Grid Services for MPI

Camille Coti[2], Ala Rezmerita[1], Thomas Herault[1], and Franck Cappello[2]

[1] Univ Paris Sud; LRI; INRIA Futurs; F-91405 Orsay France
[2] INRIA Futurs; F-91405 Orsay France
coti@lri.fr, rezmerit@lri.fr, herault@lri.fr, fci@lri.fr

1 Introduction

Institutional grids consist of the aggregation of clusters belonging to different administrative domains that build a single parallel machine. In order to run a MPI application over an institutional grid, one has to address many problems. One of the first problems to solve is the connectivity of nodes not belonging to the same administrative domain.

To protect the network from unauthorized access many sites use firewalls. On some sites firewalls are configured to allow outbound connections and to block inbound connections, often with the exception of a few well-known ports (*e.g.*, SSH). On some other sites there is a strict separation between the internal and external networks and only a front-end machine is accessible. Theses connectivity constraints limit the execution of parallel application between multiple sites.

The connectivity problems can sometimes be solved when only one site uses a firewall, since all required connections are initiated from the protected site. However this solution requires modifications to applications or communication libraries. Also, if all sites are using firewalls this approach can no longer be applied. Another solution is to configure the firewalls so that a port range is open and adapt the applications to use only theses ports. However this solution is a threat to the site security. Sometimes the only possibility for the compute nodes to communicate with the outside world is to use the front-end machine as a bridge. In addition to causing connection setup problems the use of Network Address Translation (NAT) devices complicates machine identification. The private addresses used in a NAT site are not globally unique, which causes difficulties in creating a unique identifier for every machine.

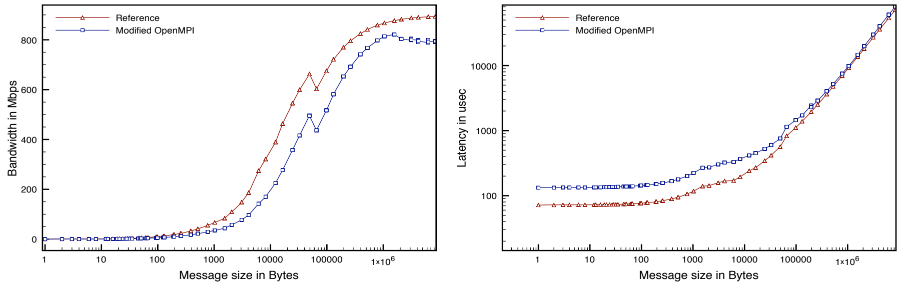
In this work, we propose a set of Grid or Web Services that provide a new level of communication for establishing connectivity of MPI applications over an experimental grid.

2 Results

We define a distributed framework to allow the grid infrastructure to provide services to the applications. In this paper we detail a brokering service that provides the computing nodes a way to communicate with each other. Other services can be implemented in our framework, such as monitoring service, spawning service and distributed storage service. The brokering service establishes a connection between nodes that would not be able to communicate with each other otherwise because a NAT and/or a firewall

F. Cappello et al. (Eds.): EuroPVM/MPI 2007, LNCS 4757, pp. 393–394, 2007.

are standing between them. When the MPI library needs to establish a communication between two nodes that don't belong to the same cluster, it invokes the brokering service that finds the best method to establish this connection (NAT and/or firewall bypassing) and returns the appropriate connection information to the initiator of the connection. Some techniques have been presented in [2]. We implemented the service using the light-weight web-services engine gSOAP[3] and interfaced it with OpenMPI[1].



(a) Bandwidth measured by NetPIPE (b) Latency measured by NetPIPE

Fig. 1. Communication performance

Figure 1 shows the impact of the framework on communication performances, measured using the NetPipe test. The nodes are interconnected by a proxy, which adds a hop between them. We can see the impact of this additional hop on bandwidth on figure 1(a) and on latency on figure 1(b). Since the service is invoked only to establish the communication, it has no effect on the performances of the communications themselves. Therefore, the other techniques that establish a direct connection between two nodes (reverse connection, traversing TCP and TCP hole punching) give the same performances as a direct connection without a firewall. The additional cost induced by the establishment of the connection is 10 ms.

References

1. Gabriel, E., Fagg, G.E., Bosilca, G., Angskun, T., Dongarra, J.J., Squyres, J.M., Sahay, V., Kambadur, P., Barrett, B., Lumsdaine, A., Castain, R.H., Daniel, D.J., Graham, R.L., Woodall, T.S.: Open MPI: Goals, concept, and design of a next generation MPI implementation. In: Proceedings, 11th European PVM/MPI Users' Group Meeting, Budapest, Hungary, pp. 97–104 (September 2004)
2. Rezmerita, A., Morlier, T., Néri, V., Cappello, F.: Private virtual cluster: Infrastructure and protocol for instant grids. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 393–404. Springer, Heidelberg (2006)
3. van Engelen, R.: Pushing the SOAP envelope with web services for scientific computing. In: proceedings of the International Conference on Web Services (ICWS), pp. 346–352 (2003)