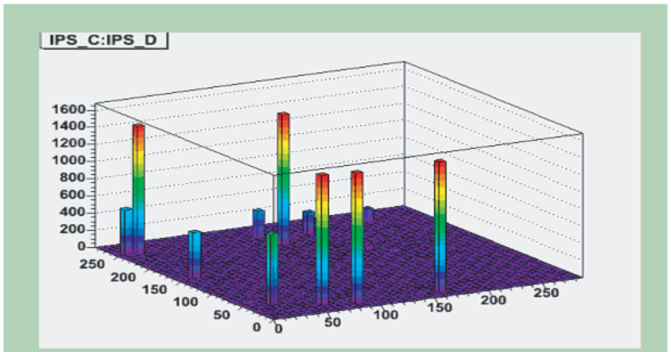
Network Traffic Analysis with Query Driven Visualization
SC05 HPC Analytics Challenge

The Challenge: *quickly find malicious and anomolous traffic in large collections of network connection data.* In the past year, about 500 terabytes crossed the boundary between the Internet and unclassified networks at LANL, NERSC, and LBNL. Monitoring tools at each of these boundaries collected summary information for approximately 10 billion distinct connections for about one TB of summary data. In addition, router-based data saved for every subnet internal to the LANL unclassified network totals 46 billion records and 2.5 TB per year, representing several Petabytes of network traffic. To meet the challenge of quickly detecting and responding to potential threats, our SC05 HPC Analytics entry combines several different technologies aimed at solving several technical challenges: large data management, efficient indexing and querying, automatic feature detection and classification, and multidimensional visualization.

Significance of Results. Our entry demonstrates new rapid and scalable data exploration capabilities enabled by the FastBit indexing and query system: an order of magnitude in performance gain when answering queries for a serial head-to-head comparison, and another order of magnitude in performance gain when using a parallel implementation. These results are significant because they reduce the "duty cycle" of hypothesis testing, which is a fundamental element of knowledge discovery. Forming queries is the basis for data exploration, and increasing the speed through better fundamental algorithms and scalable approaches is especially crucial in time-critical situations. Our results show an order of magnitude increase in performance when compared to the ROOT system, which is the "de facto standard" for data storage, retrieval and analysis employed by the High Energy Physics community. This community has some of the most challenging data management and analysis problems and some of the largest collections of data. Our multi-platform solution was developed on Linux systems, and has been deployed in production use on a diverse array of platforms: our challenge benchmark results were run on an IRIX system. Our techniques have been used successfully in knowledge discovery applications a number of different application domains, including large collections of High Energy Physics event databases, network connection data analysis, and simulation results produced by supernova and combustion models.

Excellent Serial Query Performance. For a typical multidimensional query, we compared the serial performance of ROOT and FastBit. ROOT is the "de facto standard" for data storage, retrieval and analysis in the High Energy Physics community. FastBit is the product of LBNL's Scientific Data Management research program. A serial implementation of ROOT required 2467 seconds to answer a query over three variables, while a serial FastBit query required only 309 seconds. Using state-of-the-art indexing and query technology can result in an order of magnitude increase in performance. Such performance gains are crucial in time-critical applications.

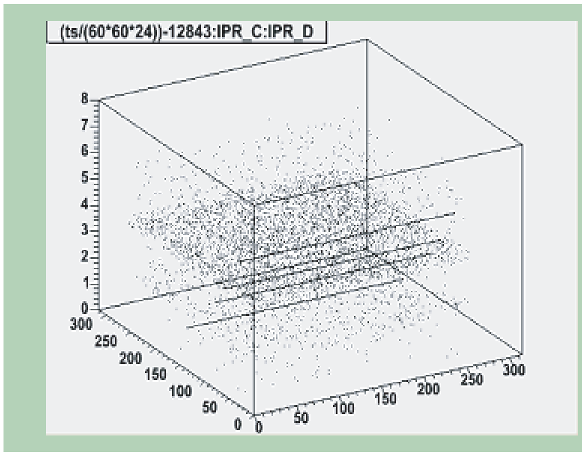
Scalable Query Performance. We conducted a modest study aimed at measuring the scalable performance of a parallel FastBit query implementation. Our results show a speedup of approximately 80%: a two-way parallel configuration runs 1.8 times faster than a serial configuration. Our twelve-way parallel configuration was able to answer a query in 22.8 seconds, compared to the 2467 seconds required for the serial ROOT implementation and the 309 seconds required for the serial FastBit implementation. These results show that it is possible to interactively answer queries on very large datasets by using larger configurations of computing hardware.



