Collaborative Computing Frameworks for Natural Sciences Research

Emory University

Abstract— **The CCF project addresses enabling technologies to support collaborative, distributed, computer-based problem solving in the natural sciences. The goal is to evolve a virtual laboratory environment for physical and biological science disciplines, based on a cooperative computing substrate, that integrally supports a shared workspace, high performance computing, distributed data management, and AV as well as graphical scientific communication. Research foci include problem solving environments and methodologies for laboratory and instrument-based scientific disciplines, and computer science issues in heterogeneous distributed systems. New approaches are being investigated and developed for fast multiway communication, robust geographically distributed data management methodologies, high-performance computational transforms inlined within collaboration sessions, and related auxiliary issues such as active documents, security, archival storage, and experiment management and control.**

Keywords— **MDC, collaborative scientific research, virtual laboratories, multiway protocols, computational transforms, distributed data management**

I. Introduction

Research methodology in laboratory and instrument based sciences is increasingly dependent upon computation, interaction, visualization, and data storage/retrieval. However, expressly designed techniques, methods, tools, and standards are not available to facilitate such collaborative work, particularly in the natural sciences. For example, in the study of the solution state structure of anti-tumor compounds with their biological target [4] the primary techniques used are NMR spectroscopy and molecular modeling calculations. Collaboration between chemists, biochemists, biologists, physicians, and pharmacologists is essential; interactive and real time viewing, inline computation of spectral data, simulation of chemical additive effects on molecular structure, and joint analysis are activities that must be conducted within a coherent and inclusive collaboratory environment. This representative scenario of natural science computer-based collaborative research can be generalized as shown in the schematic in Figure 1.

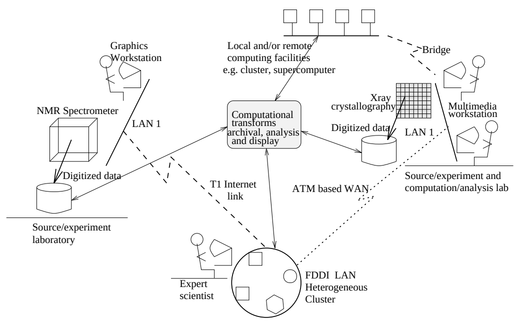

Fig. 1. Collaborative Computing Framework (CCF) Overview

The CCF project is aimed at evolving a virtual distributed computing framework that integrally incorporates computation, communication, and data management. Tools and systems that enable effective collaboration will be built upon this infrastructure, while maintaining an open and extensible interface to allow interaction with other software. This research will address fundamental issues in the three underlying technology areas, as well as in collaboration techniques, interfaces, and usability from a natural sciences perspective.

II. Research Areas

The CCF project is adopting a two-phase approach; research on the underlying technologies of network architecture and protocols, computational frameworks, and data handling comprise one phase, while user interfaces, collaboration facilities, and related aspects are being investigated in the other.

A. Communications Layer

A lightweight, responsive, and multiway communications fabric forms the cornerstone of this project. Communication performance and functionality are crucial to the type of environment that CCF will provide, especially for multimodal multiway exchange with requirements such as jitter-free, isochronous, and synchronized transmittal, low latencies, secure exchange, communications synchronization/rollback, and global operations. Existing pairwise protocols support neither the performance nor the features needed, such as security. Research on this aspect involves the formulation of a communications architecture that meets these requirements, and the implementation of a collection of *multiparty transport-level* protocols based on the notion of multiple entities forming a session or connection, with facilities for dynamic participant enrollment. Novel techniques for the provision of communication and synchronization functions, multiway delivery with selectable security, reliability, and synchrony semantics, will be developed.

B. Computational Transforms

Integral support for high-performance computational transforms, inlined within a collaborative session, is central to CCF. Several strategies are being investigated: (1) an extensible library of modular domain-specific computational transforms in distributed object form, with standardized access methods and interfaces; (2) agent-based heterogeneous computing methodologies via a combination of scripting languages and relocatable object modules; and (3) threads-based network-parallel computing methods to overcome granularity and latency obstacles to achieve high-performance.

C. Data Management

A distributed data management framework will complement the communication and computation infrastructures in CCF. Collaboration sessions involve multiple types of data and objects, including instrument data, graphical plots and images, active documents, and session archival records. Distributed storage techniques, incorporating declustering for parallelism and concurrency control for orderly access, will be investigated. In addition, methods for indexing and rapid positioning/access, compression and encryption streams during storage and retrieval, and algorithms for efficient format conversion will be developed.

III. Collaboration Infrastructure

Research in the core technologies described above will complement and buttress the actual collaboration tools that will be developed. Novel concepts in collaboratories for the natural sciences will be investigated empirically, by tightly coupling software toolkit design and development with actual use in chemistry and biochemistry laboratory settings. A primary goal of the CCF collaboratory will be to provide an inclusive environment representing the scientific apparatus, interfaces, computational transforms, and expertise of researchers in a virtual workspace. Indexable multimodal bulletin boards, and integrated data access and computational tools will be readily available.

