Community Cyber Infrastructure Enabled Discovery in Science and Engineering

Ahmed Elmagarmid[1,2] Arjmand Samuel[3] Mourad Ouzzani[2]
[1] Department of Computer Science [2] Cyber Center, Discovery Park,
[3]Department of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47906, USA

Abstract

Cyber infrastructure for discovery in science and engineering has not only ***enhanced*** *the process of discovery but has also* ***enabled*** *new venues for discovery including systems biology, ecosystem modeling, and individualized medicine. The transition from enhancing to enabling discovery in science and engineering has led to what we term the* ***Silicon Shift****. In this paper, we conceptualize the vision of a* ***community cyber infrastructure (CCI),*** *shared by the research community, that would enable a new era of multi-disciplinary discovery. We introduce the concept of cyber communities and outline the process of discovery utilizing community cyber infrastructure. The emphasis of this paper is to lay out the essential attributes of CCI including requirements, functional architecture, and examples of implementation options. We thus formalize the design of a community cyber infrastructure by elucidating the specific requirements in terms of services, resources and deployment. We then lay out the foundations for an agile and adaptable CCI's functional architecture based on the stated requirements. In this functional architecture, knowledge, models, simulation results and visualizations are shared along with computing cycles, storage, and bandwidth. The emphasis is to crystallize a design for a CCI in as general terms as possible, so that this vision can adapt and remain relevant with innovation in technology.*

1. Introduction

Traditionally, researchers have conducted scientific experiments in a laboratory environment using application specific instruments and subjects, giving rise to measurements which are manually recorded and analyzed by the scientists using mathematical and statistical techniques. Observations gathered from experiments have been instrumental in validating theories set forth by researchers, and in discovering new and at times, unexpected phenomena. This is, in short, how science has been conducted for the past few hundred years. Consequently, the two pillars of science have been ***theory*** and ***experimentation***.

However, most experiments today generate observations which cannot be recorded, let alone analyzed, manually. Examples of such experiments are the use of mass spectroscopy for analysis of cells, DNA sequencing, particle physics experiments, and astronomy. This exponential increase in the amount of information generated by scientific experiments prompted researchers to look elsewhere for efficient handling and analysis of data. With the advent of computing, researchers can now afford fast and efficient storage of experimental observations and analysis at unimaginable speeds. However, computing in the context of experimentation in science and engineering was initially viewed as yet another tool to be used in the laboratory; akin to test tubes and multi-meters. The underlying thought being that the new tool; i.e. computing, ***enhances*** the process of experimentation and observation by aiding information storage and analysis. It was now possible to store more information, analyze more records, run statistical analysis on a bigger set of records etc. Computing in many ways started to ***enhance*** the process of discovery in science and engineering.

Today experimental observations are not just stored digitally, but most are "born" digital. An interesting example is that of the three-dimensional computer model of arterial blood flow that was built to aid treating different circularly diseases [1]. This model uses multiple supercomputers in parallel through the TeraGrid (www.teragrid.org) to determine exact blood flow. Each run of this device can generate hundreds of megabytes of data which are directly saved in digital form with no analog artifact of the experiment. This profound shift allows more meaningful analysis of data, visualization based on immersive environments and modeling/simulation 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 analysis of data. Interaction between the researcher and data is not necessarily through a laboratory notebook, but through an immersive environment where the researcher interacts with data utilizing visual representation, interaction technologies and analytical reasoning [8] powered by such technologies as haptic interfaces. Consequently, computing is not only enhancing discovery, but is in fact ***enabling*** it. A case in point is that of systems biology, which is defined as the study of interaction between the components of a biological system [2]. The systems biology approach is characterized by a cycle of theory, computational modeling and experiments to describe cells and cell processes. As it is evident, systems biology exists because of computing and so, computing in many ways does not only enhance discovery but enables it. In light of the above facts, the research community has termed computing as the ***third pillar of science,*** together with theory and experimentation [3].

The evolutionary change in the way experiments are conducted is marked by a profound state that we call the ***Silicon Shift*** (Figure 1). Hence, prior to the silicon shift, most of the experiments are conducted in the lab with actual subjects and scientific equipment. Consequently, most of the discoveries are made in the laboratory. The advent of computing changed this scientific methodology gradually by aiding the researcher in gathering, aggregating, analyzing and reporting data and findings. With the increase in computing power, data storage and network bandwidth, more and more verification and aggregation of results is deputed to computing infrastructure. This results at a defining point in experimentation where in-silico (modeling and simulation) becomes the norm, and experimentation (wet lab) is solely used for validation of results.

