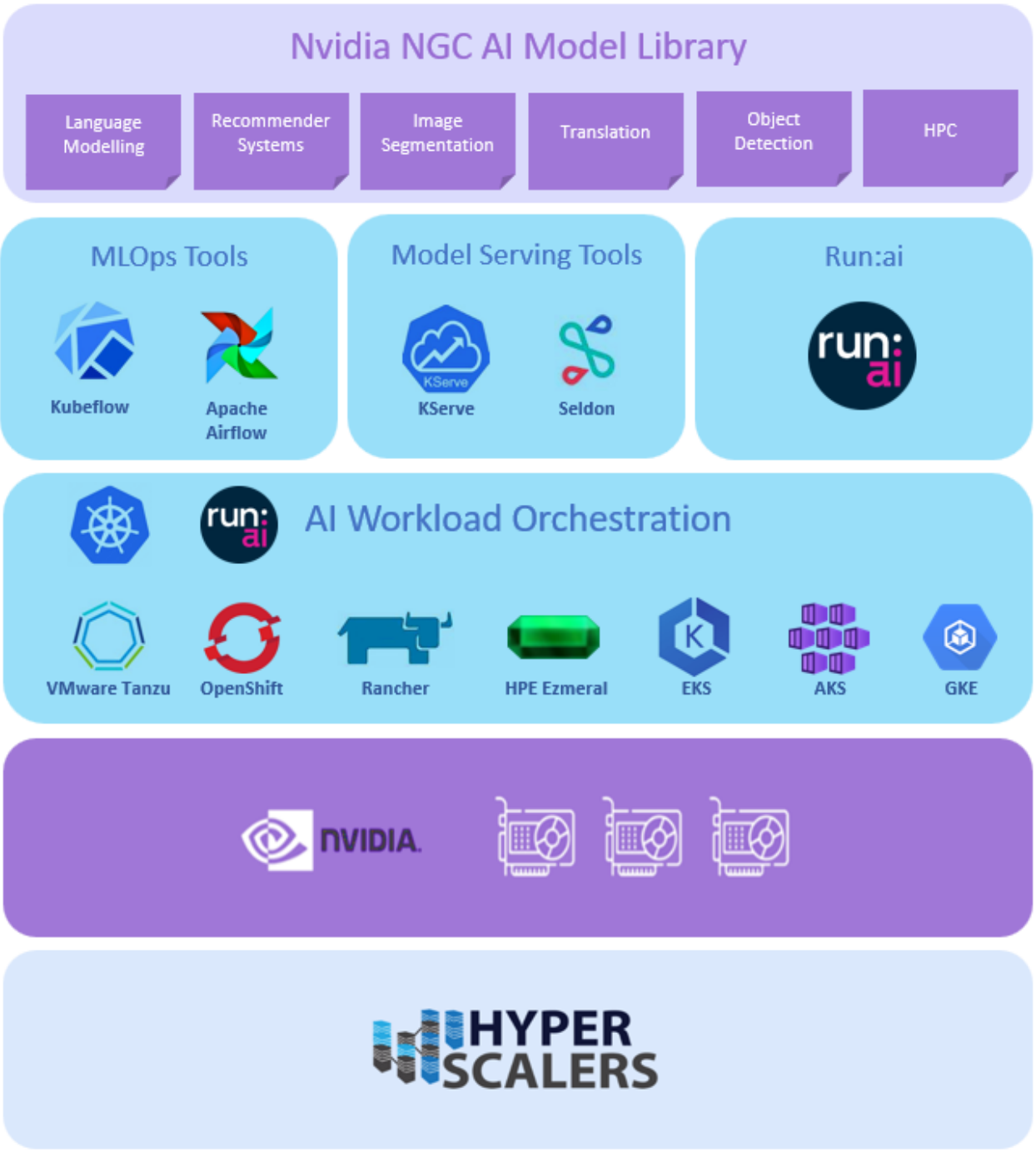

Figure 1 Hyperscalers Run:ai Appliance stack

Featured Hardware from Hyperscalers

Hyperscalers have identified the hardware configurations that can be categorised based on the customer use cases as the below. Run:ai enterprise appliance can be deployed in any of these high performance ultra-dense GPU servers.

