

# Welcome to this session: High Performance Computing for Data Science

**The session will start shortly...**

Questions? Drop them in the chat.  
We'll have dedicated moderators  
answering questions.





# What is Safeguarding?

Safeguarding refers to actions and measures aimed at protecting the human rights of adults, particularly vulnerable individuals, from abuse, neglect, and harm.



To report a safeguarding concern reach out to us via email:  
[safeguarding@hyperiondev.com](mailto:safeguarding@hyperiondev.com)

## Live Lecture Housekeeping:

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly.
- No question is daft or silly - ask them!
- For all non-academic questions, please submit a query:  
[www.hyperiondev.com/support](http://www.hyperiondev.com/support)
- To report a safeguarding concern reach out to us via email:  
[safeguarding@hyperiondev.com](mailto:safeguarding@hyperiondev.com)
- If you are hearing impaired, please kindly use your computer's function through Google chrome to enable captions.



## Learning Outcomes

---

- ❖ **Discuss the role of High-Performance Computing (HPC)** in data science and its advantages over traditional computing.
- ❖ **Explain key concepts in HPC**, including parallel and distributed computing, GPU acceleration, and cloud-based HPC solutions.
- ❖ **Evaluate different HPC architectures** and frameworks used in data science, such as MPI, CUDA, and Spark.
- ❖ **Analyse real-world case studies** where HPC has been used to solve complex data science problems.
- ❖ **Recognize challenges and limitations of HPC**, including scalability, energy efficiency, and cost.



## Large-Scale Data Processing

As data science applications become more complex, traditional computing methods struggle to handle large-scale data processing efficiently. Imagine that we could share the workload involved in a project between multiple devices.

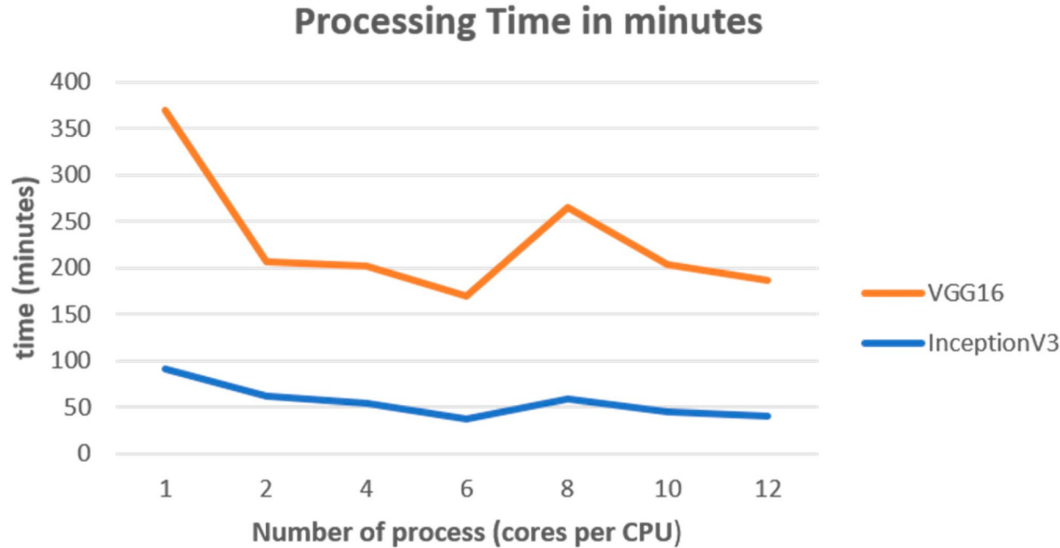
- What do we need to consider regarding the **division of work**?
- How will transitioning a problem from being solved by one device, to multiple devices simultaneously **affect the complexity of the problem** and the **cost to the environment**?

# Large-Scale Data Processing

High-Performance Computing (HPC) provides powerful solutions, but it comes with its own set of challenges and considerations. HPC plays a crucial role in areas such as climate modeling, genomic analysis, financial forecasting, and AI research. Understanding how to leverage HPC is essential for data scientists tackling large-scale problems.

