Computer Networks 54 (2010) 57–78

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Resilience technologies in Ethernet

Minh Huynh [a,*], Stuart Goose [b], Prasant Mohapatra [a]

[a] Computer Science Department, University of California at Davis, Davis, CA, USA
[b] Siemens Technology-to-Business, Berkeley, CA, USA

ARTICLE INFO

Article history:
Received 30 May 2009
Accepted 18 August 2009
Available online 22 August 2009
Responsible Editor: L. Lenzini

Keywords:
Ethernet
Resilience
Industrial Ethernet Network
Metro Ethernet Network
PROFINET
Spanning tree
Failure
Recovery

ABSTRACT

In choosing a network service technology, a subscriber considers many features such as latency, jitter, packet loss, security, and availability. The most important feature, and usually the one that determines the final selection, is the service availability. In this article, a full spectrum of applications are studied, ranging from the minimal constraints of home networks to the rigorous demands of Industrial Ethernet Networks. This is followed by a thorough examination of Ethernet layer resilience technologies. This paper provides the resilience characteristics that are key for each class of application

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Quintessentially, Ethernet is a simple networking technology to connect two endpoints at the data link layer. Using Ethernet, a local area network (LAN) can be built and configured in a short amount of time. Its success is in part due to standardization that enables the interoperability among equipment vendors. Techniques for plug-n-play and auto-negotiation means that an Ethernet LAN does not require additional equipment, such as a rate converter because a 10 Mbps interface can communicate directly with a 100 Mbps interface. In addition, Ethernet has become the aggregation protocol, allowing other network protocols to run it, such as MPLS over Ethernet and SONET over Ethernet.

Traditionally, Ethernet uses CSMA/CD technology where multiple devices sense the medium for clearance before transmitting its data. This approach works well for a LAN in the office environment that has relatively low traffic rate and no Quality of Service (QoS) requirement. However, as applications transform and to stay ahead of competing technologies, Ethernet evolves into a full duplex gigabit network with Service Level Agreements (SLA) to meet the applications' QoS requirements. Currently, Ethernet is emerging as a significant player in new territory such as Metropolitan Area Network (MAN) and Industrial Area Network where incumbent technologies are the major players. Gradually, Ethernet is replacing legacy technologies such as private lines, ATM, and Frame Relay. One of the advantages of Ethernet over the legacy technologies are the equipment expenditure and operation expenditure. Fig. 1 shows the savings of operating Ethernet over other legacy technologies in a three year period, a study by the Metro Ethernet Forum (MEF).

In choosing a service technology or service vendor, a subscriber has to consider many parameters such as latency, jitter, packet loss, committed information rates, security and availability. All of which are important for data services considering the sharing of resources. However, studies have found that the Availability SLA weighs

* Corresponding author. Tel.: +1 408 893 5049.
E-mail addresses: mahuynh@ucdavis.edu (M. Huynh), stuart.goose@siemens.com (S. Goose), pmohapatra@ucdavis.edu (P. Mohapatra).

1389-1286/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2009.08.012

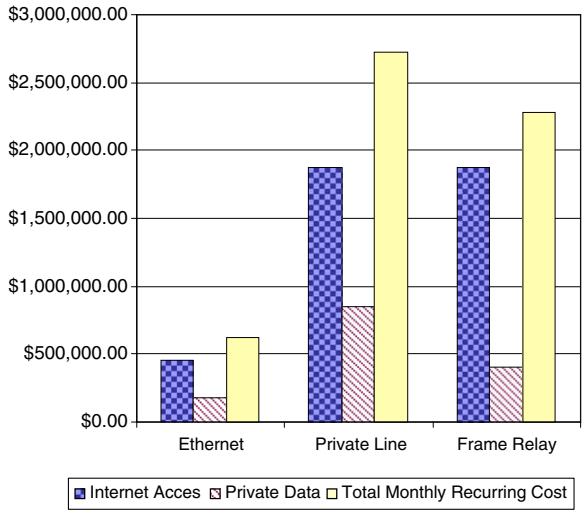

Fig. 1. Recurring Cost of Operation in a 3 year period study. Ethernet can save more than 50% over a 3 year period in a business case study from the MEF.

more than all the others in determining the market size for services and the resulting potential revenues [18]. The result of one recent market analysis shows that 50% of subscribers expect at least the 99.99% service availability. Fig. 2 shows the recovery time for different failure rate and its availability in term of the number of 9s [18]. For example, if the recovery time is 100 min and the failure rate is 10 occurrences per year, then the availability is

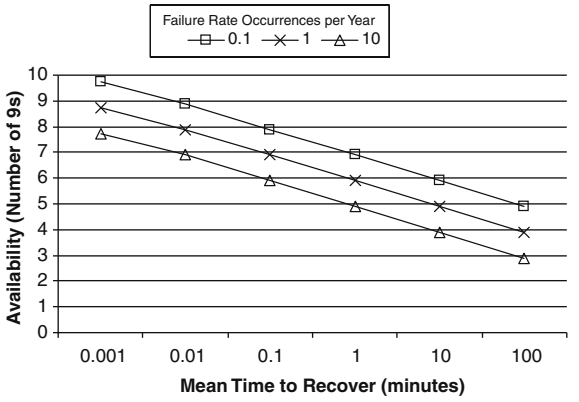

Fig. 2. Availability vs. recovery time for different frequency of failure.

99.9% (three 9s); but for a failure rate of 0.1 occurrences per year, then the availability is 99.999% (five 9s).

In addition to being competitive in term of price per Mbps and QoS, service providers also need to be competitive in terms of Availability SLA of 99.99% or higher. On top of subscribers' dissatisfactions, network downtime beyond the SLA has other tangible cost implications. Reduction in downtime translates to significant savings in maintenance costs. Therefore, in their own interests, service providers would try to achieve the availability level above the guaranteed SLA.

The primary focus of this paper is on the resiliency of Ethernet across a spectrum of constraints for a range of applications and their contingent requirements. We show also how protocols in Ethernet deal with failure detection and recovery. The protocols are abstracted and grouped into their peers to show the features that enable the appropriate response to failures that match the application needs.

2. Topology

A network topology comprises the following fundamental topologies: linear, tree, ring, star, or mesh. The linear topology and tree topology are configured without any redundancy; whereas ring and mesh topologies have redundant links built-in to protect the network. Redundancies within a network include network elements such as switches and links that exceed the minimum number for the network to operate. The redundancies create more than one path between the source and destination to reroute the traffic at the time of failure. Fig. 3 shows a typical linear topology where the switches are connected in a line. Each switch has at most two links connecting the immediate adjacent switches. The latency for any communication is proportional to the distance of the ingress and egress switch. Therefore, due to the latency requirement of some applications, when a linear topology reaches a certain size it must be branched out, as shown in Fig. 4.

Fig. 5 shows an example of a tree topology with a root switch. Within a tree topology, there is only one path between any nodes, leaf nodes or switches. Essentially, traffic tends to be forwarded toward the root on its path to the destination. In effect, this topology has the bottleneck at links around the root. A subset of the tree topology is a star topology, as shown in Fig. 6.

The ring topology is popular in network deployment because of its simplicity, deterministic behavior, and built-in

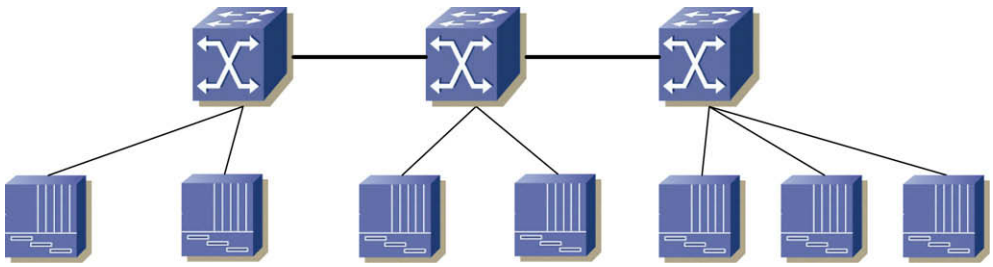

Fig. 3. Example of a linear topology with leaf nodes.

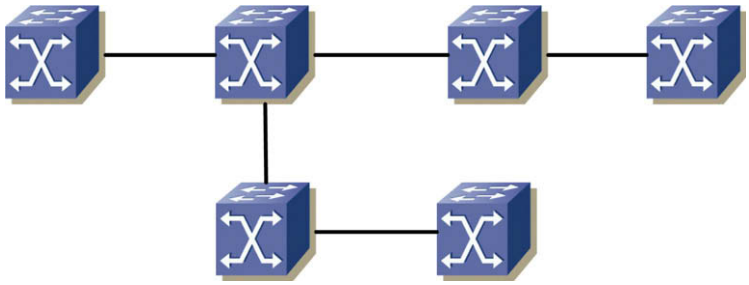

Fig. 4. Example of a branch from a linear topology.

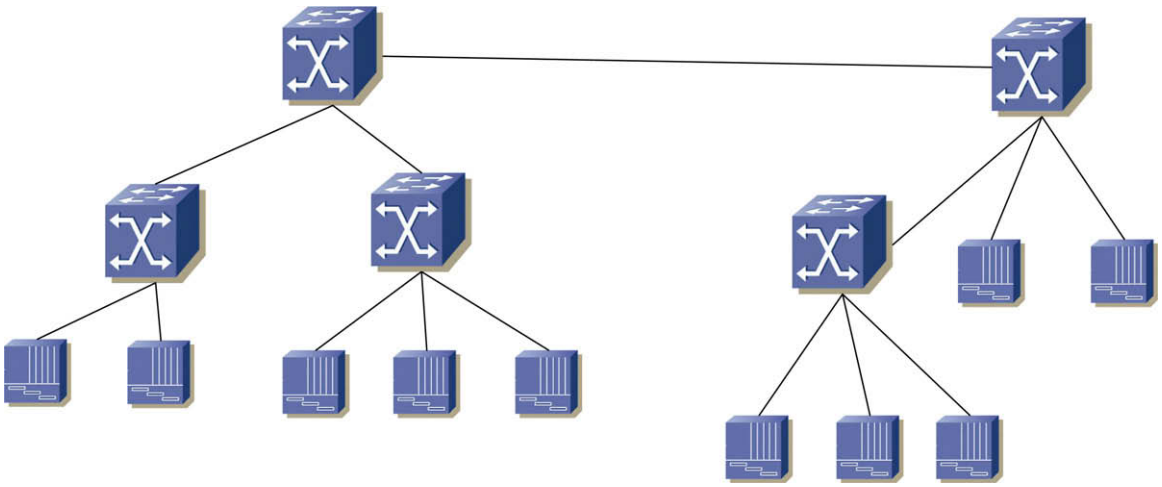

Fig. 5. Example of a tree topology.

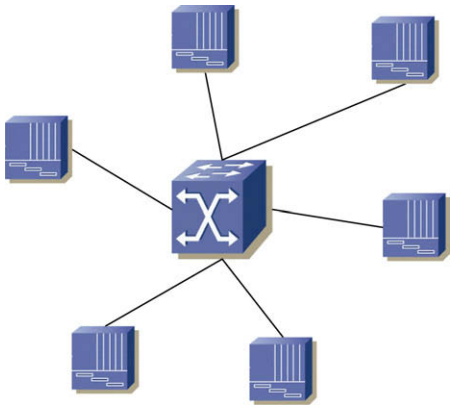

Fig. 6. Example of a star topology.

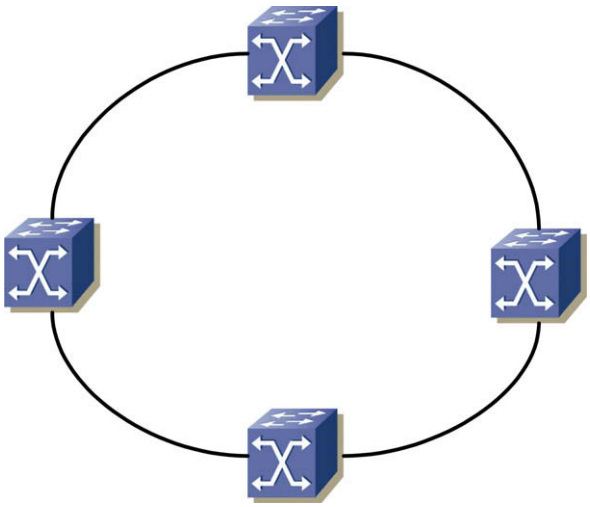

Fig. 7. Example of a single ring topology.

redundancy. Different variants of the ring topology are shown in Figs. 7–9. A ring topology creates a loop in the network that causes a frame to circulate infinitely. The management protocol must ensure that loops are eliminated while still able to exploit the advantages of the redundant links. In a multiple ring topology, all the rings can be managed by a single management instance, or different management instances that intercommunicate.

