

**Query-Driven Network Flow Data Analysis and
Visualization**

Final Report

SPAA LB05-001720 - PNNL

**E. Wes Bethel
Lawrence Berkeley National Laboratory**

14 June 2006

1 Background and Significance

This document is the final report for a limited-scope and duration project conducted through Lawrence Berkeley National Laboratory's (LBNL) Work-For-Others (WFO) program for the National Visualization and Analytics Center (NVAC) at Pacific Northwest National Laboratory (PNNL). The original statement of work, submitted and accepted in February 2005, focused on architectural issues for a broad analytics data processing framework. The statement of work was revised and accepted on approximately September 2005 and the work period extended through June 2006. The new statement of work focused on "analytics research" for new capabilities that would enable rapid analysis and visualization of large collections of Network Flow data. This document is the final report and deliverable for the project.

In a series of workshops held by the Office of Science Scientific Data Management community in 2004¹, dozens of application scientists stated that their ability, or lack thereof, to manage, analyze and understand data are limiting factors in their ability to conduct scientific research. Motivated by the desire to improve a researcher's ability to more quickly gain an understanding of data characteristics, our approach is to combine state-of-the-art scientific data management technology with visualization and analysis tools. The resulting combination is intended to fulfill several distinct objectives. First, it relies on emerging research from the field of scientific data management to store, find and retrieve data records from large data collections. Second, it focuses visualization and analysis processing only on the data deemed to "be interesting" by a researcher. A byproduct of this characteristic feature is reduced visual interpretation processing load. Third, we applied the aggregate technology to a challenging problem in network flow data analysis. Fourth, we leveraged characteristics of the underlying data management technology to help a user to formulate better queries. The overall objective is to more quickly and efficiently explore via discourse very large collections of NetFlow data.

2 Approach and Methods

In our experiments, we used what is known as "bitmap indexing" from the scientific data management community for the data indexing and searching infrastructure. Specifically, we used LBNL's FastBit implementation², which has successfully been used to rapidly return queries from large data collections generated by high energy physics experiments³. As applied to NetFlow data, queries take the form of compound Boolean expressions like "(destination IP address range == 192.101) AND (1 <= destination_port <= 1024) AND (nbytes > 0)." We explored the efficacy of using bitmap indexing as the basis for NetFlow data indexing and querying, where results are then propagated along to analysis, visualization and interaction tools. A particular feature of bitmap indexing leveraged is its ability to very quickly return "counts" of data records that would satisfy a particular query. We show that such a capability, combined with multiresolution and statistical analysis and presentation, is instrumental in helping a user to formulate more meaningful queries to accelerate the knowledge discovery process.

3 Results

The original deliverables spelled out in our August 2005 Statement of Work are listed below, along with the dates of completion and final outcome.

¹ <http://www-conf.slac.stanford.edu/dmw2004>

² <http://sdm.lbl.gov/FastBit>

³ K. Wu, W.-M. Zhang, V. Perevoztchikov, J. Lauret, and A. Shoshani. "The Grid Collector: Using an Event Catalog to Speed Up User Analysis in a Distributed Environment." In Computing in High Energy and Nuclear Physics (CHEP) 2004, Interlaken, Switzerland.

1. An SC05 HPC Analytics Challenge entry based upon the combination of technologies described above including web-based summaries of the entry.
 - a. This task was completed on 1 Aug 2005 – the date entries were due.
 - b. Our submission is posed on the web⁴ and is available as LBNL Technical Report number LBNL-58768.
 - c. The entry received a runner-up award at the SC05 HPC Analytics Challenge.
2. A submission to the IEEE Transactions on Visualization and Computer Graphics special issue on Visual Analytics. Submissions are due 28 Nov 2005, with final camera-ready copy due on 19 May 2006.
 - a. This task was completed on 2 December 2005 with a submission to the IEEE TVCG special issue on Visual Analytics.
 - b. The paper authors consisted of visualization experts (Bethel, Smith), scientific data management experts (Stockinger, Wu) and network security experts (Campbell, Dart, Lee, Tierney).
 - c. The paper described how the duty cycle in the network data analysis process could be accelerated by leveraging state-of-the-art scientific data management technology in conjunction with a visual presentation of information. We demonstrated reasonably linear parallel speedup for queries, and applied the system to a large collection of network data – larger by an order of magnitude than anything previously attempted. Our performance study focused on the time required to answer a typical query and compared the performance of FastBit and ROOT (the “gold standard” for high performance data index/query in the high energy physics community); the result is that FastBit was about an order of magnitude faster than ROOT for a realistic-sized problem – 2467 seconds vs. 309 seconds. The paper was not accepted for publication as: (1) the reviewers wanted a stronger link between the SDM and visualization technologies, and (2) there were some questions about the performance testing methodology the reviewers felt could not be trivially answered.
 - d. Our submission is posted on the web⁵ and is available as LBNL Technical Report LBNL-59166.
3. An SC05 demonstration in the LBNL booth (Nov 2005).
 - a. We demonstrated an early prototype of our application in the LBNL booth at SC05 in November 2005.