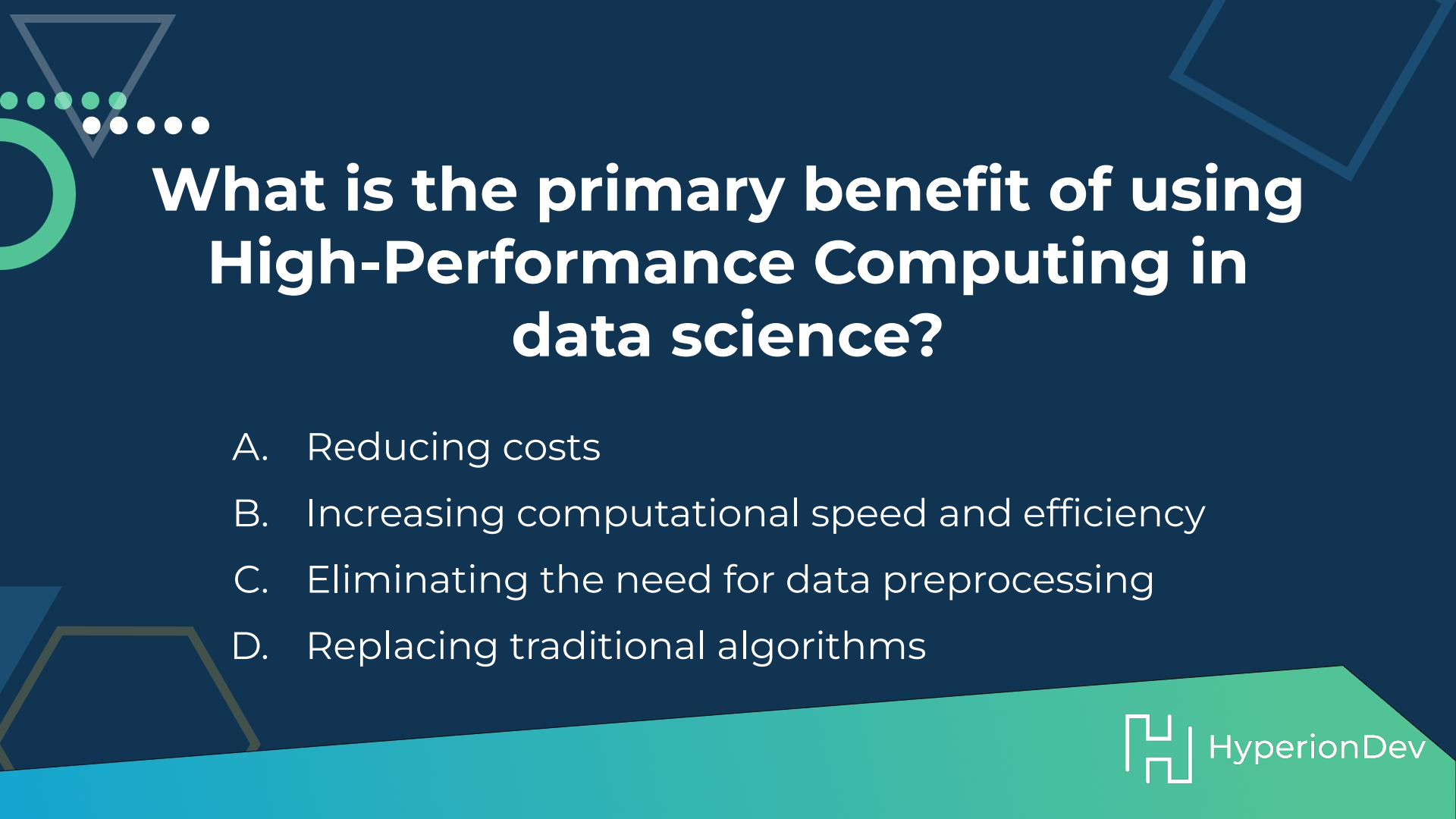
- *How long do you think it would take to train an AI model on a single CPU versus a distributed HPC system?*

