Mobile Interfaces to Computational, Data, and Service Grid Systems

Mauro Migliardi[a]
om@dist.unige.it

Muthucumaru Maheswaran, Balasubramaniam Maniymaran
Paul Card, and Farag Azzedin[b]
{*mahes, bmmaran, paul, farag*}*@win.trlabs.ca*

[a]University of Genoa, Genoa, Italy
[b]University of Manitoba and TRLabs, Winnipeg, Canada

This article briefly describes the issues related to providing mobile access to computational, data, and service Grids. Then it describes our preliminary efforts to enhance computational and service Grids to handle mobile access. In particular, we focus on how the HARNESS mobile extensions project approaches the problem of enhancing Grid computers with mobile access and we describe how the InviNet project provides access to service Grids with consistent quality of service.

I. Introduction

Grid systems [1] are very large-scale, generalized network computing systems that can scale to Internet size environments with resources distributed across multiple organizations and administrative domains. Using the design objectives and target applications of the Grid systems, they can be categorized into (a) computational Grids, (b) data Grids, and (c) service Grids. *Computational Grids* have higher aggregate computational capacity available for single applications than the capacity of any constituent machine in the system. *Data Grids* provide an infrastructure for synthesizing new information from data repositories such as digital libraries or data warehouses that are distributed in a wide area network. *Service Grids* provide services that are not provided by any single machine.

With the emergence of Grids as a suitable hosting technology for applications in various areas including data- and compute-intensive computing, online gaming, and data storage and archiving, it is becoming important to provide mobile access interfaces to Grids. For example, in a tele-health application, the patient data including Computed Tomography images can be stored in data Grids. In emergency situations, health care personal may want to access this information in different forms from the field. Due to consistency and privacy issues and resource limitations, replication of the content may be restricted. Therefore, on demand computations that are resource intensive may be performed to create the data that is most appropriate for the mobile device. Hence, it is essential that the data is manipulated on high-performance machines that are part of the data Grid forming the back-end instead of the mobile machines that are the front-end.

Providing mobile access to distributed computing resources and in particular web oriented resources is a well examined topic [2, 7, 8]. However, supporting mobile interfaces in the Grid context provides new challenges and opportunities. Following are some of the major research issues in this context:

- One of the main issues handled in Grid computing systems is resource heterogeneity. Introduction of mobile devices requires an extension of the heterogeneity model to incorporate them.
- Disconnections and the resulting network partitions should be handled efficiently. Most projects propose solutions to disconnections by treating them as failures or willing permanent changes. However, with mobile or wearable devices, these events need to modeled in a general way.
- A mapping scheme should decide how different components of an application should be mapped, i.e., which components should be assigned to the back-end and which to the front-end (mobile device). These decisions should consider various issues including power consumption, locality of access, resource requirements, throughput requirements.
- The components of an application mapped onto the back-end should be located such that load balancing and *quality of service* (QoS) considerations are met.

II. Our Approaches

In this section, we report on our preliminary efforts to enhance Grids with mobile access as presented at the *Advanced Topic Workshop on Middleware for Mobile Computing*.

In the first project that extends the HARNESS metacomputing system [6, 5], we examine how wearable computers can be used with a computational Grid. With wearable computers, the user moves with the machine or terminal. Therefore, the physical identity of the end terminal and its properties do not change as the user moves. This is often referred to as *terminal mobility* and QoS issues are caused in this instance due to variations in connection quality.

However, in general, the user can move among different end terminal devices. In this situation, the end terminal identity and properties change as the user hops between terminals (the terminals may be wearable, mobile, or stationary). This form of mobility is called *user mobility* [2]. In the second project, the InviNet, we propose an architecture for handling this aspect in the context of Grid systems.

II.A. HARNESS Mobile Extensions

To address the issues related to enhancing computational Grids with mobile accessibility, the HARNESS project

[6, 5] adopts a model based on mobile activities that migrate between wearable computers by means of active transactions. Coupled with the ability of the HARNESS system to manage multiple users and reconfigure the services provided by the virtual machines, this approach allows us to seamlessly connect the model based on mobile activities with a Grid computing system. In our project, we envision four different modes of operation: (a) single wearable computer; (b) peer-to-peer network of wearable computers; (c) a wearable computer and Grid system connected together; (d) a peer-to-peer network of wearable computers connected to a Grid system.

