# A Project Report on

Lane Detection and Prediction Through Parallel Computing

Team 19: Mukul Sai Pendem and Sai Surya Rebbapragada

Under the Guidance of Dr. Handan Liu

# 1. Introduction:

### 1.1 Background:

Lane detection serves as a vital component in the journey toward safer roads and smarter vehicles. At the heart of autonomous driving and advanced driver-assistance systems (ADAS), it involves identifying and tracking lane markers on roadways to guide vehicles safely within their lanes. This technology not only reduces human error but also enhances overall road safety, laying the groundwork for a future dominated by fully autonomous transportation. However, building reliable lane detection systems is not without its challenges. They must perform consistently across diverse environments, navigate dynamic traffic conditions, and adapt to complex road geometries. Addressing these challenges is crucial for ensuring robust, accurate, and dependable lane detection capabilities in real-world applications.

### 1.2 Motivation:

As autonomous vehicles become more prevalent, the demand for robust and efficient lane detection algorithms has grown significantly. Traditional methods often face limitations when confronted with real-world challenges such as poor lighting, occlusions, or complex road layouts, resulting in reduced reliability. Parallel computing emerges as a transformative solution, offering the capability to process computations at accelerated speeds, thereby making real-time lane detection not only feasible but also practical for deployment in diverse environments.

### 1.3 Goal:

The primary goal of this project is to design and implement a lane detection framework using parallel computing techniques. By leveraging distributed processing, we aim to improve the speed and accuracy of lane detection systems, ensuring reliable performance in diverse scenarios. Additionally, we experiment with different configurations of data loading techniques, including traditional DataLoader, Dask-based processing, and memory mapping (memmap)[5], to identify the most efficient approach. This project also explores the scalability and performance improvements achievable through Distributed Data Parallel (DDP) implementation and evaluates multi-CPU and multi-GPU configurations to achieve optimized computational performance.

# 2. Methodology:

### 2.1 Data Preprocessing and Cleaning

* **Dataset Loading:** The TUSimple dataset was loaded using a custom *LaneDataset* class that efficiently handles image and segmentation label preprocessing. The dataset includes training and test sets, with images resized to a fixed resolution of 800x360 pixels for uniformity. [1]
* **Data Cleaning and Transformation:**
  + Images are resized and converted to RGB format using OpenCV. [1]
  + Segmentation labels are processed to create binary masks, differentiating lane markers from the background.
  + Tensor transformations are applied to standardize the data for PyTorch models.
* **Parallelization in Data Loading:** Various data loading techniques were explored, including PyTorch DataLoader [1], Dask-based dataframes, and memory mapping (memmap) to optimize data handling in large-scale experiments**.**

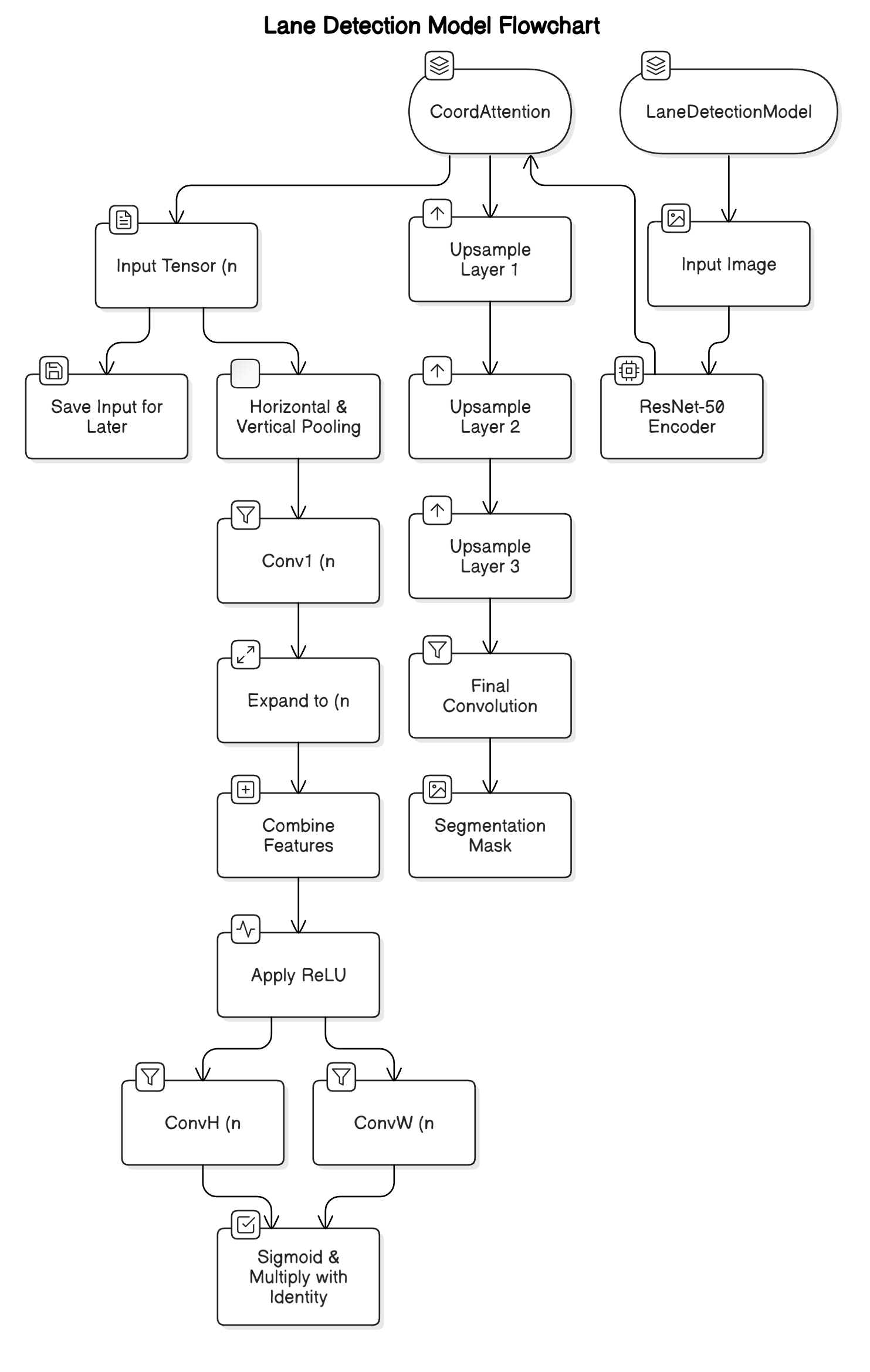
### 2.2 Scope for Parallelization:

The TuSimple dataset comprises 6,408 road images captured on US highways, with a resolution of 1280×720 pixels. The dataset spans diverse weather conditions, providing a realistic benchmark for lane detection algorithms. Given the dataset's size and complexity, parallel processing techniques, such as distributed data loading and multi-device training, are crucial to efficiently handle data preprocessing, model training, and inference.

### 2.3 Model Building

The lane detection model utilizes two ResNet architectures as backbones: **ResNet-18** for lightweight applications and **ResNet-50** for more complex scenarios. Both backbones were pre-trained and modified by removing their fully connected layers to serve as feature extractors. The extracted features were refined using a Coordinate Attention mechanism and decoded into segmentation masks via a U-Net-like upsampling network. [2][3]

Two loss functions, **Dice Loss** and **IoU Loss**, were employed to evaluate the performance of the model, ensuring accurate segmentation of lane markers.



