Advanced Techniques of Machine Translation

Tutorial Session #1

ScienceCluster: Introduction

- What is the ScienceCluster?

*ScienceCluster is a managed **high-performance computing** (HPC) **environment** designed to support a wide range of **research** workloads. Whether your tasks are CPU-intensive, require large shared memory, benefit from **GPU acceleration**, or depend on high-speed interconnects, ScienceCluster enables you to choose the optimal resources for your needs. It consists of a **network of interconnected machines** that **function as a single**, unified **system**.*

- Pilot project: ScienceCluster access for non-research purposes (lectures/exercises)

ScienceCluster: Installing stuff

- uses mamba (=conda) for package management
- *module load mamba*
  - loads the module mamba, without that command the ScienceCluster doesn't recognize mamba
  - probably does some path-shenanigans in the background ? (not an admin, so who knows)

ScienceCluster: Transferring Files

- scp:
  - Move file from local filesystem to the ScienceCluster
    scp c/path/to/local/file shortname@cluster.s3it.uzh.ch:path/to/file/on/server
  - Move directory recursively to the ScienceCluster
    scp -r c/path/to/dir shortname@cluster.s3it.uzh.ch:path/to/target/dir
- rsync:
  - Sync whole directories
  - Ignores unchanged files
    ➔ useful for backing up/mirroring data
- ScienceApp GUI
  - Probably the **most straightforward** method
  - https://apps.s3it.uzh.ch/pun/sys/dashboard

ScienceCluster: Storage Options

home

/home/$USER

15GB

for configs, etc.

data

/data/$USER

200GB

for the ‘main’ data, e.g. scripts, models, etc.

scratch

/scratch/$USER

20TB

purged after 30 days

HDD

shares

/shares/<PROJECT>

per project, shared for all members

10TB

-HDD

ln -s /shares/atomt ~/shares

Good practice: Backup important data!

ScienceCluster: SLURM

- OpenSource cluster management software
- Commands:
  - srun
  - **sbatch**
  - **sacct**
  - **scancel**

Running things on the ScienceCluster (with SLURM)

sbatch my_script.sh

examples:

- toy_example.sh
- assignment1.sh

Take not of how slurm takes parameters:

#SBATCH …

Important: Parameters have to occur between the ‘shebang’ and the first instruction!

Assignment 1

- Train a NMT model from scratch (transformer)
- CZ-EN parallel corpus
  - Training: 10M sentence pairs
  - Test/Valid: 5K sentence pairs each
- Start with a toy example, just to get used to the ScienceCluster
  - 1’000/100/100 split
- Plan enough time! Training may still take up to 14 hours, even if running on GPU

Good luck!