Today, increasing number of discoveries in all fields of science and engineering are taking place at the intersection of disciplines. A case in point is the use of nanotechnology for bio-medical research; where the fields of chip fabrication, electronics, physics and biology come together to discover new phenomena which was not possible in individual disciplines. Further, most scientific discoveries are not a result of one person conducting experiments behind closed doors of a laboratory, but involve communities of researchers, collaborating to find solutions to various aspects of the same problem. With increasing reliance on computing technology to ***enable*** discovery and the afforded high connectivity, there is a compelling necessity to develop a community-based computing infrastructure, or Community ***Cyber Infrastructure*** (CCI) which allows inter- and intra-disciplinary collaboration for discovery.

NSF's "Cyber infrastructure Vision for 21st Century Discovery" report [4] defines ***Cyber Infrastructure*** (CI) as the infrastructure which "*integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools*." ***Cyber-Communities,*** as defined by most researchers, is a group of users of a common, public, and shared cyber infrastructure used for discovery; these users include researchers, scientists, educators and students distributed across institutions, geography and disciplines.

The key role of cyber infrastructure in science and engineering has led to the launch of a new 5-year initiative by NSF to start in 2008, namely Cyber-enabled Discovery and Innovation (CDI), to develop "*a new generation of computationally based discovery concepts and tools to deal with complex, data-rich, and interacting systems*". A timely impetus in this direction has been a recent NSF sponsored workshop[1] highlighting many of the challenges of building a CI for multidisciplinary research.

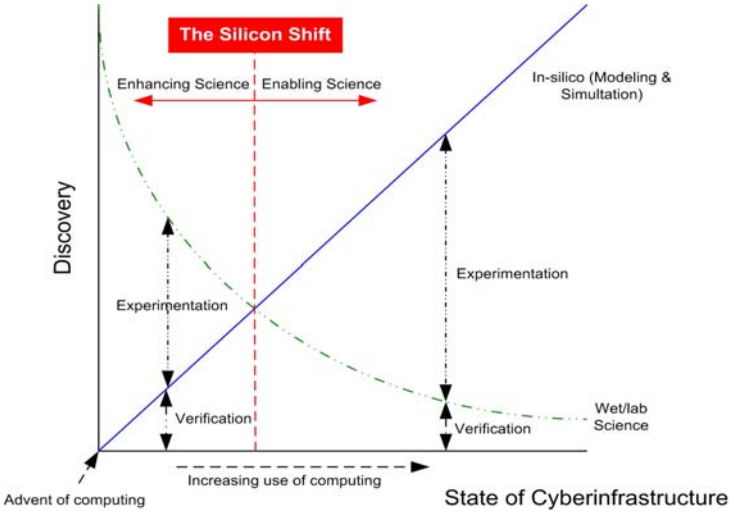

Figure 1: The defining moment in science: ***The Silicon Shift (SS).*** Prior to the SS, computing ***enhances*** discovery, whereas after SS it ***enables*** discovery.

[1] NSF Symposium on Cyber-Enabled Discovery and Innovation, September 2007, RPI, Troy, NY (http://mediasite.itops.rpi.edu/Mediasite4/Catalog/Front.aspx?cid=4c580e0d-628f-43ee-b52c-a6863a291048)

In this article, we present the prevalent **vision** for a ***Community Cyber Infrastructure*** (CCI) [9] to be used by various communities in science and engineering. CCI would enable the discovery of new phenomena in science and engineering by allowing scientists and engineers to seamlessly conduct research and collaborate across multiple disciplines. The envisioned CCI is designed to be shareable and transparent in nature; scientists are not concerned with the underlying infrastructure design and operational issues, but rather concentrate on the science inquiry which will mainly consist of devising in-silico experiments. The ubiquity and transparence of CCI will be more akin to the utility infrastructure [5] which pervades our daily lives. To translate this vision into a tangible infrastructure, we outline the process of discovery by cyber-communities using CCI (Section 2). We then review the fundamental requirements of such an infrastructure (Section 3). Based on these requirements, we present a core functional architecture that will serve as the basic framework for implementing the envisioned CCI (Section 4). In this section, we also mention examples of various options available to implement such a CCI. The aim is to present one of the many possible approaches to implement a given functional component. Additionally, we offer an overview of some recent efforts and key reports on cyber infrastructure (Sidebars) by government agencies as well as academia and the industry.

2. The Future of Discovery using CCI

CCI involves several processes and challenges in enabling cyber-communities to conduct scientific inquiries, and understand and validate scientific phenomena. We represent the scientific inquiry process through the pyramid depicted in Figure 2(a).

The users of the CCI are a community of inter- and intra- disciplinary researchers. These researchers collaborate with each other for conducting scientific inquiry. The inquiry is usually discipline specific and consists of a set of processes to be executed on data in various forms. The response to the inquiry, and the process involved, may not be known at the time of articulation; however the inquiry may be based on previous observations of data or results. Similarly, the inquiry may also lead to discovery of phenomena previously unknown and not directly related to it. This is depicted as the highest tier of the pyramid in Figure 2(a).