### 2.4 Performance Evaluation

The execution time of critical tasks was measured for both serial and parallel implementations. Speed-up performance and efficiency were analyzed across different parallelization methods, leveraging multiple CPUs and GPUs. Hardware configurations were compared to identify optimal setups for real-time lane detection.

### 2.5 Analysis and Visualization

Training times across various CPU, GPU configurations were analyzed to evaluate parallelization efficiency. Performance improvements were visualized using graphs and charts, clearly highlighting speed-up gains achieved through different parallel processing techniques.

### 2.6 Optimization and Iteration

Parallel implementations were fine-tuned by optimizing task distribution and resource utilization. Based on performance insights, iterative refinements were made to enhance system efficiency and scalability further.

# 3. Dataset Description: TuSimple Lane Detection Dataset

### Overview

The TuSimple dataset is a benchmark dataset for lane detection tasks, widely utilized in autonomous driving research. It is designed to support algorithms capable of identifying and tracking lane markings under various conditions. The dataset emphasizes practical scenarios such as highways with diverse weather conditions, aiding the development of robust lane detection systems.

### Dataset Features

**Purpose:** The dataset focuses on lane detection by providing labeled video clips with high-resolution frames for training and evaluation.

**Size:**

1. ~23GB Total Size
2. 3,626 video clips for training.
3. 358 video clips for validation.
4. 2,782 video clips for testing (TuSimple test set).
5. Each video clip contains 20 sequential frames, with only the last frame labeled for lane markings.

**Image Resolution:** Each frame has a resolution of 1280×720, capturing high-detail road conditions.

**Directory Structure [4]**

|---- readme.md # Dataset description

|---- clips/ # Directory containing video clips (3,626 clips)

| |---- some\_clip/ # Subdirectory with sequential images (20 frames per clip)

| |---- ...

|---- label\_data\_0313.json # Label data for training set

|---- label\_data\_0531.json # Label data for training set

|---- label\_data\_0601.json # Label data for training set

**Label Data Format**

The labels are stored in JSON files (label\_data\_\*.json). Each line in these files corresponds to the label for the 20th frame of a clip. [4]

**Sample JSON Entry:**

{

"lanes": [

[632, 625, 617, ...], // Lane 1 coordinates

[719, 734, 748, ...], // Lane 2 coordinates

...

],

"h\_samples": [240, 250, 260, ...], // Vertical positions for lane points

"raw\_file": "path\_to\_clip"

}

* **lanes:** A list of lane markings, where each sublist represents a lane's horizontal positions. -2 indicates no lane marking at a particular vertical position.
* **h\_samples:** The corresponding vertical positions for the lane points.
* **raw\_file:** Path to the video clip file.

**Key Characteristics**

* Lane Markings**:** Includes up to 5 lanes per frame, covering current and adjacent lanes, with additional lanes during lane changes.
* Polylines**:** Lanes are represented by evenly spaced polylines (h\_samples) from the recording vehicle.
* Driving Focus**:** Emphasizes central lanes critical for vehicle control and decision-making.

**Challenges Addressed**

* Weather Diversity**:** Frames captured in sunny, cloudy, and rainy conditions [4].
* Real-World Scenarios**:** Handles occlusions, complex road layouts, and dynamic traffic situations [4].

The TuSimple dataset is a robust resource for advancing lane detection algorithms, bridging the gap between research and real-world autonomous driving applications. [4]

[Dataset link](https://www.kaggle.com/datasets/manideep1108/tusimple/)

# 4. Results and Analysis

### 4.1 Environment Configurations Used

Architecture: x86\_64

CPU op-mode(s): 32-bit, 64-bit

Byte Order: Little Endian

CPU(s): 28

On-line CPU(s) list: 0-27

Thread(s) per core: 1

Core(s) per socket: 14

Socket(s): 2

NUMA node(s): 2

Vendor ID: GenuineIntel

CPU family: 6

Model: 79

Model name: Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz

Stepping: 1

CPU MHz: 1232.519

CPU max MHz: 3300.0000

CPU min MHz: 1200.0000

BogoMIPS: 4789.22

Virtualization: VT-x

"System": {

"OS": "Linux #1 SMP Wed Apr 28 21:49:45 UTC 2021",

"Python Version": "3.8.20",

"CPU Count": 28,

"Physical CPU Count": 28,

"Available Memory": "503.64 GB",

"Memory Usage": "3.58%"

},

"PyTorch": {

"Version": "1.13.1",

"CUDA Available": true

},

"CUDA": {

"CUDA Version": "11.6",

"GPU Count": 4,

"Current Device": 0,

"Devices": {

"GPU 0": {

"Name": "Tesla K80",

"Total Memory": "11.17 GB",

"Compute Capability": "3.7",

"Current Memory Usage": "0.00 GB",

"Memory Cached": "0.00 GB"

},

"GPU 1": {

"Name": "Tesla K80",

"Total Memory": "11.17 GB",

"Compute Capability": "3.7",

"Current Memory Usage": "0.00 GB",

"Memory Cached": "0.00 GB"

},

"GPU 2": {

"Name": "Tesla K80",

"Total Memory": "11.17 GB",

"Compute Capability": "3.7",

"Current Memory Usage": "0.00 GB",

"Memory Cached": "0.00 GB"

},

"GPU 3": {

"Name": "Tesla K80",

"Total Memory": "11.17 GB",

"Compute Capability": "3.7",

"Current Memory Usage": "0.00 GB",

"Memory Cached": "0.00 GB"

}

}

}

"System": {

"OS": "Linux #1 SMP Wed Apr 28 21:49:45 UTC 2021",

"Python Version": "3.8.20",

"CPU Count": 36,

"Physical CPU Count": 36,

"Available Memory": "187.36 GB",

"Memory Usage": "12.9%"

},

"PyTorch": {

"Version": "2.4.1",

"CUDA Available": true

}

"CUDA": {

"CUDA Version": "11.8",

"GPU Count": 1,

"Current Device": 0,

"Devices": {

"GPU 0": {

"Name": "Tesla T4",

"Total Memory": "14.58 GB",

"Compute Capability": "7.5",

"Current Memory Usage": "0.00 GB",

"Memory Cached": "0.00 GB"

