

Project Title: **IQ-ECho: Interactive Quality of Service Across Heterogeneous Hardware**

Comment:

Project Type: **SciDAC-MICS**

PI: **Karsten Schwan, Georgia Tech**

Co-PIs: **Greg Eisenhauer, Matt Wolf (GT) and via additional NSF funding, Nagi Rao (ORNL)**

1. Project Description

Executive Summary

This research addresses the problems created by the increased distribution and heterogeneity of the hardware/software platforms on which teams of researchers will conduct future scientific collaborations. The assumptions that drive this work are that on the one hand, it is difficult to constrain the future hardware and software systems such teams will use, especially given the diverse scientific tasks they will be carrying out, and that on the other hand, effective real-time collaboration demands that team members be able to interact with each other and with remote resources as if they were co-located. Two problems arise:

1. Heterogeneous underlying operating systems and software platforms create interoperability problems for the scientific simulations, their input data, and their outputs via which and their team members collaborate.
2. Heterogeneous underlying hardware and networks make it difficult to guarantee the timely transport of data and execution of software required for seamless remote collaboration.

Solution Approach

While we are pursuing solutions to Problem 1 in other research, this funded effort focuses on innovative solutions to Problem 2, that is, the provision of Quality of Service support in middleware. Specifically, we will provide middleware-level functionality that permits end users and applications to move data with the timeliness they require, despite dynamic changes in underlying platform resources. To meet this goal, our specific focus will be the timely visualization of large-scale data on heterogeneous clients and across heterogeneous networks. For instance, it should be feasible for an end user operating via a DSL-connected home PC to collaborate with another end user operating on a lab-resident high end graphics machine directly connected to a cluster-based data server. Making this work requires more than just network-level QoS support; end users must both understand and explicitly manage their applications' communications, or the middleware they are using must provide such support.

Specific Research Goals

The research we propose differs from previous work in our pursuit of an 'open systems' approach for realizing QoS functionality for future high performance computing platforms. Specifically,

our goal is to substantially enhance the support that middleware provides for runtime resource management. Toward this end, given that end users provide application-specific descriptions of their needs and the resource tradeoffs that are admissible, the proposed IQ-Echo middleware will offer the Quality of Service interfaces and mechanisms -- hence the 'Q' in 'IQ-Echo' -- via which (1) such needs and tradeoffs are made known to underlying resource management (at the OS or network levels) and (2) information about platform resources is made known to applications. In contrast to previous work on QoS in middleware, such as BBN's Quo system, our approach takes advantage of the 'open' nature of many of the current platforms used in high performance computing, such as Linux-based Beowulf clusters or the programmable network interfaces used in the ASAN project funded by DOE at Georgia Tech (Profs. Yalamanchili and Schwan, PIs). Specifically, rather than asking end users to state their QoS needs with attribute-value pairs as in Quo, for example, or with the payoff or utility functions used in our own prior work, we propose to simply let an application 'program' the underlying 'open' platform. Such programming has two goals: (1) to help the network or operating system manage resources in a manner meaningful to applications, and (2) to have system levels export to applications resource information that is comprehensible and meaningful.

Toward these ends, the IQ-Echo middleware will support client-specific (or more generally, receiver-specific) data transport, sampling, and transformation, by permitting applications to create logical communication channels with associated filter functions, using the *derived channel* concept developed in our earlier work. These functions may be dynamically defined and run by the data receivers, but receivers can also 'push' them into the address spaces of data senders and/or even into underlying operating system kernels or networks. Thus, filter functions *extend* collaborators' functionality, in a fashion meaningful to their applications.

2. Milestones and Deliverables

There are two major tasks associated with this project, the development of the IQ-Echo middleware and the creation and evaluation of a representative scientific application that demonstrates the utility of the middleware. In the first year of this grant, we developed the basic middleware methods and techniques proposed in our research and have released an initial version of IQ-Echo. It contains many facilities for customizing and adapting high-performance data flows and is briefly described in the next section. Preliminary evaluations of both used synthetic applications and emulated network testbeds. We will continue to enhance and release new versions of IQ-Echo for the duration of the project. In the second year of this grant, we have focused on applying our technologies to realistic applications on representative testbeds. Specifically, while our earlier results demonstrate the feasibility of our approach on emulated networks and machines (using Emulab) and synthetic applications, this year's work has produced (1) a realistic application program, termed SmartPointer, described in a Supercomputing conference paper, and (2) we have now applied the adaptive communication/middleware technologies to this application across a networking testbed being constructed jointly with Oakridge National Labs. Deployment of the IQ-Echo infrastructure and a sample application were the significant milestones called for in the first two years of our project.

