Design and Implementation of an Efficient Multipath for a SAN Environment

Jun Luo, Ji-wu Shu, and Wei Xue

Tsinghua University, Beijing, China
luojun03@mails.tsinghua.edu.cn

Abstract. Multipath provides multiple paths between application Servers and storage devices. Multipath can overcome single point of failure, and improve a system's reliability and availability. This paper presents a multi-layer Multipath, and describes the design and implementation of a Multipath system in a storage area network (SAN). For an application server, we implemented Multipath in the volume management layer. For a storage server, we implemented Multipath in the SCSI Middle Level layer. This system can make the most use of the storage server's characteristics to decrease the time of failure discovery and location, and it is independent of lower SCSI cards and storage devices, so it has good compatibility. This paper also proposes methods for choosing paths, automatically recovering paths and balancing the load. We tested the read performance and the average response time, and the results showed that with the load balanced, the read performance improves 17.9% on average, and the average response time decreases 15.2% on average.

1 Introduction

In the SAN environment, data in a storage device can be protected by RAID technology, but RAID cannot increase the reliability of data transfers between the application server and the storage device. If one SCSI card or something in the data transfer path fails, the application server cannot access the storage device. Multipath technology can provide multi-paths between application servers and storage devices. Multipath can ensure a storage device's availability if one path has failed. Multipath is an important part of SAN disaster recoverability. And load balance technology in Multipath can improve system performance.

At present Multipath is commonly implemented on the application server, such as EMC's Powerpath [1] and HP's Securepath [2]. And Multipath is also implemented in the storage device itself. This is often based on a special storage device, and is integrated into the storage device controller. When one port or controller fails, the I/O can access the device through another port or controller. According to the device's different response time to the I/O through different ports, storage devices can be grouped into three behavior models [3]. Holding the device, the server should set a path-choosing policy according to the device's behavior model to achieve the best performance.

G. Chen et al. (Eds.): ISPA Workshops 2005, LNCS 3759, pp. 101–110, 2005.

For the application server Multipath, the path-choosing policy is set on the application server. But in SAN environments, different application servers can connect to different behavior model devices, and application servers with the same devices may even have different device orders. So the administrator should set the server's path-choosing policy independently. Management is very complicated and it may unnecessarily reduce the performance. Load balance in the application server Multipath is only based on the local load, and does not take into account the status of other servers. So it is hard for the storage device to achieve real load balance.

Multipath can be implemented at different levels in an operating system's I/O path. Usually Multipath can be implemented on three levels: volume management, SCSI level and SCSI card driver level. Volume management includes LVM [4] [5], EVMS [6] and MD. It can configure multiple devices exposed by operating system, which point to the same storage device, to form multiple paths for failover. Multipath at this level is very mature and easy to run, but it cannot shield the bottom devices and applications can directly access the bottom devices.

Multipath can also be implemented at the SCSI level. This level is closer to the bottom devices and can shield them from applications. For example, IBM's Michael Anderson and others have implemented a Mid Level Multipath in Linux-2.5+ kernel [3]. This was limited to a special operating system edition. Its compatibility was bad.

The implementation at the SCSI card driver level binds the same two or more cards to form a Multipath. This can also shield devices from the upper level. Once a path fails, the card driver will directly send commands through other paths. Adaptec's Qlogic 5.x edition [7] and Emulex's MultiPulse [8] both implement Multipath at the card driver level. And many companies' Multipath software is based on the card driver's Multipath function as designed by Adaptec or Emulex, for example EMC's Powerpath software is based on Emulex's MultiPulse. This method is based on a special SCSI card, and can only be used for the same edition card of the same company. It does not have good compatibility.

In this paper, we describe the design and implementation of a multiple layers Multipath for the Tsinghua Mass Storage Network System (TH-MSNS) [9]. The TH-MSNS has an additional storage server between Fabric and the storage device. In the storage server, the TH-MSNS has software to simulate a SCSI TARGET Driver. We added a Multipath layer on the storage server to implement Multipath for the storage device. According to the storage server's special software architecture, we implemented the Multipath layer between the SCSI Target Middle Level and the operating system's SCSI Mid Level. It is compatible with different Linux kernels, and is independent of the lower SCSI cards and storage devices. Storage devices are directly connected to the storage server, so the storage server can set the most suitable path-choosing policy for storage devices with different behavior models. And this method can also shield storage devices' characteristics from the application server, so the application server can set path-choosing policy according to the Fabric status. As for the application