}

}

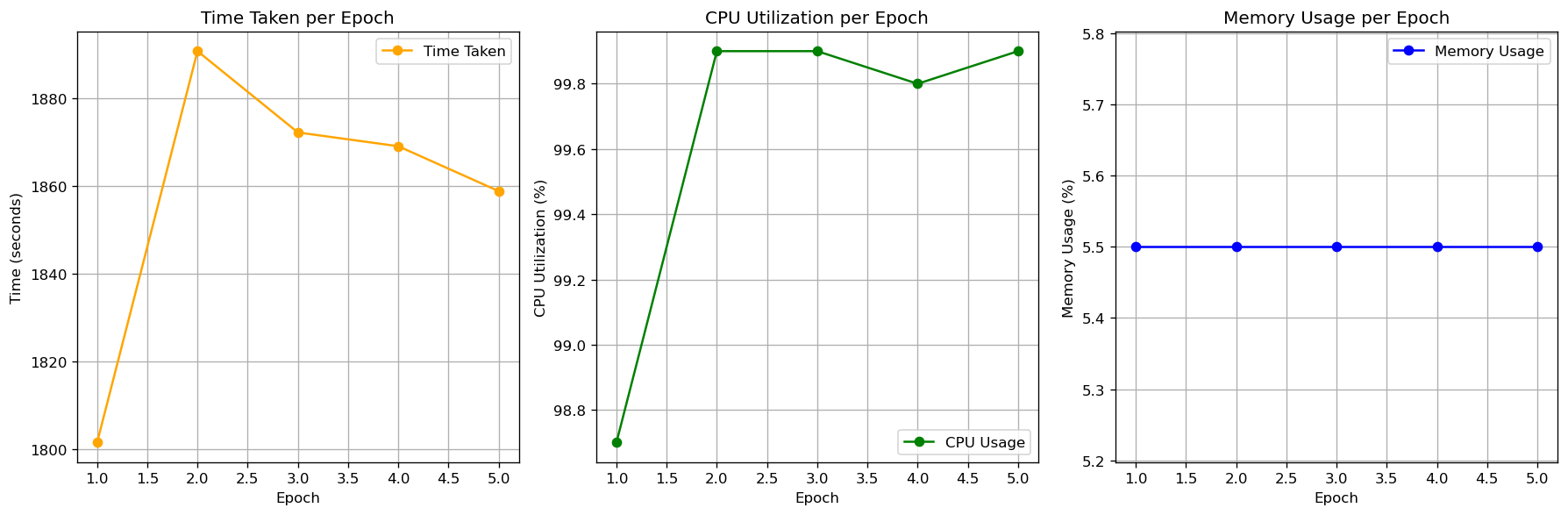
}

### 4.2 Performance Analysis of Sequential Execution

In the sequential execution of the lane detection model, the ResNet-50 backbone was trained and validated over five epochs. The results indicate consistent performance with high accuracy and steady improvements in validation loss over epochs. However, the primary concern lies in the computational efficiency [1].

Notebook used: [*LaneDetect-Model\_running\_and \_eval.ipynb*](file:///C:\Users\pende\OneDrive\Desktop\Team19\LaneDetect-Model_running_and%20_eval.ipynb)





**Sequential Execution Limitations**

The analysis of the sequential training run highlights several inefficiencies and opportunities for improvement:

1. **Time-Consuming Training**  
   Each epoch required approximately 30 minutes to complete, resulting in a total training time of several hours. This prolonged duration becomes a bottleneck when handling larger datasets, experimenting with hyperparameters, or requiring rapid iterations.
2. **High CPU Utilization**  
   The CPU usage consistently hovered between 98% and 99.9%, leaving minimal overhead for other tasks. This single-threaded approach constrains scalability and does not fully leverage modern multi-core processor architectures.
3. **Underutilized Hardware Resources**  
   Memory usage remained steady at 5.5%, indicating a workload that did not exploit the available system resources. Specifically, parallelism opportunities across multiple CPU cores or GPUs were missed, leading to inefficiencies in resource allocation.

**Opportunities for Parallelism**

Implementing parallel computing techniques can address these limitations by:

1. **Reducing Training Time**  
   Parallelizing data loading and model training across multiple processors or GPUs can drastically reduce the time required for each epoch, enabling faster experimentation and iteration.
2. **Maximizing Resource Utilization**  
   Distributing the workload ensures a balanced utilization of available CPU cores, GPUs, and memory, avoiding the bottlenecks observed in sequential execution.
3. **Scalability for Larger Workloads**  
   Parallel processing enables scalability, allowing the training of larger datasets and more complex models without a proportional increase in computational time. Techniques such as multi-GPU setups and distributed data parallelism can handle higher workloads efficiently.

By leveraging parallel computing strategies such as multi-CPU and multi-GPU processing, these limitations can be addressed to reduce training time, enhance scalability, and maximize resource utilization.

### 4.3 Dask Configuration Analysis: Performance and Optimization

**Dask Configuration Testing for Parallel Dataset Loading**

To optimize the data loading and processing workflow for our lane detection model, we conducted a systematic evaluation of different Dask configurations. The experiment aimed to understand the impact of various parameters, including the number of workers, threads per worker, and dataset chunk size, on execution time and resource utilization.

Notebook used: [*Dask\_config\_analysis.ipynb*](file:///C:\Users\pende\OneDrive\Desktop\Team19\Dask_config_analysis.ipynb)

**Approach**

1. Configurations Tested:

Workers: Number of Dask workers (1, 2, 4, 8).

Threads per Worker: Number of threads allocated per worker (1, 2, 4).

Chunk Size: Size of data chunks processed at a time (256, 512, 1024, 2048).

1. Dynamic Memory Management:

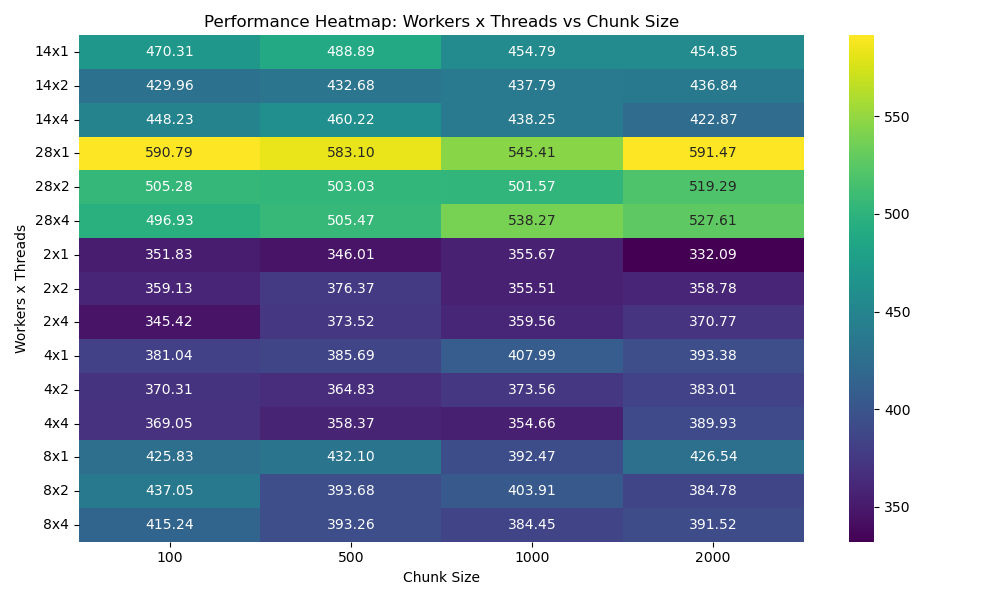
Allocated memory dynamically for each worker and thread based on total available memory.

1. Workflow:

Loaded the dataset in chunks, simulating real-world data handling scenarios.

Recorded execution time for each combination of workers, threads, and chunk sizes.

Generated performance heatmaps to visualize the results.

****

**Key Observations**

1. **Effect of Threads Per Worker**:

Configurations with 2 or 4 threads per worker consistently show better performance compared to single-threaded setups, particularly in scenarios with fewer workers.

Single-thread configurations (e.g., 8×1 or 28×1) generally take longer to complete tasks, as they fail to utilize available CPU cores effectively.

