Design and analysis of optical metropolitan networks

Lian, Ziwen

2005

Lian, Z. (2005). Design and analysis of optical metropolitan networks. Master's thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/4608>

<https://doi.org/10.32657/10356/4608>

Nanyang Technological University

Downloaded on 26 Apr 2023 19:00:20 SGT

DESIGN AND ANALYSIS OF OPTICAL METROPOLITAN NETWORKS

LIAN ZIWEN

**SCHOOL OF ELECTRICAL & ELECTRONIC ENGINEERING
NANYANG TECHNOLOGICAL UNIVERSITY**

2005

Design and Analysis of Optical Metropolitan Networks

Lian Ziwen

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Master of Engineering

2005

Acknowledgements

I would like to thank everybody who contributed in so many ways, great and small, to my ideas and work in the past two years.

First of all, I would like to express my greatest gratitude to my supervisor, Associate Professor Zhong Wende. He has been a constant source of inspiration, motivation and guidance throughout the past two years.

It is my great pleasure to thank and acknowledge Associate Professor Sanjay Kumar Bose and Doctor Wang Yixin for their continuous assistance during the past two years and their invaluable suggestions on my thesis.

Many thanks also go to my labmates in Institute for Infocomm Research (I²R) for their helpful suggestions and encouragements.

I am also grateful to my family. This thesis is dedicated to my parents and my sister whose love and continuous support are always the greatest inspiration for me.

Summary

Ethernet has grown to be the dominant technology in local area networks (LANs) to date due to its simplicity, cost-effectiveness, ease of manageability and maturity. The 10 Gigabit Ethernet (10GbE) is the latest generation of Ethernet technology which is able to provide 10Gbps line speed and at the same time inherits all advantages from traditional Ethernet. Besides the sheer amount of bandwidth, one of the significant improvements of 10GbE is the new design of the LAN and WAN (wide area network) PHYs (physical layers). With these two special PHYs, 10GbE now is able to work in MANs (metropolitan area networks) and WANs. It would be desirable to have 10GbE technology to be one of the next-generation MAN technologies not only because it will extend the advantages of traditional Ethernet from LANs to MANs, but it will also provide seamlessly connections to current existing Ethernet LANs and eliminate the need for additional layers of protocols and synchronization to convert transmissions between Ethernet and other network types. However originally designed for data transmissions in small campus networks, Ethernet technology has some limitations when it operates in MANs. With the motivation of adapting Ethernet technology from LANs to MANs, this thesis aims to design a new network scheme called optical Resilient Ethernet Ring (RER) based on 10GbE to address some of these limitations.

System designs of the proposed RER scheme such as the network architecture, frame format, frame forwarding table etc. are analyzed in this thesis. RER supports true ring topology. A simple RER header is designed specifically for the RER scheme. With the combination of the RER header and the conventional Ethernet frame header, Ethernet frames can be efficiently transmitted in the RER. A hierarchical frame forwarding scheme is also proposed for the RER to simplify the process of forwarding frames from sources to destinations. A new RER forwarding table structure is designed specifically for the hierarchical frame forwarding scheme. System operations of the RER such as frame switching, self-learning process, failure detection and recovery etc. are also presented in this thesis. RER supports flexible bandwidth allocation and spatial reuse mechanism. The ring topology of RER simplifies the decision-making process for frame forwarding and also enables the RER network to recover from a link or node failure within 50ms. Two typical types of protection schemes, the wrapping scheme and the steering scheme, are analyzed in terms of protection switching time and packet loss. A new *wrap then steer protection scheme* is proposed and examined. The new protection scheme combines the advantages of the wrapping scheme and the steering scheme to achieve fast recovery and optimal transmitting paths after recovery simultaneously. RER switch design is also analyzed. Functional blocks are designed for RER switches. The two most time-consuming switch operations, i.e. hashing operation and memory access, are discussed.

This thesis also studies the interconnection of multiple RER rings. Three different

approaches based on transparent bridging technique are proposed and examined in terms of bandwidth efficiency and size of frame forwarding table. This study has found that by allowing RER switches to learn MAC addresses residing in remote rings and making use of a globally unique RER switch identifier, good bandwidth efficiency and small size of frame forwarding table can be achieved. Protection issues of interconnected RER rings are also discussed. The spanning tree protocol (STP) needs to be applied in bridge ports (ports that interconnect rings) to form a loop-free spanning tree to prevent frames from circulating among them. However, this introduces the infamous persistent black hole problem. The proposed solution to this is to synchronize the forwarding table entries in ring switches with the changes of port states by a real-time forwarding table entry removal mechanism.

Finally, three sets of computer simulations are conducted to demonstrate the feasibility and effectiveness of the proposed RER scheme. The numerical results of the first set of simulations show that RER transmits data much more efficiently compared to other types of ring networks such as the Resilient Packet Ring (RPR) and rapid spanning tree protocol (RSTP) based ring networks. The results of the second set prove the proposed wrap then steer protection scheme combines the advantages of the wrapping scheme and the steering scheme to achieve fast recovery and optimal transmitting paths after recovery simultaneously. The results of the third set illustrate the fact that bridge nodes which interconnect RER rings can easily become bottlenecks of the network if a large amount of traffic is passed across rings.

Table of Content

Acknowledgements.....	i
Summary.....	ii
Table of Content.....	v
List of Figures.....	viii
List of Tables.....	xi
List of Acronyms.....	xii
Chapter 1 Introduction.....	1
1.1 Motivation.....	2
1.2 Objectives.....	8
1.3 Major Contributions.....	9
1.4 Organization of the Thesis.....	11
Chapter 2 Review of Network Technologies.....	13
2.1 Ethernet.....	13
2.1.1 Ethernet Data Frame Format.....	14
2.1.2 Ethernet Devices.....	15
2.1.3 Spanning Tree Protocol.....	17
2.1.4 10 Gigabit Ethernet.....	19
2.2 SDH/SONET Rings.....	21
2.2.1 Frame Format.....	22
2.2.2 Circuit-based Network.....	23

2.2.3 Self-Healing Rings.....	25
2.3 RPR.....	27
2.3.1 Physical Layer.....	27
2.3.2 Media Access Control Layer.....	28
2.4 Next-Generation MANs.....	31
Chapter 3 Optical Resilient Ethernet Ring.....	35
3.1 RER OSI Layer Model.....	35
3.2 System Architecture.....	37
3.3 Switch ID.....	39
3.4 Frame Forwarding Table.....	40
3.5 Ring Image.....	43
3.6 MAC Frame Structure.....	43
3.7 Switch ID Assignment Process.....	46
3.8 Self-Learning Process.....	47
3.9 Frame Switching.....	49
3.10 Prevention of Frame Circulation.....	52
3.11 Failure Detection and Recovery.....	53
3.11.1 Failure Detection.....	53
3.11.2 Protection Schemes.....	55
3.11.3 Summary for Protection Schemes.....	62
3.12 RER Switch Design.....	62
3.13 Summary.....	69
Chapter 4 Interconnection of Multiple RERs.....	72
4.1 Approaches for Interconnecting RER Rings.....	72
4.1.1 Interconnecting RER Rings - First approach.....	73
4.1.2 Interconnecting RER Rings - Second approach.....	76
4.1.3 Interconnecting RER Rings - Third approach.....	79
4.2 Protection Issues of Interconnected RER Rings.....	84

4.3 Summary.....	87
Chapter 5 Simulations and Performance Analysis.....	89
5.1 Simulation on Three Different Ring Approaches.....	89
5.2 Simulation on Three Different Protection Schemes.....	95
5.2.1 Scenario 1: Light Offered Load.....	96
5.2.2 Scenario 2: Moderate Offered Load.....	101
5.3 Simulation on Interconnection of Two RER Rings.....	106
Chapter 6 Conclusions and Future Work.....	109
6.1 Conclusions.....	109
6.2 Recommendations for Future Work.....	115
Bibliography.....	117
Author's Publications Related to This Thesis.....	125

List of Figures

Figure 2.1: Ethernet Basic Data Frame Format.....	14
Figure 2.2: Spanning Tree Protocol.....	18
Figure 2.3: Distances supported in 10GbE.....	21
Figure 2.4: SONET Frame Format.....	22
Figure 2.5: Circuit-switched SONET.....	24
Figure 2.6: Fully logical mesh for a 4-node SONET.....	24
Figure 2.7: RPR base data format.....	28
Figure 3.1: RER OSI layer model.....	36
Figure 3.2: Architecture of the RER.....	38
Figure 3.3: Example of frames forwarding tables.....	42
Figure 3.4: MAC frame structure.....	45
Figure 3.5: Figure 3.5: Switch ID assignment process.....	46
Figure 3.6 (a): Self-learning process in source switch.....	48
Figure 3.6 (b): Self-learning process in intermediate switch.....	48
Figure 3.7: Failure detection process.....	55
Figure 3.8: Example of fiber-cut failure.....	56
Figure 3.9: Wrapping scheme.....	57

Figure 3.10: RER switch functional diagram for handling sub-network traffic.....	59
Figure 3.11: RER switch functional diagram for handling ring traffic.....	63
Figure 4.1: Transparent bridge network reference.....	73
Figure 4.2: Transparent bridge architecture.....	73
Figure 4.3 (a): First approach for interconnection of multiple RER rings – 1.....	74
Figure 4.3 (b): First approach for interconnection of multiple RER rings – 2.....	75
Figure 4.4 (a): Second approach for interconnection of multiple RER rings – 1.....	76
Figure 4.4 (b): Second approach for interconnection of multiple RER rings – 2.....	78
Figure 4.5 (a): Third approach for interconnection of multiple RER rings – 1.....	80
Figure 4.5 (b): Third approach for interconnection of multiple RER rings – 2.....	82
Figure 4.6: Broadcast storm scenario due to interconnection of multiple RER rings..	84
Figure 4.7: Single-attachment to interconnect multiple RER rings.....	85
Figure 4.8: Dual attachment to interconnect multiple RER rings.....	85
Figure 4.9: Persistent black hole scenario in interconnection of multiple RER rings..	86
Figure 5.1: Simulation topology for different ring approaches.....	90
Figure 5.2: Average delay vs. normalized offered traffic for different ring networks with 10 nodes.....	93
Figure 5.3: Normalized carried traffic vs. normalized offered traffic for different ring networks with 10 nodes.....	94
Figure 5.4: Throughput vs. number of nodes for different ring networks with a fixed access traffic rate at each node 2.5Gbps.....	94
Figure 5.5: Carried traffic under double-fiber cut failure with access traffic rate at	

each node 1Gbps.....	99
Figure 5.6: Carried traffic under single-fiber cut failure with access traffic rate at each node 1Gbps.....	99
Figure 5.7: Traffic loss under double-fiber cut failure with access traffic rate at each node 1Gbps.....	100
Figure 5.8: Traffic loss under single-fiber cut failure with access traffic rate at each node 1Gbps.....	100
Figure 5.9: Carried traffic under double-fiber cut failure with access traffic rate at each node 2Gbps.....	104
Figure 5.10: Carried traffic under single-fiber cut failure with access traffic rate at each node 2Gbps.....	104
Figure 5.11: Traffic loss under double-fiber cut failure with access traffic rate at each node 2Gbps.....	105
Figure 5.12: Traffic loss under single-fiber cut failure with access traffic rate at each node 2Gbps.....	105
Figure 5.13: Simulation topology for interconnection of two RER rings.....	106
Figure 5.14: Normalized carried traffic vs. normalized offered traffic for different traffic patterns.....	108

List of Tables

Table 2.1: SONET/SDH hierarchy.....	23
Table 2.2: Comparison of SONET, RPR and 10GbE as a next-generation MAN technology.....	32
Table 3.1: Types defined in the utility field.....	45
Table 5.1: Average traffic delay under double-fiber/single-fiber cut failure with access traffic rate at each node 1Gbps.....	98
Table 5.2: Average traffic delay under double-fiber/single-fiber cut failure with access traffic rate at each node 2Gbps.....	103
Table 5.3: Comparison of three protection schemes.....	103

List of Acronyms

Abbreviations	Full Expressions
10GbE	10 Gigabit Ethernet
AIS	Alarm Indication Signal
ANSI	American National Standards Institute
APS	Automatic Protection Switching
ATM	Asynchronous Transfer Mode
BLSR	Bidirectional Link-Switched Ring
BPDU	Bridge Protocol Data Unit
CIR	Committed Information Rate
CoS	Class of Services
CRC	Cyclic Redundancy Check
DA	Destination Address
DSID	Destination Switch Identifier
EIR	Excess Information Rate
GFP	Generic Framing Protocol
GMII	Gigabit Media Independent Interface
GRS	GFP Reconciliation Sublayer
HDLC	High-level Data Link Control

HEC	Header Error Check
IP	Internet Protocol
ITU	International Telecommunications Union
LAN	Local Area Network
MAC	Media Access Control
MAN	Metropolitan Area Network
MII	Media Independent Interface
MPLS	Multi-Protocol Label Switching
NS2	Network Simulator version 2
OAM&P	Operations, Administration, Maintenance, and Provisioning
OC	Optical Carrier
OSI	Open Systems Interconnection
PCS	Physical Coding Sublayer
PHY	Physical Layer
PMD	Physical Media Dependent
POS	Packet over SONET
QoS	Quality of Services
RER	Resilient Ethernet Ring
RPR	Resilient Packet Ring
RSTP	Rapid Spanning Tree Protocol
SA	Source Address
SDH	Digital Hierarchy

SID	Switch Identifier
SSID	Source Switch Identifier
SLA	Service Level Agreement
SONET	Synchronous Optical NETworking
SPE	Synchronous Payload Envelope
STA	Spanning Tree Algorithm
STP	Spanning Tree Protocol
TOH	Transport Overhead
TTL	Time to Live
UPSR	Unidirectional Path-Switched Ring
VLAN	Virtual LAN
WAN	Wide Area Network
WTS	Wrap then Steer
XAUI	10 Gigabit Attachment Unit Interface
XGMII	10 Gigabit Media Independent Interface

Chapter 1 Introduction

A metropolitan area network (MAN) [1-6] is a large computer network usually spanning a city. It covers a geographic area larger than the area covered by a local area network (LAN) [7-8] but smaller than the area covered by a wide area network (WAN) [7, 9]. A MAN may interconnect several LANs by bridging them with backbone lines and it may also offer efficient connections to WANs. Nowadays fiber optic rings are widely used in MAN communication topologies. It is not difficult to appreciate why this topology has become popular. In practice, ring topologies take many advantages over other alternatives in many applications [10-13]. These alternatives include tree, star, branch, and mesh topologies. Rings are more efficient than meshes when optimizing for fiber route-miles. Rings simplify route diversity and failure protection; that is, even under a link failure, full connectivity among all ring nodes is maintained. Rings are more deterministic than meshes. Furthermore, rings have simple route calculation algorithm. A good MAN technology should be able to provide high bandwidth, low delay/jitter and high availability, high scalability, efficient bandwidth allocation, good resiliency, efficient support for data traffic transmissions and low cost.

1.1 Motivation

MANs can be based on various technologies. Synchronous Digital Hierarchy/ Synchronous Optical NETworking (SDH/SONET) [14-20] is probably the most widely deployed MAN technology today. SONET rings provide point-to-point circuits among ring nodes. With a dedicated circuit, SONET provides guaranteed bandwidth, delay, and jitter. Moreover, SONET utilizes dual-ring architecture to achieve sub-50ms system restoration after a link or node failure, thereby making link failures nearly transparent to most applications and higher layers. It also uses synchronous Time Division Multiplexing, which is bandwidth inefficient because of the fact that a fixed amount of bandwidth is allocated to each end-to-end circuit and the bandwidth cannot be shared between circuits. Static circuits preclude nodes from using unused bandwidth on other links and from bursting beyond their circuit rate. Thus, overall, circuit-switched rings are not efficient in transport data traffic which is inherently bursty and must operate at considerably lower bandwidth utilization than packet-switched rings. SONET does not support spatial reuse mechanism; that is, each frame has to circulate the entire ring before being removed. This is another waste of bandwidth because the bandwidth down the road beyond the destination cannot be reclaimed. Last but not least, SONET's resiliency is achieved at the cost of wasting half of the total bandwidth because SONET reserves one ringlet out of two to provide system protection. Resilient Packet Ring (RPR) [21-24] is an emerging MAN technology which operates in dual-ring architecture as well. Unlike SONET, both of RPR ringlets can be used to transmit data in normal operations. Protection mechanisms are defined to be able to provide sub-50ms system protection and restoration against link RPR supports destination packet removal thereby enabling spatial reuse. As contrast to

circuit-based technologies such as SONET, RPR is a packet-based technology meaning bandwidth can be assigned dynamically to each packet. This overcomes a limit of true circuits that prevents unused bandwidth from being reclaimed by other nodes or traffic classes. RPR also defines three traffic classes. Each class of traffic has different requirements in guaranteed bandwidth, end-to-end delay and jitter bounds. Besides these features, RPR defines a fairness algorithm which can throttle traffic dynamically to ensure every node gets fair access to the network resources. However, though proposed four years ago, RPR has not been completely standardized yet partly because of the complexity of its protocol.

It is not difficult to appreciate why current MAN technologies, e.g. SONET and RPR, favor dual-ring topology, but not other common network topologies such as star, tree and mesh etc. Among these topologies, star topology and tree topology are the simplest. However, these two topologies fail to provide adequate network resiliency in MANs. The failure in the hub in a star topology based network can disable large segments of the network. Likewise, the failure of a link or a node in a tree topology based network will either separate the network into two isolated parts or results in a long network recover period to re-construct a new tree. Unlike these two topologies, dual-ring topology has inherent resiliency because there always exists two paths for a frame to be transmitted from the source node to the destination node in normal operations. In case one path is no longer available due to a fiber-cut or a node failure, the other path can be quickly activated to avoid service disruption. SONET has proved that dual-ring topology is able to provide a fast network recovery within 50ms after a fiber-cut or a node failure. Mesh

topology also provides good resiliency as what dual-ring topology does by interconnecting nodes to ensure that more than one path exist between any two nodes. However, mesh topology is generally much more complex compared to dual-ring topology. Configuring or administrating a mesh topology based network could be not only difficult but also time consuming. Finally, the simplicity of dual-ring topology ensures a shorter frame transmitting time because no forwarding decisions have to be made when a node forwards frames (see Section 3.9 for more details).

