Resilient Routing Using MPLS and ECMP

A. Iselt A. Kirstädter A. Pardigon
Siemens AG, Corporate Technology
Information and Communication,
Munich, Germany
Email:{andreas.iselt|andreas.kirstaedter}@siemens.com
antoine.pardigon@m4x.org

Thomas Schwabe
Munich University of Technology,
Institute of Communication Networks
Munich, Germany
Email:thomas.schwabe@ei.tum.de

***Abstract*— The increasing commercial importance of the Internet together with a rising number of real-time and mission-critical applications make fast resilience mechanisms a major issue for IP network planning and operation. Current IP-layer rerouting is too slow to meet these requirements. Therefore we propose a new approach combining two protocols readily available in every router: The fast local reaction of the Equal-Cost-Multiple-Path extension of OSPF operating on a network with its connectivity enhanced by the introduction of a limited number of MPLS paths in critical areas of the topology. We describe an algorithm for the determination of these MPLS paths and its optimization to obtain an equal loading of the physical network links. Numerical results on the basis of real network topologies show that already a small number of MPLS paths can offer sufficient connectivity for fast protection via ECMP. Furthermore, it can be proven that the bandwidth overhead necessary for this enhanced network resilience is as small as possible.**

I. Introduction

Network resilience is becoming one of the major aspects in the planning and operation of IP networks. The main reasons for this development are the increasing commercial importance of the Internet together with a rising number of real-time and mission-critical applications operated via IP networks.

The IP layer of data networks was originally constructed to deliver as much robustness as possible. In case of nearly any possible failure, IP networks are able to recover as long as sufficient physical connectivity is provided. The only drawback is that the corresponding rerouting processes are too slow for the services operated via today's packet networks. Standard timer values in routing protocols like OSPF lead to rerouting times in the order of 30 to 40 seconds. This is evidently to slow for multimedia and mission-critical applications.

Thus new concepts for speeding up the recovery performance of data networks are required. Due to the revenue situation in the network market, they have to deliver the necessary reconfiguration speed at low operational and equipment costs for heterogeneous equipment beneath the routing layer.

One approach could be the reduction of the timer values involved. But due to their distributed operation the routing protocols require an identical status of information in every router within the network to generate consistent re-routing decisions. At any topology change (due to failures) the corresponding knowledge has first to be propagated throughout the network. This limits the speed achievable by pure timer value reduction. Also important in this respect is the fact that at low timer values the re-routing speed becomes strongly dependent on implementation issues deeply within the routers' software and hardware [1]. Therefore, the usefulness of timer value reduction is limited and re-routing speeds below 1 second are out of reach.

Re-routing performance can also be improved by employing local reaction mechanisms. The Equal-Cost-Multiple-Path mechanism (ECMP) of OSPF allows a distribution of the packets on several outgoing links at every node in the network between source and final destination as far as there exist several shortest paths (in the cost metric selected for the routing) between the distributing node considered and the final destination. As soon as a router detects (mostly by hardware mechanisms) that an outgoing link is no longer operable it will switch over locally to forwarding the packets on the remaining links having the same cost to the destination. The great advantage of ECMP is its integration in the OSPF standard making it readily available in every OSPF router. The main limitation of this approach results from the fact that in real networks the equal cost condition is only rarely fulfilled due to connectivity limitations in the network graph as it is seen by the IP layer.

Other proposals like [2] use their own multi-path routing schemes trying to distribute the packet traffic onto as much paths into the network as reasonable. Again the packets are alternatively forwarded onto several outgoing links in every router on their way to their destinations and a local reaction provides a fast switch-over in the case of a network element failure. The main drawback of this approach is the fact that the required mechanisms still would have to be implemented in the routers available today. Also no efforts in this direction are visible in the corresponding standardization bodies.

Some network operators improve their resilience behavior by using protection switching mechanism provided by MPLS as a common intermediate layer between IP and the physical network. Since MPLS is often used for traffic

engineering purposes it is implemented in the majority of today's routers together with its Fast Re-Route (FRR) mechanism (originally provided by Cisco Systems) offering a fast protection. But network operators are often reluctant to install MPLS in addition to IP. A second network management becomes necessary and a huge number of MPLS paths have to be configured: not only those end-to-end paths needed for the transport of the IP packets but also a large number of paths for the protection of links and nodes rising sharply with the network size. In the case of significant changes in the load pattern applied to the network many or all of these MPLS paths have to be re-calculated and routed. The regarding of capacity constraints converts this into a heavy task.

