

Community-Cyberinfrastructure-Enabled Discovery in Science and Engineering

A community cyberinfrastructure would enable a new era of multidisciplinary research and collaboration in science and engineering. With such an infrastructure, researchers could share knowledge and results along with computing cycles, storage, and bandwidth. A generic, transparent cyberinfrastructure would also foster more meaningful analyses of data and visualization, modeling, and simulation of real-world phenomena.

Traditionally, researchers have conducted scientific experiments in a laboratory environment using application-specific instruments and then manually recorded and analyzed their measurements using mathematical and statistical techniques. Observations gathered from such experiments were instrumental in validating theories and discovering new and, at times, unexpected phenomena. This process of theory and experimentation is, in short, how science has been conducted for the past few hundred years.

Today, however, most experiments generate observations that we can't record, let alone analyze manually. Examples include the use of mass spectroscopy for analyzing cells, DNA sequencing, particle physics experiments, and astronomy. This exponential increase in the amount of available information has prompted researchers to look elsewhere for help to efficiently handle and analyze data. With the advent of computing, researchers can now afford fast and efficient stor-

age of experimental observations and analysis at unimaginable speeds.

Computing in science and engineering has both enhanced the process of discovery and enabled new venues, including systems biology, ecosystem modeling, and individualized medicine. Hence, a growing number of discoveries in all fields of science and engineering are occurring at the intersection of disciplines. With increasing reliance on computing technology, there's a compelling necessity to develop a community-based computing infrastructure, or *community cyberinfrastructure* (CCI),¹ that allows interdisciplinary and intradisciplinary collaboration. In our vision, this CCI is shareable and transparent,² and it offers a place where scientists don't have to be concerned about underlying infrastructure design and operational issues but can concentrate on science inquiry and devising *in silico* experiments.

To translate this vision into a tangible infrastructure, we elucidate its specific requirements in terms of services, resources, and deployment. The functional architecture we describe lets multiple disciplines and researchers collaborate and conduct experiments.

The Silicon Shift

In the context of experimentation in science and engineering, computing was initially viewed as yet

another laboratory tool—akin to test tubes and multimeters—that could enhance the process of experimentation and observation by aiding information storage and analysis. The reality today is that experimental observations aren't just stored digitally but are often “born” digital. Consider the 3D computer model of arterial blood flow built to aid in the treatment of circularly diseases.³ The model uses multiple supercomputers in parallel through the TeraGrid (www.teragrid.org) to determine exact blood flow, so each run can generate hundreds of megabytes of data that are digitally saved with no analog artifact of the experiment.

This profound shift allows more meaningful analyses of data, visualizations based on immersive environments, and models and simulations of real-world phenomena. Additionally, the collaborative space afforded to researchers has created an environment not seen before in science. New phenomena are uncovered not because of an actual experiment but are based on modeling, simulation, and distributed data analysis. Interaction between the researcher and data occurs through an immersive environment in which the researcher uses visual representation, interaction technologies, and analytical reasoning⁴ powered by such technologies as haptic interfaces. Consequently, computing isn't only enhancing discovery but enabling it. A case in point is the field of systems biology, which is the study of the interaction between a biological system's components.⁵ The systems biology approach is characterized by a cycle of theory, computational modeling, and experiments to describe cells and cell processes—hence, systems biology exists because of computing. Accordingly, the research community has termed *computing* to be the third pillar of science, together with theory and experimentation.⁶

This evolutionary change in the way scientists conduct experiments is marked by what we call the *silicon shift* (see Figure 1). Prior to this period, most experiments occurred in the lab with actual subjects and scientific equipment. The advent of computing changed this scientific methodology gradually by helping the researcher gather, aggregate, analyze, and report data and findings. With the increase in computing power, data storage, and network bandwidth, more result verification and aggregation is deputized to the computing infrastructure. This happens at a defining point in experimentation, where *in silico* (modeling and simulation) becomes the norm, and researchers use experimentation (wet lab) solely to validate results.