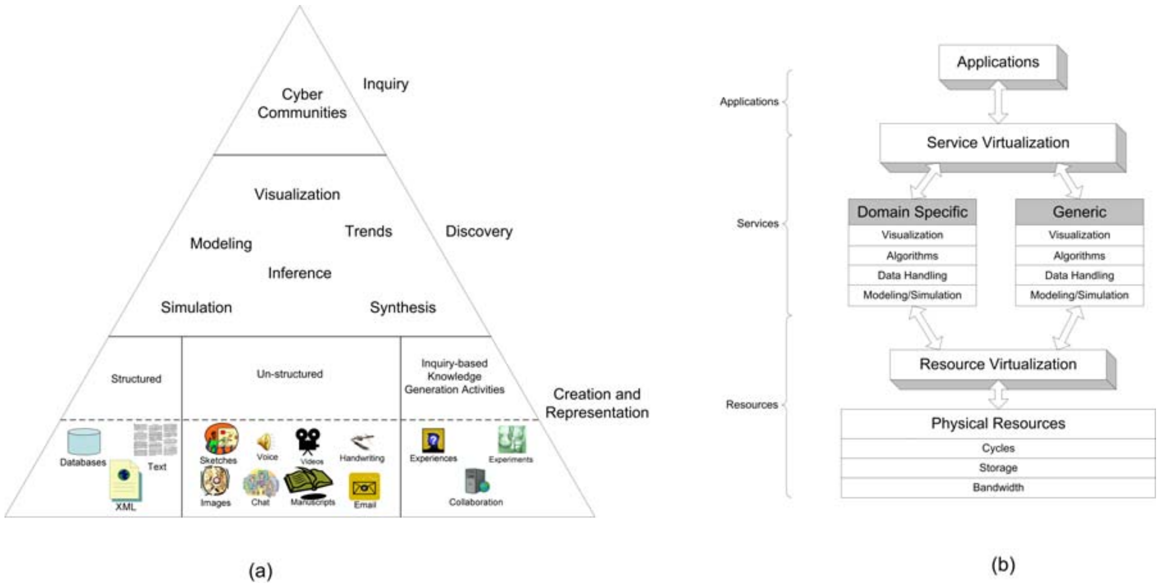

Figure 2: (a) Community-based CI for scientific inquiry.
(b) Requirements of a community-based CI

A number of computing tools are provided by CCI to respond to questions posed by the community members. These tools include visualization of data, creation of models, simulation, and extraction of trends using data mining techniques; which may exist locally or remotely in a distributed manner. Increasingly, the deployment of these tools as services allows distributed development, deployment, sharing, and usage. CCI enables distributed tools and services, provided by the community, to be employed as distributed components of an experiment which can be composed and executed both locally and remotely. An example of such a service is statistical analysis algorithms provided by one of the participating CCIs and used by a researcher to discover trends in data generated by his or her simulation. The sharing of CCI tools and services is not only in the form of community members using algorithms and analysis techniques developed by each other, but also involves researchers sharing virtual environments and interacting through shared visualizations to work live on the same problem; more like sharing virtual worlds. We broadly categorize these activities as ***Discovery*** and represent them as the middle rung of the pyramid in Figure 2(a).

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 may or may not have clear structural links within itself. The challenge in this regard is two fold; (1) relating structurally heterogeneous data for effective interpretation (2) and dealing with the huge scale of data generated by scientific experiments. The two challenges are further exacerbated by the data being generated by a community of inter- and intra-disciplinary researchers rather than by a single scientist.

The classification of data used and generated by experiments can be divided into two main categories; namely, structured (conforming to a schema) and, un-structured (not conforming to a schema). While the inherent structure in traditional data representation, such as relational databases, and XML documents, allows contemporary data mining and analysis techniques to be used, un-structured data sources do not easily allow this. Examples of un-structured data representation are video, chat logs, images, and so on. In addition, the scale of data generated by most scientific experiments poses a huge challenge, both in terms of storing it and applying fast and efficient data analysis algorithms that would produce meaningful results for the researcher. CCI addresses these challenges by techniques such as virtualization both at the services level as well as at the resource level.

Experiments composed by the researchers utilizing services and data lead to other experiments and collaborations among other researchers. While each experiment and collaboration may result in discovery of new scientific phenomena, they also result in additional data which again raise the issues of structural heterogeneity and scale. We term these activities as ***Creation and Representation*** and depict them as the last rung in the pyramid in Figure 2(a).

The ultimate goal for developing, deploying and using a CCI is that it should be transparent in nature. The researchers are not concerned with the underlying nuts and bolts of the infrastructure, but rather concentrate in using CCI for conducting scientific experiments. To achieve this goal, the structure of the discovery process (depicted in Figure 2(a)) does not have to take into account the underlying hardware and software design issues; these will be addressed by the various virtualization services at different levels of the infrastructure as we will discus in more details in Section 3.

The methodology for discovery, outlined in this section, briefly summarizes the shared vision of using a CCI for enabling discovery by cyber-communities of scientists and engineers. To realize this vision, we need to understand and define the requirements of such a CCI and subsequently propose an architectural framework which conforms to those requirements.

