April 1, 2010
Position Paper for NSF Campus Bridging Workshop

A High-Performance Campus-Scale Cyberinfrastructure For Effectively Bridging End-User Laboratories to Data-Intensive Sources

Philip Papadopoulos[1]
Larry Smarr[2]
University of California, San Diego

A Campus-Scale Dedicated 10Gbps Campus Data Utility

Enabled by nearly a decade of NSF investment, UCSD has been able to investigate, at campus scale, the use of dedicated optical fibers, or wavelengths on the fibers, to simplify the process of bridging data-intensive campus-resources from data generation or storage devices, internal or external to the campus, into end-users labs. To meet expected performance, much of this infrastructure must be on-campus, connecting the lab to the campus gateway or to campus data systems. This has been done in a fashion that is easily duplicated on other campuses.

Two major NSF awards made this cyberinfrastructure (CI) research possible. In 2002, our colleagues and we were recipients of an NSF Information Technology Research grant, the “OptIPuter” (NSF OCI-0225642)[3], that asked the fundamental question of “how would distributed systems be redesigned if bandwidth leaving campus laboratories was essentially unlimited?” In addition to the optical networking research, the OptIPuter project led to the design and software development needed to create tiled display wall OptIPortals[4]--scalable “termination devices” for ultra-speed data flows (typically 10Gbps dedicated per user) entering a user’s laboratory.

In 2004, we began developing a prototype terabit-class, campus-scale network instrument through the Quartzite[5] Major Research Instrumentation (MRI) grant (NSF CNS-0421555). The Quartzite system simultaneously supports network-intensive applications, while using the instrument to examine different hybrid network configurations. The three level central Quartzite switch supports both packet switching as well as dense wave-division multiplexing (DWDM) switching-both wavelength conserving switching or switching flows from one wavelength to another. Our work catalyzed the design of UCSD’s nascent research cyberinfrastructure overlay to the traditional shared campus Internet by utilizing inexpensive DWDM multiplexing and packet switching to provide many dedicated 10Gbps Ethernet to

[1] Director, UC Systems, San Diego Supercomputer Center(SDSC)

[2] Director, California Institute for Telecommunications and Information Technology (Calit2)

[3] www.optiputer.net

[4] See white paper submitted to this workshop by DeFanti, et al.

[5] The OptIPuter, Quartzite, And Starlight Projects: A Campus To Global-Scale Testbed For Optical Technologies Enabling LambdaGrid Computing (invited Paper) Larry Smarr, Joe Ford, Phil Papadopoulos, Shaya Fainman, Thomas DeFanti, Maxine Brown, Jason Leigh, Optical Fiber Communication Conference & Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC) 2005, Anaheim, California, March 6-11, 2005, CD ROM. [www.optiputer.net/publications/articles/Smarr-OFC-2005.pdf]

various labs throughout campus. Quartzite today has more than sixty 10Gbps paths that interconnect computing, storage, OptIPortals, and instruments at various sites around the UCSD campus. This means a total provisioned of 600 Gbps (1.2 Tbps bidirectional).

Quartzite allows for a centralized campus compute and storage complex that is connected to end-user labs via the 10Gbps dedicated optical wavelengths. This complex at UCSD is called the Triton Resource, a new facility being built at SDSC for UCSD and UC researchers. It has ~30 Teraflops of computing capacity with a total of 15TB of memory. Triton has a small Lustre parallel file system at 110TB usable capacity, which will grow to support large temporary storage (~1 Petabyte) in a new file system called DataOasis. The interconnect is 10Gbps Myrinet (416 MX ports, 32 10GbE) with 8x10Gbps channels connected to Quartzite, TeraGrid, and the Campus Research Network. The UCSD Triton Resource[6] includes the Petascale Data Analysis Facility (PDAF) with 9TB of RAM distributed among just 28 32-way SMP systems, each with large local memory(0.25 – 0.5TB memory per) and 4GB/sec of network I/O (120GB/sec aggregate). These “fat memory” nodes will be expanded even further next year when the NSF-funded Gordon[7] data analysis supercomputer is brought online at SDSC. The Quartzite network allows us to incorporate nodes of Triton as peripherals.