Furthermore, most scientific discoveries aren't the result of one person conducting experiments

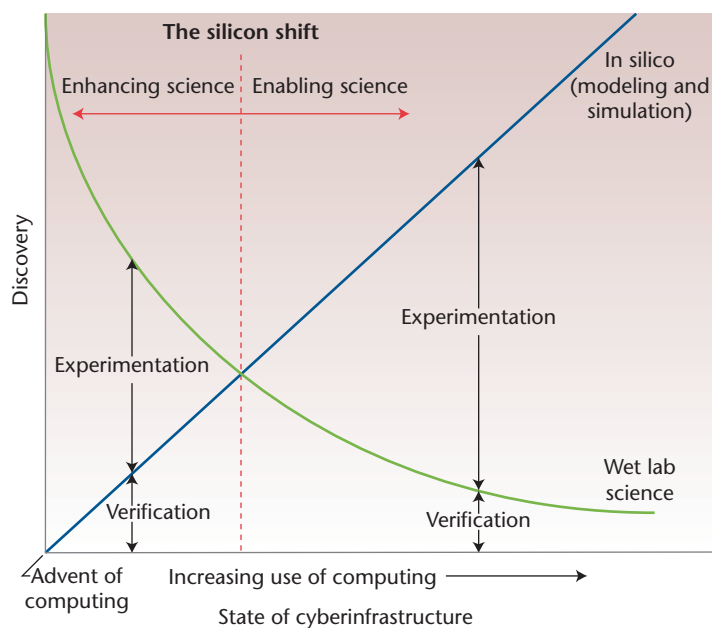


Figure 1. The silicon shift. Prior to this shift, computing enhanced discovery, whereas now it enables discovery.

in a laboratory but involve whole communities of researchers, collaborating to find solutions to various aspects of the same problem. A case in point is the use of nanotechnology for biomedical research, where the fields of chip fabrication, electronics, physics, and biology come together to discover phenomena that wasn't possible in individual disciplines (see the “Purdue Ionomics Information Management System” and “nanoHUB” sidebars for examples of CCI development in the Cyber Center at Purdue University).

The key role of cyberinfrastructure in science and engineering has led the US National Science Foundation (NSF) to launch a five-year initiative, called Cyber-Enabled Discovery and Innovation, to develop “a new generation of computationally based discovery concepts and tools to deal with complex, data-rich, and interacting systems.” The NSF Symposium on Cyber-Enabled Discovery and Innovation (www.rpi.edu/nsfcdi/), held in September 2007, helped highlight many of the challenges of building a cyberinfrastructure for multidisciplinary research (see the “Cyberinfrastructure Across Disciplines” sidebar).

The Future of Discovery

A CCI involves several processes and challenges in enabling cybercommunities to conduct scientific inquiries and understand and validate scientific phenomena. Figure 2 illustrates this process.

PURDUE IONOMICS INFORMATION MANAGEMENT SYSTEM

The rapid advances in high-throughput technology have enabled the generation of massive amounts of experimental data on biological systems. To use this data for knowledge generation and discovery, it's critical that researchers have access to every piece of related information in a digestible format. The Purdue Ionomics Information Management System (PiiMS; www.purdue.edu/dp/ionomics/) is an example of a Web-accessible, community-based cyberinfrastructure (CCI) that targets such discovery processes and related data and metadata acquisition activities.¹ Purdue University researchers developed PiiMS to promote an understanding of how plants take up, transport, and store their nutrient and toxic elements, collectively known as the ionome, which will benefit human health and the natural environment. Currently, labs around the country submit seeds or lines to be planted and analyzed at Purdue; they can then track the different stages of their orders through PiiMS. The fully

functional version of PiiMS will support a collaborative network of labs in which each lab will be able to manage the different experimental and analysis stages through a shared cyberinfrastructure.