Over twenty years' development, Ethernet has become the dominant LAN technology today. Ethernet is characterized by low cost, simplicity, ease of manageability, and simple integration with existing equipment. The emerging 10 Gigabit Ethernet (10GbE, IEEE 802.3ae Standard) [25-27] over optical fiber is rapidly becoming another solution of choice for next-generation MANs. The new 10GbE preserves the current Ethernet semantics, including minimum and maximum frame size, and frame format. It has some distinct characters from Gigabit Ethernet, which makes it ready to be deployed in MANs and WANs. 10GbE supports full-duplex transmissions only and works only with optical media while Gigabit Ethernet supports both half-duplex and full-duplex transmissions and also works over copper medium. The 802.3ae specification defines two sets of PHYs (physical layers) for 10GbE called the LAN and WAN PHYs. In the LAN PHY, a 1,550 nm serial interface with a range of just less than 25 miles over single-mode fiber is defined which is long enough for most MAN applications. The WAN PHY definition is also included to operate at a data rate compatible with a payload of OC-192 SONET. This technology is very significant because not only will Ethernet run at 10Gbps and serve as

a LAN connection, but it will also work in MANs and WANs. With 10GbE, network managers will be able to build LANs, MANs, and WANs using Ethernet as the end-to-end Layer 2 transport. In spite of these changes in 10GbE from earlier versions of Ethernet that makes 10GbE more suitable to be deployed in an MAN environment, 10GbE is still considered an Ethernet technology because it uses the same IEEE 802.3 MAC protocol, 802.3 frame format, and the same minimum and maximum frame sizes and it is also designed to run all of the familiar advanced Layer 2 Ethernet functions, such as IEEE 802.3ad link aggregation and 802.1Q VLAN etc. There are many advantages to adapt 10GbE in MANs. First of all, 10GbE is simple to implement. Ethernet is an understood technology because it has been in place for so long and the technology concepts are easy to grasp. This almost eliminates the need for training of network staff when 10GbE is deployed on the network. Its simplicity also results in highly reliable systems, because there are fewer components that can break down. High reliability greatly reduces the life-cycle cost of a 10GbE implementation. Another major strength of 10GbE is that it is Ethernet-based. This means that Ethernet and Fast-Ethernet frames being forwarded across the backbone and traversing 10GbE links do not need to be translated or encapsulated into another protocol and hence eliminates the need for additional layers of protocol and synchronization that would have required extra cost and complexity. This also makes the total delivery time of a frame across the network much faster. Last but not least, Ethernet is generally a cheap technology compared to its alternatives. The low cost characteristics of 10GbE helps carriers better manage costs in MANs. Therefore, 10GbE is a good candidate to be adapted in MANs.

However originally designed for small campus networks, Ethernet protocols may not fulfill the stringent requirements in MANs. There are some limitations for 10GbE to become a MAN technology. Traditional Ethernet technology cannot work over topologies which contain active loops, such as rings and meshes [28]. The IEEE 802.1D Spanning Tree Protocol (STP) [28-29] is therefore used in Ethernet switches to select a tree from the current topology. However, the STP is not desirable in MANs because of the slow network convergence; that is, the STP typically takes tens of seconds to complete the tree redrawing process after a topology change which is not acceptable to guarantee service-level agreements (SLA) [30]. IEEE 802.1w Rapid Spanning Tree Protocol (RSTP) [29, 31-32] is an evolution of the STP that provides faster network convergence after a topology change. However, depending on the size of the network, the convergence time of the RSTP may still vary from 10ms to a full second or more. Besides, 10GbE does not support extensive differentiated data services, quality of services (QoS) and carrier-class management [33-34].

10GbE is an OSI Layer 2 technology. There are advantages for running Layer 2 technologies over MANs. Layer 2 switching is faster than Layer 3 routing because Layer 3 routing investigates the packet header more deeply, it adds more overhead than straight Layer 2 switching. In addition, Layer 2 switches or bridges are much cheaper than Layer 3 routers. There are some disadvantages for running Layer 2 technologies over MANs as well. First of all, Layer 2 addressing assumes a flat address space with universally unique addresses. Therefore, Layer 2 is not suitable for a large network with many terminal devices. Routers operate at OSI Layer 3, so can formulate and adhere to a hierarchical

addressing structure. Routers can build their networks that maintain performance as they grow in size. Routed networks also have more flexible traffic flow because they can use the hierarchy to determine optimal paths depending on dynamic factors such as network congestion. If Ethernet rings are connected by routers instead of Ethernet switches, the size of Ethernet ring will be reduced since a router can have a larger sub-network than an Ethernet switch can have. Besides, Layer 2 devices must use a translation function to switch between dissimilar media. The translation function can result in serious problems. Layer 3 devices such as router are essentially independent of the properties of any physical media and can use a simple address resolution protocol (ARP) to resolve differences between Layer 2 and Layer 3 addresses.

Many operators have adapted Asynchronous Transfer Mode (ATM) in the network backbone. ATM and 10GbE are two quite different technologies. (a) ATM is based on a connection-oriented protocol. If a cell needs to be sent, the ATM system will determine the most expeditious route and reserve the appropriate bandwidth on the route for the duration needed to send the cell. As network traffic changes, the route is continuously reconfigured. This system gives predictable destination arrival times. 10GbE is based on connectionless protocol. This means once the transmission channel is clear, a frame will be sent down the channel with no knowledge of whether other frames might be using the channel at the same time. In this system, there are restarts, collisions and no guarantee of arrival times. (b) ATM has a multiplexing architecture which is designed to support traffic with various bandwidth and delay requirements. This allows ATM to support voice, video and data multiplexed on the same links. QoS is established when the

connection is made. In contrast, 10GbE does not offer QoS guarantee. (c) Administration of an ATM network is very difficult. Understanding all of the technology behind ATM cells and virtual circuits requires a large learning curve. Ethernet is simple and easy to learn. (d) ATM switches are more expensive than Ethernet switches. In conclusion, there is no absolutely correct answer on which system is ideal for networks of the future. All the above factors have to be considered in relation to which applications the network system is expected to run.

In this thesis, a new optical Resilient Ethernet Ring (RER) scheme which is based on the 10GbE technology is proposed in order to adapt Ethernet technology from LANs to MANs without making major changes to the conventional Ethernet technology. The proposed RER scheme is designed to operate in ring topology. The ring topology simplifies the decision-making process for frame forwarding and also enables the RER network to recover from a link or node failure within 50ms. Interconnection of RER networks is studied for greater geographical coverage and network capacity. Finally, computer simulations are carried out to examine the effectiveness of the proposed RER.

1.2 Objectives

The objectives of this research work are listed below:

- To propose a new RER scheme which is based on the 10GbE technology in order to adapt Ethernet technology from LANs to MANs
- To study inter-connections of multiple RER so as to cover a larger geographical area

- To demonstrate the feasibility and effectiveness of RER under various operating conditions by conducting a series of computer simulations

1.3 Major Contributions

The original contributions of this thesis are listed as follows:

- This thesis presents a new RER scheme which is based on the 10GbE technology. The RER scheme is the first step for 10GbE to be deployed in ring topologies and expands the Ethernet technology from LANs to MANs. This scheme will provide high speed, scalable, resilient ring based networks featuring spatial reuse and protection mechanisms (capable of sub-50ms protection switching). This thesis focuses on the medium access control (MAC) layer for the RER scheme.
- System designs of the proposed RER scheme such as network architecture, frame format, frame forwarding table etc. and system operations such as frame switching, self-learning process, failure detection and recovery etc. are analyzed in detail. A hierarchical frame forwarding scheme is proposed for the RER to simplify the process of forwarding frames from sources to destinations. Two typical types of protection schemes, the wrapping scheme and the steering scheme, are also analyzed in terms of protection switching time and packet loss. A new *wrap then steer protection scheme* is proposed and examined. The new protection scheme combines the advantages of the wrapping scheme and the steering scheme to achieve fast recovery and optimal transmitting paths after recovery simultaneously.

- A detail study of interconnection of multiple RER rings is carried out. Three different approaches based on transparent bridging technique are proposed and examined in terms of bandwidth efficiency and size of frame forwarding table. We have found that by allowing RER switches to learn MAC addresses residing in remote rings and making use of a globally unique RER switch identifier, bandwidth efficiency and small size of frame forwarding table can be achieved simultaneously. Protection issues of interconnected RER rings are also analyzed. To prevent frames from circulating among bridge ports (ports that interconnect rings), the STP needs to be applied to configure those bridge ports as a loop-free spanning tree. The infamous persistent black hole problem is possibly occurred in interconnected RER rings, we show that this problem can be solved by effectively synchronizing the forwarding table entries in ring switches with the changes of port states using a real-time forwarding table entry removal mechanism.
- Three sets of computer simulations are conducted to demonstrate the feasibility and effectiveness of the proposed RER scheme. The results of the first set show that the RER network can achieve significantly better performance on throughput and average traffic delay compared to the RSTP based network. The results of the second set prove that the wrap then steer scheme combines the advantages of the steering scheme and the wrapping scheme to achieve a fast convergence time, low traffic loss and short traffic delay after network recovery. The results of the third set illustrate

the fact that bridge nodes which interconnect RER rings can easily become bottlenecks of the network if a large amount of traffic is passed across rings.

1.4 Organization of the Thesis

The rest of the thesis is organized as follows:

Chapter 2 reviews different types of network technologies. Overviews of the dominant LAN technology – Ethernet and two MAN technologies – SONET and RPR are presented. At the end of this chapter, comparisons among 10GbE, SONET and RPR are carried out as being MAN technologies.

Chapter 3 proposes a new RER scheme which is based on the 10GbE technology and intends to adapt Ethernet technology from LANs into MANs. RER system designs and system operations will be discussed in detail.

Chapter 4 studies interconnections of multiple RER rings. Three different approaches that can be used to interconnect RER rings are proposed and examined. Protection issues of interconnected RER rings are also analyzed.

Chapter 5 examines the feasibility and effectiveness of the proposed RER scheme by conducting a series of computer simulations. Performance analysis is carried out based on the simulation results.

Finally, Chapter 6 summarizes the contents of this thesis and presents suggestions for further research.

Chapter 2 Review of Network Technologies

In this chapter, different types of network technologies are reviewed. We first present an overview of the dominant LAN technology – Ethernet. Then we review two MAN technologies – SONET and RPR. At the end of this chapter, comparisons among 10GbE, SONET and RPR are carried out as being MAN technologies.

2.1 Ethernet

Ethernet, also known as IEEE 802.3, was originally designed to work primarily in small campus networks [35-36]. Over twenty years' development, it has become the dominant technologies in LANs due to its simplicity, low cost, ease of manageability and maturity. Ethernet has become so popular today that nearly all traffic on the Internet originates or ends with an Ethernet connection. Until now, Ethernet has gone through many generations that are summarized below:

- First version of Ethernet—CSMA/CD Access Method, 1983
- Thin Ethernet—10Base2, 1985
- 10Base5 and 10Broad36, 1985
- Fiber Optic Inter-Repeater Link (FOIRL), 1987
- 10 Mbps—10Base-T Ethernet, 1990

- 100 Mbps—Fast Ethernet, 1995
- 1,000 Mbps—Gigabit Ethernet, 1998
- 10,000 Mbps—10 Gigabit Ethernet, 2002

2.1.1 Ethernet Data Frame Format

The IEEE 802.3 standard defines a basic data frame format [35] for Ethernet which contains eight fields as shown in Fig. 2.1.

Figure 2.1: Ethernet Basic Data Frame Format

The Preamble is a stream of bits used to allow the receiver to synchronize with signal. It is an alternating pattern of 56 ones and zeroes. The Start Frame Delimiter is always 1010 1011 and is used to indicate the beginning of the frame. The Destination Address is the MAC address of the network adapter to which the frame is destined. A Destination Address of all ones specifies a broadcast message. The Source address is the MAC address of the network adapter sourcing the frame. The Length/Type

indicates the length of the Ethernet frame if the value of this field is between 0 and 05DC (hex) or otherwise indicates the EtherType [37] of the frame. The Payload field is where the upper layer headers such as TCP/IP are placed if any and then followed by the actual user data. Length of this field should be between 46 bytes to 1500 bytes. If user data is longer than 1500 bytes, it will be split and carried by multiple Ethernet frames. Unused data will be padded behind if the actual user data is shorter than 46 bytes. The Frame Check Sequence is calculated using a Cyclic Redundancy Check (CRC). The FCS allows network devices to detect an error if the frame is damaged during transmission.

2.1.2 Ethernet Devices

In an Ethernet LAN, switch is commonly used to filter and forward frames between LAN segments [38-39]. Switches operate at the layer 2 (Data Link layer) of the Open Systems Interconnection (OSI) reference model. That is, it simply looks at each frame and determines from a physical or Media Access Control (MAC) address which station a data unit is intended for and switches it out toward that station. The MAC address [35] is a unique value associated with a network adapter. MAC addresses are 48-bit binary numbers. The first 24 bits of a MAC address represent the ID number of the adapter manufacturer and the next 24 bits represent the serial number assigned to the adapter by the manufacturer.

A switch learns where stations are through its continuous self-learning process. The self-learning process is simple. Once a switch receives a frame from one of its ports, it learns the station sourcing the frame can be reached from this port. This information is recorded in the forwarding table so it can be looked up later on. Hence a switch may not know locations of all stations. Only those stations that have sourced frames and any of the frames has ever been passed through the switch will be learned. To forward a frame, a switch looks up its switching table to choose an appropriate output port according to the destination address attached in the frame. In case the destination station is unknown to the switch, the frame will be broadcast via all the ports of the switch except the port the frame was received. This usually happens when the switch newly comes up.

One indication from the learning process is that loops are not allowed in a switched Ethernet network. Loops confuse the self-learning process. A switch possibly receives a single frame multiple times from different ports if loops exist in the network and the switch wonders which port it should use to reach the source station. Broadcast mechanism makes trouble in loops as well because broadcast frames will circulate in the loops endlessly and take up large amount of network capacity. Switches do not have built-in mechanism to stop such broadcast storms. Hence the current standard switches use Spanning Tree Protocol (STP) to transform any topology into a loop-free spanning tree [28-32], so as to prevent undesired loops in the network.

Router [35, 39] is another kind of popular Ethernet device. It operates at the layer 3 (Network layer) of the OSI reference model. The basic function of routers is the same as switches - to forward the received frame to the next network point toward the destination. Router performs routing function based on Internet Protocol (IP) addresses [35]. An IP address is a 32-bit number (IPv4) that identifies a network adapter on a TCP/IP network. IP addresses are hierarchical. Each IP address consists of two identifiers: an identifier of a particular network and an identifier of the particular station within that network. To move packets from one router to another along the route, only the first identifier of the address is looked at. Hierarchical address makes routers much more scalable in terms of network size compared to switches. Therefore, in wide area networks such as the Internet, routers are commonly used to interconnect networks.

2.1.3 Spanning Tree Protocol

The IEEE 802.1D Spanning Tree Protocol (STP) [28-30] is a link management protocol for media access control switches. The STP prevents undesirable loops in a network that are created by multiple active paths between network devices. Using control messages called bridge protocol data units (BPDU), switches running the STP elect a “root” switch of the spanning tree. After electing a root switch, the switches use the spanning tree algorithm (STA) to calculate loop-free paths across the entire switched network. To establish path redundancy, the STP creates a tree that spans all of the switches in an extended network, forcing redundant paths into a standby, or

blocked, state. The STP allows only one active path at a time between any two network devices (this prevents the loops) but establishes the redundant links as a backup if the initial active link should fail. In the event of a link or node failure, the STP runs to reconfigure the spanning tree. Each switch will place ports that were previously in blocked state into forwarding state where necessary, thus providing network redundancy and network reconvergence. Fig. 2.2 shows a simple example to illustrate how the STP reconfigures the network after a link cut failure (the dashed link denotes ports connected to this link are in blocked state). Originally the ports connected to the link B-C are blocked so a loop-free spanning tree is formed, as shown in Fig. 2.2(a). After a link cut failure between A and C, the STP runs to reconfigure the tree by blocking the ports connected to the link A-C and activating the ports connected to the link B-C, as shown in Fig. 2.2(b).

Figure 2.2: Spanning Tree Protocol

There are two drawbacks of using the STP. Firstly, network convergence is slow. The tree reconfiguration process is a bottom-up process for the root switch gathering all information from its leaf switches, followed by a top-down process for the root switch

to dispatch reconfiguration messages to all its leaf switches. The whole process takes from several seconds to tens of seconds depending on the network size. This slow network convergence is built into the 802.1D specification through its use of timers [28]. Secondly, network traffic is unbalanced. The links near to the root switch are easy to get saturated while rest of links will be idle or with a low utilization.

The IEEE 802.1w Rapid Spanning Tree Protocol (RSTP) [29, 31-32] is an enhancement of the STP that is designed to achieve faster convergence time. The RSTP uses a real-time handshake mechanism rather than timers to actively confirm that a port can be safely placed into forwarding state. While the legacy STP passively waits for the network to converge before turning a port into forwarding state until the timers expire. As a result, the RSTP would speed up the network convergence to between hundreds of milliseconds to 10 seconds depending on the network size. The RSTP is also backward compatible with the STP which is a helpful feature when gradually migrating STP networks to RSTP networks. However, mixed STP/RSTP networks have the slow convergence time of the STP.

2.1.4 10 Gigabit Ethernet

10 Gigabit Ethernet (IEEE 802.3ae) is the latest generation of Ethernet technologies [25-27]. This technology is very significant because not only will Ethernet run at 10Gbps and serve as a LAN connection, but it will also work in MANs and WANs. With 10GbE, network managers will be able to build LANs, MANs, and WANs using

Ethernet as the end-to-end Layer 2 transport. Compared to Gigabit Ethernet, 10GbE has some distinct characters which make it suitable to be deployed in MANs and WANs.

At Layer 2, 10GbE is much unchanged from the pre-existing Ethernet Layer 2 standards. It is still an asynchronous link protocol. It uses the same IEEE 802.3 MAC protocol, 802.3 frame format, and the same minimum and maximum frame sizes. It is also designed to run all of the familiar advanced Layer 2 Ethernet functions, such as IEEE 802.3ad link aggregation [40] and 802.1Q VLAN [41] etc. The most important change at Layer 2 is that 10GbE is the first Ethernet to fully reject any shared media applications, and operate only in full-duplex mode. In effect, this development decouples 10GbE from any transport limits other than those of the physical layer.

