Appendix D – Cyberinfrastructure

Integrated Data Management and Synthesis Support

The BCN data management infrastructure will provide the core of the system. The data used to develop the BCN products as well as the products themselves are organized and archived within the data management system. The original source data used to develop the data products are archived to ensure that data products can be regenerated and the provenance tracked. The cyberinfrastructure will require petabyte-scale storage to enable the archiving of data, models, and other components used in developing calculations as well as the BCN data products themselves.

The underlying BCN data includes remote sensing data (images), individual sensor data (time series), observational data (text strings), biological information (varied), disturbance data (text strings), papers, reports, simulation data, model data, equations, analysis products, etc. A range of data Indexes and other database tools will enable organization of the data and fast searches on the data as well as fast access to the data. Automated data ingestion routines will need to be developed to enable integration of new data sources and updated data. Data movement between storage and processing systems will enable processing on the data. A wide array of tools for managing large-scale distributed data and staging data for use in calculations have been developed (Duan et al. 2005, Callaghan et al. 2010) and the BCN cyberinfrastructure will leverage these tools.

Web interfaces, data products, programmatic access

On-line, web-based, and map-based user interfaces to BCN data products at scales ranging from neighbourhood to regional and global scale will enable users to request version of the product tailored to the region of interest which may be as small as a neighbourhood and as large as the globe. The interface will provide access to products based on time and spatial position. For example global or regional views, or time series of indexes, could be extracted and viewed. In addition, trust in the BCN will only grow if the data product provenance is easily reviewed and checked to understand the data and models involved in the calculation as well as the fidelity of the estimate. Data mining and analytics mechanisms such as data cubes will be key enablers in the development of these interfaces. Furthermore, programmatic access to these resources will allow external software systems to leverage the generated data products.

Computational Infrastructure

The computational capabilities required will depend on the scale of the model, deadline constraints, and scale of the data. There are a number of ways to enable the computational infrastructure needed to support BCN. In the past, a dedicated centralized computing facility would have been the only answer. The advantage of a centralized facility is that everything can be stored and accessed locally and the computational priorities can be set by the user. However a dedicated facility can be very expensive to purchase and to maintain. Alternatives to that include grid computing services, cloud infrastructure systems and cloud-based platforms. Grid computing is a model where many participating sites offer computation and storage. Access control and coordination of many sites is complex with this model, but the reward is that existing computational resources can be leveraged and shared with many other applications and users. High Performance Computing (HPC) resources such as WestGrid, Compute Canada, TeraGrid/XD, and the National Energy Research Supercomputing Center provide options for distributed computing.

Virtualization of computational resources enabled the solutions known as Cloud Computing, which offer a promising direction for providing the cyberinfrastructure needed (Ryu et al. 2010). Clouds provide more flexible environments where custom virtual machines can be established to support different services. With a cloud model, resources from around the network can be gathered into a single large-scale execution framework to perform calculations and then they can just as easily be brought back down. Public Clouds such as Amazon Web Services are particularly interesting for their low upfront cost. Support/maintenance of the physical cloud infrastructure is handled by the provider and idle computational cycles can be returned to the system. In this model, only the computing resources used are billed.

Another emerging model for computational infrastructure is based on providing an entire deployment platform, which is usually is built on top of a Cloud infrastructure. With this approach, not only hardware, networking and server management tasks are automated, but also the databases, security and other fundamental software systems are automated as well. This further decreases systems administration costs, but imposes a vendor lock in with the platform provider. Microft's Windows AZURE and Google App Engine are two examples following this model. They both offer a platform for deploying applications, offering on-demand and transparently scalable processing capability, different strategies for data storage (including low cost, large scale repositories) and a collection of basic software services that simplify software development and deployment within the platform.

References Cited

Callaghan S, et al. (2010) Scaling up workflow-based applications. *Journal of Computer and System Sciences* 76 SI: 428-446.

Duan RB, et al. (2005) DEE: A distributed fault tolerant workflow enactment engine for Grid computing. High Performance Computing and Communications, *Proceedings Book Series: Lecture Notes in Computer Science* 3726: 704-716.

Ryu et al. (2010) Global remote sensing in a PC: cloud computing as a new tool to scale land surface fluxes from plot to the globe. *FluxLetter* 3(3):9-13. Available online at: http://bwc.berkeley.edu/FluxLetter/