Design of Storage Area Network Based on Metro WDM Networking

Bernardi Pranggono and Jaafar M.H. Elmirghani

School of Engineering, University of Wales Swansea

Abstract: Storage Area Networks (SANs) have become an essential part of today's enterprise computing. The paper reports the design of SANs based on metro Wavelength Division Multiplexing (WDM) networks technologies. It reviews related SAN technologies. It also presents analysis and simulation results of a proposed architecture. Throughput, delay and packet dropping probability results are presented under Poisson and self-similar traffic.

1. Introduction.

In the past few years we have witnessed a tremendous growth in network traffic due to the explosive development of data-intensive applications such as enterprise resource planning (ERP), customer relationship management (CRM), multimedia, e-business, backup and recovery, data warehousing and mining. This phenomenon creates a constant demand for a dedicated storage system with a bigger capacity of data storage. This storage system needs to be scalable, robust and able to provide data in an efficient way.

Direct Attach Storage (DAS) is clearly unable to comply with these latest requirements since it depends heavily on the computer system resources and it is also prone to disaster due to single-point-of-failure (SPOF). Network Attached Storage (NAS) and Storage Area Networks (SANs) are the better solutions to fulfil the requirements. In their latest development, the difference between NAS and SAN is becoming more and more unclear and harder to be identified as the similarity in their features and capabilities grows.

IP-based storage has become more widespread in the past few years due to its simplicity and cost-effectiveness. Through the work of the Internet Engineering Task Force (IETF) a handful of new protocols have emerged, which include Fibre Channel Internet Protocol (FCIP), Internet Fibre Channel Protocol (iFCP), and Internet Small Components Systems Interface (iSCSI). However, these IP-based protocols inherit the basic limitations of IP including being: a slow transport protocol compared to Fibre-Channel (FC) technology [1], a less robust transport mechanism compared to synchronous optical network/synchronous digital hierarchy (SONET/SDH) [2], possessing large overheads [3] and having in many cases, limited quality of service guarantees.

Transport network technologies have kept in pace with the unprecedented demand on bandwidth with progress in telecommunication technologies. Photonic network technologies and WDM have become a technology of choice to increase network bandwidth in recent years and the trend is continuing. Recent developments in WDM research have led to new technologies that offer an unparalleled bandwidth as currently available systems can support up to 600 wavelengths per fiber, enabling a single fiber to transfer few terabits per second of data and information.

Due to its superior characteristics, WDM technology is not only utilized in backbone wide area networks but has also started to be utilized in metro access networks [4, 5]. This latest development makes it more and more feasible for WDM technology to be used in SANs to give SANs a more competitive, high-bandwidth, high-scalability and low-latency edge.

The unprecedented users demand and the advent of photonics technologies create great opportunities for metro WDM networks based SAN. Optical storage area networking has started to receive attention. For example the work in [6] discusses a Fibre Channel – Arbitrated Loop (FC-AL) interconnect.

The remaining sections of this paper are organized as follows. Section II reviews related SAN technologies. Section III discusses a proposed architecture for metro WDM storage area networking. Simulation results using both Poisson and self-similar traffic models are given in Section IV. Finally, the main conclusions are given in Section V.

2. Storage Area Network

SANs are devoted high-speed networks (usually independent from LANs and WANs) whose primary function is to interconnect computer systems and storage devices in an efficient way. SANs have four primary components, a communication infrastructure such as communication interfaces and switches, a management layer which manages the connections, storage devices, and computer systems. SANs utilize block-storage commands to access shared data in the storage. Block-storage commands highly depend on the Small Computer Systems

Interface (SCSI) command protocol for data transfer [3]. From the perspective of a computer system, a SAN is a network that transfers both data blocks and the commands required to retrieve and store these data blocks on disks robustly and securely.

In recent years SAN technologies have developed rapidly and a number of new IP-based protocols have emerged, i.e. FCIP, iFCP and iSCSI. FCIP is an IP-based FC tunneling protocol which can be used to interconnect SAN facilities over legacy IP networks. iFCP is an IP-based gateway-to-gateway protocol. iFCP mechanisms provide FC fabric service to FC devices at the wire speed using an IP infrastructure. iSCSI uses the SCSI protocol over a TCP/IP network. It enables any machine on an IP network (initiator) to connect to a remote dedicated server and perform block transfer work on it as if it is a local hard disk. Since the IETF ratified iSCSI as a standard in February 2003, it is probably true to conclude that IP-based storage (SAN-over-IP) will continue to develop. However, better use of the optical layer may be needed and expected.