The major developments of 10GbE are at the Physical layer (OSI Layer 1), more specifically at the physical media dependent (PMD) sub-layer. Unlike Gigabit Ethernet, which relies on the pre-existing PMDs developed for Fiber Channel networks [35], 10GbE creates its own. In this respect, entirely new LAN and WAN PHYs (physical layers) have been defined. A LAN PHY has been defined for 10GbE to support a range of distances, from as short as 65 meters to as long as over 40 kilometers, as shown in Fig. 2.3. A WAN PHY has been defined to add to the basic LAN PHY a simplified SDH/SONET framer that allows Ethernet switches to interface with the access portions of SDH/SONET optical networks. In order to do so,

the Physical Coding Sublayer (PCS) of the WAN PHY incorporates a basic SONET/SDH frame, and some of the SONET/SDH management information. The WAN PHY interfaces will also operate at a slightly different data rate, compatible with the payload rate of SONET OC-192 and SDH VC-4-64c

PMD	Fiber mode	Minimum Distance
850 nm serial	Multi	65 m
1310 nm WDM	Multi	300 m
1310 nm WDM	Single	10 km
1310 nm serial	Single	10 km
1550 nm serial	Single	40 km

Figure 2.3: Distances supported in 10GbE

2.2 SDH/SONET Rings

SDH/SONET [14-20] defines a set of interface standards at the Physical layer (OSI layer 1) for synchronous data transmissions over fiber optic networks. SONET is the United States version of the standard published by the American National Standards Institute (ANSI). SDH is the international version of the standard published by the International Telecommunications Union (ITU). SDH/SONET is expected to provide the transport infrastructure for worldwide telecommunications. Although there are a lot of similarities between SONET and SDH, there are some significant differences, especially in terminology. In this thesis we focus primarily on SONET and SONET terminology in order to avoid getting confusing.

2.2.1 Frame Format

SONET uses a basic transmission rate of Synchronous Transport Signal level-1 or simply STS-1, equivalent to 51.84Mbps. Higher-level signals are integer multiples of the base rate. For example, STS-3 is three times the rate of STS-1 (155.52Mb/s). The frame format of the STS-1 signal is shown in Fig. 2.4.

Figure 2.4: SONET Frame Format

As can be seen, the STS frame format is composed of 9 rows of 90 columns of 8-bit bytes, or 810 bytes. The byte transmission order is row-by-row, left to right. With a constant frame length of 125 μ s (8000 frames per second), STS-1 has a bit rate of 51.84Mbps. The optical equivalent of STS-1 is known as Optical Carrier level-1 (OC-1). In general, the frame can be divided into two main areas: Transport overhead (TOH) and Synchronous Payload Envelope (SPE). The first three columns of the

STS-1 frame are for the TOH. The remaining 87 columns constitute the SPE. The TOH is composed of section overhead and line overhead. The SPE also consists of two parts: path overhead and the payload. SONET provides substantial overhead information, allowing simpler multiplexing and greatly expanded OAM&P (Operations, Administration, Maintenance, and Provisioning) capabilities. Path overhead is carried from end-to-end along entire transmitting paths. Line overhead is for the STS-N signal between STS-N multiplexers. Section overhead is used for communications between adjacent network elements, such as regenerators. The SDH/SONET standard defines a hierarchy of interface rates that allow data streams at different rates to be multiplexed. The most common rates for SONET and SDH are given in Table 2.1.

SONET signal	SDH signal	Line rate (Mbps)	SPE rate (Mbps)
STS-1, OC-1	None	51.84	50.112
STS-3, OC-3	STM-1	155.52	150.336
STS-12, OC-12	STM-4	622.08	601.344
STS-48, OC-48	STM-16	2,448.28	2,405.38
STS-192, OC-192	STM-64	9,953.28	9,621.50
STS-768, OC-768	STM-256	39,813.12	38,486.02

Table 2.1: SONET/SDH hierarchy

2.2.2 Circuit-based Network

SONET was designed to optimize for point-to-point, circuit-switched applications such as voice traffic [14-15, 42]. In circuit-switched networks, a fixed amount of bandwidth is allocated and reserved for each circuit connection. On a point-to-point circuit, contention for resources is resolved during circuit setup on a call-by-call basis.

Once the circuit is established, the full circuit bandwidth is dedicated for the connection. No congestion control scheme is needed to adjust the sending rate on the circuit. Sending rates should in fact be matched as closely as possible to the circuit rates to keep the pipe full. With dedicated point-to-point circuit bandwidth, SONET is able to provide guaranteed end-to-end packet delivery with well-controlled latency, jitter, bandwidth and packet-loss characteristics to real-time mission-critical services such as voice traffic. Fig. 2.5 illustrates the logical circuits established in a simple SONET network where each dashed line indicates a point-to-point circuit connection.

Figure 2.5: Circuit-switched SONET

Figure 2.6: Fully logical mesh for a 4-node SONET

Despite the guaranteed packet delay, latency and jitter circuit-switched networks can provide, most of the limitations of SONET stem from these fixed optical circuits. First of all, SONET provisions point-to-point circuits between ring nodes. Each circuit is allocated a fixed amount of bandwidth that is wasted when not used. For the SONET network that is shown in Fig. 2.5, every circuit connection on the ring is allocated only one third of the ring's total bandwidth. The fixed allocation of bandwidth puts a limit on the maximum burst traffic transfer rate between endpoints. This is a

disadvantage for data traffic, which is inherently bursty. Secondly, if all nodes require circuits to all other nodes, an N -node ring will have to provision totally $N*(N - 1)$ circuits. Even for moderate values of N , the total ring capacity is quickly exhausted. For example, the network design for a 4-node SONET ring calls for a fully logical mesh as shown in Fig. 2.6, the network designer must divide the ring bandwidth into 6 provisioned circuits. Therefore, provisioning the circuits necessary to create a logical mesh over a SONET ring results in extremely inefficient use of ring bandwidth. In addition, the provisioning of circuits is a difficult and lengthy process. Due to these reasons, SONET is not scalable in network size. Traditional SONET solutions can contain only up to 16 nodes. As the amount of data traffic that stays within metro networks is increasing, a fully meshed network that is easy to deploy, maintain, and upgrade is becoming an important requirement. Thirdly, multicast traffic on a SONET ring requires each source to allocate a separate circuit for each destination. A separate copy of the packet is sent to each destination. The result is multiple copies of multicast packets traveling around the ring, wasting bandwidth.

2.2.3 Self-Healing Rings

In SONET, protection with self-healing rings is called Automatic Protection Switching (APS) [14-20, 43-47] and comes in two flavors: Unidirectional Path-Switched Ring (UPSR) and Bidirectional Link-Switched Ring (BLSR).

UPSR is the simpler of the two. A sending node injects identical traffic in both fibers

of the dual counter-rotating ring. The receiving node monitors the two fibers and selects the better of the two signals based on criteria such as bit error rate and Alarm Indication Signal (AIS). Advantages of UPSR include the fact that the receiving node makes all decisions with no interaction with either local or remote transmitters. UPSR also offers the fastest recovery, often without any data loss which results in virtually uninterrupted service. The downside includes 50% bandwidth utility as well as the introduction of asymmetric delay.

BLSR is frequently used in core network applications and is more complicated. In BLSR, both fibers of the dual counter-rotating ring can be used to send data. A sending node injects only one copy of data in either fiber based on criteria such as shortest path. BLSR do not improve any bandwidth utility, however. In order to be able to provide backup, every fiber of the dual counter-rotating ring can only be used to up to half its capacity – the second half of the capacity is reserved for protection purposes. In the event of a link cut failure, the node adjacent to the failed link redirects traffic from original fiber to another fiber in the reverse direction so that traffic still can reach its destination.

The SONET standards specify that the service interruption time should not exceed 50ms, which is low enough to be unnoticeable by customers who participate in a live conversation when voice is carried over a SONET network. While service recovery with this technique meets the 50ms goal, SONET requires a substantial amount of

dedicated backup resources as half of the total link capacities are used for system protection purpose only.

2.3 RPR

Resilient Packet Ring (RPR), known as IEEE 802.17, is a network topology being developed as a new standard for fiber optic rings [21-24, 48-49]. The IEEE began the RPR standards development project in December 2000 with the intention of creating a new Media Access Control layer to meet the requirements of a packet-based metropolitan area network. RPR covers layer 1 (Physical layer) and layer 2 (Data Link layer) of the OSI Reference Model.

2.3.1 Physical Layer

At present, the RPR Working Group does not plan to define physical layers unique to 802.17 MACs. In fact, it intends to leave RPR Layer 1 open so RPR can be compatible with Ethernet, SONET, and DWDM physical layer standards. Although RPR does not have its own physical layers, it can make use some other accepted and proven physical layers. In order to do so, specific physical reconciliation sublayers need to be defined to translate between each particular physical interface and the common RPR MAC interface. The standard currently defines such sublayers for both Ethernet and SDH/SONET physical layers [23-24]. There are two reconciliation sublayers defined for SDH/SONET physical interfaces. The SDH/SONET physical reconciliation sublayer defines an HDLC-like (High-level Data Link Control)

encapsulation for RPR frames with a SONET payload and is similar to POS (packet over SONET) encapsulation. The GFP reconciliation sublayer (GRS) defines an encapsulation mechanism using Generic Framing Protocol (GFP) [50-51]. There are also two reconciliation sublayers defined for Ethernet physical interfaces. The Gigabit reconciliation sublayer defines an interface between the RPR MAC and the Gigabit Media Independent Interface (GMII). The 10 Gigabit reconciliation sublayer defines interfaces between the RPR MAC and the 10 Gigabit Media Independent Interface (XGMII) and 10 Gigabit Attachment Unit Interface (XAUI) interfaces.

2.3.2 Media Access Control Layer

A new set of MAC protocols are defined to allow RPR to provide the SONET-level protection and restoration with using bandwidth more efficiently like the Ethernet. Fig. 2.7 shows RPR base data frame format [49].

Figure 2.7: RPR base data format

The Time-To-Live is decremented at each hop to ensure packets do not circulate forever. The Base Ring Control defines sub-fields such as ringlet ID, frame type,

frame priority etc. for basic ring controls. The Destination Address is IEEE assigned 48-bit MAC address of the destination station. The Source Address is IEEE assigned 48-bit MAC address of the source station. The ttlBase is set to the initial TTL value upon transmission. The Extended Ring Control defines sub-fields such as extended data frame indication, flooding mechanism indication etc. for advanced ring controls. The Header Error Check is 16-bit Cyclic Redundancy Check (CRC) covering previous fields. The Protocol Type is the same as the Length/Type field in Ethernet frame which indicates the bytes in payload if the value of this field is between 0 and 05DC (hex) or otherwise indicates the IEEE Type Field Register [37]. The Payload contains transmitted data. The Frame Check Sequence is 32-bit CRC covering the Payload.

RPR is data optimized networks, with at least two counter-rotating fiber ringlets in which multiple nodes share the bandwidth without the requirement of provisioning circuits. The nodes on the ring can automatically negotiate for bandwidth among themselves via a Fairness Algorithm [22-23, 52-54]. Each station has a topology map of the ring and can send data on the optimal ringlet towards its destination. Both ringlets can be used to carry working traffic. Similar to SONET, RPR is able to provide sub-50ms system protection and restoration against fiber or node failures but without reserving half of the ring capacity for this purpose.

PRP MAC defines a fairness protocol to ensure a fair distribution of available bandwidth across all stations on a ring even during times when the ring is heavily utilized or congested. The RPR fairness algorithm is a local fairness mechanism that dynamically throttles traffic to enforce fairness among the stations on the ring. The RPR algorithm includes the following functions – (a) determining when the congestion threshold is crossed and when the congestion has subsided; (b) determining the fair rate for advertisement; (c) determining the station's allowed rate; (d) sourcing and consuming fairness messages; (e) communicating the allowed rate to the traffic shapers for controlling access to the medium. In a congested network, a congested downstream node conveys its congestion state to upstream nodes such that they will throttle their traffic and ensure that there is sufficient spare capacity for the downstream station traffic. To achieve this, the congested downstream node transmits fairness messages containing its local fair rate upstream. The fairness algorithm adjusts the allowed sending rates of upstream nodes. Each node has a traffic shaper to ensure that the actual sending rate complies with the allowed sending rate. After a convergence period, congestion is alleviated once all nodes' rates are set to the fair rate. Likewise, when congestion clears, stations periodically increase their sending rates to ensure that they are receiving their maximal bandwidth share.

Although RPR is optimized for data traffic, it also defines a minimum of intermediate circuit layer to support legacy voice services. In order to do so, the RPR MAC supports three service classes, i.e. Class A, Class B and Class C. Class A has high

priority. Services of class A are Committed Information Rate (CIR) services. This service supports guaranteed bandwidth and low latency/jitter applications. Voice, video, and circuit emulation applications can utilize this class of traffic. Class B has medium priority. Services of class B below the provisioned bandwidth are CIR services that have less stringent (but still bounded) jitter/latency requirements. Services of class B above the provisioned bandwidth are considered Excess Information Rate (EIR). Business data applications can utilize the medium class of traffic. Class C has low priority. Services of class C are best-effort services. Consumer Internet access is an application for this class of traffic. Class A and Class B-CIR services are not controlled by the fairness algorithm. Traffic shapers ensure bandwidth for these services up to their provisioned values. However, Class B-EIR and Class C services are controlled by fairness algorithm. Traffic shapers shape the bandwidth for these services according to the allowed fair rate determined by the fairness algorithm [55].

2.4 Next-Generation MANs

Next-generation MANs should be able to meet the following requirements. First of all, they should be able to scale to bandwidths of tens of Gigabits per second (Gbps) per link and large numbers of nodes. Secondly, they should be low cost per Gbps, low installation cost, and low operational costs. Thirdly, they should support differentiated services such as extensive Quality of Service (QoS) and Service Level Agreement (SLA) controls. Fourthly, they should be robust enough to provide redundant

hardware and sub-50ms system protection and restoration against fiber or node failure.

Last but not least, they should be optimized for IP transport. Up to 95% of all traffic in the metro area is carrying by Internet Protocol (IP). IP will be the dominant type of traffic transported in next-generation MANs. Optimizing for IP transport allows flexible creation and support of new services, and minimizes equipment and operating costs.

Table 2.2 shows the comparison of SONET, RPR and 10GbE as a next-generation MAN technology.

Feature	SDH/SONET	RPR	10GbE
Support ring topology	Yes	Yes	No
Sub-50ms system protection	Yes	Yes	No
Differentiated data services	Yes	Yes	No
Carrier-class management	Yes	Yes	No
Fast network provision	No	Yes	Yes
Efficient transport of IP data	No	Yes	Yes
Bandwidth efficiency	No	Yes	Yes
Access network technology	No	No	Yes
Simplicity	No	No	Yes
Low cost	No	No	Yes
Maturity	Yes	No	Yes

Table 2.2: Comparison of SONET, RPR and 10GbE as a next-generation MAN technology

SDH/SONET dominates current-generation MANs. It offers a number of benefits for both service providers and end users such as carrier-class management and high level of resiliency. However, there are also some well-known drawbacks to SDH/SONET ring networks. Cost is a major concern, particularly in the enterprise sector. SDH/SONET is 10 times more expensive than Ethernet technology being used today. What's more, SDH/SONET is not good at transporting data traffic. Furthermore, provisioning for a large SDH/SONET ring network could be a difficult and lengthy process. Lastly, SDH/SONET does not use bandwidth efficiently. RPR is emerging as a new technology to provide the SONET-level protection and restoration with using bandwidth more efficiently like the Ethernet. Its fairness algorithm ensures a fair distribution of available bandwidth across all stations on a ring even during times when the ring is heavily utilized or congested. However, RPR has its own limitations. Although RPR equipment may offer reliable transport solutions, it does not have built-in support for access services. To date, there are relatively few RPR chip suppliers hence RPR equipment remains expensive compared to SONET and Ethernet equipment. Lastly, since RPR includes some advanced operations in the MAC protocol such as fairness algorithm, congestion control, traffic shaping and support of intermediate circuit layers etc., it could become a complex technology. Though proposed four years ago, RPR has not been completely standardized yet partly because of the complexity of its protocol.

10GbE can be a flexible and cost-effective candidate for the next-generation MANs [56]. First of all, the LAN PHY of 10GbE allows 10GbE networks to extend over much longer distances than campus. Running over single-mode fiber at 1,550-nm, 10GbE links can extend to beyond 40km. The WAN PHY allows 10GbE to operate at a slightly different data rate, compatible with the payload rate of SONET OC-192 and SDH VC-4-64c. In other words, 10GbE is interoperatable with current existing SDH/SONET networks. Secondly, as Ethernet is the dominant access network technology, 10GbE based MAN is able to provide seamlessly connections with current existing LANs. Thirdly, the simple and cost-effective 10GbE helps carriers better manage costs. Because it is a simpler technology, it is easier to troubleshoot and manage. Lastly, Ethernet technology is tested and well-known, which makes 10GbE user-friendly and shortens network operators' learning curve.

As shown in Table 2.2, the major limitations for 10GbE to become a MAN technology will be – (a) it does not support ring topology; (b) it does not provide sub-50ms system protection and restoration against fiber or node failures; (c) it does not support differentiated services; (d) it does not provide carrier-class management. A new Resilient Ethernet Ring (RER) scheme, which is based on the 10GbE technology, will be proposed in Chapter 3 to address some of these limitations. The RER scheme is designed to allow 10GbE technology to be deployed in next-generation MANs.

Chapter 3 Optical Resilient Ethernet Ring

In the previous chapter, we have reviewed the dominant LAN technology – Ethernet, and two MAN technologies – SONET and RPR. In this chapter, a new optical Resilient Ethernet Ring (RER) scheme is proposed. This scheme is based on standard 10GbE (802.3ae) technology to allow Ethernet to operate in ring topologies for data transport in metropolitan area networks (MANs). RER system designs such as OSI layer model, system architecture, MAC frame structure etc. and system operations such as frame switching, failure detection and recovery etc. will be discussed in detail in this chapter.

3.1 RER OSI Layer Model

The proposed RER scheme is based on the standard 10GbE (802.3ae) technology. It occupies layer 1 and layer 2 in the OSI layers as shown in Fig. 3.1.

Figure 3.1: RER OSI layer model

At Layer 2, 10GbE is essentially unchanged from the pre-existing Ethernet Layer 2 standards, as described earlier at Section 2.1.4. The most important change at Layer 2 is that 10GbE is the first Ethernet to fully reject any shared media applications, and operates only in full-duplex mode. In effect, this development decouples 10GbE from any transport limits other than those of the physical layer. The proposed RER scheme has exactly the same features as 10GbE technology at layer 2 except that minor modifications will be made in the MAC protocol so that RER becomes more suitable to work in the MAN environment. As described later, these modifications include new mechanisms for frame switching, topology discovery and protection.

