Introduction to AI Cloud
Slurm and Singularity Training

March 2023

CLAAUDIA, Aalborg University

Outline

2

AI Cloud Background

System design

Fair usage

Where to go from here

AI Cloud Background

System design

Fair usage

Where to go from here

Background

4

AI Cloud is for GPU-accelerated computations.

▶ Typically training of artificial neural networks, but also any other computations that can utiilise GPUs
▶ For research purposes at AAU.
▶ Admit students based on confirmation from supervisor.
▶ Free to use.

Usage guide here: https://aicloud-docs.claaudia.aau.dk/

What is AI "made of"?

- ▶ Heterogeneous cluster – meaning it consists of several servers with different hardware configurations.
- ▶ You can use different NVIDIA GPUs in AI Cloud: T4, A10, A40, V100
- ▶ Three NVIDIA DGX-2 in the AI Cloud cluster (NIVIDIA's top-level GPU server from two generations ago).
- ▶ GPU system. CPU-primary computations should be done somewhere else. Strato, UCloud, or DeiC throughput HPC.
- ▶ FYI it is possible to get access to much larger facilities outside AAU, for example the supercomputer LUMI. Email claaudia@aau.dk to get access.

Which data can be processed?

▶ Shared access to multi-user system.
▶ Users' data separated by ordinary file system access restrictions.
▶ Not suitable for sensitive/secret data. Only usable for data classification levels
  levels 0 and 1
▶ Contact CLAAUDIA for a custom solution for research with confidential/sensitive
  data (levels 2 and 3). You can get a virtual machine with a GPU reserved for your
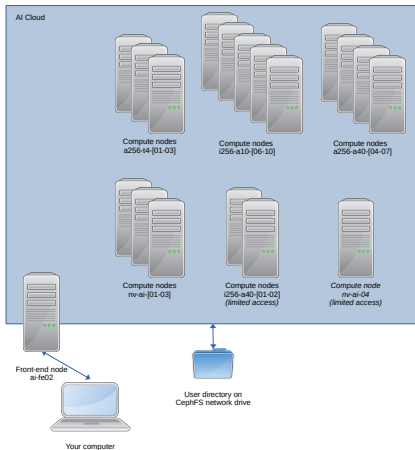 project.

AI Cloud Background

System design

Fair usage

Where to go from here

High level design

AI Cloud

Compute nodes
a256-t4-[01-03]

Compute nodes
i256-a10-[06-10]

Compute nodes
a256-a40-[04-07]

Compute nodes
nv-ai-[01-03]

Compute nodes
i256-a40-[01-02]
(limited access)

Compute node
nv-ai-04
(limited access)

Front-end node
ai-fe02

Your computer

User directory on
CephFS network drive

Resource management

9

AI Cloud is a multi-user environment

▶ Resources (CPUs, GPUs, memory) must be shared fairly between all users
▶ Solution: a resource management system (queue system)
▶ AI Cloud uses Slurm - a well-known resource management system in many HPC environments
▶ For the curious: Slurm documentation (it is quite extensive…)

Slurm provides you access to the computational resources.

Software environment

Users generally have different requirements for software in the AI Cloud. For example:
TensorFlow, PyTorch, CUDA, CUDNN, etc.

▶ Different users' requirements may be in conflict
▶ A shared selection of software for everyone would require a lot of maintenance
▶ Solution: personal containers for individual software environments
▶ AI Cloud uses Singularity (similar to Docker; newer versions are called AppTainer)
  to manage containers for individual users

Workflow in AI Cloud

You must use both the queue system Slurm and in most cases the container tool Singularity to be able to run computations in AI Cloud.

▶ Download or build a container to run your software in
▶ Singularity can only run on the compute nodes, so this must be run through Slurm
▶ Once you have a container, define your jobs to run in the container and start them via Slurm

Demonstration in AI Cloud

More tools

▶ The tool "nodesummary" can help you get an overview of how busy the AI Cloud's compute nodes currently are: https://git.its.aau.dk/CLAAUDIA/aicloud-tools

▶ DeiC have built the tool Cotainr to help build custom containers more easily: https://cotainr.readthedocs.io/en/latest/

▶ Both will be installed for all in the AI Cloud soon. Until then you can download the tools yourselves.

AI Cloud Background

System design

Fair usage

Where to go from here

Fair usage

We kindly ask that all users consider the following guidelines:

▶ Please be mindful of your allocations and refrain from allocating more resources than you know, have tested/verified that your jobs can indeed utilise.

▶ Please be mindful and de-allocate the resources when you do no use them. This ensures better overall utilisation for everyone.

▶ It is not OK to allocate a GPU to a job and then not use it. When we see jobs doing this, we will stop them and may have to do so without warning if the cluster is very busy.

▶ Resource discussion in the steering committee — contact your faculty representative.

AI Cloud Background

System design

Fair usage

Where to go from here

Where to go from here

▶ The user guide (the link from the first slide)
▶ You can request additions and clarifications by opening an issue here:
  https://github.com/aau-claaudia/aicloud-docs/issues
▶ Copying data to a local drive for higher I/O performance (only some of the
  compute nodes support this
▶ Support: support@its.aau.dk
▶ Use the resource and give feedback (we love feedback). Share with us your
  success stories (including benchmarks, solved challenges, new possibilities, etc.)
▶ Share with other users on the Yammer channel.