Application of High-performance Visual Analysis Methods to Laser Wakefield Particle Acceleration Data

Oliver Rübel [*†‡], Prabhat [*], Kesheng Wu [*], Hank Childs [§], Jeremy Meredith [¶], Cameron G.R. Geddes [‖], Estelle Cormier-Michel [‖], Sean Ahern [¶], Gunther H. Weber [*], Peter Messmer [**], Hans Hagen [†], Bernd Hamann [‡†*] and E. Wes Bethel [*‡]

Abstract

Our work combines and extends techniques from high-performance scientific data management and visualization to enable scientific researchers to gain insight from extremely large, complex, time-varying laser wakefield particle accelerator simulation data. We extend histogram-based parallel coordinates for use in visual information display as well as an interface for guiding and performing data mining operations, which are based upon multi-dimensional and temporal thresholding and data subsetting operations. To achieve very high performance on parallel computing platforms, we leverage FastBit, a state-of-the-art index/query technology, to accelerate data mining and multi-dimensional histogram computation. We show how these techniques are used in practice by scientific researchers to identify, visualize and analyze a particle beam in a large, time-varying dataset.

Index Terms: data mining, visual data analysis, accelerator modeling, parallel visualization, large data visualization, temporal visualization, temporal data analysis

1 Introduction

Laser WakeField Accelerators (LWFAs) promise to be a compact source of high-energy electron beams and radiation. Using an intense laser pulse fired into a plasma, LWFAs have been shown to generate electric fields thousands of times stronger than those in conventional particle accelerators. LWFAs accelerate particles to high energies of 1GeV within $3cm$ compared to $>\approx 5m$ in traditional electromagnetic accelerators [1].

In order to gain deeper understanding of phenomena observed in LWFA experiments, researchers at the LOASIS facility at LBNL [1] and at Tech-X[**] use the VORPAL [4] simulation code to computationally model their experiments. Gaining scientific insight is often challenging due to the size and complexity of the data output by the simulation. To support scientific knowledge discovery and hypothesis testing on this kind of data, we combine and extend two different technologies aimed at enabling rapid, interactive visual data exploration and analysis. We extend and adopted histogram-based parallel coordinates as a vehicle for visual information display and as an interface for data selection via multi-dimensional thresholding. We employ state-of-the-art index/query technology to accelerate the data mining process as well as generation of conditional histograms. We execute these activities on large, parallel machines to achieve interactive data mining and visual analysis rates on the order of seconds for terabyte-size datasets.

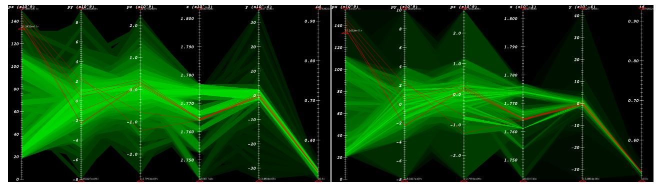

Figure 1: Histogram-based parallel coordinates using $32x32$ uniform binning (left) and adaptive binning (right). The adaptive binning preserves more details in dense areas while discarding some details in sparse areas of the data which may ease comparison of selections (red) with density-features of the data in low-level-of-detail views.

Historically, a scientific researcher investigates time-varying data of this type by first creating an animation, then viewing it to identify a timestep and multivariate threshold values that will result in isolating particles that form a beam. Having defined the beam, one would run scripts searching at each timestep for particles of interest to perform particle tracing. In the past, these activities have typically required hours to complete. Using our techniques and implementation, this activity is executed at interactive rates: a scientific researcher can interactively select a beam and trace it within seconds, greatly reducing the duty cycle in visual data exploration and mining while improving accuracy of the analysis [6].

2 Related Work

Parallel coordinates are a common information visualization technique for visual display of high-dimensional data sets. In the case of extremely large data, standard, line-based parallel coordinates suffer from problems like excessive clutter and occlusion. We have built on and extend the work of Novotný and Hauser [5] who binned parallel coordinates to overcome these limitations, which are especially acute with very large datasets.

