UC Research Cyberinfrastructure Meeting

October 10 – 11, 2005

Summary Notes and Recommendations

The UC Research Cyberinfrastructure Meeting, held October 10-11, 2005, at UC San Diego, brought together a community of faculty researchers and IT professionals committed to positioning UC to utilize the full potential of advanced information technologies to strengthen UC's research competitiveness in multiple disciplines. The meeting was sponsored by the Council of Research Vice Chancellors, the UC Information Technology Leadership Council, and the California Institute for Telecommunications and IT (Calit2) at UCSD with additional support from the Center for IT Research in the Interest of Society (Citris), the Industry University Cooperative Research Program and UC Discovery Grant Program, and the Corporation for Education Network Initiatives in California (CENIC).

Seventy people attended the event, including several participants via video conference. Participants represented nine of the ten UC campuses; the UC Office of the President; the Lawrence Berkeley, Lawrence Livermore, and Los Alamos National Laboratories; CENIC; and the National Science Foundation.

The meeting's agenda, presentations, and other materials are collected at http://www.ucop.edu/irc/itlc/cyber.htm. This document serves to summarize discussions and recommendations relative to IT / cyberinfrastructure needs that could be addressed on an inter-campus or Universitywide basis.

Summary of Major Recommendations Made during the Meeting:

Major recommendations put forward during the meeting brainstorming sessions include the following:

Planning

- Develop a 10 year vision/plan for UC cyberinfrastructure which accounts for stable, evolutionary technology with services and support to end users.
- Develop and communicate a non-technical business case for strategic investment in cyberinfrastructure (for leadership support and funding).
- Create a Research Computing Group under the IT Leadership Council to facilitate collaboration among UC campus, medical center, and lab organizations responsible for delivering and supporting research computing infrastructure and services. (Note: The group is being formed.)

Funding

- Submit an NSF proposal (with private sector partners) to prototype state-of-the-art regional cyberinfrastructure.

- Benchmark cyberinfrastructure spending in other "best practice" states.
- Identify stable funding streams for research cyberinfrastructure (for example, through redirection of contract and grants overhead – a "1% solution").
- Identify a seed funding strategy to launch new initiatives as demonstration projects while long term funding strategies are developed.

Network

- Create a UC advanced research network (dedicated lambdas operated by CENIC) to provide inter-campus, state, regional, national, and international advanced networking capability to the UC research community.
    - Identify faculty on each campus to be early adopters of this strategy.
    - Develop secure, reliable networking services for this dedicated network.
    - Leverage existing collaborations in the medical sciences (e.g., California health care and telemedicine initiatives).
- Create a demonstration project to develop a "sensor internet" connecting distributed devices for data collection.

Data Stewardship

- Develop a networked, layered approach for data preservation and exploitation.
- Form academic/industry partnerships to facilitate academic use of large, distributed federated data repositories.
    - Leverage existing competency centers and resources (e.g., CDL, SDSC).
    - Study/survey existing preservation solutions.
- Increase UC's representation in forums that influence the policy and funding affecting UC's potential as a digital steward.

High-Performance Computing

- Identify strategies to improve utilization of existing computational capabilities at the campus and Universitywide levels (e.g., campus-based clusters, San Diego Supercomputer Center (SDSC), National Energy Research Scientific Computing Center (NERSC).
    - Develop statements of best practices that campuses can adopt to provide appropriate levels of support to their local cyberinfrastructure.
- Create "UC Grid," a blueprint for distributed, networked, high performance computing capabilities designed to address inter-campus or Universitywide research computing needs.

MEETING SUMMARY: Day 1

The first day of the meeting consisted of presentations by providers of cyberinfrastructure services and researchers in various disciplines concerning current research trends and their reliance on information technology, as well as future cyberinfrastructure needs. The following is a brief summary those presentations. The full presentations are available at http://www.ucop.edu/irc/itlc/cyber.htm.