A mesh topology is a general topology of a network. There are two types: a partial mesh and a full mesh, as shown in Fig. 10. Typically, there is more than one path between any pair of source destination because of the redundant links in the mesh. The path with the best cost is used

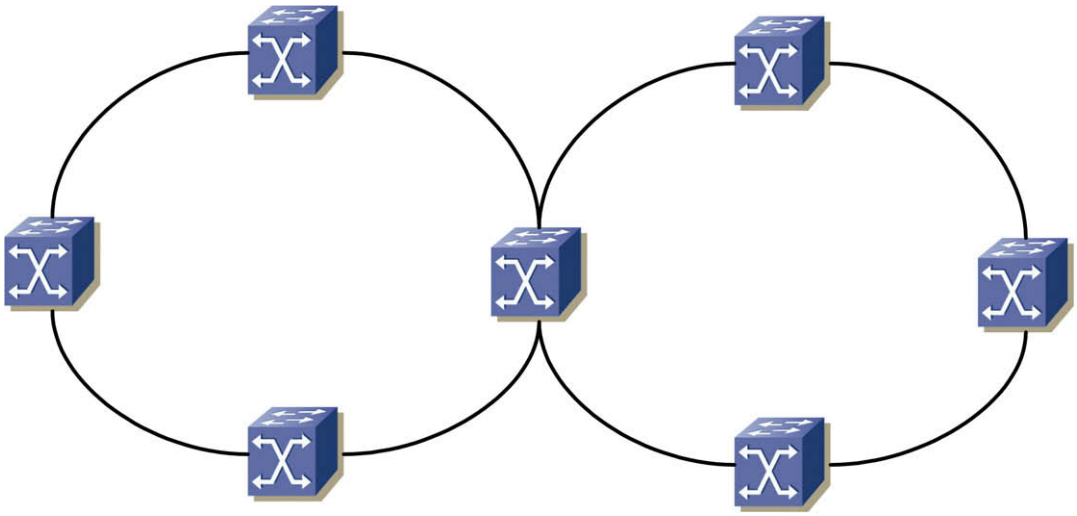

Fig. 8. Example of a multiple rings topology.

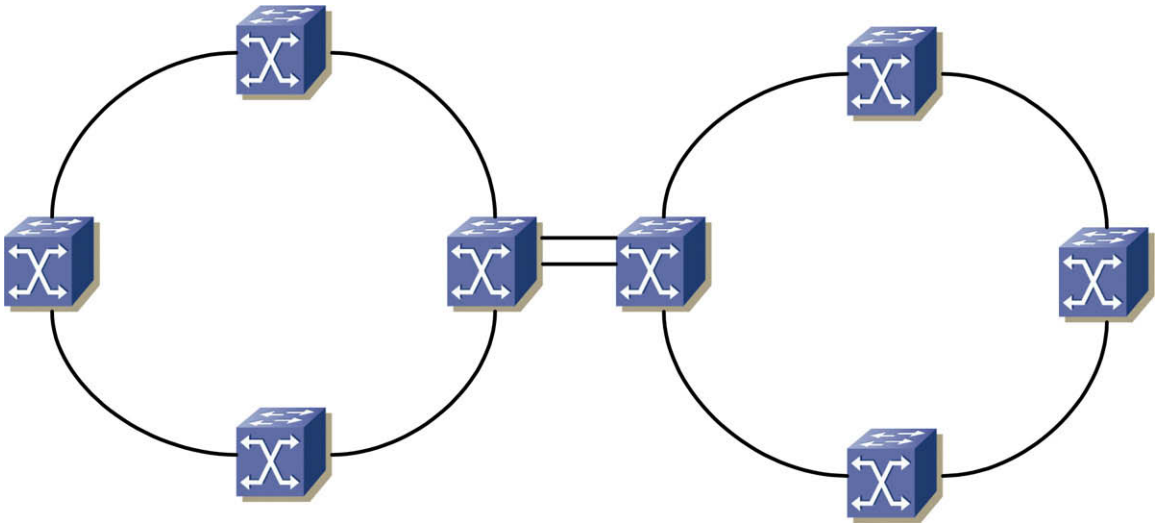

Fig. 9. Example of a multi-ring with redundancy between rings.

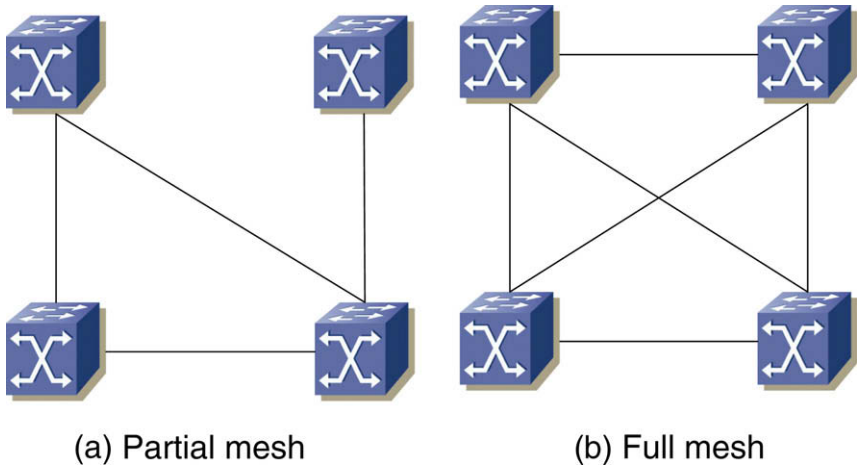

Fig. 10. Example of a mesh topology.

as the primary path until a failure occurs affecting the delivery of the packets. In a fully meshed topology, each switch has direct links to every switch in the topology. To prevent infinite looping in the topology during a flood, each switch uses the split-horizontal approach to forward a packet. A switch would forward a packet to all the switches in the network; however, it would not forward packets that it had received from another switch.

Different applications prefer a certain topology to fit their quality of service requirements. For example, typical topologies used in industrial networks are rings and linear topologies because of their deterministic behavior. At the field level of an industrial network, linear topologies and ring topologies s are deployed. However, ring topologies are majority found at the control level of an Industrial Ethernet Network. Moving up the network hierarchy, such as the management level or an enterprise LAN, mesh topologies and star topologies are preferred because the network can tolerate high latency and best-effort behavior. In a larger network, such as the metropolitan area network, the metro core network deploys the ring architecture while the metro access network runs on a meshed architecture.

3. Failure types

Network failures account for more than one third of IT related failures [1]. These failures can occur across all of the seven OSI layers. Fig. 11 shows the distribution of errors in a LAN across the OSI model. Misconfigurations are generally the main cause of failures in the link layer that resulted in corrupted forwarding tables, while a link failures and node failures are the main causes in the physical layer. A link failure occurs when a cable damaged or when errors occur at the network interface. Usually this type of failure is localized and can be fixed quickly via the backup path by the protocols managing the forwarding topology. A node failure is more severe in that all links connected to this node, including the connected leaf nodes, lose their connections. These leaf nodes have no way to reroute their traffic, unless they have redundant links that connect to another switch or router. Another failure that occurs at the physical layer is when corrupted packets arrive at the receiver. Corruption occurs during the transmission or propagation of the packet on the link when one or more of the bits inside the packet is modified. After examining the error correction checksum of the packet, the receiver discovers the error and discards the packet.

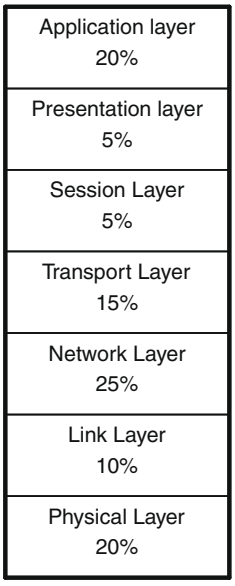

Fig. 11. Frequency of network related errors in a LAN across the OSI model.

4. Protection mechanism

To achieve a high level of service availability, a network architecture can provide a system of physical redundancy in parallel with software for efficient management. The physical redundancy is needed to eliminate the single point of failure syndrome on the routing path. There are reserved resources in a system, such as redundant links and redundant nodes, which after a failure occurs these standby resources are used to reroute the traffic. There are different levels of protection ranging from 1+1, offering 100% protection, to m:n where the protection resources are shared offering only partial protection of the traffic.

The protection type 1+1 is the most expensive mechanism, but it guarantees 100% protection. At the ingress node, the traffic is replicated and is sent to the destination via two disjoint paths. The egress node is responsible for forwarding one frame and discarding the duplicate. The decision is performed on a per frame basis and is triggered by an event such as missing frames from the primary flow. As there are always two flows carrying identical traffic, the bandwidth utilization is very inefficient.

In contrast to the 1+1 protection, the m:n schemes protect the network using a shared set of reserved resources. Specifically, n working resources are being protected by m protection resources. The protection resources are activated in the event of a failure. There is no mapping between the protection resources and the working resources as in the 1+1 case. Any one of the m resources can be used to reroute any of the n working resources. There are some special cases of m:n setting like 1:1, 1:n, and n:1

During the recovery phase, different approaches have been developed depending on the required reaction time to a failure. The speed of recovery time is dictated by the configuration of the standby resources.

In Cold Standby, the backup paths are predetermined offline but not activated until there is failure detected on the primary path. Once a fault is detected, the source node establishes the backup path to continue forwarding the traffic. The delay of the control message that detects the failure is directly proportion to the delay of the recovery time.

By contrast, a Hot Standby activates the predetermined backup path at the same time as the primary path. This is an example of 1+1 protection. The backup path forwards the same traffic as the primary, consuming network re-

sources inefficiently. It has the fastest recovery since the only delay is the failure detection time, which is the dominant latency in service restoration for any type of protection mechanism. Hot Standby can also be combined with Cold Standby so that few resources are needed. Hot Standby restores the high priority traffic while Cold Standby is restores all other traffic classes in the network.

A compromise between Cold Standby and Hot Standby is Shared Redundancy. The backup paths are determined on the fly when a failure detected. The traffic is then rerouted around the faulty links. Shared Redundancy can use redundant resources efficiently to forward traffic during normal operation. Ring and mesh are example topologies that can exploit this approach.

5. Resilience requirements and their respective applications

Network deployments are typically tuned to the requirements of the applications that they support. Table 1 shows a summary of the requirements and recommendations for three different categories of network and their respective recommended recovery performance [2–7].

Category 1 includes the low end of the network performance spectrum that includes end-user applications, home LAN, and small businesses. The applications include web-browsing, e-mail, file transfer, e-commerce transactions, and non-interactive video and audio streaming. Since these streaming applications are non-interactive, application-level buffering helps to mitigate the performance degradation in case of failure on the intermediate nodes. In addition, a plethora of error correction algorithms exist to ameliorate the perceptual effect on end-users. As for the other applications, the recommended recovery time for a disrupted service is deemed tolerable by the users.

The applications in category 2 are interactive media streaming and the core network performance of the Metropolitan Area Network (MEN). The difference between the streaming applications in this category and those in category 1 is the bi-directional interactive nature. Such interactivity demands faster response times in both directions. Meanwhile, the metro core networks inherit the de facto recovery time of the optical network of less than 50 ms.

Category 3 applications have the most demanding performance requirements of Ethernet Networks. These applications are used in factory automation and the precise motion control of drives. Such applications are used to control high precision industrial machinery and to provide a reliable and safe environment. Depending on the specific applications, nodes in production facilities are synchronized to within microseconds to milliseconds. Consequently, these deployment scenarios have the highest constraints on fault detection and recovery. For example, PROFINET IO does not tolerate delays above 10 ms nor jitter above 1 ms.

5.1. End-user LANs and small business offices

The applications in this category are uni-directional interaction applications characterized by the request-response pattern of the end-users. The expected delay is in the second range, categorizing these services as non-real-time. Examples are file transfers, web-browsing, emails, e-commerce transactions, audio and video streaming. The audio and video streaming in this category lack the conversational dimension, therefore, the recovery delay can be relaxed. Table 2 shows the requirements for applications in this category.

