FeT' 2009

8th IFAC International Conference on Fieldbuses & Networks in Industrial & Embedded Systems

May 20-22, 2009
Hanyang University, Ansan
Republic of Korea

PREPRINTS

TABLE OF CONTENTS

Session Ⅰ : Wireless Sensor Networks

Session Ⅱ : Control Systems

Session Ⅲ : PHY & MAC

Session Ⅳ: ZigBee

Session Ⅴ: Distributed Systems and Middleware

Industrial Session Ⅰ

Industrial Session Ⅱ

Invited Speaker Ⅱ

Session Ⅵ : Network Dependability

Session Ⅶ : Applications

Work in progress - Ⅰ

Work in progress - Ⅱ

Work in progress - Ⅲ

Highly Available and Reliable Networks based on Commercial-off-the-shelf Hard- and Software

Thomas Turek*, Heimo Zeilinger*, Berndt Sevcik*, Edgar Holleis*, and Gerhard Zucker*

**Vienna University of Technology, Institute of Computer Technology*
Gußhausstraße 27-29, 1040 Vienna, Austria
{turek, zeilinger, sevcik, holleis, zucker}@ict.tuwien.ac.at

Abstract: The steady replacement of circuit-switched networks by packet-switched networks pave the way for Voice over Internet Protocol (VoIP) into domains that long have been the market for proprietary and thus, closed communication systems e. g. for safety-critical purposes. Because the quality of a call is influenced by delay and jitter, fault detection and failover play an important role. Hence, the paper deals with these aspects and perspectives to build highly available and reliable networks based on Commercial-off-the-shelf (COTS) hard- and software. To fulfill the requirement of a sub-second network convergence time, a redundant end system and an access network are studied. The presented tests and results of Layer-2 and Layer-3 protocols show that in most albeit not all cases fault detection and failover times in the range of a few milliseconds are possible.

Keywords: Availability, Reliability, Redundancy, Replication, Protocols

1. INTRODUCTION

The past decade has seen three key developments with relevance to future communication systems for safety-critical purposes: Packet-switched networks surpassing circuit-switched networks, the emergence of a global market of powerful, but low-cost Commercial-off-the-shelf (COTS) components, and the appearance of free software and open source movements.

Packet switched networks surpassing circuit switched networks means that current and future communication systems will be based on VoIP (Voice over Internet Protocol). Advantages arise by introducing an architecture consisting of a common data transport based on the Internet Protocol (IP), an independent signaling mechanism which is presently based on the Session Initiation Protocol (SIP) and application and services located above these technologies. Up to now, VoIP implementations were found primarily in private and commercial installations. Future application fields are seen in the area of safety-critical scopes like emergency call centers. Safety can be seen in this context as the ability of a system not to cause environmentally harming events, due to loss of critical data e. g. loss or delay of signaling information and media packets.

Failures on packet transmission based on link or device failures will inevitably result in interruption of the communication. Therefore, the network has to provide redundancy to continue communication even after such troublesome events by finding alternative paths for the packet transmission – preferably without notable interruption of the communication. Thus, failover from the “old” to the “new” path must be completed in the range of a few milliseconds.

Vendors of network devices for the use in safety-critical environments want to reap the benefits of the economies of scale inherent in the mass market that developed around network equipment, even at the price of discarding their own, proprietary solutions. Building a highly available and reliable VoIP communication system out of COTS components requires careful examination of every single component. Safety aspects have to be investigated for the backbone network which connects different sites and is not under control of the customer, the access network connecting various types of end systems and for the end systems themselves. Redundancy has to be added to all sections of the communication path.

Because contemporary COTS equipment is typically outfitted with powerful processing capabilities, communication systems can now implement many functions in software that were previously realized by circuitry, including safety related features. The free software and open source movements readily provide many of the needed pieces, such as operating system, VoIP functionality, failure detection and response, and data replication.

This paper takes a look at the common problems of this safety discussion. A topology based on COTS hard- and software is presented allowing the coupling of highly available networks with commercial networks of lower availability. A redundant end system consisting of two general purpose platforms of identical design which run in warm standby mode and its interfacing to the access network is outlined. For the access network Layer-2 and Layer-3 protocols are investigated for use in the field of time-critical VoIP applications. Investigated Layer-2 protocols are Rapid Spanning Tree Protocol (RSTP) (IEEE 2001) and Link Aggregation Control Protocol (LACP) (IEEE 2000). The investigated Layer-3 protocol is Open Shortest Path First (OSPF) (Moy). Timing constraints between the access networks are not within the scope of this article. The tests are done using a specially developed device called "Link Cutter" which allows to automatically test different link error scenarios by disturbing the physical channel with, for example, resistors or capacitors.