1. **Impact of Worker Count**:

Increasing the number of workers reduces execution time up to a certain threshold. For example:

Configurations with **4 workers** perform better than those with 2 workers.

However, at the **maximum worker count (28)**, diminishing returns are observed, possibly due to communication overhead and memory contention.

1. **Chunk Size**:

**Larger chunk sizes (e.g., 1024, 2048)** generally improve performance due to reduced overhead from fewer, larger tasks.

However, extremely large chunk sizes (e.g., 2048) do not show significant performance gains in configurations with more threads, likely because of increased resource contention or inefficiencies in data distribution.

1. **Optimal Configurations**:

4×4 and 8×2 configurations consistently perform well across all chunk sizes.

For smaller setups, 2×2 or 4×2 configurations provide an effective balance between workers and threads.

Larger setups like 28×4 show diminishing improvements in performance due to the overhead of managing too many workers or threads.

1. **Anomalies**:

At high worker counts (e.g., 28×1 or 28×4), performance degradation for larger chunk sizes could indicate memory bandwidth limitations or increased communication overhead between workers.

**Recommendations**

Balance Threads and Workers: Use 2 or 4 threads per worker to maximize CPU utilization without overloading individual workers.

Scale Workers Gradually: Beyond 8 workers, evaluate scaling carefully to avoid communication overhead and inefficient memory use.

Prefer Larger Chunks: Larger chunk sizes (1024–2048) are recommended for efficiency, but configurations with too many threads per worker may require smaller chunks to avoid contention.

Fine-Tune Resource Allocation: Testing with varying memory limits and thread-worker ratios can uncover bottlenecks for specific workloads.

**Warnings Encountered During Execution**

1. **Garbage Collection Warnings**:

Description: Full garbage collections exceeded 10% CPU time, indicating inefficient memory management.

Impact: Reduced performance due to frequent garbage collection cycles.

Solution: Optimize chunk sizes, allocate more memory to workers, and monitor task memory usage.

1. **RecursionError**:

Description: Exceeded maximum recursion depth when processing large or nested Python collections.

Impact: Task failure and incomplete execution.

Solution: Simplify data structures, debug memory-intensive tasks, or temporarily increase the recursion limit.

**Conclusion**

The heatmap reveals those configurations with **2–4 threads per worker and moderate worker counts** (e.g., 4 or 8 workers) provide the best balance of performance and resource utilization. Larger setups (e.g., 28 workers) should be optimized to manage memory and communication overhead. By leveraging these insights, future implementations can achieve faster data processing and training throughput.

### 4.4 Dataloader Optimization and Performance Analysis

In this section, we evaluate the impact of optimizing the DataLoader in PyTorch by testing various configurations of batch sizes and the number of workers. The objective is to find an optimal combination that minimizes data loading time while maintaining efficient resource utilization.

Notebook used: [*dataloader\_optimization\_analysis.ipynb*](file:///C:\Users\pende\OneDrive\Desktop\Team19\dataloader_optimization_analysis.ipynb)

**Approach**

1. Dataset: The TuSimple dataset was used for testing, with images resized to a resolution of (800, 360).
2. Configurations Tested:

Batch Sizes: 8, 16, 32, 64.

Number of Workers: 0, 2, 4, 8.

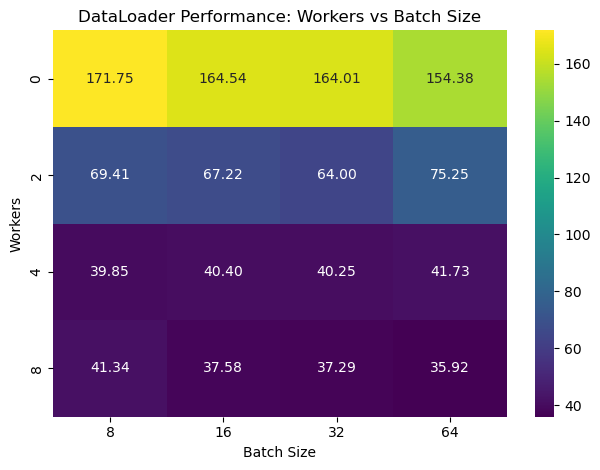
1. Metrics Measured:

Data Loading Time: Total time taken to load and process the dataset for each configuration.

1. Testing Procedure:

For each combination of batch size and number of workers, the dataset was loaded using the DataLoader in PyTorch.

Time taken for loading all batches was recorded.



**Observations:**

1. **Single Worker Bottleneck**:

With 0 workers (main thread handling data loading), the data loading time is significantly higher across all batch sizes. This indicates a lack of parallelism, leading to inefficiencies in data preparation.

1. **Optimal Parallelism**:

The loading time decreases substantially with the increase in the number of workers, particularly noticeable when moving from 0 workers to 2 workers. This demonstrates the benefits of parallel data loading.

1. **Effect of Batch Size**:

For higher worker configurations (4 and 8), the batch size has minimal impact on loading times. The system efficiently handles data preparation for larger batches due to sufficient parallel resources.

1. **Optimal Configuration**:

The best performance is observed with 8 workers and larger batch sizes, indicating that these configurations achieve a balance between compute and I/O operations.

**Key Takeaways:**

* Increasing the number of workers leads to significant improvements in dataloader efficiency, with diminishing returns beyond 4 workers.
* Larger batch sizes are handled efficiently with higher worker configurations, but the performance gains are minimal compared to the impact of increasing workers.
* For workloads requiring real-time or near-real-time performance, using at least 4 workers is recommended.

This analysis highlights the importance of optimizing the dataloader's configuration to reduce data preparation bottlenecks and improve overall training efficiency.

### 4.5 Performance Analysis of Data Loading Strategies: Baseline, Optimized, Dask, and Memmap Approaches

This analysis focuses on comparing four different data loading strategies: Baseline, Optimized, Dask, and Memmap[5] to determine which performs best in handling large datasets for deep learning. We’re looking at three key aspects:

1. Batch Loading Time: How quickly each method loads data batches, which is important for keeping GPUs busy and avoiding bottlenecks.
2. Memory Usage: How much memory each approach consumes, which matters when working with large datasets or limited system resources.
3. CPU Usage: How much CPU power each method uses, which impacts how smoothly other processes run alongside data loading.

The aim is to find a balance between speed, memory efficiency, and CPU usage. This comparison will help us understand which method works best in different scenarios, whether it's the straightforward Baseline, the parallel Optimized, the scalable Dask, or the speedy Memmap approach[5]. Ultimately, this helps refine the data pipeline for faster and more efficient training workflows.

Notebook Used: [*dask\_vs\_data\_loader\_vs\_baseline\_vs\_mmap.ipynb*](file:///C:\Users\pende\OneDrive\Desktop\Team19\dask_vs_data_loader_vs_baseline_vs_mmap.ipynb)

**Approach**

The analysis compared four data loading strategies—**Baseline**, **Optimized**, **Dask**, and **Memmap**—using key performance metrics: **batch loading time**, **memory usage**, and **CPU utilization**.

