Rosen Center for Advanced Computing

October 10, 2008

The Hub Concept for Scientific Collaboration

Michael McLennan, Senior Research Scientist and Software Architect for Hubs

Background

Starting in October 2002, the Network for Computational Nanotechnology (NCN) was formed by the National Science Foundation (NSF) and charged with a mission to create, deploy, and operate a national resource for theory, modeling, and simulation in nanotechnology, to connect users in research, education, design, and manufacturing.[1-2]

In the course of this work, the NCN created a cyberinfrastructure embodied by the web site at nanoHUB.org. The popularity of nanoHUB has skyrocketed since its inception. During the period October 2007 to September 2008, more than 81,000 people accessed nanoHUB to view a collection of seminars, tutorials, animations, publications, and simulation tools submitted by more than 550 contributors from all over the world. A little more than half of all users come from the United States, and less than 15% are from Purdue; the others come from 172 countries all over the world. The nanoHUB has users at all of the Top 50 Engineering Schools,[3] at 333 international educational institutions, and at 14% of all available `.edu` domains.

Some people compare nanoHUB to the highly successful Open Courseware Initiative from MIT.[4] But the nanoHUB is more than just a repository for course materials. It is a place where researchers and educators can meet and accomplish real work. The nanoHUB offers integrated, online web meetings via Macromedia Breeze, source code collaboration through its nanoFORGE area, event calendars, and many other services designed to connect researchers and build a community. But most importantly, the nanoHUB connects users to the simulation tools they need for research and education.[5-7] Users can access more than 120 interactive, graphical tools, and not only launch jobs, but also visualize and analyze the results, all via an ordinary web browser. The NCN's emphasis on usability has produced a clean interface that makes it easy to use

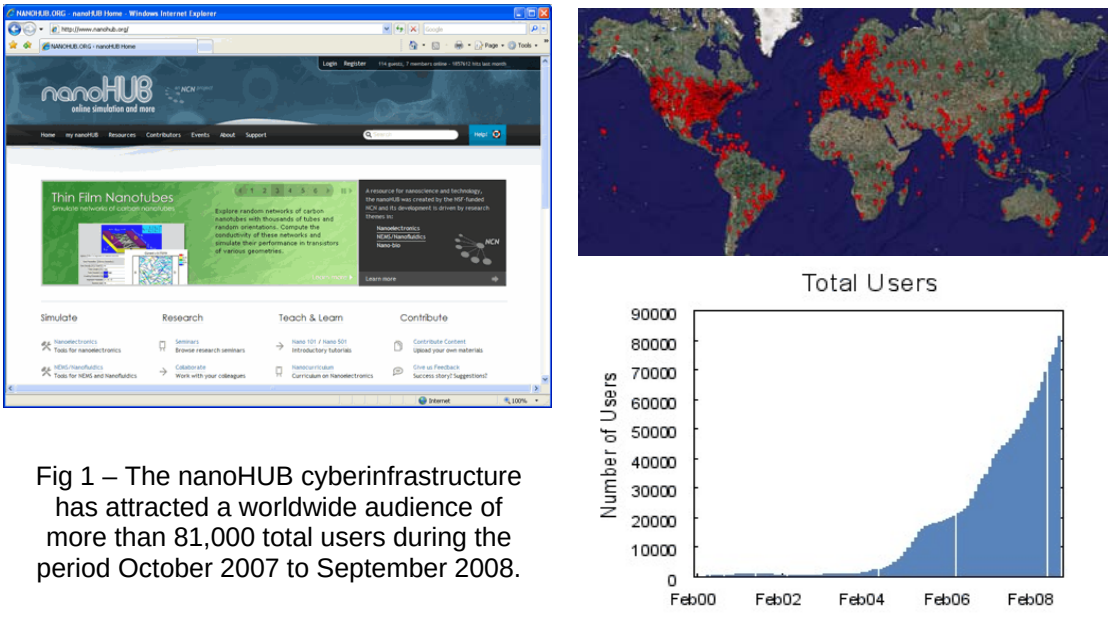
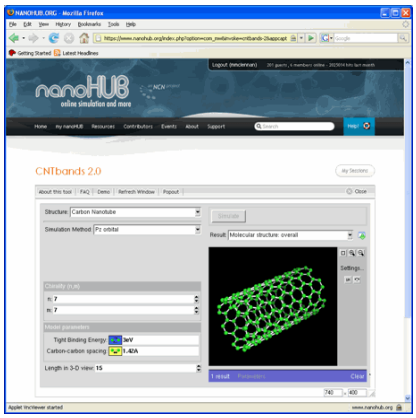