We utilize FastBit [2], a state-of-the-art index/query system for data extraction and subsetting. It implements the fastest-known bitmap compression technique, and has been demonstrated to be effective in a number of data analysis applications. In particular, it has a number of efficient functions for computing conditional histograms, which are crucial for this work.

To deliver our research results to scientific researchers, our new techniques are implemented in VisIt [3]. VisIt is a production-quality, parallel capable visual data analysis application that runs on virtually all modern HPC platforms.

3 System Design

To accelerate data mining operations, we use FastBit, a state-of-the-art data management technology for indexing and searching. We use FastBit to perform data subsetting/selection and to compute conditional histograms. We implemented these operations using FastBit directly in the file-reader stage of the processing pipeline in VisIt. The conditional histograms serve as basis for the visual presentation of data vis-a-vis histogram-based parallel coordinates.

[*]Computational Research Division, Lawrence Berkeley National Laboratory (LBNL), One Cyclotron Road, Berkeley, CA 94720, USA.

[†]International Research Training Group "Visualization of Large and Unstructured Data Sets," University of Kaiserslautern, Germany.

[‡]Institute for Data Analysis and Visualization, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA .

[§]Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550

[¶]Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831, USA.

[‖]LOASIS program of Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA.

[**]Tech-X Corporation, 5621 Arapahoe Ave. Suite A, Boulder, CO 80303.

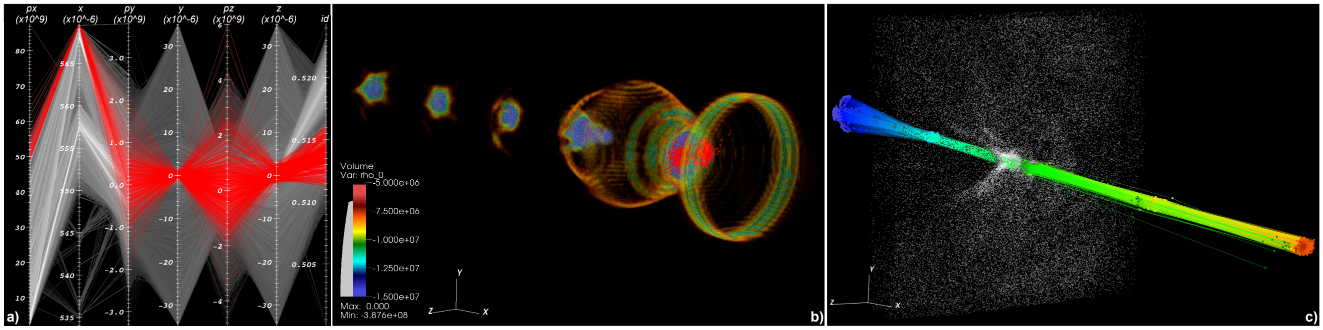

Figure 2: a) Parallel coordinates of timestep $t = 12$ of the 3D dataset. Context view (gray) shows particles with x-direction momentum component $px > 2 * 10^9$. The focus view (red) shows particles with $(px > 4.856 * 10^{10}) \&\& (x > 5.649 * 10^{-4})$ which form a compact beam in the first wake period following the laser pulse. b) Volume rendering of the plasma density and the focus particles (red). c) Traces of the focus particles for timestep $t = 9$ (when most particles entered the simulation window) to timestep $t = 14$. Color indicates px. The context particles are shown in gray for timestep $t = 12$ where the original selection was performed. The increasing px values show that the selected particles are constantly accelerated.

In contrast to earlier work, we employ a histogram-based parallel coordinates rendering for both context and focus views of large, complex data. Via FastBit, we can recompute conditional histograms fast, thus enabling support for fast data selection and smooth drill-down into finer level of detail in very large datasets. As a further improvement, we also support use of adaptively binned (equal weight) histograms. As illustrated in Figure 1, adaptively binned histograms are especially useful for display of low-level-of-detail views where the number of bins per variable is much smaller than the number of pixels per parallel axis.