The major developments of 10GbE are at Layer 1, as described earlier in Section 2.1.4. 10GbE defines a LAN PHY to support a range of distances, from as short as 65 meters to as long as over 40 kilometers. It also defines a WAN PHY to add to the basic LAN PHY a simplified SDH/SONET framer that allows Ethernet switches to interface with the access portions of a SDH/SONET optical network. The WAN PHY interfaces will also operate at a slightly different data rate, compatible with the payload rate of OC-192c and SDH VC-4-64c. RER has exactly the same features as 10GbE technology at layer 1.

3.2 System Architecture

The architecture of RER is shown in Fig. 3.2. It consists of two counter-directional rings as shown in Fig. 3.2(a), where each ring node represents a 10GbE switch. RER serves as a 10GbE MAN that interconnects various locations. At each location, the 10GbE capable switch multiplexes the traffic from one or more hosts and/or sub-networks onto the RER. RER can also be connected to WANs through routers.

(a) Ring structure

(b) Node structure

Figure 3.2: Architecture of the RER

Other than the proposed ring topology, there are some other alternative topologies available in Ethernet based MANs. These alternatives include star, tree and mesh. Ring topology has advantages over other alternatives, as stated earlier in Chapter 1. Some network developers suggest that RSTP should be used in Ethernet based MANs instead of STP, because the convergence of STP is too slow. If RSTP is applied in a ring topology, the ring will be configured as a straight line with the root of the tree in the center. This violates the ring nature and hence incurs problems such as long mean

transmitting paths and unbalanced network traffic. Simulations have been conducted to show the performance difference between true rings and RSTP based rings (see Section 5.1 for more details). Comparison is also carried out based on the simulation results. In this project, RER will be designed to operate in a true ring topology.

The structure of an individual ring node is shown in Fig. 3.2(b). Each switch has at least three input ports and three output ports. These ports are grouped in pairs, with each pair consisting of one input port and one output port. Out of these pairs, two participate in ring connection and are referred to as *Ring Pair 1* and *Ring Pair 2*. The remaining pairs, called *Local Pairs* are used to load traffic onto the ring from local hosts or sub-networks or to deliver ring traffic to local hosts or sub-networks.

A ring switch may be involved in one of three roles during a transmission – (a) as a *source switch*, (b) as a *destination switch* or (c) as an *intermediate switch*. A source switch is the entrance point where frames from local hosts/sub-networks enter the RER. A destination switch is the exit point where frames leave the RER. Intermediate switches lie in the middle and relay frames between the source and the destination switches. A ring switch is also referred to as the *boundary switch* of a host if this is the ring switch which connects the host to the RER.

3.3 Switch ID

To facilitate forwarding of frames from source switches to destination switches, each

ring switch is assigned an 8-bit integer as a Switch ID to represent itself. This is locally unique and is used to identify an individual switch. Switch ID with value FF is reserved for the broadcast purpose. A frame with destination switch ID FF indicates the frame should be broadcast. This allows a total node space of 255 in a single RER network, which is reasonably large for a MAN. The *Switch ID Assignment Process* (see Section 3.7 for more details) runs automatically to assign a Switch ID to each ring switch when the RER network starts up. The globally unique 48-bit IEEE address assigned in each switch can also be used to identify a switch, which is the way taken by the RPR approach [49]. Our approach for the RER results in less frame overhead and simplifies frame forwarding; however, switch IDs need to be re-assigned whenever a ring switch insertion or removal takes place (see Section 3.7 for more details).

3.4 Frame Forwarding Table

To simplify the process of forwarding frames from source hosts to destination hosts, we present a hierarchical frame forwarding scheme here. In this scheme, the transmission path of a frame, from the point where it enters the ring to reaching the destination host, may be divided into two parts – (a) from the source switch to the destination switch and (b) from the destination switch to the destination host. Correspondingly every ring switch contains a frame forwarding table which also consists of two parts - (a) *Sub-forwarding Table 1* and (b) *Sub-forwarding Table 2*. Sub-forwarding table 1 is used to forward frames from source switches to destination

switches (the first part of transmission path), whereas sub-forwarding table 2 is employed to forward frames from destination (boundary) switches to destination hosts (the second part of transmission path). Sub-forwarding table 1 records all the hosts and their corresponding boundary switches within an RER network. Hence it contains two entries - (a) a host MAC address and (b) the boundary ring switch ID of that host. Source switch uses sub-forwarding table 1 to determine the destination switch of a received frame. Sub-forwarding table 2 also contains two entries - (a) a host MAC address and (b) a corresponding egress port of the boundary switch. After a frame arrives at its destination switch, the destination switch uses sub-forwarding table 2 to determine the egress port via which the frame can reach its destination host. Sub-forwarding table 1 is common to all ring switches within an RER network, while sub-forwarding table 2 is dedicated to individual ring switches. Forwarding tables are built up through a self-learning process conducted in each ring switch (more details can be found in Section 3.8). Details on the usage of sub-forwarding tables are presented in Section 3.9.

Fig. 3.3 gives an example of how switch forwarding tables are configured. Fig. 3.3(a) shows a simple ring network where Hosts a1 and a2 are connected to Switch A, and Hosts g1, g2 and g3 are connected to Switch G. For simplicity, a1_mac_addr is used to represent the MAC address of Host a1. A_SID is used to represent the ID of Switch A. Fig. 3.3(b) shows the forwarding table contained in Switch A and Fig. 3.3(c) shows the forwarding tables contained in Switch G, respectively.

(a) A Simple ring network

Sub-forwarding Table 1		Sub-forwarding Table 2	
Host MAC Address	Switch ID	Host MAC Address	Egress Port
a1_mac_addr	A_SID	a1_mac_addr	port-to-a1
a2_mac_addr	A_SID	a2_mac_addr	port-to-a2
g1_mac_addr	G_SID		
g2_mac_addr	G_SID		
g3_mac_addr	G_SID		

(b) Forwarding table in Switch A

Sub-forwarding Table 1		Sub-forwarding table 2	
Host MAC Address	Switch ID	Host MAC address	Egress Port
a1_mac_addr	A_SID	g1_mac_addr	port-to-g1
a2_mac_addr	A_SID	g2_mac_addr	port-to-g2
g1_mac_addr	G_SID	g3_mac_addr	port-to-g1
g2_mac_addr	G_SID		
g3_mac_addr	G_SID		

(c) Forwarding table in Switch G

Figure 3.3: Example of frames forwarding tables

3.5 Ring Image

As described in previous section, RER employs a hierarchical frame forwarding scheme. Source switch uses sub-forwarding table 1 to determine the destination switch of a received frame in the first part of transmission path. Besides, it also needs to decide in which direction the frame should be forwarded to the destination switch since an RER network consists of two counter-directional rings. To facilitate the decision making, all ring switches store a copy of the ring image. The ring image refers to the collection of information on the number and arrangement of all ring switches in the RER network. Source switches will determine the optimal direction to be used for forwarding data frames according to the information contained in the ring image.

3.6 MAC Frame Structure

Frames that are transmitted within sub-networks (a sub-network may be a LAN or just a host) of the RER remain the same as the conventional Ethernet frame format as shown in Fig. 3.4(a). Conventional Ethernet frame format was designed for frame transmissions within LANs in a tree topology. In order to efficiently forward frames from one LAN to another via the RER ring, we propose to insert an RER header in the conventional Ethernet frame when a frame enters the RER. This RER header is removed at the destination switch before the frame leaves the RER and will therefore be transparent to the end-users. All other fields from the conventional Ethernet frame remain unchanged, as shown in Fig. 3.4(b). The RER header contains 5 fields. These

are Destination Switch ID (DSID) field, Source Switch ID (SSID) field, Utility (U) field, Header Error Check (HEC) field and Time-To-Live (TTL) field. Fig. 3.4(c) shows the details of an RER header. The destination switch ID and the source switch ID are both locally unique 8-bit integers representing the destination switch and the source switch respectively. The Utility field is an 8-bit field that is further divided into a Type field, a Wrap field, a Priority field and a Reserve field as shown in Fig. 3.4(d). The Type field is a 2-bit field used to identify the type of the frame. Table 3.1 shows the various frame types defined in the RER. The Wrap field is a 1-bit field used to indicate whether the frame has been wrapped due to system failure (see Section 3.11 for more details). The Priority field is a 1-bit field used to indicate the priority level of the frame. The remaining 4 bits in the utility field are reserved for future use. Header error check (HEC) is an 8-bit checksum computed over the DSID, SSID and U fields only. When a switch receives a frame from its neighbor switch, it will regenerate the HEC based on the DSID, SSID and U fields of the received frame to confirm if the RER header is valid or not. The frame will be discarded if its RER header is not valid. The time-to-live is an 8-bit field that must be decremented every time the frame is forwarded to the next ring switch. The frame is stripped off the ring when its TTL value reaches zero. The TTL field is necessary to prevent frames from circulating in the ring. The TTL field is kept outside the HEC protection. This expedites frame forwarding, as otherwise the HEC has to be re-written at every intermediate switch.

Preamble	SFD	DA	SA	L/T	Data	FCS
----------	-----	----	----	-----	------	-----

SFD: Start Frame Delimiter

L/T: Length/Type

DA: Destination Address

FCS: Frame Check Sequence

SA: Source Address

(a) Conventional Ethernet frame format

(b) RER frame format

DSID: Destination Switch ID

U: Utility

SSID: Source Switch ID

TTL: Time-to-Live

HEC: Header Error Check

(c) RER header

(d) Utility field

Figure 3.4: MAC frame structure

Value (bin)	Type Description
00	Data frame
01	Keep-alive frame
10	Image discovery frame
11	OAM control frame

Table 3.1: Types defined in the utility field

3.7 Switch ID Assignment Process

Assignment of switch IDs (SID) is done automatically as shown in the figure below:

Figure 3.5: Switch ID assignment process

Step 1: When an RER network starts up, all ring switches agree to designate the switch with the lowest MAC address to be the first ring switch, denoted as S0. In the example above, Switch A is selected to be S0.

Step 2: After S0 is selected, the other SIDs may be assigned sequentially along either ring direction. In the example above, SID is assigned sequentially along counter-clock direction. Switch C, Switch D and Switch B are assigned SID 1, 2, 3 respectively.

Assigning SIDs sequentially simplifies the decision-making process in a source switch to determine the direction in which to forward frames. For a ring containing m nodes, hop counts from switch A to switch K will be simply calculated as $(k-a)$ in one direction and $(m-k+a+1)$ in the opposite direction where a and k represent the SIDs of switches A and K, respectively, with $k > a$. Following a shortest-path first rule, the

direction with the smaller hop count will be chosen as the actual direction of transmission.

When a new switch is inserted between $S(n)$ and $S(n+1)$, $S(n)$ and $S(n+1)$ detect the switch as a new neighbor by receiving its keep-alive frame with source MAC address different from that of the previously observed neighbor (see Section 3.11 for more details). The detection of a new switch on the ring triggers the ring image discovery process so that every switch is aware of the new switch, and the ring image is updated. The newly inserted switch will be assigned an SID $S(n+1)$ and the other switches with SID greater than $S(n)$ will have their SIDs increased by 1, e.g. $S(n+1)$ becomes $S(n+2)$, $S(n+2)$ becomes $S(n+3)$ and so on. The sub-forwarding table 1 stored in every ring switch also needs to be updated accordingly. In case a switch is removed, this process is simply reversed. Note that the MAN topology is not expected to change frequently. Therefore, the operations required to handle ring switch insertion or removal are not expected to happen too often in MANs.

3.8 Self-Learning Process

All Ethernet switches have self-learning capability to learn how to reach a specific host from the frames they receive [38-39]. This self-learning process helps to build up forwarding tables so that the information learnt can be reused later to forward frames. Similarly in the RER, a ring switch also builds up its sub-forwarding tables 1 and 2 through a continuous self-learning process.

- When a ring switch receives a frame from its sub-network, it learns that the source host is available on the receiving port by recording the SA field in the frame and the receiving port. This learning process helps to build up sub-forwarding table 2.

This self-learning process is illustrated in the figure below:

Figure 3.6 (a): Self-learning process in source switch

- When a ring switch receives a frame from its neighboring ring switch, it learns the hosts in other ring switches by observing (SSID, SA) pairs in the received frame. This learning process helps to build up sub-forwarding table 1. This process is illustrated in the figure below:

Figure 3.6 (b): Self-learning process in intermediate switch

Entries in sub-forwarding tables 1 and 2 are kept in a cached manner and will be eventually removed either if the size of forwarding tables exceeds the switch capacity

or if too much time has elapsed since the last update for a particular entry in the tables.

3.9 Frame Switching

Frames can be transmitted in either direction of the ring in the RER based on the shortest-path rule. The RER is flexible in bandwidth assignment since there is no fixed bandwidth allocated between any two end points. The spatial reuse mechanism is also used to conserve ring bandwidth in the RER by allowing the destination switch to strip a received frame off the ring.

As stated in Section 3.4, the transmission path of a frame, from the point where it enters the ring to reaching the destination host, may be divided into two parts – (a) from the source switch to the destination switch and (b) from the destination switch to the destination host. Ring image and sub-forwarding table 1 are involved in the first part. Sub-forwarding table 1 is used to find out the destination switch based on the DA field in the received frame. Ring image is used to determine the direction in which the frame should be forwarded to the destination switch. Sub-forwarding table 2 is involved in the second part to determine the egress port to be used by the destination switch to forward the frame to the destination host.

A source switch inserts an RER header in every frame arriving from its sub-network before it is transmitted into the RER. The operations carried out at a source switch are

summarized below:

- 1) Regenerate the FCS and compare it with the original FCS value in the received frame. The frame is discarded if they do not match.
- 2) Update its sub-forwarding table 2 by recording the SA field and the port receiving the frame in the switch as part of the self-learning process.
- 3) Look up its sub-forwarding table 1 to determine the DSID based on the DA field.
- 4) Select a direction to forward the frame based on the ring image.
- 5) Set the type of frame.
- 6) Set the Wrap field to 0 to indicate that the frame has not been wrapped.
- 7) Set the priority field to 0 for a regular frame or 1 for a high-priority frame.
- 8) Calculate HEC over the DSID, SSID and Utility fields.
- 9) Set the TTL field to be equal to the number of ring switches in the ring.
- 10) Insert the RER header (including DSID, SSID, Utility, HEC and TTL fields) into the frame.

When a ring switch receives a transit frame from its neighbor, it has the following three options – it may (a) discard the frame, (b) accept the frame or (c) forward the frame. If the frame is invalid, it is discarded. If the frame is valid and the ring switch itself is the destination, the frame is accepted. Otherwise, a valid frame is forwarded in the same direction in which the frame was received. The ring topology simplifies the decision making process for forwarding frames because intermediate switches can simply follow the forwarding direction chosen by the source switch (i.e. the direction

along which the frame was received) instead of checking its forwarding table to make its own forwarding decisions. The following operations take place at intermediate switches:

- 1) Decrement TTL. The frame is discarded if it reaches zero.
- 2) Regenerate HEC and compare it with the original HEC value in the received frame. The frame is discarded if they do not match.
- 3) Update its sub-forwarding table 1 by recording (SSID, SA) in the switch as part of its self-learning process.
- 4) Compare the DSID field with its switch ID. If they do not match, the frame will be forwarded.

Note that only the HEC will be regenerated to confirm the validity of the received frame. The FCS will not be regenerated in intermediate switches in order to save processing time and hence expedite frame forwarding. The reason for keeping the TTL field outside HEC protection is because, in that case, the HEC field does not have to be re-written at every intermediate switch.

The following operations take place at a destination switch:

- 1) Decrement TTL. The frame is discarded if it reaches zero.
- 2) Regenerate HEC and compare it with the original HEC value in the received frame. The frame is discarded if they do not match.

- 3) Update its sub-forwarding table 1 by recording (SSID, SA) in the switch as part of its self-learning process.
- 4) Compare the DSID field with its switch ID. If they match, the frame will be accepted.
- 5) Remove the RER header from the frame.
- 6) Strip the frame off the ring if it is a unicasting frame. Multicast or broadcast frames are not stripped and continue to flow in the ring.
- 7) Check sub-forwarding table 2 to determine which egress port can be used to reach the destination host.

3.10 Prevention of Frame Circulation

Without the use of the TTL field, frames would continue to circulate in the ring forever, if transmitted frames fail to get stripped off properly from the ring. Unicast frames are normally stripped by the destination switch. These may also be stripped by the source switch if the destination switch fails to do the stripping. Multicast and broadcast frames are only stripped by the source switch. If both the source and destination switches drop out of the ring while a unicast frame is in flight, or if the source switch drops out while its multicast or broadcast frame is in flight, the frame might circulate around the ring endlessly. These problems are solved by including a TTL field in each frame. Each node forwarding the frame decrements the TTL field and discards the frame whenever it reaches zero. By setting the TTL field to an initial value corresponding to the number of switches in the ring, we can ensure that the

frame will reach each switch but will be limited to a transit of only one complete rotation around the ring.

3.11 Failure Detection and Recovery

As designed to be a MAN technology, RER should provide network resiliency to detect a network fault such as a link or node failure and recover the network from the failure rapidly. This section discusses failure detection and protection schemes for the RER.

3.11.1 Failure Detection

Dual-ring networks have built-in natural resiliency because there are always two paths between any two ring switches during normal operations. These two paths are in opposite directions and will not usually be of the same length in terms of hop-counts. Normally, frames are switched to follow the shorter path of the two. In case of a link or node failure, affected frames will be directed to the alternative path to avoid service disruption.

Failure detection is achieved through periodical exchanges of keep-alive frames between ring switches. Keep-alive frames are control messages containing non-data payload. Each ring switch keeps monitoring the health of its two neighbors. Receiving keep-alive frames from a neighbor indicates that both the neighbor and the link to the neighbor are still functioning. The frequency of sending keep-alive frames can be

configured based on operating requirements. A high frequency implies the switch can detect loss of connectivity with a neighbor rapidly. A keep-alive frame sent once every 5 millisecond is proposed. Keep-alive frames are only sent to the neighbors (TTL=1). RER keeps a timer for each of its neighbors to count the time elapsed since the reception of last keep-alive frame. Timer expires if no keep-alive frame is received from that link for three consecutive keep-alive frame sending periods and the link to this neighbor is considered to be down.