Fig 1 – The nanoHUB cyberinfrastructure has attracted a worldwide audience of more than 81,000 total users during the period October 2007 to September 2008.

powerful research tools. Simulation jobs can be dispatched on national Grid resources, including the NSF TeraGrid and the Open Science Grid. The nanoHUB middleware hides much of the complexity of Grid computing, handling authentication, authorization, file transfer, and visualization, and letting the researcher focus on research. This approach also helps educators bring these tools to the classroom, letting them bypass the mire of Grid computing and focus instead on the physics they are trying to teach.

(a)



(b)

(c)

Fig 2 – The nanoHUB offers more than 1,300 resources, including (a) simulation tools that you can access via a standard web browser, along with (b) seminars, tutorials, and (c) podcasts, to help explain the science behind the tools.

Hub Technology for Your Project

Much of the cyberinfrastructure developed for this project has nothing to do with nanotechnology per se, but can be applied to many scientific disciplines with a need for simulation, modeling, and collaboration. The Rosen Center for Advanced Computing (RCAC), the computing research division of Information Technology at Purdue (IT@P), has helped to develop and maintain the nanoHUB throughout its existence. RCAC has a solid track record for production-quality support. Last year, RCAC maintained more than 21 TeraFlops of high-performance computing equipment, and kept nanoHUB running with better than 99.5% uptime!

RCAC is now offering "hub" technology to everyone on Purdue's West Lafayette campus. This whitepaper outlines the process of setting up an empty hub using a software package that we will refer to as HUBzero™. Each new hub will have the same capabilities as nanoHUB, but will have its own content, its own tools, and its own community of scientific users.

NOTE: What we are offering is a "hosting service," whereby we will setup a hub site and keep it running. Your team must be responsible for creating tools, tutorials, and other content and uploading that content onto the hub. If you would like some assistance in creating content, we have staff available to help. However, we would need to discuss your applications and budget extra time for the appropriate level of support.

Your team should also include a ***hub administrator***, who will fill in "about" pages and contact information, approve contributions, respond to support tickets, and generally keep an eye on your new hub. This is often a graduate student, post doc, or administrative assistant associated with the project. You may even designate multiple administrators to help share the load.

Your hub will *NOT* include web meeting capabilities or any other commercial software unless you pay an extra charge to license the appropriate software.

How Does a Hub Differ From a Web Site?

At its core, a hub is a web site built with many familiar open source packages—an Apache web server, PHP web scripting, Joomla content management system, and a MySQL database for storing content and usage statistics. The HUBzero™ software builds upon that infrastructure to create an environment in which researchers, educators, and students can access tools and share information. Specifically, we define a "hub" as a web-based collaboration environment with the following features:

- **Interactive Simulation Tools**

  The signature service of a hub is its ability to deliver interactive, graphical simulation tools through an ordinary web browser. In the world of portals and cyber-environments, this capability is completely unique. Unlike a portal, the tools in a hub are interactive; you can zoom in on a graph, rotate a molecule, probe isosurfaces of a 3D volume—interactively, without having to wait for a web page to refresh. You can visualize results without having to reserve time on a supercomputer or wait for a batch job to engage. You can deploy new tools without having to rewrite special code for the web.

