
NYU
NYU HPC Big Data
hpc@nyu.edu
https://hpc.nyu.edu
NYU

Who We are?

Research Technology Services Team, IT

- High Performance Computing (HPC) for research and courses
- Support Big Data, ML/DL/AI
- Cloud Computing with Google Cloud Platform (GCP), Amazon Web Service (AWS)
- JupyterHub for courses
- Security Data Research Environment (SDRE)
- High Speed Research Network (HSRN)
- Other research related requests for IT supports

Outline

- Dataproc cluster for courses
  - Available services with Dataproc
  - Backend setups for Dataproc
- Big data on HPC Greene cluster
  - Spark with single node and multiple nodes cluster mode
  - Dask on Greene cluster
- Q&A

What is Dataproc

- Dataproc is a managed Hadoop and Spark service in Google Cloud Platform. It takes advantage of open source tools for big data processing and machine learning.
- Its components include Hadoop Distributed File System (HDFS), MapReduce, Spark, Hive, Trino, Pig etc.
- The default block size in HDFS is 128 MB. The file system is not good for storing many small sized files.
- YARN is used for resource management, and job scheduling and monitoring. Its functionalities are very preliminary, comparing to its HPC analog such as SLURM.

Dataproc For Courses I

- The cluster serves many hundreds of course takers, for practice and doing exercise, not for real big data projects. The cluster is created at the beginning of a semester and destroyed at the end of a semester
- Spark runs on top of YARN
- Spark structured streaming is feasible

Dataproc For Courses II

- User entrance
  - https://dataproc.hpc.nyu.edu/
- Getting started user guide
- Home filesystem
  - When you first login, you land at /home/<netid>_nyu_edu directory
  - Where you can run the commands: ls, pwd, cd etc
- HDFS filesystem
  - 500GB quota per user
  - Access by run the commands: hdfs dfs -ls, hdfs dfs -put ...

Demos

NYU Dataproc

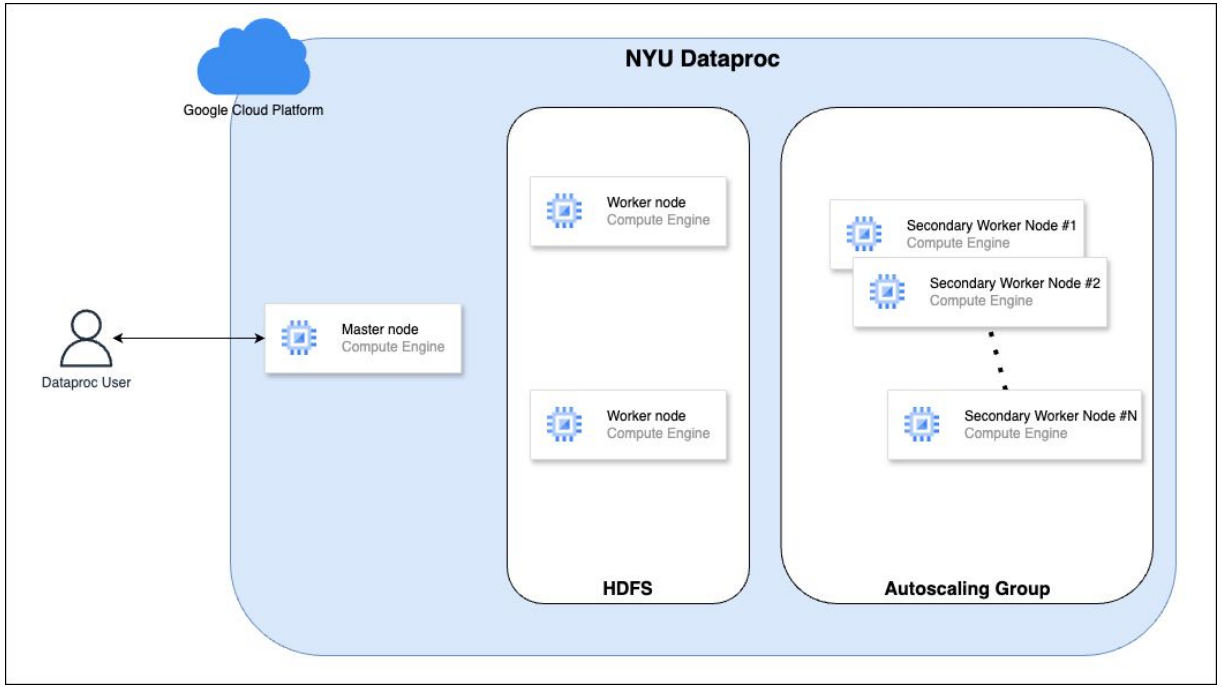
Google Cloud Dataproc is a cloud-based Hadoop service that NYU HPC provides for courses. It comes with the following interfaces for interacting with Hadoop:

Interface	Description	URL
Command Line Interface Console	This is how you log into Dataproc.	https://dataproc.hpc.nyu.edu/ssh
Data Ingest Console	This is how you upload data into Dataproc.	https://dataproc.hpc.nyu.edu/ingest
MapReduce Job History	A web interface where you can see all MapReduce jobs that have run on the cluster.	https://dataproc.hpc.nyu.edu/jobhistory/
Spark History Server	A web interface where you can see all Spark jobs that have run on the cluster.	https://dataproc.hpc.nyu.edu/sparkhistory/
YARN Application Timeline	A web interface where you can see all applications (both Hadoop MapReduce and Spark) that have run on the cluster.	https://dataproc.hpc.nyu.edu/apphistory/
YARN ResourceManager	YARN is the resource manager and job scheduler used by the Dataproc cluster. YARN allows you to use various data processing engines for batch, interactive, and real-time stream processing of data stored in HDFS. The YARN web interface allows you to see all current and recently executed applications, as well as information about the current state of the cluster (such as number of vCores and RAM free).	https://dataproc.hpc.nyu.edu/yarn/

Dataproc For Courses III

- Maximum YARN application lifetime is **5** hours. Vast majority finish in <5 mins.
- The default Spark deploy mode is cluster, that is recommended for production run. For debugging purposes, run spark in the client mode:
  - spark-shell --deploy-mode client
- login node is for debugging and submitting jobs, not for computing heavy work. When done for the day, please log out.
  - Memory limit per user is **3** GB
  - User process lifetime is **48** hours.
- pyspark is available, and the python and its package are from conda install.
- container log size check enabled. App outputs should be written to HDFS, not to stdout.

Dataproc System Setup I



Dataproc System Setup II

- Login to master node with your NYU Google account
  - home folder name: **netid_nyu_edu**
- Start with 1TB Hadoop Distributed File System (HDFS). Scale up to **8**TB.
- Ingress storage bucket to upload large input data
- All files in home and HDFS will be **deleted** at the end of the semester

Big Data On Greene HPC Cluster

Greene Cluster

- Greene is a general purpose HPC cluster for NYU public use
- Greene hardware
  - ~38K CPU cores, 220TB memory
  - 768 GPUs, 13TB GPU memory
  - 12PB parallel storages
  - HDR 200Gb/s Infiniband network for computation communications and storages

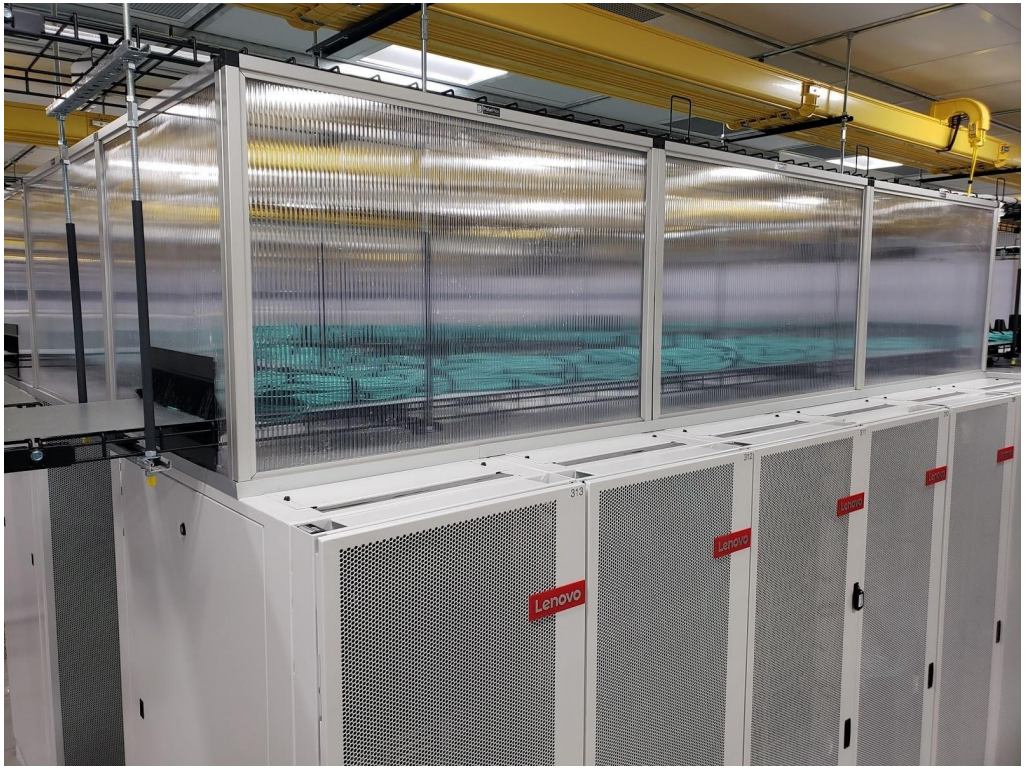
https://hpc.nyu.edu

Greene Cluster in Research Computing Center (RCDC)

