A Survey on Improved Performance of Database Related Operations using GPUs

Abstract

The growth of size in datasets and the demand for tools to perform scientific data-analysis have been rising in the recent years. Many database researchers are looking for ways to accelerate various database operations in order to maintain good or better performance with today’s larger datasets. A lot of the work to accelerate these operations has been done for parallel database systems that involve loosely-coupled clusters. However, with the arrival of the GPU, and the techniques offered by the General Purpose Graphics Processing Units (GP-GPUs), new opportunities are offered for the database community. Some researchers are starting to use these devices to improve database operations using the relatively inexpensive tightly-coupled system that the GPUs offer instead of the more expensive loosely-coupled systems found in clusters. These tightly coupled systems also offer additional advantages in that there is almost no overhead in creating threads and their main memory bandwidth is higher than a CPU’s. Many researchers in the database community have found ways to exploit this high thread-level parallelism and higher memory bandwidth in order to accelerate several database operations. This article presents an overview of the current techniques and algorithms that have been successfully implemented in GPUs in order to accelerate database operations.

Keywords: GP-GPU, high-performance in database operations, GPU, high thread-level parallelism.

# Introduction.

GPUs are both powerful and inexpensive devices that could be used in order to obtain more computational throughput. Not only can they perform faster than CPUs in number of float point operations per second but they also have a higher growth rate in performance. Over the last decade the growth rate of GPU performance has been higher than that of the CPUs. Figure [#x] shows how the NVIDIA’s GPU’s floating point performance has increased dramatically between the years 2003 and 2008, and how the gap in performance between the CPU and GPU has widened over those years. GPUs were primarily designed and used for performing algorithms around graphic computation, such as transforming, rendering and texturing geometric primitives such as triangles. However, many have sought techniques for using the GPU to perform tasks that are not related to graphics and that are typically handled by the CPU.



Figure . Taken from [#x NVIDIA Programming guide]

Many researchers view GPUs as today’s most powerful computational hardware for the dollar. Researchers and developers in areas not related to graphics have also become interested in harnessing this power for general-purpose computing. This effort is collectively known as GPGPU (for “General-Purpose computing on the GPU”). [#x:Survey Paer]

Two types of GPU programming languages have been used in the work discussed in this paper. There are graphic APIs such as DirectX, and OpenGL, and GPGPU languages such as CUDA. The graphic APIs process textures through a programmable hardware pipeline. To drive the computation vertices and pixels, processors are employed. These APIs allow one to directly utilize the hardware features related to rendering and visualization. In a lot of work that is presented, which was previous to the development of the GPGPU language CUDA, developers had to use the graphics APIs to map their programs to the graphics rendering mechanism, and so, such this work has been expressed in terms related to that programming. In order to clarify common terminology, an overview of important terms will be presented here. Often the terms kernels, and fragment programs can be used interchangeably, they both refer to a piece of code that we want to compute in parallel.

Even though GPUs are some of the fastest devices out there for computation, they are not suitable for many of the tasks that are commonly handled by the CPU. Currently, GPUs lack some useful computational constructs, such as support for various instructions. Also support for data types such as integers has recently arrived, but it has not been around for previous years of GPU research. They also do not commonly support 64-bit precision yet, and also sacrifice some performance in order to support it [3]. GPUs are also unable to replace CPUs in various types of applications where data dependency is so high that it is only possible to parallelize few tasks, or none of them. As a rule of thumb, if given an application one cannot create at least thousands of threads; there is no benefit in performance in using a GPU. Many applications are also dominated by both memory communication, and memory serialization effects associated with indirect memory addressing, which are also not a good match for GPUs. Finally, there are also some applications that computationally perform better in GPUs, but the cost of transferring the data from the CPU to the GPU would be greater than the benefit in performance obtained.