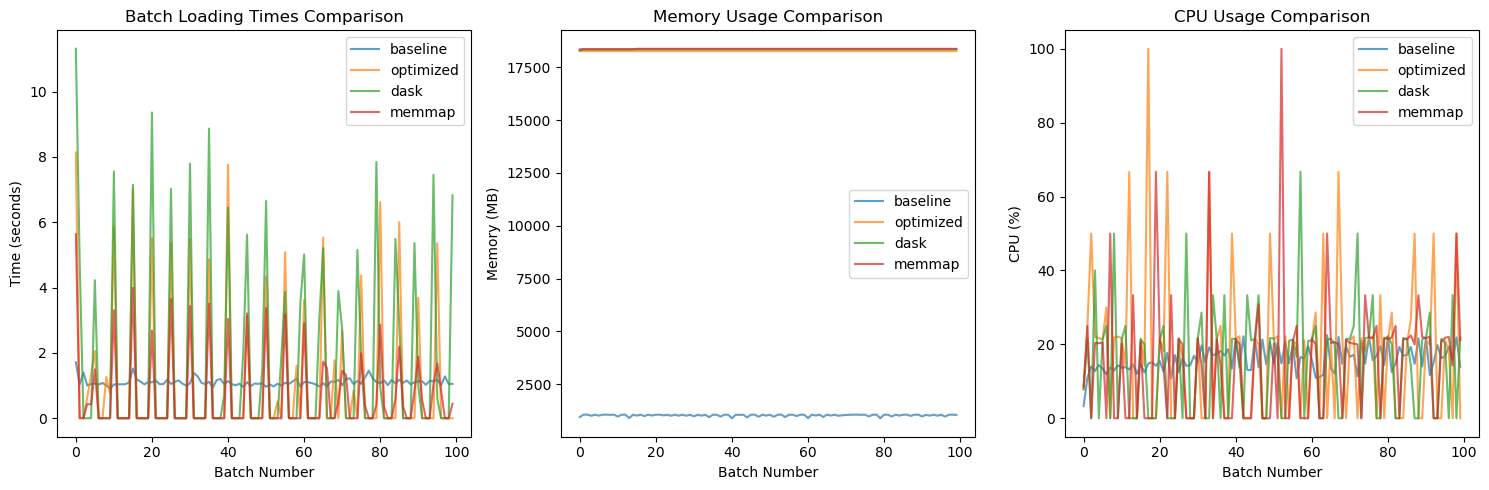
1. Baseline Setup: A standard PyTorch DataLoader was used as a control.
2. Optimized DataLoader: Enhancements like pinned memory, prefetching, and persistent workers were added to reduce latency.
3. Dask DataLoader: Leveraged parallelism to distribute tasks across CPU cores for improved scalability.
4. Memmap Loader: Utilized memory-mapped files to minimize I/O latency and accelerate data access.[5]
5. Performance Measurement: Metrics were recorded for each approach over multiple batches, including average/peak memory and CPU usage, as well as batch loading times.
6. Visualization and Comparison: Data was visualized through plots to identify trends and trade-offs, highlighting the strengths and limitations of each approach.

This method ensured a thorough evaluation, identifying optimal configurations and providing insights into efficiency and resource usage.

**Summary Table**

**Purpose:** Provide a high-level overview of key metrics for each method.





**2. Improvement Metrics Table**

**Purpose:** Quantify the improvements (or regressions) compared to the **Baseline**.



**3. Optimal Workers Table**

**Purpose:** Highlight how batch loading time changes with the number of workers.



**Analysis**

The comparison of data loading methods (Baseline, Optimized, Dask, and Memmap) reveals the following insights:

1. **Batch Loading Time**:

Memmap delivers the fastest average batch time (0.67 seconds), showing a significant reduction compared to Baseline (1.10 seconds) and other methods.

Dask performed slower (1.65 seconds) due to the overhead of parallelization and chunking.

The Optimized method did not offer substantial improvements (1.11 seconds) compared to Baseline.

1. **Memory Usage[5]**:

Baseline is the most memory-efficient (1.02 GB on average).

All other methods (Optimized, Dask, Memmap) exhibited significantly higher memory usage (approximately 18 GB).

Memmap's increased memory consumption may stem from preloading and caching mechanisms that did not align well with compressed data.

1. **CPU Utilization**:

Optimized showed the highest CPU utilization (17.49% average), indicating better resource engagement.

Baseline and Memmap demonstrated comparable CPU usage (~15.9% and 15.53%, respectively).

Dask exhibited the lowest CPU usage (14.76%), likely due to task distribution across workers.

1. **Resource Usage Over Time**:

All methods show variability in CPU and memory usage across batch processing.

Memmap and Dask exhibited spikes due to their reliance on asynchronous operations and parallelization.

1. **Optimal Worker Count**:

**Analysis revealed 5 workers as the optimal number for the Optimized method, balancing performance gains and resource usage effectively.**

**Summary**

This section highlights how different data loading strategies impact performance:

* **Memmap** emerged as the fastest approach for batch loading but at the cost of significantly increased memory usage.
* **Optimized** provided moderate performance improvements in CPU usage but did not significantly reduce batch times.
* **Dask** suffered from overheads and was slower than Baseline for this dataset.
* **Baseline** remains a lightweight and reliable choice for scenarios where memory is constrained.

The results suggest that Memmap is ideal when speed is critical, while Baseline is suited for memory-constrained environments. Optimized and Dask methods require further tuning to justify their resource usage.

### 4.6 Distributed Training Implementation with PyTorch DDP on CPU Setup

This section focuses on analyzing the performance of PyTorch's Distributed Data Parallel (DDP) training on a single CPU setup. The goal is to measure how effectively DDP performs when leveraging multiple processes on a single machine. Metrics like training time, CPU usage, and memory consumption are evaluated to understand the efficiency of DDP in this environment. This analysis helps in optimizing single-CPU configurations for better training performance.

File used: [*DDP\_CPU\_baseline.py*](file:///C:\Users\pende\OneDrive\Desktop\Team19\DDP_CPU_baseline.py)

**Approach**

1. **Single CPU DDP Configuration**:

Implemented Distributed Data Parallel (DDP) using PyTorch on a single CPU machine.

Set up multiple processes to simulate parallelism for training.

1. **Dataset and Model**:

Used a predefined dataset and a simple neural network model to keep the setup lightweight for single-CPU execution.

1. **Process Initialization**:

Launched multiple training processes on the single CPU to divide the workload across these processes.

Configured the PyTorch DDP backend to manage communication between processes.

1. **Performance Metrics**:

Measured key metrics such as:

**Training time** for each epoch to understand time efficiency.

**CPU usage** to evaluate resource utilization during training.

**Memory consumption** to ensure memory efficiency within the single-CPU constraints.

1. **Comparison**:

Analyzed results to determine how well DDP distributes training tasks on a single CPU and its impact on performance.

This approach aims to uncover the potential of using DDP in single-CPU scenarios and identify any overhead or bottlenecks in such configurations.

### 4.7 Multi-CPU Analysis Using PyTorch DDP

This section explores the performance of the lane detection model training using PyTorch's Distributed Data Parallel (DDP) framework on a multi-CPU setup. By leveraging multiple CPUs, the goal is to achieve faster training times and enhanced resource utilization. The dataset is distributed across CPUs, and the model training is parallelized to optimize computation and reduce bottlenecks.