PiiMS satisfies the data-handling and modeling requirements of a CCI using an elaborate data and metadata upload and gathering mechanism. Visualization and analysis requirements are satisfied by a search portal that lets users view summaries and plots as well as compose queries. PiiMS's modularity, agility, and commoditization requirements are also key design features: each element of data analysis is modular because users can retrieve it at any stage of analysis, even by applying a different line of analytics. Using common components such as LDAP, persistence layers, computation and statistics packages, and graphing and reporting tools renders PiiMS even more agile.

Reference

1. I. Baxter et al., "Purdue Ionomics Information Management System: An Integrated Functional Genomics Platform," *Plant Physiology*, vol. 143, Feb. 2007, pp. 600–611.

A CCI's users will be researchers conducting scientific interdisciplinary and intradisciplinary inquiries (see the top rung of the pyramid in Figure 2). The response to their inquiries might be unknown at the time of articulation, even though it might be based on previous observations of data or results. Similarly, an inquiry could lead to the discovery of previously unknown and, possibly tangentially related, phenomena.

A CCI will offer computational tools for data visualization, model creation, simulation, and trend extraction. The CCI will also enable distributed tools and services—for example, one participating CCI might provide statistical analysis algorithms, that researchers at other CCIs can use to discover trends in data generated by their simulations. We broadly categorize these activities as *discovery* and represent them as the middle rung of the pyramid in Figure 2.

Today, most scientific experiments rely on the creation, usage, and interpretation of huge amounts of data under different formats and structures. Data generated by experiments might not have clear structural links within itself, so the challenge here is twofold: relating structurally heterogeneous data for effective interpretation and dealing with the huge scale of data generated by scientific experiments. These two challenges are further exacerbated because the data is generated by a community of interdisciplinary and intradis-

ciplinary researchers rather than a single scientist.

We can classify data as either structured (conforming to a schema) or unstructured (nonconforming). Although the inherent structure in traditional data representation, such as relational databases and XML documents, lets us use contemporary data-mining and analysis techniques, unstructured data sources aren't as easy to manipulate. Examples of unstructured data representation include video, chat logs, images, and so on. In addition, the scale of data generated poses an additional challenge, both in terms of storing it and applying analysis algorithms to it. A CCI could address these problems by using techniques such as virtualization at the service and resource levels.

The ultimate goal for developing, deploying, and using a CCI is that it should be transparent: researchers shouldn't be concerned with the infrastructure's underlying nuts and bolts. To achieve this goal, the discovery process won't have to account for the underlying hardware and software design issues; various virtualization services at different levels of the infrastructure will address these.

Requirements

To realize this vision, we need to understand and define CCI requirements and propose an architectural framework for handling them.^{7,8} Although we state these requirements in general terms, they translate into concrete functional architectural

NANOHUB

nanoHUB (<http://nanoHUB.org>), a Web-based resource for research, education, and collaboration in nanotechnology, is a collaborative cyberinfrastructure built by the US National Science Foundation's Network for Computational Nanotechnology (NCN). NCN is a network of universities with a vision to pioneer the development of nanotechnology from science to manufacturing through innovative theory, exploratory simulation, and novel cyberinfrastructure. NCN students, staff, and faculty are developing the nanoHUB science gateway and using it in their own research and education. Collaborators and partners across the world have joined NCN in this effort and have created a community of researchers. nanoHUB connects computer scientists and applied mathematicians to problem-driven scientists and engineers to address large-scale problems and develop community codes for nanotechnology.

nanoHUB enables simple resource browsing and launching of interactive simulation tools from any Web browser, and users can also access and share all kinds of resources, including live simulations. nanoHUB provides computing resources, simple experiment setup and analysis interfaces,

and state-of-the-art research models in nanotechnology to its community of researchers. From March 2007 to February 2008, more than 6,200 users executed over 270,000 simulations, and the number of annualized users, who run simulations and explore nanoHUB content such as tutorials, seminars, and classes delivered as interactive lectures, podcasts, and PDF files, now exceeds 60,000.

nanoHUB's users don't need to download, install, or configure software components, with the exception of a few browser plug-ins. The nanoHUB back-end virtualization service¹ manages all jobs by locating suitable machines or clusters of machines and running in a virtual machine with suitable resources that match the user's requirements. Users only know that they're running a specific application but are masked from implementation details such as machine type and configuration. nanoHUB adheres to the requirements of service and resource virtualization by virtualizing the physical hardware and software.