Cyberinfrastructure Direction at NSF/NIH and Implications for UC/DOE

Computation facilities such as the San Diego Supercomputer Center (SDSC), the National Energy Research Scientific Computing Center (NERSC), the TeraGrid, and the Open Science Grid provide high-end computing resources within UC that provide a third-level tier in a hierarchy of resources that includes small-scale and home computers (tier 1), as well as computing clusters at the campuses (tier 2).

UC can lead in the creation of a regional cyberinfrastructure, supporting the campuses with links to national and international cyberinfrastructures.

UC/National Network Infrastructure Update

CENIC (Corporation for Education Network Initiatives in California) is positioned to provide a regional, high-performance research network for UC, as well as links from that regional infrastructure to national and international networks, such as Internet2, the National Lambda Rail, Pacific Wave, and the Global Lambda Integrated Facility.

CENIC services are structured in three tiers to achieve economies of scale while addressing the varied needs of research and education. The first tier (CalREN-DC) provides high-quality, stable service for the K-20 community. The second tier (CalREN-HPR) provides high-performance research networking for large applications. The top tier (CalREN-XD) provides “bleeding edge” services at the very high end of research networking.

UC Campus/Lab Research Computing Cluster Survey Results

Preliminary results of a survey of research computing cluster capability within UC were presented. The results included 130 clusters at 6 campuses and one national lab. The information presented included the research discipline served, the types of hardware and software resources utilized within the clusters, the clusters' Internet connectivity, the clusters' size and capacity, and the source of system support.

Campus/Lab Research Computing Support Strategies

The demand for high-end computational resources is growing rapidly on the campuses and labs. UCLA, UCI, UCD, and LBNL, among others, are addressing this increased demand through the formation of groups and managed technology services to address campus/lab research computing needs. These units are addressing the following and other issues:

- provision of high-end computing resources
- provision of high-capacity storage resources

- provision of data center facilities (space, power, air conditioning)
- provision of system administration services and assisting other campus system administrators
- development of custom software
- support for commercial software, as well as licensing
- support for system acquisition
- scientific visualization tools and capabilities
- collaboration tools
- cybersecurity
- assistance with grant writing
- financial support models

Requirements for Information and Data Curation

Information in physical form has long been preserved in well-protected storage facilities by libraries, museums, and universities. Historically, data stewardship has been a largely independent activity of each institution. The advent of information in digital form, however, has introduced the opportunity for large-scale collaboration among data stewards, using a common infrastructure providing

- file storage,
- security,
- network,
- migration and emulation strategies,
- common tools and services for ingest, automated metadata augmentation, etc.), and
- appropriate policy environment(s).

Such an infrastructure can empower and introduce cost efficiencies into local efforts, such as

- selection,
- documentation,
- curation,
- access authorization,
- presentation/visualization,
- presentation planning, and
- community building, outreach, training (with data creators and users).

The California Digital Library has built a prototype digital preservation service. For UC to achieve these goals fully, it must overcome technological, organizational, linguistic/emotional, and financial challenges.

Faculty Overviews of Research Directions and Cyberinfrastructure Requirements

Mathematical and Physical Sciences

Faster and larger clusters are required for parallel computation and image processing. Correspondingly faster network end-to-end performance is also required for communication among multiple clusters. The end-to-end performance requirements will likely require institutions to rethink their network designs and bandwidth allocation strategies.

Humanities and Social Sciences

The human sciences require cyberinfrastructure to "record, annotate, extract/identify and comprehend cultural expressions whether historical or contemporary, and in all their relations." In particular, the following technologies are needed:

- search engines
- translation tools, both linguistic and cultural
- concordances
- visualization tools and GIS mapping devices
- gallery builders and tele-immersive environments
- mobile phones and PDAs
- grid ware

Data used by research in the social sciences has a number of unique properties:

- variety of kinds of objects – quantitative, video, text, geographic, administrative
- complexity of the objects – intrinsically complex (e.g., interpersonal interaction) and extrinsically complex (context of the interaction)
- incomprehensibility of the objects – problem of coding and understanding
- intractability of the objects – privacy, secrecy, confidentiality, and copyright

Because of this, social scientists require:

- new kinds of data collection methods
- detailed metadata and indexing systems that deal with the complexity of the objects
- software that transforms data into useful measures and that searches for patterns
- timely and broad ranging archiving efforts
- sensitive policies regarding the use and archiving of data
- institutions that support these efforts

Biological / Life Sciences

The size of populations used in life sciences research is increasing, and data maintained for each member of the population is large. This results in the following challenges:

- pooling/sharing data
  - very large data sets
  - multiple institutions
  - confidentiality / security
- robust image computing algorithms for large populations
- computing imaging-based statistics
- predicting outcomes & optimizing treatment (simulation) demands high capacity computing and network resources
- access to experts who know how to conduct the study and analyze the results
  - highly multidisciplinary
- discipline-focused software tools
- extended time to pursue research & development across disciplines
- visualization is important to guide analysis
- processing generates data that is interpreted and must be stored for future reference

More specifically, the life sciences have need of cyberinfrastructure support in the following areas:

- facilities (space, power & cooling)
- networking
- high-performance computing systems (hardware, software, and support)
- high-performance computing user support
  - application development
  - parallel code optimization
- data support (support for large-scale, heterogeneous, & distributed scientific databases
- software infrastructure for system integration

Collaborative studies require that cyberinfrastructure become as ubiquitous as the telephone and Internet.

Engineering and Computer Science

Two things are happening that impact cyberinfrastructure needs:

- Systems are becoming increasingly complex.
- Algorithms are becoming increasingly sophisticated.

A deep understanding of performance issues is required at all levels.

MEETING SUMMARY: Day 2

Following the presentations from Day 1, attendees shared observations in summary:

- Institutions are needed to make things happen; organizational service capabilities are required
- Benchmarking strategies are needed, as are metrics for success.
- Need to identify opportunity costs
- Business models need to be developed to demonstrate economies of scale (e.g., co-location)
- UC should pursue engagement with industry (identify collaborators)
- Need frameworks for moving from prototypes to services
- Clarify UC goals regarding research cyberinfrastructure
- Highlight demonstration projects that address national needs
- Timeframes for goals should be 5-10 years (stability/evolution)
- Identify efficiencies to achieve now
- Develop the value proposition of UC research cyberinfrastructure – targeted messages to sponsors and funding sources
- control vs. influence

In order to provide more focus to specific issues, four breakout groups were formed to discuss:

Networking (Group A): Initial topics identified for discussion:

- 10 gigabit connectivity
- intra-campus capability
- disaster recovery
- on-ramps to CENIC XD and NLR
- cyber security
- campus management infrastructure to support XD

Computational Capacity (Group B) Initial topics identified for discussion:

- cluster management strategies
- collective resources (UC Grid)
- data capabilities
- medicine – tech transfer (needs)
- on ramp from computational clusters to Universitywide facilities and resources
- visualization
- computation-driven networking needs
- strategies to maximize research and education drivers

Data Stewardship (Group C) Initial topics identified for discussion:

- data collection
- data repositories
- digital preservation

- data flow
- virtual reality, visualization, data mining / access / search
- data integration
- data capture
- policy (security, ID management)
- human subjects / confidentiality

Harvesting Externally Funded Faculty Entrepreneurship for UC (Group D) Initial topics identified for discussion:

- initiative with state (funding similar to Instructional Technology)
- federal funding targets / UC initiative
- leverage California national leadership through UC efforts
- galvanize congressional leaders (state /federal) re: UC agenda
- harvesting research competitiveness
- collaboration incentives

Following are summaries of these break out group discussions.

Networking (Group A)

California has been a leader in developing advanced networking initiatives; however, California is no longer keeping pace with technology advances. Other states have seized the leadership position. Some examples are the Texas Learn Initiative, Louisiana LONI, Illinois I-WIRE, and Indiana I-Light initiatives which provide advanced statewide networking capabilities such as optical overlay networks (Illinois) and on-ramps to high-speed backbone networks such as National Lambda Rail. These initiatives translate into competitive advantage for their states in pursuing advanced research opportunities.

