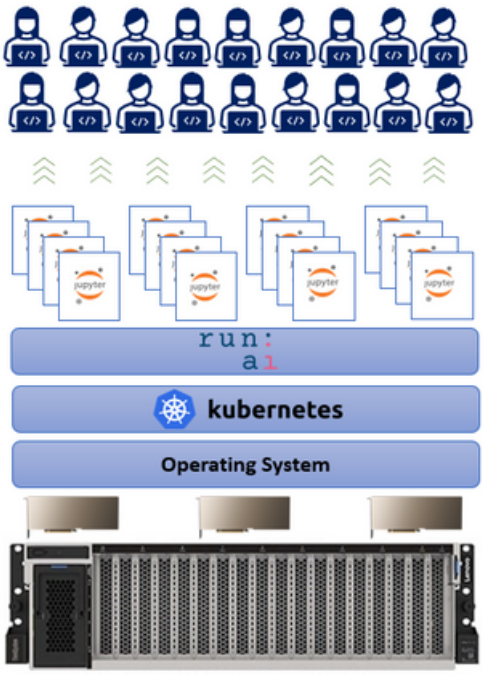
SERVER BUILT FOR AI COURSES



Have you got a problem of a greater number of students but limited GPU's?

AI/DL/ML

GPU ALLOCATION

MEMORY MANAGEMENT

FRACTALISING GPU

The most common discussion point we come across higher education are that GPU's are expensive and it's difficult to prove value for money especially in teaching universities other than research projects. Teaching AI courses with limited compute resources is though job, mainly when the purpose of the course is to learn how quickly AI/ deep learning jobs are executed in a GPU. Many universities are using CPUs or low powered cluster as an alternative to GPU. The real value of deep learning jobs are achieved when the students get to use the allocated GPU to run their deep learning jobs.

It is important to get the fundamentals right when it comes to sizing the server and choosing the software stack. OCF together with our niche partners have built a solution which can address the problem with limited GPU and excess student demand. The cost-effective solution will help to overcome the challenge by fractionalising the GPU's and automating the allocation of these resources. The fractional GPUs are assigned automatically to individual students or to a cohort based on demand, the solution ensures that no student overuses the allocated quota or will have to wait for others to finish jobs.

The solution comprising of Lenovo 2U server with 3*A100 80G with formidable I/O capabilities. Kubernetes containers are installed to make use of the latest technology in the AI advancements. Run:ai on the server acts as an intelligent orchestration layer to schedule jobs and to manage GPU memory by users and their privilege. Widely used software packages (such as Jupyter Notebook, Tensor flow, Pytorch and Keras) are readily installed. This plug and play AI server solutions are ready and optimised to be installed on your departmental AI/ ML courses.