Reference

1. M. Lundstrom and G. Klimeck, "The NCN: Science, Simulation, and Cyber Services," *Proc. IEEE Conf. Emerging Technologies: Nanoelectronics*, IEEE Press, 2006, pp. 496–500.

components. We divide the key requirements into three main categories: service, resource, and deployment; see Figure 3. We divide the overall requirements into service-layer, resource, and deployment requirements.

We can further divide the service-layer requirements into community-specific and cross-community requirements. The former relate to services developed and deployed for use in a specific research community—for example, a specialized genetic marker detection algorithm designed with a specific intent. A cross-community service might be a statistical analysis algorithm designed to reveal statistical qualities of data, irrespective of its application domain. The service layer's key requirements are as follows:

- **Data handling.** A research community should be able to share data and metadata, as well as the annotation history of its creation and processing, and the infrastructure should provide storage and query capabilities for data in any experiment-centric and application-specified format. Diverse disciplines might require domain-specific structure, format, and representation but should also be able to translate and share it with other disciplines. We must couple wider availability of data to communi-

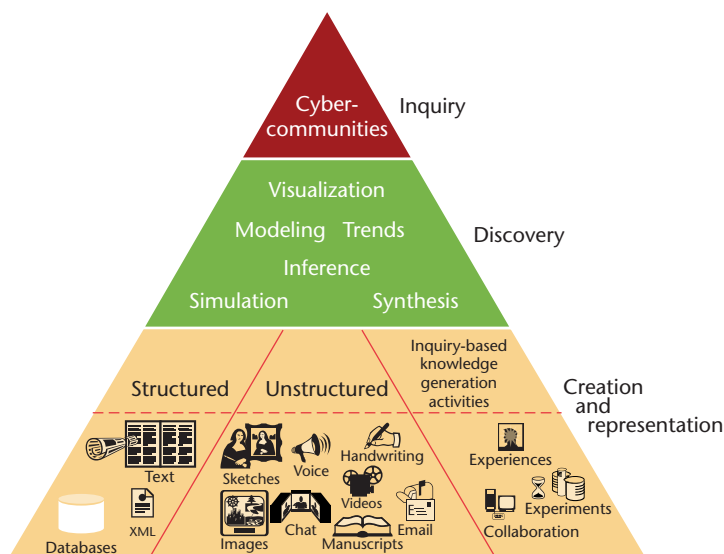


Figure 2. Community-based cyberinfrastructure for scientific inquiry. Researchers articulate a knowledge inquiry that triggers several discovery activities on diverse data.

ties with preservation and integrity. This raises issues of curation when researchers are generating, storing, and analyzing data from different disciplines.

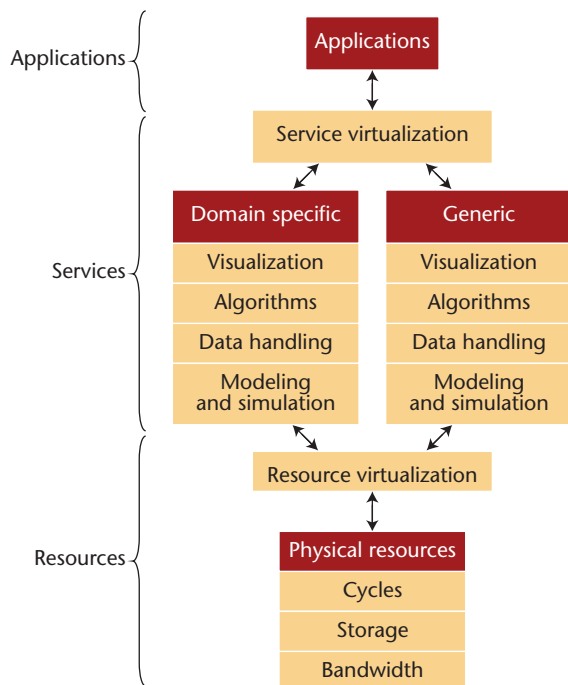


Figure 3. Requirements of a community-based cyberinfrastructure.

