A Scavenger Grid for Intranet Indexing

Ndapandula Nakashole and Hussein Suleman

Department of Computer Science, University of Cape Town
Private Bag, Rondebosch, 7701, South Africa
{nnakasho,hussein}@cs.uct.ac.za

Abstract. Digital library services, such as searching and browsing, are increasingly needed in more restricted environments than the public Web. This paper proposes a scavenger Grid of idle desktop workstations to support computationally-intensive indexing services. A prototype software system was developed using commodity Grid middleware and information retrieval tools. This system demonstrated that the overhead incurred by Grid middleware is manageable and that performance gains are significant.

1 Introduction

Companies and educational institutions produce large numbers of electronic documents on a daily basis. Some documents, however, embody the intellectual property of the organization and should not be accessible to the outside world. These documents need to be indexed for rapid retrieval within the organization, but cannot be made accessible to the outside world. Thus, Web search engines may not be appropriate and third party indexing service providers can be used only if there are strong trust relationships and information is guaranteed to be secure. Typically, such a trust relationship already exists within the organization. The natural solution then is to use an in-house information retrieval appliance or system to index documents and enable retrieval.

2 Scavenger Grid-based Indexing

The premise of this work is the use of idle workstations in a scavenger Grid environment to handle medium to large-scale search engine indexing, using standard tools and techniques. A scavenger Grid or "cycle-stealing" Grid is a distributed computing environment made up of underutilized computing resources in the form of desktop workstations, and in some cases even servers, that are present in most organizations. The scavenger Grid proposed in this study is the use of a combination of dedicated processing nodes (static nodes) and non-dedicated processing nodes (dynamic nodes) for indexing. A prototype search engine designed for a scavenger Grid environment was developed as a proof of concept. The Grid used in this research makes use of a local scheduler and "cycle stealing" technology, Condor [3, 2], and a storage middleware solution, the Storage Resource

Broker (SRB) [1]. SRB is client-server middleware that virtualizes data space by providing a unified view of multiple heterogeneous storage resources over a network.

3 Evaluation

Tests were carried out to assess performance of Grid-based indexing in an actual scavenger Grid. In the first scenario, Fig. 1 part A, the number of static nodes was kept constant at 4 while the number of dynamic nodes was varied. In the second scenario, Fig. 1 part B, in addition to the dynamic nodes, the number of static nodes were varied also. The results in Fig. 1 show that the task of indexing can be performed by idle workstations without relying on a large fixed set of dedicated resources.

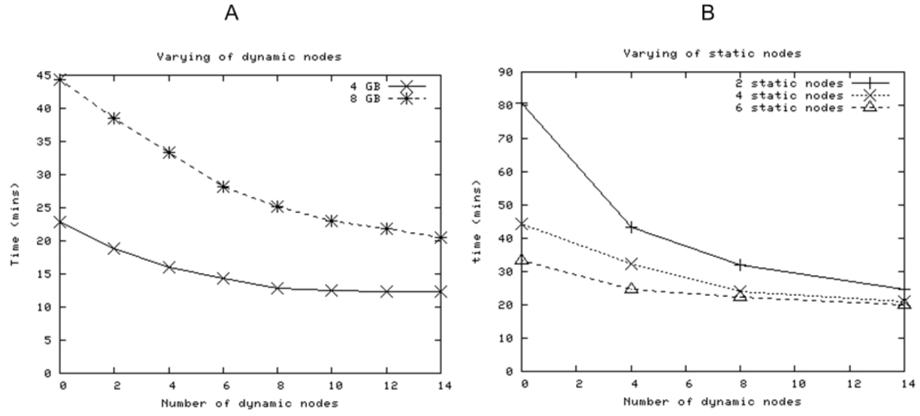

Fig. 1. Indexing performance of different numbers of dynamic nodes (part A) and Indexing performance of different numbers of static nodes (part B)

4 Conclcusions

This study has looked at the use of resources already at an organization's disposal in the form of a scavenger Grid to provide cost-effective scalability within an intranet while retaining control over who can access the data. The results show that as additional resources become available, there are performance gains which make up for the inevitable grid middleware overhead.

References

1. C. K. Baru, R. W. Moore, A. Rajasekar, and M. Wan: The SDSC storage resource broker. Proceedings of the conference of the Centre for Advanced Studies on Collaborative Research.1998.
2. M. Litzkow and M. Livny: Experience with the condor distributed batch system. In Proceedings of the IEEE Workshop on Experimental Distributed Systems,1990.
3. Condor: High Throughput Computing. 2007. http://www.cs.wisc.edu/condor/.