BigData, Big Network

Areas of focus for the workshop included but were not limited to science gateways, bioinformatics, oceanography, seismography, astronomy, nano-materials sciences, digital media, and scientific instruments. **All areas of research that involve current or potential cross-border collaboration between Mexico and the United States were welcome** if they are or can be enabled by the 10-Gigabit network connectivity between Mexico and the US.

List of Abstracts for BDBN Workshop

* * * * *

Focus Area: Science Gateways

1) Title: The Science Gateway Institute and opportunities for international collaboration

Submitter: Nancy Wilkins-Diehr, UCSD

Summary: This presentation will report on a 2-year study entitled "Opening Science Gateways to Future Success" that evaluated the characteristics of successful gateways and analyzed sustainability issues. That work served as a foundation for the Science Gateway Institute proposal.

2) Title: HUBzero: An Open Source Platform for Building Science Gateways

Submitter: Michael McLennan PURDUE University

Summary: HUBzero is an open-source software platform used to create science gateways or "hubs" for scientific collaboration, research, and education. It has a unique combination of capabilities that support science and engineering. A little like YouTube.com, HUBzero allows people to upload content and "publish" to a wide audience, but instead of being restricted to short video clips, it handles datasets, analysis tools, and other kinds of scientific content.

3) Title: nanoHUB: A Science Gateway with global usage and impact

Submitter: Gerhard Klimeck, PURDUE University

Summary: This Presentation will provide an overview of these processes and their impact as they are supported on nanoHUB.org today. Annually, nanoHUB provides a library of 3,000+ learning resources to 260,000+ users worldwide. Its 270+ simulation tools, free from the limitations of running software locally, are used in the cloud by over 12,000 annually. Its impact is demonstrated by 1030+ citations to nanoHUB in the scientific literature with over 11,000 secondary citations, yielding an h-index of 50, and by a median time from publication of a research simulation program to classroom use of less than 6 months. Cumulatively, over 19,000 students in over 1,000 formal classes in over 185 institutions have used nanoHUB simulations.

4) Title: Towards a long-term e-Science infrastructure in Latin America

Submitter: Jesús Cruz Guzmán, UNAM

Summary: The collaborative experience in Latin America had built an e-Infrastructure, but the fast emergency of new technologies and the difficult in using for many users present a big challenge for their support. Science Gateway is a way of provide access to many services, one effort in LA is realized by the SCALAC (Servicios de Computo Avanzado Para Latino América y el Caribe) community, organized inside CLARA Net, and other similar services for Mexican Institution for Research and Education. The principal aim is to build advance computing services based on standard technologies, with flexibility adapted to the community requirements. The science gateway is the bridge between the scientific applications and e-Infrastructure resources devoted to support the Virtual Research Communities. Here the services architecture is presented and the status of the services supporting the work of the Virtual Research Communities.

5) Title: A Standard-based Science Gateway Framework to Seamlessly Access HPC, Grid and Cloud Resources Distributed Worldwide

Submitter: Roberto Barbera, INFN Italy

Summary: Proposes to present the Catania Science-Gateway Framework and the results of the program outlined by the CHAIN-REDS project to demonstrate standard-based interoperability amongst high performance computing, grid and cloud resources distributed in Latin America, in the US and the rest of the world and based on several different middleware stacks.

* * * * *

Focus Area: Digital Media

6) Title: Big Data, Big Networks, Big Displays -- the Next WAVE

Submitter: Tom DeFanti, UCSD

Summary: Video wall collaborations between CICESE and UCSD have been ongoing for several years. At UCSD, both 2D and stereo 3D walls are in use, displaying big data up to 64 megapixels resolution. Specific effort has been invested to optimizing high-speed local and distance file serving and collaboration using multiple 10Gb/s and 40Gb/s networks and software tuned to synchronized image sharing (SAGE), extremely high-resolution static and streaming image viewing (MediaCommons), and immersive virtual reality experiences (CalVR), as well as to driving surround sound and focused SoundBender audio speaker arrays. Recent results adapting flash memory big data technology championed by the San Diego Supercomputer Center to “FIONA” PCs driving 2D/3D big data displays locally with 40Gb/s network interfaces to impedance match multiple 10Gb/s connections, along with their applications will be presented, with focus on omics and archaeology, as two examples with great cross-border potential. The latest big displays at UCSD, the WAVE and WAVElet, and use of emerging UHDTV (4K) panels will also be described.

