Project Proposal for Data Compression Analysis in the ATLAS Experiment

Danielius Kundrotas
advisor: Tomasz Bold

Introduction

Experiments at CERN, especially those conducted with the Large Hadron Collider (LHC), generate enormous amounts of data. For example, the ATLAS experiment at the LHC produces about 10 petabytes of raw (detector signals) data annually from particle collisions. Managing and storing this data requires substantial financial and environmental resources. High-capacity storage systems and data centers must be employed, leading to significant operational costs. These systems consume vast amounts of electricity, contributing to carbon emissions and other environmental impacts.

Reducing the size of these datasets through efficient data compression techniques can mitigate these issues. Smaller datasets require less storage space, which in turn reduces the need for extensive physical infrastructure. This reduction lowers both the energy consumption of data centers and the costs associated with maintaining them. Moreover, less data being transmitted and processed can lead to further energy savings, thereby decreasing the overall carbon footprint.

By focusing on data compression, data-intensive experiments can make their operations more sustainable and cost-effective, aligning with global goals for environmental conservation and financial efficiency.

Takeaway message: CERN's experiments generate huge datasets, and reducing their size through compression lowers costs and environmental impact by decreasing storage needs and energy consumption.

Lossy Compression Techniques

Lossy compression techniques reduce data size by approximating and discarding some parts of the original data that are considered less critical. However, this process can lead to information loss, which may affect the integrity and usability of the data, especially in scientific research where precision is paramount.

Before implementing lossy compression, a thorough examination of the potential information loss must be conducted. This involves identifying the specific

types of data that can tolerate some loss without compromising the validity of the research findings. The examination should include:

1. **Quantitative Analysis:** Measuring the extent of data reduction and the corresponding loss in information content.
2. **Qualitative Analysis:** Assessing how the loss of certain data affects the overall conclusions drawn from the experiments.
3. **Error Tolerance Studies:** Determining acceptable error margins within which the scientific outcomes remain reliable.
4. **Impact Assessment:** Evaluating the potential risks and benefits of lossy compression on ongoing and future experiments.

By carefully examining these aspects, researchers can introduce lossy compression in the experiments in a controlled manner that minimizes negative impacts and ensures the integrity of the experimental data.

Takeaway message: Before using lossy compression, it's crucial to carefully examine potential information loss to ensure scientific data remains reliable and accurate.

Case Study: Online Tracking Performance at Hadron Colliders

At hadron colliders, with an overwhelming background of hadrons, the tracking of charged particles is needed for online data filtering. Their performance needs to be continuously monitored. Its performance is established with respect to the results of charged particle tracking performed with high precision offline (using much more time than in the online filter). The analysis leading to the evaluation of the online tracking performance will be used in a case study of lossy compression. It will involve:

1. **Baseline Performance Measurement:** Establishing a performance baseline using uncompressed data to understand the current accuracy and efficiency of the online tracking algorithms.
2. **Compression Implementation:** Applying various levels of lossy compression to the data and tracking the changes in performance metrics.
3. **Comparative Analysis:** Comparing the performance of the tracking system with compressed data against the baseline to identify any degradation in tracking accuracy and efficiency.
4. **Error Characterization:** Analyzing the types of errors introduced by compression and their impact on the tracking results.

5. **Optimization Strategies:** Developing methods to mitigate the negative effects of compression, such as algorithmic adjustments or selective data preservation.

This example study aims to provide empirical evidence on how lossy compression affects an example analysis, helping to make informed decisions on its broader application in experimental data.

Takeaway message: A study will be conducted to assess how lossy compression affects online tracking performance estimation, comparing compressed data against uncompressed benchmarks.

Detailed Expansion on Implementation Details

In the realm of high-energy physics, continuous advancements in data processing tools are crucial for managing the vast volumes of data generated by experiments. A new tool for tracking performance evaluation (Inner Detector Track Performance Monitor - IDTPM) that is currently under development will be utilized as a testbed for evaluating online filtering and tracking performance under different data compression scenarios. The tool is part of the ATLAS Athena software framework and is publicly available at the CERN gitlab repository: `https://gitlab.cern.ch/atlas/athena/-/tree/main/InnerDetector/InDetValidation/InDetTrackPerfMon`. Extension needed in order to study the impact of the compression are limited to a single file that implements matching function between online and offline reconstructed charge particle tracks.

This tool features:

- **Modular Design:** A flexible architecture that allows easy testing of various compression algorithms.
- **Performance Metrics:** A comprehensive set of performance plots indicating the impact of the compression on the performance analysis.
- **Adaptive Algorithms:** Incorporation of adaptive algorithms that can adjust filtering and tracking parameters based on the incoming data.

Using this new tool as a testbed will provide valuable insights into how lossy compression affects the evaluation of online filter tracking performance. The findings can guide further development and optimization of both the tool and the compression techniques used in data processing at the CERN ATLAS experiment. The use of the official tool would facilitate discussion with experts on the results of these studies.

Takeaway message: A tool that is being developed within the ATLAS software framework will be used to test how lossy compression impacts online filtering and tracking performance analysis results, helping refine both the tool and inform about the impact of potential data compression.

Project Timeline

June 20th - July 5th

Project Proposal and Familiarization with ATLAS Software

- Write the project proposal.
- Get familiar with ATLAS software and its documentation.

July 5th - July 15th

Setup and Environment Configuration

- Set up the working environment to work with the Inner Detector Track Performance Monitor (IDTPM).
- Install ATLAS software and understand the development process.

July 15th - August 1st

Building Understanding

- Understanding development process.
- Understanding IDTPM and evaluation of performance for uncompressed data.
- Evaluate the performance for uncompressed data as a baseline.

August 1st - August 10th

- Develop and introduce an example of a compression emulation.
- Apply various levels of lossy compression to the data.
- Track changes in statistical performance metrics to find the relationship between compression level and its impact on tracking accuracy and efficiency.

August 10th - August 25th

Optimization Strategies and Adaptive Compression

- Systematic study of compression effect.
- Development of adaptive compression scenario.

August 25th - September 5th

Presenting Results

- Creating YouTube lecture with context, code briefing and main results.
- Creating slides and preparing presentation for other participants.