MINI-HPC AND HYBRID HPC-BIG DATA CLUSTERS

ASHRAQAT MOHAMED 221000836
RAGHAD MOHAMED 221001278
FARID MAGED 221000545
MALAK HAITHAM 221001396

SUPERVISED BY: DR. MOHAMED MAHMOUD EL SAYEH

COURSE CODE:CBIO312

INTRODUCTION

This project explores distributed machine learning using:

1. A Mini-HPC Cluster with MPI
2. A Hybrid Big Data Cluster with Docker Swarm and Spark

We tested then trained the ML model on:

1. A testing dataset
2. A real bioinformatics gene expression dataset

The goal was to compare performance, scalability, and architecture of traditional vs. modern clusters.

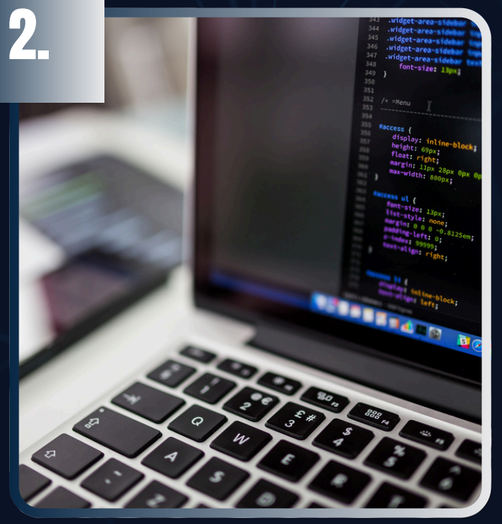
MATERIALS & ENVIRONMENT

1.

HARDWARE/TOOLS USED

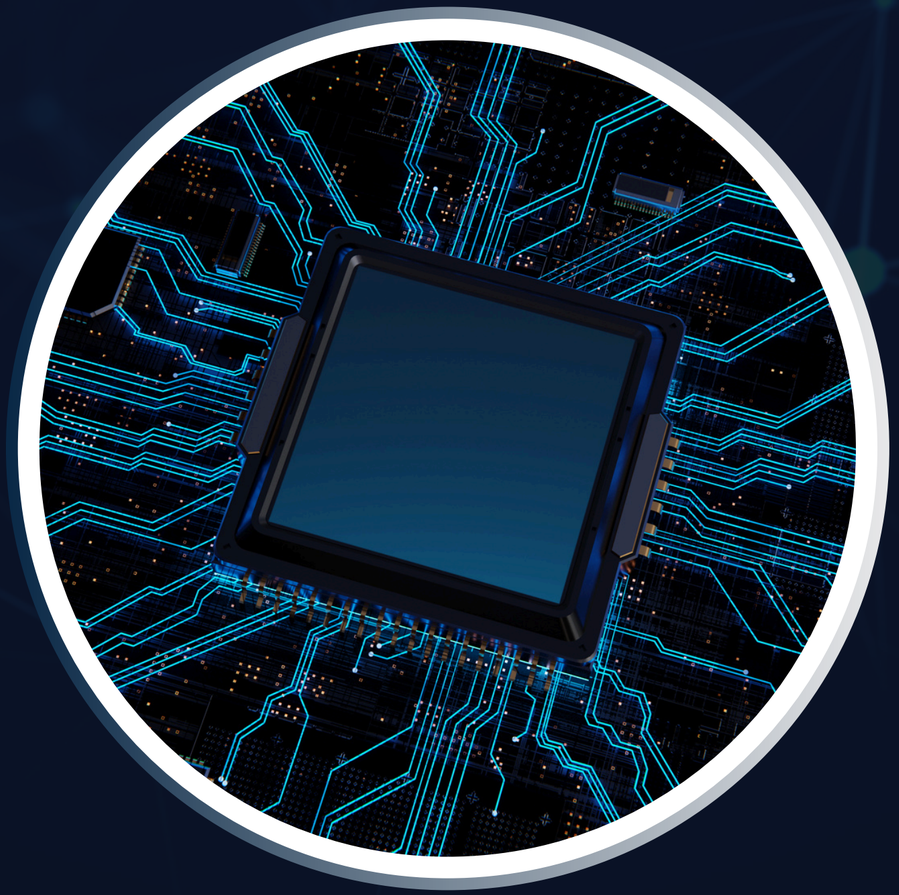
- 3 VirtualBox VMs (1 master + 2 workers)
- Ubuntu Server 20.04 on each VM
- Docker, OpenMPI, mpi4py, PySpark

2.