# Large-Scale Data Processing



VGG16 and InceptionV3 are two different Convolutional Neural Network models

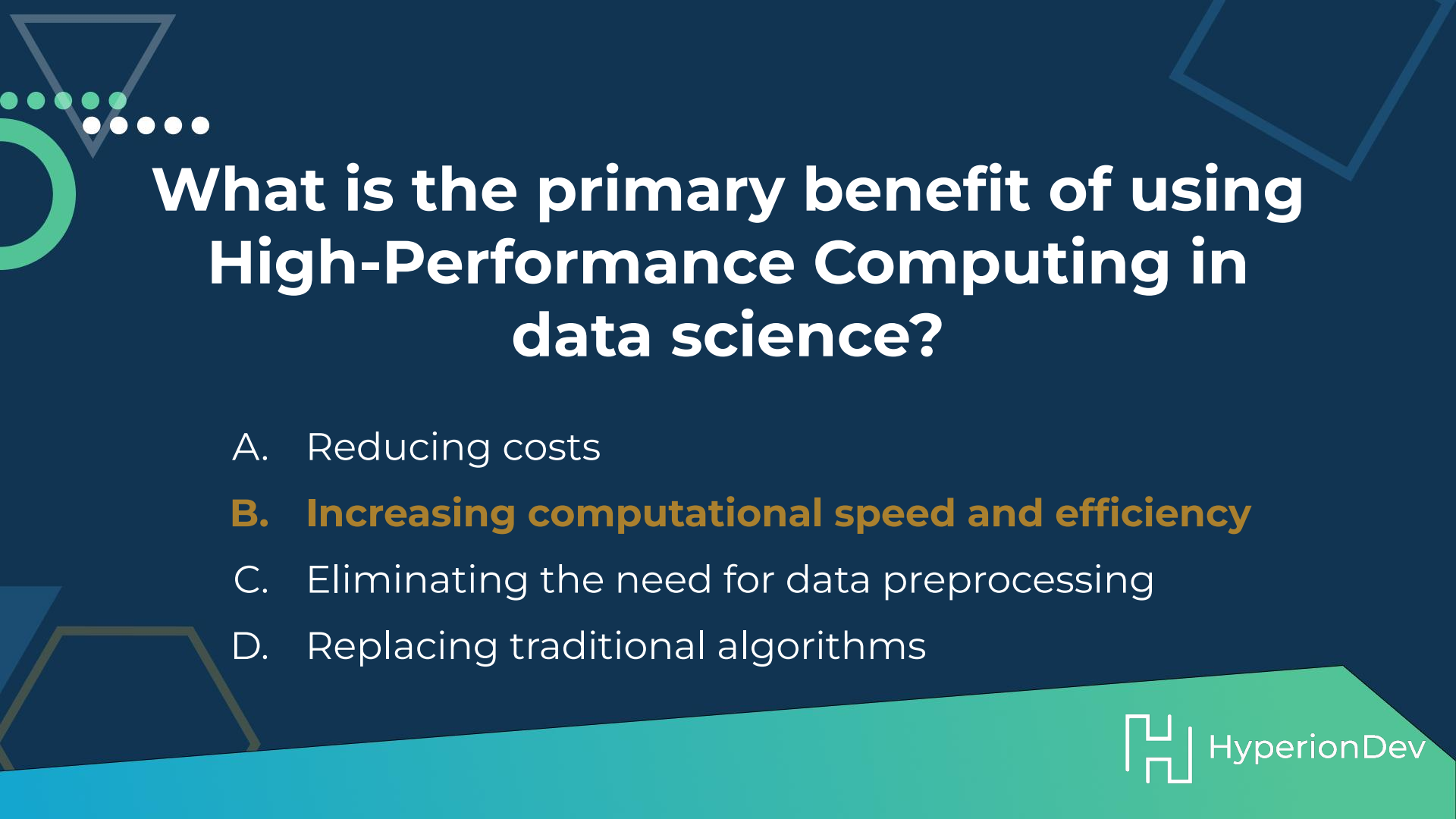
Source: [Weed Classification using Convolutional Neural Networks](#)



# What is the primary benefit of using High-Performance Computing in data science?

- A. Reducing costs
- B. Increasing computational speed and efficiency
- C. Eliminating the need for data preprocessing
- D. Replacing traditional algorithms