2. STATE OF THE ART

The achievement of high availability is mainly done by applying redundancy concepts. For Ethernet networks IEEE defined protocols for adding alternate paths to address network failures. Since several years STP (802.1d) is available which ensures a loop free topology based on Layer-2 by disabling redundant links. In error scenarios these links are brought back online again. Later RSTP (802.1w) specification was released which implements a faster design of STP by keeping full compatibility with it. The important addendum was the definition of different link types like "point-to-point" and "edge". While RSTP is deactivating links LACP (802.1ad) another protocol issued by IEEE is aggregating links to a virtual link aggregation group and distributing the traffic across them. The building of such groups is also possible between different switches based on a proprietary stacking mechanism which is required to allow device failures. Proprietary solutions like HiPER-Ring (Schaub et al.) defined by Hirschmann or EAPS (Sha et al.) designed by Extreme Networks try to solve the fast convergence problem by using ring topologies. Very short timings are achieved but the downside of risks and costs associated with them have to be considered.

Protocols for allowing redundancy of Layer-3 of the OSI model are defined by the IETF. Dynamic routing protocols like OSPF can be used for path selection issues. The weak point of this protocol is the "Hello" protocol which is used for neighbor establishment and afterwards for aliveness detection. Improvements are possible by the use of Bidirectional Forwarding Detection (BFD) (Katz et al.) as a protocol independent Hello protocol in combination with Layer-3 routing. However, it will not be in the scope of this article.

Alternative approaches are the transmission of duplicated data-streams as shown in (Ogawa et al., Manousos et al.). However, for safety-critical applications the redundancy of Real-time Transport Protocol (RTP) data packets will not fulfill the requirements; instead the physical channels have to be constructed redundant, too.

3. REDUNDANT END SYSTEM

The last decade has seen the creation of a mass marked of digital equipment geared at consumers, as well as small and large businesses. Of interest are networking appliances and telephony products. Each of those is typically equipped with network ports coupled with CPUs powerful enough to run feature rich embedded operating systems like Linux or Wind River's VxWorks. Those appliances are not developed from scratch, but based on System-on-Chip (SoC) solutions. Appliance vendors only slightly customize the SoC's reference designs in their products, sometimes it amounts to merely branding the device.

While the network of the safety-critical communication system can use COTS appliances unmodified, the end systems need to be modified to support safety-critical applications. However, they can still take advantage of the economies of scale, because they can be based on the same SoC designs like the COTS appliances. The authors argue that with careful component selection SoC technology is able to provide a viable, low-cost and powerful hardware basis for safety-critical systems. So powerful indeed, that safety related functionality, like fault detection, state replication and failover, can be implemented as software.

That evokes another key development of the last ten years: The emergence of the free software and open source movements. They provide many of the software mechanisms needed: The Linux HA project (Linux) provides a framework for managing fault detection and failover. Replication can be provided by open source database products, cluster file systems or Distributed Replicated Block Device (DRBD).

We propose a redundant hardware design like the one depicted in Fig. 1, featuring two identical processor boards based on commodity SoC products, each with multiple on-chip Ethernet controllers and redundant power supplies. Storage is provided by solid storage devices e. g. Compact Flash. Both nodes are able to fulfil their function all by themselves and are visible to remote systems as separate hosts. They have their own power supplies and network ports. Internally, the two nodes are coupled by an additional high-speed serial link (i. e. Ethernet) to facilitate state replication and heartbeat exchange.

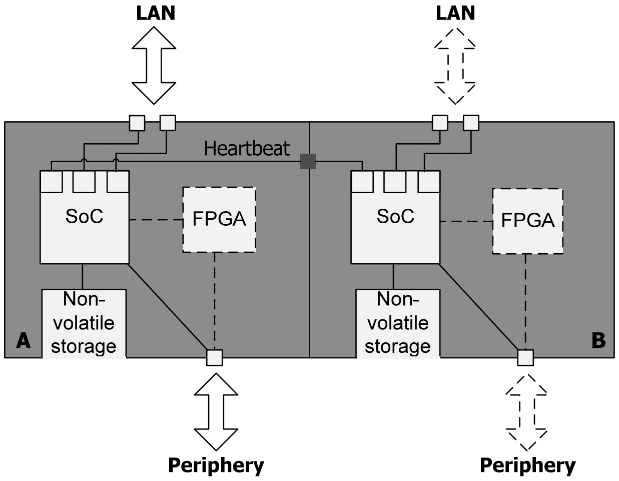

Fig. 1. Redundant end system

The two nodes operate in a warm standby configuration, where both nodes are powered up and standing by for takeover of the target applications, so each other is informed about the proper operation of the partner.