3. Requirements of a Community Cyber Infrastructure

In this section, we outline the key requirements for building the envisioned CCI based on the cyber infrastructure requirements defined in [4, 6], with special emphasis on a community based CI. While these requirements have been stated in the most general terms, they translate to concrete functional architectural components as discussed in Section 4. We divide these key requirements into three main categories, namely; service, resource and deployment requirements. Further, we define the overall characteristics of a CCI. The three categories of requirements, together with the characteristics, provide a holistic view of a CCI.

As discussed in Section 2, a CCI is created to aid a community of researchers in discovery, for which they use or create software artifacts for scientific inquiry utilizing an underlying infrastructure of services and physical resources. The interrelationship between the various requirements is depicted in Figure 2(b). The overall requirements are divided into service layer requirements, resource requirements and deployment requirements. The overall characteristics stipulate the infrastructural requirements of the CCI.

The service layer requirements can be divided into community specific and cross-community (applicable across disciplines) requirements. The community specific requirements relate to services which are developed and deployed to be used by a specific research community. An example of such a service is a specialized genetic marker detection algorithm which is designed with a specific intent. A cross-community service may be a statistical analysis algorithm which is designed to reveal statistical qualities of data, irrespective of the application domain. The following are the key requirements of the service layer:

- ***Data handling*** is the ability to provide storage and query capabilities for data in any experiment-centric and application specified format. Specific data handling requirements of a community of researchers is the ability to share not just data, but also meta-data and annotation history of its creation and processing. Diverse disciplines may require domain specific structure, format and representation; but may also require the ability to translate and share it with other disciplines. Wider availability of data to communities has to be coupled with preservation and integrity. Issues of curation are also central when inter- and intra-disciplinary data is being generated, stored and analyzed.
- ***Modeling and simulation*** is the ability afforded by CCI to create, execute and analyze mathematical and algorithmic models of real world phenomena. The complexity of inter- and intra-disciplinary modeling and simulation may increase many-fold, with multiple disciplines bringing their unique perspective of the real world phenomena into the model. `
- ***Algorithms,*** both, local and distributed are created and deployed on CCI for use by researchers as processors of data in their experiments. CCI’s challenge in creating algorithms is to devise specialized algorithms with general interfaces, to be used for inter- and intra-disciplinary discovery so that multiple disciplines can utilize them with ease.
- ***Visualization*** is the ability of CCI to allow researchers to visualize data in a variety of formats and visual constructs. For inter- and intra-disciplinary discovery, the key challenge is to be able to share visualizations between disciplines with different perspectives of the world. The collaborative nature of visualization allows the community of researchers to interact with data, and with each other, in virtual environments.

The CCI service layer requires the seamless and transparent availability of several physical resources on which it is built. Thus, the resource requirements can be defined as follows:

- ***Cycles*** represent the processing power in the form of CPU resources available to CCI users. The community-based CI challenge is to be able to share cycles between researchers separated by geography and/or institution. CPU resources also need to be flexible and modular to support upgrades in technology.

- ***Storage*** of data in application and service specified formats is required by any CCI. Storage may be in the form of data bases, text, audio, video files etc.
- ***Bandwidth*** of networked resources provides connectivity between nodes of the same CCI or with other collaborating CCIs. For a community based CCI, high availability bandwidth are essential as most interactions between community members take place over physical networks.

The requirements that will guide the deployment of CCI based on the concept of virtualization which we consider at two levels are:

- ***Service virtualization*** is the ability of the CCI to abstract and share services among applications (both local and remote). At the heart of this requirement is the realization that CCI is designed for inter- and intra-disciplinary sharing and may require abstraction of services for various disciplines to be used effectively. Specifically, the data handling, modeling and simulation, algorithms and visualizations may be used by multiple researchers' applications at varying levels of abstraction. Service virtualization may also be used for aggregation of services before being used in an experiment. The key idea here is to hide unnecessary details wherever possible and abstract services to the highest possible level.
- ***Resource virtualization*** is the ability of the CCI 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 which use them. Note that setting up a CCI is a cost and time intensive activity and it is not feasible to setup CCIs separately for each domain of knowledge. Sharing of virtualized physical resources allows the community of researchers to use the same physical infrastructure while saving on precious investments on hardware.

