DDSS: A Low-Overhead Distributed Data Sharing Substrate for Cluster-Based Data-Centers over Modern Interconnects

Karthikeyan Vaidyanathan, Sundeep Narravula, and Dhabaleswar K. Panda

Department of Computer Science and Engineering
The Ohio State University
{vaidyana, narravul, panda}@cse.ohio-state.edu

Abstract. Information-sharing is a key aspect of distributed applications such as database servers and web servers. Information-sharing also assists services such as caching, reconfiguration, etc. In the past, information-sharing has been implemented using ad-hoc messaging protocols which often incur high overheads and are not very scalable. This paper presents a new design for a scalable and a low-overhead *Distributed Data Sharing Substrate* (DDSS). DDSS is designed to support efficient data management and coherence models by leveraging the features of modern interconnects. It is implemented over the OpenFabrics interface and portable across multiple interconnects including iWARP-capable networks in LAN/WAN environments. Experimental evaluations with networks like InfiniBand and iWARP-capable Ammasso through data-center services show an order of magnitude performance improvement and the load resilient nature of the substrate. Application-level evaluations with Distributed STORM achieves close to 19% performance improvement over traditional implementation, while evaluations with check-pointing application suggest that DDSS is highly scalable.

1 Introduction

Distributed applications in the fields of nuclear research, biomedical informatics, satellite weather image analysis etc., are increasingly getting deployed in cluster environments due to their high computing demands. Advances in technology have facilitated storing and sharing of the large datasets that these applications generate, typically through a web interface forming web data-centers [1]. A web data-center environment (Figure 1) comprises of multiple tiers; the first tier consists of front-end servers such as the proxy servers that provide services like web messaging, caching, load balancing, etc. to clients; the middle tier comprises of application servers that handle transaction processing and implement business logic, while the back-end tier consists of database servers that hold a persistent state of the databases and other data repositories. In order to efficiently host these distributed applications, current data-centers also need scalable support for intelligent services like dynamic caching of documents, resource management, load-balancing, etc. Apart from communication and synchronization, these applications and services exchange some key information at multiple sites (e.g, timestamps of cached copies, coherency and consistency information, current system load). However, for the sake of availability, high-performance and low-latency, programmers use

Y. Robert et al. (Eds.): HiPC 2006, LNCS 4297, pp. 472–484, 2006.

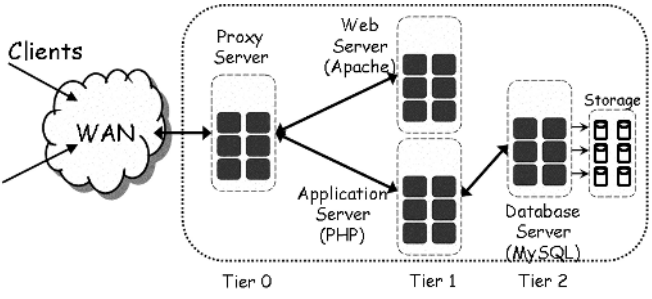

Fig. 1. Web data-centers

ad-hoc messaging protocols for maintaining this shared information. Unfortunately, as mentioned in [2], the code devoted to these protocols accounts for a significant fraction of overall application size and complexity. As system sizes increase, this fraction is likely to increase and cause significant overheads.

On the other hand, System Area Network (SAN) technology has been making rapid progress during the recent years. SAN interconnects such as InfiniBand (IBA) [3] and 10-Gigabit Ethernet (10GigE) have been introduced and are currently gaining momentum for designing high-end computing systems and data-centers. Besides high performance, these modern interconnects provide a range of novel features and their support in hardware, e.g., Remote Direct Memory Access (RDMA), Atomic Operations, Offloaded Protocol support and several others. Recently OpenFabrics [4] has been proposed as the standard interface that allows protable implementations over several modern interconnects like IBA, and iWARP capable ethernet interconnects including [5] Chelsio, Ammasso [6], etc., both in LAN/WAN environments.

In this paper, we design and develop a low-overhead distributed data sharing substrate (DDSS) that allows efficient sharing of data among independently deployed servers in data-centers by leveraging the features of the SAN interconnects. DDSS is designed to support efficient data management and coherence models by leveraging the features like one-sided communication and atomic operations. Specifically, DDSS offers several coherency models ranging from null coherency to strict coherency.

Experimental evaluations with IBA and iWARP-capable Ammasso networks through micro-benchmarks and data-center services such as reconfiguration and active caching not only show an order of magnitude performance improvement over traditional implementations but also show the load resilient nature of the substrate. Application-level evaluations with Distributed STORM using DataCutter achieves close to 19% performance improvement over traditional implementation, while evaluations with checkpointing application suggest that DDSS is scalable and has a low overhead. The proposed substrate is implemented over the OpenFabrics standard interface and hence is portable across multiple modern interconnects.

2 Constraints of Data-Center Applications

Existing data-center applications such as Apache, MySQL, etc., implement their own data management mechanisms for state sharing and synchronization. Databases