As a consequence of the discussion above we propose an alternative approach that only operates in the IP layer and only relies on a interesting combination of protocols readily available in every router. We use the Equal-Cost-Multiple-Path mechanism (ECMP) of OSPF. The limitations are alleviated by enhancing the connectivity through the introduction of virtual links via MPLS path at selected and limited locations within the network. Thus, by making an intelligent use of ECMP and MPLS we are able to provide high network resilience with very short reconfiguration intervals while at the same time keeping the administrative overhead very low: since the routing and packet forwarding is still handled purely by the IP layer the network can be managed by well-known principles. Furthermore, failing MPLS path configurations won't impair the overall network operation as they would be corrected by the IP routing mechanisms.

II. Existing Protocols

In this section the main protocols that are necessary for the mechanism proposed later are described and evaluated with respect to their main advantages and shortcomings. These protocols are MPLS and OSPF with the ECMP extension. They are both available in most current router implementations.

A. OSPF

1) Basic Mechanisms: OSPF (Open Shortest Path First) is the most used interior gateway routing protocol (IGRP) in IP networks. Its current version v2 is defined in RFC2328 [3] where also several extensions to the initial version are included.

OSPF is a layer 3 link state routing protocol. Link state protocols rely on a distributed map concept. Each router maintains an identical database describing the area's topology. On this basis, each router calculates and constructs individually the shortest path(s) from itself towards any destination in the area.

The distribution of the topology information is done via Link State Advertisement (LSA) and Link State Update (LSU) messages, which are flooded to the whole area when there is a change in the state of the neighborhood (new interface appears, neighbor link or router failure) [4] [1].

Failure detection is either based on lower layer alarm escalation or on the Hello protocol. In the Hello protocol each router periodically sends Hello packets on all its outgoing interfaces. The adjacent routers detect these packets. If a router has not received Hello packets from an adjacent router within the "Router Dead Interval", the link between the interfaces of the two routers is considered down until two Hello packets are received again.

2) ECMP Extension of OSPF: With ECMP (Equal cost multi-path) a router evenly distributes the load over the fan of all available shortest paths with equal lowest cost. The main advantages of this approach are that it allows a better load distribution and a faster failure reaction. For the routing of the packets either round robin based or flow based distribution is possible. Hereby the flow based routing avoids the problems of missequenced arrival of packets. In the case of a link failure the preceding router

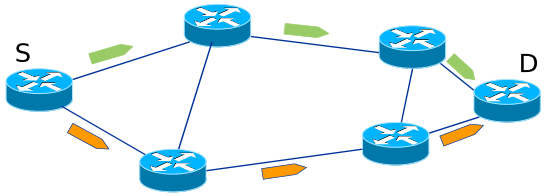

Fig. 1. ECMP routing between S and D.

automatically switches the flows to one of the remaining links of the multi-path fan (fig. 1 and 2). The fact that the reaction is local is the reason why this scheme is very fast.

Unfortunately, the fraction of cases, where multiple shortest paths are available in typical network infrastructures, is quite limited. For example in the COST example network [5], 68 out of 110 paths do not have multiple shortest paths.

Thus the fast local ECMP reaction is only possible for a limited fraction of flows. Even without the last hop problem, it is quite difficult to accommodate at least two paths on a regular network.

3) OSPF Protocol Evaluation: The main advantage of OSPF is that the protocol is completely autonomous. Once that the network administrator has set up the interface output costs of the routers, the protocol, which is active in each router of the system, will discover the topology, discover the changes, and react to these changes on its

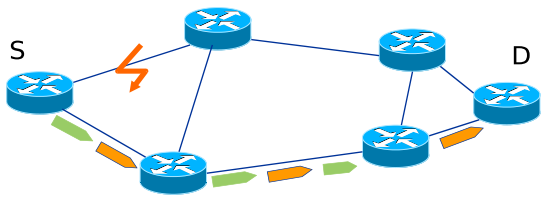

Fig. 2. When a link fails, all packets are rerouted on the remaining path.

own; the network recovery is ensured with this protocol, as long as some physical connectivity exists.

