Geospatial Operations Performance of Data Processing Units

Derda Kaymak
Department of Computer Science
Marquette University
derda.kaymak@marquette.edu

Satish Puri
Department of Computer Science
Marquette University
satish.puri@marquette.edu

ABSTRACT

It is becoming increasingly important to effectively handle the growing amount of data in various fields, including geospatial operations. One approach is to leverage specialized hardware called data processing units (DPUs) alongside high-performance computing. To explore this, we developed a benchmarking tool to measure the performance of geospatial operations on different hardware configurations. The tool was used to compare the results obtained from running the operations on CPUs and DPUs using different data and parameters. The study found that initially, the CPU outperformed the DPU, but performance could be improved by offloading certain tasks from the CPU to the DPU. By harnessing the parallel processing capabilities and optimized architecture of DPUs, offloading specific computations can enhance overall performance. However, it is important to consider workload characteristics, data properties, and the capabilities of the DPU when determining the effectiveness of offloading tasks. Continued research and benchmarking will help optimize the distribution of workloads and maximize the benefits of using DPUs for geospatial operations.

Keywords

Geospatial Operations; Data Processing Units (DPU); HPC

1. INTRODUCTION

.

The global volume of data is rapidly increasing, necessitating efficient processing methods. High-performance computing (HPC) plays a vital role in managing and analyzing large datasets. Improving communication and load balancing are key areas for enhancing HPC performance. Nvidia's BlueField-2 and BlueField-3 DPUs are recent hardware solution for optimizing communication and computation processes. In geospatial analytics, these processing units can offload CPU tasks and reduce memory redundancy, improving overall performance. However, performance evaluation is crucial due to the hardware's recent release and limitations. For this, we created a benchmarking tool to measure the performance BlueField-2 and BlueField-3 and compare it with existing processors. By running the tool with different datasets and parameters, the performance of processors can be assessed and analyzed. In this paper, the performance of various architectures including CPU and DPU will be compared. In addition, the features of the benchmarking tool created for performance evaluation will be shared.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

Dataset	Size	Records
Cemetery	56 MB	193 K
Sports	590 MB	1.8 M
Lakes	9 GB	8.4 M

Table 1: Datasets

2. EXPERIMENTAL SETUP

.

In the benchmarking tool, the input consists of two files containing geospatial data, Base and Query layers. Geospatial operations are then executed on these data, and the time taken for these operations is measured. The process includes creating an R-tree using the Base Layer, performing queries for each geometry in the Query Layer using the Minimum Bounding Rectangles (MBRs) from the R-tree, and executing specific operations (e.g., intersection, overlap, touch, equality, covering) on the candidate geometries obtained after filtering. The elapsed time encompasses all these steps. The algorithm of the benchmarking tool is shown in Figure 1.

The benchmark was tested on the Thor cluster of the HPC Advisory Council, which has 32 Intel Broad-well E5-2697A chipset as CPU nodes, and 32 BlueField-2 and 16 BleuField-3 devices as DPU nodes. For communication between similar architectures, the MPI library was employed. On the other hand, communication between the CPU and DPU was facilitated using gRPC.

For benchmarking experiments, we utilized cemetery, sports, and lakes data sourced from the UCR-STAR dataset to conduct geospatial operations. Information about these data is given in Table 1.

3. BENCHMARK RESULTS

.

In a series of experiments, different approaches were tested and compared:

