Introduction to AI Cloud

Slurm and Singularity Training

May 2022

CLAAUDIA, Aalborg University

Outline

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 I

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 recommendation from staff/supervisor/researcher.
- Free—but the system does attempt to balance load evenly among departments.

Background II

- Two NVIDIA DGX-2 in the AI Cloud cluster
  - Shared. Users' data separated by ordinary file system access restrictions. Not suitable for sensitive/secret data. Usable for levels 0 and 1
- One DGX-2 set aside for research with confidential/sensitive (levels 2 and 3) data.
  - Sliced (vitual machines). There are projects, and more are coming with requirements on data protection.
- GPU system. CPU-primary computations should be done somewhere else. Strato, uCloud, or DeiC throughput HPC.
- A lot of things are happening both in DK and at EU level. The HPC landscape is being reshaped, for example with the new supercomputer LUMI. It is possible to get access to much larger facilities outside AAU. Email claaudia@aau.dk for more information.

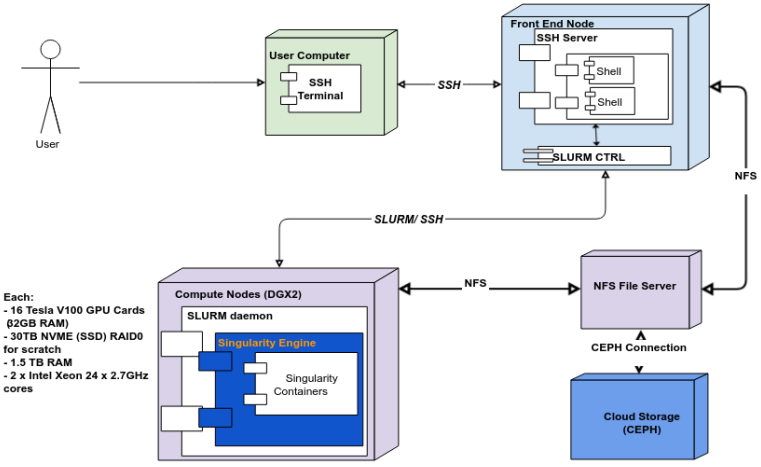
AI Cloud Background

System design

Fair usage

Where to go from here

High level design



Resource management

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

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 administration
- Solution: personal containers for individual software environments
- AI Cloud uses Singularity (similar to Docker) to manage containers for individual users

Workflow in AI Cloud

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

- Get or build a container to run you 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

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.

We see challenges towards the end of semesters (cyclic):

- More HW (NVIDIA T4, A10, A40) is on the way in "new AI Cloud".
- It is for research ... administration intends to interfere as little as possible ... but we do try to help and do something.
- 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 documentation
  - More workflows
  - Copying data to the local drive for higher I/O performance
  - Inspecting your utilization
  - Matlab, PyTorch, ...
  - Fair usage/upcoming deadline
  - Links and references to additional material
  - Support (fastest response): support@its.aau.dk
  - Advisory (slower response – longer time span): claaudia@aau.dk
- Use the resource and give feedback. Share with us your success stories (including benchmarks, solved challenges, new possibilities, etc.)
- Share with other users on the Yammer channel.