There are four tiers of users who can be targeted by a new initiative in network infrastructure. These tiers can be thought of as a pyramid.

- The **first tier** of the pyramid corresponds to networking to the home. A recent initiative in this area was the Gigabit or Bust initiative targeted to demonstrate the value of bringing high-speed connectivity to homes and businesses in California. .
- The **next tier** is K-12. Several states have initiated statewide programs targeting high speed network connectivity to the schools. CENIC provides networking services to the K-12 community in California.
- College campuses are the **next tier** and although they typically have good bandwidth, the network capacity, services, and technology used in the campus infrastructure is typically out-of-date and insufficient to satisfy many emerging instructional and research needs. Although a few UC campuses have already upgraded their networking infrastructure, a majority of the campuses need resources to develop the justification and fund / execute a plan for high performance networking at the campus.
- The **top tier** of the pyramid includes researchers in need of advanced networking capabilities and technologies beyond what has been deployed. In order to justify such

a network, a business case with cost/benefit analysis must be developed. In each of the four tiers, there are opportunities to pursue and a cost associated with not pursuing them.

There are several opportunities which were identified by the group as potential projects to pursue. These are outlined below:

Advanced Network Infrastructure. Target the top tier of research computing users and build the use case for an advanced UC research network which provides capabilities beyond what is available in any other network or test bed today. The UC campuses would need to survey their various researchers to find a few who can articulate break-through capabilities and advances which would be enabled by the advanced network capabilities. Once the use cases and needed capabilities are identified, they can be matched with the appropriate computer science researchers to make a convincing case for the research and development required to provide the network. If possible, the network should be designed to be dynamically configurable to easily address new opportunities to serve additional researchers. Such a project could be undertaken with seed funding to connect a sample set of sites and demonstrate the value of such an infrastructure.

Disaster Recovery. Put together a state-of-the art UC disaster recovery system strategy (both the infrastructure and application). This system could be designed with the appropriate levels of infrastructure redundancy to ensure continued operation during a disaster and access to the necessary data and applications to assess damage and conduct recovery operations. All the UC campuses already have redundant network connects but this redundancy would need to be throughout critical campus infrastructure and would require extensive cooperation between campuses to ensure continued operation even when an entire campus is off-line. Such a disaster recovery system and capability could serve as a model for the state and the nation. Note: The IT Leadership Council has tasked the campus data center managers with collaborating to produce a systemwide data center disaster recovery strategy.

Open Collaboration and Cyber security. Develop high-bandwidth network services which can provide appropriate levels of cybersecurity while supporting open collaborative science. Today most high-performance science networks by-pass cybersecurity measures leaving these networks vulnerable to cyber attacks. Today's science collaborations are often global in nature and involve extensive trust relationships between sites and individuals, these trust relationships can serve as an attack vector from site-to-site. Development of a network which can support global collaborations while preserving and protecting the resources and the ability to conduct science is needed. Such a secure network would serve as a model for the many large science collaborations today and government and industry in the future.

Sensor Networks. Develop a demonstration project wiring together all the sensors and sensor networks throughout the state such as environmental sensors, traffic sensors, geological sensors, and others into a "sensor Internet."

Other Potential Projects. Several additional projects which could demonstrate advanced networking and significant benefit to the UC system were also discussed. These projects include such endeavors as:

- Take advantage of the Pacific Wave portion of the National Lambda Rail which links the Pacific coast together and build on the success of iGrid 2005 by providing a networking capability and enabling innovation in new applications to take advantage of the networking capabilities.
- Tie together the data and distributed computing centers across the UC campuses so as to provide a UC-wide computing and data storage fabric containing a range of capabilities from supercomputers to clusters.
- Move to IPv6 networking capabilities throughout the campuses was also discussed as a project which might enable advanced applications on the network.

Regardless of which of these projects are chosen to be pursued, every effort should be made to include the UC managed national laboratories in the projects and to leverage industrial partnerships whenever possible.