The ability of the CCI to cater for the needs of an ever-changing discovery landscape requires certain characteristics including in particular:

- ***Modular.*** The services provided, and the resources being used by the CCI need to be modular in the sense that new services or resources may be added seamlessly and old ones removed. Modularity in terms of resources also allows simpler upgrades to technology as aging components can be replaced with new ones as long as they conform to certain co-existence standards or upward compatibility such as instruction sets in case of processing units. The modular design of services also allows enhancements in their design and capabilities over time.
- ***Agile.*** Each component within CCI needs to be agile so that it can adapt to changing user expectations and advancing technology. With changes in operating environments such as research focus and nature of experiments, an agile CI needs to cater for this change. Examples of new requirements may be larger modeling spaces or enhanced visualization methods.
- ***Commoditized.*** Commoditization of resources and services helps to create a CCI which can be assembled without regard to physical configuration and structure of the underlying technology. A commoditized data handling service available at a remote CCI can be utilized by a researcher without regard to the underlying data

formats, database engines, storage spaces, or physical memory. In case of commoditized CPU cycles, a researcher requiring additional cycles for an experiment he/she is conducting simply offloads tasks to a remote CPU. Commoditization of services and resources is provided by the respective virtualization requirements.

- ***Secure.*** Security cuts across all layers of the CCI. The ability to establish authentication and authorization of users to utilize certain resources is an important characteristic of an advanced CCI.

4. Architectural Components of a Community Cyber Infrastructure

In this section, we present the functional architecture of CCI that will serve as a general framework for the actual implementation. This architecture is based on the requirements set forth in Section 3. We generically define each architectural component while offering some specific examples and options for implementing each of these components. The aim in this regard is to define an architecture which can be translated into an actual implementation utilizing any technology or product. Use of these technologies and products depends on the domain of research, long term goals, available resources, etc.

Figure 3 depicts an architectural overview of the functional components of a CCI that has been adapted from the one proposed in [2]. The overall architecture has been divided into two major layers; namely, the *discovery layer* and the *cyber infrastructure layer*. Within the discovery layer resides the cyber communities comprising of researchers from various disciplines. These researchers would utilize the CCI to compose experiments and discovery processes. The cyber infrastructure layer is home to the various architectural components which make the discovery layer possible. We now shed some light on each of these components.

Enabling Technologies: The lowest layer of CCI consists of the enabling technologies which facilitate key functionalities such as computation, storage, networking operating environment, etc. These functionalities correspond to the Cycles, Storage and Bandwidth requirements set forth in Section 3. In addition, modularity, agility and commoditization of physical resources are important design considerations that need to be taken into account while deploying these technologies. Security at this level is also a cardinal design consideration and includes use of secure network protocols, authentication mechanisms and authorizations models.

Enabling technologies may range from a variety of CPU types, memory models, operating systems (Linux, Windows, etc), storage systems, to networking technologies such as TCP/IP protocols and peer-to peer networking paradigms.

Services: The next major component of the CCI utilizing the enabling technologies is the services layer. This layer is further divided into two distinctly featured sub-services namely; *Infrastructural* services and *Executive* services.

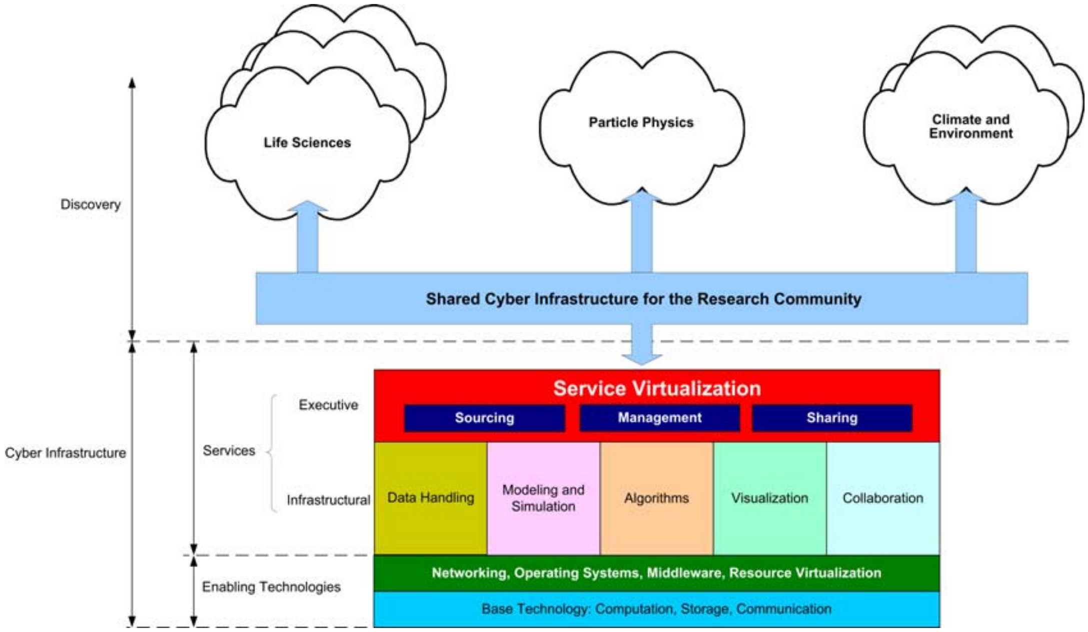

Figure 3: Multi-disciplinary discovery using shared services of CCI.