  The HUBzero™ infrastructure includes a tool execution and delivery mechanism based on Virtual Network Computing (VNC).[8] Any tool with a graphical user interface can be installed on the hub and deployed with a few hours. For legacy tools and other codes without a graphical interface, an interface can be quickly created by using the Rappture toolkit that comes with HUBzero™.[9] The Rappture interface helps to set up jobs and visualize results. The jobs themselves can be dispatched to the TeraGrid, the Open Science Grid, and other participating cluster resources via Condor[10], as shown in Figure 3. Using this architecture, the nanoHUB has brought over 120 different simulation tools online in just 4 years, with 80 more tools currently under development. We expect similar growth and performance for other hubs.

- **Online Presentations**

  In order for users to make the most of the tools on a hub, they need to understand the limitations of each tool and its underlying science. Along with the tools, each hub features a series of online presentations, which are PowerPoint slides combined with voice and

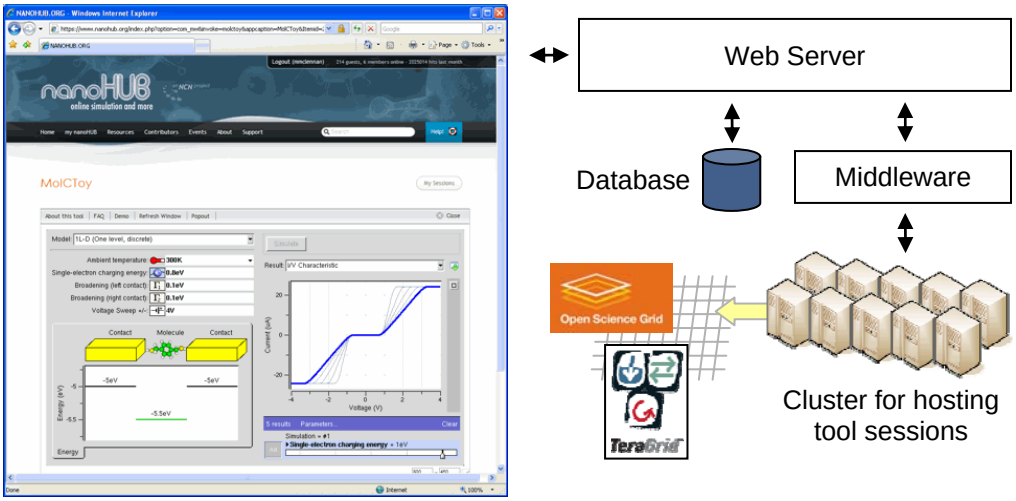


Fig 3 – In the world of portals and Grid computing, the HUBzero™ cyberinfrastructure is completely unique. Users access interactive, graphical tools through a web browser. The tools themselves run on a cluster at Purdue, where jobs can be dispatched to more powerful computers in the national Grid infrastructure, including the TeraGrid and the Open Science Grid.

animation. Listening to a presentation on a hub is much like being in the room during a standard seminar. But unlike a seminar, the material is available 24x7, and you can skip through the talk by browsing the table of contents or searching for important keywords. HUBzero™ currently uses Macromedia Breeze® to deliver the online presentations in a very compact format using Flash®, which is installed on 98% of the world's internet-enabled desktops.[11] Unlike streaming video, the Flash® format can be viewed over a dial-up connection, so presentations have greater reach into dorm rooms, K-12 classrooms, and other places where network bandwidth may be limited. Online presentations can also be distributed as podcasts, so your users can access them on-the-go via their video or audio iPod.

- **Mechanism for Uploading New Resources**

  Each hub is a place for users to come together and share information. One important way to accomplish this is by encouraging all users to upload their own tools, presentations, and other materials onto the hub. The HUBzero™ software includes a self-service area that guides the user through the process—much like purchasing something on the web. On the first screen, the user enters a title and an abstract, clicks *Next*, then uploads associated files, clicks *Next*, acknowledges a list of contributors, clicks *Next*, and so forth. At the end of the process, the resource is submitted for posting. Your own hub administrator approves all submissions, keeping out spam or any other inappropriate materials. Newly published items automatically appear on the *What's New* page of the hub, and are also available on a corresponding news feed for Really Simple Syndication (RSS) readers.

