Infrastructure for Scalable and Interoperable Visualization and Analysis Software Technology

A National Visual Analytics Center Position Paper

E. Wes Bethel
Visualization Group
Lawrence Berkeley National Laboratory
Berkeley, California
June 2004

1 Our Vision

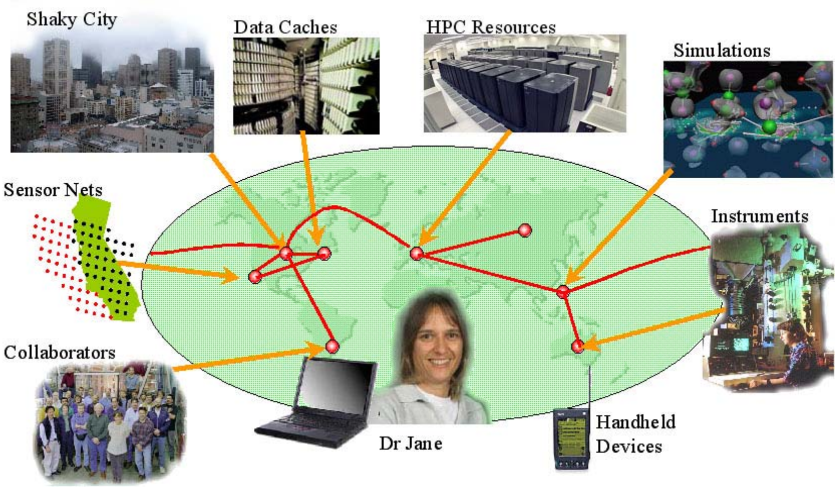
Dr. Jane is a geophysics researcher who studies the effects of how a new construction material can mitigate the effects of earthquakes on urban structures. In the course of her studies, she runs experiments at the molecular scale, where she observes molecule-molecule interactions, and also at the urban scale, where she uses large equipment to induce shock waves into the ground. She has access to instrumentation attached directly to urban structures in Shaky City, where real-time data is collected and stored at a nearby data center.

Dr. Jane runs large-scale simulations around the world wherever cycles are available. Because the simulations routinely generate terabytes of results, simulation results are stored on data caches located close to the computational resource and are later analyzed and visualized. She routinely uses experimental data to seed her simulations, and in turn, uses simulation results to modify her experimental parameters. She uses data fusion techniques to compare the results of experiment and simulation, and uses data analysis and mining techniques to discover the combinations of material properties and construction techniques that produce the most resilient buildings and for validation of simulation models.

When an earthquake occurs, she is able to immediately connect to a vast network of sensors to observe first-hand how her handiwork has reduced direct shock damage caused by earthquake shaking. Such observation can occur using a desktop workstation, or using a handheld telephone. Dr. Jane is part of a large, multidisciplinary team that is scattered around the world. The collaborators have access to subsets of the collection of experimental and simulation data depending upon security policy. The data may reside at any one of the member institutions.

As you may have guessed, this description is fiction (and is adapted from [1]). However, the scope of activities performed by Dr. Jane is not far fetched, and is reasonably characteristic of many modern research scientists.

The vision of scientists – and also analysts and decision makers in general – is the ability to gain insight in a timely fashion from data that is large, diverse and at times distributed over a wide area. Whether the data source is a sensor network or a scientific instrument, the general principles of the scientific process apply across many domains: hypothesize; experiment or simulate; analyze, test and validate; refine and repeat. It is our position that the primary challenges facing large-scale science or national security concerns are quite similar: diverse and heterogeneous data sources provide the input to advanced analysis software employed by analysts to glean insight.



2 Vision Foundation

Our vision represents a large federated collection of technologies organized into a well-orchestrated, end-to-end solution. A successful National Visual Analytics program will include similar activities that span research and development for the fundamental technologies as well as integrative and deployment efforts designed to produce capabilities in use to solve the nation's most difficult science and related problems.