3. Detailed Progress to Date

IQ-ECho

To attain high performance in the real-time exchange of data across collaborating machines and end users, this project is developing and evaluating middleware methods and techniques for coordinating application-level with network transport-level adaptations of data communication. Complementing previous work on TCP-friendly communication and on adaptive transport protocols, our approach is to use middleware to strongly coordinate application-level with transport-level changes in communication behavior, so as to best meet application needs without violating fairness in network resource usage. The approach is embodied in the IQ-ECho middleware, which implements the distribution of scientific data to remote collaborators. Using IQ-ECho, application-level adaptations like selective data down-sampling are triggered by transport-level information provided by the instrumented IQ-RUDP protocol underlying IQ-ECho's communications. The application- to network-layer exchange of information necessary for such coordinated adaptations is implemented with quality attributes, which provide a lightweight way for an application to provide quality of service information and to describe its adaptation to the transport layer, and for IQ-RUDP to share network status information with an application.

In addition to triggering application-level adaptations and reacting to certain changes in network state, IQ-RUDP also re-adapts its own communication behavior after an application adaptation has been performed, in part to remain fair to other network flows. Such transport-level reactions can be performed at higher rates and with smaller overheads than possible at application level.

SmartPointer

SmartPointer explores a common problem with data portals, which is the heterogeneous nature of the machines and platforms on which they must operate. Specifically, we ask how end users may meaningfully interact across highly heterogeneous machines and network connections. This issue has arisen, for example, for the Terascale Supernova Initiative in which end users from the national labs operating across Gigabit links must interact in real-time with collaborators on PCs operating across standard Internet links. Taken to an extreme, we ask how end users operating with very large data sets displayed only on high end systems like Immersadesks driven by large SMP machines can usefully interact via low end engines like laptops and PDAs.

The specific interaction paradigm explored is one in which a low end device essentially acts as a 'smart pointer' into the large data space existing in the distributed system and/or displayed on devices like a CAVE or Immersadesk. That is, while a CAVE may be used to render to an end user the entire data space (or large portions thereof), the handheld device cannot hope to render any meaningful subset of this data in the same fashion. Instead, its role is to provide alternative views of specific data elements, to activate analyses meaningful for these elements and display analysis results, and to track and interact with collaborators. When located in the same room as the immersive device, the smart pointer may present certain details or complementary information about the data displayed in its entirety. When operating in a distributed system, a smart pointer may be viewed as presenting similar information about the large, distributed, and shared data space that defines end users' distributed collaboration. In both cases, a smart pointer permits an end user to interact with the large, shared data space as per his/her current interests and needs, where most such interactions entail the activation of services that transform, analyze, filter, and sample shared data.

The idea is to support distributed collaboration and steering of computational science, with a strong emphasis on providing the personalized data portals that the smart pointers represent to every collaborator. The collaborative environment is provided already by the AccessGrid toolkit. By adding the IQ-ECho event channel infrastructure, the middleware that underlies the current implementation, as a parallel data transport to the Access Grid, we can utilize IQ-ECho's high performance communications infrastructure for heterogeneous binary transport of the scientific data. In addition, IQ-ECho's facilities for runtime, source-based filtering of data help to optimize performance of the clients as well as enhancing their customizability.

How IQ-ECho Supports SmartPointer

For the SmartPointer application, we require a communications infrastructure that can be flexible, adaptive, and yet support high performance. Traditional HPC-style communications systems like MPI offer the required high performance, but rely on the assumption that communicating parties have a priori agreements on membership lists and on the basic contents of the messages being exchanged. For the sort of system we have described, however, both data types and subscription lists of communicating parties must be flexible. This need for flexibility has led some designers to adopt techniques like Java's RMI or meta-data representations such as XML+SOAP. These methods have high costs that interfere with total performance, because data marshalling becomes a key issue.

