WCER Working Paper No. 2010-12
October 2010

Cyberinfrastructure and Scientific Collaboration: Application of a Virtual Team Performance Framework with Potential Relevance to Education

Sara Kraemer
Wisconsin Center for Education Research
University of Wisconsin–Madison
sbkraeme@wisc.edu

Christopher A. Thorn
Wisconsin Center for Education Research
University of Wisconsin–Madison
cathorn@wisc.edu

Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison • http://www.wcer.wisc.edu/

WCER working papers are available on the Internet at http://www.wcer.wisc.edu/publications/workingPapers/index.php. Recommended citation:

> Kraemer, S., & Thorn, C. A. (2010). *Cyberinfrastructure and scientific collaboration: Application of a virtual team performance framework with potential relevance to education* (WCER Working Paper No. 2010-12). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingPapers/papers.php

The research reported in this paper was supported by grants from the Chicago Public Schools (O6-1025-PR16), the Children First Fund: The Chicago Public Schools Foundation (C2005-01251 and C2006-01405), the Spencer Foundation (20070013), the Joyce Foundation (06-2980), and the New York City Teacher Data Initiative (907507), and by the Wisconsin Center for Education Research, School of Education, University of Wisconsin–Madison. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies, WCER, or cooperating institutions.

Cyberinfrastructure and Scientific Collaboration: Application of a Virtual Team Performance Framework with Potential Relevance to Education

Sara Kraemer and Christopher A. Thorn

High throughput computing (HTC)—and grid technologies in general—are ubiquitous in, and mission-critical to, interdisciplinary research that requires large amounts of computing power and access to expert advice (Thain, Tannenbaum, & Livny, 2005). We define HTC as an environment that can deliver large amounts of processing capacity over long periods of time. In addition to computational power delivered, there is a second, critical measure of system quality: HTC systems are designed to be extremely fault-tolerant and to require minimal human intervention (Livny & Raman, 1998). These and other characteristics enable and support distributed teams, driving the interactions and forms of collaboration that emerge when users from various scientific domains use HTC resources to work on computational problems.

In this paper, we explore the sociotechnical facets of scientific collaborations using HTC resources through the lens of a virtual team performance framework. HTC systems that effectively foster interdisciplinary virtual team collaboration feature resource flexibility, end-user control, open-ended planning, and distributed resource management (Thain et al., 2005). The Condor Project[1] at the University of Wisconsin–Madison (UW–Madison) is one such HTC system. HTC generally, and Condor specifically, are drivers of leading-edge science collaborations in co-located research teams and in large-scale, internationally distributed production environments.

The examination of collaborations and virtual teams using HTC has potential relevance to the field of education. One of the core needs for K–12 education reform is the effective development and maintenance of district-wide longitudinal data systems. An Institute of Education Sciences (IES) panel conducted a systematic review of literature on the effective use student achievement data to support instructional decision making and recommended that school districts develop and maintain high-quality data systems (Hamilton et al., 2009). High-quality, district-wide data systems are necessary to provide teachers, principals, school staff, and district staff the information they need to make decisions for student learning.

As districts transition from using data for accountability purposes to using it for ongoing improvement and decision making, data systems will need to handle an ever greater volume (e.g., student outcome scores, benchmark assessment data, locally developed formative assessments, attendance records, finance information, demographic information) and be configured to permit ever more sophisticated analysis (e.g., using multiple data systems for multiple purposes, refining data to a more granular—e.g., classroom—level). The current

We thank Miron Livny at the Condor Project, University of Wisconsin–Madison (UW–Madison), for input into the development of these ideas. We also thank Francis Halzen at IceCube and David Schwartz at the Laboratory for Molecular and Computational Genomics at UW–Madison for their participation in this research project.

[1] http://www.cs.wisc.edu/condor/

infrastructure design in districts presents a significant challenge; these accountability-based designs are outmoded for the demands of data-driven decision making. HTC is poised to support districts in the development of their data systems because it supports environments that require a high volume of computing cycles with minimal human oversight or support.

Overview and Research Questions

We report here on an exploratory case study of two scientific collaborations at the Grid Laboratory of Wisconsin (GLOW),[2] one of several Condor computing resource pools at UW–Madison. GLOW is an interdisciplinary effort that spans 11 scientific domains: biostatistics and medical informatics, chemical and biological engineering, chemistry, computer sciences, engineering physics, genomics, genetics, materials science and engineering, medical physics, physics, and astrophysics. The resources are distributed among six laboratory sites. The laboratories provide the necessary hardware, software, and support infrastructure for the development and experimental evaluation of HTC applications. Each of the sites addresses local computational needs and maintains full control over local resources while sharing unused computing power and storage space across site boundaries according to a group-defined policy. The goal of GLOW is to bring together domain and computer scientists to make HTC computing an effective tool for scientific research by harnessing and sharing the power of commodity resources. GLOW members collaborate in the development, implementation, testing, sharing, and deployment of grid-enabled capabilities while cultivating interdisciplinary science.

We conducted exploratory interviews with team members from two GLOW projects that differed across a number of characteristics such as membership size, research purpose, and range of scientific disciplines: (a) *IceCube*,[3] a particle physics collaboration that is constructing and currently operating a neutrino telescope in the Antarctic ice; and (b) the *Laboratory for Molecular and Computational Genomics* (LMCG),[4] whose investigations contribute to the creation of new systems such as optical mapping, optical sequencing, and nanocoding. Our aim was to explore the viability of applying a virtual team performance framework to collaboration using distributed cyberinfrastructure technologies and to assess the potential relevance of the application to education reform and research.

Two research questions guided the study:

- What are the sociotechnical characteristics of virtual teams using Condor and HTC technologies?
- How do the characteristics of the Condor HTC technology affect virtual team performance and collaboration?