Most database operations are typically data intensive operations, and are not a good match for a GPU. However there are some of them that are also very computationally intensive and could benefit from the GPUs high-level parallelism. Among them are some operations that are performed in database queries, which could be accelerated when, streaming several queries of a dataset [#x]. Also some GPU-based sorting algorithms have better cost-effective performance than CPU-based algorithms [#x]. Join Operations are computationally expensive, and have been accelerated by using the GPU-based sorting algorithm in order to sort the records based on the join key [#x]. Finally query evaluation using indices may take advantage of the numerous threads in a GPU in order to examine several records of an index simultaneously in order to answer queries quickly [#x]. The rest of the paper will discuss each of these operations and approaches taken to implement them in more detail and organized as follows. Section 2 provides a brief overview of the basic database operations that can be performed efficiently on the GPU.

GPU-based algorithms perform computations on 2D arrays of 32-bit floating point data values known as textures. Each array element corresponds to a pixel. Pixels are transformed by programmable fragment processors, each executing the same fragment program on each pixel. The multiple GPU fragment processors perform data parallel computations on different pixel arrays simultaneously. This simple data-parallel architecture avoids write-after-read hazards while performing parallel computations.

**Database operations when performing data queries**

Govindaraju et al. presented some novel algorithms for accelerating various database operations on the GPU. They considered algorithms that aimed at operations such as conjunctive selections, aggregations, and semi-linear queries. These operations are often essential in database, data warehousing, and data mining applications. They took advantage of the inherent pipelining and parallelism of GPUs. The single instruction and multiple data (SIMD) capabilities, and vector processing functionalities were also useful.

Their algorithms took into consideration some of the limitations of the programming model of current GPUs and perform no data re-arrangements. They compared their performance with an optimized implementation of CPU-based algorithms and their experiments indicated that the graphics processor available on commodity computer systems is an effective co-processor for performing database operations.

The database operations they focused on mostly revolved around select queries which were of the following form:

SELECT **A**

FROM **T**

WHERE **C**

Here, **T** is the table that is being queried. **A** could be a list of attributes or aggregations such as SUM, COUNT, AVG, MIN, MAX defined on individual attributes. **C** is the condition of the query which is defined as a Boolean combination of predicates. A boolean combination is a set of predicates combined using relational operators such as AND, OR, EXIST, NOT EXIST among others. Predicates have the form op where op is the operator and it could be any of the following: .

Essentially predicates, boolean combinations and aggregations are considered the three different categories of database operations which could be accelerated using GPUs.

To speed up predicates in the form of they were evaluated via a depth test and stencil test. To use the depth test for performing comparisons, attribute values need to be stored in the depth buffer. They use a simple fragment program for copying the attribute values from the texture memory to the depth buffer.

Predicates in the form of, are first transformed into a semi-linear query which is then evaluated as a predicate with a constant. This algorithm performs the operations using the vector processing units on the GPUs.

To solve Boolean combinations in the GPU efficiently they are first rewritten in conjunctive normal form (CNF) omitting the NOT operators. Then a stencil test is used repeatedly for evaluating a series of logical operators recursively with the intermediate results stored in the stencil buffer.

The Aggregations are all implemented using the counting capability of the occlusion queries on GPUs. Because the aggregations Min and Max are considered special cases of the kth largest number problem, a single algorithm was generated.

The algorithms for the database operations in the predicates, boolean, and aggregation categories were implemented on a GPU and their performance was compared with optimized implementations of CPU-based algorithms. Tests were performed by testing different types of queries that would use one or more of the database operations described in the three different categories. Predicate evaluation queries only involved queries were a predicate was included in the condition. Boolean combination database operations were tested through range since these can be expressed as a boolean combination of two predicates. For more complex boolean combinations of predicates, queries with more than one relational operator, multi-attribute queries were used. Finally, the algorithm kth largest number was tested for queries that involved the MIN and MAX aggregations, and an accumulator implementation in the GPU using a mipmap texture was used in order to test the SUM and AVG aggregations. The results of these tests are summarized in the following table [x].