Computational Capability (Group B)

There are many issues in cyberinfrastructure computational capability that could be improved dramatically by a strategy of cooperation that is an acceptable to researchers and central IT at the local level. At the UC systemwide level, incentives might be used to facilitate local-level implementations, for example grants, funding, preferential resource allocations, etc. Any system wide actions would need to be closely paralleled with campus grass-roots efforts which would likely have greater impact.

Identifying best practices is a good place to start. Publishing practices used at relatively advanced campuses, detailing how they are dealing with computational problems, would certainly benefit those campuses that are less developed in these areas. Everyone recognizes that economies of scale are important, but how does one achieve them in practice? What pricing strategies are campuses using currently to attract research groups into a centralized management structure?

Campuswide Issues

Despite the best intentions of central IT administrators to help researchers, there are the usual differences of opinion expressed by researchers that make it difficult to converge on a consensus support strategy. Central IT organizations have demonstrated their abilities to scale resources effectively, while providing support that is cheaper, better, and more secure, than in the case of individual researchers who use student labor to maintain their research infrastructure. IT administrators need to stay focused on the researchers' needs as the solution is proposed and developed; those needs are to save money, to enable science that couldn't be done any other way, to do this in a manner that is time critical, and to provide impeccable service.

Proposed Action: Develop statements of best practices that campuses can adopt to provide appropriate levels of support to their local cyberinfrastructure.

UC Systemwide Issues

Greater central coordination of campus resources might obviate the need for some researchers to purchase new equipment. For example, if a clustering software package were developed to better utilize "idle cycles" within UC, UC systemwide cluster capability would put us in the top 50 supercomputer sites. Even at the local level, the ability to grab cycles from laboratories, classrooms, and even the computers used by administrators, a la SETI at home, would give researchers unprecedented computational capability at little additional cost.

Proposed Action: Investigate systemwide opportunities for sharing high-performance computing capabilities within UC.

Other Issues

- Researchers find it difficult to give up control over their equipment to the centralized approach which usually is less responsive to their needs. Thus, if there were systemwide disincentives (i.e., more expensive) to "go it alone," then there might be a greater use of central resources. A specific example occurs when researchers expect their Deans to provide needed space, cooling, power, and support for their research computing resources, for example by paying to upgrade HVAC and electrical outlets, when central IT groups already have such facilities.
- If a local, or systemwide, action were to be taken toward better integration of cyberinfrastructure resources, it would be important to pick a scientific issue of high-current interest, for example one in: life science, terrorism, global warming, data mining, etc., where needed integration would provide a significant dividend when the event eventually occurs.
- There are several approaches that could be taken to obtain better cyberinfrastructure integration. A "bottom-up approach" would focus on the needs of the researcher and preserve the flexibility and control that they expect. A "top-down approach" would focus on preserving campuswide cost of ownership, energy and space costs, etc. A middle approach would focus on integrating existing high-profile/high-strength projects into a broader collaboration. In all cases, we need to promulgate greater awareness within the researcher community by talking with them, or by conducting surveys, to identify science-driven opportunities.
- In short, the objective should be to solve timely/high-impact scientific problems by developing a seamless, system-wide computational capability.

Digital Stewardship (Group C)

This group addressed potential actions the University can take to improve its ability to provide appropriate stewardship of information in electronic form.

Data Preservation and Exploitation

Focusing on disciplines with high-impact or high-demand needs, such as social sciences and health sciences, develop a networked and layered approach for data preservation and exploitation services. This approach should foster a scalable, extensible infrastructure by exploring practical solutions to key preservation problems including:

- development of data capture and collection development tools
- networking to connect front-end data units with back-end preservation facilities
- replication, redundancy, and federation among back-end facilities

Existing infrastructure and experience (e.g., data libraries, UCSF, SDSC, and CDL) can be leveraged, as can high-probably funding sources, such as the State of California and NIH. Potential outcomes include:

- an exploration and articulation of benefits to science, business partnerships, and revenue generation that flow from data access, visualization, federation, and other data exploitation services
- the beginning of a policy environment within UC and amongst key agencies that have an interest in using these facilities
- the start of a growing confidence in UC as a trusted repository