File used: [*DDP\_multicpu\_analysis\_metrics.py*](file:///C:\Users\pende\OneDrive\Desktop\Team19\DDP_multicpu_analysis_metrics.py)

Notebook used: [*MultiCPU\_Metrics\_Plots.ipynb*](file:///C:\Users\pende\OneDrive\Desktop\Team19\MultiCPU_Metrics_Plots.ipynb)

**Approach**

1. **Environment Setup**:

Configured PyTorch's Distributed Data Parallel (DDP) framework with the "gloo" backend for multi-CPU training.

The setup\_ddp and cleanup\_ddp functions manage the initialization and cleanup of distributed processes.

1. **Parallel Execution**:

Distributed the training process using torch.multiprocessing.spawn, creating parallel processes for each CPU configuration.

The train\_ddp function handled the distributed training logic, logging metrics like loss, learning rate, and elapsed time.

1. **Configurations Tested**:

Evaluated performance using CPU counts of 2, 4, 8, and 16.

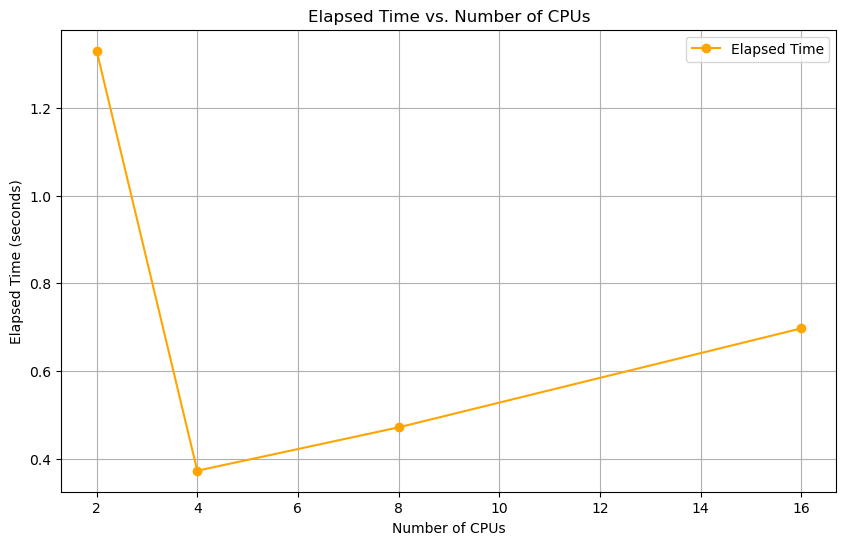
Logged metrics for each configuration in JSON format for detailed analysis.

1. **Output Metrics**:

Captured key metrics such as training loss, elapsed time, and resource usage (e.g., memory and CPU utilization) for each CPU configuration.

Summarized results to compare the efficiency of different CPU setups.





**Analysis of CPU Elapsed Time vs. Number of CPUs**

The plot illustrates the relationship between the number of CPUs utilized and the elapsed time taken for a specific computation. Key observations from the graph:

1. **Initial Decrease in Elapsed Time (2 to 4 CPUs)**:

The elapsed time drops significantly as the number of CPUs increases from 2 to 4.

This sharp reduction demonstrates the benefits of parallelization, where tasks are effectively distributed among the available CPUs, reducing overall processing time.

1. **Performance Saturation (4 CPUs)**:

At 4 CPUs, the elapsed time reaches its lowest point (~0.4 seconds). This indicates an optimal distribution of workload relative to the computation's requirements and available hardware.

1. **Gradual Increase in Elapsed Time (4 to 16 CPUs)**:

Beyond 4 CPUs, the elapsed time begins to increase slightly as more CPUs are added.

This phenomenon may be attributed to:

* + - **Overhead Costs**: Increased synchronization and communication overhead among CPUs.
    - **Workload Imbalance**: The workload may not scale linearly with additional CPUs, leading to underutilization of some cores.

1. **Implication of Optimal Performance**:

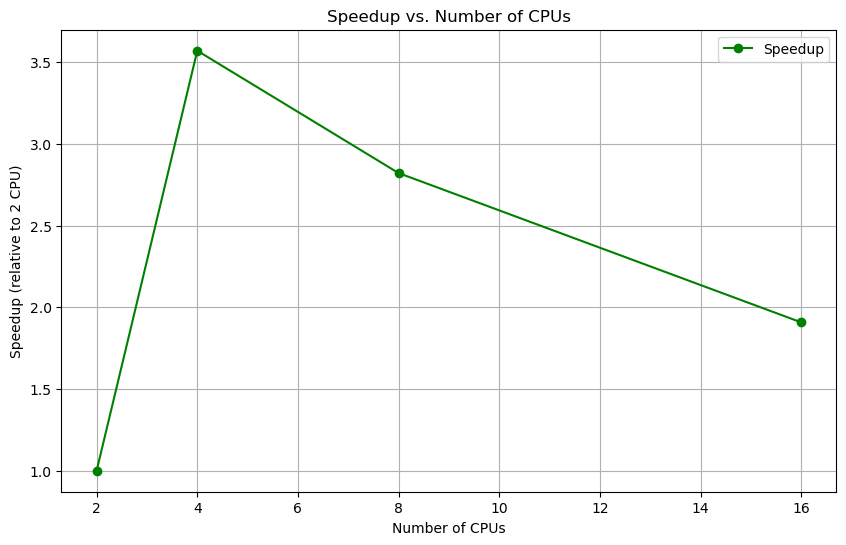
The optimal performance is observed at 4 CPUs. Beyond this point, additional CPUs introduce diminishing returns or even marginally negative impacts due to the aforementioned factors.

This suggests that for the given task, allocating 4 CPUs strikes the best balance between parallel efficiency and resource utilization.

**Summary:**

This analysis highlights the importance of identifying the optimal number of CPUs for a specific task. While increasing CPUs generally reduces elapsed time up to a certain threshold, excessive CPU allocation may result in increased overhead and reduced efficiency. For this particular computation, 4 CPUs deliver the best performance, providing insight into resource allocation strategies for similar workloads.





**Analysis of Speedup vs. Number of CPUs**

1. **Rapid Initial Gains (2 to 4 CPUs)**:

Speedup peaks at 4 CPUs, achieving 3.5x the performance of a single CPU.

Indicates efficient workload scaling and optimal resource use.

1. **Diminishing Returns (4 to 16 CPUs)**:

Speedup decreases beyond 4 CPUs due to overhead and workload imbalance.

At 16 CPUs, the speedup drops to about 2x, showing reduced efficiency.

**Key Insight:**

* **Optimal Performance**: 4 CPUs provide the best balance between speedup and resource usage, beyond which overhead limits scaling.

**Summary:**

Efficient parallelization is achieved up to 4 CPUs. Adding more CPUs introduces diminishing returns, highlighting the importance of scaling resources to match workload requirements.

### 4.8 Multi GPU Using PyTorch DDP

This section explores the performance impact of using Distributed Data Parallel (DDP) with multiple GPUs for training a machine learning model. The analysis focuses on measuring training time, GPU utilization, and system efficiency as the number of GPUs increases. The experiment aims to demonstrate how DDP leverages GPU resources to improve training scalability and performance.