In our system we make use of two main types of data selection, both of which are directly accelerated using FastBit: i) multivariate thresholding, and ii) identifier-based (ID-based) selection. Multivariate thresholding is used for defining "interesting" data subsets. ID-based selection is the basis for tracing of particle over time.

4 Beam Analysis

LWFA simulations model the effects of a laser pulse being applied to a hydrogen plasma. Similar to the wake of a boat, the radiation pressure of the laser pulse displaces the electrons in the plasma, and with the space-charge restoring force of the ions, this displacement drives a wave (wake) in the plasma. Electrons can be trapped and accelerated by the longitudinal field of the wake, forming electron "bunches" of high energy. In order to gain a deeper understanding of the acceleration process, scientists need to answer complex questions like: i) which particles become accelerated; ii) how are particles accelerated, and iii) how was the beam of highly accelerated particles formed and how did it evolve.

Figure 2 shows an analysis of a 3D particle dataset with an overall size of 210GB, consisting of 30 timesteps, each having a size of $\approx 7GB$ ($\approx 2GB$ for the index), and ≈ 90 million particles per timestep. A scientist defines the beam using multi-dimensional thresholding, which is implemented by a combination of fast data subset/selection, histogram computation, and display/interaction with parallel coordinates. This process provides direct visual feedback about properties of a selection directly in the parallel coordinates plot, enabling more accurate selection of the beam (see Figure 2a). Having identified a particle subset of interest, the scientist typically performs further analysis of subset properties using different types of plots. As shown in Figure 2b, a combination of volume rendering of the plasma density, along with a rendering of the user-defined beam particles reveals the location of the electron beam within the wake. By tracing the selected particles through time one can effectively analyze the temporal history of a beam. In such analysis, additional beam-substructures are often revealed. By refining the initial selection based on information from different timepoints, a user can effectively detect and analyze also beam-substructures. Analysis of particle paths then supports the analysis of the evolution of the beam (see Figure 2c). Using a series of ID-queries one can compute particle paths efficiently even for very large datasets.

5 Conclusions

The system we have developed enables rapid knowledge discovery from large, complex, multivariate and time-varying scientific datasets. We have presented a novel approach for efficiently creating histogram-based parallel coordinate displays. Using this form of visual information display as basis for performing complex multivariate data, the scientist can effectively explore and analyze large and complex data. Using state-of-the-art index/query technology, we are able to efficiently extract data subsets that meet a set of conditions as well as quickly compute conditional histograms. We have shown how the combination of these technologies can effectively be used for exploration of large, time-varying laser wakefield particle acceleration data. Our future work will focus on improving the performance of our system by parallelizing the most expensive parts of our visual data analysis framework.

Acknowledgements

This work was supported by the Director, Office of Science, Offices of High Energy Physics and Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 through the Scientific Discovery through Advanced Computing (SciDAC) program's Visualization and Analytics Center for Enabling Technologies (VACET) and the COMPASS project. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We also thank the VORPAL development team for ongoing efforts to develop and maintain this parallel C++ framework on a variety of supercomputing platforms.

References

[1] LOASIS: `http://loasis.lbl.gov/`.

[2] FastBit is available at `https://codeforge.lbl.gov/projects/fastbit/`.

[3] VisIt is available at `https://wci.llnl.gov/codes/visit/`.

[4] C. Nieter and J. R. Cary. VORPAL: A Versatile Plasma Simulation Code. *J. Comput. Phys.*, 196(2):448–473, 2004.

[5] M. Novotný and H. Hauser. Outlier-preserving focus+context visualization in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):893–900, 2006.

[6] O. Rübel, Prabhat, K. Wu, H. Childs, J. Meredith, C. G. R. Geddes, E. Cormier-Michel, S. Ahern, G. H. Weber, P. Messmer, H. Hagen, B. Hamann, and E. W. Bethel. High performance multivariate visual data exploration for extemely large data. In *SuperComputing 2008 (SC08)*, Austin, Texas, USA, Nov 2008. (to appear).