- *Modeling and simulation.* Researchers should be able to create, execute, and analyze mathematical and algorithmic models of real-world phenomena. The complexity of interdisciplinary and intradisciplinary modeling and simulation could increase many-fold, with multiple disciplines bringing their unique perspective of real-world phenomena into the model.
- *Algorithms.* The CCI should help researchers create and deploy algorithms that others can use to process the data in their experiments. The challenge is to devise specialized algorithms with general interfaces so that multiple disciplines can use them with ease.
- *Visualization.* Using the CCI, researchers should be able to visualize data in various formats and visual constructs and interact with that data, and with each other, in virtual environments. The key challenge is to be able to share visualizations between disciplines with different perspectives of the world.

The CCI service layer requires the seamless and transparent availability of several physical resources. Thus, we can define its resource requirements as follows:

- *Cycles.* The CCI challenge is to share cycles between researchers separated by geography or

institution. CPU resources also must be flexible and modular to support technology upgrades.

- *Storage.* The CCI will need data storage in application- and service-specified formats. Storage will be in multiple formats, such as database, text, audio, and video files.
- *Bandwidth.* Researchers will require networked connectivity between nodes of the same CCI and with other collaborating CCIs. For a community-based CCI, high-availability bandwidth is essential because most interactions between community members will occur over physical networks.

The previous requirements will be supported by providing virtualization at two levels:

- *Service virtualization.* The CCI will need to abstract and share services among applications (both local and remote). Specifically, multiple researchers' applications might use data handling, modeling and simulation, algorithms, and visualizations at varying levels of abstraction. Service virtualization might also be used for aggregation of services before being used in an experiment. We want to hide unnecessary details wherever possible and abstract services to the highest possible level.
- *Resource virtualization.* The CCI will also need to abstract and share physical resources among multiple services. Resource virtualization provides an abstraction between the physical resources (cycles, storage, and bandwidth) and the services that use them. Because setting up a CCI is a cost- and time-intensive activity, it isn't feasible to set up CCIs separately for each knowledge domain. Sharing virtualized physical resources lets the community of researchers use the same physical infrastructure while reducing hardware investments.

Finally, a CCI that caters to the needs of an ever-changing discovery landscape must exhibit certain characteristics:

- *Modularity.* The services provided and the resources being used by the CCI must be modular in the sense that new services or resources can be added seamlessly and old ones removed. Modularity in terms of resources also lets us perform simple technological upgrades, replacing aging components with new ones as long as they conform to certain coexistence standards.
- *Agility.* Each component in a CCI needs to adapt

CYBERINFRASTRUCTURE ACROSS DISCIPLINES

In addition to the US National Science Foundation's effort to highlight the need for a multidisciplinary cyberinfrastructure,^{1,2} other government agencies have expressed their cyberinfrastructure vision as well. The US National Archives and Records Administration has outlined its commitment to create a cyberinfrastructure to electronically archive and preserve all information related to the US government.³ Similarly, a US National Science and Technology Council report summarizes the security and information assurance challenge as "measures for protecting computer systems, networks, and information from disruption or unauthorized access, use, disclosure, modification, or destruction...with the purpose of providing Integrity, Confidentiality and Availability."⁴

The US Department of Defense's vision for Net-centric operations and warfare outlines its vision of a cyberinfrastructure as "a world in which information is virtual and on demand with global reach. Information is protected by identity-based capabilities that allow users to connect, be identified, and access needed information in a trusted manner."⁵

The UK's National E-Science Center has set out its requirements for a national e-infrastructure to help ensure that the UK maintains and enhances its global standing in science and innovation.⁶

Lastly, the GEONgrid (www.geongrid.org/) project's vision is to interlink and share multidisciplinary data sets and computational environments to understand the complex dynamics of Earth systems. GEON has developed a shared cyberinfrastructure for Earth sciences research as well as education at the K-12 and professional levels.