Data Acquisition. Early in the pipeline, data must first be acquired. Data acquisition may involve using any of a number of devices ranging from magnetic loops embedded in highways to biochemical sensors placed in static locations or on mobile platforms. Once acquired, raw data could be processed at the source to reduce downstream bandwidth and processing requirements. After acquisition, data is transferred via wired or wireless networks to storage locations. Storage locations might be loosely assembled into a federation for shared access across programmatic boundaries.

Data Management. Once acquired, data is accumulated into larger collections for subsequent processing and analysis. Large data collections should be "prepared" for application-specific access. Some applications might need to perform query-based access to find data values or events that match a specific set of constraints. Other applications may wish to have access to collections of data to perform large-scale analysis. Federated collections of data form the basis for wide-area data fusion and mining applications. When the size of source data ranges into the terabytes, it becomes imperative to employ layout and summary optimization techniques to minimize the time required to access data.

Data Fusion, Analysis and Visualization. After being readied for analysis in repositories, data are then used as input to data analysis and visualization tools. Data analysis and visualization tools provide the means to detect and track features in data, and to present complex data in visual form suitable for use by analysts and decision makers. They include assimilation of data arriving from different sources, different data acquisition modalities, and different delivery formats: streaming data from passive sensor nets, actively controlled instruments (eg. robots), computer simulations, and even searchable data archives. In some instances, analysis and visualization are performed interactively to find an answer to a specific question. In other instances, a "drag net" of analysis capabilities are run in an offline fashion to find new and interesting features in data. In yet other instances, sequences of analysis and visualization tools are executed in response to dynamic data sources to present "just in time" information to analysts and decision makers.

Applications: Cross-Program Technology Integration. The technology areas above are the "raw materials" that are then integrated into finished applications or application toolkits. The process of creating and deploying applications consists of careful functional and performance requirements analysis coupled with technology integration by a multidisciplinary team of experts working closely with end users. Careful programmatic planning and execution will help to ensure ongoing support and maintenance for new technologies as they are deployed into the field.

3 LBNL Contributions

Lawrence Berkeley National Laboratory is well positioned to play a central role in the National Visual Analytics Center program as our programmatic goals share a common vision and common technical challenges. Two particular areas of emphasis at LBNL are ripe for dual-use in the NVAC program.

Visualization. LBNL's visualization effort spans the range of algorithmic research for novel and useful visualization techniques through high performance software architectures and implementations. Most recently, LBNL has begun a project called the Distributed Visualization Architecture (DiVA). DiVA's ultimate objective is to foster the definition and emergence of a "virtual standard" for high performance scientific visualization. Our research focuses on visualization algorithms, infrastructure and architecture as applied to DOE-funded science projects requiring the ability to perform scientific visualization and data analysis on large and complex scientific data using remote and distributed visualization (RDV) resources. Our project combines research activities from two complementary areas that will result in new capabilities needed by science programs. The two research areas are (1) graphics and visualization algorithms targeted especially for RDV, and (2) the infrastructure needed to effectively deploy these new RDV capabilities. DiVA represents, in one view, a testbed for trying out new ideas, and an integrative environment for creating new applications.

Scientific Data Management. LBNL's Scientific Data Management Group, led by Arie Shoshani, performs research and development into tools and techniques for data modeling, data storage and retrieval, Grid-based access and access optimization. Their recent work in the Earth System Grid project demonstrates query-based access to distributed, wide-area, federated data resources. The ESG is a culmination of several related research activities that include parallel and Grid-based I/O infrastructure, exploratory analysis and data mining, distributed and heterogeneous data integration, and efficient processing access of very large data sets.

The two LBNL groups – Scientific Data Management and Visualization – have a close working relationship. In addition, both groups have strong ties with related academic programs that would serve as partners as part of a larger, multi-institution effort.

[1] J. Shalf and W. Bethel, "How the Grid Will Affect the Architecture of Future Visualization Systems," IEEE Computer Graphics and Applications, Volume 23, Number 2, March/April 2003, pp 6-9.