Upon receiving a keep-alive frame, a ring switch first performs ring switch ID consistency check by comparing the SSID attached in the frame with its own ring image. If the SSID does not agree with the ring image, ring switch ID consistency check fails and this triggers the ring image discovery process. Otherwise, the ring switch resets its timer for that sending neighbor.

The failure detection process can be summarized in the figure below:

Figure 3.7: Failure detection process

3.11.2 Protection Schemes

Borrowing from the self-healing ring idea for SDH/SONET rings, two protection schemes, i.e. wrapping and steering, have been proposed for the RPR [15]. These two schemes are also applicable to the RER with minor modifications. Unlike SDH/SONET reserves half of the ring capacity to provide system protection, RER does not reserve any bandwidth for protection purpose.

Steering protection scheme:

When the steering protection scheme is applied to the RER, the two switches adjacent

to the failure (referred to as edge switches in the following discussion) that detect a link or node failure are required to report this topology change to other switches by initiating a ring image discovery process. The ring image in every ring switch then gets updated. After all ring images in ring switches have converged to a new set, the source ring switches forwards frames based on the new set of ring image. Fig. 3.8 shows the ring re-configuration after a dual-fiber cut and single-fiber cut failures when the steering scheme is in use.

Figure 3.8: Steering scheme

An example is shown in Fig. 3.9. Switch S1 sends frames to switch S5 via the path S1-S2-S3-S4-S5 in the normal situation, as illustrated in Fig. 3.9(a). After a link failure occurs between S4 and S5, S1 redirects the frames via path S1-S0-S9-S8-S7-S6-S5 (6 hops), as illustrated in Fig. 3.9(b). Before the completion of the ring image discovery process, frames destined to a switch beyond the point of failure are dropped at the edge switches.

Figure 3.9: Example of fiber-cut failure

We examine here the protection switching time and traffic dropped due to fiber cut failures. The switching time denotes the time period from the time the failure occurs to the time the network is recovered. Protection switching time for the steering scheme can be calculated as:

$$t_{switching} = t_{detect} + t_{discovery} + t_{steer} \quad (3-1)$$

where t_{detect} denotes the time to detect the failure. In RER, t_{detect} is set to three consecutive keep-alive frame sending periods. After failure is detected, switches that

detect the failure report this topology change to other switches by initiating a ring image discovery process. $t_{discovery}$ denotes the time to complete the ring image discovery process, t_{steer} denotes how fast a ring switch can switch itself to the re-steering mode after completion of the ring image discovery process. Normally t_{steer} is very small compared to t_{detect} or $t_{discovery}$ and hence negligible. This equation is applicable for both dual-fiber cut and single-fiber cut failures.

The amount of traffic dropped due to fiber cut failures can be calculated as:

$$\begin{aligned} D &= \lambda * t_{switching} + D_{buffer} \\ &= \lambda * (t_{detect} + t_{discovery} + t_{steer}) + D_{buffer} \end{aligned} \quad (3-2)$$

The total traffic dropped D contains two parts. The first part is $\lambda * t_{switching}$ which represents the traffic dropped at the two edge switches during the protection switching time where λ denotes the line rate. After the protection switching time, there might still be some traffic destined to destination switches beyond the point of failure that has been put in the ring switch transit buffers and has not yet reached the point of failure. This traffic will eventually reach the point of failure and be dropped there (the second part of traffic dropped). D_{buffer} represents this amount of traffic dropped. This equation is applicable for transmissions along both A-B and B-A directions (see Fig. 3.8) in dual-fiber cut failures. For single-fiber cut failures, this equation is only applicable for transmissions along the A-B direction (see Fig. 3.8). There will be no drop of frames for transmissions along B-A direction. As can be seen in (3-2), traffic loss in the steering scheme is affected by the ring size. A larger ring size incurs a

longer $t_{discovery}$, and hence leads to a larger amount of traffic dropped.

Wrapping protection scheme:

With the wrapping protection scheme, a link failure causes frames to be wrapped (looped back) at the two edge switches. Fig. 3.10 shows the ring re-configuration after a dual-fiber/single-fiber cut failure when the wrapping scheme is in use.

Figure 3.10: Wrapping scheme

Using the same example as in the steering scheme, S1 sends the traffic via path S1-S2-S3-S4-S3-S2-S1-S0-S9-S8-S7-S6-S5 (12 hops) after a link failure between S4 and S5, as illustrated in Fig. 3.9(c). In the wrapping scheme, only the edge switches are involved while the other switches are not aware of the failure. Hence this scheme is simple and very fast. Fast recovery time ensures that very few frames are dropped and consequently few re-transmissions are required. However as shown in the example, frames may have to be forwarded in a much longer path as compared to the situation where the steering scheme is used (12 nodes vs. 6 nodes). Under the wrapping scheme, the edge switches need to rewrite the TTL field to the number of switches in the ring before the frame is wrapped in order to maintain the same node

space as originally supported. Edge switches also need to rewrite the Wrap field to 1 indicating that the frame has been wrapped. Multicast or broadcast frames can be stripped by source switches if the Wrap field value is 0 or by the edge switches if the Wrap field value is 1.

We here again examine protection switching time and traffic dropped due to fiber cut failures. Protection switching time for the wrapping scheme can be simply calculated as:

$$t_{switching} = t_{detect} + t_{wrap} \quad (3-3)$$

where t_{detect} again denotes the time to detect the failure which is set to three consecutive keep-alive frame sending periods in RER. t_{wrap} denotes how fast an edge switch can switch itself to the wrapping mode after detecting the failure. Normally t_{wrap} is very small compared to t_{detect} and hence negligible. This equation is applicable for both dual-fiber cut and single-fiber cut failures.

The amount of traffic due to fiber cut failures can be calculated as:

$$D = \lambda * t_{switching} = \lambda * (t_{detect} + t_{wrap}) \quad (3-4)$$

where λ again denotes the line rate. Again as in the steering scheme, there might still be some traffic destined to destination switches beyond the point of failure that has been put in the ring switch transit buffers and has not yet reached the point of failure after the switching time. This traffic will eventually reach the edge switches and be wrapped there (instead of being dropped in the steering scheme). This equation is

applicable for transmissions along both A-B and B-A directions (see Fig. 3.10) in dual-fiber cut failures. For single-fiber cut failures, this equation is only applicable for transmissions along the A-B direction (see Fig. 3.10). There will be no drop of frames for transmission along B-A direction.

Hybrid protection scheme – wrap then steer scheme:

As described above, either the wrapping scheme or the steering scheme can be used to recover the RER network from a link/node failure. The wrapping scheme is faster but transmitting paths after wrapping may not be optimal. The steering scheme is slower but transmitting paths after steering are optimal. Since there are advantages and disadvantages for each of the two schemes, we proposed a new protection scheme, which is a combination of the two, and it is called *wrapping then steering scheme*. In this scheme, the nodes detecting the failure firstly perform wrapping as soon as possible to ensure the system is rapidly recovered. They then inform other nodes to start a new ring image discovery. After the new ring image has been discovered, all source ring switches forwards frames according to the new ring image so frames can be transmitted in the optimal paths. This new scheme combines advantages of both wrapping and steering schemes. The drawback is that it makes the engineering of the network a bit more complicated, compared to the situation only the wrapping scheme or the steering scheme is used. Simulation has been done to show the performance of these three protection schemes. Simulation results and analysis are presented in Section 5.2.

3.11.3 Summary for Protection Schemes

Although any of the wrapping, steering and wrap then steer protection schemes can be selected when an RER network is deployed, all ring switches must converge to a single scheme. Ring image needs to be updated in every ring switch in the steering scheme as well as the wrap then steer scheme, while it does not need to be updated in the wrapping scheme. Since all protection schemes are transparent to higher-layer services such as IP, no noticeable interruption in service is expected to occur in any case. Unlike SONET/SDH rings, the self-healing RER architecture does not require to set aside half of the ring capacity for protection purpose. This allows ring bandwidth to be used more efficiently. However, care needs to be taken when failure occurs in the RER. Traffic originally flowing in two ring directions may add up and flow in one ring direction after failure. This aggregate traffic may possibly exceed the link bandwidth capacity and result in dropping of frames.

3.12 RER Switch Design

The macro-view of an RER switch is shown in Fig. 3.2(b). The switch contains at least two pairs of 10Gb ports for ring connection and at least one pair of 10Gb/1000Mb ports to aggregate traffic from sub-networks to the RER. The 1000Mb port supports 10Mbps/100Mbps with Media Independent Interface (MII) and 1000Mbps with Gigabit Media Independent Interface (GMII) [35]. The 10Gb port supports 10Gbps with the 10 Gigabit Media Independent Interface (XGMII) [25].

Predominantly shared buffer architecture is chosen for gigabit Ethernet switches [57-59]. Shared buffer architecture gives many advantages over the distributed buffer architecture [60]. The RER switch is also designed to be shared buffer architecture. The micro-view of an RER switch is illustrated in the following two functional diagrams. One is for handling sub-network traffic and the other is for handling ring traffic.

Figure 3.11: RER switch functional diagram for handling sub-network traffic

Fig. 3.11 shows the RER switch functional diagram for handling sub-network traffic.

The media access controller implements the IEEE 802.3 standards including packet

delineation, flow control generation and response, preamble and inter-frame gap regeneration, and cyclic redundancy check (CRC) checking and generation. The parser will extract the layer 2 MAC addresses including source MAC address and destination MAC address. In the searching engine, sub-forwarding table 1 will be searched for a match with destination MAC address. If it is a hit, destination switch ID will be returned. Ring image will be then resorted to determine which direction to forward the frame based on the returned destination switch ID. Otherwise, the frame will be broadcast to both ring output ports. The CRC calculator simply calculates the checksum of the RER header. Finally, the inserter will insert the RER header into the received frame. In the mean time, the switch self-learning process is running to learn the (source_MAC_address, receiving_port) pair and update sub-forwarding table 2 if needed. The queue manager allocates frames into different output queues based on the priority of the frames.

Figure 3.12: RER switch functional diagram for handling ring traffic

Fig. 3.12 shows the RER switch functional diagram for handling ring traffic. This functional diagram shares some common functional blocks such as the media access controller, parser and queue manager as shown in Fig. 3.11. After the layer 2 MAC addresses are extracted by the parser, the destination comparator compares the frame destination switch ID with its own ID. If they don't match, the received frame will be put to the output queues directly. If they match, sub-forwarding table 2 will be searched in the searching engine for a match with destination MAC address. If it is a hit, an output port will be returned. Otherwise, the frame will be broadcast to all sub-network output ports. Before forwarding the frame to the sub-network, the

remover will remove the RER header from the received frame. In the mean time, the self-learning process runs to learn the (source_MAC_address, source_switch ID) pair and update sub-forwarding table 1 if needed.

The design of RER switches is quite different from conventional gigabit Ethernet switches. First of all, conventional Ethernet switches use a single functional diagram for all incoming frames [59], i.e. all frames are processed in the same way. While in RER switches, frames coming from ring neighbors and sub-networks are processed differently. Furthermore, conventional Ethernet switches search forwarding tables to make forwarding decisions for all frames received. While RER switches don't have to do so for frames coming from ring neighbors and the destination switch ID doesn't match its own ID.

For 10GbE, the shortest time interval between any two arrivals of frames can be calculated as:

$$t = \frac{l_{\min frame} + l_{preamble} + l_{IFG}}{\lambda} = \frac{64Bytes + 8Bytes + 12Bytes}{10Gbits/s} = 67.2ns$$

where $l_{\min frame}$ denotes the minimum frame length which is 64-byte, $l_{preamble}$ denotes the length of Preamble field which is 8-byte, l_{IFG} denotes the length of inter-frame gap which is 12-byte. In another words, for non-blocking operations, a 10Gbps switch has to process any frame within 67.2ns, which includes lookup on destination and source addresses, frame forwarding and address learning. Among all contributors to the switch chip latency, the most time-consuming operations are hash calculation and

memory access [59].

Hashing is used to do the multiple-to-one mapping to search a large table efficiently.

The forwarding table is organized in terms of buckets with an index corresponding to the result of the hash function. As the bucket size is much smaller than the table size, the lookup is done much faster. In the RER, ring image is a small table containing up to several hundreds of entries. For such small tables, searching is easily done even without involving any hash functions. Several hash functions for address lookup in moderate tables have been discussed in [61]. Among them, CRC polynomials were proved to be excellent hashing functions. In RER, sub-forwarding table 2 is a small-moderate table containing up to several thousands of entries and can be searched using these hashing functions. However as MANs covers a much larger geographical area than LANs, the number of hosts in a single MAN could be huge, possibly over 10 thousands. This means, sub-forwarding table 1 in RER may contain over 10 thousands of entries. In fact, 10GbE switches such as Summit 400 Series [62] from Extreme Networks Inc. and Cisco Catalyst 6500 Series from Cisco Systems Inc. [63] support a forwarding table containing up to 16k MAC addresses. Such large table would require either a more complex hash function or a multi-level search, both of which would result in more processing. As can be seen, ring image, sub-forwarding table 1 and sub-forwarding table 2 can be searched using different hashing functions according to the table size. Another option is to use a single hashing function to search ring image to maintain software simplicity. This dedicated hashing function should be able to handle efficient search in a table containing up to 16k entries.

Memory access is another time-consuming operation. To minimize switch chip latency, memory access has to be minimized. Conventional Ethernet switch forwards a frame to a proper output interface based on the destination MAC address. At the same time the switch must also learn the association between source MAC address and ingress interface in order to build up forwarding tables for future forwarding. Hence each frame must search the forwarding table twice, once for destination lookup and once for source lookup. If the source lookup is a miss, a third memory access is needed to learn the new source MAC address. Hence every conventional Ethernet switch requires 2 or 3 memory accesses to process a frame.

In RER switches, the number of memory accesses varies among source switch, intermediate switch and destination switch. In source switches, frames are received from one of its sub-network. One memory access is required for destination lookup in sub-forwarding table 1 to determine the destination switch. If it is a hit, another memory access is required to search the ring image to decide which direction to forward the frame. Self-learning requires a memory access for source lookup in sub-forwarding table 2. If it is a miss, another memory access is required to update sub-forwarding table 2. Therefore, a total of 2 or 3 or 4 memory accesses are required in source switches. In intermediate switches, there is no destination lookup because no forwarding decision is required to be made. Self-learning process required a memory access for source lookup in sub-forwarding table 1. If it is a miss, a second memory

access is required to update sub-forwarding table 1. Therefore, a total of 1 or 2 memory accesses are required in intermediate switches. In destination switches, frames are dispatched to their sub-networks. One memory access is required for destination lookup in sub-forwarding table 2 to determine which output port a frame should be forwarded to. Self-learning process requires a source lookup in sub-forwarding table 1. Again if it is a miss, a third memory access is required to update sub-forwarding table 1. Therefore, a total of 2 or 3 memory accesses are required in destination switches.

We conclude from above that to process a data frame, RER source switches, on the average, require more memory accesses than conventional Ethernet switches; RER intermediate switches, on the average, require less memory accesses than conventional Ethernet switches; RER destination switches, on the average, require the same memory accesses as conventional Ethernet switches. This gives an advantage to RER switches because there is only one source switch during a transmission while there can be many intermediate switches. Overall, RER switches require less memory accesses than conventional Ethernet switches to process a frame.

3.13 Summary

A new optical Resilient Ethernet Ring (RER) scheme is proposed in this chapter. This scheme is based on standard 10GbE (802.3ae) technology to allow Ethernet to operate in ring topologies for data transport in MANs.

We have described in detail the topics including RER OSI layer model, system architecture, switch ID, frame forwarding table, ring image, MAC frame structure, switch ID assignment process, self-learning process, frame switching, prevention of frame circulation, failure detection and recovery, and finally RER switch design. The proposed RER supports true ring topology. It also supports flexible bandwidth allocation and spatial reuse mechanism. We have proposed a hierarchical frame forwarding scheme for ring topology to simplify the process of forwarding frames from sources to destinations. Based on the hierarchical frame forwarding scheme, the forwarding table in an RER node also consists of two corresponding parts – (a) sub-forwarding table 1 and (b) sub-forwarding table 2. The simplicity of ring topology speeds up frames switching in the RER because once a forwarding decision has been made when the frame enters the RER, other RER nodes receiving the frame can simply follow the forwarding decision instead of making their own. In order to efficiently forward frames in RER, we have proposed to insert a 5-byte RER header in the conventional Ethernet frame when a frame enters the ring. The RER header will be removed when the frame leaves RER so it will be transparent to the end-users. RER is able to detect a link or node failure quickly and recover the network from failure within 50ms by continuously monitoring keep-alive frames from neighbors. The wrapping scheme or the steering scheme can be used to recover RER from a link or node failure. We have proposed a new scheme called wrap then steer scheme which is the combination of the wrapping scheme and the steering scheme. The new scheme

combines the advantages of the two schemes to achieve both fast recovery and optimal transmitting paths after recovery. Finally, we analyzed the RER switch design. Functional blocks are presented specifically for RER switches. The two most time-consuming switch operations, i.e. hashing operation and memory access, are also discussed.

Chapter 4 Interconnection of Multiple RERs

In the previous chapter, we have proposed a new RER scheme that is based on the 10GbE and suitable to be deployed in MANs. In practice, multiple rings can be expected in MANs. These rings are required to be interconnected to form a unified network for greater geographical coverage and network capacity. The general issues for interconnection of multiple RER rings will be how the RER MAC handles transmissions where the destination addresses are not in the domain of the RER local ring address space and how interconnected RER rings is protected against network failures. This chapter examines three different approaches to interconnect multiple RER rings and protection issues of interconnected RER rings.

4.1 Approaches for Interconnecting RER Rings

IEEE 802.1D transparent bridging technique [28, 64-65] used in the original Ethernet may also be applied to interconnect multiple RER rings. With this technique, frames with unknown destinations are broadcast and bridges relay broadcast frames from one ring to another. Fig. 4.1 and Fig. 4.2 show the network reference and bridging architecture of interconnecting two RER rings respectively. We here examine three different approaches for interconnecting RER rings.

Figure 4.1: Transparent bridge network reference

Figure 4.2: Transparent bridge architecture

4.1.1 Interconnecting RER Rings - First approach

Fig. 4.3 illustrates the most basic transparent bridging technique to interconnect two RER rings. The RER frame format remains unchanged when a frame passes through transparent bridges. The entire network is assumed to be newly up so all forwarding tables are empty. The self-learning process on frames from remote rings is disabled in order to keep forwarding tables small.