Additionally, a replication agent keeps application state synchronized between primary and secondary node. From the outside, both nodes are addressable by themselves, as well as by means of a cluster address, which at any time is held by the currently active (primary) node. This basic design can be enhanced by optionally adding a Field Programmable Gate Array (FPGA) or Digital Signal Processor (DSP) to each node. Both approaches follow common industry practice.

The module can serve as a base for many different devices that make up a safety-critical communication network. It can be used in end systems and it can implement network services. As proof of concept we realized a redundant SIP proxy, running Linux as operating system, Linux HA 1.2.5 for fault detection and failover, OpenSER 1.3.0 (OpenSER) as SIP proxy and Quagga 0.99.9 (Quagga) as routing software.

Linux HA comes in two major versions (1 and 2), both providing very different capabilities. Version 1 supports no more than two hosts, consists of two daemon processes (heartbeat and stonith) and supports a simple resource model (start, stop, monitor). Version 2 is geared towards multi-node clusters, supports a sophisticated resource model and uses an election mechanism to select the primary cluster node. Due to its simplicity, version 1 is the better match for the presented redundant end system.

State replication represents a challenge under such conditions. Neither hardware nor software was designed from scratch to function in a redundant configuration. Support for state replication is only added afterwards.

There is a fundamental trade-off involved in deciding at what level of data abstraction to replicate. It is easiest to replicate on file-system level using a highly available cluster file system or Linux Distributed Replicated Block Device (DRBD). However, in such a setting the two nodes would be but very loosely coupled in a sense that during fail-over all uncommitted state would be lost and it would be up to the communication partners to handle such a situation. The two nodes would thus appear as separate hosts.

Since a wide range of software, especially web technology, uses Relational Database Management Systems (RDBMSes) for state management, it is another option to use database technology for replication. That however bears the disadvantage that all databases operate under the assumption that data is primarily held in background storage. This assumption is violated in the presented embedded node where background storage is a scarce and vulnerable resource, both in terms of durability and Input/Output (I/O) bandwidth.

We therefore opted for a different path: We wrote a custom "database module" against OpenSER's database abstraction layer, which keeps OpenSER's state in main memory and additionally replicates it to the secondary node via the internal link. In this setting, fail-over is immediate, or at least no slower than Linux HA can handle the situation. However, the reference implementation does not include facilities for replication restart after failure of one of the nodes. Care has to be taken for the necessary processor resources and bandwidth to be available during restart procedures.

This setting cannot migrate live Transmission Control Protocol (TCP) sessions, which would be necessary for a perfectly transparent fail-over. Due to limitations in OpenSER's transaction module, it cannot either migrate ongoing SIP transactions, which are similar in concept, but simpler. However, there is nothing to suggest that the approach couldn't be extended to migrating SIP transactions using a modified software setup.

4. PROTOCOLS – TESTS AND RESULTS

In the section, we look at the test of Layer-2 and Layer-3 protocols with respect to fault detection and failover and discuss their results.

4.1 Link Cutter

The network set-up for Layer-2 and Layer-3 protocols consists not only of network devices such as switches and routers but also of network cables interfacing components with each other.

For optimal operation network cables need to work under special conditions and in case of non-compliance, severe consequences on data transmission can be observed. In order to take these considerations into account a so-called Link Cutter – a remotely controllable and configurable device – is used within our test environment. Its task is to simulate failures e. g. short circuits, loose contacts, crosstalk behavior, broken wires, or the like of a 100base-TX network cable.

4.2 Layer-2 protocols

The analysis starts with the Layer-2 protocols RSTP and LACP. RSTP is a protocol which has been designed to ensure a loop-free network topology within a bridged LAN. It is an evolution of the STP standard and achieves a reduction of the network convergence time, which can be up to a minute for STP. RSTP is designed for meshed network topologies but is also used for industrial Ethernet in ring network topologies. As the protocol is designed to manage LANs exceeding the size of the analyzed network structure by far, restrictions like the size of the network have not to be considered in the current test scenarios.