[2] http://www.cs.wisc.edu/condor/glow/index.html

[3] http://icecube.wisc.edu/

[4] http://www.lmcg.wisc.edu/

Condor, HTC, and Collaboration

The Condor Project embodies a philosophy of flexibility, and this philosophy has allowed the Condor design to flourish in a highly unpredictable distributed operating environment (Thain et al., 2005). International distributed systems are heterogeneous in numerous ways: they are composed of many types and brands of hardware; they run various operating systems and applications; they are connected by unreliable networks; they change configuration constantly as old components become obsolete and new components come online; and they have many owners with private policies and requirements that control their participation in the community. Condor has adopted a five-component philosophy of flexibility to address these challenges and enable virtual team collaboration (Thain, Tannenbaum, & Livny, 2003, pp. 301–302):

- *Let communities grow naturally.* Given tools of sufficient power, people will organize the computing structures they need. However, human relationships are complex, people invest their time and resources to varying degrees, and relationships and requirements change over time. Therefore, Condor design permits but does not require cooperation.

- *Leave the owner in control, whatever the cost.* To attract the maximum number of participants to a community, the barriers to participation must be low. Owners of computing resources will not donate their property for the common good unless they maintain some control over how it is used. Therefore, Condor gives owners the tools to set policies and retract resources for private use.

- *Plan without being picky.* It is critical to plan for slack resources as well as resources that are slow, misconfigured, disconnected, or broken. The designers of Condor spend more time and resources contemplating the consequences of failure than the potential benefits of success.

- *Lend and borrow.* The Condor Project has developed a large body of expertise in distributed resource management and aims to give the research community the benefits of that expertise while accepting and integrating knowledge and software from other sources. It has also instituted a mechanism for collective problem sharing and problem solving among its users.

- *Understand previous research.* The Condor Project continually updates its organizational knowledge with previous research to apply both well-known fundamentals and cutting-edge techniques to emergent problems. The inclusion of current user innovations keeps the work focused on the edge of discovery rather than wasting effort remapping known territory.

The Condor Project is more than a complex set of computational resources. The Condor team maintains a close intellectual partnership with computer and domain scientists working together on the challenges of HTC in the context of breakthrough science. Condor has advanced HTC technology through improvements in their software coupled with innovations in computational approaches developed by a wide range of domain scientists. These interactions have acquainted Condor team members with numerous sociotechnical problems affecting interdisciplinary virtual team performance.

Research at the Condor Project suggests that interdisciplinary collaboration using HTC resources has important sociotechnical implications for collaborative research and technology

design (Thain, Tannenbaum, & Livny, 2006). Interesting interactions arise when the barriers to computational resource access are removed. For example, some virtual teams using HTC embody what is known as *tool-based specialization*—that is, they view their interaction with HTC as a function of their research areas—whereas other teams exemplify a *perspective-based specialization*—that is, the HTC ceases to be part of their research area, and problem situations define their research space (Schmidt, Rasmussen, Brehmer, & Leplat, 1991; Whitely, 1974). Those operating in perspective-based specialization may be more likely to share local research methods and solutions with other team members or teams and view HTC as an enabling technology, rather than simply as resource sharing.

As HTC technologies mature, they move out of a tool-based environment and become a part of the research infrastructure. This change may lower perceived risk and enhance trust among users, thereby enabling new models of financial support and social engagement. As use of the tools expands from basic to applied research and production work, new risks and benefits are likely to appear. For example, computer scientists may perceive a risk of being identified as merely an infrastructure provider. On the other hand, for those contemplating the adoption of HTC, the emergence of HTC as a commodity service may greatly reduce perceived risk. As HTC systems provide increasingly ubiquitous access to robust computing environments, potential users will see HTC as a relevant and important aspect of their work. Fault resistance—the other significant element of HTC system success and a key characteristic of the Condor Project—is another critical enabler at the margin of adoption (Livny & Raman, 1998).

Perspective-based collaboration is particularly common in experimental research involving complex instrumentation, such as telescopes, particle accelerators, or CT scanners (Katz & Martin, 1997). For example, high-energy particle physics is a domain that has embodied perspective-based collaboration characterized by collectivism, erasure of the individual as an epistemic subject, nonbureaucratic mechanisms of work, lack of overbearing formal structures, and an absence of rigid rules (Chompalov, Genuth, & Shrum, 2002). These characteristics are representative of possible sociotechnical components of a typology of interdisciplinary virtual team scientific collaboration.

Virtual Teams in HTC

A virtual team is a group that works across time and distance and whose interactions are mediated by technology (Driskell, Radtke, & Salas, 2003). Although the types of technology and communication enabled in virtual team environments vary, the core feature of a virtual team is that its members work together on a common task while they are spatially separated. Team members may be geographically and temporally dispersed; permanent or nonpermanent; members of different organizations, countries, or cultures. They can meet partially or fully in cyberspace. Virtual teams in HTC may vary across time and geography, domains of science, team size, background or culture, type of task, type of research problems (e.g., applied, basic), computational needs, fluidity of membership in the HTC community, and degree of interdisciplinarity within their scientific domain and/or across research projects.

Research on virtual teams has emphasized the study of electronic communication technologies and processes as mediators of team performance (Martins, Gilson, & Maynard, 2004). Communication technology and virtuality contribute to the transformation of teamwork in

three important ways: (a) they introduce new dimensions of communication among members by breaking down traditional barriers of space and time; (b) they modify traditional group processes; and (c) they enhance a team's ability to access, share, manipulate, retrieve, and store information (Godar & Ferris, 2004). Virtual team communication technologies may include videoconferencing, Internet chat rooms, e-mail, and bulletin boards (Avolio, Kahai, Dumdum, & Sivasubramaniam, 2001). In addition to these previously studied performance-mediating technologies, other forms of technology such as HTC link distributed team members, mediate their performance, and give their teams unique characteristics.

As noted earlier, the core characteristic of HTC is its ability to provide large amounts of computing for sustained periods of time. However, HTC has a number of additional characteristics that provide greater access to a wide range of disciplines and toolsets. For example, because Condor runs on many computing platforms and can execute any software that does not require user interaction, a wide range of tools is readily available—from commercial research software to scripting engines and compilers. In addition, the abundance of available scientific tools allows individual scientists or teams to engage with the Condor HTC environment using tools familiar to them. The enabling of existing tools in an HTC setting provides critical social and technological gateways for new adopters. Access to the HTC environment also exposes new adopters to tools and methods used by others to address similar computational problems. In this way, scientists' skills and knowledge are affected by the capabilities and characteristics of HTC technologies and tools.