Figure 4.3 (a): First approach for interconnection of multiple RER rings - 1

Fig. 4.3(a) shows a scenario where Host a sends a frame to Host c using the first approach. The figure in the middle represents the RER network. The bold numbers inside the network denote the transmitting steps (described later). The rows on the top represent simplified frame formats showing only address fields where c denotes the MAC address of Host c (DA), a denotes the MAC address of Host a (SA), A denotes the locally unique switch ID of Switch A (SSID), FF indicates the frame is broadcast (DSID). The tables at the bottom represent the forwarding tables contained in the specific ring node after the transmission from Host a to Host c is completed. For example, $N1_FT$ denotes the forwarding table contained in Bridge N1. All figures in the following discussions use the same notations.

The transmitting steps are described as follows:

1. Host a in RER1 starts sending a frame addressed to Host c in RER3.
2. Switch A receives the frame. It inserts an RER header in the frame and broadcasts the frame over RER1 because the destination MAC address is unknown.

3. Bridge N1 receives the frame and broadcasts it over RER2. Self-learning is performed at Bridge N1.
4. Bridge N2 receives the frame and broadcasts it over RER3. Self-learning is not performed at Bridge N2 because the frame is from remote rings.
5. Switch C receives the frame and redirects it to its sub-networks. Finally the frame reaches the destination Host c.

When Host c replies with a frame to Host a, as shown in Fig. 4.3(b), the transmitting steps are simply reversed. The frame is broadcast over RER3 and RER2 and then unicast (instead of being broadcast) in RER1 because Bridge N1 has learned the destination MAC address in the previous transmission.

Figure 4.3 (b): First approach for interconnection of multiple RER rings - 2

All subsequent transmissions from Host a to Host c result in frames being broadcast over RER1, RER2 and being unicast in RER3. All subsequent transmissions from Host c to Host a result in frames being broadcast over RER3, RER2 and being unicast in RER1.

In this approach, forwarding tables are kept small because switches are allowed to learn MAC addresses only within its local ring. As a result, the burden of searching and computing in each ring switch or transparent bridge is light. However, the penalty is poor bandwidth utilization because frames originating at one RER and terminating at another RER are broadcast.

4.1.2 Interconnecting RER Rings - Second approach

To improve bandwidth utilization, broadcast should be avoided as much as possible. This can be done by allowing ring switches to learn MAC addresses in both local and remote rings (instead of disabling learning of MAC addresses in remote rings in the first approach). This approach is illustrated in Fig. 4.4. The entire network is again assumed to be newly up so all forwarding tables are empty.

Figure 4.4 (a): Second approach for interconnection of multiple RER rings - 1

Concepts of current destination switch and current source switch are introduced here which represent the exit and entrance of current ring respectively. For example, Host a sends a frame to Host c in the network as shown in Fig. 4.4(a). The real destination

switch is Switch C and the real source switch is Switch A. When the frame is transmitting in RER2, the current destination switch will be Bridge N2 because N2 is the exit of current ring (RER2) and the current source switch will be N1 because N1 is the entrance of current ring (RER2). However when the frame is transmitting in RER3, the current destination switch will be Switch C, which is the same as the real destination switch, and the current source switch will be N2 because N2 is the entrance of current ring (RER3). Real DSID and real SSID denote the switch IDs of real destination switch and real source switch respectively. Similarly, current DSID and current SSID denote the switch IDs of current destination switch and current source switch respectively. Since the ring switch ID is a locally unique 8-bit integer (see Section 3.3 for more details), the current DSID and current SSID are used in frame transmissions across rings instead of the real DSID and SSID.

The transmitting steps are described as follows:

1. Host a in RER1 starts sending a frame addressed to Host c in RER3.
2. Switch A receives the frame. It includes an RER header in the frame and broadcasts the frame over RER1 since the destination MAC address is unknown.
3. Bridge N1 receives the frame. It replaces the SSID in the frame with its own SSID (current SSID) and then broadcasts the modified frame over RER2. Self-learning is performed at Bridge N1.

4. Bridge N2 receives the frame. It replaces the SSID in the frame with its own SSID (current SSID) and then broadcasts the modified frame over RER3. Self-learning is performed at Bridge N2.
5. Switch C receives the frame and redirects it to its sub-networks. Self-learning is performed in Switch C. Finally the frame reaches the destination Host c.

When Host c replies with a frame to Host a, as shown in Fig. 4.4(b), switch C unicasts (instead of broadcasting as in the first approach) the frame directly to Bridge N2 because switch C has learned the MAC address of Host a. For the same reason, Bridge N2 unicasts the frame to Bridge N1; Bridge N1 unicasts the frame to Switch A and finally the frame reaches Host a. When the frame passes through Bridge N2, Bridge N1 and Switch A, the MAC address of Host c is learned.

Figure 4.4 (b): Second approach for interconnection of multiple RER rings - 2

All subsequent transmissions between Host a and Host c result in frames being unicast in RER1, RER2 and RER3. No broadcast of frame is required.

In this approach, bandwidth efficiency is maintained over RER rings because broadcast of frame is effectively avoided. However, the forwarding tables in ring switches, and transparent bridges become large because the MAC address in every frame (no matter where it comes from) needs to be learned when the frame is passing through. As a result, this increases the burden of searching and computing in ring switches and transparent bridges. In this case, the transparent bridges may easily become the bottleneck of the network if there are large amount of transmissions across RER rings. To address this problem, more powerful searching and computing capabilities would be required for the transparent bridges.

4.1.3 Interconnecting RER Rings - Third approach

The first two approaches described are unable to simultaneously achieve the two desirable features of high bandwidth efficiency and small forwarding tables. A compromise approach is to make use of a globally unique ID to represent a ring switch. This can be done by assigning each ring with a ring ID. The combination of ring ID and switch ID is globally unique and can be used to distinguish a ring switch in multiple rings. Fig. 4.5 illustrates this approach. The entire network is again assumed to be newly up so that all forwarding tables are empty.

Figure 4.5 (a): Third approach for interconnection of multiple RER rings - 1

In Fig. 4.5(a), A represents a globally unique ID to represent Switch A which is a combination of a ring ID and a locally unique 8-bit switch ID. As can be seen, not only the current DSID and current SSID (the first two columns of a simplified frame format) but also the real DSID and real SSID (the next two columns of a simplified frame format) appear in the RER frame format. In the first two approaches, bridge nodes perform self-learning on the MAC addresses of source hosts. However in this approach, with globally unique real SSIDs, bridge nodes can perform self-learning on the real SSIDs. This helps to reduce the size of forwarding tables in bridge nodes. In this approach, forwarding tables in ring switches (not in bridge nodes) are in the format (host_mac_addr, real_DSID, current_DSID). For example, as shown in Fig. 4.5(a), one entry in *C_FT* is (a, A, N2) which represents that the real DSID for Host a is A, and the current DSID for Host a is N2. Forwarding table format in ring switches are more complex compared to the first two approaches. Forwarding tables in bridge nodes have the format (SID, current_DSID). For example, one entry in Bridge N2 is (A, N1) which represents that the current DSID for Switch A is N1.

The transmitting steps are described below:

1. *Host a* in RER1 starts sending a frame addressed to *host c* in RER3.
2. Switch A receives the frame. It includes an RER header in the frame and broadcasts the frame over RER1 because of unknown destination MAC address.
3. Bridge N1 receives the frame. It replaces the current_SSID with its own SSID and then broadcasts the modified frame over RER2. Self-learning is performed at Bridge N1.
4. Bridge N2 receives the frame. It replaces the current_SSID with its own SSID and then broadcasts the modified frame over RER3. Self-learning is performed at Bridge N2.
5. Switch C receives the frame and redirects it to its sub-networks. Finally the frame reaches the destination *host c*.

When Host *c* replies a frame to Host *a*, as shown in 4.5(b), Switch C unicasts the frame to Bridge N2 because Switch C has learned the MAC address of Host *a*. For the same reason, Bridge N2 unicasts the frame to Bridge N1; Bridge N1 unicasts the frame to Switch A and finally the frame reaches Host *a*. When the frame passes through Bridge N2, Bridge N1 and Switch A, the MAC address of Host *c* is learned.

Figure 4.5 (b): Third approach for interconnection of multiple RER rings - 2

All subsequent transmissions between Host a and Host c result in frames being unicast in RER1, RER2 and RER3. No broadcast of frame is required.

It may be noted that, in this approach, the transparent bridge forwarding tables are indexed by the globally unique ring switch IDs instead of the host MAC addresses that are used in the first two approaches. This effectively reduces the size of the forwarding tables in the bridge nodes so that the searching and computing burdens in the bridge nodes are reduced. At the same time, bandwidth efficiency is maintained because broadcast is effectively avoided. However, this approach does have some drawbacks. It adds complexity in both frame header and forwarding tables in ring switches. More importantly, it is hard to assign the ring ID automatically. This will probably have to be done manually; however, that would mean that RER would no longer have the “plug and play” feature. One possible solution to avoid ring ID assignment is to use the 48-bit MAC address of ring switch as a globally unique ring

switch ID. However, this will require longer overhead in frame headers.

The bridges may also be replaced with routers which would then perform layer 3 routing based on destination IP addresses. The routing functions are invoked only when frames need to go across from one ring to another ring. This section compares bridging and routing with regard to the following network design issues:

- **Loops:** To prevent looping and replication, topologies that may contain loops need to run the Spanning-Tree Protocol to construct topologies that do not contain any loops (see Section 4.2 for more details). Because the spanning-tree algorithm places certain ports in blocking mode, only a subset of the network topology is used for forwarding data. In contrast, routers provide freedom from loops so optimal paths can be used.
- **Convergence:** In transparent switching, Spanning-Tree Protocol needs to be applied to prevent looping and replication. However, the convergence of re-constructing a new tree could be slow (see Section 4.2 for more details). In a routed environment, sophisticated routing protocols, such as Open Shortest Path First (OSPF) and Enhanced Interior Gateway Routing Protocol (Enhanced IGRP), maintain concurrent topological databases of the network and allow the network to converge quickly.
- **Media Dependence:** Generally, bridging mechanism can be used to interconnect similar LANs. To interconnect dissimilar LANs, routing mechanism should be utilized.
- **Sub-networking:** Layer 2 addressing assumes a flat address space with

universally unique addresses. Therefore, bridging mechanism is not suitable for a large network with many terminal devices. Routers operate at Layer 3, so can formulate and adhere to a hierarchical addressing structure. Routers can build their networks that maintain performance as they grow in size. Routed networks have more flexible traffic flow because they can use the hierarchy to determine optimal paths depending on dynamic factors such as network congestion.

4.2 Protection Issues of Interconnected RER Rings

Multi-ring interconnection can induce broadcast storm if bridge ports form a closed loop. One broadcast storm scenario is illustrated in Fig. 4.6.

Figure 4.6: Broadcast storm scenario due to interconnection of multiple RER rings

Suppose a frame is broadcast over RER1 in the network as shown in Fig. 4.6. Possibly the broadcast frame will be relayed to RER2 and RER3 and comes back to RER1 via Bridge N3 or Bridge N1. If no specific action is taken, this may lead to endless frame circulation causing a broadcast storm. To avoid such broadcast storms due to multi-ring interconnections, the STP/RSTP needs to be applied on bridge ports to

form a loop-free spanning tree.

Redundancy needs to be provided to protect the multi-ring system against any network failure. Fig. 4.7 shows the simplest way to interconnect two RER rings. Connectivity between the two RER rings is not protected. Single-attachment becomes a single point of failure.

Figure 4.7: Single-attachment to interconnect multiple RER rings

Dual-attachment provides network redundancy. As shown in Fig. 4.8, the connectivity between the two RERs is protected.

Figure 4.8: Dual attachment to interconnect multiple RER rings

As discussed above, the STP/RSTP needs to be resorted and applied in bridge ports to avoid broadcast storm in multi-ring interconnections. With the STP/RSTP, only one link (either Bridge1 or Bridge2) can be active at a time. If the active link fails, the

other link will become active. However in such networks, persistent black hole problem may occur. This problem is illustrated in Fig. 4.9.

Figure 4.9: Persistent black hole scenario in interconnection of multiple RER rings

At the steady state, the link between S2 and S4 is active. The forwarding table in ring switch S1 associates Host b's MAC address with ring switch S2. In RER1, frames addressed to Host b will be forwarded directly to S2, as shown in Fig. 4.9 (a). Later due to some changes in network parameters such as link state, the STP may be involved to re-construct a new tree. After the STP converges, the link between S2 and S4 moves from forwarding state to blocked state while the link between S3 and S5 moves from blocked state to forwarding state. If S1 does not realize this change of bridge port states, it continues to forward frames addressed to Host b directly to S2, resulting in continuous frame dropping in Bridge1, as shown in Fig. 4.9 (b). Persistent black hole occurs. Similar persistent black hole exists in RER2 for traffic from Host b

to Host a.

To avoid the persistent black hole in this scenario, forwarding table entries in ring switches need to be synchronized with the changes of bridge port states. Using the example illustrated in Fig. 4.9, this means forwarding table entries in S1 need to be updated accordingly to reflect the change from forwarding state to blocked state in Bridge1 after network reconstruction. One possible way is to remove the forwarding table entry in S1 which associate Host b's MAC address with S2. After the removal of this entry, S1 broadcasts every frame addressed to Host b over RER1. The broadcast would cease once the association between Host b's MAC address and S3 is re-learned. After that, S1 forwards all frames addressed to Host b directly to S3. Ring switch forwarding table has a built-in timer to age table entries periodically. However the aging period is typically several minutes. Large amount of traffic could have been dropped before the table entries are aged. A much quicker removal mechanism can be realized by triggering a removal of table entries based on the reception of Bridge Protocol Data Units (BPDUs) which are sent by the STP/RSTP during tree re-construction.

4.3 Summary

In this chapter, issues for interconnection of multiple RER rings are discussed. Three different approaches to interconnect multiple RER rings have been presented and compared. In the first approach, ring switches are restricted to learn MAC addresses

only within their local rings. This results in a small forwarding table but poor bandwidth utilization. In the second and third approach, ring switches are allowed to learn MAC addresses in both local and remote rings. The second approach uses a locally unique 8-bit switch ID to identify a ring switch. As a result, bandwidth efficiency is realized but forwarding tables become large. The third approach makes use of a globally unique switch ID to identify a ring switch. Proposed as a compromise to the first and the second approach, this approach maintains bandwidth efficiency and also keeps forwarding tables in transparent bridges small. However, this approach increases complexity in both frame header and forwarding table which makes the implementation of the network more complicated.

Protection issues are also discussed in this chapter. The STP/RSTP needs to be applied in bridge ports to prevent frames from circulating among them. Doubly bridged links could be used other than single bridged link to interconnect RER rings in order to avoid single point of failure. However, the use of doubly bridged link together with the STP/RSTP could incur the infamous persistent black hole problem. To solve the problem, synchronization of the forwarding table entries in ring switches with the changes of bridge port states needs to be ensured. This can be done by a real-time forwarding table entry removal mechanism triggering by the reception of BPDUs sent by the STP/RSTP during tree re-construction.

Chapter 5 Simulations and Performance Analysis

In this chapter, we report on the simulations conducted to examine the performance of the proposed RER and the performance analysis carried out based on these simulation results. Three sets of simulations are presented. The first simulation is to compare the performance difference among three different ring approaches, i.e. RER, RSTP and RPR. The second simulation compares the performance of three different protection schemes, i.e. the wrapping scheme, the steering scheme and the wrap then steer scheme. Finally, the third simulation examines the performance of two interconnected RER rings. NS2 (Network Simulator version 2) [66] is the simulation tool used in this project to conduct the simulations. The NS2 simulator is open source software that can be downloaded from website [67].

5.1 Simulation on Three Different Ring Approaches

The objectives of this simulation are (a) study the performance difference between the proposed RER and one of the MAN technologies RPR which are both employed dual-ring topology; (b) study the performance difference between dual-ring topology and tree topology in MANs. We have conducted simulations to study the performance differences among three different ring approaches, i.e. RSTP, RPR and RER. Three

networks employing ring topology have been simulated. RSTP, RPR and RER protocols are applied in the same physical ring network. Note that RSTP will configure the ring topology as a tree by blocking certain ports of a ring switch so as to prevent frames from circulating in a closed loop. Root node lies in the center of the RSTP network. Fig. 5.1 shows the network topology after applying the three protocols.

Figure 5.1: Simulation topology for different ring approaches

For each network simulated, the link rate was set to 10Gbps and each link between two adjacent ring switches was set to 5km. We assume that every ring switch has two buffer stores, i.e. transit buffer and transmit buffer. Transit buffers are used to receive/deliver frames from/to neighboring ring nodes, while transmit buffers are used to load/deliver frames from/to sub-networks. In our simulation study, the transit and transmit buffers in each ring switch were set at 50KB and 100KB, respectively. Frame arrivals follow a Poisson process and frame sizes follow a uniform distribution from 64 bytes to 1518 bytes (both inclusive). Frames are randomly addressed to various destinations with equal probability. We assume that all traffic is unicast UDP traffic of the same priority. No network failures occur during simulations. Ring switches have

learned all host MAC addresses in the network. The same simulation settings are applied in all the following simulations unless otherwise specified.

When frames enter the RPR network from subnets, RPR overhead of 18 bytes (according to the latest RPR draft [49]) will be added to the original frames. Likewise, an RER header of 5 bytes will be added to the original frames when the frames enter the RER network.

In this thesis, the normalized offered traffic is defined as the total bit rate generated by all the nodes divided by the link speed. The normalized carried traffic is defined as the total bit rate successfully transmitted by all the nodes divided by the link speed. The average delay is defined as the average time span required to move a frame, i.e. from entering the ring to leaving the ring. The throughput is defined as the ratio of the normalized carried traffic to the normalized offered traffic.

In the following discussion, we will show the results for this set of simulations. Firstly, the number of nodes is fixed at 10 and the offered load is gradually increased. Fig. 5.2 shows the average delay performance versus the offered load for three different ring networks. In the absence of congestion, the most important component of the delay is the mean distance between end nodes. The propagation delay between adjacent nodes is $25\mu\text{s}$ since each link is 5km. In the RSTP network, the average delay experienced by frames generated at the root node and at the edge node (the farthest from the root

node) is quite different. We denote RSTP Root and RSTP Edge to measure the average delay of frames sourcing by the root node and the edge node of the RSTP network, respectively. The RSTP Root shows the best average delay performance while RSTP Edge shows the worst in the RSTP network. As shown in the figure, the RSTP Edge exhibits high delay ($\sim 128\mu\text{s}$) even under low offered load while RSTP Root, RPR and RER exhibits much lower delay ($\sim 70\mu\text{s}$). As the offered load increases, all the three networks show increasing average delay because of queuing of frames in switch buffers. RSTP starts queuing frames at 1.2 (normalized offered traffic load) which is much lower than RPR (~ 2.2) and RER (~ 2.4). As can be seen from Fig. 5.3, RER achieves highest throughput among the three networks. RER saturates at a higher offered load (~ 3.7), while RSTP saturates at the much lower (~ 2.7). Secondly, the access traffic rate at each node is fixed to 2.5Gbps and the number of nodes is gradually increased. The result is illustrated in Fig. 5.4. As can be seen, throughput in all the three networks decreases as the number of nodes increases. RER again performs best and RSTP performs worst.