# What is the primary benefit of using High-Performance Computing in data science?

- A. Reducing costs
- B. Increasing computational speed and efficiency**
- C. Eliminating the need for data preprocessing
- D. Replacing traditional algorithms



# Which of the following best describes parallel computing?

- A. A single process executing multiple instructions at once
- B. Dividing tasks across multiple processors to execute simultaneously
- C. Running multiple applications at the same time
- D. Using cloud storage to enhance computation



# Which of the following best describes parallel computing?

- A. A single process executing multiple instructions at once
- B. Dividing tasks across multiple processors to execute simultaneously**
- C. Running multiple applications at the same time
- D. Using cloud storage to enhance computation



....  
**Which framework is commonly used  
for distributed computing in data  
science?**

- A. TensorFlow
- B. Apache Spark
- C. NumPy
- D. PostgreSQL



Which framework is commonly used  
for distributed computing in data  
science?

- A. TensorFlow
- B. Apache Spark**
- C. NumPy
- D. PostgreSQL

## Lecture Overview

---

- Introduction to HPC
- Key Concepts
- Frameworks
- Case Studies in HPC
- Challenges and Future Trends in HPC





# Introduction to High Performance Computing

# High Performance Computing

- ❖ The use of powerful processors, networks and parallel supercomputers to tackle problems that are very computationally or data-intensive
- ❖ A HPC system **divides workloads into smaller tasks** and assigns them to **multiple resources** for **simultaneous processing**.
- ❖ HPC systems can solve problems that are either **too large** for standard computers to handle individually or would take **too long to process**.
  - For this reason, it is also sometimes referred to as **supercomputing**.





# High Performance Computing

- ❖ HPC's parallel computing capabilities can greatly **accelerate iterative processes** compared to traditional computing.
  - HPC systems typically run at speeds **more than one million times faster** than the fastest commodity desktop, laptop or server.
- ❖ HPC's ability to **quickly process massive amounts of data** powers some of the most fundamental aspects of today's society.
  - Reduce the time to **train deep learning models** from days to hours
  - Banks can verify **fraud** on millions of credit card transactions at once
  - Automakers test their car designs for **crash safety**
  - Simulations to help us **predict the weather or molecular structure**



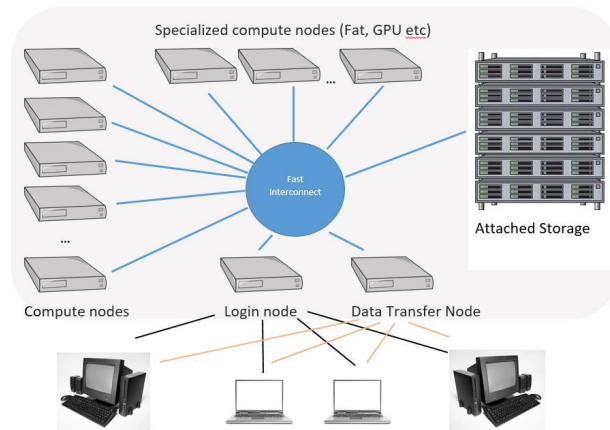
# Key Concepts in HPC

## Key Concepts in HPC

- ❖ HPC can take the form of **custom-built supercomputers** or **groups of individual computers** called **clusters or nodes**.
- ❖ **An HPC cluster** comprises multiple **high-speed computer servers** which use **high-performance multi-core CPUs** or **GPUs**.
- ❖ GPUs are **specialized computer chips** designed to process **large amounts of data** in parallel, making them ideal for some HPC, and are currently the standard for ML/AI computations.
- ❖ A single HPC cluster can include 100,000 or more nodes.