|  |  |  |
| --- | --- | --- |
| **Type of Query or Database Operation** | **Approximate Speed up** | **Approximate Speed up (Computational Time only)** |
| Predicate Evaluation | 3 | 20 |
| Range Query | 5.5 | 40 |
| Multi-attribute Query | 2 | 20 |
| Semi-linear query | 9 | not mentioned |
| kth largest number | 2 | 3 |
| Accumulator | 0.05 | not mentioned |

Overall they conclude in their research that algorithms for semi-linear queries and selection queries had the highest performance gain when implemented in GPUs. A medium performance gain was observed for the kth largest algorithm and no performance gain was observed for the accumulator algorithm which ended up being 20 times slower than its CPU-based implementation.

# Sort operations.

The Sort operation in databases or datasets transforms an unordered set of data into an ordered set of data with a given criteria. Different techniques have been developed to sort on a CPU. However, most of these techniques are not good candidates for a GPU implementation because most of them have data-dependencies difficult to get around. For a sorting algorithm to be GPU-friendly, it has to be oblivious to the input data.

Algorithms for sorting on GPUs are based on sorting networks. A sorting network consists of a network of wires and comparator modules which are used to sort a sequence of numbers. Each comparator will connect two wires and compare two values. It will then output the smaller value to one wire and the larger value into another wire. Sorting networks sort input data in a fixed number of steps for any given data, meaning they are independent of the input data. They also have a fixed communication pattern, making gather operations possible (gather operations have been implemented in GPUs before). Given these two reasons sorting networks are good candidates for GPUs because they can be implemented without data-dependent branching unlike CPU algorithms. The computational complexity of a sorting network algorithm in a GPU is.

The bitonic sort works by sorting bitonic sequences which are composed of two monotonic sequences. The sorting algorithm proceeds in stages, where it starts with an array of elements with which it constructs bitonic sequences. The number of steps performed at each stage is equal to the stage number it is in. So for example, if the sorting network is in the third stage it will require three steps. Each step proceeds by merging two bitonic sequences of sizeinto new bitonic sequence of size. The steps are performed in descending order until reaching the first step, and elements are compared in pairs, where the maximum and minimum are swapped if they are not in the correct order. Eventually, at each stage the array is divided into sorted data regions (highlighted green in and red in the illustration) which are then also sorted until the entire array is sorted. As indicated by the arrows in the illustration, bitonic sequences that are adjacent are merged, and the elements within them are sorted. [#x]



Figure . This figure illustrates the

bitonic sorting network on 8 different values

Fang et al. [#x] implement in their software GPUQP the GPUSort which is an implementation of the bitonic sort for GPUs. They store the input array by mapping it to a 2D texture with four color channels. Using a pixel program pairs of elements are compared in parallel, and the minimum or maximum of each pair is stored locally. This procedure continues until the array completely sorted.

Many researchers have also proposed variants of this sorting network algorithm to improve performance, but the global strategy is the same for all. [#x:survey-of-gpu]

Many of the algorithms based on the bitonic sorting network are not designed to handle very large databases, because they cannot work with databases that do not fit in the GPU’s memory. However, Govindaraju et al. present the GPUTeraSort, a sorting algorithm developed to sort database rows based on keys. It is able to sort billion-record wide-key databases that do not fit in the GPU video memory or in the main memory using a GPU. Their algorithm uses the GPU as a coprocessor that uses its data and task parallelism to perform memory-intensive and compute-intensive tasks ­­­while the CPU is used to perform I/O and resource management. The GPUTeraSort has a task pipeline with two phases. In the first phase, it reads disk asynchronously and builds keys in the CPU. At the same time it sorts using the GPU and then it generates runs and writes back to disk. This first phase is illustrated in image [x]. The sorter in this first phase of the algorithm makes use of the bitonic sort network, which will sort the data that is transferred from the CPU. Finally, in the second phase it reads, merges and writes the runs, back to the CPU, completing the sort [#x].