7) Title: Big Media, Big Networks

Submitter: Laurin Herr, Cine GRID

Summary: This session will introduce CineGrid, a non-profit membership organization that has nurtured a community of members from many countries over the past 9 years who collaborate to promote research and exchange of ideas in this area. This session could also include a panel discussion with participants from Mexico and the USA (to be decided) who would explore potential future projects between the two countries using a 10Gbps research network link.

* * * * *

Focus Area: Oceanography

8) Title: Numerical modeling for tropical cyclone landfall over northwestern Mexico

Submitter: Luis Farfán, Ismael Villanueva, Julian Delgado, CICESE

Summary: We studied ten tropical cyclones that, during the period 1996-2010, made landfall in northwestern Mexico from the Eastern Pacific Ocean. Our goal is to present a synthesis of the predicted tracks, wind fields and rainfall patterns around the landfall area. The model simulations start in advance of the actual landfall and results are also used to make an analysis of the interaction between the incident cyclone and coastal topography. Datasets issued by the National Hurricane Center are used to compare tracks and intensity with respect to observations as well as the official forecasts issued on a real-time basis.

9) Title: Transport and dispersion of hydrocarbons in estuaries of the Gulf of Mexico: an example of a proposed data-intensive application

Submitter: Arnoldo Valle-Levinson, UFL

Summary: The Gulf of Mexico Research Initiative was established, after the Deepwater Horizon oil spill in April 2010, to improve understanding of hydrocarbon transport in the Gulf of Mexico. In response to an upcoming call for proposals, a group of researchers from U.S. and Mexico plans to create a Research Consortium. This Consortium will concentrate in the study of hydrocarbon transport in estuarine systems around every Gulf State: Florida, Alabama, Mississippi, Louisiana, Texas, Tamaulipas, Veracruz, Tabasco, Campeche, and Yucatan. Participant researchers will not only come from these Gulf States, but also from other regions in the US and Mexico. Investigations will include field measurements and numerical simulations. Field measurements will consist of surface drifter deployments, dye dispersal studies, and shipboard measurement of environmental variables. Measurements and simulations will produce many Terabytes of data that will require effective storage, transfer, and access for a seamless international collaboration. The participation of CICESE and XSEDE in this Consortium will ensure such a seamless and effective international collaboration. This contribution will present the idea of the Consortium to seek feedback and possible strengthening of the collaborative effort.

10) Title: AMLIGHT, Simulation Datasets, and Global Data Sharing

Submitter: Jean-Bernard Minster SCRIPPS, John J. Helly, UCSD, Steven Day, San Diego State University, Raul Castro Escamilla, CICESE, Philip Maechling, Thomas H. Jordan, Southern California Earthquake Center, Amit Chourasia, San Diego Supercomputer Center

Summary: The advent of continental scale and intercontinental high-bandwidth networks opens up new options for the architecture of ICS U World Data System. No longer must large datasets be duplicated at many nodes and, in the concept of a system of systems, server nodes and client nodes can in fact be vastly different in terms of their scope, usage, and commitments, and may not necessarily maintain vast data holdings. At the same time, consideration must also be given to failure-modes and fault tolerance of increasingly critical network infrastructure. AMLIGHT should be a very practical testbed to explore the usefulness and applications of such concepts. We propose to initiate a broadly based discussion of these topics in conjunction with other stakeholders worldwide, especially in Latin America.

* * * * *

Focus Area: Nano-materials Science

11) Title: 1H-MoS2 nanoparticles grown on graphene and 1H-BN monolayers

Submitter: Donald Galvan, G. Alonso, S. Fuentes, UNAM

Summary: A theoretical study has been performed on a sandwich made by MoS2 nanoparticles grown on graphene and BN monolayers. The calculations reported in this study have been carried out by means of the tight binding method within the extended Hückel framework using YAeHMOP (Yet Another Extended Hückel Molecular Orbital Programs) computer package. The analysis performed includes Energy Bands, Total and Projected Density of States and Mulliken Population Analysis.

* * * * *

Focus Area: Bioinformatics

12) Title: Comparative Human Microbiome Analysis

Submitter: Larry Smarr, Calit2, Weizhong Li, UCSD

Summary: We are carrying out very deep metagenomic sequencing of human gut microbiomes from healthy subjects and from people with the autoimmune Inflammatory Bowel Disease. We compare one subject with IBD to metagenomic datasets downloaded from the NIH Human Microbiome Project repository, including 35 healthy subjects and 20 with IBD. We also analyze the changes in this one subject over multiple times, including comparing before and after drug therapy. The dataset of Illumina short reads for one person is ~10GB. The total comparison dataset contains ~0.5 trillion DNA bases. These Big Data had to be moved across the network to the San Diego Supercomputer Center where over 200,000 cpuhours were consumed in the analysis and then back to Calit2 where a 64 megapixel wall was used for visual analysis. This approach could be extended for cross-border comparisons of human gut microbiomes to examine differences in food intake and various disease states.