We argue that distributed computational cyberinfrastructure generally—and HTC specifically—are enabling technologies for virtual team collaboration. Virtual teams using HTC introduce new and novel ways of accomplishing research objectives that were previously unattainable due to limitations of computational power. With HTC, the obstacles to accessing computational power are removed, and scientists have access to other scientists all over the world, linked through their use of HTC.

Virtual Team Performance Framework

Our research framework (Figure 1) is an adaptation of Powell, Piccoli, and Ives' (2004) *input-process-output* (IPO) model. This model, which draws on the authors' review of the virtual team research literature, includes four general categories of variables:

- *Inputs* (design, culture, technical expertise, training);
- *Socioemotional processes* (relationship building, cohesion, trust);
- *Task processes* (communication, coordination, task-technology-structure fit); and
- *Outputs* (performance, satisfaction).

A key assumption of the IPO model is that the input states affect group outputs through the interactions that take place among members.

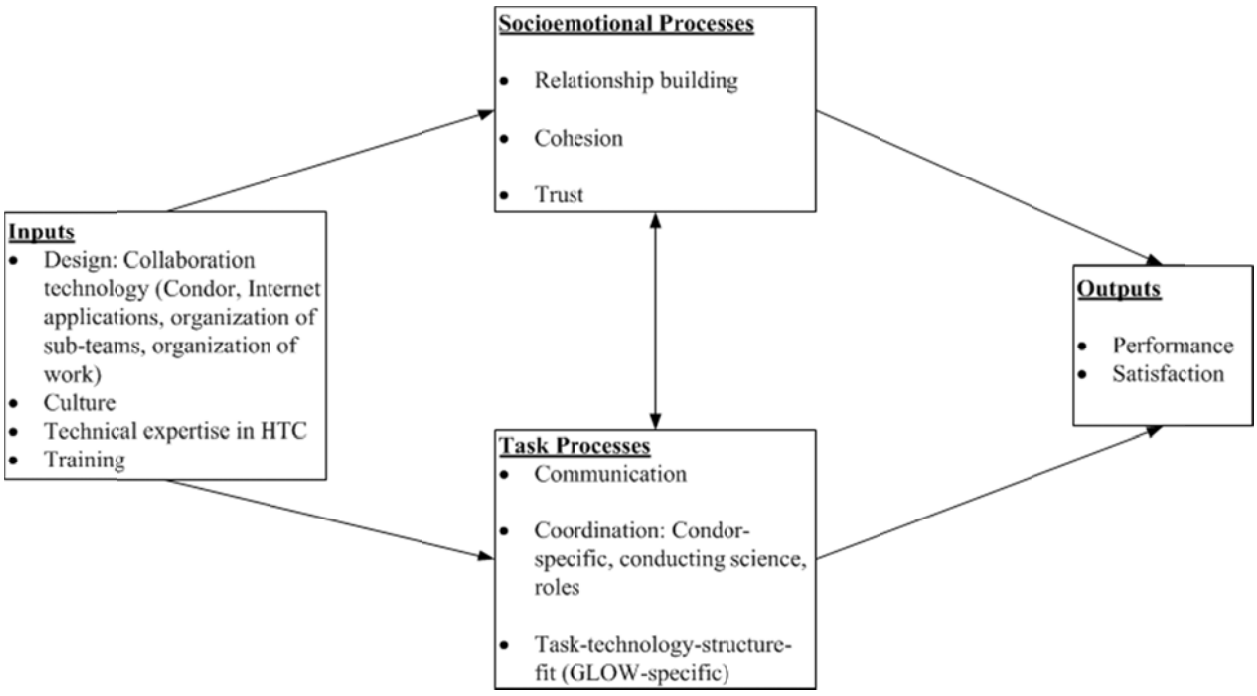

Figure 1. **Virtual team performance framework in HTC. Adapted from "Virtual Teams: A Review of Current Literature and Directions for Future" by A. Powell, G. Piccoli, and B. Ives, 2004*, ACM SIGMIS Database, 35*(1), p. 8.**

The IPO model has been identified as a qualified framework for modeling virtual team performance. Schiller and Mandviwalla (2007) approached the IPO framework from a theory synthesis perspective, analyzing the frequency, pattern of use, and ontological basis of virtual team–relevant theories. Other literature reviews have found the IPO model to be the dominant framework used in the study of teams and virtual teams and have provided a sound basis for organizing and describing the various factors that underlie and mediate virtual team performance (Cuevas, Fiore, Salas, & Bowers, 2004; Martins et al., 2004).

Methods

Qualitative techniques are often useful for gathering rich, detailed explanations of complex, intricate phenomena in order to reveal their context (Trochim, 2001). Our study design consisted of an exploratory, qualitative, case study analysis of two interdisciplinary virtual teams using HTC technologies. Focus groups were conducted with scientists and research staff at each of the two laboratories.

Sample

Our sample consisted of two GLOW interdisciplinary scientific research teams. The teams varied across several characteristics: size, duration of collaboration, scientific disciplines represented, organization of team members, and research topics and goals. The first team was IceCube, a collaboration supporting a neutrino detector at the South Pole; the second team was the Laboratory for Molecular and Computational Genomics (LMCG). We chose IceCube

because it represents a traditional approach to large-scale, distributed collaboration using infrastructure. For comparison, we chose LMCG, a much smaller team with a more limited scope of work in an interdisciplinary science that is relatively new compared to IceCube's field of particle physics. Below, we discuss some of the characteristics of these GLOW teams that led us to judge them qualified sources of information for this exploration of virtual team performance.