The GPUTerasort has five stages that were designed to be executed sequentially. However, some stages could be executed using multi-buffered pipeline-parallel independent threads:

The first stage involves the reader, which asynchronously reads the input file into a main memory buffer with size of 100 MB approximately. The reading bandwidth is improved by striping the input file across different disks so that the data is transferred from all disks in parallel. The I/O bandwidth and the CPU usage of the reader depend on the number of overlapping asynchronous I/O requests.

The second stage involves a Key-generator which computes the (key, record-pointer) pairs from the input buffer. In practice this stage is not computationally intensive but can be memory intensive it is because reading each key from main memory. It then sequentially writes a stream of keypointers pairs to the main memory of the GPU.

The third stage is the actual sorter which reads and sorts the key-pointer pairs. This stage is computationally intensive and memory-intensive on large buffers with wide keys.

The fourth stage is the reorder stage which rearranges the input buffer based on the sorted key-pointer pairs to generate a sorted output buffer (a run). On large databases, re-order is expensive because it randomly reads and writes long records from the input buffer and so it has many memory stalls.

In the fifth stage a writer asynchronously writes the run to the disk. Striping a run across many disks is not efficient fro Phase 2 reads; therefore the GPUTerasort cyclically writes the phase 1 runs to individual disks in very large transfers. The writer thread requires less than 10% of the CPU to achieve near-peak I/O performance.



Figure 3. First phase of the GPUTeraSort algorithm.

The performance of the GPUTeraSort was evaluated on billion-record files against optimized standard CPU-based algorithms. It was concluded that the overall performance of GPUTeraSort was in the mid-range GPU (costing around $300) which is comparable to that of a CPU-based algorithm running on a high-end dual Xeon processors (costing around $2,200). “In practice, GPUTeraSort achieves a good price-performance and outperforms the current PennySort benchmark.”



**Relational Joins on Graphics Processors**

In databases a relational join combines records from two or more tables. A new set is created that can be saved as a table or used as is. To perform a join operation, the records of multiple relations must be combined with a common key attribute. Join operations are computationally expensive, and can be accelerated by sorting the records based on common key attribute.

He et al. [#x] presented a novel design and implementation of basic relational join algorithms for GPUs. In their implementation they took advantage of the most recent GPU features which include support for writing to random memory locations, efficient inter-processor communication, and the new programming model for general-purpose computing provided by CUDA. Additionally, in their work, they used data-parallel primitives for performing map, prefix-scan and split operations to simplify the development of their algorithms.

The algorithms that were implemented included the non-indexed and index-nested loop join, the sort-merge join and the hash join. A set of data-parallel primitives split and sort was designed in order to implement them. The sort primitive they used was an improved version of the bitonic sorting network discussed in the sorting section. For this primitive they did the following optimizations:

They improved memory bandwidth utilization by guarantying coalesced access to global memory.

They moved repetitive fetches in the bitonic sort to local memory in order to improve speed of the algorithm.

The indexed nested loop join algorithm uses an additional data structure. They adapt a cache-optimized search tree, CSS-Tree [#x] to the GPU. This index helps it perform a greater number of concurrent index searches when performing the join operation.

They evaluated their work against optimized parallel counterparts of the algorithms on an Intel quad-core CPUs. Their work achieved a performance gain of 2-7X faster on the different joins when compared to the CPU-based approaches.

|  |  |  |  |
| --- | --- | --- | --- |
| **Joins** | **CPU(sec)** | **GPU (sec)** | **Speedup (sec)** |
| NINLJ (Non Indexed Nested Loop Join) | 528.0 | 75.0 | 7.0 |
| INLJ (Indexed Nested Loop Join) | 4.2 | 0.7 | 6.1 |
| SMJ (Sort-Merge Join) | 5.0 | 2.0 | 2.4 |
| HJ (Hash Join) | 2.5 | 1.3 | 1.9 |

Fang et al. [#x] also propose the Min-Max Join (MMJ) a new GPU-based algorithm to execute join operations. This algorithm uses hashing and sorting primitives previously designed in the work of He et al. [#x], as well as GPU-specific features, such as scattering and min-max blending. The Min-Max Join (MMJ) operation was implemented in their software GPUQP. Unfortunately there is no performance data published about the implementation.