- **Tool Development Area**

  Uploading an online presentation or a PDF document is fairly straightforward, as described above. Uploading a tool, however, is a little more complicated. Tools must be uploaded, compiled, tested, fixed, compiled again, and tested again—often many times—before being published. This requires a little more support from your own hub administrator, but HUBzero™ helps to automate the process. Each hub comes with a companion site for source code development based on the open source Trac package for project management. (Think of this as your own private version of SourceForge.net, the supporting infrastructure for the open source development community.) Each tool will have its own project area within this site, with a Subversion repository for source code control, a ticketing system for bug tracking, and a wiki area for project documentation.

- **Ratings and Citations**

  The hub philosophy is not to judge the quality of each resource before deciding to post, but rather, to post resources and let the community judge the quality. Registered users are allowed to post 5-star ratings and comments for each resource. Registered users can also post citations that reference the resource in the literature, so everyone can see other work that builds upon the resource. The ratings and citations for each resource are combined with web statistics (measuring the popularity of the resource) to produce a single number on a scale of 0 to 10, called the *ranking*, which defines the overall quality of the resource. Resources with the highest ranking appear at the top of the list for searching and browsing operations; resources with the lowest ranking are much harder to find on the hub.

- **Content Tagging**

  Each of the resources on a hub is categorized by a series of tags, which are arbitrary strings defined by the user when uploading content. Each tag has an associated page on the hub where its meaning is defined and its resources are listed. For example, the tag "carbon nanotubes" might help users find all tools, seminars, and tutorials related to carbon nanotubes. Another tag "K-12" might help educators find content appropriate for elementary education. Tags are not only defined by the contributor, but also by your own hub administrator, and can even be added by other users when they rate the resource.

- **Wikis and Blogs**

  Each hub supports the creation of "topic" pages, which are similar to the Google "knol" model for knowledge articles. Each topic page is created with a standard wiki syntax by a specified list of authors. Other users can be allowed to add comments to the page or even suggest changes. The original authors are notified of changes suggested by other users. The changes can be incorporated, and the users suggesting them can be added as co-authors for the page, so they can make further changes without approval. The ownership of the page can also given to the entire community, so anyone can make changes without approval, in a wiki-like manner.

  Topic pages act as lightweight (1-2 page) articles that help to describe various resources on the hub and pull them together in a coherent way. Each topic page that includes a reference to a resource is listed at the bottom of the resource page under a "see also" heading. This helps users who have found a resource see other articles related to it, and discover other related resources.

- **User Groups for Private Collaboration**

  During the operation of the nanoHUB, we have found that many users like to limit their collaborations to a smaller audience. For example, a researcher might upload a presentation intended only for other members of his research group. An employee at IBM might upload a tool, but only for use by other IBM employees. The HUBzero™ software supports this by letting users create and manage their own groups of users. Any registered user can create a group and invite others to join it. The creator can accept or reject group members, and can promote various members to help manage the group. Resources associated with a group can be kept private, meaning that their access is limited to other members of the group.

- **User Support Area**

  From time to time, users will have problems with logins, questions about tools, and may otherwise need assistance. The HUBzero™ software comes with a built-in user support area. Users can click on the *Help* link near the top of any page and fill out a form to file a support ticket. If a tool encounters an unexpected error, a ticket is filed automatically. By examining, updating, and closing the tickets, your own hub administrator can keep track of the people having problems and the resolution of each issue.

  Some of the questions that users will ask are beyond the understanding of a single hub administrator, or even a hub team! For this reason, HUBzero™ includes a question-and-answer forum patterned after http://answers.yahoo.com and http://askville.amazon.com. Any registered user can post a question, and other users can provide answers. At some point, the best answer is chosen as the "final" answer by the person who originally asked the question. The list of past questions/answers forms a knowledge base upon which your community can draw for immediate help with a similar problem.