IceCube. UW–Madison is the lead institution for the construction and operation of the IceCube telescope at the South Pole. The project is an international collaboration of more than 250 faculty, scientists, postdocs, and students from 35 institutions. The IceCube telescope encompasses a cubic kilometer of ice and uses a novel astronomical messenger called a neutrino to probe the universe. It will search for neutrinos from the most violent astrophysical sources: events like exploding stars, gamma ray bursts, and cataclysmic phenomena involving black holes and neutron stars. The telescope is a powerful tool for searching for dark matter and could reveal the new physical processes associated with the enigmatic origin of the highest energy particles in nature.

GLOW has contributed to all aspects of the commissioning and operation of the detector, especially the challenging production of real-time detector simulations, which rely completely on GLOW resources. GLOW has enhanced the timeliness of delivering simulation and experimental results to IceCube team members, yielding a competitive advantage. Since the IceCube collaboration is relatively small for a particle physics collaboration (about one quarter the typical size), team members are able to more easily communicate and share findings, which fosters an overall culture of cohesiveness.

For our case study of the IceCube team, we conducted a focus group interview with the IceCube principal investigator and with a computational resource scientist who manages the HTC for the project.

Laboratory for Molecular and Computational Genomics. LMCG investigates single molecule phenomena for the creation of new systems in the biological sciences. Except for two to three external collaborators, the LMCG team consists of 13 people who work at UW–Madison. Within this group, many disciplines are represented, including but not limited to chemistry, statistics, bioinformatics, engineering, and genetics. LMCG research focuses on a range of genomics topics, including the human genome. GLOW has contributed to the effectiveness of computational processing at LMCG and made it possible to write algorithms without considering the tradeoff between computer efficiency and program efficiency—a huge time-saver.

We conducted a focus group interview with the LMCG principal investigator, two research scientists, and one postdoctoral fellow. The LMCG team previously included an in-house computational resource expert who coordinated LMCG work with Condor, but the position was vacant at the time of the focus group.

Data Collection

The data were collected through semistructured focus group sessions conducted with scientists and research staff at IceCube and LMCG in June 2009. The interviews addressed two

topics (see focus group protocol in the appendix). The questions for the first topic—the sociotechnical factors of virtual team performance—were adapted from the Powell et al. (2004) IPO framework. The questions for the second topic—the sociotechnical characteristics of Condor technology and their impact on virtual team performance—were adapted from Condor's 5-point philosophy of flexibility (Thain et al., 2005). Follow-up probes were used only when interviewees did not cover all of the elements in their responses. Standardized probes were used to avoid potential bias.

The focus groups were conducted at IceCube and LMCG offices at UW–Madison. The focus group session at IceCube was audiotaped and transcribed. The participants in the LMCG focus group declined to be audiotaped, and thus handwritten notes were taken by the interviewer (first author) and transcribed electronically. The IceCube focus group session lasted 60 minutes; the LMCG session lasted 50 minutes.

Data Analysis

We performed a content analysis on each focus group discussion. The content analysis used a defined coding structure to capture the critical content of the data (Ryan & Bernard, 2003). The coding structure consisted of nodes and subnodes that represented categories and subcategories of the sociotechnical factors in the Powell et al. (2004) framework. This constituted an a priori framework. The transcribed discussions provided the analytic content associated with each node. Each piece of the coding scheme was defined, and comments could be coded more than once.

The NVivo qualitative software package was used to capture the coding content. Each focus group session was analyzed separately, and then the two sessions were combined into one cohesive node framework. The first round of the thematic content analysis was exhaustive; the second and third were more selective, combing prior analyses for redundancies and overdifferentiation.

Once all of the interview comments were coded in one framework, we performed within-case and between-case analyses. NVivo allowed us to systematically separate responses and make side-by-side comparisons of comments and themes. For within-case analyses, we summarized and interpreted the fundamental aspects of the comments in each group, based on the literature and our conceptual framework. For between-case analyses, we compared and contrasted the responses of the two groups, actively seeking differences and similarities in relation to virtual team performance, collaboration, and implications for HTC.

We also addressed differences between the audiotaped (IceCube) and non-audiotaped (LMCG) group sessions. Although the transcript of an audiotaped focus group session contains more detail than handwritten notes taken by a focus group interviewer, in this case the first author both collected and analyzed the data, thereby overcoming at least in part the limitations of the LMCG focus group data.

Results

Below, we report on each category of the IPO framework of virtual team performance. For each category, we provide both a within-case and a between-case analysis.

Input Factors

We identified four categories of input factors: culture, technical expertise, and training (summarized in Table 1) and design (summarized in Table 2).

Table 1
Input Factors: Culture, Technical Expertise, and Training

Categories	Subcategories	IceCube	LMCG
Culture	Differences in collaborator/team member background	X	–
	Historical/current differences in attitudes and goals	X	–
	Interdisciplinary team and field	–	X
Technical expertise	Description of Condor use	–	X
	Condor liaisons	X	X
Training		–	X

Culture was defined as differences among team members, mainly in nationality, ethnicity, and scientific discipline. IceCube participants described the differences between the cultures and mindsets of particle physicists and astronomers. The principal investigator of IceCube, a particle physicist, described the field as being composed of "explorers" willing to take large risks to build and operate instruments such as the IceCube telescope, without knowing if their assumptions are entirely correct or if the technology or infrastructure will work. The counterpoints in the IceCube project were the astronomers, scientists who historically do not join a collaboration team until the instrument is working and producing data. The LMCG participants did not cite many cultural differences, perhaps reflecting the interdisciplinary dimension of the genomics field generally and the laboratory specifically. The principal investigator of LMCG was hired under the UW–Madison Cluster Hiring Initiative, a program designed to promote multidisciplinary collaboration, and held faculty positions in the chemistry and genetics departments. Further, as a relatively young field, genomics may lack the historical markers and established cultural patterns, language, or models characteristic of older disciplines.