5.2. Interactive multimedia

The majority of applications in this category are subject to the human perception of real-time, such as Voice over IP, video conferencing, and gaming. The requirements are subject to the acceptable level of delay of an average person and, therefore, more onerous than the previous category of streaming applications. The ITU-T recommends an acceptable delay to be in the range of 0–150 ms. Within this range, any delay below 30 ms is not noticeable. Delays

Table 1
A summary of the requirements and recommendations for each type of application.

Category	Services	Medium	Bandwidth	Delay/recovery	Jitter	Error
1	Interactive	Audio	4–13 kbit/s	<1 s (playback); <2 s (record)	<1 ms	<3% FER
		Data	NA	<4 s	NA	0
	Streaming	Audio	5–128 kbit/s	<10 s	<2 s	<1% pkt loss
		Video	20–384 kbit/s	<10 s	<2 s	<2% pkt loss
		Data	<384 kbit/s	<10 s	NA	0
2	Conversation voice	Audio	4–25 kbit/s	<150 ms	<1 ms	<3% FER
		Video	32–384 kbit/s	<150 ms	NA	<1% FER
		Data	NA	<250 ms	NA	0
	MEN	Bulk	NA	<50 ms, <200 ms, <2 s, <5 s	NA	0
3	Industrial Ethernet Network: PROFINET	Data	6.4–96 kbit/s	5–10 ms	<1 ms	0
		Data, Video	64 kbit/s–3.2 Mbit/s	150 μs–1 ms	1 μs	0
	Industrial Ethernet Network: SERCOS III	Data	96 kbit/s–2 Mbit/s	31.25 μs–1 ms	1 μs (hw), 50 μs (sw)	0

Table 2
Requirements for end-users applications define in the ITU-T Recommendation G.114 [3].

Services	Medium	Application	Bandwidth	Delay/recovery	Jitter	Error
One way interactive	Audio	Voice message	4–13 kbit/s	<1 s(playback); <2 s(record)	<1 ms	<3% FER
	Data	Web-browsing	NA	<4 s per page	NA	0
	Data	Transaction (e-commerce)	NA	<4 s	NA	0
	Data	Email	NA	<4 s	NA	0
Streaming	Audio	Speech, music	5–128 kbit/s	<10 s	<2 s	<1% pkt loss
	Video	Movie, flash video	20–384 kbit/s	<10 s	<2 s	<2% pkt loss
	Data	FTP	<384 kbit/s	<10 s	NA	0

Table 3
Performance requirements for multimedia applications define in the ITU-T Recommendation G.114.

Services	Medium	Applications	Bandwidth	Delay/recovery	Jitter	Error
Conversation voice and two-way interactive	Audio	Conversation voice, VoIP	4–25 kbit/s	<150 ms preferred; <400 ms max	<1 ms	<3% FER
	Video	Video conference	32–384 kbit/s	<100 ms lip-synch; <150 ms preferred; <400 ms max	NA	<1% FER
	Data	Gaming	NA	<250 ms	NA	0
	Data	Telnet	NA	<250 ms	NA	0

between 100 ms and 150 ms can be mitigated with echo cancellation algorithms. However, delays between 150 ms and 400 ms may be acceptable, but they exacerbate the degradation in quality. Table 3 shows the requirements for applications in this category.

5.3. Metropolitan area networks

In Metro Area Networks (MEN), the following protection approaches that can coexist are under consideration by the Metro Ethernet Forum (MEF) [8]:

1. Aggregated Line and Node Protection (ALNP).
2. End-to-End Path Protection (EEPP).
3. MP2MP Protection.

Aggregated Line and Node Protection (ALNP) provides protection for local link and local nodes via a detour mechanism. The detour path temporarily traverses the point of failure and merges back onto the primary path. ALNP supports 1:1 and 1:n protection on the detour routes. Since the protection is local, it has a quicker recovery rate than the End-to-End Path Protection (EEPP). ALNP can also be invoked before EEPP.

Unlike ALNP, EEPP provides disjoint backup paths from the source to the destination than the primary path supporting 1+1, 1:1, and 1:n protection. The number of the backup paths depends on the policy requirement of the network in question.

MP2MP Protection is used to protect the E-LAN service in a MEN where ALNP and EEPP are insufficient. An E-LAN service in MEN is a multipoint to multipoint service connectivity between all of its User to Network Interface (UNI). For a MP2MP protection, three approaches are in use and discussed in the VPLS and Spanning Tree Protocol sections:

1. Split Horizon Forwarding.
2. Spanning Tree Protocol family.
3. Link Redundancy.

Within MEN, the different service restoration time requirement depends on the service level specification of various applications. The MEF has defined the following category of network recovery times:

- Sub 50 ms recovery time.
- Sub 200 ms recovery time.
- Sub 2 s recovery time.
- Sub 5 s recovery time.

For example, some real-time or soft real-time applications require a sub 200 ms recovery time, while some TCP-based application can tolerate a sub 5 s recovery time before triggering the Spanning Tree Protocol to reconverge the topology.

5.4. Industrial Ethernet Networks

Industrial Ethernet Network has the most stringent QoS requirements because of the high precision required to perform measurements and to control the plant reliably and safely. Many industrial machines necessitate a real-time synchronization between the master node and the slave nodes. This synchronization constrains the request and response cycle time to a few milliseconds time, and in some cases microseconds range. Operating in safety-critical and hazardous environments, it is key that Industrial Ethernet Networks deliver:

- Real-Time and deterministic behavior.
- High availability.
- Rugged and durable operation over extended periods of time.

Table 4
The typical grace time in an Industrial Network from IEC 62439.

Applications	Typical grace time
Enterprise management system	20 s
Automation management, for example, manufacturing, discrete automation	2 s
General automation, for example, process automation, power plants	200 ms
Time-critical automation, for example, synchronized drives	20 ms

An operational plant can tolerate a failure in the automation system only for a short amount of time, called grace period. For the plants to be in continuous operation, the recovery time has to be shorter than the grace period. Table 4 shows the typical grace time from the International Electrotechnical Commission (IEC) [27]. However, different factory plants have different requirements that could be stricter the requirements from the IEC.

There exist numerous different Industrial Ethernet technologies because each vendor tailors its proprietary protocols to fit the needs of its customers. As a result, a new concept of Real-Time Ethernet (RTE) arose. The International Electrotechnical Commission (IEC), a standards body, is working to bring together the different Industrial Ethernet under a common platform and set of requirements. For the scope of this paper, we focus on protocols that enhance redundancy and resilience in Ethernet Networks. These protocols belong to the PROFINET and SERCOS III family as shown in Table 5 with the corresponding requirements.

6. SERCOS III

SERCOS III family is designed only for the line and ring topologies with a maximum of 511 slave nodes per network at 100 Mbps rate [7]. SERCOS III synchronizes between the master nodes and the slave nodes through customized hardware and it can integrate non-real-time traffic in between the scheduled time slots. Similar to EtherCat, SERCOS III processes Ethernet frames on the fly. However, there are some differences such as rigid frame format preventing any changes at runtime; minimum of two frames per cycle to separate the input and output data; and non-real-time data is inserted in gaps between frames. As a result,

- Lower bandwidth utilization than EtherCat.
- Topology independent slave-to-slave communication.
- Fragmentation of Ethernet frame if the non-real-time gap is shorter than the maximum Ethernet frame length.

7. PROFINET I/O

Initially, PROFINET was developed as the answer to the hype of Ethernet to protect the investments in Profibus. PROFINET has three different flavors: Component based Automation (CbA), Soft Real-Time (SRT), Isochronous Real-Time (IRT) [7]. Fig. 12 shows the comparative cycle time and jitter rate among the three approaches.

7.1. Component based automation (CbA)

CbA uses standard unmodified Ethernet hardware and standard TCP/IP software. It is a best-effort approach where the performance is unpredictable. Via a proxy device, CbA can have access to the Profibus network. Topology recognition is not needed here.

7.2. Soft real-time (SRT)

The next level in PROFINET I/O is the soft real-time approach where Programmable Logic Controller (PLC) applications run on standard unmodified hardware and standard TCP/IP for processing data communication. However, SRT uses specialized data process protocols and bypasses TCP/IP layers for real-time data in order to achieve

Table 5
Performance requirement for PROFINET and SERCOS III (*in reality, 960 kbit/s and 108 kbit/s is the maximum instead of 3.2 Mbit/s and 408 kbit/s, respectively).

Services	Medium	Applications	Bandwidth	Delay/recovery	Jitter	Error
Industrial Ethernet Network: PROFINET	Data	Drive control, factory automation	6.4 kbit/s 12.8 kbit/s 48 kbit/s 96 kbit/s	5–10 ms	<1 ms	0
	Data, video	Motion control	64 kbit/s 427 kbit/s 3.2 Mbit/s* 408 kbit/s*	150 μs–1 ms	1 μs	0
Industrial Ethernet Network: SERCOS III	Data	Factory automation	2 Mbit/s 1.5 Mbit/s 1 Mbit/s 384 kbit/s 1 Mbit/s 192 kbit/s 400 kbit/s 256 kbit/s 96 kbit/s	31.25 μs 62.5 μs 125 μs 250 μs 250 μs 500 μs 1 ms 1 ms 1 ms	1 μs (hw), 50 μs (sw)	0

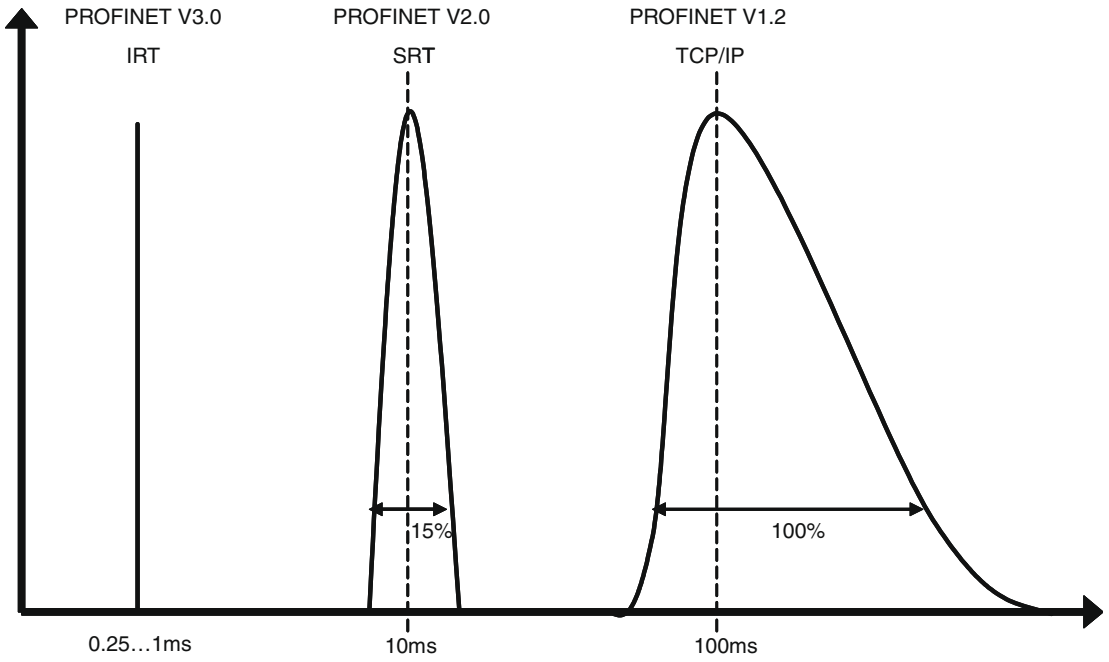

Fig. 12. Communication cycle time and their jitter [7].

a cycle time of 5–10 ms and 15% jitter rate. The drawbacks in SRT are that it is influenced by TCP traffic (non-real-time data traffic) and unpredictable queuing delay. Media Redundancy Protocol (MRP) [27] from Siemens is an example of a SRT protocol. MRP runs a topology recognition protocol, such as SNMP and LLDP.