Data Handling Services (DHS), Modeling and Simulation Services (MSS), Algorithm Services (AS), Visualization Services (VS) and Collaboration Services (CS) constitute the infrastructural services. Capabilities and constraints of the infrastructural services depend on the research discipline, organizational goals such as budget, etc. While a certain design attribute may be important for one specific discipline, it may not be of much significance to another. Design of visualization techniques is one example of such a design choice. While weather modeling may require a three dimensional view of data, the same visualization design may be an overkill in a genetic search environment. Sub-services within each service are modular and commoditized so that they can be used individually within a specific application. Further, services may be abstracted and a service may be a combination of multiple services assembled to represent a specific functionality.

Examples of DHS may include a simple file system, a database management system such as MySQL and SQL Server, etc. Several modeling and simulation software have been developed for both specific application domain as well for general purpose usage across disciplines. An example of discipline specific simulation software is Earth Simulator[2] and NS2[3], while that of a general purpose simulation software is ASCEND[4]. The algorithm service can range from simple and generic statistical analysis algorithms to more sophisticated genetic discovery algorithms designed specially for a domain. An interesting example of the visualization service is the CAVE[5] which has been utilized in weather visualization and other applications. Collaboration software can range from

[2] http://www.es.jamstec.go.jp/index.en.html
[3] http://www.isi.edu/nsnam/ns/
[4] http://ascend.cheme.cmu.edu/
[5] http://cave.ncsa.uiuc.edu/

simple exchange of documents employing emails, FTP, wiki and so on to more sophisticated Computer Supported Collaborative Work (CSCW) applications.

The executive services are designed to allow virtualization of the underlying infrastructural services, consequently addressing the deployment requirements set forth in Section 3. Executive services include three important and necessary constituent services, namely, *Sourcing*, *Management* and *Sharing*.

Sourcing service allows the CCI to discover and update existing infrastructural services held by the CCI and eventually other CCIs. It maintains knowledge, including constraints and semantics, of all services currently available to the scientific community using the CCI. The sourcing service also allows the CCI to advertise its own services which it shares with other CCIs.

The management service allows managerial functions to take place at the service level. It controls provisioning of resources to individual applications, balances loads among multiple applications and releases resources once experiments are completed. The management service also creates a secure environment for infrastructural services to be shared by local and remote users. Another function of the management service is to provide the CCI administrators with loading and usage reports for optimal reconfiguration and load balancing.

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

The design of sourcing, management and sharing services depends heavily on the underlying technology as well as the long term goals and usage of the CCI. However, a modular design, agility in operation and deployment, ability to commoditize as well as a secure architecture remain the guiding principles for their design. The use of a service oriented architecture (SOA) is one promising option to implement these executive services.

The CCI, conforming to the set of requirements spelled out in Section 3, allows multiple researchers to simultaneously access CCI resources through virtualized services at all times. Further, it also allows the researcher community to conduct research by sharing CCI artifacts. As depicted in Figure 3, multiple researchers conducting climate and environmental research can co-exist with researchers conducting particle physics analysis and other researchers using the CCI for life sciences research. In this example, the three disciplines use the same CCI and may share discipline specific artifacts such as models and algorithms among themselves and conduct multi-disciplinary research.

It is important to note that the functional architecture described in this section can be implemented by utilizing a number of off-the-shelf products (as mentioned above) as well as customized implementations. However, the aim is to adhere to the guiding functional architecture presented in this section and satisfying the requirements set forth in Section 3. It is also worth noting that this general framework does not specify how the different resources and services are distributed and through which channels they are consumed. In particular, using a portal centric design or a distributed scheme are all possible options depending on the needs and specific requirements and resources of a particular CCI.

5. Conclusion

In this article, we formalized the shared vision for design considerations and use of a community cyber infrastructure. This vision is based on the need for computationally enabled multi disciplinary discovery by community of researchers to embark in new science inquiry and discover new phenomena. The aim is to provide a high level relationship between the process of discovery in science and engineering and the utility of a generic cyber infrastructure to enable this activity. The design methodology is to investigate the process of discovery using a CCI and formulating the set of requirements for services, resources and deployment based on this process. We also present design considerations for such a CCI. The functional architecture we described allows multiple disciplines and researchers to collaborate and conduct experiments. The proposed high level functional architecture possesses the potential to adapt and remain relevant with innovation in technology.

References

[1] S. Dong, G.E. Karniadakis, N.T. Karonis, “Cross-site computations on the TeraGrid” Computing in Science & Engineering, 7(5), 14-23, 2005.

[2] Snoep J.L. and Westerhoff H.V.; Alberghina L. and Westerhoff H.V. (Eds.) (2005.). "From isolation to integration, a systems biology approach for building the Silicon Cell". Systems Biology: Definitions and Perspectives: p7, Springer-Verlag.