13) Title: Protein-protein interactions on viral self-assembly analysis

Submitter: Mauricio Carrillo-Tripp, LANGEBIO, Vijay Reddy, SCRIPPS

Summary: Our current collaboration involves the generation, and eventual transfer, of Big Data related to the atomic structure of viral capsids, the protein shell that protects and transports the viral genome from one host to the other. The main goal is to identify and understand the molecular mechanisms involved in the building of the these particles, which go through an spontaneous self-assembly process inside the cell. The data analysis needed in our research collaboration, which is not currently using the cross-border 10 Gigabit network connection, could greatly benefit from its use.

* * * * *

Focus Area: Seismography

14) Title: Current state of permanent GPS network in northern Baja California, Mexico: Challenges and opportunities for next decade

Submitter: Javier Gonzalez-Garcia, Alejandro Gonzalez-Ortega, CICESE

Summary: Currently, the permanent GPS network in northern Baja California consist of 19 stations, which produce ~40Mb/day of raw data with sampling rate at 15 seconds. As more GPS stations are planned to be installed during next couple of years and the sampling rate would be change to 1-10 hz for real time applications, it will become imperative to have appropriate data flow. With this, the data flow will be increased by ~10^3 more data than today, with low latency and real time communications would made possible Alert for big earthquakes giving at least 10 sec. to make important decisions before the arrival of the first seismic waves. With no doubt the topographic/survey community and engineering in general and navigation in particular all this and more will be benefit with this living revolution.

15) Title: Possible enhancement of UABC's cross border collaborations in seismology, seismic and fault mapping using geophysical data and LIDAR

Submitter: Octavio Lazaro-Mancilla, UABC

Summary: As researcher of Laboratorio de Sismología y Geofísica Aplicada de la UABC I have collaborated with researchers of Seismolab of CALTECH. Seismological collaboration include the deployment of temporary seismometers in field after April 4, 2010 El Mayor-Cucapah earthquake. For storage of data we have used flash cards that after were sent by mail to the Seismolab for analysis. On the other side, I have collaborated with researchers of CICESE, CALTECH, Virginia Tech and USGS in an active–source seismic imaging experiment in the Mexicali and Imperial Valleys in 2011. Nowadays CICESE, CALTECH researchers and I are working in a urban fault mapping project in the City of Mexicali using a variety of geophysical data and aerial images. In this project it is possible to do studies of LIDAR in urban areas in Mexicali and repeat them if possible after a certain time for detection of differences that can indicate subsidence and / or fault movement. For education and research I think that better facilities are needed for real time transmissions of large data sets such as LIDAR and satellite images including seismology data.

16) Title: Possible enhancement of Caltech's cross-border collaborations in Geology and Geophysics

Submitter: Joann Stock, Robert Clayton, CALTECH

Summary: Caltech Seismological Laboratory personnel have both formal and informal collaborations with researchers at CICESE, UNAM (Instituto de Geofisica), UABC-Mexicali, UABC-Ensenada, and University of Sonora. Seismological collaborations are extensive and varied and would benefit from improved speed of transmission between UNAM and Caltech, which may prove to be particularly crucial in the event of a major earthquake. For deployment of temporary or permanent seismometers, the ability to directly connect instruments to the internet and transmit data in real time is a strong advantage and is key to newer instrument designs. However data transmission was compromised in the Mexicali Valley area by the 2010 earthquake, which caused interruption of internet and wireless telephone services as well as infrastructural damage. For the educational and research collaborations, better facilities are needed for video conferencing including the ability for simultaneous presentation of large 3D data sets such as LiDAR point clouds, DEMs, aerial and satellite imagery, and high resolution bathymetry. Ideally we would like to jointly examine and work with these in real time, involving investigators on both sides of the border, for both one-on-one research collaborations and thesis committee meetings. A high-speed network is essential for future earth-science collaborations between Mexican and American scientists.

17) Title: Earthquake Monitoring and sharing data in real time in the North Baja California Region, Mexico

Submitter: Victor Wong-Ortega, CICESE

Summary: This project pretend to set up new instrumentation to measure seismic activity in the earthquake prone zone of Mexicali Valley, Baja California, México, to the Northwest Seismic Network of México (RESNOM) of CICESE and to the National Center for Prevention of Disasters (CENAPRED). Real-time data provided from these devices installed in the northwest of Mexico will be used to protect lives and property in north Baja California region. USGS, CENAPPRED and CICESE will collaborate to execute this project.