In addition to those deliverables, we leveraged other funding within the LBNL Visualization Group to make more headway than was originally covered by the SOW to PNNL.

4. Submission to IEEE VAST06. We revised our TVCG submission taking into account the reviewer comments: we used a more rigorous methodology for conducting performance tests, and we developed a custom visual analytics application that emphasizes the creation and display of n -dimensional histograms. This approach is unique in that we are using only the statistical characteristics of the underlying data for the visual presentation and for assisted query formulation. Our performance results show that our approach is scalable on a modest-sized SMP platform, and exhibits speedups ranging from three to five orders of magnitude speedup over traditional techniques for network data analysis depending upon degree of parallelism (156381 seconds vs. 5 seconds, 2237 seconds vs. 2 seconds), and speedups ranging from about two to four orders of magnitude faster than when using ROOT and projection indices for finding “interesting network data” in a large collection (1357 seconds vs. 5 seconds, 99 seconds vs. 2 seconds). A prepublication version of the paper is available

⁴ <http://vis.lbl.gov/Publications/2005/SC05-HPCAnalytics-LBNL-58768.pdf>

⁵ <http://vis.lbl.gov/Publications/2005/LBNL-59166-QueryDrivenNetworkDataAnalysis.pdf>

upon request. Authors on this paper include a visualization expert (Bethel), scientific data management experts (Stockinger, Wu) and network security experts (Campbell, Dart). Status: *accepted for publication, camera-ready copy due 1 August 2006*.

5. Submission to Supercomputing 2006. We took the IEEE VAST06 paper ideas another step further to demonstrate temporal, multiresolution browsing and feature detection. In this paper, we identify the temporal characteristics of a distributed scan that occurs within a 4-week window from a 42-week collection of network connection data collected at NERSC in 2004. This paper also presents new work in scientific data management driven by needs posed by our visual analytics application – the ability to quickly compute 1- and 2-dimensional histograms. A prepublication version of this paper is available upon request. Authors on this paper include a visualization expert (Bethel), scientific data management experts (Stockinger, Wu) and network security experts (Campbell, Dart). Status: *acceptance decision notification 5 July 2006*.
6. Online Vignette. We created a web page⁶ describing the temporal, multiresolution visual analytics process employed in our SC06 paper.

4 Conclusion and Future Work

The main focus of our visual analytics research has been on testing the thesis that the duty cycle in knowledge discovery can be dramatically reduced by leveraging state-of-the-art scientific data management technology: by reducing the time required to find data that satisfy an analyst's criteria, less time will be required to confirm the expected or discover the unexpected. To that end, our IEEE VAST06 submission drives home the point: traditional tools for network data analysis – awk, grep, sed – simply are not usable for large network collections; interactive analysis is both feasible and demonstrably possible when using current SDM technologies combined with a well-designed visual analytics interface. The IEEE VAST06 paper shows we can expect two to five orders of magnitude in time required to discover and display “interesting data” when we combine state-of-the-art scientific data management technology with visualization and analytics infrastructure. Lessons from the IEEE VAST06 paper in turn drove new developments in SDM technology resulting in new capabilities for rapidly computing 1- and 2-D histograms. In all cases, our visualization and visual analytics methodology used statistical information about the data – distribution of counts satisfying an analyst's criteria – rather than using the raw data itself.

These features represent major new technological leaps for visual analytics: (1) using state-of-the-art scientific data management tools in conjunction with visual interfaces results in completely new capabilities and unprecedented performance gains for tackling knowledge discovery in real-world datasets; (2) such visual analytics can occur completely in “data statistics space” rather than requiring direct access to the underlying data.

⁶ <http://vis.lbl.gov/Vignettes/QDV-NetworkTraffic/qdv-vignette.html>