Technical expertise has an impact on team performance and individual satisfaction. With regard to Condor and HTC, both IceCube and LMCG cited the use of staff researchers with computational/cyberinfrastructure expertise. These researchers also served as liaisons between the Condor Project and GLOW computational scientists (other projects on GLOW had similar positions). The reasons for having on-staff computational expertise at the laboratories were that the project scientists were not cyberinfrastructure experts and the computational issues are large and complex enough for full-time researcher positions. Team members felt that the time and resources of staff scientists were better spent on technical research related to the missions of their projects.

Training within the group was cited by LMCG as necessary and dependent on project goals in both the short and long term.

Design—the structuring of virtual team interactions—was the largest category within the input factors. Five main subcategories were cited within design: collaboration technologies

(Condor and HTC; Internet resources and phone), leadership, organization of subteams, organization of work, and membership size (Table 2).

Table 2
Input Factors: Design

Categories	Subcategories	IceCube	LMCG
Collaboration technologies: Condor and HTC	Freedom in writing algorithms	–	X
	Arranging of large jobs for distributed resources	X	–
	Ease of use	X	–
	Efficiency of data delivery	X	–
	Enabling of new discoveries	X	–
	Cost-efficiency	X	–
Collaboration technologies: Internet resources and phone	DocuShare	X	–
	E-mail	X	–
	Web pages	X	–
	Wikis	X	–
	Phone calls	X	–
Leadership	Chairs of intracollaboration subteams	X	–
	Group or institution level of leadership	X	X
	Individual level of leadership	X	X
	Spokesperson for the collaboration	X	–
Organization of subteams	Committees at each institution	X	–
	Executive committee	X	–
	Large collaboration activities	X	–
Organization of work	Collaboration over time in multiple scientific domains	X	–
	Collaborations on science	X	–
	Data delivery to collaborators	X	–
	Interdependent work and tasks	X	–
	Operations and physical infrastructure work	X	–
	Need for cyberinfrastructure expertise	X	X
Membership size		X	X

Both laboratories cited Condor as a *collaboration technology*. For IceCube, Condor—and HTC technologies in general—allowed UW–Madison, the project's home institution, to efficiently deliver a large amount of complex data to collaborating scientists. IceCube team members were able to accomplish large computational jobs via distributed resources and meet norms of collaborative practice. Further, they were able to spend less money on computational resources—a substantial factor, given the size of the project. Condor enhanced collaboration by making a range of research tasks possible and supporting a large distribution of task scale.

For LMCG, a key factor was Condor's assistance with retooling complex algorithms to run on Condor. The LMCG computational resource staff person would work with Condor and the domain scientists to figure out a way to retool algorithms. Because of the availability of computational resources, revised algorithms did not have to be written in such a way as to conserve resources. As a result, the time needed to complete new algorithms was considerably

reduced, providing, as one mathematician put it, "the gift of time" to work on other problems. The difference in the two teams' responses to questions about Condor and HTC reflected the fact that computational resource support was provided by IceCube to many scientists around the world and by LMCG to only two to three distributed scientists on a limited basis; most LMCG work was done at the laboratory.

IceCube used a range of Internet and other resources for communication, coordination, and project management, including DocuShare, e-mail, web pages, wikis, and phone calls. IceCube team members cited these technologies as important mechanisms for the collaboration they coordinated with more than 250 staff and scientists on a yearly basis. LMCG did not cite many technologies for collaboration; again, this may be because most of the laboratory team members were located on the UW–Madison campus.

Leadership was cited as a design factor for virtual team performance. Leadership occurred at two levels—at the group or institution level and at the individual level. The group or institution level refers to the lead institution for the collaboration and team. The individual level refers to the specific individual leading the collaboration and team. The individual may not necessarily be from the lead institution; this would occur most frequently in larger collaborations in which multiple groups are assigned leadership responsibilities that rotate over time. IceCube leadership positions included leader of collaboration at each institution, spokesperson for collaboration (a rotating position), and chairs of intracollaboration subteams formed around specific topics or goals. Leadership played a role in effective collaboration at IceCube, and formalized leadership structures were necessary, given the size of the project and the complexities of the scientific problem set, the high-profile scientific work, and the distributed teams doing parallel and nonparallel work.

LMCG was both the lead institution for the team and the home institution of the individual leader (principal investigator of the laboratory). This organizational design of leadership is indicative of a smaller team size and a flat organization where most team members are co-located.

Subteams within the IceCube project were fairly static structures, although the subteam membership and leadership evolved over time. IceCube had committees at each institution, an executive committee for the entire collaboration, and other collaboration structures such as topic-specific subteams that examined a particular area or problem subset. In contrast, LMCG had a flat organizational structure and did not cite groupings as part of their overall team design.

The *organization of work* at IceCube reflected a long-time collaboration that supported a range of research activities across multiple domains of science. Areas of IceCube work included data analysis and simulations; data delivery; telescope construction, operation, and maintenance; production; and cyberinfrastructure-specific work. The IceCube collaboration consisted of multiple sites (other than the research institutions) that were involved in operations and physical infrastructure work, such as the operation of the telescope in the Antarctica and the construction of the ice drill and telescope at the UW–Madison-affiliated Physical Sciences Laboratory in Stoughton, Wisconsin. IceCube also cited the interdependent nature of the collaboration's tasks: one group's tasks depended on the completion and delivery of another's. For example, one of the roles of the UW–Madison group was to deliver data from the telescope operation to scientists in

other locations. At LMCG, by contrast, the organization of work was less formally organized and more fluid and person-specific, and it did not require physical infrastructure or operations. These differences may have been due to LMCG's smaller size and the nature of its research.

For Condor work specifically, both groups highlighted a strong need for cyberinfrastructure expertise within their collaboration or laboratory and considered such expertise an essential function within the organization. The individuals who served in this capacity worked on cyberinfrastructure within their respective teams while also serving as liaisons with the Condor team via the overall GLOW collaboration.

The difference in the teams' *membership size* directly and indirectly affected collaboration and the way cyberinfrastructure was used. As noted, IceCube supported a collaboration of 250 people and many institutions as well as infrastructure in Antarctica, whereas LMCG supported 13 people at UW–Madison and a few distributed collaborating scientists. Membership size had an impact on design factors such as the formalization of organizational structures and roles and the need to invest resources in coordination. Condor can outfit projects with a range of membership sizes, but it may have special relevancy for large collaborations that need to deliver copious amounts of complex data to distributed team members.