- **Usage Statistics**

  Each hub reports statistics about how its resources are being used, including the total number of users in a given period, the number of web hits, simulation jobs launched, CPU hours used, etc. Statistics are reported down to the level of each individual resource, so you can see how many users have accessed a particular tool, or how many times an online presentation has been viewed. Usage numbers are rolled up to provide an overview of usage for interesting categories. For example, you can see how many users access all tools, how many accessed the resources from a single contributor, how many are located in the US, how many are working in industry, and so forth.

- **News and Events**

  Each hub includes a calendar and a mechanism for any registered user to post events. This helps your hub become a focal point for the community. Each hub also includes a news area, where your hub administrator can post short stories that describe the progress being made by researchers on your hub.

- **Feedback mechanisms**

  Each hub includes a feedback area where users can respond to a poll question, share a success story, or provide other comments and suggestions.

Hub Configuration

Since RCAC has already developed the appropriate software, the cost of setting up a new hub is minimal. The cost of operating a hub may increase over time as the number of users grows from tens, to hundreds, to thousands, and beyond. In this section, we outline the costs of establishing an empty hub for thousands of users, assuming that no more than 200 users will be running simulation tools at any given time.

Each hub will be installed with the standard HUBzero™ source distribution, an empty content database, and an empty list of users. Once the hub is open for business, users will be able to create accounts, log in, and upload content. Your own hub administrator—or team of administrators—will have special privileges to approve uploaded content and customize the look of the site. If your administrators choose to open a user poll, for example, they can do so at any point by activating the poll component and entering a question; if they don't want to use polls, they can turn that feature off. If they want to move the poll to another page or a non-standard spot, they can do so by editing the HTML code within a template or on a static content page, with advice as needed from the HUBzero™ team.

The basic configuration of each hub is as follows:

Standard Features:

- ✓ One standard cluster node in nanoHUB rack (Dell 1950, 8GB RAM, 2x750GB disks RAID1)
- ✓ Disk space for 500 users (500 GB total space)
- ✓ Apache / PHP / Joomla + customized HUBzero™ components
- ✓ LDAP server for registered users and tools
- ✓ MySQL database for all content
- ✓ Trac project management and subversion repository (developer area for tools)
- ✓ Customized CSS (graphic design) for HUBzero™ site
- ✓ Access to tool sessions for 200 simultaneous users
- ✓ Workspaces (Linux machines accessible within a web page)
- ✓ Off-site backup of database and all content files

The following features are *NOT INCLUDED* in the standard HUBzero™, but can be added later at additional cost. If you are interested in these features, contact us to generate a quote.

Extra Features:

- ✓ Online web meetings with Macromedia Breeze (requires additional licenses)

As your user base grows, it may be necessary to expand beyond one cluster node, beyond 200 simultaneous tool users, and so forth. Such expansions are outside of the scope of the current document, and will be detailed elsewhere. As a rule of thumb, however, you should budget an additional $30,000 for hardware refresh (for additional CPU and disk space) after about 3 years of operation.

Cost

The total cost for setting up each starter hub is estimated as follows:

Hardware	$6,000
Off-site backup ($10/GB)	$7,500 (one-time cost)
OS installation/machine configuration	1 week
Web server + HUBzero™ + MySQL + LDAP	1 week
Middleware + network configuration	1 week
Customized CSS development	3 weeks
Ongoing training for hub administrators	1 week/year
Ongoing log/statistics processing	1 week/year
Ongoing machine maintenance/support	1 week/year
Ongoing OS/security updates	1 week/year
Ongoing HUBzero™ software updates	1 week/year
Ongoing customizations and tool development	1 week/year