7.3. Isochronous real-time (IRT)

Running on specialized a ASIC, Isochronous Real-Time (IRT) is defined to have cycle times in the range of 150 μs to 1 ms and 1 μs jitter with the synchronization of all nodes. However, the fastest time supported by commercial equipment starts from 500 μs [7]. IRT is deployed on tree or line topologies where it can support a maximum of 25 devices per line. To achieve deterministic behavior and low cycle time, IRT schedules real-time data at regular interval and inserts best-effort in between, as shown in Fig. 13. Each time slot reserves a certain bandwidth for the IRT data. Scheduling is complex because of the interdependencies between the topology and the performance. Each topology has its own set of parameters to achieve the desired results. Any small tweak in the configuration or the physical topology could result in unpredictable behavior. The remaining 50% of the cycle time goes to the best-effort data.

8. Category 1 – End-user applications

Most protocols in this category recover in less than the recommended recovery time of 1–3 s, except for STP. In other words, for applications other than the interactive voice message, all other services will operate without interruption during a failure. The few applications that recover in less than 1 s can satisfy without interruption. As STP was designed well before the emergence of the mod-

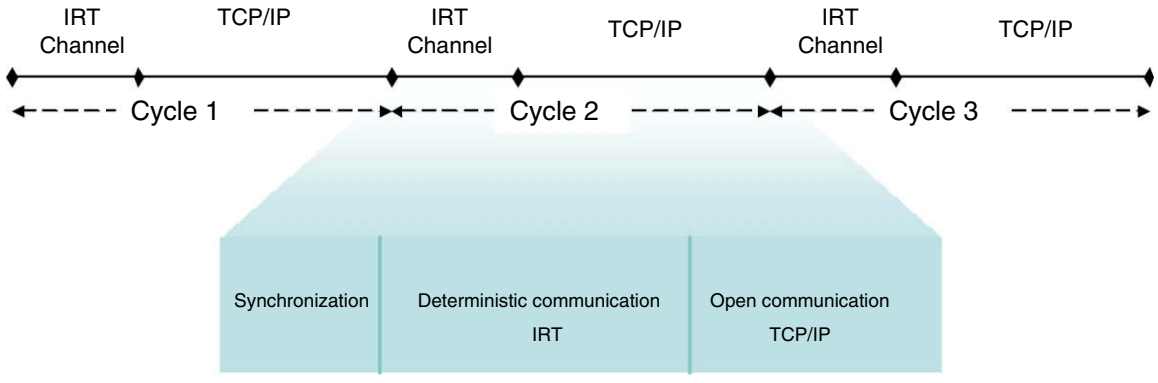

Fig. 13. Time division for IRT communication [7].

Table 6
Comparison charts for resilient protocols operating in end-user environment.

Protocols	LDD latency	LDD method	Global reconvergence latency	deterministic	Frame loss	Topology
EAPS	NA	Hello pkt + NI detection	<1 s	n	y	Ring/multi-ring
MRP (Foundry Networks Inc.)	NA	Hello pkt + NI detection	<1 s	n	y	Ring/multi-ring
STP	PHY detection	NI detection	30–60 s	n	y	Mesh
RSTP	PHY detection	NI detection	1–3 s	n	y	Mesh
MSTP	PHY detection	NI detection	1–3 s within partition	n	y	mesh
ESRP	1 s	Hello from master	3 s	n	y	Mesh
VSRP	NA	Hello from master	<1 s	n	y	Mesh
VRRP	1s	Hello from master	3 s	n	y	Mesh
RRSTP	NA	Hello pkt + NI detection	500 ms to 1 s	n	y	Ring

ern applications its recovery time was acceptable, but it is now obsolete. The ring topology boasts the protocols with the fastest recovery time. Since the behavior on a ring is more predictable, it is easier to optimize the management protocol than with mesh networks. However, the recovery time of protocols managing ring networks with a central redundancy manager is directly proportional to the size of the ring. As the ring size grows, the failover time also grows making it difficult to sustain a failover time below 1 s. Tables 6 and 7 summarizes the protocols that are suitable for applications in this class of network performance.

8.1. STP

Historically, an Ethernet-based network used the Spanning Tree Protocol (STP) as the de-facto protocol to manage its topology. STP is standardized in IEEE 802.1d [12] to forward layer 2 frames. Using the shortest path to the central root, STP forms a tree that is overlaid on top of a mesh Ethernet Network as shown in Fig. 14. Unlike IP packets, Ethernet frames do not have a time-to-live field. Therefore, the Spanning Tree blocks redundant links in the topology to avoid a broadcast storm that can bring down the network. The drawback of this approach is that the links around the root will be heavily congested, leaving it at risk of failure and unbalance loads. Upon a failure, STP takes 30–60 s to recover.

8.2. RSTP

To improve the recovery time of the Spanning Tree, the IEEE standardized the Rapid Spanning Tree Protocol (RSTP)

Table 7
Comparison charts for resilient protocols operating in end-user environment (continue).

Protocols	Centralized/ distributed	Backup path computation	Scalability	Standard/ industry	Synchronization
EAPS	Redundancy manager	Open blocked port	4096 VLANs, 64 EAPS domains	RFC 3619	Yes (complete flushing of FDB before restarting forwarding)
MRP (Foundry Networks Inc.)	Redundancy manager	Open blocked port	NA	Foundry Networks Inc.	Yes (complete flushing of FDB before restarting forwarding)
STP	Distributed	On the fly	Max 7 hop	IEEE 802.1	No
RSTP	Distributed	On the fly	Max 7 hop	IEEE 802.1w	No
MSTP	Distributed	On the fly	max 7 hop	IEEE 802.1s	No
ESRP	Redundancy manager	Switch to backup node	3000VLANs	Extreme Networks	Yes (master and slave nodes)
VSRP	Redundancy manager	Switch to backup node	NA	Foundry Networks Inc.	Yes (master and slave nodes)
VRRP	Redundancy manager	Switch to backup node	NA	RFC 3768	Yes (master and slave nodes)
RRSTP	Distributed	Open blocked port	Max 7 hop	Riverstone	No

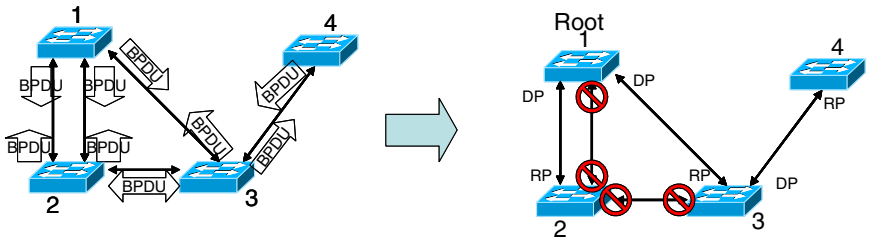

Fig. 14. STP's process of selecting root node and block redundant links to create a loop free topology.

as specified in IEEE 802.1w [12]. RSTP reduces the recovery time by cutting down the number of port states to three: discarding, learning, and forwarding. In addition to faster aging time and rapid transition to forwarding state, the reconvergence time was trimmed to between one and three seconds contingent on the topology. The topology change notification was accelerated by using the switch that discovers the fault to perform a broadcast notification, as opposed to STP where the notification traverses first via the root. However, RSTP still shares other drawbacks of STP, such as network underutilization, congestion near the root, and no load balancing.

8.3. MSTP

The most recent enhancement to STP is the Multiple Spanning Tree Protocol (MSTP) [13], as defined in IEEE 802.1s. MSTP partitions the topology into different regions that are connected together by a common Spanning Tree, called the Internal Spanning Tree (IST). The regions in MSTP are instances of RSTP with each with their own regional root. The regional roots are connected to the common root from the IST, as shown in Fig. 15. One or more VLAN can be assigned into an instance of the RSTP. By distributing and directing traffic over different VLANs, it is possible to achieve a more balanced load across the network.

8.4. EAPS

Ethernet Automatic Protection Switching (EAPS) is a ring resilience protocol in an Ethernet Network, as specified in the IETF RFC 3619 [9]. An EAPS ring contains at least two switches, where one of the nodes acts as the master. Multiple EAPS domains can exist on the same physical ring, where each EAPS domain is configured to protect a group of VLANs. A control VLAN is reserved for sending only EAPS control messages. The master node sends out periodic polls from the primary port on the control VLAN to be received on the backup port to check the connectivity of the ring. A non-master switch can notify the master switch of a failure via a link down message. Initially, traffic is sent on the primary port of the master node and the backup port is blocked, as shown in Fig. 16. If the polling timed out or a link down message is received, the master forces the forwarding database of all nodes on the ring to be flushed. The backup port is then unblocked to mend the ring for forwarding traffic. The master continues to poll the ring until the ring is restored, whereupon all nodes flush their forwarding databases and the backup port blocked to prevent loops. By tuning the switches, the fault detection and recovery can be sub-second. Initial testing of EAPSv2 showed that a failover is less than 50 ms for 10,000 layer2 flows and 100 protected VLANS [10]. EAPS is limited by the number of VLAN space and a maximum of 64 EAPS domains on a single ring.

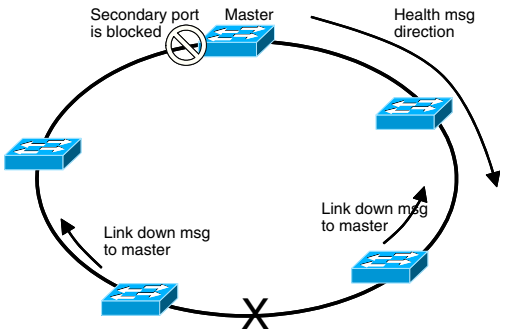

Fig. 16. Fault detection in EAPS.

8.5. MRP (Foundry)

As an alternative to Spanning Tree Protocol (STP), Foundry Networks Inc. developed Metro Ring Protocol (MRP-Foundry)[11]. This is not to be confused with Media Redundancy Protocol (MRP-IEC) specified by in IEC 62439 [27] for Industrial Ethernet Network. MRP-Foundry is designed to provide fast recovery in a ring topology. Similar to EAPS, MRP-Foundry requires a master node on the ring that initially forwards traffic on its primary port while blocking its backup port. A hello message is sent from the primary port to be received on its backup port. Multiple rings can be merged to create a large topology, but an MRP instance can only run on one physical ring and not the entire topology as shown in Fig. 17.

8.6. RRSTP

Rapid Ring Spanning Tree Protocol [18], developed by RiverStone Networks, leverages MSTP and RSTP to improve

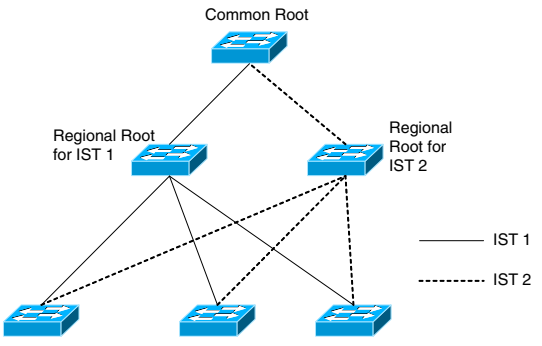

Fig. 15. An example of a MSTP configuration.

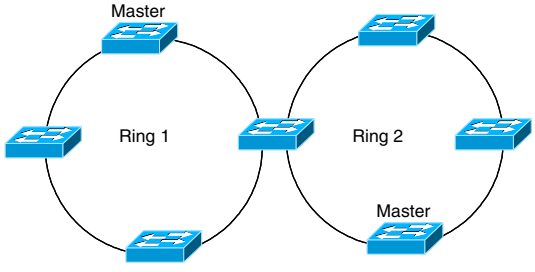

Fig. 17. A multi-ring topology in MRP-Foundry.

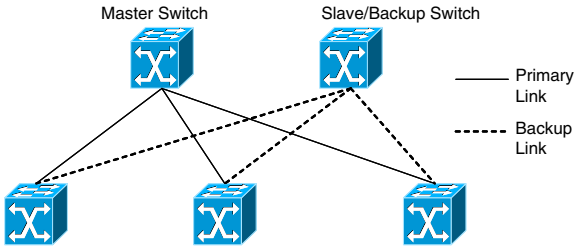

Fig. 18. A switch redundancy example.

the failover time to sub-second range. Restricted to a ring topology, each instance of the Spanning Tree manages its designated ring. Each root has a primary and secondary port. Traffic is sent initially on the primary port and the secondary port is opened for use if a link is broken on the ring. The recovery time after a failure is proportional to the BPDU hello time which is between 500 ms and 1 s.