For example, the Community Cyberinfrastructure for Advanced Marine Metagenomics (CAMERA) provides a targeted “computational science destination” which has a 512-processor cluster and about 200TB of project storage on the 1st floor of Calit2. This resource is used by over 3500 remote users from over 75 countries. When the local CAMERA-owned infrastructure is insufficient to meet user demands, the Quartzite campus-scale infrastructure is used to directly mount at 10Gbps (or multiples thereof) the CAMERA data resources onto the SDSC Triton Resource. CAMERA computations can overflow into Triton without any explicit movement of data, providing a relatively seamless integration of resources.

The Dawn of inexpensive 10G Ethernet

Quartzite was a multi-million dollar investment to help us investigate how the next-generation of campus networks should be designed and implemented. In 2005, incremental port cost on a Force 10 E1200 switch-router (used as the core of the network), was approximately $5000-$7000. Today, mid-scale (48-port) 10GbE switches with modest routing capability are about $500/port. Larger switches (100s of ports) with more complete routing capabilities are entering the Market in the first half of 2010 with expected pricing of about $1000/port. In other words, laboratories on campuses can be connected with significant bandwidth for the price of one or two high-end servers. This changes the economics and fundamentally allows remote campus resources to be brought “virtually” into laboratories, via the switched optical fiber infrastructure. Storage and computing (the fundamental elements of cloud computing) therefore no longer need to be located in the end-user lab, but can be elsewhere, allowing for economies of scale in these common resources.

Whither the Grid and Enter the Cloud?

The Grid as envisioned in the mid 1990s sold itself as a way to knit together distributed resources to form low-cost supercomputers. Our community has learned a great deal from the extended grid experiment, and perhaps the greatest lesson was that, in general, most lab scientists found the Grid too difficult to use and simply refused to expend the effort needed to

[6] Triton is directed by Papadopoulos

[7] www.sdsc.edu/News%20Items/PR110409_gordon.html

get “over the hump.” We argue that another issue with the Grid was that infrastructure and data did not appear as local or locally-controlled resources by the user.

It's not a surprise that cloud computing has captured the imagination of scientists, because they could control the definition of their resources while not owning actual hardware. There are probably a large number of small-data needs (email, social networking, photos, etc.) for which commercial clouds will be very useful, since the shared Internet and commercial cloud systems are well engineered for megabytes of data.

However, the cloud (especially in its current commercial forms) does not necessarily meet the needs of data-intensive scientists, for which terabyte-sized data manipulation is the norm. In CAMERA, for example, an annotation data set directly mounted on Triton is 1.6TB and is re-processed in multiple passes. Modern gene sequencers, which are appearing in increasing numbers on campuses, can easily produce a terabyte per run in less than a day. The problem is that a terabyte would take more than ten days to move from the lab to a remote cloud at the usual 10 Mbps achieved on heavily shared wide area networks. Using Amazon published transfer prices, it would also cost about $300 to copy a Terabyte data set in and out of a commercial cloud just once. Neither of these numbers makes the cloud “easy to use” or time practical for large data projects.

The Need for Data-Intensive Campus “Clouds”

On the other hand, with a well wired campus with many 10Gbps optical paths, such as UCSD, the same terabyte takes only ten minutes to transfer from lab to campus cloud. Furthermore, the campus cloud can be engineered for high I/O storage and tightly coupled high performance computing clusters, as in Triton, neither of which make financial sense for commercial clouds. As we mentioned CAMERA (and several other projects at UCSD using Triton) can have their locally-owned storage mounted directly onto such a community resource. What about reverse? That is, having centralized storage directly mounted onto lab resources.

In theory, this is no harder, but security and service scalability become important for practical implementation. This is where elements of the Grid, namely identity management and identity proxy are technically quite important. A major step forward is the Indiana University “Data Capacitor” which provides temporary file space at a campus scale. But from our view, scientists need permanent online storage that can used directly in their labs with sufficient performance. The pricing of 10Gbps makes it financially practical to provide the wiring at campus scale to make data transfer times minimal. However, there are significant technical issues to solve to implement centralized storage that meets performance, security, and data integrity requirements.

Software Integration of Community and Local Resources

Most users build local infrastructure (eg. domain scientists build their own cluster and storage), because they often need to control the software structure. They need their analysis codes (many of which are home grown) to work on their data. It's clear that having every lab “roll their own” is both highly inefficient and creates islands of data. Yet, if the CI community that understands and has a track record of building scalable infrastructure wants to impact users, special attention has to be paid to “ease-of-use,” meaning making the remote infrastructure behave and perform as close to local infrastructure as possible. This is still an open challenge.