FastBit – *Helps Finding the Proverbial Needle in a Haystack*

Kesheng "John" Wu, Kurt Stockinger, Arie Shoshani, Wes Bethel
Lawrence Berkeley National Laboratory
Scientific Data Management Center

Summary

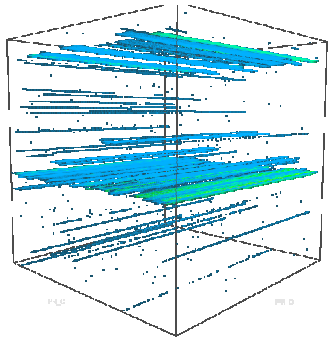
FastBit is a software package designed to meet the searching and filtering needs of data intensive sciences. In these applications, scientists are trying to find nuggets of information from petabytes of raw data. FastBit has been demonstrated to be an order of magnitude faster than comparable technologies. In this brief report, we highlight how we work with a visualization team, a network security team and a DNA sequencing center to find the nuggets in their data.

A common problem facing the data intensive applications is that they need to quickly and accurately locate key pieces of data out of the terabytes (10^{12} bytes) or petabytes (10^{15} bytes) of raw data. For example, out of hundreds of millions of collision events in a high-energy physics experiment called STAR, the physicists might expect less then 100 events with clear and unambiguous signatures of the Quantum-Gluon Plasma (QGP), a state of matter that the experiment aims to find. This type of searching problem presents a tremendous challenge for the existing searching technologies.

The technology for speeding up a search is generally called indexing. Most of the existing indexing methods are designed for transactional data where new records are usually generated by modifying the existing records. However, in data intensive sciences, most new records are fresh additions to be appended to the existing dataset. Our FastBit technology takes advantage of this "append-only" feature to provide a much more efficient searching tool for scientific data. In a series of performance tests, we observed that our technology improves the speed of searches by up to a factor of 12 compared with a commonly used commercial database system. In the remainder of this report, we briefly summarize three new features that we recently added to FastBit to support a number of DOE funded projects.

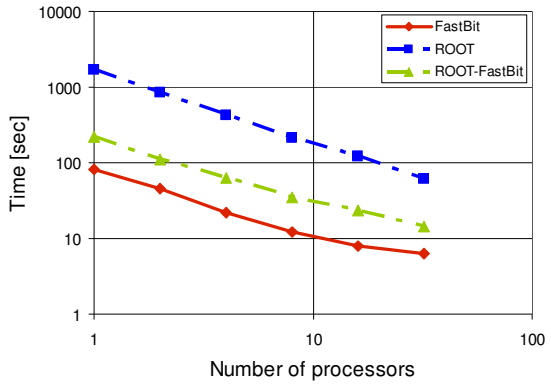
Supporting Conditional Analysis

Building on the searching capability, we have recently added functions to compute conditional multivariate histograms, such as the number of active internet connections over a time period. Working with the Visualization Group and the network security team at NERSC, we have applied this new feature to the analysis of network traffic by generating histograms dynamically based on specifications of the security analysts. For example, the figure below represents a two hour period of traces, where the blue color indicates a single connection to a particular IP address and port and the green indicates multiple connections.



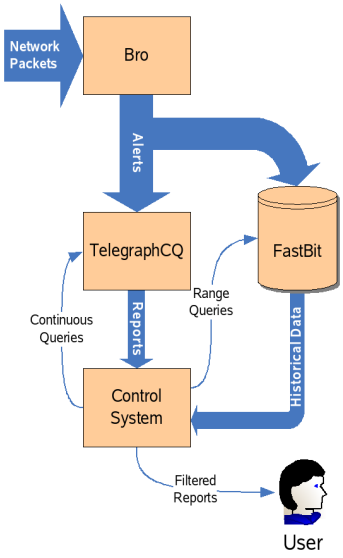
Therefore the blue sheets are likely to represent a coordinated effort to probe for network vulnerabilities (one connection to each address at a time) while the green is likely to represent a legitimate access (many connections to the same address).

In a performance measurement, we have compared our histogram algorithm against that of a common data analysis tool in high-energy physics called ROOT. We observe that our histogram algorithm is an order of magnitude faster than that of ROOT as shown in the next figure. The process of generating these dynamic histograms can be divided into two distinct steps, a searching step to identify the records that satisfy the required conditions and a counting step to put the selected records into the desired bins. We have also integrated the searching capability into ROOT (shown as ROOT-FastBit in the following figure), which significantly reduces the searching time needed; however, FastBit is still nearly a factor of two faster.



Supporting Top-K Queries

In a collaborative effort with the Database Group at UC Berkeley, we are integrating FastBit with TelegraphCQ to provide support of combined query on real-time streaming data and archival data in a single system as illustrated by the next figure. As a first application of the new system, we are currently working on analyzing a set of network traffic records. In this application, it is common to query for the top-K sources or destinations of network traffic, a form of top-K queries. Our system has been shown to be so efficient that it is able to handle many weeks of NERSC traffic data on a single desktop computer.



Supporting Aggregate Computations

FastBit employs a vertical data organization where each variable of a dataset (also can be viewed as a column of a data table) is stored in a separate file. This allows us to compute common aggregate functions such as *sum*, *minimum*, *maximum* and *average* more efficiently than the commonly used horizontal data organization. This feature is useful for many applications though it was initially implemented to meet the data analysis need of a quality assurance project jointly started by the Scientific Data Management Center and the Quality Assurance Group of the Joint Genome Institute (JGI). In the first week of using FastBit, we observed an anomaly in the temperature reported by one of the DNA sequencing machines. This anomaly was not observed by other quality control measures in-place at that time. This demonstrated that FastBit was able to handle the volume of the data produced by JGI.

For further information on this subject contact:
John Wu
Lawrence Berkeley National Laboratory
KWu@lbl.gov
(510)486-6609