These three IP-based protocols offer IP simplicity, cost-effectiveness and ubiquitous deployment. However, they all inherit the well-known drawbacks of IP as a slow transport protocol compared to FCP, particularly when delivering block-storage type commands like SCSI commands [1]. It is shown in [7] that IP-SAN has only high throughput (> 10 Mbytes/s) up to medium distance (up to 300 km) and also prone to packet loss. On the other hand, legacy SANs i.e. FC-SONET/SDH have drawbacks such as the FC distance limitation (up to 10 km) and SONET inefficiency in transporting bursty data traffic as it designed to support voice traffic.

It is therefore of interest to explore systems and protocols that may enable WDM to be effectively used to enhance the capacity, access speed and capabilities of distributed SANs.

3. SAN Based on Metro WDM Network

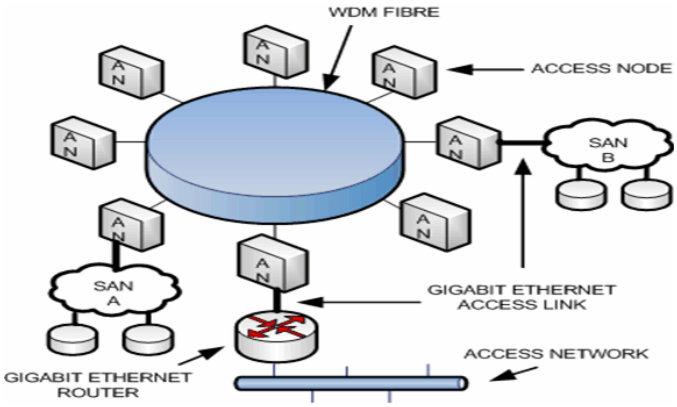

Fig 1. Network architecture

Metro WDM may be considered for SANs as it has high-bandwidth, high-scalability and low-latency. A MAN type network has to be relatively flexible and scalable in the sense that it is expected to satisfy different user demands and possible configurations. In addition, the physical architecture is strongly reliant on the geographical relative positions of the access nodes (AN).

A simple MAC protocol for metro WDM access ring networks proposed in [8] provides good results in terms of throughput, queueing delay, and packet dropping probability. When a node wants to send a packet on its assigned wavelength, it probes the activity of the channel. If there is an empty slot, the node simply sends the data into the slot unit (it is assumed that the packet size is always fixed and equal to the slot size, i.e. MTU). The node goes into wait mode (until next round) if the slot is found to be already used. The proposed architecture uses a destination-stripping mechanism where the destination node is responsible to mark the slot empty once it receives the packet correctly. To maintain fairness, a simple restriction is introduced i.e. a node cannot re-use the slot it just emptied. We evaluate this protocol in a metro setting for SAN applications, where SANs co-exist with other network users.

The architecture is a single-fibre multi-channel slotted ring which is designed to interconnect Access Nodes (ANs) (in this scenario some of the ANs serve SANs) on a regional scale. Each AN has add-and-drop capabilities to access the ring slots and is used to link an access network to the ring. The architecture uses Gigabit Ethernet (GbE) networking operating at 1 Gb/s to link access networks to the metro network. The size of the slot on the ring is fixed and is equal to Ethernet Maximum Transmission Unit (MTU) frame size, i.e. 12,000 bits. Each AN is equipped with one fixed transmitter and four fixed receiver (FT-FR^4) or one fixed transmitter and one tuneable receiver (FT-TR) [9]. The proposed architecture is relatively close to TT-FR model discussed

by Marsan *et al*. in [10] in the sense that wavelengths are shared by a fixed number of nodes. However, in [10] each node can use any wavelength for transmission but it receives data on a fixed wavelength. Therefore, if there are less wavelengths than nodes, a fixed number of nodes will share a wavelength for reception, and all the wavelengths will be shared for transmission. Bononi [11] analytically evaluated the performance of both FT-TR (equivalent to our FT-FR^4) and TT-FR slotted architectures and concluded that they have a similar theoretical networking performance. The proposed SAN network architecture using metro WDM ring network is shown in Fig. 1.