RSTP measurements are obtained by the use of the network topology in Fig. 2. The network consists of a redundant network structure, which includes two end systems – A and B – two managed switches (Dell Powerconnect 3424) – SWI1 and SWI2 – and two edge routers (Netgear 7312) – RTR1 and RTR2. RTR1 and RTR2 provide the gateway to the core network. In order to avoid the single point of failure problem, VRRP (Virtual Router Redundancy Protocol) (Hinden) is used. The traffic sink – labeled with C – represents the backbone network. As long as no component or link fails any data packets will be directed across the main route, illustrated in Fig. 2.

In case of failure the redundant counterpart will take over. SWI1 is automatically configured as root node. Both end systems form a Linux-HA cluster run on a Linux kernel 2.6.20. The master – system A – represents the traffic source. As it is focused on the protocols, the failover behavior between traffic source A and B is not discussed here.

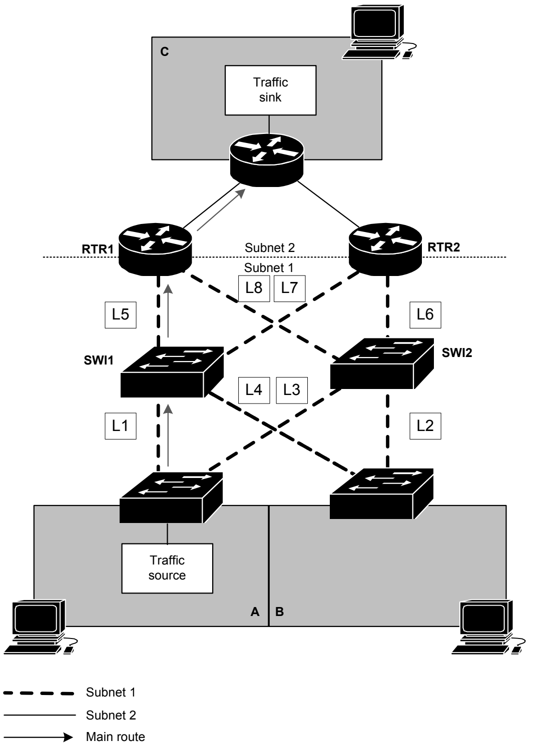

Fig. 2. Network topology – RSTP tests

The network behavior is evaluated by randomly creating different link or device failures. A User Datagram Protocol (UDP) stream is generated by the traffic source and sent across the main route. Achieved results are summarized in Table 1.

Table 1. RSTP results

	Link/Device	Failover time (s)
Link failure	L1, L5	~1.9
Device failure	SWI1 (Root node)	~2.3

Link failures of L1 and L5 – main route – result in an average downtime of 1.9 s while the failure of the root node causes a breakdown of 2.3 s. The results show that RSTP is unemployable for the stated problem. The aim of a subsecond failover stays out of range as path finding depends on protocol timing parameters which are defined by a second as a minimum in case of a root node failure.

LACP is generally used for increasing bandwidth by aggregating redundant links to one Link Aggregation Group (LAG). This section deals with the redundancy attributes of LACP.

Fig. 3 shows side A of the test network in Fig. 2. It consists of the end system A, representing the traffic source, two stacked switches, SWI1 and SWI2, the router RTR1 and the traffic sink C. The redundant side – side B – is used in case of the breakdown of RTR1 or traffic source A only. A discussion of side B will not yield additional results, regarding the usability of LACP. L1 and L3 are aggregated to a LAG as well as L5 and L8. The shown network part forms a Virtual Local Area Network (VLAN). The system is configured to transmit traffic only across the main route – SWI1. Thereby reference flaws should be avoided. For enabling the node redundancy SWI1 and SWI2 are stacked to a virtual one. A LAG between both is configured. The stacking mechanism is not standardized and therefore depends on the implementation by the hardware manufacturer. All network nodes within a link aggregation have to support LACP. At the end system L1 and L3 are aggregated by the Linux bonding kernel module.