Overall, the poor performance of the RSTP network is mainly due to the following two reasons. First, the RSTP network has largest mean distance between end nodes. The transmitting paths in RSTP are always larger than or equal to the ones in the RPR or RER network because of the open-loop structure of the RSTP network. Second, traffic becomes unbalanced in the RSTP network also due to its open-loop structure, even though uniform traffic is assumed during the simulations. As a result, the root

switch gets saturated quickly and degrades the network performance. The performance differences between the RER and RPR are primarily because of the longer header overhead in the RPR. In practice, RPR requires more complicated implementation than the RER. The RER is shown to be the best solution among these three networks. As in the case of the RPR, the RER ensures smaller mean distance between ring switches by employing the dual ring architecture. The RER protocol is simpler than the RPR protocol resulting in less bandwidth consumption due to frame overheads. In addition, the RER uses bandwidth effectively and has an efficient switching mechanism. As a conclusion, the proposed RER has better performance compared to the MAN technology RPR because of the protocol simplicity of RER. Ring topology achieves better performance compared to tree topology in MANs.

Figure 5.2: Average delay vs. normalized offered traffic for different ring networks with 10 nodes

Figure 5.3: Normalized carried traffic vs. normalized offered traffic for different ring networks with 10 nodes

Figure 5.4: Throughput vs. number of nodes for different ring networks with a fixed access traffic rate at each node 2.5Gbps

5.2 Simulation on Three Different Protection Schemes

As discussed in Section 3.11, the protection schemes that may be applied to RER are the steering scheme, the wrapping scheme and the wrap then steer scheme (denoted as the wts scheme in the following discussion). The objectives of this simulation are to study the performance differences among the three different protection schemes under various offered load.

This set of simulations used the same simulation settings as the previous set of simulations (see Section 5.1 for more details) except that fiber cut failures were scheduled to occur in this set of simulations. We assume that the traffic was started at $t = 0.15\text{sec}$ and that a fiber cut failure between switch5 and switch6 occurred at $t = 0.2\text{sec}$. In addition, the period of sending keep-alive messages was set to 5ms.

The offered traffic is defined as the total bit rate generated by all the nodes. The carried traffic is defined as the total bit rate successfully transmitted by all the nodes. The traffic loss is defined as the total bit rate dropped by all ring nodes.

We have simulated two scenarios for the three different protection schemes. In the first scenario, the offered load was light. In the second scenario, the offered load was moderate.

5.2.1 Scenario 1: Light Offered Load

In this scenario, access traffic rate at each node was set to 1Gbps. Double-fiber and single-fiber cut failures were simulated. The performance of the three protection schemes are studied and compared.

Fig. 5.5 and Fig. 5.6 show the carried traffic under the double-fiber cut failure and the single-fiber cut failure respectively. Traffic is started at $t = 0.15\text{sec}$. The carried traffic is approximately 10Gbps before the fiber cut failure occurs ($0.15\text{sec} < t < 0.2\text{sec}$), which is equal to the offered traffic ($10 \times 1\text{Gbps}$). The carried traffic drops sharply when the failure occurs ($t = 0.2\text{sec}$). After the failure is detected and the network recovers, the carried traffic goes back to 10Gbps which is the same level as before the failure. The wrapping scheme and the wrap then steer (wts) scheme both recover the network from failure at around $t = 0.215\text{sec}$. The steering scheme recovers the network from failure at around $t = 0.23\text{sec}$. As can be seen, the wrapping scheme and the wts scheme both recover the network from failure rapidly ($\sim 15\text{ms}$), while the steering scheme takes a longer time ($\sim 30\text{ms}$) to do so. This can be explained by using Eqn. (3-1) and Eqn. (3-3) (see Section 3.12.2 for more details). According to Eqn. (3-3), the protection switching time for the wrapping scheme is calculated as $t_{\text{switching}} = t_{\text{detect}} + t_{\text{wrap}}$ where t_{detect} (time to detect a failure) is equal to three consecutive keep-alive frame sending periods which is 15ms ($3 \times 5\text{ms}$) in this case, t_{wrap} (time to switch to the wrapping mode) is generally much smaller compared to

t_{detect} and hence negligible. According to Eqn. (3-1), the protection switching time for the steering scheme is calculated as $t_{switching} = t_{detect} + t_{discovery} + t_{steer}$ where t_{detect} (time to detect a failure) is again 15ms as in the wrapping scheme. However, $t_{discovery}$ (time to perform a complete ring image discovery process) is not negligible. As a result, the steering scheme takes a longer time to recover a network from failure compared to the wrapping scheme and the wts scheme.

Fig. 5.7 and Fig. 5.8 show the traffic loss under the double-fiber cut failure and the single-fiber cut failure respectively. The area under each curve represents the total traffic loss if the specific protection scheme is used. Although the instantaneous traffic loss in each protection scheme is approximately the same (~2Gbps/~1Gbps, for the double-fiber/single-fiber cut failure respectively), the protection switching time for the steering scheme is longer (~30ms) compared to the wrapping scheme and the wts scheme (both ~15ms). As a result, the steering scheme incurs a larger total traffic loss compared to the wrapping scheme and the wts scheme. Another observation is that the traffic loss during the single-fiber cut failure (~1Gbps) is approximately half of the traffic loss during the double-fiber cut failure. The reason is that under a single-fiber cut failure only traffic flowing in the specific ring direction is affected, while traffic flowing in both ring directions is affected under a double-fiber cut failure.

Table 5.1 shows the average traffic delay under the double-fiber/single-fiber cut failure. In the absence of congestion (light offered load), the most important

component of the delay is the mean distance between end nodes. The network without any failure achieves the lowest average traffic delay ($\sim 127\mu\text{s}$) because every transmitting path is optimal, i.e. shortest mean distance between end nodes. Both the steering scheme and the wts scheme have a shorter average traffic delay ($\sim 153\mu\text{s}/\sim 142\mu\text{s}$, for the double-fiber/single-fiber cut failure respectively) compared to the wrapping scheme ($\sim 223\mu\text{s}/\sim 176\mu\text{s}$, for the double-fiber/single-fiber cut failure respectively). This is because the affected transmitting paths after network recovery are sub-optimal for the steering scheme, and non-optimal for the wrapping scheme. It may be noted that the values for the wts scheme here are taken from the steady state where the steering scheme is dominating. Before reaching the steady state, the wts scheme went through a transient state where the wrapping scheme was dominating. In the transient state, the average traffic delay is long ($\sim 223\mu\text{s}/\sim 176\mu\text{s}$, for the double-fiber/single-fiber cut failure respectively). The average traffic delay under a single-fiber cut failure is shorter than under a double-fiber cut failure because fewer transmitting paths get affected under a single-fiber cut failure.

Conditions	Average delay (μs)
No_failure	127
Doublefiber_cut_failure_wrapping	223
Singlefiber_cut_failure_wrapping	176
Doublefiber_cut_failure_steering	153
Singlefiber_cut_failure_steering	142
Doublefiber_cut_failure_wts	153
Singlefiber_cut_failure_wts	142

Table 5.1: Average traffic delay under double-fiber/single-fiber cut failure with access traffic rate at each node 1Gbps

Figure 5.5: Carried traffic under double-fiber cut failure with access traffic rate at each node 1Gbps

Figure 5.6: Carried traffic under single-fiber cut failure with access traffic rate at each node 1Gbps

Figure 5.7: Traffic loss under double-fiber cut failure with access traffic rate at each node 1Gbps

Figure 5.8: Traffic loss under single-fiber cut failure with access traffic rate at each node 1Gbps

5.2.2 Scenario 2: Moderate Offered Load

In this scenario, access traffic rate at each node was set to 2Gbps. Double-fiber and single-fiber cut failures were simulated. The performance of the three protection schemes are studied and compared.

Fig. 5.9 and Fig. 5.10 show the carried traffic under the double-fiber cut failure and the single-fiber cut failure respectively. Compared to the light offered load scenario (Fig. 5.5 and Fig. 5.6), the protection switching time is approximately the same. One obvious difference is that in this moderate offered load scenario the carried traffic is lower than the offered traffic after the network recovers from failure, while in the light offered load scenario the carried traffic is equal to offered traffic. As shown in the two figures, the carried traffic after the network recovers from failure is ~18.3Gbps/~19.2Gbps (for the double-fiber/single-fiber cut failure respectively) for the wrapping scheme, ~19.4Gbps/~19.8Gbps for the steering scheme, and ~19.4/~19.8Gbps for the wts scheme which is less than the offered traffic (10*2Gbps).

The offered/carried traffic difference can be explained by Fig. 5.11 and Fig. 5.12 which show the traffic loss under the double-fiber cut failure and the single-fiber cut failure respectively. Compared to the light load scenario (Fig. 5.7 and Fig. 5.8), the traffic loss in the network is no longer zero for each of the protection scheme after the network recovers from failure. The non-zero traffic loss may be due to the fact that

traffic originally flowing in two ring directions adds up and flow in one ring direction after the failure. This aggregate traffic possibly exceeds the link bandwidth capacity and results in dropping of frames. The traffic loss after network recovery is highest for the wrapping scheme ($\sim 1.7\text{Gbps}/\sim 0.8\text{Gbps}$, for the double-fiber/single-fiber cut failure respectively) because the affected transmitting paths are non-optimal after recovery resulting in heavy congestion in the network. The traffic loss after network recovery is lowest for the steering scheme ($\sim 0.6\text{Gbps}/\sim 0.2\text{Gbps}$) because the affected transmitting paths are sub-optimal after recovery resulting in light congestion in the network. For the wts scheme, the traffic loss after network recovery is high in the transient state ($\sim 1.7\text{Gbps}/\sim 0.8\text{Gbps}$) and quickly drops to low in the steady state ($\sim 0.6\text{Gbps}/\sim 0.2\text{Gbps}$).

Table 5.2 shows the average traffic delay under the double-fiber/single-fiber cut failure. Compared to the light load scenario (Table 5.1), the traffic delay increases pretty much for the wrapping scheme (from previous $\sim 223\mu\text{s}/\sim 176\mu\text{s}$ to now $\sim 285\mu\text{s}/\sim 206\mu\text{s}$). This is due to the heavy congestion in the network and hence frames start queuing in ring switches. The steering scheme and the wts scheme achieve relatively low traffic delay (both $\sim 174\mu\text{s}/\sim 150\mu\text{s}$) because the network after recovery is under light congestion.

Conditions	Average delay (μ s)
No_failure	127
Doublefiber_cut_failure_wrapping	285
Singlefiber_cut_failure_wrapping	206
Doublefiber_cut_failure_steering	174
Singlefiber_cut_failure_steering	150
Doublefiber_cut_failure_wts	174
Singlefiber_cut_failure_wts	150

Table 5.2: Average traffic delay under double-fiber/single-fiber cut failure with access traffic rate at each node 2Gbps

The performance of the steering scheme, the wrapping scheme and the wrap then steer scheme is compared in the Table 5.3.

Protection schemes	Convergence time	Traffic loss	Traffic delay after recovery
wrapping	fast	low	long
steering	slow	high	short
wrap then steer	fast	low	short

Table 5.3: Comparison of three protection schemes

Overall, the wrap then steer scheme combines the advantages of the steering scheme and the wrapping scheme to achieve a fast convergence time, low traffic loss and short traffic delay after network recovery. As a conclusion, the wrap then steer scheme is more suitable to run in MANs environment. The RER can employ the wrap then steer scheme to achieve better performance in system resiliency.

Figure 5.9: Carried traffic under double-fiber cut failure with access traffic rate at each node 2Gbps

Figure 5.10: Carried traffic under single-fiber cut failure with access traffic rate at each node 2Gbps

Figure 5.11: Traffic loss under double-fiber cut failure with access traffic rate at each node 2Gbps

Figure 5.12: Traffic loss under single-fiber cut failure with access traffic rate at each node 2Gbps

5.3 Simulation on Interconnection of Two RER Rings

As discussed in Chapter 4, multiple RER rings can be interconnected using IEEE 802.1D transparent bridging technique. The objectives of this simulation are to study the performance on a network which consists of two interconnected RER rings under various local/remote load distributions. The RER network simulated is shown in Fig. 5.13. In case a host located in RER1 sends traffic to a host located in RER2, the source switch in RER1 first unicasts traffic to the bridge node. The bridge node relays traffic to RER2 and then unicasts the traffic to the destination switch, and vice versa.

Figure 5.13: Simulation topology for interconnection of two RER rings

This set of simulations used the same simulation settings as the first set of simulations (see Section 5.1 for more details). For simplicity, we further assume that traffic is generated only in RER1.

Fig. 5.14 shows the carried traffic for different traffic patterns under various offered loads. Traffic pattern (x, y) denotes that x percent of traffic is destined to the local ring and y percent of traffic is destined to the remote ring. As shown in the figure, traffic loss for pattern $(0, 100)$ occurs first while increasing the normalized offered traffic, followed by patterns $(40, 60)$ and $(70, 30)$. Traffic loss for pattern $(100, 0)$ occurs at

the highest normalized offered traffic.

The bridge node is the only linkage between the two RER rings. All frames destined to remote ring are therefore forwarded to the bridge node. However the bridge node may become the bottleneck of the network as it has only one 10Gbps output port to the remote ring. Moreover, network traffic is unbalanced traffic destined to the remote ring is aggregated to the bridge node. As a result, the links near to the bridge nodes become congested. Due to the bottleneck and unbalanced traffic problems, traffic loss for the pattern (0, 100) occurs even when the normalized offered traffic is very low (~ 1). As less and less traffic is destined to the remote ring, the bottleneck and unbalanced traffic problems have less and less effects on the network traffic. Traffic loss for (40, 60) and (70, 30) occur at higher normalized offered traffic, (~ 1.6 and ~ 2 respectively). For (100, 0), i.e. no traffic destined to remote ring, the bottleneck problem is completely removed and network traffic is totally balanced. As a result, traffic loss occurs at the highest normalized offered traffic (~ 2.4). As a conclusion, the bridge interconnects two LANs tends to be easily saturated if remote traffic ratio is high. In actual implementations of such a network, local/remote traffic distribution needs to be studied carefully. If remote traffic ratio is high, a more powerful bridge node is needed.

Figure 5.14: Normalized carried traffic vs. normalized offered traffic for different traffic patterns

Chapter 6 Conclusions and Future Work

6.1 Conclusions

Ethernet technology was designed to work primarily with LANs. Over twenty years' development, it has become the dominant technologies in LANs due to its simplicity, low cost, ease of manageability and maturity. The emerging 10GbE would not only provide data rates of 10Gbps in LANs but is also designed to work with MANs and WANs, where SDH/SONET currently dominates. Service providers are desired to push Ethernet into MANs and WANs. Firstly, this would simplify network management and make operations more efficient and less expensive by eliminating the need to convert transmissions between Ethernet, SONET, and other network types. Secondly, 10GbE equipment would also be less expensive if it would be based on a widely used technology. Since there are still not a lot of compelling applications for 10GbE in LANs, pushing 10GbE into MANs and WANs would be a good choice. Therefore, we have proposed a new optical Resilient Ethernet Ring (RER) scheme which is based on the standard 10GbE technology and is very well suited to MANs.

In chapter 2, a literature survey of different types of network technologies was presented. An overview of the dominant LAN technology, Ethernet, has been given.

We also review two MAN technologies, SONET and RPR. SDH/SONET dominates current-generation MANs. However, there are some well-known drawbacks to SDH/SONET ring networks. First of all, SDH/SONET is expensive. Secondly, SDH/SONET is not good at transporting data traffic. Thirdly, provisioning for a large SDH/SONET ring network could be a difficult and lengthy process. Lastly, SDH/SONET does not use bandwidth efficiently. RPR is emerging as a new technology to provide the SONET-level protection and restoration with using bandwidth more efficiently like the Ethernet. But RPR has its own limitations. RPR equipment does not have built-in support for access services. RPR equipment remains expensive compared to SDH/SONET and Ethernet equipment due to relatively few RPR chip suppliers. Lastly, RPR might become a complex technology by including advanced operations in the MAC protocol such as fairness algorithm, congestion control etc. Though proposed four years ago, RPR has not been completely standardized yet partly because of the complexity of its protocol. We also compare the 10GbE, SONET and RPR technologies. The comparison shows 10GbE can be a flexible and cost-effective candidate for the next-generation MANs.

In chapter 3, we have described in detail the basic RER system design and operation issues such as architecture, frame format, frame forwarding table design, frame switching, protection schemes etc. The proposed RER supports flexible bandwidth allocation and spatial reuse mechanism. Unlike RSTP based rings, RER supports true ring topology. Every RER node stores a copy of the ring image which refers to the

collection of information on the number and arrangement of all RER nodes in the network. We proposed a hierarchical frame forwarding scheme for ring topology to simplify the process of forwarding frames from source to destination. In this scheme, the transmission path of a frame entering RER is divided into two parts – (a) from the entrance of RER to the exit of RER and (b) from the exit of RER to the destination. With this scheme, a forwarding decision is made at the entrance of RER to determine the exit of RER and the direction in which the frame will be forwarded towards the exit. The main job of an RER node becomes to forward the frame to the correct exit of RER (part (a)). At the exit of RER, another forwarding decisions will be made to further forward the frame to the destination host (part (b)). Based on the hierarchical frame forwarding scheme, the forwarding table in an RER node also consists of two parts correspondingly – (a) sub-forwarding table 1 and (b) sub-forwarding table 2. Sub-forwarding table 1 and ring image are involved in forwarding decision making for part (a) of the hierarchical scheme and sub-forwarding table 2 is involved in forwarding decision making for part (b) of the hierarchical scheme. The simplicity of ring topology expedite frames switching in the RER because once a forwarding decision has been made when the frame enters the RER, other RER nodes receiving the frame can simply follow the forwarding decision instead of making their own. In order to efficiently forward frames in RER, we proposed to insert a 5-byte RER header in the conventional Ethernet frame when a frame enters the ring. The RER header will be removed when the frame leaves RER so it will be transparent to the end-users. The use of 8-bit switch ID to represent an RER node keeps the RER header

small. RER is able to classify frames into two priority classes by making use of the priority bit. Broadcast or multicast frames are stripped by the RER node which initiated the broadcast or multicast. Initial value of TTL for a broadcast or multicast frame is set to the number of RER nodes in the network so the frame will reach each RER node but will be limited to a transit of only one complete rotation around the ring. RER is able to detect a link or node failure quickly and recover the network from failure within 50ms by continuously monitoring keep-alive frames from neighbors. The wrapping scheme or the steering scheme can be used to recover RER from a link or node failure. The wrapping scheme is faster but transmitting paths after wrapping may not be optimal. The steering scheme is slow but transmitting paths after steering are optimal. We proposed a new scheme called *wrap then steer scheme* which is the combination of the wrapping scheme and the steering scheme. The new protection scheme combines the advantages of the two schemes to achieve fast recovery and optimal transmitting paths after recovery simultaneously. Finally, RER switch design is considered. Functional blocks are presented specifically for RER switches. The two most time-consuming switch operations, i.e. hashing operation and memory access, are also discussed.

