BEYOND NETWORK: *REGIONAL ORGANIZATIONS* AS EFFECTIVE WORKFLOW INTEGRATORS

PREPARED BY DON DUROUSSEAU AND JARDA FLIDR
RESEARCH TECHNOLOGY SERVICES
DIVISION OF INFORMATION TECHNOLOGY

GW COLONIAL ONE CAAREN

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Beyond Network - Regional Organizations as effective workflow integrators

Since their inception, *Regional Organizations* have been tasked predominantly with providing network connectivity in their respective regions, aggregating Internet research traffic of multiple local institutions, and connecting them to the national R&E backbones. While availability of bandwidth, and high-performance data throughput have certainly remained a necessary condition for efficient regional and national collaboration, the *Regionals* – facing tough competition from local commercial network providers – can no longer justify their existence solely by relying on network-centric services. In order to prevail, they must look beyond the network domain.

XSEDE, Advanced CyberInfrastructure - Research and Education Facilitators (ACI-REF), National Supercomputing Centers, Open Science Grid (OSG), to name a few, have been the traditional end system destinations interconnected by the Regional and National R&E networks. However, along with the evolution of these infrastructures, the related research carried out in these environments has evolved as well. Computational researchers are no longer just the scientists who run large-scale, computationally intensive simulations on thousands of big-iron HPC cores. They can just as easily be domain scientists wishing to process a large amount of data from a scientific instrument, either continuously or in an on-demand fashion, where they might require access to a temporary database, or to deploy a short-term data server for collaboration and data distribution with an attached large volume storage array. None of the above-mentioned computational facilities can fully address such requirements in the time frames or functionalities required by these tasks and workflows, particularly when inexperienced researchers are involved. We classify such end systems as *hard*[1]. They offer powerful, yet slow-to-acquire, preconfigured resources with largely predefined and predetermined intended use, and steep learning curves[2].

Conversely, *Regional Organization* – thanks to their inherent network capabilities, optimal topological positions, and staff skill levels combined with their ongoing interactions with local researchers – have a unique opportunity to fill this functional gap between the *hard* resources and the agile requirements of modern, dynamic computational research. The *Regionals* can deploy high-performance, shared, reconfigurable and on-demand resources with relatively few restrictions (and few guarantees), while opening access to virtually any authorized user on a first-come first-served basis. As a result, the *Regionals* can offer a highly effective staging area for the National Computational Infrastructures by stitching together arbitrary scientific workflows with the high-end resources. Because of their functional and geographical mobility, we call these resources *fluid*.

For example, a scientific workflow, which moves and preprocesses hundreds of terabytes of telescope data in undetermined time intervals cannot be addressed effectively by the proposal-driven big iron. Long-term resource allocations to support intermittent, high-intensity, yet low-duty cycle workloads would be simply uneconomical and wasteful. Addressing such requirements locally on the organizational level would prove prohibitively expensive even for big institutional connectors. This is exactly where the *Regional Organizations* can step up to help smaller or less developed research institutes in their vicinity, because they:

- are on the crossroads of data flows
- are usually unhindered by institutional policies
- connect a large number of research organizations
- peer with the national R&E backbones

[1] hard as in *hard*ware: rigid, immutable

[2] our focus is on Domain Scientists without Computer Science expertise

As a concrete example, Capital Area Advanced Research and Education Network (CAAREN) is planning to deploy the following set of *fluid*, self-service resources to support smaller research institutions with short-term, high-intensity workflows and intermittent, on-demand resource requirements. Such services might include:

- high-performance Virtual Machine deployments in the OpenStack environment with VM-FEX, high-throughput technology and arbitrary OS and application workloads
- bursting into public or private cloud resources over direct peering, while creating hybrid cloud infrastructures whenever local resources become insufficient
- deployment of small services such as databases, web servers, or data transport frameworks
- short term, high-performance and high-throughput cluster storage, based on a geographically distributed *ceph* file system capable of sinking up to 100Gbps of data
- new technologies adoption (*e.g.* long-haul InfiniBand for a long-distance HPC cluster integration)

Some of these services are based on previously NSF-funded CC*NIE projects such as SDNX; others are logical extensions of core NSF programs like Globus, OSC OnDemand, GENI and XSEDE. Thus, the *Regional Organization* can provide an effective "*wrapper*" around the National Computational Infrastructure by terminating arbitrary scientific workflows and providing services that the *hard* resources are not well positioned, or willing, to offer.

In order to facilitate access to its resources, CAAREN will work directly with the regional R&E institutions to get them connected to Internet2, as well as on assisting researchers in the local schools with using a high performance computing system for their analytical needs. Our plan is to develop a set of recommendations capable of guiding each school through the steps needed to make the last-mile connections onto the campus, purchase the appropriate networking and storage equipment as needed and to connect their researchers to the infrastructure and fill the gaps in informatics training to more effectively coordinate educational needs in the small schools with the resources available for sharing at the large school. A side product of our efforts will be a best practices guide for research cooperation between larger research institutions like GWU and the smaller schools that have few internal resources to exploit training in HPC and advanced networking. We plan to test our solution for shared HPC at the local level in the Nation's Capital with the schools listed in the endnote[i] and then scale the approach of a unified training environment from a few campuses to a regional and potentially national level.

We have been working with the DC Government to provide last mile connections from the schools to the CAAREN infrastructure (e.g., DC-Net is the fiber optic provider to schools, offices, museums, libraries, etc. in Washington DC). Additionally, GWU will partner with the Office of the Chief Technology Officer and the Mayor's Office in DC to develop a strategic plan for each school to participate in upcoming Smart Cities research and development activities currently being incentivized through a joint partnership memorandum of understanding between GWU and the Mayor's Office to cooperate on the development of training curricula and research partnerships with R&E institutions in DC.

---

[i] list of potential schools in Washington DC not currently on CAAREN that are capable of participating in a shared infrastructure experiment targeting the delivery of specialized training for network engineers and researchers wishing to access to cloud-based HPC resources:

> Gallaudet University • Catholic University • Trinity Washington • De La Salle College • Washington International School • Institute of World Politics • Graduate School USA • Levine Music • Westwood College