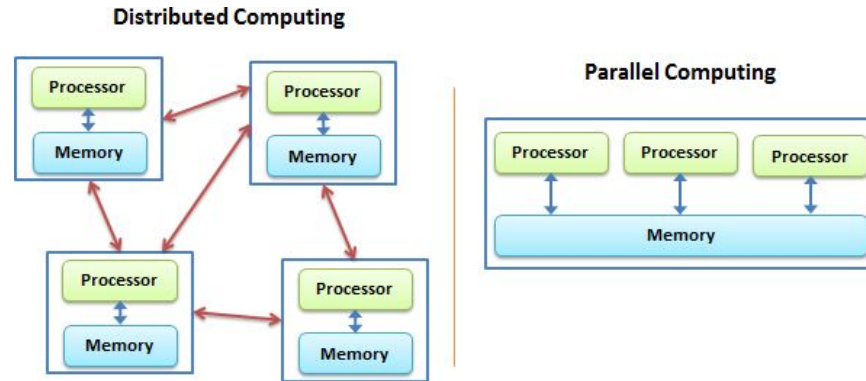
# Key Concepts in HPC

- ❖ High performance computing has **three main components**:
  - Compute, Network, Storage
- ❖ The nodes **(compute)** of the HPC system are connected to other nodes to run algorithms and software simultaneously, and are then connected **(network)** with data servers **(storage)** for output.



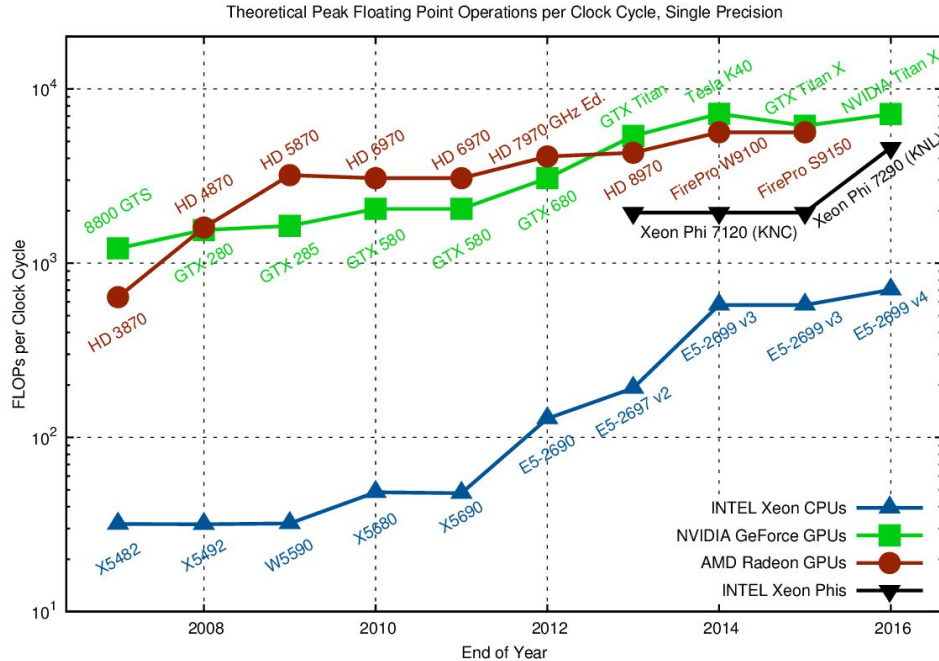
Source: [Iowa State University](https://www.iastate.edu/~hpc/)

- ❖ HPC can typically be broken down into two general design types:
  - **Cluster or parallel computing:** runs multiple tasks simultaneously on multiple, closely-connected processors. Best for computationally complex problems.
  - **Distributed computing:** connects computers via a network so that they can act as one powerful machine. Suits problems with massive datasets.



Source: [O'Reilly](#)

# CPU vs GPU for ML



Source: [Comparing CPUs and GPUs for Machine Learning](#)



BREAK



# HPC Frameworks

- ❖ There are three main programming paradigms/frameworks for HPC:

OpenMP	CUDA	MPI
Open Multi-Processing	Compute Unified Device Architecture	Message-Passing Interface
Shared-memory parallel programming model	Parallel computing platform for NVIDIA GPUs	Message-passing interface for distributed computing
Easy to use and integrate into existing code	Based on threads for parallel execution	Suitable for communication between multiple processing units
Suitable for multi-core processors	Suitable for massive parallel processing	Suitable for distributed computing environments