References

1. NSF Cyberinfrastructure Council, *Cyber Infrastructure Vision for 21st Century Discovery*, Mar. 2007; www.nsf.gov/pubs/2007/nsf0728/index.jsp.
2. D.E. Atkins et al., *Revolutionizing Science and Engineering through Cyberinfrastructure*, tech. report, US Nat'l Science Foundation, Jan. 2003; www.pnl.gov/scales/docs/cyberinfra_2003.pdf.
3. K. Thibodeau, *Preserving Electronic Records: Developments at the National Archives and Records Administration*, Electronic Records Archives Program, June 2004; www.archives.gov/era/pdf/thibodeau-040617.pdf.
4. *Federal Plan for Cyber Security and Information Assurance Research and Development*, Interagency Working Group on Cyber Security and Information Assurance, Apr. 2006; <http://handle.dtic.mil/100.2/ADA462532>.
5. *Surety, Reach, Speed: The Disa Strategy*, Defense Information Systems Agency, Mar. 2007; www.disa.mil.
6. OSI e-Infrastructure Working Group, *Developing the UK's E-Infrastructure for Science and Innovation*, UK Nat'l E-Science Center, 2004; www.nesc.ac.uk/documents/OSI/index.html.

to changing user expectations and advancing technology. Examples of new requirements might be larger modeling spaces or enhanced visualization methods.

- **Commoditization.** Commoditization of resources and services helps create a CCI that can be assembled without regard to the underlying technology's physical configuration and structure. A researcher can then use a commoditized data-handling service available at a remote CCI without regard to the underlying data formats, database engines, storage spaces, or physical memory. For commoditized CPU cycles, researchers requiring additional cycles would simply offload tasks to a remote CPU.
- **Security.** Security cuts across all layers of a CCI, so an advanced one must be able to authenticate and authorize users to access certain resources.

These key requirements are essential for designing the envisioned CCI and must be taken into account in defining the architectural components of CCI.

Architectural Components

In this section, we generically define each architectural component of a CCI while offering some specific examples and options for implementing each component. Our aim is to define an architecture that can be translated into an actual implementation with any technology or product. Use of these technologies and products depends on the domain of research, long-term goals, available resources, and so on.

Figure 4 depicts an architectural overview of a CCI's functional components, which we've adapted from an earlier work.⁵ We've divided the overall architecture into discovery and cyberinfrastructure layers.

Enabling Technologies

The lowest CCI layer consists of the enabling technologies that facilitate key functionalities such as computation, storage, networking operating environment, and so on. These functionalities correspond to cycle, storage, and bandwidth requirements. Modularity, agility, and commoditization of physical resources are also impor-