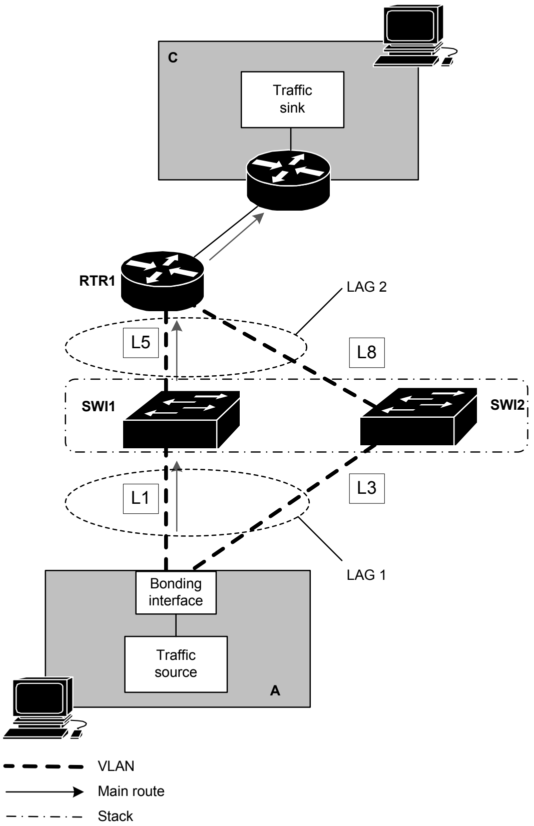

Fig. 3. Network topology – LACP tests

The network behaviour is evaluated as described for the RSTP tests. The test results are shown in Table 2.

Table 2. LACP results

	Link/Device	Failover time (s)
Link failure	L1, L5	<0.01
Device failure	SWI1 (Stack master) SWI2 (Stack slave)	~10 -

In case of link failures the convergence time depends on the detection duration at the physical layer. Hence the traffic is switched from the main route to the backup route within 10 ms. The breakdown of the master node – SWI1 – results in network downtimes up to 10 s. The reason for this long delay is located in the stacking mechanism that uses a procedure for finding a new root node, which exceeds the required time limits. As the stacking mechanisms are not standardized different manufacturers or future firmware updates may increase the performance – it is doubtful if these changes lead to a satisfying failover behavior. Hence further work focuses on Layer-3 protocols which tend to provide developments in the discussed area.

4.3. Layer-3 protocols

After having discussed the Layer-2 protocols RSTP and LACP, the analysis ends with the Layer-3 protocol OSPF.

OSPF is a protocol that belongs to the group of link-state routing protocols offering network convergence times of a few seconds. Theoretically as well as practically, the time to resume the transmission of packets after a possible link or device failure, depends on timing parameters such as the time to detect the link failure, the time to calculate an alternative path through the network, the time to propagate the LSA (Link State Advertisements) within the network and the time to update the routing tables.

OSPF implements its own failure detection mechanisms that not only provide advantages but also disadvantages with respect to high network convergence times and other limitations e. g. configurability and manageability. In this context, consider a Layer-2 switch which is connected to two neighbor devices running a Layer-3 protocol (see Fig. 4 Traffic Source – SW1 – RTR1). The Layer-2 switch hides a possible link failure to the redundant end systems. Although, the Layer-3 protocol enables the detection of the failure, the span of time depends on the Hello protocol. Its timers are in the range of one second. In the above scenario, it seems as if the use of Layer-3 protocols does not provide the proper means to reach the required sub-second link failure detection time.

Measurements are obtained by using the network topology in Fig. 4. The failover behavior of the network is studied by triggering different link or device failures. Therefore, the two customer edge routers (RTR1/2 Model Cisco 1841) are connected via Layer-2 switches (Dell Powerconnect 3424) with the redundant end system (A/B) using the OSPF routing daemon named Quagga. The measurements are performed using the Linux kernel 2.6.20 and Quagga 0.99.9.

Quagga builds on the Linux routing subsystem, which by default introduces a delay of 2 s before committing updates to the routing table. This mechanism exists, because Linux is designed to accept routing updates from multiple, uncoordinated sources, and the routing table's calculation is a costly operation. For the purpose of the measurements this mechanism is disabled. Link or device failures are simulated by shutting down ports and rebooting or turning on/off network devices.

The results for the test are listed in Table 3 showing the time required to reestablish connectivity after a failure has occurred.