SOFTWARE/ML LIBRARIES

- Python 3, scikit-learn, pandas, numpy
- Spark MLlib, Docker Compose
- Datasets



METHODOLOGY

We set up three virtual machines with static IP networking and configured passwordless SSH to enable seamless communication between nodes. Using mpi4py, we ran distributed machine learning tasks across the cluster. After that, we installed Docker, initialized a Swarm, and deployed a Spark cluster using spark-stack.yml. Finally, we ran distributed ML jobs on the gene expression dataset using PySpark.

RESULTS

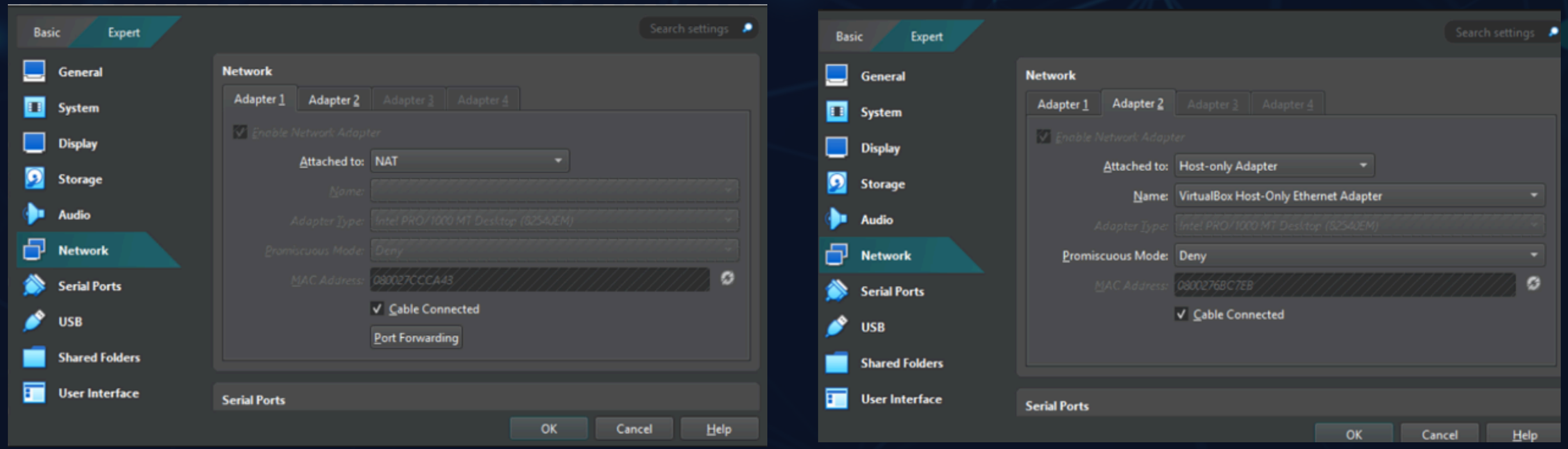


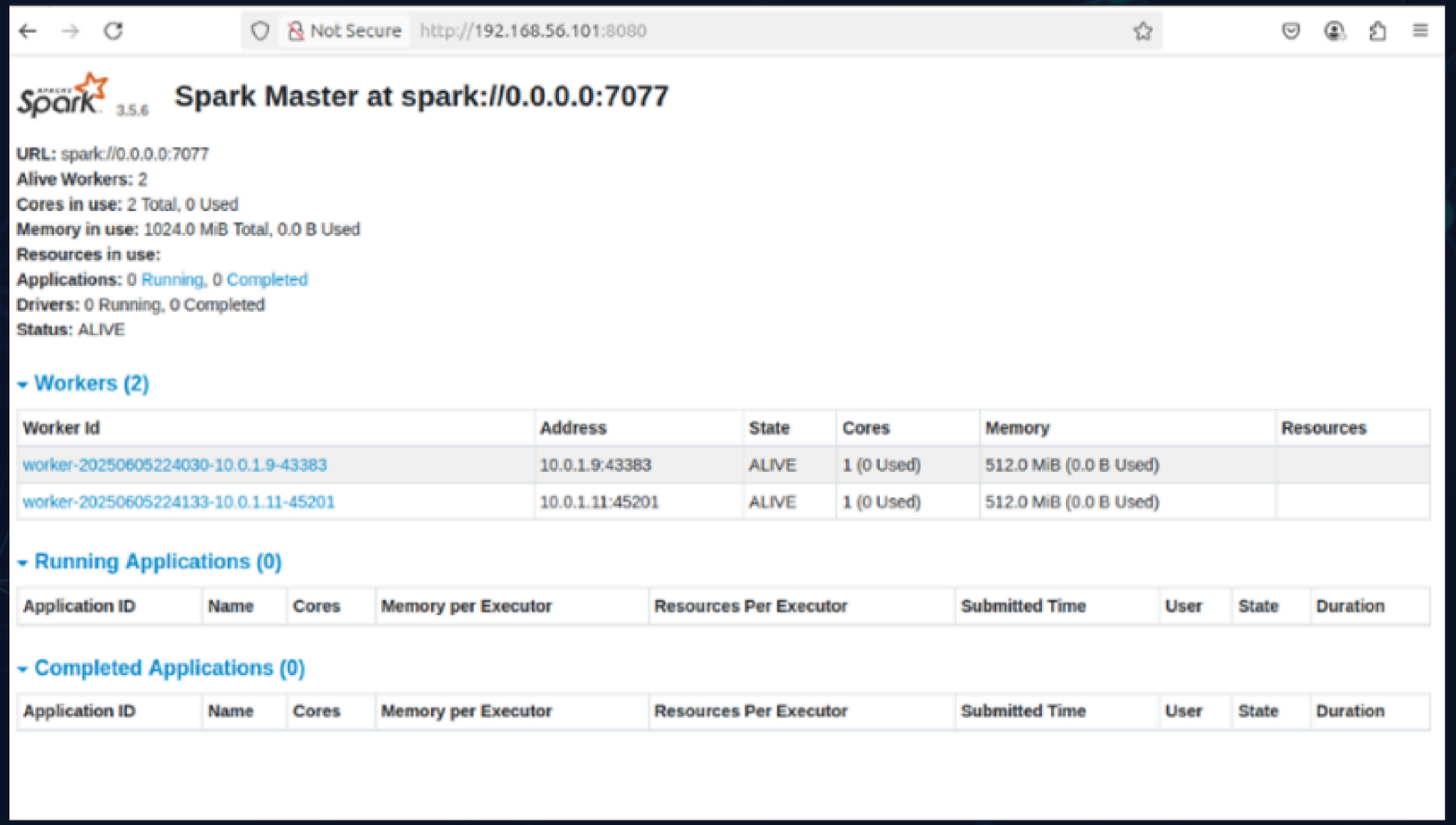
Fig 1.0 Network setup for 3VM

RESULTS

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Gene_1_Expression	Gene_2_Expression	Gene_3_Expression	Gene_4_Expression	Gene_5_Expression	Gene_6_Expression	Protein_1_Level	Protein_2_Level	Protein_3_Level	Protein_4_Level	Protein_5_Level	Protein_6_Level	Disease_Status
2	6.39	5.58	6.09	4.87	5.21	5.93	2.14	1.98	2.05	1.89	2.11	2.03	Healthy
3	3.12	2.98	3.25	3.01	3.15	2.87	4.23	3.89	4.12	4.01	3.95	4.15	Diseased
4	5.67	4.92	5.34	5.12	4.78	5.45	1.92	2.07	1.85	2.13	1.99	2.04	Healthy
5	2.89	3.21	2.95	3.08	3.14	2.76	3.78	4.12	3.95	4.03	4.21	3.87	Diseased
6	5.23	6.01	5.47	4.95	5.32	5.76	2.09	1.87	2.12	1.94	2.01	2.08	Healthy