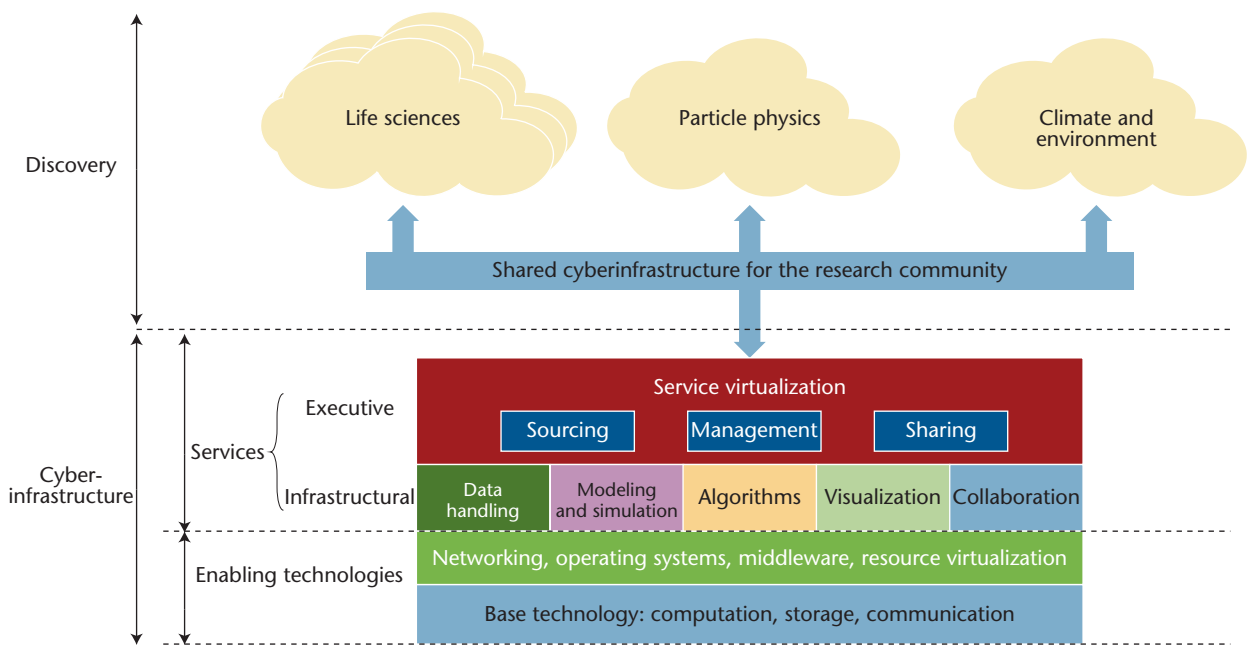


Figure 4. Multidisciplinary discovery using shared services of CCI. The discovery layer holds the cybercommunities composed of researchers from various disciplines. The cyberinfrastructure layer is home to the various architectural components that make the discovery layer possible.

tant design considerations that we must account for when deploying these technologies. Security at this level is likewise a cardinal design consideration and includes the use of secure network protocols, authentication mechanisms, and authorization models.

Enabling technologies might range from various CPU types, memory models, operating systems (such as Linux or Windows), storage systems, networking technologies (such as TCP/IP protocols), and peer-to-peer networking paradigms.

Services

We divided the service layer into two distinctly featured subservices: infrastructural and executive services.

Data-handling, modeling, simulation, algorithm, visualization, and collaboration services constitute the infrastructural services. The capabilities and constraints of these services depend on the research discipline and organizational goals (such as budget). Although a certain design attribute might be important for one specific discipline, it has little significance for another—for example, although weather modeling might require a 3D data view, the same visualization design might be overkill in a genetic search environment. Subservices within each service must be modular and commoditized so that they can be used in-

dividually in a specific application. Furthermore, the infrastructure might abstract services and use them in combination with other assembled services to represent a specific functionality.

Examples of data-handling services might include a simple filesystem, a database management system such as MySQL, and an SQL server. Researchers have developed several modeling and simulation software applications for specific domains as well as for general-purpose usage across disciplines. Examples of discipline-specific simulation software include the Earth Simulator (www.es.jamstec.go.jp/index.en.html) and the Network Simulator NS-3 (www.nsnam.org); Ascend (<http://ascend.cheme.cmu.edu/>) is a general-purpose simulation software application.

The algorithm service can range from simple, statistical analysis algorithms to more sophisticated genetic discovery algorithms designed specially for a domain. CAVE (<http://cave.ncsa.uiuc.edu/>), which researchers have used in weather visualization and other applications, is an example visualization service. Collaboration software can range from simple document exchanges using email, FTP, or wikis to more sophisticated computer-supported collaborative work applications.

Executive services should allow virtualization of the underlying infrastructural services, to address any deployment requirements. Such


services include three important and necessary constituent services: sourcing, management, and sharing.

The sourcing service lets a CCI discover and update existing infrastructural services held by that CCI and eventually other CCIs—specifically, it maintains knowledge, including constraints and semantics, of all the services currently available to the scientific community using the CCI. The sourcing service also lets the CCI advertise its own services, which it can share.