Greene Cluster in Research Computing Center (RCDC)

NYU

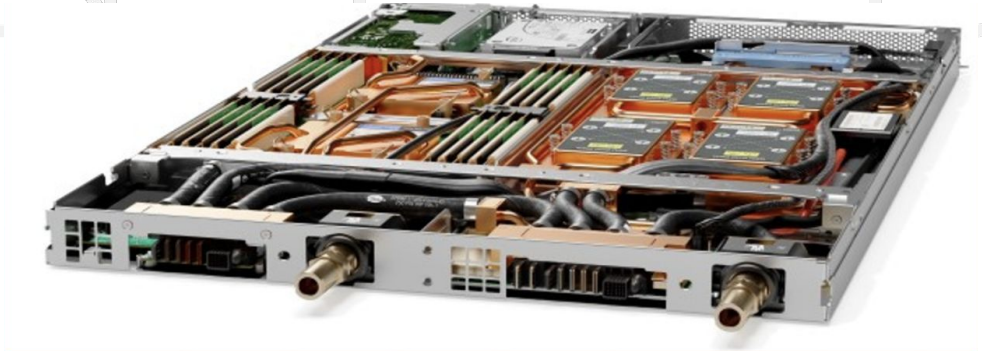
Greene Cluster in Research Computing Center (RCDC)

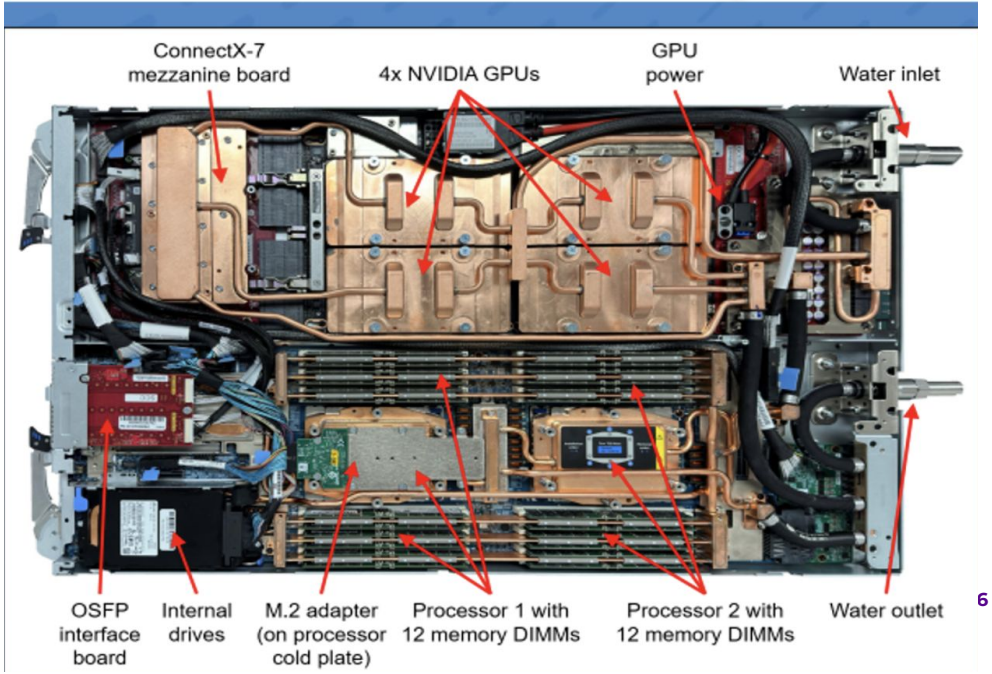


NYU

Greene Cluster in Research Computing Center (RCDC)

Liquid cooling NVIDIA H100 GPU server SD650N-V3





NYU

Before Logging In

Check your Network Location

NYU Campus:

From within the NYU network, that is, from an on-campus location, or after you VPN inside NYU's network, you can follow the next page for instructions to login to the HPC clusters directly.

Off Campus:

From a commercial network outside of NYU (from an off campus location). Before following the next page to log into the clusters, you would need to connect to the NYU VPN or the HPC Gateway Server.