We identify previously unknown potentially hostile hosts by querying on a combination of failed connections originating from a range of suspect IP addresses. The image above shows the results of a query presented as a 2D histogram. The vertical axis indicates the number of failed connection attempts, and the last two octets of the IP address correspond to the two histogram axes. Hosts with a large number of failed connection attempts are immediately visible as "tall spikes." In a twelve-way parallel configuration, our system answers this query in 22.8 seconds on a 241GB dataset compared to the 2467 seconds required using a serial implementation of another commonly used query and analysis engine. Being able to quickly answer such queries is crucial for time-critical network connection data analysis.

The Challenge Source Data. The data we use in our challenge entry consists of twenty-four weeks of network connection summary information. Each record consists of source and destination internet address and port numbers, start time and duration of each connection, protocol, packet and byte count, along with a handful of other attributes. The data was collected from the borders of a major DOE facility, and represents information from 1.1 billion connections and occupies approximately 241GB of space. While our methods are scalable to larger data sizes, the amount of storage resources available to our research effort is limited in size.

Fast Queries. This first step in performing queries on large collections of data is to construct searchable indices. For our experiment, the size of the index structures was about 73GB, which is only about 30% the size of the raw data. In contrast, index structures for tree-based methods may be as large as 400% the size of the original data, and require O(n log n) processing complexity. The worst-case scenario for compressed bitmap indices is 200% times the size of the original data and require O(n) processing complexity. In FastBit, a typical query for our network connection data would be a compound Boolean expression like "select IPS_B, IPS_C, IPS_D where IPS_B < 100 AND IPS_C < 100 AND IPS_D = 128 AND sourcePort = 22". This particular query finds network connections that originate from a specific range of network addresses on the port nomrally used by SSH.



Once we have extracted a subset of network connections that are from suspicious hosts, we can visually present their behavior as a 3D scatterplot. The example above shows what appears to be scanning by some of the hosts. The scans are immediately visible as lines in the scatterplot. The line structure occurs as a hostile host sequentially steps through destination IP address space. Such patterns are typically visible only in large amounts of network data. Our HPC Analytics Challenge entry demonstrates interactive querying and processing capabilities suitable for use with large collections of network data.

State-of-the-Art Data Management. LBNL has developed a unique indexing, storage and retrieval system known as FastBit, which uses extremely efficient compressed bitmap indices. FastBit has been effectively used for managing huge collections of scientific data. For this challenge, we adapted this system, which has already integrated with the ROOT data analysis and graphing system developed at CERN, for the purpose of network data analysis. FastBit's performance has been shown to exceed that of similar (commercial) systems for multi-dimensional queries, scaling linearly with the size of the data set and sublinearly with the number of dimensions. With the help of the ROOT developers, we parallelized the index evaluation and data retrieval. This allowed us to increase ROOT-FastBit's performance by another order of

Automatic Feature Detection, Classification and 3D Presentation. Our collaborators at Los Alamos National Laboratory bring to bear several technologies in this HPC Analytics Challenge. Their work focuses on tools that aim to automatically find anomolous patterns in data, along with visualization applications tailed for high-dimensional data presentation and exploration.

Contacts and Contributors.

Lawrence Berkeley National Laboratory: Kurt Stockinger, Kesheng (John) Wu, Scott Campbell, Stephen Lau, Eli Dart, Brian Tierney, Jason Lee and E. Wes Bethel.

Los Alamos National Laboratory: Steve Smith, Mike Fisk, Eugene Gavrilov, Alex Kent, Christopher Davis, Rick Olinger, Rob Young, Jim Prewett, and Paul Weber.

University of New Mexico: Tom Caudell.

BERKELEY LAB

Office of Science
U.S. DEPARTMENT OF ENERGY

Los Alamos
NATIONAL LABORATORY
EST. 1943