[3] “Towards 2020 Science”, the 2020 Science Group, http://research.microsoft.com/towards2020science, accessed December 2007

[4] NSF Cyberinfrastructure Council, “Cyber infrastructure Vision for 21st Century Discovery”, March 2007, http://www.nsf.gov/pubs/2007/nsf0728/index.jsp, accessed December 2007

[5] “Understanding Infrastructure: Dynamics, Tensions, and Design”, Report of a Workshop on “History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures” Version 1.0 — 13 December 2006, http://connect.educause.edu/Library/Abstract/UnderstandingInfrastructu/37182?time=1197911478, accessed December 2007

[6] D. E. Atkins, K. K. Droegemeier, S. I. Feldman, H. Garcia-Molina, M. L. Klein D. G. Messerschmitt, P. Messina, J. P. Ostriker, M. H. Wright, “Revolutionizing Science and Engineering Through Cyberinfrastructure”, Report of the National Science Foundation, Blue-Ribbon Advisory, January 2003

[8] James J. Thomas, Kristin A. Cook, “Illuminating the Path: The Research and Development Agenda for Visual Analytics”, Report of National Visualization and Analytics Center, 2005, http://nvac.pnl.gov/agenda.stm, accessed December 2007

[9] Ahmed Elmagarmid, “Cyber Communities: Innovation in Science and Engineering”, The Cyber Center, Purdue University, http://www.purdue.edu/dp/cybercenter/cci.pdf, accessed December 2007

Sidebar: PiiMS

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 is critical that every single piece of related information be collected and presented to the researcher in a digestible format. The Purdue Ionomics Information Management System (PiiMS) (http://www.purdue.edu/dp/ionomics/) [1] is an example of a web-accessible and community-based cyber infrastructure targeting such discovery process and related data and metadata acquisition activities. It is being used by various researchers from different disciplines and locations. Currently, different 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 working where each lab will be able to manage the different experimental and analysis stages through a shared CI.

PiiMS has been developed to promote 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. PiiMS' main functionalities include: (i) collecting and managing elemental profiling data and associated metadata on the experimental treatment, sample preparation, and instrument settings necessary to interpret the results, (ii) supporting the entire process of planting, growing, harvesting, drying, and analyzing of plants; (iii) providing integrated workflow control, data storage, and analysis to facilitate high-throughput data acquisition, along with integrated tools for data search, retrieval, and visualization for hypothesis development.

The CCI requirements of data handling and modeling (cf. Section 3) are satisfied by an elaborate data and metadata upload and gathering mechanism. Visualization and analysis requirements in PiiMS are satisfied by the search portal which allows users to view summaries and plots as well as compose queries. Requirements of modularity, agility and commoditization (cf. Section 3) are also key design features of PiiMS. Each element of data analysis is modular in the sense that data can be retrieved at any stage of analysis and a different line of analytics applied to it. The use of common components such as LDAP, persistence layer, computation and statistics packages, and graphing and reporting tools render PiiMS more agile as well as commoditized[6].

References

[1] Ivan Baxter, Mourad Ouzzani, Seza Orcun, Brad Kennedy, Shrinivas S. Jandhyala and David E. Salt, "Purdue Ionomics Information Management System. An Integrated Functional Genomics Platform", Plant Physiology, February 2007, Vol. 143, pp. 600-611

[6] See http://www.purdue.edu/cybercenter for more details and similar projects

Sidebar: NanoHUB

NanoHUB (http://www.nanohub.org), a web-based resource for research, education, and collaboration in nanotechnology, is a collaborative CI of the NSF-funded Network for Computational Nanotechnology (NCN). The 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 cyber infrastructure. NCN students, staff, and faculty are developing the nanoHUB science gateway while making use of it in their own research and education. Collaborators and partners across the world have joined the 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 sophisticated interactive simulation tools from any Web browser. Users can access and share all kinds of resources including live simulations. NanoHUB provides three resource components to its community of researchers: (i) Compute resources- Local, Teragrid, and Open Science Grid, (ii) Interfaces - Simple experiment setup and analysis, and (iii) Models- State of the art research in nanotechnology. In the past 12 months, over 5,500 users have executed over 211,000 simulations. The total annualized user number now exceeds 23,400 users, who run simulations and explore nanoHUB content such as tutorials, seminars, and classes, delivered as interactive lectures, podcasts and pdf files.

NanoHUB's users log on to a web portal and access state of the art models, run graphical simulations and view results online. Users do not need to download, install and configure any software component with the exceptions of few plugins to their favorite browser. In-VIGO [1], the NanoHUB back-end virtualization service, manages all jobs by locating a suitable machine or a cluster of machines and booting a virtual machine with suitable resources matching the user's requirements. Users only know that they are running a specific application but are masked from the implementation details such as machine type and configuration. NanoHUB truly adheres to the requirements of service and resource virtualization (cf. Section 3) by virtualizing the physical hardware and software required by the user. Another interesting feature of NanoHUB is the ability of a community of researchers to interact with each other in terms of sharing of data, models and knowledge in general.