Process Factors

Process factors are grouped into two main categories: socioemotional processes and task processes. *Socioemotional processes* include relationship building, trust, and cohesion among team members (see Table 3). *Relationship building* at IceCube was affected by the project's relatively small size as a particle physics collaboration. For instance, the principal investigator of IceCube—in contrast to leaders of larger collaborations—was able to have a relationship with many members of the team. The IceCube collaboration was also characterized by an unusual level of *trust*, perhaps because of the absence of competition with other scientists (the IceCube telescope is one of a kind). Focus group participants cited particle physicists as being known for their "sharp elbows" and noted that IceCube did not reflect that lack of trust or collegiality. Neither IceCube nor LMCG addressed *cohesion* explicitly, although some of their perceptions about group cohesion may have been revealed in discussions of culture and coordination or may have been task-specific. The issue of cohesion would be an element to explore further in a future study. LMCG did not address the elements of socioemotional processes at all. This could be a limitation of the study design—that is, respondents might have been more willing to talk about socioemotional issues in one-on-one interviews or in smaller focus groups.

Table 3
Socioemotional Processes

Categories	Subcategories	IceCube	LMCG
Relationship building	Interpersonal relationships with collaborators	X	–
	Lack of competition	X	–
Trust	Collegiality	X	–
Cohesion		–	–

Task processes can include various levels of coordination, task-technology-structure fit, and communication (see Table 4). *Coordination* represents the degree of functional articulation and unity of effort between different organizational parts and the extent to which the work

activities of team members are logically consistent and coherent (Cheng, 1983). Within coordination, the *conduct of science* was relatively autonomous at IceCube. Collaborators ultimately made their own decisions as to what kind of science to do and how (for example, by splitting time among research tasks). For *Condor-specific* coordination tasks, both IceCube and LMCG relied on Condor for (a) expertise in distributed resource management to deliver the computational power needed to execute their scientific algorithms, models, and simulations; (b) support for the use of Condor generally; (c) deep engagement with the team and its projects, including the defining and redefining of research problems; and (d) technical assistance. LMCG noted that the Condor system had become progressively easier to use over the years, attributing this improvement to the continuous feedback Condor invited from user groups as well as to current research from the field of study. LMCG reported that the quality of the team's interaction with Condor was high overall and that it was very easy to gain access to more computational resources when needed. IceCube, on the other hand, mentioned that some GLOW team members displayed "opportunistic" behaviors intended to capture more computational resources by tricking the system (for example, by submitting jobs on an individual basis rather than as a group).

IceCube made the most *general comments* about coordination. Team members noted that within their collaborative structure, failure to meet deadlines or work goals had consequences—for example, not receiving data to continue one's research studies. The coordination of many IceCube activities was formalized, with task coordination formally documented to avoid repetition of work or goals. An example of this formal documentation is a memorandum of understanding stipulating roles, responsibilities, and deliverables. Yet informal collaboration also occurred at IceCube, without formal documentation. Such informal collaboration typically still required some level of coordination of subteams at one or more research institutions. IceCube noted that although a democratic work structure existed in theory, in practice issues were never decided by vote because "science is not owned by democracy." Moreover, one team member ventured that if the IceCube collaboration had to coordinate a vote to resolve an issue, "something [would be] very wrong." However, lobbying or peer pressure was acknowledged as an informal mechanism for persuading other team members to agree on an issue. Collaboration at IceCube was often planned, but sometimes it was initiated in response to interesting research issues or at semiannual "all hands" meetings. Both IceCube and LMCG mentioned engaging in long-distance collaboration and coordination of research activities, although for LMCG these were on a much smaller scale than for IceCube.

IceCube also differentiated *roles* according to task function within the group, identifying the roles of analysis coordinator, data delivery coordinator, engineer (primarily related to construction of the telescope), and scientist. LMCG did not cite roles for groups within its coordination structure.

Most comments related to *task-technology-structure fit* referred to GLOW team coordination. For example, a critical organizational design piece of GLOW is the monthly meeting of GLOW teams. At this monthly meeting, decisions about resource allocation and usage are made among the group; attendees are typically those who organize computational resources for each physical site. Individual research sites can also request resources and expertise directly from the Condor team. IceCube and LMCG both reported that their computational resource needs had always been met (although on rare occasions they had observed differences

among GLOW team members on the question of how resources should be divided). IceCube credited the GLOW intra-group structure and organization of slack resources with the absence of resource competition among the group.

The nature of scientific tasks within an HTC environment may have an impact on collaboration within teams and enhance workflow among groups within each team. For example, IceCube physicists deliver data to astronomers, who interpret it. HTC may enhance collaboration for projects like IceCube because Condor makes data delivery efficient and seamless to astronomers. For laboratories like LMCG, HTC has reduced the time scale of making significant scientific contributions by improving the throughput of laborious tasks, such as gene sequencing. Since there is a lack of history of what is "doable" in genomics, the LMCG team can harness the capability of HTC and enhance their ability to evolve and innovate. LMCG may be more apt to tolerate evolution in roles because of the lack of perceived historical viewpoints or boundaries.

Communication was not singled out as a salient factor in the group discussions in this study, although coordination technologies that facilitate communication were noted. The dimension of communication is one to follow up on in future research on virtual teams in HTC.