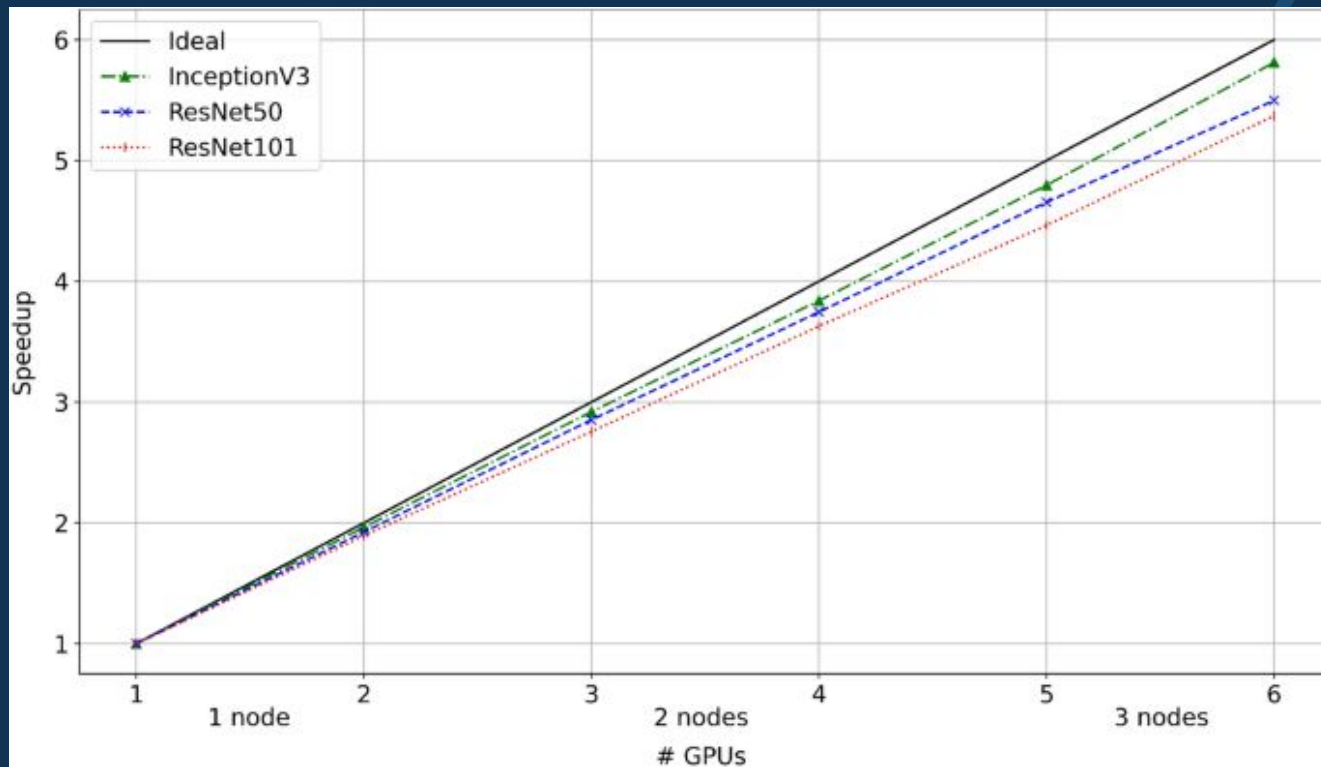


# HPC Frameworks

- ❖ We will often find ourselves in scenarios where we'd like to implement parallel or distributed computing in our code.
- ❖ There are many packages which we can use to implement it in our code:
  - **Spark:** open-source, distributed processing system used for big data workloads.
  - **Hadoop:** framework that allows for the distributed processing of large data sets across clusters of computers.
  - **Dask:** Python module and Big-Data tool that enables scaling pandas and NumPy.

## Case Studies

- ❖ **AI & Machine Learning:** The paper “[A Container-Based Workflow for Distributed Training of Deep Learning Algorithms in HPC Clusters](#)” presents a methodology that leverages containerization and the Horovod library to facilitate distributed training of deep learning models across multiple GPUs in HPC environments. This approach addresses challenges related to multi-user HPC systems and varying cluster configurations, offering a scalable and efficient solution for training complex models.