7	3.34	3.02	3.19	2.91	3.27	3.11	4.15	3.92	4.07	3.85	4.13	4.02	Diseased
8	6.12	5.45	5.89	5.67	5.34	6.01	1.95	2.03	1.88	2.14	1.97	2.06	Healthy
9	2.78	3.15	2.92	3.04	2.99	3.23	4.09	3.76	4.18	3.94	4.05	3.81	Diseased

leukemia_expression.csv

RESULTS



Spark integration

RESULTS

Metric	MPI (Task 1)	Spark (Task 2)
Training Time	~0.05–0.2s	~15–20s
Test Accuracy	63–67% (digits), ~70% biodata	~68–70% (bio data)
Scalability	Manual (slots per node)	Auto-scaling via Swarm
Feature Selection	Manual / top-k features	ChiSqSelector (auto)
Fault Tolerance	Low (restart if failed)	High

DISCUSSION

- MPI cluster gave faster training times but lacked fault tolerance.
- Spark took longer to start but offered robust job handling and better recovery.
- Troubleshooting SSH and networking gave hands-on experience in cluster configuration.
- Spark UI was helpful for monitoring jobs.
- We successfully distributed ML tasks on real gene expression data, showing both setups were effective.

CONCLUSION

- Built and evaluated two distributed computing architectures:
  - A Mini-HPC cluster using MPI/mpi4py
  - A hybrid HPC-Big Data cluster using Docker Swarm and Spark
- Demonstrated real-world applicability by running distributed machine learning tasks on a bioinformatics gene expression dataset, achieving accurate classification results.

THANK YOU!