Greene Cluster Logins

- Hostname: greene.hpc.nyu.edu
- 3 login nodes are configured with load balancer, open to NYU network and NYU VPN
- HPC clusters are using NYU netID and associated password
- Mac/Linux/Windows Subsystems from terminal

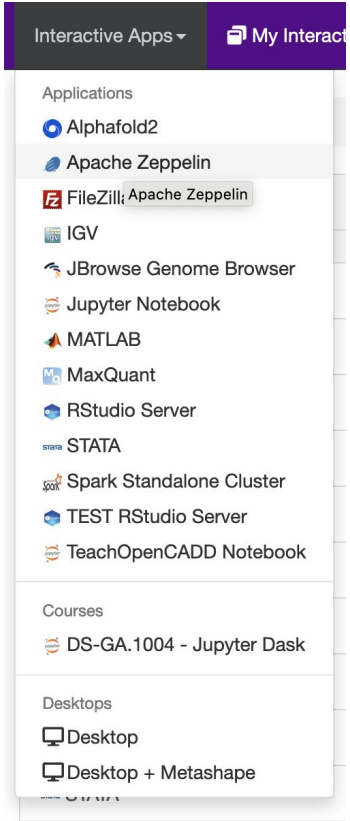
```
ssh NYUnetID@greene.hpc.nyu.edu
```

- For Windows: Putty, MobaXterm
- VScode

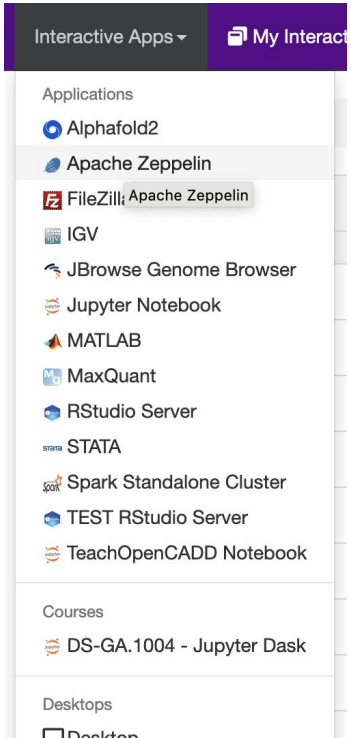
https://sites.google.com/nyu.edu/nyu-hpc/accessing-hpc

Greene Open OnDemand Server

- HPC cluster in browser
  - https://ood.hpc.nyu.edu
- Provide GUI interface to Jupyter notebooks, Matlab, **Spark Standalone Cluster**, **Dask with Jupyter**, Remote Desktop

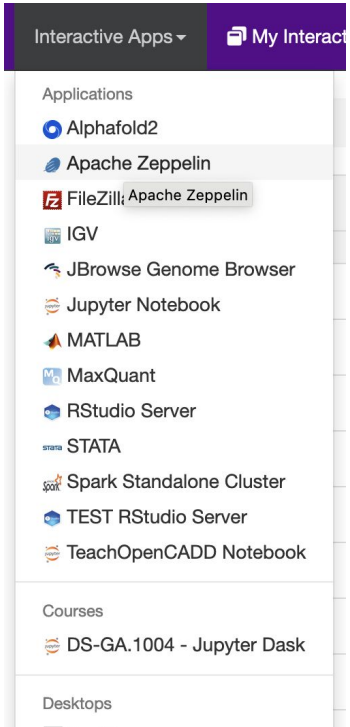


Spark on OOD Demons



NYU

Dask on OOD Demos



NYU

Spark Jobs in Batch Mode

Examples are available:

/scratch/work/public/apps/pyspark/3.5.0/examples/spark

Greene Compute Nodes

Nodes	Number	# of CPU cores	Available memory (GB)	GPUs
Standard memory	524	48	180	
Medium memory	40	48	369	
Large memory	4	96	3014	
V100 NVIDIA GPU	11	48	369	4 32GB/16GB PCIe NVIDIA V100
RTX8000 NVIDIA GPU	73	48	369	4 48GB PCIe NVIDIA RTX8000
A100 NVIDIA GPU	9	64	500	4 80GB Nvlink NVIDIA A100
A100 NVIDIA GPU	34	80	1024	4 80GB Nvlink NVIDIA A100
AMD MI100 GPU	3	128	512	8 32GB PCIe AMD MI100
AMD MI250 GPU	2	128	512	8 64GB AMD MI250
NVIDIA H100 GPU	15	96	1510	4 80GB Nvlink NVIDIA H100

NYU

Greene Hardware Specifications

Q/A