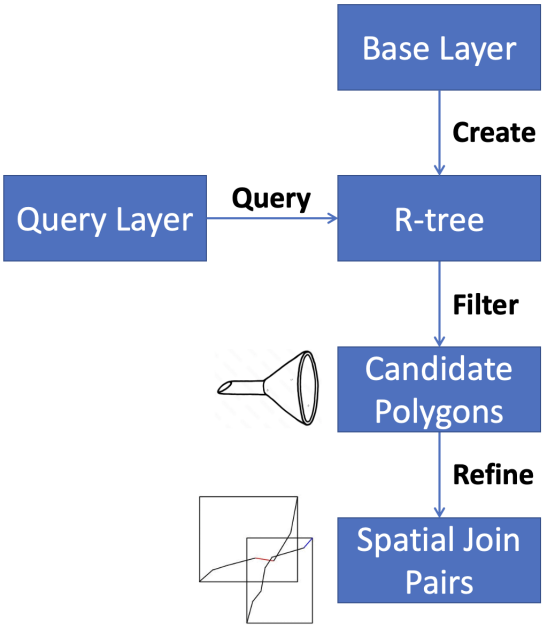

Figure 1: Geospatial Filtering and Refinement Algorithm

The base layer data was divided into 128 equal pieces, and the round robin method was used to test the performance. The CPU consistently outperformed the BlueField-2 DPU by approximately 2.7 times for the intersection operation. However, as the number of processes increased, BlueField-3 showed greater performance improvement compared to the CPU, indicating better scalability for the DPU. The results are shown in Figure 2. Also, using different dataset pairs for the base and query layers, the CPU exhibited around 2.7 times better performance than BlueField-2 and 1.9 times better performance than BlueField-3 for the Sports-Cemetery file pair. This performance advantage slightly increased with larger data pairs. In a similar experiment, the base layer data was divided according to the number of processes. The CPU-DPU performance ratio remained the same, but the performance increase was limited to 1.7 when doubling the number of processes. This limitation was due to the larger R-tree generated with less data splitting, leading to reduced query time complexity.

When the CPU and DPU were used together with dynamic load balancing, the processing time significantly improved. While the same process took about 76 seconds on BlueField-2 and approximately 28 seconds on the CPU, using both together reduced the processing time to around 16 seconds. The results are shown in Figure 3.

To explore performance with different numbers of processes and nodes, tests were conducted on BlueField-3. As shown in Figure 4, it was observed that employing multiple nodes provided better performance than using multiple processes on a single node.

4. CONCLUSIONS

.

Based on these experimental results, it is evident that the CPU generally possesses greater computing power compared to DPUs. However, particularly in the domain of geospatial

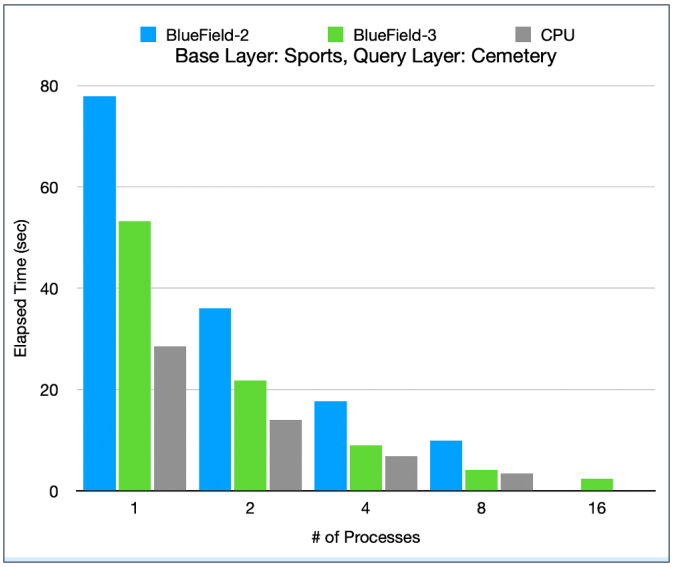

Figure 2: Intersects performance of single node using data divided into 128 partitions

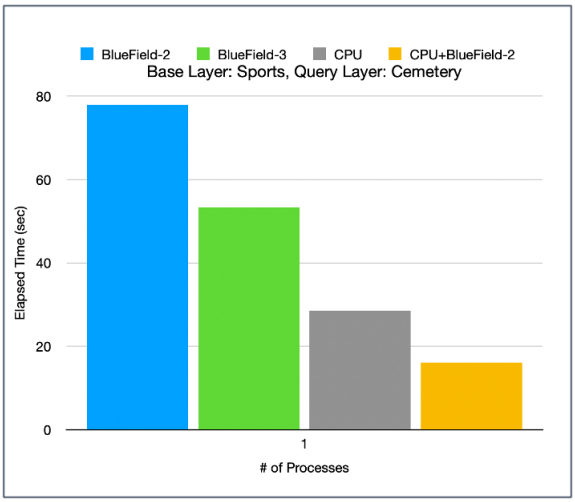

Figure 3: Intersects performance of different systems

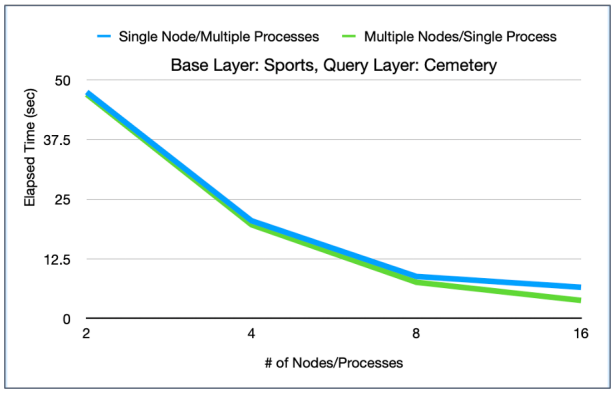

Figure 4: Performance difference between multiple nodes and multiple processes

data analytics, DPUs can enhance performance by offloading certain tasks from CPU.

5. REFERENCES