8.7. Master-slave paradigm

In topologies where switches are connected to an upstream node, the critical point of failure is at the upstream node where aggregated traffic converges. Therefore, it is necessary to protect this node with a standby backup node as shown in Fig. 18. A slave node operates in standby mode blocking all incoming traffic while monitoring the communication with the master node. When the master node fails, the slave immediately assumes all functions of the master. While not a requirement, it is possible to avoid topology reconvergence by enabling seamless frame forwarding and MAC address learning for the new node.

8.8. ESRP

Extreme Networks developed a master and node architecture called Extreme Standby Router Protocol (ESRP) [14]. The recovery time is contingent on the communication between the master and the slave node. In the face of failure, ESRP master node sends the Extreme Discover Protocol (EDP) messages to announce the new master. The downstream nodes then discard all forwarding entries to relearn them on the new port connecting to the slave node. At the same time, the slave node can reuse the same forwarding information as that on the master node. If the downstream nodes are not from Extreme Networks, another mechanism is required to make the transition.

8.9. VSRP

A similar design to ESRP for the master-slave paradigm is the Foundry Networks' Virtual Switch Redundancy Protocol (VSRP) [15]. The failover can occur in sub-second range if all of the switches are VSRP-aware. To select the next master node in VSRP, the master and slave nodes have their priority values initially set. Each time a port fail on the master node, its priority is reduced. Over time, when the master's priority is lower than that of the backup node, the failover will then take place.

8.10. VRRP

The Virtual Router Redundancy Protocol (VRRP) is a master-slave architecture that has been standardized in RFC 3768 [16] to increase the availability of the default gateway. VRRP evolved from Cisco's proprietary HSRP [17]. The backup routers and the master router are advertised as one virtual router to the nodes in the network. The virtual router does not propagate its IP routes beyond the subnet to which it belongs. Missing three consecutive broadcasts from the master triggers a replacement of the master node and the next highest priority backup node takes over. The backup router can be used also for load sharing if desired. VRRP can be routed over Ethernet, MPLS, and token ring networks.

9. Category 2 – Interactive applications and MAN

This category includes bi-directional interactive streaming applications and Metro Area Networks. Approximately only half of the protocols are able to meet the requirements to operate without interruption during a failure. The majority of the recovery time is consumed by the failure detection step. For Ethernet, the Gigabit IEEE 802.3 [31] specification states that detecting a loss at the physical layer in 1000BASE-T requires at least 750 ms. To reach the de facto standard of <50 ms recovery time, new protocols either run over optical network (PESO) or implement their own detection protocol rather than relying on the physical Network Interface (NI). As seen in the previous category, a centralized manager instead of a distributed solution is common among these fast recovery schemes. To enable fast recoveries, the backup paths tend to be computed before hand. Tables 8 and 9 show the summary of protocols under this category.

Table 8
Comparison charts for resilient protocols operating in metro area network and multimedia environment.

Protocols	LDD latency	LDD method	Global reconvergence latency	Deterministic	Frame loss
Viking	Relies on other fault detection mechanism, then report the fault to the manager via SNMP	Send traffic update to manager	300–400 ms	n	y
Ethereal	220 ms	Hello msg	(220 + 31) ms	n	y
SmartBridge	PHY detection	NI detection	10–20 ms (10 Mbps)	n	y
PESO	NA	NI detection	<50 ms	n	y
VPLS	NA	Using LDP or BGP	<50 ms	n	y

Table 9
Comparison charts for resilient protocols operating in metro area network and multimedia environment (continue).

Protocols	Centralized/distributed	Topology	Backup path computation	Scalability	Standard/industry
Viking	Central manager	Mesh	Offline (100–200 s)	Scalable	Academia
Ethereal	Distributed	Mesh	On the fly	216 guaranteed connections	Academia
SmartBridge	Distributed	Mesh	On the fly	Low (required global knowledge of the topology)	Academia
PESO	Central manager	Mesh	Over provision	Scalable	Academia
VPLS	Centralized	Mesh	Static	Scalable with H-VPLS	RFC 4761-4762

9.1. Viking

Viking [19], proposed by Sharma et al., aimed to improve the resilience of STP ad RSTP by pre-computing multiple Spanning Trees such that in the event of a failure switching to a backup Spanning Tree can be rapid and hence maintain the quality of service. Initially, Viking finds k-shortest primary path and k backup path for each primary path. Each path computation avoids the heavily used link via a weight assignment scheme. Spanning Trees are then created by merging these paths together. A Viking server monitors network conditions through information sent from the nodes in the network and acts accordingly. Viking can guarantee bandwidth and delay requirements of current flows and disallow new flows if the network nears capacity.

9.2. Ethereal

Varadarajan et al. proposed Ethereal [20], a connection oriented architecture, to support assured service and best-effort service at the Ethernet layer. Ethereal uses the Propagation Order Spanning Tree for fast reconvergence once a failure has been detected. Utilizing periodic hello messages to immediate neighbors, a switch can detect a failure if there are missing consecutive hello messages. Once a fault has been detected, all best-effort traffic is discarded. The established QoS-assured flows are maintained unless part of the path is affected by the fault. The best-effort flows behave consistent with the STP protocol, while requests to reserve paths with the required QoS parameters are required for QoS-assured traffic. Ethereal design is directly aiming at real-time multimedia traffic via hop-by-hop reservation. Similar to a MPLS, each switch makes a request to its immediate downstream hop for the flow reservation, whereupon the penultimate node sends a reply indicating whether the reservation was successful. The scalabilty of Ethereal is limited as only 65536 connections can be supported.

9.3. SmartBridge

Realizing that congestion on the links surrounding the root node is problematic for STP, SmartBridge [21] combines the advantages of STP and IP routing to forward frames along the shortest paths. Exploiting full knowledge of the topology, frames traverse along the host of known locations on a calculated shortest path. Frames with unknown source address are discarded automatically, triggering a topology acquisition process. Frames with unknown destination address are flooded akin to standard STP, but with a minor modification to update the host location table. Frames with known source and destination addresses are guaranteed to be forwarded on the shortest path based on an assignment of weights, such that any least-weight path from source to destination is a shortest path and the least-weight path from source to destination is unique.

9.4. PESO

To protect Ethernet over SONET with a low overhead, Acharya et al. proposed PESO [22]. Traditional SONET uses a 1+1 protection, but this can be considered excessive since data traffic can tolerate failure and operate at a reduced rate. Depending on the protection requirements, PESO will compute an optimum routing path that uses virtual concatenation (VC), as shown in Fig. 19, and Link Capacity Adjustment Scheme (LCAS) to make the necessary recovery. For the scenario where a single failure should not affect more than x% of the bandwidth, PESO transforms the link capacity in the topology to the equivalent of y lines. Each chosen line out of y cannot carry more than x% protected bandwidth. PESO determines the number of members in the VC. Using a path augmentation maximum flow algorithm, such as Ford and Fulkerson [23] or Edmonds and Karp [24], PESO determines the routes that the virtual concatenation group (VCG) will take. Upon failure, LCAS removes the failed member resulting in a continuous connection with the destination but the throughput

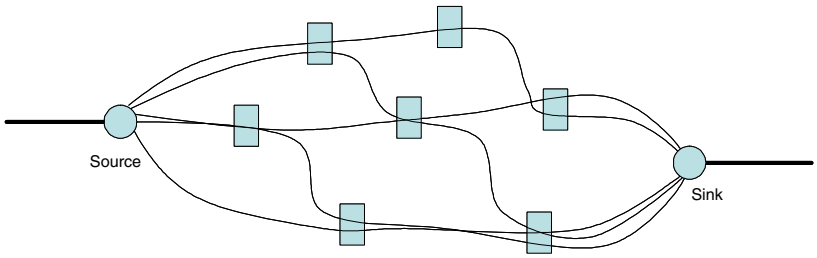

Fig. 19. Virtual concatenation (VC).

has been reduced not less than x% protected bandwidth. With PESO, VC and LCAS provide high resilience for the network with a fast recovery time between 2 ms and 64 ms.

9.5. VPLS

To integrate multiple legacy services such as ATM, Frame Relay and private line, the IETF drafted Virtual Private LAN Services (VPLS). VPLS is a multipoint-to-multipoint service that unifies remote sites running on different technologies onto a common platform in a Virtual LAN (VLAN). VPLS, also known as Layer 2 MPLS, is similar to MPLS in that it uses multipoint tunneling scheme to create the VLAN. The difference between VPLS and MPLS is the interface connecting the customer edge equipment (CE) and the provider equipment (PE). In MPLS, the PE uses routers for IP traffic, while the PE in VPLS uses Ethernet switches. In addition, VPLS emulates the behavior of an Ethernet LAN for broadcasting unknown MAC address and address learning.

The IETF defined two versions of VPLS, VPLS-LDP [25] and VPLS-BGP [26]. The differences between them are the approaches each one takes to establish the full knowledge of the topology. VPLS-LDP creates a full mesh of tunnels by first using UDP to determine neighbors, then establishing a TCP session to request for label mapping. VPLS ID must be defined before establishing virtual circuit (VC) labels for LSP. Both labels are prepended to Ethernet frames for fast switching. By contrast, VPLS-BGP uses the BGP approach to discover the topology and to obtain the labels. Both VPLS versions use the split-horizontal technique to broadcast a frame.

10. Category 3 – Industrial Ethernet Networks

Because of the strict constraints and highly specialized nature of Industrial Ethernet Networks, each protocol is tailored to meet the requirements of the applications that it serves. Therefore, there is a performance range that encompasses the three different classes of performance: CbA, Soft Real-Time, and Isochronous Real-Time. Deterministic behavior is a crucial requirement in this category of protocols. All nodes in the network synchronize their actions into regular intervals enabling rapid recovery that is one to three magnitudes faster than protocols from previous categories. The predominant topology here is ring or double rings because of the deterministic nature of the ring. As an extra measure, some protocols provide over protection to prevent frame loss to meet the network specifications. Tables 10 and 11 display the characteristics of these protocols.

10.1. MRP (IEC), HSR, HiPER-Ring

To manage redundancy in a ring topology, the IEC specified the Media Redundancy Protocol (MRP) under clause 5 of IEC 62439 [27,28]. Siemens and Hirschmann collaborated to show pre-versions of MRP on a ring topology for Ethernet-based networks. These were called HiPER-Ring and High Speed Redundancy (HSR) exhibiting a recovery time between 200 m and 300 ms. Each ring in a MRP managed network has a redundancy manager (RM). Initially, the RM blocks the secondary port on the ring and forwards traffic only on the primary port. The RM sends a test packet periodically around the ring. A loss of three consecutive test packets constitutes a failure on the ring. Following the fault detection, MRP has a transition time where data throughput is completely halted during which switches change state and flush their forwarding database. All nodes on the ring must be synchronized to flush the forwarding database before the nodes can resume forwarding. This delay incurs a cost of at most one trip around the ring. Optionally, the RM can also react directly to a link down notification from an intermediate node instead of waiting for the timeout from the missing test packets. If an intermediate node fails, the RM detects the ring opened state via missing test frames or link down notification frames from the two adjacent nodes, before initiating a topology change. If the RM fails, the ring transforms into a line topology. The IEC specifies two profiles: a 500 ms recovery

Table 10
Comparison charts for resilient protocols operating in Industrial Ethernet Networks.

Protocols	LDD latency	LDD method	Global reconvergence latency	Deterministic	Frameloss
MRP (IEC)	20 ms*3	Hello pkt around the ring every 20 ms, loss of 3pkt constitute a link down	200–500 ms	Yes	Yes
HSR	10 ms*3	Hello pkt around the ring	200 ms (80 ms for FDB flush)	Yes	Yes
HiPER-Ring	NA	Hello pkt around the ring	200–300 ms	Yes	Yes
PRP	NA	Varies depend on topology	0	Yes	No
HASAR	NA	Hello pkt around the ring	0	Yes	No
DRP	50 ms	Ringcheck pkt and link check pkt	<100 ms (84.5 ms for 100 Mbps)	Yes	Yes
CRP (at end node)	Vary depend on network size	Control msg and NI detection	Max of 1–2 s for 512 nodes	Yes*	Yes
BRP (at end node)	Vary depends on network size, 4.8 ms (100 Mbps, 500 nodes)	Control msg and NI detection	10 μs/1 ms/5 ms	Yes*	Yes

* Need customized tuning for each individual network.

