

Assignment No. 5.

Title : HPC Application.

Problem statement: Implement HPC Application in AIML domain.

Software & Hardware Requirement:

Multicore processor, 4GB RAM, Jupyter notebook, python, libraries.

Theory :

HPC applications in the field of AIML involve leveraging parallel computing architectures to train and deeply complex models on large datasets effectively. Here's a theoretical overview of implementing an HPC application for AIML domain.

Parallel Processing :

HPC applications harness the power of parallel processing to accelerate AIML tasks such as model training and inference.

Data Parallelism :

In this the dataset is divided into smaller batches & multiple computing units process these batches simultaneously.

Each unit performs the same operations on its assigned batch of data, allowing for efficient training.

Model Parallelism:

It involves distributing different parts of a model across units. This approach is useful for training large models that cannot fit into memory of a single unit.

Task Parallelism:

It involves executing independent tasks concurrently to maximize resource utilization. Can be applied to tasks such as data preprocessing, feature extraction, model training & inference.

Optimized algorithms and libraries.

HPC application for AI/ML domain utilize optimized algorithms and libraries tailored for parallel computing. These include optimized linear algebra (e.g. Intel MKL, NVIDIA CUBLAS) distributed deep learning frameworks (e.g. Tensorflow, pyTorch) and parallel optimization algorithm (e.g. stochastic gradient descent with momentum).

Hardware Acceleration:

Hardware accelerations such as GPUs and TPUs play a crucial role in accelerating AI/ML workload. These offer massive parallelism and are optimized for deep learning tasks enabling sufficient speedups compared to CPUs.

Scalability and Performance Optimization:
HPC application in AI/ML domain focus on scalability and performance optimization to handle large scale datasets. Techniques such as asynchronous arithmetics are employed.

Resource management and Job scheduling.
Efficient resource management and job scheduling are essential for maximizing the utilization of HPC resources.

Job scheduling such as SLURM, PBS, kubernetes, are used to allocate compute resources, manage job queries.

Monitoring and profiling.

These are employed to analyze the performance of AI/ML applications on HPC systems. These help identify bottlenecks, optimize resource utilization and fine tune parameters to achieve optimal performance.

Conclusion:

Thus we have successfully implemented HPC application for all AI/ML domain.