1. Cost efficient (Up to 4 x PCIe 80GB GPUs)
2. Balanced performance (Up to 4 x PCIe 80GB GPUs)
3. High Performance Compute (Up to 8 x SXM 80GB GPUs)

High Performance Compute (Up to 8 x SXM 80GB GPUs)

Balanced performance solution← (Up to 4 x PCIe 80GB GPUs)→ Cost efficient solution

Hyperscalers Run:ai Appliance

Assign the Right Amount of AI Compute Power to Users, Automatically

The Hyperscalers Run:ai Appliance is a Kubernetes-based software platform for orchestration of containerized AI workloads that enables GPU clusters to be utilized for different Deep Learning workloads dynamically - from building AI models, to training, to inference. With Run:ai, jobs at any stage can obtain access to the compute power they need, automatically [7].

Run:ai's compute management platform speeds up data science initiatives by pooling available resources and then dynamically allocating resources optimally as needed. These powerful capabilities maximise AI compute power utilisation and therefore return on investment for your organisation.

Key Features

- Fair-share scheduling to allow users to share clusters of GPUs easily and automatically
- Fractional GPU allocation for interactive/ training workloads
- Simplified workflows for building, training (including multi-GPU and distributed training) and deployment of AI models
- Visibility into workloads and resource utilization to improve user productivity
- Control for cluster admin and ops teams, to align priorities to business goals
- On-demand access to Multi-Instance GPU (MIG) instances for the A100 GPU

Key Benefits

Advanced Kubernetes-based Scheduling Eliminates Static GPU Allocation

The *Run:ai Scheduler* manages tasks in batches using multiple queues on top of Kubernetes, allowing system admins to define different rules, policies, and requirements for each queue based on business priorities. Combined with an over-quota system and configurable fairness policies, the allocation of resources can be automated and optimized to allow maximum utilization of cluster resources.

Because it was built as a plug-in to K8s, Run:ai's scheduler requires no advanced setup, and is certified to integrate with any number of Kubernetes "flavors" including Red Hat OpenShift and HPE Ezmeral.

The following diagram illustrates key architecture and functional elements of the fully integrated Hyperscalers Run:ai Appliance:

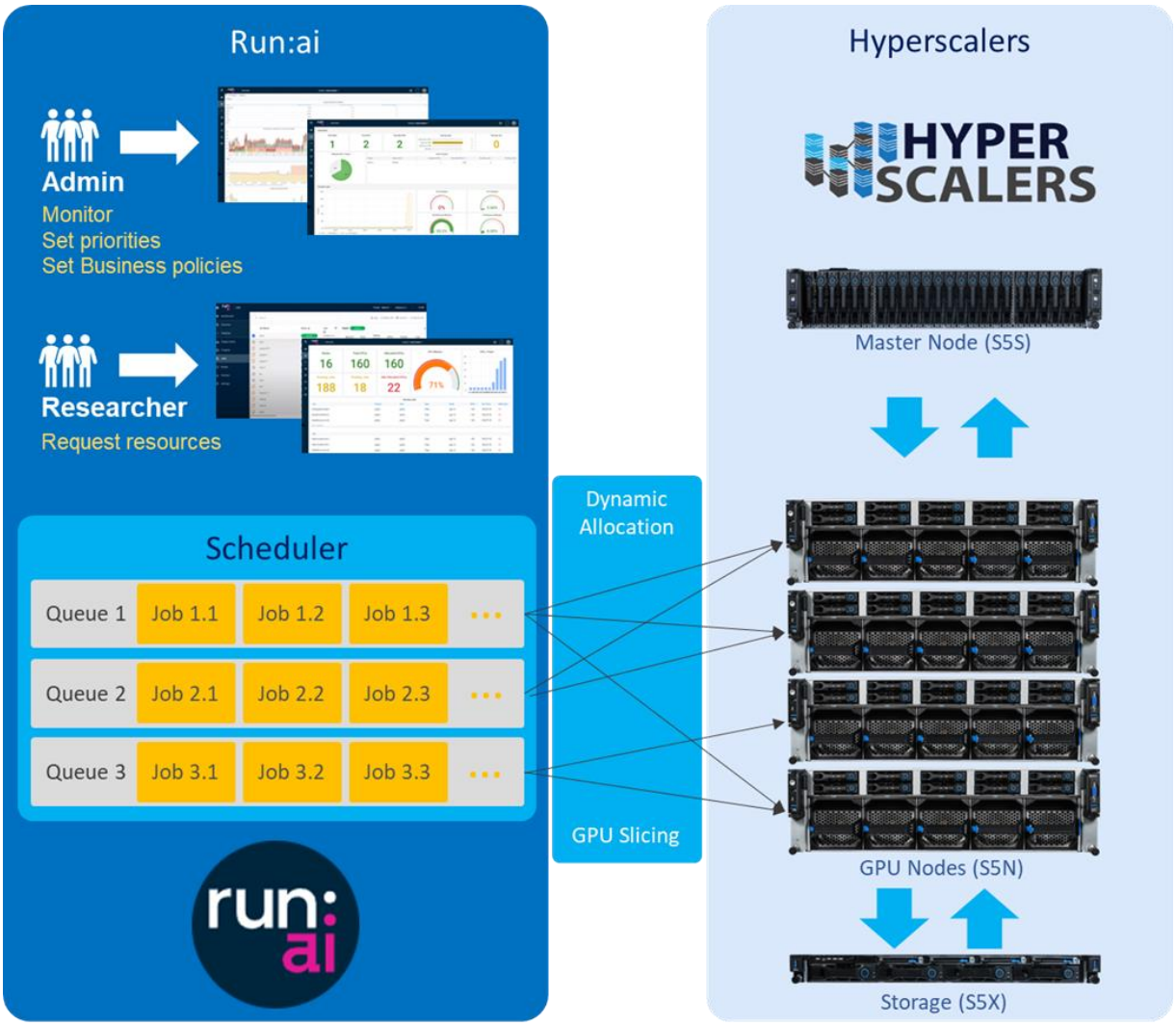

Figure 2 Hyperscalers Run:ai Appliance Workflow Architecture

The following illustration shows Run:ai management dashboard monitoring UI capabilities supporting both high-level overview and detailed observation of AI workload status and resource utilisation:

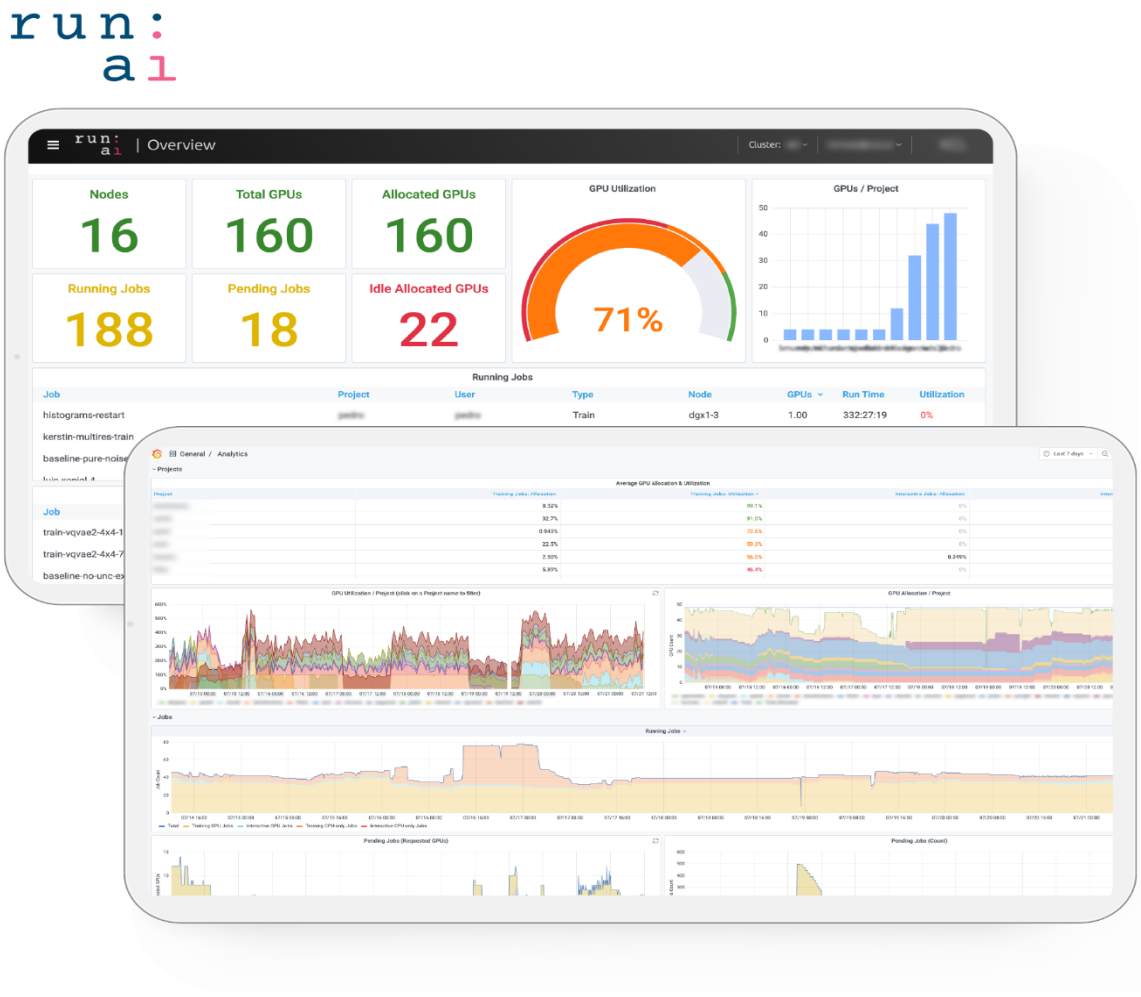

Figure 3 Overview of the Hyperscalers Run:ai Management Dashboard

No More Idle Resources

Run:ai's over-quota system allows users to automatically access idle resources when available based on configurable fairness policies. The platform allocates resources dynamically, for full utilization of cluster resources. Our customers see improvements in utilization from around 25% from when we start working with them to over 75% once more fully optimised.

Bridge Between HPC and AI

The Run:ai Scheduler allows users to easily make use of integer GPUs, multi-node/multi GPUs, and even GPU Multi-Instance GPU (MIG) instances for distributed training on Kubernetes. In this way, AI workloads run based on needs, not available capacity. Run:ai empowers you to combine the benefits and efficiency of High-Performance Computing with the simplicity of Kubernetes.

Accelerate AI

By using Run:ai resource pooling, queueing, and prioritization mechanisms, researchers are shielded from infrastructure management hassles and can focus exclusively on data science. Many workloads can be run in parallel without compute bottlenecks. Run:ai delivers real time and historical views on all resources managed by the platform, such as jobs, deployments, projects, users, GPUs and clusters.

The Hyperscalers Run:ai Appliance can accelerate your time to reach productive AI results, in particular as it is delivered as a turn-key solution including all master, worker and storage nodes pre-configured and ready to start running your AI workloads.

Streamline AI

Run:ai can support all types of workloads required within the AI lifecycle (build, train, inference) to easily start experiments, run large-scale training jobs and take AI models to production without ever worrying about the underlying infrastructure. The Run:ai Atlas platform allows MLOps and AI Engineering teams to quickly operationalize AI pipelines at scale and run production machine learning models anywhere while using the built-in ML toolset or simply integrating their existing 3rd party toolset.

Productize AI

Run:ai's unique GPU Abstraction capabilities effectively "virtualize" all available GPU resources to maximize infrastructure efficiency and increase ROI. The platform pools expensive compute resources and makes them accessible to researchers on-demand for a simplified, cloud-like experience.

A distinction should be made between the capabilities of Run:ai and other GPU virtualisation products that do not support suballocation/percentage-based splitting of GPUs across multiple AI workloads, or dynamic reclamation and re-assignment of GPU resources on the fly during AI workload job progress.

Run:ai dynamic resource allocation capabilities prevent GPU resources from becoming unusable for other new and ongoing workloads even though the original workload for which they were allocated no longer needs them. (Run:ai can disable dynamic allocation if required for specific workloads and fall back to static allocation).

Who we are?

Hyperscalers [1] is the world's first open supply chain Original Equipment Manufacturer- OEM, solving Information Technology challenges through standardization of best practices and hyperscale inspired practices and efficiencies. Hyperscalers offers choice across two open hardware architectures:

- Hyperscale - high efficiency open compute equipment as used by macro service providers
- Tier 1 Original – conventional equipment as per established Tier 1 OEM suppliers.

Each architecture is complete with network, compute, storage, and converged GP GPU infrastructure elements, and is open / free from vendor lock-in.

Hyperscalers' appliance solutions are packaged complete with hardware, software and pre-built (customisable) configurations. These were all pre-engineered using an in-house IP Appliance Design Process and validated in partnership with associated major software manufacturers. Many can be "test-driven" using Hyperscalers Lab as a Service (LaaS). Hyperscalers appliance solutions are ideally suited to IaaS, PaaS and SaaS providers looking to implement their services from anywhere.

Run:ai [8] helps organizations accelerate their AI journey - from fast entry into building initial models to scaling AI in production. Using Run:ai's Atlas software platform, companies streamline the development, management and scaling of AI applications. Researchers gain on-demand access to pooled resources for any AI workload. An innovative operating-system supports management of AI equipment resources beginning from fractions of GPUs up and into large-scale distributed training.