4. Simulation Results

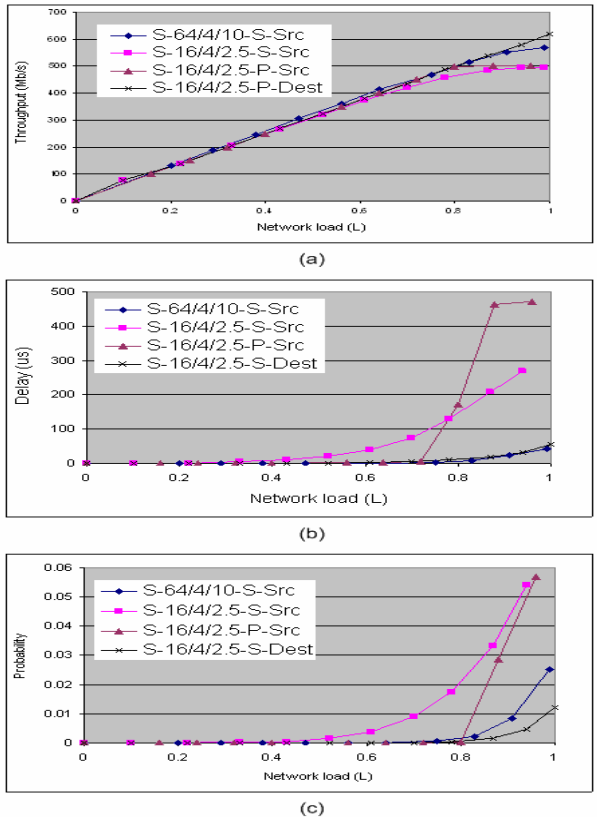

Fig. 2 Network architecture performance (a) Average throughput; (b) Average delay; (c) Average probability

Two comparable networking environments are considered. In both environments the length of the ring is 138 km (with a ring diameter of around 44 km). Four wavelengths were used in both architectures. The first network interconnects 16 nodes with each wavelength transferring data at 2.5 Gb/s (S-16/4/2.5). In the second environment, 64 nodes are interconnected by the network where each wavelength data rate is 10 Gb/s (S-64/4/10).

The architecture was evaluated by an event driven simulation using Poisson and self-similar traffic. A specific notation is used to specify the traffic model and slot release and re-use scheme applied in the simulation. The first suffix is used to specify the traffic model, i.e. S for self-similar traffic and P for Poisson traffic. A second suffix (Src or Dest) is used to specify source and destination stripping, respectively.

As can be seen in Fig. 2(a), the throughput of destination stripping architecture (DSA) is better than source stripping architecture (SSA). For S-16/4/2.5, at normalized network load L=1, the throughput for SSA is around 500 Mb/s while it reaches 600 Mb/s for DSA. In fact, the throughput of the DSA reaches 880 Mb/s when the normalized network load is at 1.42. A normalized network load greater than 1 is possible due to the slot-reuse policy in the destination-stripping mechanism. The queueing delay of the architecture is five to eight times better than with the SSA and it only reaches 55 µs when the normalized network load equals 1. This phenomenon is a direct result of destination stripping as slots are freed and re-used earlier than with source stripping (Fig. 2(b)). Furthermore, since the buffer load is decreased drastically, the packet dropping probability is also decreased significantly (Fig. 2(c)). The packet dropping probability (PDP) is greatly reduced as it reaches 0.01 for L=1 but is still only equal to 4e-04 for L=0.8. More importantly, when the destination-stripping scheme is used in a larger network, i.e. the S-64/4/10 case, the packet dropping probability was very small. This is understood by observing that a larger number of nodes leads to earlier slot reuse, as the distance between nodes decreases, resulting in reduced queueing delay, buffer load and PDP. This scalability feature is suitable for SAN in metro settings where many nodes are expected to be involved. This PDP is also comparable to IP-based SAN which is around 1e-04 - 1e-03 [7]. It is also worth mentioning than in this latter case, and due to the space-reuse capabilities of optical ring networks [11], the proposed architecture could handle a network load $L = 1.5$ without noticeable deterioration in performance. Indeed, this is suitable for the SANs' environment where scalability, reliability and performance are more crucial than any other factors.