18) Title: Space Image Repository

Submitter: Enrique Pacheco Cabrera, AEM

Summary: Mexico presents natural and anthropogenic disasters of different categories and magnitudes, such as earthquakes, floods, forest fires, volcanic eruptions, emergencies in oil fields, etc. The Mexican Space Agency (AEM) proposes to develop a national system for storage and data processing, Geomatics, spatial and astrophysicists (SNAP-DGEA) seek to integrate, develop and consolidate infrastructure, storage, processing and distribution of such data, which represents BIG DATA and therefore an obligated need for BIG STORAGE and BIG NETWORKS for the distribution systems as well as capabilities of computation of very high performance for processing and visualization.

* * * * *

Various Focus Areas

19) Title: Integrating distributed LIDAR data repositories into the OpenTopography service infrastructure

Submitter: Chaitan Baru, Viswanath Nandigam, UCSD

Summary: Operating from the San Diego Supercomputer Center, the OpenTopography facility provides online access to high-resolution LIDAR topography data and tools by hosts Lidar datasets contributed by data acquisition projects and serving as a metadata hub for remote Lidar datasets. OpenTopography's multi-tiered software architecture—consisting of an infrastructure tier, services tier and applications tier—supports scalability and extensibility.

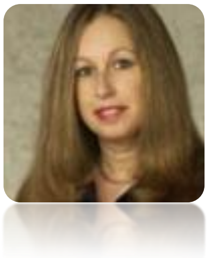
For example, CICESE in Ensenada, Mexico could host and operate data repositories with local data management procedures and data distribution timelines, while making these data available to a much larger user community via OpenTopography. Processing of these data could be performed at the remote location, or by services hosted by OpenTopography. This is technically accomplished by extending the OpenTopography service layer over the different data providers hosting the datasets, which requires packaging and distributing a customized instance of the Opal toolkit that would wrap the remote software application into OPAL-based web services and then integrating these services into the OpenTopography framework. The availability of a highspeed network links between CICESE in Ensenada and SDSC in San Diego would also enable caching of datasets on OpenTopography's storage servers, which can improve performance and make the system more robust.

20) Title: The Present and Future Development of Observatorio Astronomico Nacional in San Pedro Martir (OAN-SPM)

Submitter: Michael Richer, William Lee, UNAM

Summary: The site of OAN-SPM, operated by Instituto de Astronomia-UNAM, is among the best in the world for optical-infrared astronomy. Several medium and large scale projects in various stages of construction and planning will generate large amounts of data, requiring stable and broad connections for download and upload. Further, the raw and processed data will need to be analyzed and transmitted to sites depending on the particular collaboration partners. The ReIonization and Transients InfraRed project (operating), the Trans-Neptunian Automated Occultation Survey-II (under construction), and the San Pedro Martir Telescope (in planning) with the US among partners, all represent an internationalization of the Observatory, and a true bi-national astrophysics laboratory which will require, and greatly benefit from, a stable, high speed and broadband connection. We will describe the needs of these projects and their potential to advance cross-border collaborations in science, technology and the development of human resources.

21) Title: The Open Science Data Cloud: A Wide Area 10G Science Cloud Supporting Researchers in Science, Medicine, Health Care and the Environment



Submitters: Robert Grossman UChicago, Heidi Alvarez FIU

Summary: The Open Science Data Cloud (OSDC) is a six petabyte science cloud for researchers to manage, analyze and share their data and to get easy access to data from other scientists. The OSDC is operated by the not-for-profit Open Cloud Consortium (OCC), which operates cloud computing infrastructure for the research community. The OCC includes academic institutions, companies, and government agencies. The OSDC has international partnerships with scientists in the United Kingdom, Brazil, Canada,The Netherlands, Japan and China, and is always interested in expanding its international partnerships. The OSDC plans to begin interoperating with science clouds in other countries beginning in 2014.

22) Title: BigData-aware scheduling with uncertainty in Cloud Computing

Submitters: Andrei Techernykh, Jose Lozano Rizk, CICESE

Summary: organizations with terabytes of digital content. The scheduling of jobs on multiprocessors is generally well understood and has been studied for decades. Many research results exist for different variations of the scheduling problem. However, the big data communication-aware scheduling problems have rarely been addressed.

In this talk, we will discuss a model for cloud computing applications taking into account a variety of communication resources used in real systems. This communication-aware model, called CA-DAG, allows making separate resource allocation decisions, assigning processors to handle computing jobs, and network resources for data transmissions. We will discuss the benefits, weaknesses, and performance characteristics of such a model and resource allocation strategies in presence of uncertainty.