In the first mode, the capabilities of a wearable computer are augmented by the additional services provided by the land-based stations. These services include: repositories of modules; storage for data and activities; remote, persistent message boards; and activity arenas. The second mode of operation explores the concept of communities of users cooperating by means of wearable computers. These communities engage in collaborative activities of resource and application planning without needing a stable interconnection topology. On the contrary, the components of a community connect to each other only occasionally to exchange activities and results. The users in these communities also exploit the services available in the single user mode to enhance their capabilities, e.g., storing messages on persistent boards to setup a rendezvous or uploading currently running activities to free resources needed by cooperating applications.

In the third mode, for applications requiring computational power unsupportable by wearable computers, the wearable computers use back-end Grid systems as computational engines. The fourth mode combines the cooperative activities of mode two with the computationally intensive services provided by the back-end Grid systems in mode three. In all modes, the wearable computer users benefit from the functionality to engage in collaborative computing with the community of wearable computing users as well as the ability to tap into the large computational power provided by the back-end Grid system.

II.B. InviNet

The InviNet (Invisible Network) [3] is an ongoing project for providing pervasive access to services, content, and users that uses Grid computing as a back-end technology. The basic entity managed by the InviNet is a service (i.e., users and content are represented by the corresponding services). The InviNet is organized in a two-stage architecture. The first stage (front-end) called the InviNet kernel (INK) primarily provides service location and access facilities that handle mobility. The second stage (back-end) called the MetaGrid [4] is based on Grid computing and provides a resource orchestration framework for the services. The INK is being built to include the following features (a) unified human friendly namespace, (b) secure dissemination and discovery, and (c) user and terminal mobility. The human friendly names are based on the domain names for maximum backward compatibility. These names are constructed such that they provide “hints” on the expected mobility patterns of the named entities. The names are used to locate the entity at connection creation. Once the connection is created, the connection maintenance algorithm is responsible for maintaining the connection despite mobility. The MetaGrid back-end performs adaptive resource allocations so that services can continue to operate at acceptable levels as the demands fluctuate. For example, services that are overloaded are replicated and thus provided more resources. The replica created by the MetaGrid is used by the INK when a request for that service arrives. Similarly, replicas are deleted if services are found to be underloaded. The decoupled nature of InviNet enables the resource allocations to expand and contract as necessary.

References

[1] FOSTER, I., AND KESSELMAN, C. *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, San Fransisco, CA, 1999.

[2] LANDFELDT, B., LARSSON, T., ISMAILOV, Y., AND SENEVIRATNE, A. SLM, A Framework for Session Layer Mobility Management. In *10th IEEE International Conference on Computer Communications and Networks* (1999).

[3] MAHESWARAN, M., MANIYMARAN, B., CARD, P., AND AZZEDIN, F. Invisible network: Concepts and architecture. In *2002 International Workshop on Invisible Computing* (May 2002).

[4] MAHESWARAN, M., MANIYMARAN, B., CARD, P., AND AZZEDIN, F. Metagrid: A scalable framework for wide-area service deployment and management. In *16th International Symposium on High Performance Computing Systems and Applications (HPCS 2002)* (June 2002).

[5] MIGLIARDI, M., AND SUNDERAM, V. Plug-ins, layered services and behavioral objects: Application programming styles in the HARNESS metacomputing system. *Future Generation Computer Systems 17*, 6 (Apr. 2001), 795–811.

[6] MIGLIARDI, M., SUNDERAM, V., GEIST, A., AND DONGARRA, J. Dynamic reconfiguration and virtual machine management in the harness metacomputing system. In *International Symposium on Computing in Object-Oriented Parallel Environment (ISCOPE ’98)* (1998), pp. 127–134.

[7] SNOEREN, A. C., ANDERSEN, D. G., AND BALAKRISHNAN, H. Fine-grained failover using connection migraton. In *3rd USENIX Symposium on Internet Technologies and Systems* (Mar. 2001).

[8] SNOEREN, A. C., BALAKRISHNAN, H., AND KAASHOEK, M. F. Reconsidering Internet mobility. In *8th Workshop on Hot Topics in Operating Systems (HotOS-VIII)* (2001).