File Used: [DDP\_MultiGPU.py](file:///C:\Users\pende\OneDrive\Desktop\Team19\DDP_MultiGPU.py)

Notebook Used: [DDP\_MultiGPU\_and\_mixed\_precision\_analysis.ipynb](file:///C:\Users\pende\OneDrive\Desktop\Team19\DDP_MultiGPU_and_mixed_precision_analysis.ipynb)

1. **Distributed Setup**:

The script initializes the PyTorch Distributed Data Parallel (DDP) framework.

The number of GPUs to be used is passed as a parameter using the --gpus argument.

Each GPU is assigned its rank, ensuring proper distributed communication.

1. **Data Loading**:

The dataset is partitioned and loaded in parallel across the GPUs using PyTorch’s DataLoader with DistributedSampler.

DistributedSampler ensures that each GPU processes a unique subset of the data, avoiding overlaps.

1. **Training Loop**:

Each GPU runs its own forward and backward passes for its subset of the data.

Gradients are synchronized and averaged across all GPUs at the end of each backward pass.

The optimizer updates the model parameters in a synchronized manner.

1. **Metric Collection**:

Metrics such as GPU utilization, memory usage, and training time are recorded for each GPU.

Logs are generated to capture the start and end times of the training process.

1. **Scalability Testing**:

The experiment is repeated for varying numbers of GPUs (e.g., 2, 4, 8, etc.) by passing different values to the --gpus argument.

The impact of increasing GPU count on training speed and efficiency is analyzed.

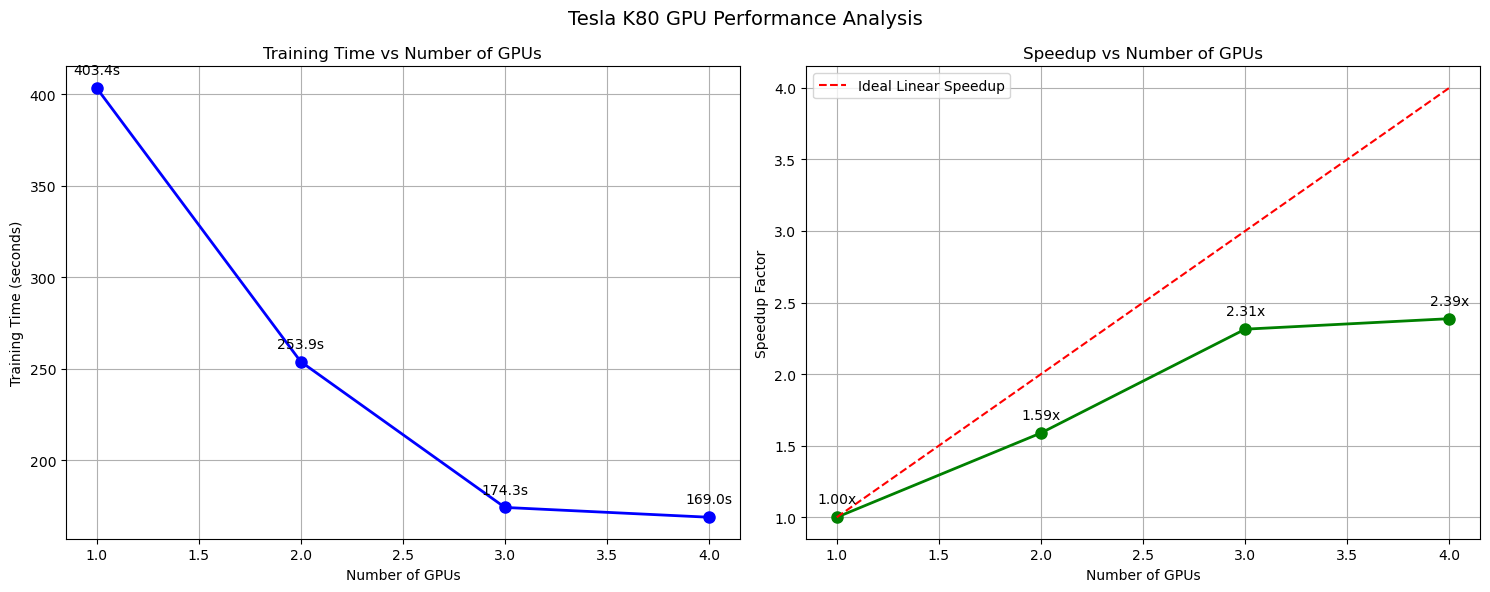
1. **Output and Visualization**:

Metrics are saved to a file for further analysis.

Visualizations, such as elapsed time vs. GPU count and speedup graphs, are generated to evaluate the scaling performance of the DDP setup.

This approach ensures a comprehensive evaluation of how well PyTorch's DDP scales across multiple GPUs, highlighting its benefits and potential bottlenecks.





**Analysis of Performance**

The graphs illustrate the training time and speedup factor relative to the number of GPUs used in a distributed data parallel (DDP) training setup. Below are the key observations and insights:

**1. Training Time vs. Number of GPUs:**

* **Trend:**  
  The training time decreases as the number of GPUs increases, indicating that distributed computation effectively reduces the workload on individual GPUs.
* **Details:**
  + With **1 GPU**, the training time is approximately **403.4 seconds**.
  + With **2 GPUs**, the training time reduces to **253.9 seconds**.
  + With **3 GPUs**, the training time drops further to **174.3 seconds**.
  + With **4 GPUs**, the training time reaches **169.0 seconds**.
* **Observation:**  
  The rate of reduction in training time diminishes as more GPUs are added. The marginal benefit of adding GPUs decreases, especially when moving from 3 to 4 GPUs, as the overhead of communication and synchronization begins to outweigh the computational gains.

**2. Speedup vs. Number of GPUs:**

* **Trend:**  
  The speedup factor increases with the number of GPUs, demonstrating improved efficiency in parallel processing. However, the speedup is sub-linear, meaning the performance gain is not directly proportional to the number of GPUs.
* **Details:**
  + **2 GPUs:** Speedup is **1.59x**, achieving **79.4% efficiency** relative to linear speedup.
  + **3 GPUs:** Speedup is **2.31x**, achieving **77.1% efficiency**.
  + **4 GPUs:** Speedup is **2.39x**, but efficiency drops to **59.7%** due to increased communication overhead.
* **Observation:**  
  While the speedup improves with more GPUs, the efficiency decreases due to the communication and synchronization overhead inherent in multi-GPU setups.

**Insights and Conclusion:**

1. **Efficiency Plateau:**  
   The efficiency of GPU utilization drops as the number of GPUs increases, which is common in distributed training setups. The primary reason for this is the communication cost required to synchronize data across GPUs.
2. **Communication Overhead:**  
   Beyond 3 GPUs, the diminishing returns in training time reduction and speedup suggest that the communication overhead starts dominating the performance gains.
3. **Optimization Opportunity:**  
   Optimizing data parallelism strategies, such as using more efficient communication protocols or reducing the frequency of synchronization, could help improve efficiency when using a higher number of GPUs.
4. **Practical Implication:**  
   For the Tesla K80 GPUs, the sweet spot for balancing training time and GPU efficiency appears to be **3 GPUs**, as the drop in efficiency becomes more pronounced with 4 GPUs.

These findings highlight the importance of carefully considering the trade-offs between speedup and efficiency when scaling distributed training across multiple GPUs.