The management service lets managerial functions occur at the service level by controlling resource provisioning to individual applications, balancing loads among multiple applications, and releasing resources once experiments are complete. The management service also creates a secure environment in which local and remote users can share infrastructural services. Another function of the management service is to provide CCI administrators with loading and usage reports for optimal reconfiguration and load balancing.

Finally, the sharing service lets the community of CCI users share data, models, visualizations, and simulations among themselves; it also maintains lists of all shareable resources and lets member researchers select the required resource.

The functional architecture we describe here can be implemented with off-the-shelf products and customized implementations, but the goal should be to adhere to the guiding functional architecture and satisfy a requirements set. This general framework doesn't specify how to distribute the different resources and services and through which channels they're consumed—in particular, using a portal-centric design or a distributed scheme are all possible options, depending on a specific CCI's needs and specific requirements and resources.

Major efforts are under way worldwide to build community cyberinfrastructures including work in the Cyber Center at Purdue University and elsewhere (see the related sidebars for more details). However, several fundamental issues need to be solved, in particular those related to the requirements that we've elucidated here, before the envisioned CCI becomes reality. 

References

1. A. Elmagarmid, *Cyber Communities: Innovation in Science and Engineering*, The Cyber Center, Purdue Univ., 2007; www.purdue.edu/cybercenter/cci.pdf.
2. P.N. Edwards et al., *Understanding Infrastructure: Dynamics, Tensions, and Design*, tech. report, Workshop on History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures, Jan. 2007; <http://hdl.handle.net/2027.42/49353>.
3. S. Dong, G.E. Karniadakis, and N.T. Karonis, "Cross-Site Computations on the TeraGrid," *Computing in Science & Eng.*, vol. 7, no. 5, 2005, pp. 14–23.
4. J.J. Thomas and K.A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, tech. report, US Nat'l Visualization and Analytics Center, 2005; <http://nvac.pnl.gov/agenda.stm>.
5. J.L. Snoep et al., eds., "From Isolation to Integration: A Systems Biology Approach for Building the Silicon Cell," *Systems Biology: Definitions and Perspectives*, Springer-Verlag, 2005, p. 7.
6. 2020 Science Group, *Towards 2020 Science*, 2005; <http://research.microsoft.com/towards2020science>.
7. *Cyber Infrastructure Vision for 21st Century Discovery*, US Nat'l Science Foundation Cyberinfrastructure Council, Mar. 2007; www.nsf.gov/pubs/2007/nsf0728/index.jsp.
8. D.E. Atkins et al., *Revolutionizing Science and Engineering Through Cyberinfrastructure*, Blue-Ribbon Advisory, US Nat'l Science Foundation, Jan. 2003.

Ahmed K. Elmagarmid is the director of the Cyber Center at Purdue University. His research interests include a large spectrum of foundational and application-oriented database research, including video databases, data quality and confidentiality, data integration, Web services, bioinformatics, and multidatabase systems. He received the Presidential Young Investigator Award from the US National Science Foundation and the distinguished alumni awards from Ohio State University and the University of Dayton. Elmagarmid has a PhD in computer science from Ohio State University. He is a member of the ACM, the AAAS, and a senior member of the IEEE. Contact him at ake@cs.purdue.edu.

Arjmand Samuel is working on his PhD in the School of Electrical and Computer Engineering at Purdue University. His research interests include the use of computing in science and engineering research and security of information in collaborative environments. Samuel has an MS in electrical engineering from the Beijing University of Aeronautics and Astronautics, China. He is a member of the IEEE. Contact him at amsamuel@purdue.edu.

Mourad Ouzzani is a research assistant professor with the Cyber Center at Purdue University. His research interests include database and Web services with a focus on life sciences and database integration, schema matching, database system support for biological data, and access control for Web services. Ouzzani has a PhD in computer science from Virginia Tech. He is a member of the ACM and the IEEE. Contact him at mourad@cs.purdue.edu.