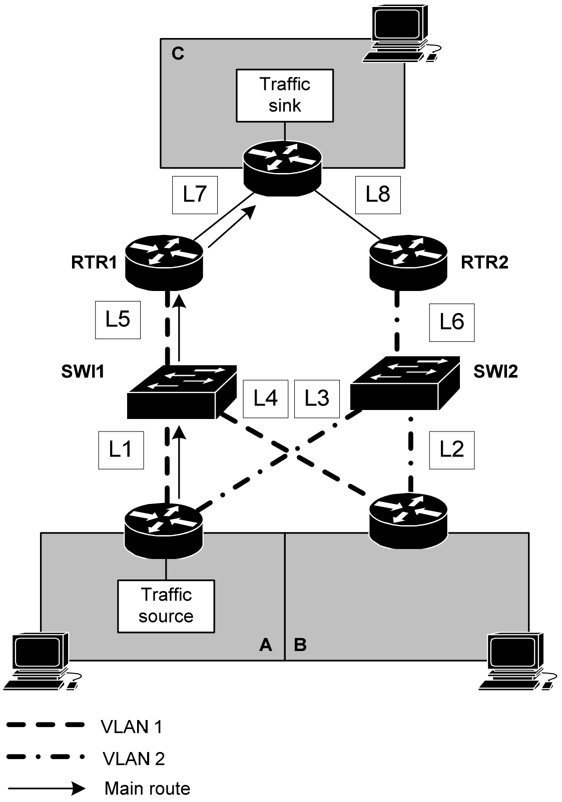

Fig. 4. Network topology – OSPF tests

The results for the OSPF configuration show that the failover time fulfills the required timing constraints whenever link status detection is possible.

Table 3. OSPF results

	Link/Device	Failover time (s)
Link failure	L2	0.006
	L3	0.040
	L6	1.139
	L8	0.066
Device failure	RTR2	1.078
	SWI2	0.055

As can be seen, a network convergence time below one second is achieved for most tested link or device failures. Problematic are the cases, where the sending node relies on expiry of the dead interval of the Hello protocol to detect link failures.

The IETF launched a draft in 2004, which focuses on the issue of sub-second link failure detection times – the BFD protocol. Today the draft is released in version 8 and will presumably be published as a Request For Comment (RFC) soon. BFD can be used for both Layer-2 and Layer-3 protocols supporting IPv4 and IPv6. However, further research is necessary to prove its performance and stability.

5. CONCLUSION

We have presented an approach to build safety-critical communication networks, taking advantage of COTS hard- and software, all the while ensuring network availability and reliability. The focus is on both, the redundant end system and the access network. Using adequate components and protocols, the proposed solution is able to reach the requirement of fault detection and failover times in the range of a few milliseconds.

As for the redundant end system, the paper covers data replication. As for the network, the Layer-2 protocols RSTP as well as LACP and the Layer-3 protocol OSPF are discussed. The Layer-2 protocols show inadequate failover performance: the RSTP protocol due to conservative timing parameters, and the LACP protocol due to unstable stacking mechanisms. Self-made changes within these protocols may lead to a better performance and stability, but conflicts with the premise of using COTS hard- and software. The Layer-3 protocol OSPF shows inadequate failover performance, too, but its use in combination with BFD seems to be new and promising. Therefore, it will attract our attention. However, BFD is still an IETF draft and its use in communication systems for safety-critical purposes has to be studied in more detail.

REFERENCES

Hinden, R. (2004), Virtual Router Redundancy Protocol (VRRP), IETF Request for Comments 3768.

IEEE Standard (2001), IEEE Std. 802.1w-2001 Media Access Control (MAC) Bridges - Rapid Reconfiguration.

IEEE Standard (2000), IEEE Std. 802.3ad-2000 Aggregation of multiple link Segments.

Katz, D., Ward, D. (2008), Bidirectional Forwarding Detection draft-ietf-bfd-base-08.txt, IETF Draft.

Linux HA Project, Available: http://linux-ha.org, 2008

Manousos, M.G., Tavoularis, A. and Economou, D. (2004) Evaluation of transmission of duplicates for packet loss recovery, Electronics Letters, volume 40 (No. 6).

Moy, J. (1998), OSPF version 2, IETF Request for Comments 2328.

Ogawa A., Sugiura K., Nakamura O., and Murai J. (2003), Enhancing the Quality of DV over RTP with Redundant Audio Transmission, ICOIN 2003, LNCS 2662, 367–375.

OpenSER Project, Available: http://www.openser.org, 2008

Quagga Routing Suite, Available: http://www.quagga.net/, 2008

Schaub, M., and Kell, H. (2003), Produkt-Analyse: HiPER-Ring vs. RSTP Redundanzverfahren mit Hirschmann Switches, ComConsult Research.

Shah, S., and Yip, M. (2003), Extreme Networks' Ethernet Automatic Protection Switching (EAPS) Version 1, IETF Request for Comments 3619.