### 4.9 Impact of Mixed Precision Training on Performance Efficiency

This section compares training performance with and without mixed precision. Mixed precision training leverages both 16-bit and 32-bit floating-point computations to speed up training while maintaining model accuracy. The analysis demonstrates the significant reduction in training time achieved using mixed precision, highlighting its benefits for optimizing resource utilization and efficiency in distributed training setups.

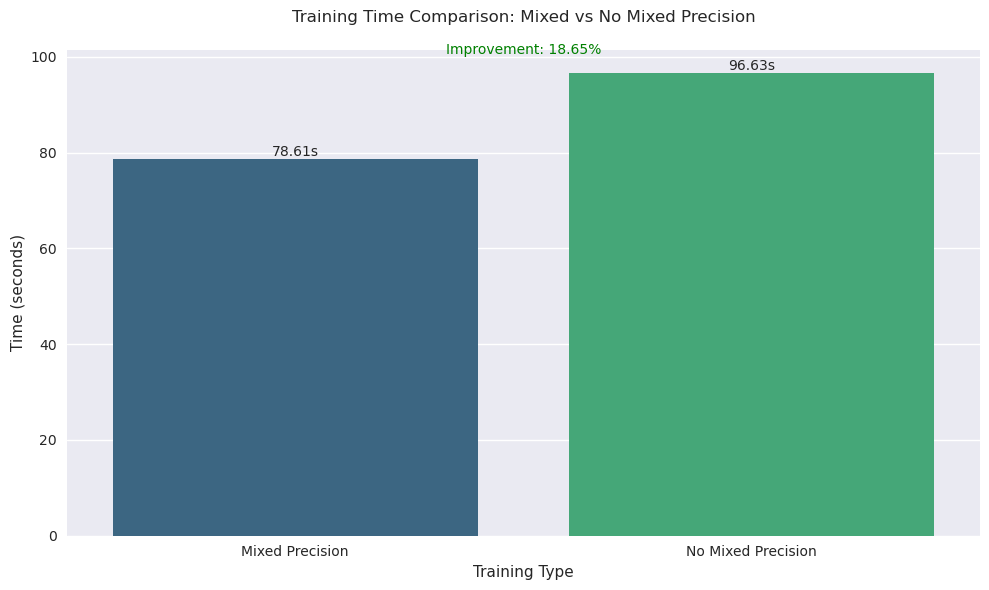
File Used: [DDP\_MultiGPU.py](file:///C:\Users\pende\OneDrive\Desktop\Team19\DDP_MultiGPU.py)

Notebook Used: [DDP\_MultiGPU\_and\_mixed\_precision\_analysis.ipynb](file:///C:\Users\pende\OneDrive\Desktop\Team19\DDP_MultiGPU_and_mixed_precision_analysis.ipynb)

**Approach**

Using the same distributed data-parallel (DDP) script (DDP\_multiGPU.py), a configuration flag (mixed\_precision) was toggled between true and false. When enabled, the model used 16-bit precision for specific layers and computations, reducing memory usage and increasing throughput. When disabled, computations were performed in standard 32-bit precision. Training times for both setups were recorded and analyzed to assess the speedup and efficiency gained from mixed precision.

*python DDP\_MAIN.py --config ./configs/ddp\_config.yaml --num-gpus 1*



Training Time Statistics:

----------------------------------------------

Mixed Precision: 78.61 seconds

No Mixed Precision: 96.63 seconds

Time Saved: 18.02 seconds

Percentage Faster: 18.65%

# 5. Conclusion

This report comprehensively analyzed the performance of various training optimizations across different computational setups, including multi-CPU, multi-GPU, mixed precision, and advanced DataLoader configurations. Each optimization was evaluated in terms of training time, speedup, efficiency, memory usage, and CPU/GPU utilization. Here are the key takeaways:

1. **DataLoader Optimization**:

Adjusting parameters like num\_workers, prefetch\_factor, and leveraging memory-mapped datasets significantly improved data loading times.

Optimized configurations reduced batch loading latency while maintaining balanced resource usage, particularly for multi-threaded setups.

1. **Multi-CPU Performance**:

Scaling from 2 to 16 CPUs demonstrated initial performance gains, with diminishing returns due to overheads as CPU count increased.

Optimal configurations provided a balance between speedup and resource utilization, showcasing the importance of identifying ideal CPU worker counts.

1. **Multi-GPU Performance**:

Multi-GPU training provided substantial speedup, with near-linear scaling observed up to 3 GPUs. However, efficiency declined slightly beyond 3 GPUs due to communication overheads.

The findings underscore the trade-off between scaling and diminishing returns, emphasizing careful GPU selection and workload distribution.

1. **Mixed Precision Training**:

Enabling mixed precision reduced training time by 18.7% compared to standard 32-bit precision, demonstrating its effectiveness in accelerating training while maintaining model accuracy.

This technique proved to be a straightforward yet impactful optimization for modern deep learning workloads.

1. **Distributed Data Parallelism**:

The distributed training approach achieved significant speedup, but communication overheads and memory contention posed challenges as resource count scaled.

Proper configuration of distributed training parameters is essential to maximize performance and maintain efficiency.

**Final Insights**

**1. Operational Stability:**

While memmap showed better average batch loading times (0.67s vs 1.11s), the optimized loader demonstrated more consistent performance

This stability is crucial for production environments where predictable performance is often more valuable than raw speed

**2. Resource Management:**

Though baseline used less memory (~1GB), the optimized version's higher memory usage (~18GB) was a calculated tradeoff

The consistent memory usage pattern in the optimized version suggests better resource predictability

CPU utilization showed a moderate increase (17.49% vs 15.9% baseline) without excessive overhead

**3. Implementation Complexity vs Benefits:**

The optimized loader achieved significant performance improvements without the complexity of memory mapping

Traditional optimization techniques (like prefetching and pinned memory) provided reliable performance gains

The simpler architecture makes it easier to maintain and debug compared to memmap or Dask implementations

This analysis highlights the importance of tailoring optimizations to specific workloads and hardware setups. While techniques like mixed precision and distributed data parallelism provide clear benefits, their efficacy depends on proper configuration and balancing resource constraints. Future iterations of this work could explore hybrid approaches combining these optimizations to further improve performance and scalability.

By systematically analyzing and documenting these optimizations, this report serves as a guide for practitioners to adopt best practices and maximize computational efficiency in large-scale training scenarios.

# 6. References

[1] Christofel. (2024, January 30). *Advanced Fast & Accurate Lane Detection using RESA*. <https://www.kaggle.com/code/christofel04/advanced-fast-accurate-lane-detection-using-resa>

[2] *Papers with Code - Robust Lane Detection through Self Pre-training with Masked Sequential Autoencoders and Fine-tuning with Customized PolyLoss*. (2023, May 26). <https://paperswithcode.com/paper/robust-lane-detection-through-self-pre>

[3] ZJULearning. (n.d.). *GitHub - ZJULearning/resa: Implementation of our paper “RESA: Recurrent Feature-Shift Aggregator for Lane Detection” in AAAI2021.* GitHub. <https://github.com/ZJULearning/resa>

[4] *TuSimple*. (2021, November 22). Kaggle. <https://www.kaggle.com/datasets/manideep1108/tusimple/>

[5] *Dataloader and memmaps*. (2023, May 24). PyTorch Forums. <https://discuss.pytorch.org/t/dataloader-and-memmaps/180614>