Table 4
Task Processes

Categories	Subcategories		IceCube	LMCG
Coordination	**Conduct of science**	Independence/autonomy in deciding science to pursue	X	–
		Resolution of disputes over data	X	–
		Splitting time	X	–
	Condor-specific	Coordination with Condor for more resources	–	X
		Facilitation of interactions	–	X
		Opportunistic scheduling of Condor jobs	X	–
		Quality of interactions	–	X
		Tricking of system for more resources	X	–
	General comments	Consequences of not meeting task goals	X	–
		Documentation of coordination	X	–
		Informal collaborations	X	–
		Intra-team coordination	X	–
		Voting or democratic structures	X	–
		Peer pressure	X	–
		Self-initiated collaborations	X	–
		Semiannual collaboration meetings	X	–
		Long-distance collaboration	X	X
	Roles	Analysis coordinator	X	–
		Data delivery coordinator	X	–
		Engineer	X	–
		Scientist	X	–
Task-technology-structure fit		Description of GLOW monthly meetings	X	–
		Division of GLOW resources across institutions	X	–
		Absence of resource competition	X	–
		Workflow facilitated by HTC	X	–
Communication			–	–

Output Factors

Output factors consisted of two categories: performance and satisfaction (Table 5). Both teams reported high *performance* in the technical output of their research and attributed a portion of that performance to HTC generally and the Condor team in particular. IceCube cited publishing research and obtaining grants as another part of their research performance metrics.

IceCube and LMCG reported *satisfaction* with both the quality of their teams and the performance of the Condor system. This satisfaction extended to the quality of interactions with Condor staff and Condor expertise in assisting the teams with computational issues and problems.

Table 5
Output Factors

Categories	Subcategories	IceCube	LMCG
Performance	Technical output	X	X
	Publication of research; grant awards	X	–
Satisfaction	Quality of team	X	X
	Performance of Condor system	X	X
	Quality of interactions with Condor staff	X	X
	Quality of Condor technical expertise	X	X

Discussion

The purpose of this exploratory study was to identify and describe some of the dimensions of scientific collaborations using HTC through the lens of a virtual team performance framework. A secondary purpose was to assess the viability of using a virtual team performance framework to study scientific collaborations using HTC. We chose to study two scientific collaborations, IceCube and LMCG, that differed across a number of characteristics such as membership size, research purpose, and range of scientific disciplines. We adapted an IPO framework developed from a synthesis of virtual team performance studies (Powell et al., 2004), using it as the basis for our analysis of virtual teams using HTC. The IPO model offers unconstrained conceptual categories, facilitating an exploratory approach to specifying the dimensions of virtual team performance.

Some of the richest dimensions of virtual team performance included culture, coordination, and design factors. *Culture* encompassed differences across disciplines, such as the historical attitudes and beliefs of particle physicists and astronomers mentioned by the IceCube project. LMCG did not mention such differences, which may reflect the nature of their field. *Design* factors were multidimensional and emphasized collaboration technologies (both HTC and communication/resource-sharing technologies) as key mechanisms for collaboration. IceCube stressed the computational power of Condor as an enabler to deliver large amounts of data to distributed team members. *Coordination* was also multidimensional, comprising categories related to the conduct of science, work with Condor, and staff roles reflecting functional purposes of the project.

The virtual team performance framework appears to be a viable tool for studying collaboration in distributed cyberinfrastructure teams. The responses of focus group participants fit within IPO framework categories, with the exception of *communication* and *cohesion*. The group discussions related to communication and cohesion focused on communication technologies, coordination, and culture rather on the specifics of communication and cohesion in the team. It is possible that the retrospective character of focus group discussion is not conducive to capturing these elements; observation of group processes over time and in different settings might be more promising. Two new subcategories emerged within the design category: *leadership* and *membership size*. However, more research is needed to expand and eventually validate key performance dimensions across various types of teams.

This work demonstrates some potential application to education reform. HTC is an extremely fault-tolerant system—that is, it performs well without significant human intervention or interaction—and school district systems such as data warehousing, accountability reporting, video encoding or re-encoding, and analytics (e.g., business intelligence in dashboards) need to be fault-tolerant and update regularly. One of the challenges facing districts is that their current infrastructure designs are outmoded, while at the same time they have a growing need to implement longitudinal data systems to collect, manage, and use student, school, and teacher data (Thorn, Meyer, & Gameron, 2007). HTC has the capacity to handle the demands for parallel processing and robust systems—core aspects of sound education infrastructure.

The sociotechnical aspects of the virtual team performance model also highlight potential applications to education reform. Districts are distributed systems: they consist of multiple roles within schools and regions, they interact with other districts within their state, and they interact with state-level systems. Thus, education, like science, needs effective team performance characterized by coordination, communication, and collaboration around computational infrastructure. For example, the IES panel review of the literature on data-driven decision making emphasized that district commitment to data quality and use requires engagement of multiple stakeholders, effective communication, clear roles and structures, and the use of teams such as a data system advisory council (Hamilton et al., 2009, pp. 39–40). The virtual team performance model of HTC could be one way to conceptualize and organize the social and technical infrastructure of data systems in school districts.

We acknowledge several limitations to this exploratory study. First, data collection was limited to two focus groups in two research collaborations. A larger sample and more investigational approaches (e.g., multiple interviews and focus groups with different team members, observation of team activities, document and/or technology analysis) would be needed to more fully define and describe the dimensions of virtual team performance and metrics in the HTC context. Second, the two teams we studied were embedded in another team—GLOW—that was not included in our analysis. It would be interesting to investigate team performance at that level as well and make linkages to the virtual team dimensions found in participating GLOW teams. Third, the current study relied on a retrospective analysis of team member viewpoints to substantiate the virtual team framework. Virtual team performance would need to be observed and recorded over a longer period of time to capture its evolution and fully describe its context. Additional possibilities for future research include expanding the within- and between-case comparisons of collaboration and virtual team performance across multiple areas of cyberinfrastructure, including HTC, high-performance computing, and cloud/grid computing.

Notwithstanding these limitations, this exploratory analysis offers an initial approximation of a thematic topography that could be leveraged to design a larger study of virtual teams using distributed cyberinfrastructure, including those in the domain of education.

References

Avolio, B. J., Kahai, S., Dumdum, R., & Sivasubramaniam, N. (2001). Virtual teams: Implications for e-leadership and team development. In M. London (Ed.), *How people evaluate others in organizations* (pp. 337–358). Mahwah, NJ: Erlbaum.