References

[1] Lundstrom, M.; Klimeck, G., "The NCN: Science, Simulation, and Cyber Services," *Emerging Technologies - Nanoelectronics, 2006 IEEE Conference on* , vol., no., pp. 496-500, 10-13 Jan. 2006

Sidebar: Cyber Infrastructure across Disciplines

The crucial role and need of cyber infrastructure to support discovery has been the subject of several prominent publications and journals and also the launch of several large scale CI projects around the world. The vision of a cyber infrastructure for discovery in science and engineering has been outlined by the National Science Foundation in the form of two reports titled *Cyber infrastructure Vision for the 21st Century Discovery* [1] and *Revolutionizing Science and Engineering through Cyber infrastructure: Report of the National Science Foundation, Blue-Ribbon Advisory* [2]. The latter envisions cyber infrastructure to be used "…for the empowerment of specific communities of researchers to innovate and eventually revolutionize *what they do, how they do it, and who participates*." *Science and Engineering Infrastructure for the 21st Century, The Role of the National Science Foundation* [3] explores the current state and future direction of the science and engineering infrastructure and highlighting the role of the National Science Foundation. *Understanding Infrastructure: Dynamics, Tensions and Design* [4] provides a historical and philosophical insight into the development of cyber infrastructure by comparing its development with the more conventional infrastructure such as electricity, rail, telephone etc.

Other government agencies also expressed their cyber infrastructure vision in the form of a number of reports. *Preserving Electronic Records: Developments at the National Archives and Records Administration* [5] outlines the National Archives and Records Administration's (NARA's) commitment to create a cyber infrastructure to electronically archive and preserve all information related to the government of United States. NARA's vision states that *"Advancement and discovery in the 21st century are driven by data, … preserving our most valuable digital assets is critical for leadership and competitiveness in research and education…."* Similarly, the National Science and Technology Council report titled *Federal Plan for Cyber Security and Information assurance Research and Development* [6] 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*."

The Department of Defense's vision for net-centric operations and warfare [7] outlines the DoD vision of a cyber infrastructure 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…"*

Developing the UK's e-infrastructure for science and innovation [8] - sets out the requirements for a national e-infrastructure to help ensure the UK maintains and enhances its global standing in science and innovation in an increasingly competitive world.

Our Cultural Commonwealth [9] discusses in detail the positive impact of cyber infrastructure on the research and inquiry in humanities and social sciences.

The *GEONgrid* (http://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 cyber infrastructure for Earth Sciences research as well as learning at the K-12 and professional levels.

References

[1] NSF Cyberinfrastructure Council, "Cyber infrastructure Vision for 21st Century Discovery", March 2007, http://www.nsf.gov/pubs/2007/nsf0728/index.jsp, accessed December 2007

[2] D. E. Atkins, K. K. Droegemeier, S. I. Feldman, H. Garcia-Molina, M. L. Klein D. G. Messerschmitt, P. Messina, J. P. Ostriker, M. H. Wright, "Revolutionizing Science and Engineering Through Cyberinfrastructure", Report of the National Science Foundation, Blue-Ribbon Advisory, Jan. 2003, http://www.pnl.gov/scales/docs/cyberinfra_2003.pdf, accessed December 2007

[3] National Science Board, "Science and Engineering Infrastructure for the 21st Century: The Role of the National Science Foundation", National Science Board, Dec. 4, 2002, https://www.rand.org/pubs/monograph_reports/MR1728/MR1728.ch7.pdf, accessed December 2007

[4] P. Edwards, S Jackson, G. Bowker, and C. Knobel. "Understanding Infrastructure: Dynamics, Tensions, and Design", Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures" January, 2007, http://connect.educause.edu/Library/Abstract/UnderstandingInfrastructu/37182?time=1197911478, accessed December 2007

[5] K. Thibodeau, "Preserving Electronic Records: Developments at the National Archives and Records Administration", Electronic Records Archives Program June, 2004, http://www.archives.gov/ear/pdf/thibodeau-040617.pdf, accessed December 2007

[6] Interagency Working Group on Cyber Security and Information Assurance, "Federal plan for cyber security and information assurance research and development", Networking and Information Technology Research and Development, April 2006, http://handle.dtic.mil/100.2/ADA462532, accessed December 2007

[7] Defense Information Systems Agency "Surety, Reach, Speed, The Disa Strategy", March 2007, http://www.disa.mil/, accessed December 2007

[8] Report of the OSI e-Infrastructure Working Group, "Developing the UK's e-infrastructure for science and innovation" National e-science Center, 2004, www.nesc.ac.uk/documents/OSI/index.html, accessed December 2007

[9] "Our Cultural Commonwealth", The Report of the ACLS Commission on Cyberinfrastructure., http://www.acls.org/cyberinfrastructure/acls.ci.report.pdf, accessed December 2007