In chapter 4, we have examined three different approaches to interconnect multiple RER rings. The first approach does not allow ring nodes to learn MAC addresses from remote rings. This approach results in a small forwarding table but poor bandwidth efficiency. The second approach allows ring nodes to learn MAC

addresses from remote rings and makes use of a locally unique 8-bit switch ID to identify a ring switch. This approach achieves bandwidth efficiency but forwarding tables become large. The third approach also allows ring nodes to learn MAC addresses from remote rings as in the second approach but uses a globally unique switch ID to identify a ring switch. Compared to the first and the second approaches, the third approach maintains bandwidth efficiency and keeps forwarding tables in transparent bridges small. However, this approach increases complexity in both frame header and forwarding tables which makes the implementation of the network more complicated. Protection issues of interconnected RER rings are also analyzed in this chapter. To prevent frames from circulating among bridge ports, the STP/RSTP needs to be applied in bridge ports to form a loop-free spanning tree. To avoid single point of failure, doubly bridged links could be used other than single bridged link. However, the use of doubly bridged link together with the STP/RSTP could incur a persistent black hole problem. To solve the problem, synchronization of the forwarding table entries in ring switches with the changes of port states needs to be ensured. This can be done by a forwarding table entry removal mechanism based on the reception of BPDUs.

In chapter 5, we have conducted three sets of simulations to examine the performance of the proposed RER. The first simulation is to compare the performance difference among three different ring approaches, i.e. RER, RSTP and RPR. Simulation results show that the RER approach can achieve significantly better performance on

throughput and average traffic delay compared to the RSTP approach. The second simulation is to compare the performance of three different protection schemes, i.e. the wrapping scheme, the steering scheme and the wrap then steer scheme. Simulation results prove that the wrap then steer scheme combines the advantages of the steering scheme and the wrapping scheme to achieve a fast convergence time, low traffic loss and short traffic delay after network recovery. The third simulation is to examine the interconnected two RER rings. Simulation results show that bridge nodes can easily become bottlenecks of the network if a large amount of traffic destined to remote rings and network traffic is unbalanced due to the aggregated traffic to the bridge node. Network performance is better if less traffic destined to remote rings.

The important issues which have not been addressed in this paper are those connected with global fairness, quality of service (QoS) or other such issues that may also be essential to MANs. These issues need to be examined in conjunction with higher-layer protocols such as Multi-Protocol Label Switching (MPLS) and Virtual LAN (VLAN) and would be beyond the scope of this project. In this thesis, we have given a preliminary RER design for extending 10GbE to MANs and have focused on the related MAC layer design issues to show the advantages of the proposed approach.

To conclude from above, we have proposed a new optical Resilient Ethernet Ring (RER) which is based on the standard 10GbE technology. The proposed RER is

constrained to layer 2 and complies with IEEE 802 standard. It extends Ethernet's simplicity, familiarity and low cost features beyond LANs and into MANs, while simultaneously meeting the requirements one needs from MANs. These requirements are scalability to the new size, efficient bandwidth utilization, basic class of services (CoS) and increased robustness against failures. Furthermore, RER remains simple and easy to use by allowing value-added features to be deferred to upper layers whenever possible. The proposed RER scheme would therefore be an attractive option in MANs.

6.2 Recommendations for Future Work

At the current stage, RER is constrained to layer 2. In the next stage, RER may be explored to higher layers. In conjunction with higher layer protocols such as MPLS and VLAN, global fairness, quality of service (QoS) and other such issues can be realized.

In the interconnection of multiple RER rings, we have proposed to apply the STP/RSTP to bridge ports to prevent broadcast frames from circulating among them. However, the convergence time for the STP/RSTP could be quite long. In next stage, studies can be carried out on how to achieve a true sub-50ms system protection and restoration for interconnected RER rings.

We have proposed to apply the bridging technique to interconnect multiple RER rings. Besides the bridging technique, routers can also be used which perform layer 3 routing based on the destination IP address. In the next stage, the performance of using routers to interconnect RER rings can be studied and compared to the performance of using the bridging technique.

Interconnection of multiple RER rings has been studied in our work. In the next stage, studies on interconnection of RER ring and other types of networks can be carried out, e.g. interconnection of RER ring and SONET ring, interconnection of RER ring and RPR ring etc.

Bibliography

- [1] Y. Chen, et al., "Metro optical networking," *Bell Labs Technical Journal*, vol. 4, no. 1, pp. 163-186, Jan-Mar, 1999.
- [2] T. Ioannis, "Metropolitan area optical networks," *IEEE Circuits and Devices Magazine*, vol. 19, no. 4, pp. 24-30, July, 2003.
- [3] P. A. Humblet, "Metro area optical networks: technologies and trends," *14th Annual Meeting of the IEEE Lasers and Electro-Optics Society (LEOS 2001)*, vol. 2, pp. 429-430, Nov. 2001.
- [4] E. Isobe, J. Cho, et al., "Linear metro transport systems for the 21st century," *Hitachi Review*, vol. 48, no. 3, pp. 144-148, June, 1999.
- [5] I. Habib, D. Awduche and A. Fumagalli, "Advances in metropolitan optical networks (architectures and control)," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 8, pp.1393-1395, Oct. 2004.
- [6] M. Kuznetsov, M. M. Froberg, et al., "A next-generation optical regional access network," *IEEE Communications Magazine*, vol. 38, pp. 66-72, Jan. 2000.
- [7] D. Tavangarian, "Local-area and wide-area computing: architectures and tools," *Proceedings Ninth Euromicro Workshop on Parallel and Distributed Processing*, pp. 40-48, Feb. 2001.

- [8] O. Ibe, "Introduction to local area networks for manufacturing and office systems," *IEEE Control Systems Magazine*, vol. 7, no. 3, pp. 36-42, Jun 1987.
- [9] K. Thompson, G. J. Miller and R. Wilder, "Wide-area internet traffic patterns and characteristics", *IEEE Network*, vol. 11, no. 6, pp. 10-23, Nov.-Dec. 1997.
- [10] T. Y. Ho, L. H. Hsu and T. Y. Sung, "Transmitting on various network topologies," *Networks*, vol. 27, no. 2, pp. 145-57 March 1996.
- [11] P. A. Morreale and G. M. Campbell, "Metropolitan-area networks," *IEEE Spectrum*, vol. 27, no. 5, pp. 40-42, May 1990.
- [12] T. Miyoshi, Y. Tanaka and K. Sezaki, "Topological design comparison for multicast network," *GLOBECOM*, vol. 3, pp.1899-1904, 1999.
- [13] A. Hutton and S. Bose, "Comparison of star and ring topologies for entanglement distribution," *Physical Review A (Atomic, Molecular, and Optical Physics)*, vol. 66, no. 3, p 32320-8, Sept. 2002.
- [14] H. Fujita, K. Sakai and T. Aoyama, "SONET fiber-optic transmission systems," *Hitachi Review*, vol. 44, no 4, pp. 187-192, Aug. 1995.
- [15] D. Woods, "SONET from scratch," *Network Computing*, vol. 11, no. 9, pp. 129-136, 15 May 2000.
- [16] M. J. Klein, "The synchronous digital hierarchy: principles, variants and applications", *Philips Telecommunication Review*, vol. 48, no. 4, pp. 20-27, Dec. 1990.
- [17] A. P. Pillai, "Synchronous digital hierarchy (SDH)", *Telecommunications*, vol. 45, no. 4, pp. 61-71, Aug. 1995.

- [18] ITU-T, "Recommendation G.707: Network Node Interface for the Synchronous Digital Hierarchy," October. 2000.
- [19] ITU-T, "Recommendation G.784: Synchronous Digital Hierarchy (SDH) management," June 1999.
- [20] Bellcore Publication GR-253-Core "Synchronous Optical Network (SONET) Transport Systems: Common Generic Criteria," January 1999.
- [21] Resilient Packet Ring Alliance, <http://www.rpralliance.org>.
- [22] P. Yuan, V. Gambiroza and E. Knightly, "The IEEE 802.17 media access protocol for high-speed metropolitan-area resilient packet rings", *IEEE Network*, vol. 18, no. 3, pp. 8-15, May-June 2004.
- [23] IEEE Std 802.17, "IEEE standard for information technology-telecommunications and information exchange between systems-local and metropolitan area networks-specific requirements-part 17: resilient packet ring (RPR) access method and physical layer specifications," 2004.
- [24] F. Davik, M. Yilmaz, S. Gjessing and N. Uzun, "IEEE 802.17 resilient packet ring tutorial," *IEEE Communications Magazine*, vol. 42, no. 3, pp. 112-118, Mar 2004.
- [25] 10 Gigabit Ethernet Alliance, <http://www.10gea.org>.
- [26] J. V. Bogaert, "E-MAN: Ethernet-based metropolitan area networks," *Alcatel Telecommunications Review*, no. 1, pp. 41-45, 2002.
- [27] O. Ishida and H. Ichino, "10 Gigabit Ethernet technologies," *NTT Review*, vol. 13, no. 4, pp. 42-47, July, 2001.

- [28] ISO/IEC 15802-3:1998 (IEEE Std 802.1D-1998). "Information technology - telecommunications and information exchange between systems - local and metropolitan area networks - common specifications. Part 3: Media Access Control (MAC) bridges," 10 Dec. 1998.
- [29] "Local and Metropolitan Area Networks – Common Specifications – Part3: Media Access Control (MAC) Bridges – Amendment 2: Rapid Reconfiguration," ammendment to IEEE Std 802.1D, 1998ed., June 2001.
- [30] M. Munafo, F. Neri, C. C. Scarpatti and A. Vasco, "Analysis of the spanning tree and source routing LAN interconnection schemes", *Internetworking: Research and Experience*, vol. 5, no. 3, pp 121-149, Sept. 1994.
- [31] IEEE Rapid Spanning Tree Protocol, <http://www.ieee802.org/1/pages/802.1w.html>.
- [32] Cisco Systems Inc., "Understanding rapid spanning tree protocol (802.1w)," <http://www.cisco.com/warp/public/473/146.pdf>.
- [33] Y. F. Wong, C. Y. Wong, L. H. Ngoh and W. C. Wong, "Operation, management and performance issues of LAN technologies applied to WAN architecture," *Proceedings Eighth IEEE International Symposium on Computers and Communication (ISCC 2003)*, vol. 2, pp. 1379-1386, June-July 2003.
- [34] M. C. Nuss, "Optical Ethernet in the Metro," *The 14th Annual Meeting of the IEEE Lasers and Electro-Optics Society (LEOS 2001)*, vol. 1, pp. 289-233, Nov. 2001.
- [35] Rich Seiferr, *Gigabit Ethernet*, Addison Wesley, Aug. 1999.

- [36] Held, Gilbert, *Ethernet networks [e-book] : Design, implementation, operation, management*, 4th Edition, Wiley, 2003.
- [37] IEEE Std 802a, "IEEE standard for local and metropolitan area networks: overview and architecture amendment 1: ethertypes for prototypes and vendor-specific protocol development," 2003
- [38] R. Seifert, *The Switch Book - The Complete Guide to LAN Switching Technology*, Wiley, 2000.
- [39] Perlman, Radia. *Interconnections, Second Edition: Bridges, Routers, Switches, and Internetworking Protocols*. Boston: Adisson Wesley, 1999.
- [40] IEEE Std 802.3ad, "Amendment to carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications-aggregation of multiple link segments," 2000.
- [41] IEEE Std 802.1Q, "IEEE standards for local and metropolitan area networks: virtual bridged local area networks," 1998.
- [42] B. E. Smith and C. Yackle, "SONET bidirectional ring capacity analysis: a pragmatic view," *Conference Record - International Conference on Communications*, vol. 1, pp. 489-493, 1994.
- [43] P. Anelli and M. Soto, "Evaluation of the APS protocol for SDH rings reconfiguration," *IEEE Transactions on Communications*, vol. 47, no. 9, pp.1386-1393, Sept. 1999.
- [44] C. C. Shyur, Y. M. Wu and C. H. Chen, "Capacity comparison and cost analyses for SONET self-healing ring networks," *IEICE Transactions on*

- Communications*, vol. E77-B, no. 2, pp. 218-225, Feb, 1994.
- [45] G. Ellinas and T. E. Stern “Automatic protection switching for link failures in optical networks with bi-directional links,” *GLOBECOM*, vol. 1, pp. 152-156, Nov. 1996.
- [46] G. Suwala and G. Swallow, “SONET/SDH-like resilience for IP networks: a survey of traffic protection mechanisms,” *IEEE Network*, vol. 18, no. 2, pp. 20-25, Mar-Apr 2004.
- [47] T. H. Wu and M. E. Burrowes, “Feasibility study of a high-speed SONET self-healing ring architecture in future interoffice networks,” *IEEE Communications Magazine*, vol. 28, no. 11, pp. 33-42, Nov. 1990.
- [48] N. Ghani and Z. S. Zhang, “Resilient packet rings: next-generation data transport,” *Optical Networks Magazine*, vol. 4, no. 6, pp. 146-148, Nov.-Dec. 2003.
- [49] <http://www.ieee802.org/17/>, IEEE 802.17 Resilient Packet Ring Working Group, draft v. 2.2, Oct 2003.
- [50] Special Issue of IEEE Communications Magazine on “Generic Framing Procedure (GFP) and Data over SONET/SDH and OTN,” May 2002.
- [51] ITU-T, “Recommendation G.7041: Generic Framing Procedure (GFP),” October 2001.
- [52] Y. Peng, Z. J. Liu and J. Liu, “High performance fair bandwidth allocation algorithm for resilient packet ring,” *The 17th International Conference on Advanced Information Networking and Applications (AINA 2003)*, pp. 415-420,

March 2003.

- [53] X. B. Zhou, G. W. Shi, H. B. Fang and L. G. Zeng, "Fairness algorithm analysis in resilient packet ring," *International Conference on Communication Technology Proceedings (ICCT 2003)*, vol. 1, pp. 622-624, April 2003.
- [54] G. W. Shi, Q. Wang, X.B. Zhou and L. G. Zeng, "Research on bandwidth management mechanism of resilient packet ring," *IEEE International Conference on Communications, Circuits and Systems and West Sino Expositions*, vol. 1, pp. 763-766, June-July 2002.
- [55] Y. Robichaud, C. C. Huang, J. M. Yang and H. Peng, "Access delay performance of resilient packet ring under bursty periodic class B traffic load," *IEEE International Conference on Communications*, vol. 2, pp. 1217-1221, June 2004.
- [56] D. Cavendish, "Operation, administration, and maintenance of Ethernet services in wide area networks," *IEEE Communications Magazine*, vol. 42, no. 3, pp. 72-79, Mar 2004.
- [57] M. V. Lau, et al., "A packet-memory-integrated 44Gb/s switching processor with a 10Gb port and 12Gb ports," *IEEE International Solid-State Circuits Conference, (ISSCC 2002)*, vol. 2, pp. 34-388, Feb 3-7, 2002.
- [58] D. Y. Kim, et al., "Trends of 10 Gigabit Ethernet switch development in Korea," *IEEE Pacific RIM Conference on Communications, Computers, and Signal Processing - Proceedings*, vol. 2, pp. 1032-1035, 2003.
- [59] M. V. Lau, et al., "Gigabit Ethernet switches using a shared buffer architecture,"

- IEEE Communications Magazine*, vol. 41, no. 12, pp. 76-84, Dec. 2003.
- [60] J. W. Causey and H. S. Kim, "Comparison of buffer allocation schemes in ATM switches: complete sharing, partial sharing, and dedicated allocation," *Serving Humanity Through Communications/IEEE International Conference on Communications (SUPERCMM/ICC 1994)*, Conference Record, vol. 2, pp. 1164-1168, May 1994.
- [61] R. Jain. "A comparison of hashing schemes for address lookup in computer networks," *IEEE Transactions on Communications*, vol. 40, no. 3, pp. 1570-1573, October 1992.
- [62] Summit 400 Series Data Sheet, Extreme Networks Inc.
http://www.extremenetworks.com/libraries/prodpdfs/products/Summit400_DS.asp.
- [63] Cisco Catalyst 6500 Series Switch Data Sheet, Cisco Systems Inc.,
http://www.cisco.com/warp/public/cc/pd/si/casi/ca6000/prodlit/c6000_ds.htm.
- [64] F. Backes, "Transparent bridges for interconnection of IEEE 802 LAN's," *IEEE Network*, pp. 5-9, Jan. 1988,.
- [65] C. K. Kwok and B. Mukherjee, "On transparent bridging of CSMA/CD networks," *Proceedings IEEE GLOBECOM*, pp. 185-190, 1989.
- [66] Network Simulator - ns-2, <http://www.isi.edu/nsnam/ns>.
- [67] NS2 current release 2.27, <http://www.isi.edu/nsnam/dist/ns-allinone-2.27.tar.gz>.

Author's Publications Related to This Thesis

1. Ziwen Lian, Wende Zhong, Sanjay K. Bose, Yixin Wang, "Resilient Ethernet Ring for Metropolitan Area Networks," *The Ninth IEEE International Conference on Communications Systems (ICCS 2004)*, pp. 316-320, Singapore, 2004.
2. Ziwen Lian, Wende Zhong, Sanjay K. Bose, Yixin Wang, "Optical Resilient Ethernet Ring for High-Speed Metropolitan Area Networks," submitted to *Photonic Network Communications*.