The IQ-ECho middleware addresses these concerns in several different ways. In particular, it provides the following:

- Information flows are represented as **event streams**, using the publish/subscribe model.
- There is no centralization; channels are represented by distributed data structures.
- Connections are managed so as to preserve transparency of local versus remote receivers, with implementations for underlying peer-to-peer communications over TCP, UDP, Wireless, multicast, and others.
- Efficient, fully typed binary data transmission based on dynamically defined event formats (PBIO).
- Dynamic extension of existing formats and discovery/operation on format contents is enabled through run-time dynamic code generation of subscriber-specified filters.
- Interoperation with CORBA and Java (JECho) via IDL and XML is enabled through conversion of data at the far endpoint.

Further information is available at <http://www.cc.gatech.edu/systems/projects/IQECho>.

Publications

Zhongtang Cai, Qi He, Greg Eisenhauer, Karsten Schwan, Matthew Wolf, "IQ-Services: Network-Aware Middleware for Interactive Large-Data Applications", Submitted to Supercomputing 2003, ACM/IEEE, May 2003

4. Future Accomplishments and Milestones

SmartPointer Framework for Remote Collaboration

Future work with the SmartPointer framework will focus on building a portable and flexible infrastructure for scientific visualization. This work has multiple components, including

graphics work, middleware support, and network/protocol optimization. The ideal is to provide an initial implementation of a visualization server infrastructure, which would allow for clients of widely ranging capabilities to access high-quality visualization data over wide-area links.

An integral component of that visualization service will be integrating with existing standards such as the CCA effort (funded by SciDAC) and OGSA (funded by NSF and SciDAC). Most of the currently planned work focuses on a CCA-based interface to the IQ-Echo visualization pipeline, using the MxN interface. This work is being pursued jointly with James Kohl at ORNL.

Additionally, work will be proceeding to enhance the scientific application layer, adding support for more generic computational chemistry codes. This work will require adding additional data layers (corresponding to additional data needed for ab-initio codes), additional annotations packages, and visualization interfaces that can exploit all of this in appropriate user-directed fashions.

Network Testbed and Evaluation

Our future work with the evaluation of IQ-Echo services for remote collaboration will (1) compare service-level adaptations that utilize different network-level techniques for assessing current network bandwidth, (2) use overlay networks to combine the lightweight data filtering and downsampling methods used in this paper with heavier-weight methods for data transformation and summarization executed by additional machines interposed into the path between data providers and consumers, and (3) consider the dynamic deployment of lightweight IQ-Echo services to dynamically utilize alternative network and machine paths from data providers to consumers. Such work will focus on high end links using the large data volumes produced by applications like the DOE Supernova Initiative. The intent is to use actual Supernova data across the 10GB link and other links connecting GT with ORNL and other DOE sites.

Immediate milestones

Network testbed and evaluation

- Establishing connectivity across machines that use 10GB link
- Install IQ-Echo middleware
- Impose large-data traffic for network evaluation and measurement
- Utilize existing measurement methods to assess available network bandwidth
- Use IQ-Echo services with large-data across high end network link

Remote collaboration

- Deployment of initial implementation of MxN interface with IQ-Echo internals
- Additional filtering code to do protocol adaptation based on IQ-Echo and IQ-RUDP.
- Evolve the internal data representation to allow for later inclusion of ab-initio information

5. Project Management

This project is managed by Prof. Schwan, jointly with Drs. Eisenhauer and Wolf. Dr. Eisenhauer is developing key IQ-Echo technologies, including the integration of modern network measurement technologies into IQ-Echo middleware. Jointly with Dr. Wolf, he is also involved in integrating IQ-Echo technologies into common standards (e.g., CCA standards). Dr. Wolf's principal role is to manage our relationship with ORNL, and to apply IQ-Echo technologies to realistic DOE applications. This role has recently been substantially strengthened by his appointment as a researcher at ORNL, thereby formalizing his role as a liaison between the GT research team and the developers and science and CS researchers at ORNL. 6 months of Dr. Wolf's annual salary are covered by this appointment.