The initial design was modified from FT-FR^4 to FT-TR with collision avoidance mechanism (FT-TR-CA) in order to make the architecture more scalable. The CA mechanism used in the new architecture simply allows one of the colliding packets to do another loop round the ring. The network simulation results from the new architecture (FT-TR-CA) confirmed that while the transmission delay (sum of the queueing delay and the propagation delay) is increased by the CA mechanism, overall network performance (throughput, queueing

delay, packet dropping probability and buffer load) is not affected in a significant manner by the CA mechanism as the ring latency is very low, i.e. 691.2 µs. The transmission delay is seen to be very acceptable in both architectures. The transmission delay touches 500 µs for the FT-FR^4 and 800 µs for the FT-TR-CA network when the normalized network load just passed 1. Both architectures (FT-FR^4 and FT-TR-CA) would also be appropriate in carrying storage traffic and more so real-time multimedia traffic since the end-to-end user delay (sum of the transmission delay and the processing time) for such traffic must usually be under 100 ms [12]. In term of scalability, the architecture is shown to be scalable as the performance improved when the number of access nodes and the wavelengths rate were proportionally increased.

5. Conclusions.

Over the past few years, the role of SANs in enterprise computing has become more and more significant, with customers demanding more secure, more reliable and more flexible storage. Significant research and development has been directed towards pushing the use of IP-based SAN, and a number of new protocols have emerged, namely FCIP, iFCP, and iSCSI. IP-based storage is attractive since it is simple, cost-effective and relatively easy to use. However, it inherits the IP basic limitations including slow transport. On the other hand, data traffic is transported inefficiently by FC-SONET based SANs due to the voice-centric characteristics of these networks.

The article has proposed a metro WDM network architecture for storage area networking. The architecture uses a multi-channels slotted ring with a simple MAC protocol to manage access to the network. Simulation results were presented for throughput, delay and packet dropping probability. The simulations showed that the proposed architecture is suitable for SANs applications which demand low queueing delay, low packet dropping probability and high throughput. The simulations also demonstrated that the architecture is scalable and the performance improved when the number of access nodes and the wavelengths rate were proportionally increased.

Acknowledgments.

The work on the INSTANT project is funded through an EPSRC/DTI LINK Grant.

References.

[1] H. Simitci, C. Malakapalli, and V. Gunturu, "Evaluation of SCSI over TCP/IP and SCSI over fibre channel connections," in Hot Interconnects 9, Aug., 2001.

[2] P. Molinero-Fernandez, N. McKeown, and H. Zhang, "Is IP going to take over the world (of communications)?," *ACM SIGCOMM Computer Communications Review*, vol. 33, no. 1, pp. 113-118, Jan., 2003.

[3] P. Sarkar, K. Voruganti, K. Meth, O. Biran, and J. Satran, "Internet protocol storage area networks," *IBM Systems Journal*, vol. 42, no. 2, pp. 218-231, 2003.

[4] K. Noguchi, "Field trial of full-mesh WDM network (AWG-STAR) in metropolitan/local area," *IEEE JLT*, vol. 22, no. 2, pp. 329-336, Feb., 2004.

[5] D. Stoll, A, B, and C, "Metropolitan DWDM: A dynamically configurable ring for the KomNet field trial in Berlin," *IEEE Commun. Mag.*, vol. 39, Feb. 2001.

[6] J. R. Heath and P. J. Yakutis, "High-speed storage area networks using a fibre channel arbitrated loop interconnect," *IEEE Network*, vol. 14, no. 2, pp. 51-56, Mar-Apr. 2000.

[7] R. Telikepalli, T. Drwiega, and J. Yan, "Storage area network extension solutions and their performance assessment," *IEEE Commun. Mag.*, pp. 56-63, Apr. 2004.

[8] C. S. Jelger and J. M. H. Elmirghani, "A slotted MAC protocol for efficient bandwidth utilization in WDM metropolitan access ring networks," *IEEE JSAC*, vol. 21, no. 8, pp. 1295-1305, Oct., 2003.

[9] B. Mukherjee, "WDM-based local lightwave networks Part I: Single-hop systems," *IEEE Network*, vol. 6, no. 3, pp. 12-27, Mar., 1992.

[10] M. Marsan, A, B, and C, "All-optical WDM multi-rings with differentiated QoS," *IEEE Commun. Mag.*, vol. 37, no. 2, pp. 58-66, Feb. 1999.

[11] A. Bononi, "Scaling WDM slotted ring networks," in Proc. Information Sciences and Systems Conf., 1998.

[12] M. Baldi and Y. Ofek, "End-to-end delay analysis of videoconferencing over packet-switched networks," *IEEE/ACM Trans. Network.*, vol. 8, no. 4, pp. 479-492, Aug., 2000.