Speedup plot for the TensorFlow Benchmark experiment using udocker over the different GPU configurations

## Case Studies

- ❖ **Climate Science:** In “[ExtremeWeather: A Large-Scale Climate Dataset for Semi-Supervised Detection, Localization, and Understanding of Extreme Weather Events](#)”, researchers introduce a comprehensive climate dataset designed to enhance the detection and analysis of extreme weather phenomena. The study demonstrates that semi-supervised learning techniques applied to this dataset can improve the identification and understanding of extreme weather patterns, aiding in more accurate predictions.

## Case Studies

- ❖ **Genomics:** The dissertation “[High-Performance Computing in Next-Generation Sequencing Read Alignment](#)” explores innovative computational strategies and hardware acceleration to optimize the alignment process of short DNA sequences to reference genomes. By integrating HPC techniques, the study addresses the challenges posed by the massive data output of next-generation sequencing, enhancing the efficiency and accuracy of genomic analyses.

## Case Studies

- ❖ **Finance:** The study “[Evaluating the Efficacy of Machine Learning Models in Credit Card Fraud Detection](#)” systematically evaluates the performance of various machine learning algorithms in detecting fraudulent credit card transactions. Utilizing a dataset of over 555,000 transactions, the research compares models such as Logistic Regression, Support Vector Machines, Random Forest, Gradient Boosting, and Neural Networks. The findings highlight that ensemble learning methods and neural networks, when powered by HPC, significantly enhance the accuracy and reliability of fraud detection systems.

## Challenges

- ❖ Although HPCs are undoubtedly useful, they do face challenges:
  - **Cost:** HPC system requires a large budget for hardware, data center, and skilled technicians.
  - **Power consumption:** Newer processors require more energy and generate more heat. High-density CPUs and GPUs used in AI applications consume a lot of power.
  - **Infrastructure complexity:** Legacy data centers may not be able to support the demands of HPC. Integrating multiple processors and accelerators increases system complexity



## Future Trends

- ❖ In the future, HPC systems may leverage quantum computing to achieve unprecedented processing power.
- ❖ Quantum algorithms create multidimensional computational spaces that are a much more efficient way of solving complex problems.
- ❖ As the capacity of HPC processing continues to expand, so too will the ability of systems to tackle our most complex engineering, scientific, and AI-related challenges.





# Which of the following is NOT a typical challenge in High-Performance Computing?

- A. Scalability
- B. High energy consumption
- C. Faster processing
- D. Cost



# Which of the following is NOT a typical challenge in High-Performance Computing?

- A. Scalability
- B. High energy consumption
- C. Faster processing**
- D. Cost



# Why is Apache Spark preferred for big data processing in HPC?

- A. It supports deep learning models
- B. It efficiently handles large-scale distributed data processing
- C. It replaces the need for databases
- D. It is the only available HPC framework



# Why is Apache Spark preferred for big data processing in HPC?

- A. It supports deep learning models
- B. It efficiently handles large-scale distributed data processing**
- C. It replaces the need for databases
- D. It is the only available HPC framework



# What is a potential future trend in HPC that could revolutionize data science?

- A. Increased reliance on traditional CPUs
- B. Quantum computing
- C. Elimination of parallel computing
- D. Replacement of all data science models with heuristics



# What is a potential future trend in HPC that could revolutionize data science?

- A. Increased reliance on traditional CPUs
- B. Quantum computing**
- C. Elimination of parallel computing
- D. Replacement of all data science models with heuristics

## Summary

---

- ★ High-Performance Computing (HPC) is essential for large-scale data science applications.
- ★ Parallel and distributed computing techniques enable efficient data processing.
- ★ GPUs and frameworks like Apache Spark accelerate big data analysis.
- ★ HPC is used in AI, genomics, climate science, and finance.
- ★ Future trends include quantum computing and energy-efficient architectures.

# Q & A SECTION

**Please use this time to ask  
any questions relating to the  
topic, should you have any.**



Thank you  
for attending



HyperionDev