Table 11
Comparison charts for resilient protocols operating in Industrial Ethernet Networks (continue).

Protocols	Centralized or distributed	Topology	Backup path computation	Scalability/ nodes support	Standard/industry	Synchronization
MRP (IEC)	Redundant manager	Ring	Open blocked port on ring	50 (guaranteed performance)	IEC	Yes
HSR	Redundant manager	Ring, double rings	Open blocked port on ring	NA	Siemens and Hirschmann	Yes
HiPER-Ring	Redundant manager	Ring, double rings	Open blocked port on ring	NA		Yes
PRP	Distributed	Linear, star, ring	Overprovision by running a parallel network	NA	IEC	No
HASAR	Distributed	Single ring	Overprovision by sending a duplicate traffic	NA	IEC	No
DRP	Distributed/moving manager	Ring, double ring	Reverse dir on ring	50	IEC	Yes
CRP (at end node, not in switch)	Distributed	Doubly mesh	None	2047	IEC	No
BRP (at end node)	Centralized	Doubly connected to star, line, ring	None	~500+	IEC	No

and a 200 ms recovery. Both profiles are guaranteed for up to 50 nodes in a network.

10.2. MRRT

In class three of the Industrial Ethernet Networks, the IRT class in PROFINET IO serves networks that tolerate minimal down time and almost no packet loss. This is the 1+1 protection type that exists in SONET network. From this requirement emerges the MRRT group of protocols, also known as "bumpless" isochronous real-time redundancy, which includes Parallel Redundancy Protocol (PRP) and Highly Available Substation Automation Ring (HASAR) described below. The term bumpless indicates zero packet loss in the event of a link failure.

10.3. PRP

Parallel Redundancy Protocol (PRP), as specified in IEC 62439 [27], requires two disjoint parallel networks to support redundancy. Each individual network can run its own topology management protocol, such as RSTP or MRP, and the topology can be linear, star, or ring. A substation or end host is connected to two disjoint networks running in parallel, as shown in Fig. 20. Both networks do not have any type of connection between them in order to isolate the fault into one network while continuing to forward traffic on the other. A substation sends a frame onto one network and a duplicate frame on the other network. The substation has the same MAC address on both interfaces. Both frames are sent simultaneously. When two frames arrive at the destination, ideally at the same time, one frame is forwarded to the upper layer while the duplicate frame is discarded. Under the assumption that both networks will not fail at the same time, the destination will always receive frames with zero loss in the face of failure.

10.4. HASAR

In some scenarios, a 1+1 protection like PRP is excessive and expensive. Therefore, a proposal for a cost effective yet bumpless redundancy solution was advocated in 2008 called Highly Available Substation Automation Ring (HASAR) [29]. Unlike PRP, HASAR requires only one network and confined to a ring topology. It is used for cost sensitive applications demanding bumpless redundancy. However, it is not a replacement for PRP, which addresses the high-end markets and topologies beyond rings. In HASAR, each node has two interfaces into the ring: port A and port B as shown in Fig. 21. A frame is always sent onto both ports at the same time traversing both directions. Each node forwards the frames it receives from port A to port B, except the originating node of the frame. The receiver accepts the first frame of the pair that arrives and discards the duplicate frame (if it arrives). This, a HASAR network can survive any single fault on the ring with zero frame loss.

10.5. CRP

The Cross-network Redundancy Protocol (CRP) [27], documented in IEC 62439, specifies a redundancy protocol at the end hosts in addition to redundancy protocols running in the switch. In this sense, CRP obviates the need for a redundancy manager, as with MRP. All end-nodes operate in a distributed manner and the switches are not aware of CRP, thus able to run their own redundancy protocols. Each end-host can be attached to two different switches on a single LAN as shown in Fig. 22; or they can be attached to two LANs similar to Fig. 20. However, unlike PRP, the two LANs are not necessary disjoint, as they are allowed to have connections between them. Periodically, each end-host sends out diagnostic frames on both of its interfaces to assess the network condition. The frames also contain the node's view of the network condition. When it receives a diagnostic frame on one port, it also expects the second on the other port. If a node receives no diagnostic frame, or if it does not receive the second diagnostic frame on the other port before receiving more diagnostic frames on the same port, the fault for the corresponding node is record in the Network_Status_Table. The node then uses the Network_Status_Table to decide which interface to send on to the destination.

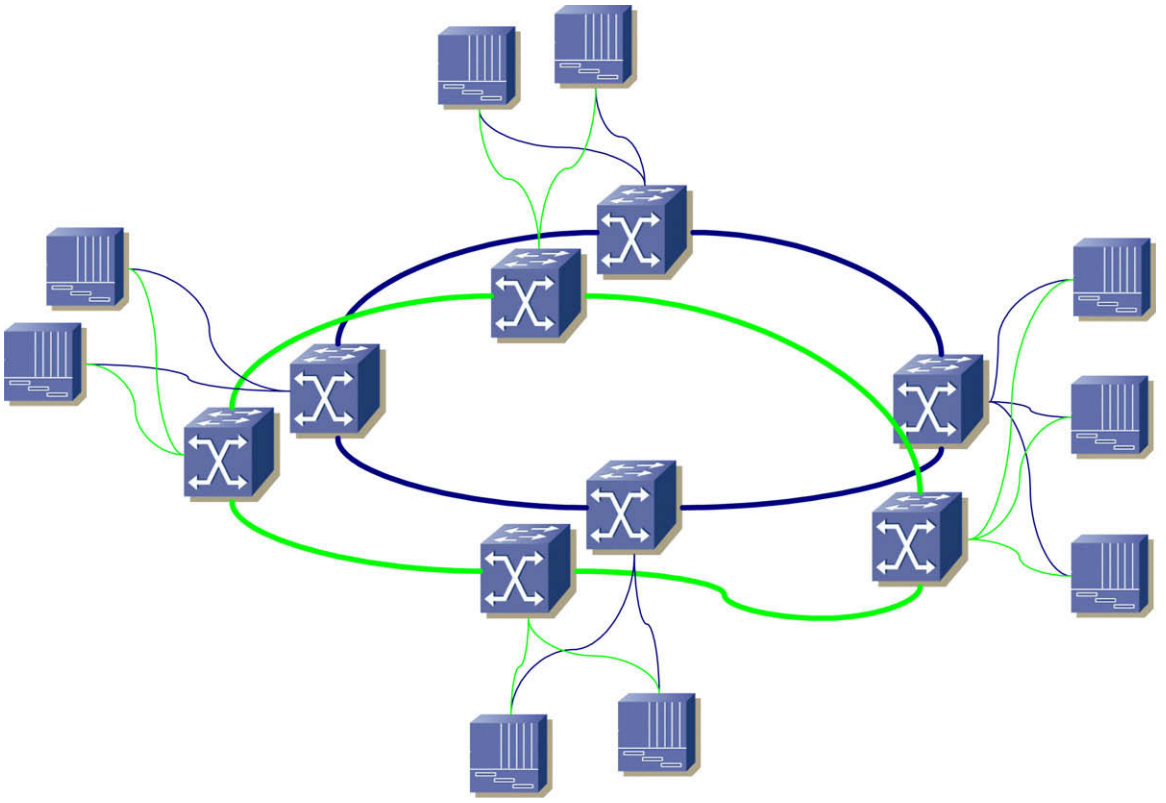

Fig. 20. PRP Ring example.

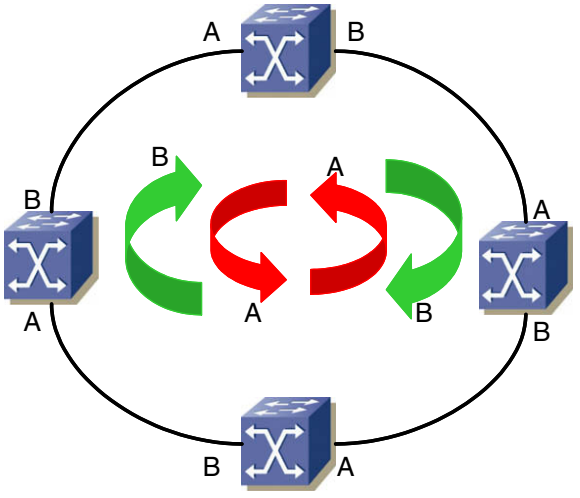

Fig. 21. HASAR ring example.

The maximum recovery time from a fault for CRP is:

$$tr = (1 + \text{Max\_Sequence\_Number\_Difference}) \times \text{tdmi} + \text{tpath} + \text{tproc}\ [27],$$

where

tr = the recovery time;
tdmi = the interval time of diagnostic frames;
tpath = the latency the packet takes to travel through the network;
tproc = the processing time of the receiving LAN redundancy entity.

The delay in a single switch is:

$$td = Nsp \times tdr \times Sp \times 8\ [27],$$

where

td = the delay in a single switch;
Nsp = the number of switch ports on a single switch;
tdr = 1/data rate;
Sp = the maximum packet size in bytes.

The following example are recovery times for various end node speeds.
Given:

tdmi = 400 ms,
$tproc = 15\mu s$,
Max_Sequence_Number_Difference = 1,
Nsp = 24, and
Sp = 1522 bytes

For all end-nodes with 10 Mbps bandwidth:

$$td = 24 \times 10^{-7} \times 1522 \times 8 = 0.029\ s.$$

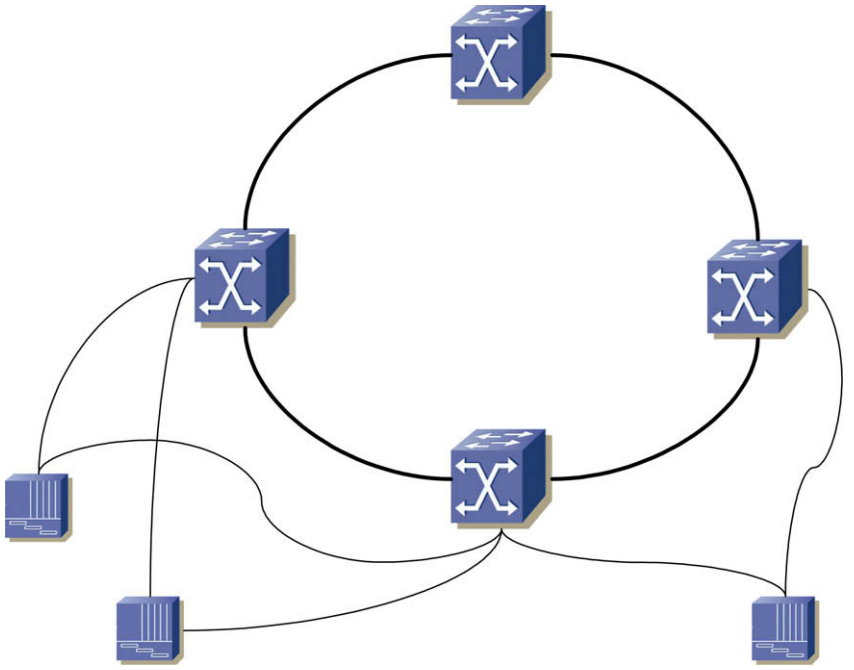

Fig. 22. CRP single LAN topology.

In a single redundant LAN with 6 switches of 24 ports each,

$$\text{tpath} = 6 \times \text{td} = 0.174 \text{ s}.$$

Thus,

$$\text{tr} = 2 \times 0.40 + 0.174 + 15\ \mu\text{s} = 0.974 \text{ s}.$$

For all end-nodes with 100 Mbps bandwidth,

$$\text{td} = 24 \times 10^{-8} \times 1522 \times 8 = 0.0029 \text{ s},$$
$$\text{tpaths} = 6 \times \text{td} = 0.0146 \text{ s}.$$

Thus

$$\text{tr} = 2 \times 0.40 + 0.0174 + 15\ \mu\text{s} = 0.8174 \text{ s}.$$

10.6. BRP

In addition to CRP, IEC 62439 also specifies another redundancy protocol at the end host called Beacon Redundancy Protocol (BRP) [27]. The difference in BRP is that it uses a central management approach as opposed to the decentralized approach in CRP. The redundancy management in this case is performed by beacon nodes. In a BRP, there are two top interconnected switches. Under each of these switches is a topology of ring, linear or star, as shown in Figs. 23–25, respectively. There are special end-nodes, called beacon nodes, which attach to the top switches and emit beacon messages on the topology intermittently. All end-nodes connected to the network on two interfaces using the same MAC address. At any time during operation, one interface is blocked while the other sends and receives data traffic, with the exception of receiving the beacon and the failure notification messages. Once a fault is detected on the active interface, the end-nodes switch to the alternate interface.