The problem is that the detection and signaling of the failures is slow when the Hello mechanism is used [6]. Moreover, the path computation and update of the forwarding information bases (FIB) take another few hundreds of milliseconds [1].

B. MPLS

1) MPLS Basic Mechanisms: When a packet is forwarded from router to router in a connectionless network, each router makes a forwarding decision independently of all the other routers, based on the packet header (i.e. destination address).

The idea of MPLS is to introduce Label Switched Paths (LSPs), that are connections across the network. The headers of all packets that use this connection are extended with a label when passing at the ingress router. Each label switched router (LSR) will just route the packet based on this label. This allows to route packets having the same destination on different paths. For the establishment of LSPs several protocols (e.g. RSVP-TE and CR-LDP) have been proposed.

MPLS can be extended with resilience mechanisms. Important schemes are end-to-end protection switching, link protection switching (e.g. [7]) and rerouting.

2) MPLS Protocol Evaluation: The main advantages of MPLS are that it allows differentiation of connections or services and it allows achieving good distribution of flows in the network. On the other hand the main drawback is that administrating a network with MPLS requires to setup a lot of working paths, $O(n^2)$, and at least as many backup paths. The planning and management of these paths is not trivial and includes still a considerable amount of human interaction.

Therefore, MPLS is still not used very often in core networks today, although most of the routers do have MPLS capabilities (see [8]).

III. Basic Concept

As stated above, ECMP can complement the efficient and robust operation of OSPF by a fast local reaction to failures but its use is limited by the physical connectivity within real IP networks. Therefore, we set up additional virtual links via MPLS paths (tunnels) between the IP routers. On the IP layer these virtual links are treated the same as physical links and OSPF/ECMP gets more degrees of freedom. Thus, the idea is to compute the placement of these tunnels, to install them and then re-run the OSPF routing algorithm, providing us with a resilient and fast reacting network. We therefore name this approach POEM (Protection using OSPF-ECMP with MPLS).

A. Where Do We Install Tunnels?

The main problem with ECMP is on the last hop: ECMP will see two different paths if we build a tunnel between A and B through C (see fig.3):

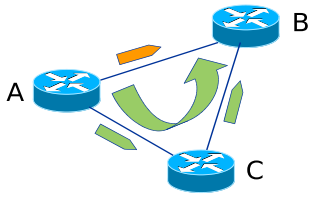

Fig. 3. How to protect a link?

- The direct one through the link A–B.
- The virtual one through A–C–B.

ECMP will split the traffic if the inserted tunnel has the same layer 3 costs as the direct link. Since the traffic inside a tunnel is always routed until the end of the tunnel no shortcuts leading to routing loops will appear. We propose to use the flow-based traffic distribution option of ECMP to avoid packet re-ordering.

Following a similar scheme, virtual MPLS links will also protect against router failures: Each OSPF path comprising having two hops long needs to be protected from a failure of the router in the middle. Since now all single-hop paths (links) and double-hop paths (nodes) in the network are protected, we are able to protect all paths in the OSPF topology.

B. Expected Recovery Time

Both, hardware-based failure detection at all physical links and RSVP-TE-Hello-based failure detection [8] at all virtual links operate in time intervals below 50ms. Therefore, ECMP can react very fast and the expected recovery times will be in the area of a few hundreds of milliseconds: 50ms for failure detection, 50ms for decision processes, and 200ms for the re-writing the forwarding database of OSPF.

IV. Algorithms

A. Basic Algorithm

The installation of a MPLS tunnel for the protection of certain link is necessary if we can't find a second path between the source and destination of the considered link in the physical topology having the same lowest costs. In this case we use a shortest path algorithm to compute a new path from source to destination avoiding the direct link. Along this path a MPLS tunnel is then set up and assigned the same costs (in the viewpoint of OSPF) as the direct link.

Similarly, to identify the tunnels for router protection, the algorithm runs through all double-hop-paths and checks whether another lowest cost path is already available. If this is not the case a new path not passing the router in the middle is computed. Its OSPF costs are

set to the original OSPF distance between the source and destination routers.