Academic Use of Massive, Distributed, Federated Data Repositories

Form academic and industry partnerships to develop and evaluate solutions to key access/discovery/ontology issues, such as:

- resource discovery/location
- data mining
- automated metadata enrichment and generation
- use of social search
- dynamic ontology
- visualization, content mapping/clustering

through the use of large, extant, Web-based materials generated by, for example:

- in response to specific events (presidential elections, gubernatorial recall, Katrina),
- in reflection of specific themes (environmental, agricultural), or
- by specific institutions (the University of California)

Survey of Existing Solutions

Conduct a study of well-served disciplines (e.g., astrophysics) to identify possible preservation solutions and to learn what impacts might accrue to local infrastructure provision and practice.

Policy and Funding

Institute a proactive approach to UC representation in forums that influence the policy and funding landscape affecting the University's potential as digital steward. For example:

- federal and other funding agencies that have influence over demand for preservation services and

- state and federal legislative efforts that affect the University's ability as a digital steward.

Harvesting Externally Funded Faculty Entrepreneurship for UC (Group D)

This group addressed the need not only to deploy and support the cyberinfrastructure that is needed now, but also to position the University to be able to deploy and support future cyberinfrastructure needs. It requires a balancing between the needs of most researchers and the needs of innovators in the use of cyberinfrastructure, recognizing that the innovators point the way toward what will be common requirements in the future. (This is not just for the elite.)

Funding

Funding strategies must be both sustained and opportunistic. The people and technology needed to support the stability needs of most researchers must have stable funding, while creativity and opportunism is required to support new, innovative applications. Funding sources include federal and state agencies, industry, and the University itself.

While the sustained funding requirements are large, there is likely a "1% solution," for bringing additional financial resources to bear (for example, redirect 1% of UC's current overhead charges to common cyberinfrastructure development).

Proposed Action: Benchmark the research cyberinfrastructure spending in other, competitive, states, such as Texas, Illinois, and North Carolina.

Strategic Communications

It is important to "galvanize" potential funders of UC's cyberinfrastructure. They need to understand that funding the cyberinfrastructure is critical for:

- Competitiveness. A viable cyberinfrastructure makes UC competitive for attracting top-notch researchers and funding for research. It is also fuel to drive economic competitiveness at both the state and federal levels.
- Solving compelling social problems. The cyberinfrastructure enables research to help avoid or mitigate diseases, such as the avian flu, and natural disasters, such as Hurricane Katrina.

Proposed Action: Develop a strategic communication strategy to get the message out that cyberinfrastructure is crucial.

Analogues to Cyberinfrastructure

While cyberinfrastructure is relatively new to the University, it has some similarities with other infrastructures.

- The telephone system is a mature technology, and it is accepted that all researchers should have access. The infrastructure is generally funded through recharge to departmental and grant funds.
- The electrical power system is a very mature technology. It is accepted that all

researchers should have access. The infrastructure is generally funded centrally by the campus. The cost of that infrastructure is likely greater than that for a robust cyberinfrastructure.

Both the telephone and electrical power systems, however, differ from cyberinfrastructure in that there are (no longer) significant innovators. We have a good idea of required capacity for nearly all applications.

Systemwide Planning

The group discussed, but did not resolve, the issue of what systemwide planning should continue with respect to research cyberinfrastructure. A proposed UC Information Technology Guidance Committee, comprised of academic and administrative leaders across UC, will be a key forum in which to advance this planning process.

It is clear that campuses need plans, and there may be a systemwide complement to those plans. Examples of the result of such planning processes include:

- **CENIC**. Every campus is responsible for their network planning, but systemwide planning resulted in the creation of CENIC, California's network provider for higher education, which continues to be a key component of UC's cyberinfrastructure.
- **The California Digital Library (CDL)**. Again, each campus is responsible for their library, but a systemwide planning resulted in the creation of the CDL, which provides a crucial set of resources and services in support of all campus libraries.