Cheng, J. (1983). Interdependence and coordination in organizations: A role-system analysis. *Academy of Management Journal, 26,* 156–162.

Chompalov, I., Genuth, J., & Shrum, W. (2002). The organization of scientific collaborations. *Research Policy, 31*(5), 749–767.

Cuevas, H. M., Fiore, S. M., Salas, E., & Bowers, C. A. (2004). Virtual teams as sociotechnical systems. In S. H. Godar & S. P. Ferris (Eds.), *Virtual and collaborative teams: Process, technologies, and practice* (pp. 1–19). Hershey, PA: Idea Group.

Driskell, J. E., Radtke, P. H., & Salas, E. (2003). Virtual teams: Effects of technological mediation on team performance. *Group Dynamics: Theory, Research, and Practice, 7*(4), 297–323.

Godar, S. H., & Ferris, S. P. (2004). *Virtual and collaborative teams: Process, technologies, and practice*. Hershey, PA: Idea Group.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., & Wayman, J. (2009). *Using student achievement data to support instructional decision making* (NCEE 2009-4067). Washington, DC: U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy, 26*(1), 1–18.

Livny, M., & Raman, R. (1998). High-throughput resource management. In I. Foster & C. Kesselman (Eds.), *The grid: Blueprint for a new computing infrastructure* (pp. 311–337). San Francisco, CA: Kaufmann.

Martins, L. L., Gilson, L. L., & Maynard, M. T. (2004). Virtual teams: What do we know and where do we go from here? *Journal of Management, 30*(6), 805–835.

Powell, A., Piccoli, G., & Ives, B. (2004). Virtual teams: A review of current literature and directions for future research. *ACM SIGMIS Database, 35*(1), 6–36.

Ryan, G. W., & Bernard, H. R. (2003). Techniques to identify themes. *Field Methods, 15*(1), 85–109.

Schiller, S. Z., & Mandviwalla, M. (2007). Virtual team research: An analysis of theory use and a framework for theory appropriation. *Small Group Research, 38*(1), 12–59.

Schmidt, K., Rasmussen, J., Brehmer, B., & Leplat, J. (1991). Cooperative work: A conceptual framework. In *Distributed decision-making: Cognitive models for cooperative work* (pp. 75–110). Chichester, England: Wiley.

Thain, D., Tannenbaum, T., & Livny, M. (2003). Condor and the grid. In F. Berman, A. Hey, & G. Fox (Eds.), *Grid computing: Making the global infrastructure a reality* (pp. 299–335). New York, NY: Wiley.

Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: The Condor experience. *Concurrency and Computation: Practice and Experience*, *17*(2–4), 323–356.

Thain, D., Tannenbaum, T., & Livny, M. (2006). How to measure a large open-source distributed system. *Concurrency and Computation: Practice and Experience*, *18*(5), 1989–2019.

Trochim, W. M. (2001). *The research methods knowledge base*. Cincinnati, OH: Atomic Dog.

Thorn, C., Meyer, R. H., & Gamoran, A. (2007). Evidence and decision-making in education systems. In P. Moss (Ed.), *106th yearbook of the National Society for the Study of Education* (Vol. 1). Malden, MA: Blackwell.

Whitely, R. (1974). Cognitive and social institutionalization of scientific specialties and research areas. In R. Whitely (Ed.), *Social processes of scientific development* (pp. 69–95). London, England: Routledge & Kegan Paul.

Appendix
Focus Group Protocol

Topic 1: Sociotechnical Factors and Virtual Team Performance

Organizational/Input Questions

1. How is work organized in your project?
    a. Is the organization participatory, bureaucratic, hierarchical, vertical, leaderless, or . . . ?
    b. Is it a formal, semiformal, or informal process?
    c. To what degree is it centralized versus decentralized?
2. How are subteams within your project organized? What resources or people do you rely on to coordinate?
3. Are there leadership roles in the project? Are there different kinds of leadership for different tasks or goals?
4. Does the membership size affect performance? Is it affected by virtual collaboration?
5. How would you characterize the culture of the team? Are there differences/similarities in culture within and/or across the various disciplines within the team?
    a. Describe the level of trust among team members. Do relationships vary across or within disciplines?
    b. Describe the cohesiveness of the team. Are some areas of the project more cohesive than others? Why?
6. What kinds of training are required either before or during the project? What does it consist of?

Process Questions

1. How is work accomplished across the virtual team? What are the structures (or support services) in place to facilitate the process of science?
2. How are the interactions among the team members structured?
    a. Face-to-face? Via technology? Phone? Or . . . ?
    b. How would you describe the quality of virtual interactions? In-person interactions? Mix of both?
    c. Does Condor facilitate or hinder the interactions of the team members? If so, how?

3. How do you coordinate research activities, goals, and timelines? Do computational resources assist in any way?

Output Questions

1. How would you characterize the quality of your scientific output? To what factors would you most attribute that characterization?

    a. How do computational resources and cyberinfrastructure affect that quality?

    b. Does the quality of scientific output vary in any way across the team?

2. How would you characterize the quality of the team performance? Does it vary across and/or within the various disciplines represented in the team?

Topic 2: Condor

1. Do you think that using Condor and interacting with Condor team members have contributed to breakthroughs in your science? If so, how? Give examples.

2. Does Condor allow you to organize your computational resources—and by extension, your team and work—in a way that best reflects your research needs and project tasks?

    a. If not, why not? If so, how? Give examples.

    b. Has using Condor changed the way you work with other scientists, both within and outside the team? How?

3. How do you set Condor policies and settings within your project?

    a. Do you feel any aspect of your work is affected by sharing computational resources with other Condor users?

    b. Does sharing Condor use facilitate or hinder any aspect of your science? Does it vary at all across different scientific disciplines?

4. Have you encountered any problems with using Condor as a computational resource?

    a. If so, how?

    b. What are the best aspects of using Condor for virtual teamwork and interdisciplinary scientific collaboration?

5. Has interacting with the Condor team affected the quality of scientific output and/or the collaboration among team members?

    a. If so, how?

    b. If not, why not? Give examples.