Query Evaluation when using Indexing.

Gosink et al. proposed a new data structure that reduces the size of the data that needs to be transferred from the CPU to the GPU. They use a binning mechanism in order to encode the data such that it takes less space. In their paper they refer to the actual data from the projected column as high resolution data, and the encoded data as low resolution data. To illustrate this idea with an example, take arbitrarily a column of integers for example where each element is four bytes long in a 32-bit Intel architecture. When each integer is encoded using their binning scheme, each element is reduced to occupy only one byte, that in all saves three bytes of space for each element that needs to be transferred to the GPU. In this case, our original integer values would be the high-resolution data, and our encoded values would be the low-resolution data. Because by changing the encoding one cannot completely answer a given query, a second data structure is required to resolve the query where the encoded data is not enough to determine which attribute values fulfill and do not fulfill the condition. For these boundary bin values, they use a data structure that contains the high resolution data called the Data Parallel OrBic Structure. With this structure they carry out a candidate-check and effectively resolve the query.



Taken from [#x].

What is obtained in the end is an encoded data table, which is also called the low resolution data and They name this structure the Data Parallel OrBiC structure. This structure is composed of encoded data tables and an OrbiC. [#x]

Fang et al. [#x] implement in their software GPUQP, GPU-based CSS-Tree indices. These are a static, in-memory, cache-sensitive variants of the B+-Tree index. These indices are organized so that traversing each level of the tree yields to good data reference locality reducing the number of cache misses. In the GPU, they are organized as an array without pointers; as a result, searches are resolved via address arithmetic as opposed to pointer chasing which are inherently bad candidates for GPUs. Unfortunately there is no performance data published about the implementation.

**Conclusions**

A survey of the current algorithms that have been implemented in GPUs is presented.

One of the conclusions of this paper is that one of the greatest bottlenecks in applying GPUs to database operations, in particular the case where query processing is involved, is the fact that most of the time is spent transferring data from the CPU to the GPU.

The GPU is viewed by many of the database researchers as an effective co-processor to the CPUs. While it is clear that the GPU can outperform the CPU in some database operations they are not a complete substitute for the CPU, and only some operations may be accelerated with a GPU.

It is also important to add that a lot of the research may not affect databases directly but they provide primitives that can be used for databases, such as sorting primitives or search primitives.

Also GPUs are advancing. There is more that can be done with every fragment or multiprocessor. Many of them are starting to offer operations they didn’t offer before because they are advancing so a lot more algorithms will be possible to implement in GPUs as well.

Context switch is not as expensive as in CPUs

improvements on programmability

References

1. Luke J. Gosink, Kesheng Wu, E. Wes Bethel, John D. Owens, Kenneth I. Joy: Data Parallel Bin-Based Indexing for Answering Queries on Multi-core Architectures. SSDBM 2009: 110-129

2. Govindaraju, N.K., Lloyd, B., Wang, W., Lin, M.C., Manocha, D.: Fast computation of

database operations using graphics processors. In: Proc. of SIGMOD. (2004) 215–226

3. Govindaraju, N., Gray, J., Kumar, R., Manocha, D.: GPUTeraSort: high performance graph-

ics co-processor sorting for large database management. In: Proc. of SIGMOD. (2006) 325–

336

4. Owens, J.D., Luebke, D., Govindaraju, N., Harris, M., Kruger, J., Lefohn, A.E., Purcell, T.: A survey of general-purpose computation on graphics hardware. Computer Graphics Forum 26 (2007) 80–113