The total direct cost is about $39k for the first year, and $14k for each additional year. Assuming 52% for overhead, the total cost for 2 years of operation including indirect costs would be approximately $75k, the cost for 3 years of operation would be $95k, and so forth. ***Note that these approximate costs are included here only for illustrative purposes. If you are writing a grant proposal, contact us and we will prepare a more detailed COEUS budget for the hub portion of your project.***

Support

The HUBzero™ team will provide:

- training on the operation of your hub
- consultation with hub administrators during normal business hours to resolve unexpected issues
- software upgrades and machine maintenance
- backup of the content database and user data files

Your own hub administrators will be responsible for things like:

- front-line support of all tickets filed by your user community
- approval of all content posted to your hub
- posting of news items and events on your hub
- installation of simulation tools on your hub
- creation of podcasts from online presentations

Upgrades

From time to time, the HUBzero™ team will make improvements to the core software. Each team of hub administrators will decide for themselves when to upgrade to the latest HUBzero™ software distribution. The HUBzero™ team will perform the actual upgrade at a time mutually agreed upon. Each upgrade may introduce bugs or other instabilities, so hub administrators are encouraged to choose a time for the upgrade most convenient to their own schedule (*i.e.,* not on the day before an important demo).

Customizations

Your hub may require additional features that are not available in the current HUBzero™ distribution. We are happy to collaborate with you to help implement new features, particularly if they might be useful for other hubs. Your proposal to a funding agency should include appropriate support for the RCAC staff members who will be developing your customized feature.

IT@P is Your Partner

IT@P can be a powerful partner in your funding proposals. Our technical staff consists of some of the world's leading experts in cyberinfrastructure, and we have developed unique expertise to create and deploy innovative solutions for the scientific community. Our staff looks forward to working with your hub team, to leverage our expertise in creating another successful, web-based virtual organization.

References

1. M.S. Lundstrom and G. Klimeck, "The NCN: Science, Simulation, and Cyber Services," *2006 IEEE Conference on Emerging Technologies - Nanoelectronics*, 10-13 Jan. 2006 Page(s):496 - 500.
2. G. Klimeck, M. McLennan, S.P. Brophy, G.B. Adams III, M.S. Lundstrom, "nanoHUB.org: Advancing Education and Research in Nanotechnology," *Computing in Science and Engineering*, 10(5), pp. 17-23, September/October, 2008.
3. Available on the World Wide Web at: http://grad-schools.usnews.rankingsandreviews.com/usnews/edu/grad/rankings/eng/brief/engrank_brief.php
4. Available on the World Wide Web at: http://ocw.mit.edu/
5. E. Howell, C. Heitzinger, and G. Klimeck, "Investigation of Device Parameters for Field-Effect DNA-Sensors by Three-Dimensional Simulation," in *Proceedings of IEEE Nanotechnology Materials and Devices Conference*, October 22-25, pg 154, 2006.
6. W. Qiao, M. McLennan, R. Kennell, D.S. Ebert, G. Klimeck, "Hub-based simulation and graphics hardware accelerated visualization for nanotechnology applications," *IEEE Trans Vis Comput Graph*, 12(5):1061-8, Sept-Oct 2006.
7. P. Ruth, X. Jiang, D. Xu, and S. Goasguen, "Virtual distributed environments in a shared infrastructure," *IEEE Computer*, 38(5):63–69, May 2005.
8. T. Richardson, Q. Stafford-Fraser, K. R. Wood, and A. Hopper, "Virtual network computing," *IEEE Internet Computing*, 2(1):33–38, 1998.
9. M. McLennan, The Rappture Toolkit (2004) on World Wide Web at http://www.rappture.org
10. D. Epema, M. Livny, R. van Dantzig, X. Evers, and J. Pruyne, "A Worldwide Flock of Condors: Load Sharing among Workstation Clusters," *Future Generation Computer Systems*, 12, 1996.
11. According to data from Adobe on the World Wide Web: http://www.adobe.com/products/player_census/flashplayer/