There are different kinds of fault in a BRP network. Firstly, if the leaf link faults are detectable in the end node physical layer, the recovery time is less than 10 μs [27]. Secondly, if the faults occurred in the direction of flow of beacon messages and those that are detectable in the node/switch physical layer, then the recovery time is less than 1 ms (two beacon timeouts) [27]. Lastly, if the faults occurred in the opposite direction to the flow of beacon messages, but are not detectable in the node/switch physical layer, the recovery time is the worse case:

$$\text{tfr} = \text{tnr} + \text{tid} + \text{tpcr}, \quad [27]$$

and

tfr = the recovery time;
tnr = the Node_Receive timer time out;
tid = the infrastructure propagation delay of the Failure_Notify message;
tpcr = the path check request timer time out.

For example, a network consists of 500 nodes with 8-port switches, 100 Mbps line, tnr = 2 ms, Path_Check_A_Request timer = Path_Check_B_Request timer = tpcr = 2 ms, data frame size of 1522 bytes, and the Failure_Notify message size of 68 bytes. The data frame's transmission time plus the inter-frame gap time is about 124 μs. The Failure_Notify message's transmit time plus the inter-frame gap time is about 8 μs.

In the worse case, the Failure_Notify message delay in each switch is:

$$124\ \mu\text{s} + 8\ \mu\text{s} = 132\ \mu\text{s},$$

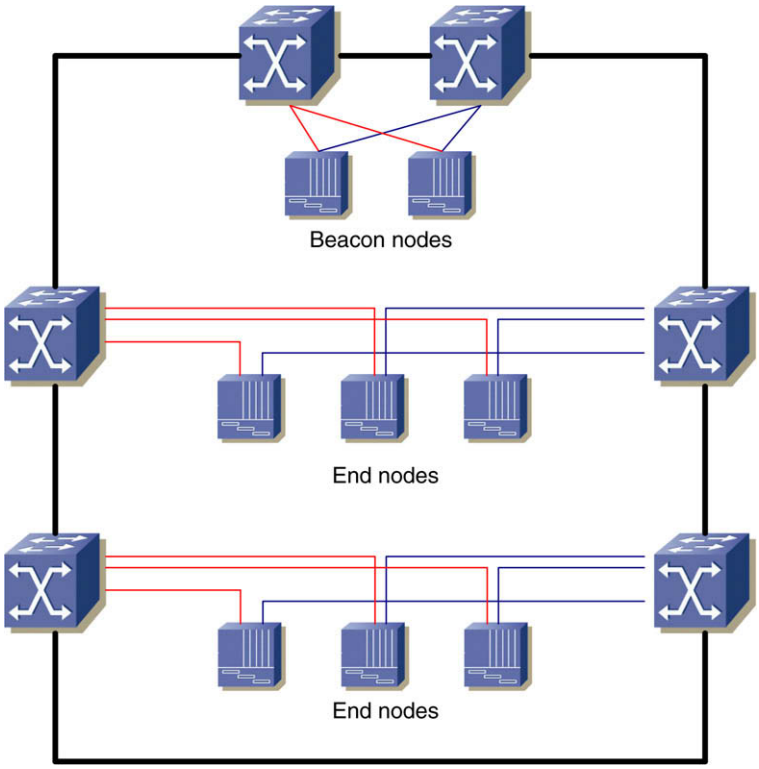

Fig. 23. BRP ring network example.

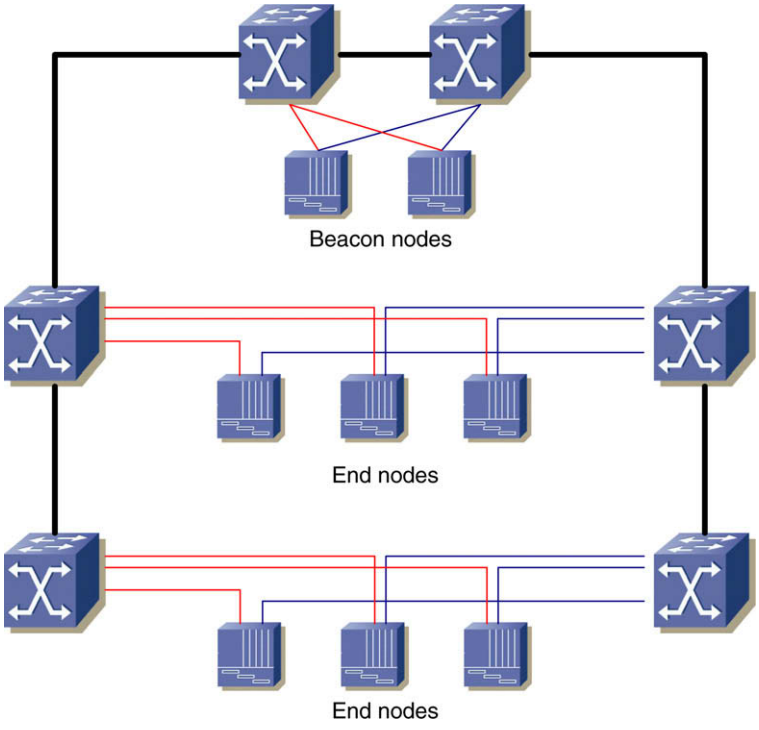

Fig. 24. BRP linear network example.

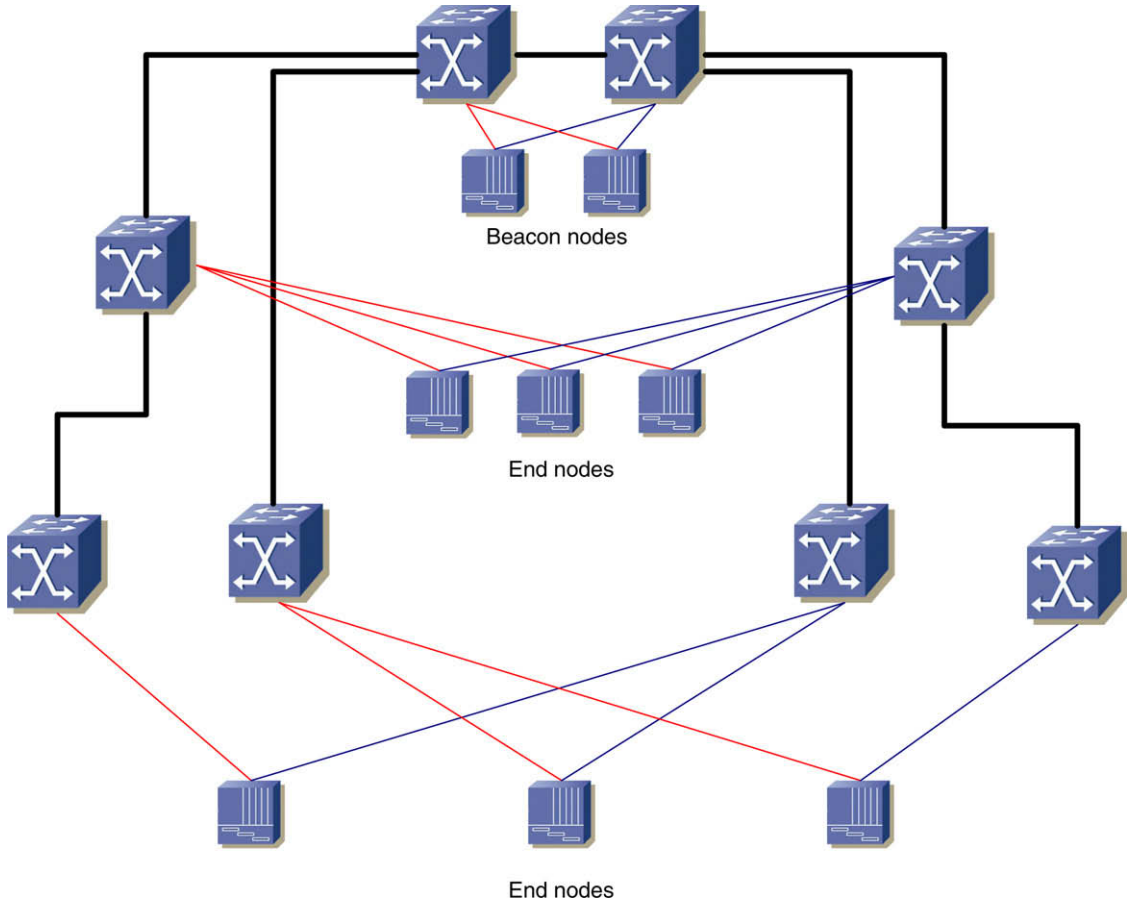

Fig. 25. BRP star network example.

then

$$\text{tid} = 8\ \mu\text{s} + (132 \times 6)\ \mu\text{s} + 8\ \mu\text{s} = 808\ \mu\text{s} = 0.81\ \text{ms}.$$

Therefore,

$$\text{tfr} = 2\ \text{ms} + 0.81\ \text{ms} + 2\ \text{ms} = 4.81\ \text{ms}.\ [27]$$

10.7. DRP

Proposed by Feng et al. as an addendum to IEC 62439 under clause 8, Distributed Redundancy Protocol [30] is a high availability network solution for a ring topology to detect a single failure and to recover in a deterministic time period. DRP synchronizes all nodes in the ring so that the scheduling can be divided into intermittent periods, called a macrocycle. Within each macrocycle, only one node is allowed to send a Ring Check frame that is used to detect a fault on the ring. Each node in the DRP ring will take a turn to send out the Ring Check frame on two of its active ring ports as shown in Fig. 26. In addition, each node sends Link Check frames to its immediate neighbor to detect any fault on the adjacent fault per macrocycle. If enough diagnostic frames are missing, the node changes the faulty link to BLOCKING mode, multicast a link down notification messages (Link Alarm frames) and Link Change frames, and flushes its forwarding database (FDB). Instead of having a central redundancy manager, by rotating the manager role among the nodes, DRP prevents single point of failure if the redundancy manager node fails.

The following example shows the maximum recovery time for a DRP network with 50 nodes at 100 Mbps on the ring ports. The parameters for the calculation of recovery time are given in Table 12 [30].

The maximum recovery time is:

$$\begin{aligned} T_r &= T_{\text{ti}} + T_{\text{to}} + T_{\text{pf}} + T^*_{\text{tt}}\text{DRPDeviceNumber} + T^*_{\text{ph}}L_{\text{ph}} \\ &= T_{\text{ti}} + T_{\text{to}} + (T_{\text{sLA}} + T_{\text{rLA}} + T_{\text{sLC}} + T_{\text{rLC}} + T_{\text{cFDB}}) \\ &\quad + (\text{Tt}_{\text{LA}} + \text{Td}_{\text{LA}} + \text{Tt}_{\text{LC}} + \text{Td}_{\text{LC}})^*\text{DRPDeviceNumber} \\ &\quad + (\text{Tph}_{\text{LA}} + \text{Tph}_{\text{LC}})^*\text{Lph} = 50 + 5 + (1+1+1+1+5) \\ &\quad + (0.005 + 0.125 + 0.005 + 0.125)^*50 \\ &\quad + (0.03 + 0.03)^*2^*50 = 50 + 5 + 9 + 17.5 + 3 \\ &= 84.5\ \text{ms}.\ [30] \end{aligned}$$

11. Future directions

Ethernet technology has come a long way since its debut thirty years ago, evolving from a simple CSMA/CD

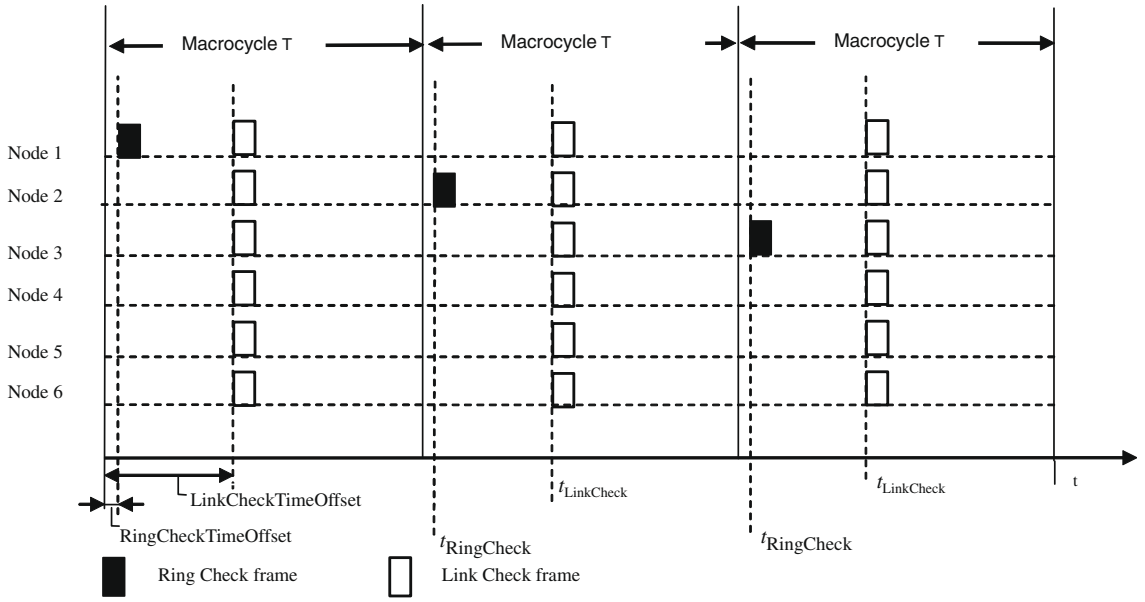

Fig. 26. DRP protocol.

