# High Performance Visualization using Query-Driven Visualization and Analytics

E. Wes Bethel[*], Scott Campbell, Eli Dart, John Shalf, Kurt Stockinger, Kesheng Wu
Lawrence Berkeley National Laboratory

## Summary

***Query-driven visualization and analytics is a unique approach for high-performance visualization that offers new capabilities for knowledge discovery and hypothesis testing. The new capabilities – akin to "finding needles in haystacks – are the result of combining technologies from the fields of scientific visualization and scientific data management. This approach is crucial for rapid data analysis and visualization in the petascale regime. This article describes how query-driven visualization is applied to a "hero-sized" network traffic analysis problem.***

It is generally accepted that as sciences move into the tera- and peta-scale regimes that one major limitation is the ability to analyze and understand phenomena "hidden" in vast and complex collections of scientific data. In recent years, much of the work in computer and computational science has aimed to produce so-called "scalable technologies" that aim to produce software algorithms and implementations capable of using hundreds or thousands of processors on supercomputers. The premise is that greater computational horsepower must be brought to bear to solve the growing data analysis and understanding problem.

We take a different approach to high performance visualization using a methodology known as "Query-Driven Visualization and Analytics." The basic idea is that instead of making a single image of a terabyte's worth of data, we compute an image that contains only data deemed to be "scientifically interesting." This approach can reduce the processing load – computational, visual and cognitive – by multiple orders of magnitude when compared to conventional scalable visualization techniques.

In our research experiments, "scientifically interesting" is defined as a combination of boolean range predicates. In other words, "interesting" is defined in terms of combinations of data range values. For example, a combustion researcher may be interested in focusing scientific inquiry and study on locations in the computational domain where there is a high temperature gradient that presumably corresponds to the location of the flame front. Query-driven visualization and analytics makes it possible to first define "interesting," then extract and process "interesting data." The major new capability is having to find and process only the small subset of data that is "interesting" – the "needle in the haystack."

Query-driven visualization and analytics is realized by combining technologies from the fields of scientific visualization, visual analytics and scientific data management. Storing, retrieving, indexing and querying data are research and engineering challenges from the field of scientific data management. For our work here, we employ LBNL's patented FastBit software. FastBit provides an implementation of compressed bitmap indexing. Visual analytics and visualization techniques are used to provide easy-to-understand visual representations of data and interactive "drill-down" capability for rapidly exploring and analyzing large and complex collections of scientific data. The

[*] (510) 486-7353, ewbethel@lbl.gov

example we show below focuses on query and display of statistical features in data rather than the data itself.

We introduced this combination of concepts in a 2005 publication[1], then follow up with a 2006 publication describing its application to a "hero-sized" network traffic analysis problem[2]. Our performance experiments showed that for this particular application, our approach was able to answer and display queries four orders of magnitude faster than conventional techniques for network traffic analysis, and two orders of magnitude faster than when using current state-of-the art systems for high energy physics data management. This type of performance gain represents a significant new capability for data understanding in the petascale regime.

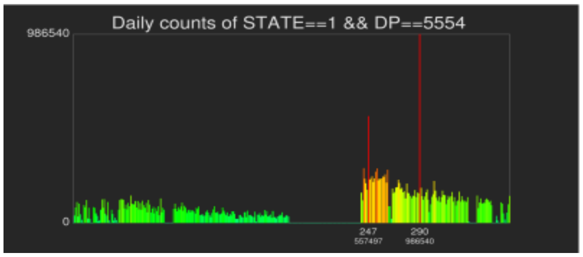

**Figure 1.** Histogram of suspicious activity over a one-year period at one-day granularity. Coordinated activity appears in the weeks around day 247.

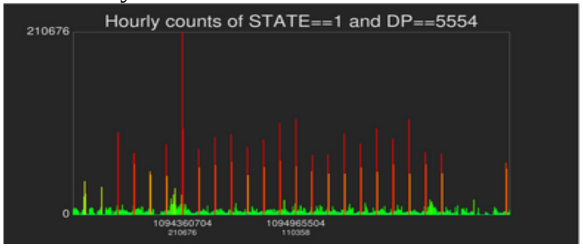

**Figure 2.** Drilling into the data, we compute and display a histogram of suspicious activity over a four-week period at one-hour temporal resolution. Evidence of regular activity becomes apparent.

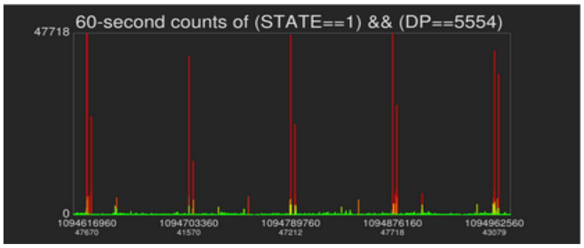

**Figure 3.** Drilling even deeper into the data, the suspicious activity occurs daily at 21:15 local time.

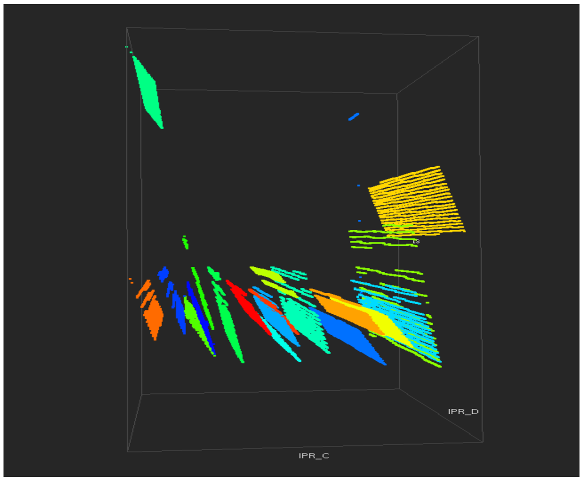

**Figure 4.** After several more query/display iterations, we determine the IP addresses of the twenty remote hosts participating in a distributed scan of a major DOE computing facility. This image shows the destination addresses being attacked by the twenty different hosts during one of the attack cycles.

Future work in query-driven visualization and analysis will include adapting and applying the concepts to science stakeholders in fusion, combustion, accelerator modeling and astrophysics. Each of these areas presents a unique set of data understanding challenges. A long-term objective is to provide such capabilities to DOE's science community for production use in its open computing facilities.

**For further information on this subject contact:**

Name: E. Wes Bethel.
Organization: Lawrence Berkeley National Laboratory.
Email: ewbethel@lbl.gov
Phone: (510) 486-7353

Note: This work is a joint effort by the Visualization, Scientific Data Management, ESnet and NERSC programs at Lawrence Berkeley National Laboratory.

[1] *Query-Driven Visualization Visualization of Large Data Sets*, IEEE Visualization 2005.

[2] *Accelerating Network Traffic Analysis using Query-Driven Visualization*. IEEE Symposium on Visual Analytics Science and Technology, IEEE Visualization 2006.