Science-specific collaboratory requirements will be addressed, including accurate depiction methods for mathematical and physical models, inlined graphing tools with distributed zoom and focus features, active hyper-links between numeric data, graphical/image representations thereof, and annotations, and archival lookup mechanisms for related and prior results, citations, and reference materials. Research at both the conceptual and implementation levels is required to determine methodologies and techniques that facilitate collaborative computer-based work, such as drag-and-drop interfaces to computational processes, three-dimensional distributed visualization with multiple independent and linked perspectives, multi-input interactive steering and manipulation of computations and graphical displays, etc. Some of these issues have been studied in other contexts [2]; the CCF project will extend these, investigate other unexplored areas, and most importantly, focus on an integrated but modularly extensible suite of methods, features, and tools to enhance collaborative computing. The emphasis will be on the use of the technology rather than the technology itself; by continually iterating the design and CCF toolkits based upon use and experience for cooperative scientific work, it is hoped that a both quantitatively and qualitatively better paradigm for collaborative computing in the natural sciences will emerge.

IV. Related Areas and Projects

The CCF project addresses research issues in several areas. Within the CCF kernel, work on the communications fabric will make an impact on multicasting, session level protocols, atomic and causally consistent message passing schemes, and auxiliary issues such as encryption and secure data transmission [1]. Other projects investigating these topics include the Mbone tools, ISIS/Horus[3], and the RMP[8] effort. In the portion dealing with high-performance computation, new techniques for effective network-based parallel computing and heterogeneous computing with threads will evolve. These advances will extend the models and results of projects such as PVM, Ariadne, Converse, and Chant [6], [7]. In terms of distributed data management, research will be conducted in declustering techniques for distributed heterogeneous file systems, concurrency control issues and multilevel consistency semantics, indexing methods, and heterogeneous data storage and retrieval [5].

CCF research on the collaboration system itself will provide insight into the effectiveness as well as the tangible and intangible benefits of scientific collaboratories. This aspect is related to CSCW systems and conferencing systems but has a number of significant distinguishing features – domain specific focus, inlined computation, open extensible interfaces, security, and a number of characteristics deriving from a shared workspace in a literal sense. The overall combination of these goals, research topics and emphases, and facilities is unique to CCF although various subsets thereof are currently being addressed in a few instances, e.g. ICE, and the EMSL effort [9], [10]. CCF will be a fully comprehensive and inclusive collaboratory.

V. Contact Information

Further information on the CCF project may be found at `http://www.mathcs.emory.edu/ccf/` or by sending email to `ccf@mathcs.emory.edu`, or by contacting Vaidy Sunderam, Dept. of Math and Computer Science, Emory University, Atlanta, GA 30322; Phone: (404) 727-5926, Fax: (404) 727-5611.

References

[1] Injong Rhee, Shun-Yan Chung, and Vaidy Sunderam, *Design of a Communication Substrate for Collaborative Computing*, http://www.mathcs.emory.edu/ccf/cctl
[2] Robert Simon, Robert Sclabassi, Taieb Znati, *Communication Control in Computer Supported Cooperative Work Systems*, Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, NC, Oct 1994, pp. 311 –321.
[3] Kenneth P. Birman and Thomas A. Joseph, *Reliable Communication in the Presence of Failures*, ACM Transactions on Computer Systems, Vol. 5, No. 1, Feb 1987, pp. 47 – 76.
[4] Marian Iwamoto, Srinivasan Mukundan, Jr. and Luigi G. Marzilli, J. Am. Chemical Soc., Vol. 116, 1994, pp. 6238–6244.
[5] Ravi Jain, John Werth and J. C. Browne eds *Proceedings – Workshop on I/O in Parallel and Distributed Systems (IOPADS) at IPPS'95*, Kluwer Press, 1996.
[6] L. Smarr, C. Cattlett, *Metacomputing*, *Communications of the ACM*, Vol. 35, No. 6, pp. 42–53, June 1994.
[7] E. Mascarenhas and V. Rego, *Parallel Programming with Ariadne* http://www.cs.purdue.edu/research/PaCS/ariadne.html
[8] B. Whetten, T. Montgomery and S. Kaplan, *A High Performance Totally Ordered Multicast Protocol*, in *Proc. Dagstuhl Distributed Systems Workshop*, Sept. 1994, ftp://research.ivv.nasa.gov/pub/doc/RMP/RMP_dagstuhl.ps.
[9] http://www.sandia.gov/ICE/Main.html
[10] http://www.emsl.pnl.gov:2080/docs/collab/CollabHome.html