Table 12
Parameters for a recovery example.

Parameter	Max. time (ms)	Descriptions
Tti	50	The time interval between two LinkCheck frames
Tto	5	The receiving timeout delay of a LinkCheck frame
Tpf	1	T_{sLA}, the transmission delay for a LinkAlarm frame in sending switch node
	1	T_{rLA}, the processing delay for a LinkAlarm frame in receiving switch node
	1	T_{sLC}, the transmission delay for a LinkChange frame in sending switch node
	1	T_{rLC}, the processing delay for a LinkChange frame in receiving switch node
	5	T_{cFDB}, the time delay to clear FDBs
Ttt	0.005	T_{tLA}, the transferring delay of LinkAlarm frame through two ring ports of a switch node
	0.125	T_{dLA}, the time delay to wait for a regular Ethernet frame with maximum size of 1518 bytes transmission before transferring a LinkAlarm frame
	0.005	T_{tLC}, the transferring delay of LinkChange frame through two ring ports of a switch node
	0.125	T_{dLC}, he time delay to wait for a regular Ethernet frame with maximum size of 1518 bytes transmission before transferring a LinkChange frame
T_{ph}	0.03	T_{phLC}, the transferring time delay of LinkChange frame on physical media
	0.03	T_{phLA}, the transferring time delay of LinkAlarm frame on physical media

LAN technology connecting nodes on a bus to a more sophisticated protocol providing Quality of Service with assured SLA. However, Ethernet has yet to mature in the new applications domains, specifically Metro Area Networks and Industrial Area Networks. In the US landscape, the legacy technologies still hold a majority of the market. Before being replaced, ISPs wish to harvest as much value as possible from legacy technologies due to the significant capital investments. Nevertheless, in many regions (particularly in Asia), Ethernet Networks are being deployed widely and rapidly in preference to legacy technologies. Further work is needed to transform Ethernet into a complete network solution. Currently, other technologies are used to fill the gaps in the service portfolio. Ethernet appears set to remain the dominate technology for office LAN networks but with the potential to dominate also in the Metro Area Networks and Industrial Area Networks.

Acronym

ASIC	Application Specific Integrated Circuit
BRP	Beacon Redundancy Protocol
CbA	Component based Automation
CRP	Cross-network Redundancy Protocol
CSMA/CD	Carrier Sense Multiple Access/Collision Detection
DRP	Distributed Redundancy Protocol
EAPS	Ethernet Automatic Protection Switching
ESRP	Extreme Standby Router Protocol
HASAR	Highly Available Substation Automation Ring
HSR	High Speed Redundancy
IA	Industrial Area Network
IEC	International Electrotechnical Commission

IEEE	Institute of Electrical and Electronics Engineers
IRT	Isochronous Real-Time
ISP	Internet Service Provider
LAN	Local Area Network
MAN	Metropolitan Area Network
MEN	Metro Ethernet Network
MRP-Foundry	Metro Ring Protocol, developed by Foundry
MRP-IEC	Media Redundancy Protocol, co-invent by Siemens and Hirschmann
MSTP	Multiple Spanning Tree Protocol
PLC	Programmable logic control; is programmed with the electronic commands for operating a machine through the various stages of its cycle
PRP	Parallel Redundancy Protocol
RRSTP	Rapid Ring Spanning Tree Protocol
RSTP	Rapid Spanning Tree Protocol
RT	Real-Time
RTE	Real-Time Ethernet
SLA	Service Level Agreement
SRT	Soft Real-Time
ST	Spanning Tree
STP	Spanning Tree Protocol
VRRP	Virtual Router Redundancy Protocol
VSRP	Virtual Switch Redundancy Protocol

References

[1] O. Kyas, Network Troubleshooting, Agilent Technologies, Palo Alto, California, 2001.
[2] ITU-T Recommendation G.1010, End-user multimedia QoS categories. URL: <http://www.itu-t.org>.
[3] ITU-T Recommendation G.114, One-way transmission time. URL: <http://www.itu-t.org>.
[4] 3GPP, Technical specification group services and system aspects service aspects; services and service capabilities, TS 22.105 V6.0.0 (2002–2009) (Release 6). URL: <http://www.3gpp.org>.
[5] Springer US, Resource management in satellite networks, QoS Requirements For Multimedia Services, ISBN 978-0-387-53991-1, Published 2007.
[6] Y. and T. Akima, A test trend of industrial real-time Ethernet, in: Proceeding of SICE-ICASE International Joint Conference, 2006.
[7] M. Rostan, Industrial Ethernet Technologies: Overview, ETG Industrial Ethernet Seminar Series, Nuremberg, Nov 2008.
[8] MEF, requirements and framework for ethernet service protection in metro ethernet networks, Technical Specification MEF2, Feb 2004. URL:<http://www.metroethernetforum.org>.
[9] S. Shah, M. Yip, Extreme networks' Ethernet automatic protection switching EAPS, RFC 3619.
[10] Extreme networks ethernet automatic protection switching evaluation report, Technical Report Reference: 80056 Issue 1.1 (30/7/03). <http://www.extremenetworks.com/technology/competitive/Default.asp>.
[11] Foundry networks foundry switch and router installation and basic configuration guide, Configuring Metro Features (Chapter 13). <http://www.foundrynet.com/services/documentation/sribcg/Metro.html>.
[12] IEEE Information technology – telecommunications and information exchange between systems – local and metropolitan area networks – common specifications, Part 3: Media Access Control (MAC) Bridges, ISO/IEC 15802-3, ANSI/IEEE Std 802.1D, 1998.
[13] IEEE standards for local and metropolitan area networks virtual bridged local area networks – amendment 3: multiple spanning trees amendment to IEEE Std 802.1Q™, 1998 Edition, IEEE Std 802.1s-2002.
[14] Extreme networks, extreme standby router protocol™ and virtual routing redundancy protocol, Whitepaper. <http://www.extremenetworks.com>.
[15] Foundry networks, foundry switch and router installation and basic configuration guide, Configuring Metro Features (Chapter 13). <http://www.foundrynet.com/services/documentation/sribcg/Metro.html#61625>.
[16] R. Hiden, Virtual router redundancy protocol (VRRP), RFC 3768 Apr. 2004. <http://tools.ietf.org/html/rfc3768>.
[17] Cisco, Hot Standby Router Protocol. <http://www.cisco.com/application/pdf/paws/9234/hsrpguidetoc.pdf>.
[18] G. Holland, Carrier class metro networking: the high availability features of Riverstone's RS metro routers, Riverstone Networks Technology Whitepaper #135. <http://www.riverstonenet.com>.
[19] S. Sharma, K. Gopalan, S. Nanda, T. Chiueh, Viking: a multi-spanning-tree ethernet architecture for metropolitan area and cluster networks, Proceedings of IEEE INFOCOM 2004.
[20] S. Varadarajan, T. Chiueh, Automatic fault detection and recovery in real time switched ethernet networks, in: Proceedings of IEEE INFOCOM, 1999.
[21] T.L. Rodeheffer, C.A. Thekkath, D.C. Anderson, SmartBridge: a scalable bridge architecture, in: Proceedings of ACM SIGCOMM, 2000.
[22] S. Acharya, B. Gupta, P. Risbood, A. Srivastava, PESO: low overhead protection for ethernet over SONET transport, in: Proceedings of IEEE INFOCOM, 2004.
[23] J.L.R. Ford, Flows in Network, Princeton University Press, 1962.
[24] J. Edmonds, R.M. Karp, Theoretical improvements in algorithmic efficiency for network flow problems, Journal of ACM 19 (2) (1990).
[25] IETF VPLS (LDP). <http://www.ietf.org/internet-drafts/draft-ietf-l2vpn-vpls-ldp-03.txt>.
[26] IETF VPLS (BGP). <http://www.ietf.org/internet-drafts/draft-ietf-l2vpn-vpls-bgp-02.txt>.
[27] IEC 62439, High availability automation networks, Feb 2008.
[28] M. Felser, Media redundancy for PROFINET IO, in: Proceeding of IEEE International Workshop on Factory Communication Systems, 2008, WFCS 2008.
[29] IEC TC57 WG10, Highly available substation automation ring, 2008.
[30] D. Feng, B. Xue, DRP-distributed redundancy protocol, Submission to IEC SC65C/MT9-HA (IEC 62439) Nov. 2007.
[31] IEEE standards for LAN/MAN CSMA/CD access method, ISO/IEC 8802-3, ANSI/IEEE Std 802.3.

Minh Huynh received his B.S. degree in computer science from the University of California, Davis in 2002. He is currently a Ph.D. candidate at UC Davis. His research interest is in Metro Ethernet Network. The focus of his research is on resilience, network load balancing, and QoS.
He had worked at Siemens Technology-to-Business on Metro Ethernet Networks and Industrial Ethernet Network. His current work at AT&T involved developing a Metro Ethernet Design Toolset. He was on the student volunteer committee for Globecom 2006 in San Francisco.

Stuart Goose has B.Sc. (1993) and Ph.D. (1997) degrees in computer science both from the University of Southampton, United Kingdom. Following a postdoctoral position at the University of Southampton, he joined Siemens Corporate Research Inc. in Princeton, New Jersey, USA. He held various positions in the Multimedia Technology Department, leading a research group exploring and applying various aspects of Internet, mobility, multimedia, speech, and audio technologies. His current position is Director of Venture Technology at Siemens Technology-To-Business Center in Berkeley, California, USA. This involves scouting for disruptive technologies from universities and startups, running projects to validate the technical and business merit of technologies, and, if successful, transferring the technologies to the business units within Siemens for commercialization.

Prasant Mohapatra received his Ph.D. in computer engineering from the Pennsylvania State University in 1993.

He is currently a Professor in the Department of Computer Science at the University of California, Davis. He has held Visiting Scientist positions at Intel Corporation, Panasonic Technologies, Institute of Infocomm Research (I2R), Singapore, and National ICT Australia (NICTA). His research interests are in the areas of wireless networks, sensor networks, Internet protocols and QoS. His research has been funded though grants from the National Science Foundation, Intel Corporation, Siemens, Panasonic Technologies, Hewlett Packard, and EMC Corporation.

He was/is on the editorial board of the IEEE Transactions on computers, IEEE Transaction on Parallel and Distributed Systems, ACM WINET, and Ad Hoc Networks. He has been on the program/organizational committees of several international conferences. He was the Program Vice-Chair of INFOCOM 2004, and the Program Co-Chair of the First IEEE International Conference on Sensor and Ad Hoc Communications and Networks (SECON 2004). He has been a Guest Editor for IEEE Network, IEEE Transactions on Mobile Computing, and the IEEE Computer.