B. Optimization

Setting up the tunnels in the way described above only leads to suboptimal results if we look at the per-link bandwidth utilization: Some physical links will become overloaded. This can be avoided using a simulation of the OSPF routing behavior after the MPLS paths have been set up. The result in terms of bandwidth utilization is then feed back, e.g. as new algorithm costs (to be differentiated from the OSPF costs) in the form of a virtual cost array **VCost** into the algorithm used for path computation (see fig.4).

As the result the sources, destinations and OSPF costs of the MPLS tunnels stay unchanged; only the path layout it self evolves and a better load distribution is reached after a limited number of iterations.

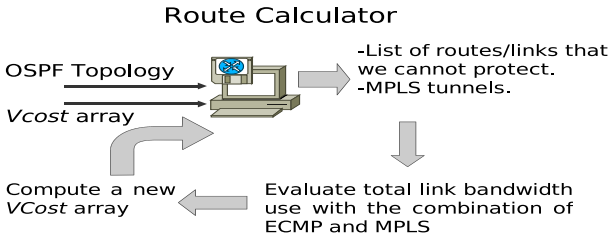

Fig. 4. Iterative optimization

V. Results

We will now provide some of the results we obtained using the various algorithms in the case of the COST 239 network ([5], Fig.5)), having 11 nodes and 52 simplex-links. Two aspects of our protocol are covered: The reduction of the number of MPLS tunnels necessary for complete network protection and the bandwidth utilization of the single links both in the presence and absence of failures. Both were investigated for the different protocol alternatives OSPF, OSPF with ECMP, POEM, and iterated POEM (with 10 iterations found to be very sufficient).

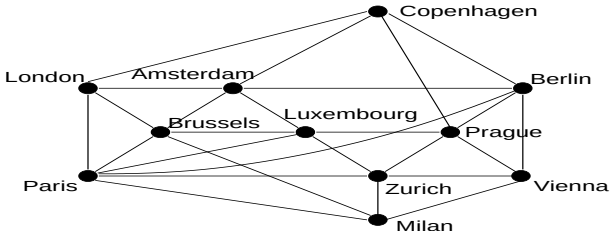

Fig. 5. COST 239 topology.

A. Number of Tunnels

The table I presents the number of MPLS tunnels that are needed to complete the topology for link protection and for node protection, compared to a network that would be entirely administrated with MPLS, with a complete setup (a backup for each working path) or with the FRR optimization (one backup path for each link). The number

Network	Link/Node protection	Complete	MPLS complete	MPLS FRR
COST	52/16	68	220	162

TABLE I

Number of MPLS tunnels to set up for protection

of MPLS tunnels to install is significantly less important with POEM than with a fully MPLS administrated network. The reason behind this result is that in a complete MPLS administration, for a topology with n nodes, the number of tunnels to install evolves in $O(n^2)$ [1] whereas in POEM, it evolves in $O(n)$ [2].

From now on, all results will be shown with link and node protection enabled.

B. Bandwidth Results

The results obtained for the load on the network with POEM are compared to those using OSPF or OSPF-ECMP, for a network in the normal state (no failures), or for a single failure (link or router). The traffic matrix used in all the tests comprises a 1 Gigabit per second demand between each of the nodes of the topology. The OSPF costs of the network had be set to one for each real link.

1) Normal State: The results are shown in table II. As expected, the average link load in the network is equivalent for OSPF and ECMP, and about 50% higher for the POEM schemes - as a result of using longer paths with the MPLS tunnels. The best load balancing is achieved using ECMP, and non optimized POEM performs quite badly as far as balancing is concerned. But the iterated optimized scheme is nearly as good as the ECMP scheme and better than classic OSPF.

We see that the distribution of the load is slightly better for ECMP than for POEM. In the case of larger topologies, the existence of more possibilities to route the paths may lead to a better load distribution for POEM than for ECMP. As expected, the optimization does not decrease the overall bandwidth use.

	OSPF	OSPF-ECMP	POEM	POEM iterated
StDev	1.59	0.84	1.92	1.09
Average	3.31	3.31	5.05	5.05

TABLE II

Bandwidth statistics.

2) Single Failure: Now we consider the maximum load of the network links in the situation after a failure is recovered. We compute the load in the network for each possible single failure and analyze the results; again, the demand matrix is 1 Gigabit per second between all nodes.

[1] complete MPLS: $n(n-1)$ working paths, $n(n-1)$ backup paths, hence $O(n^2)$, MPLS and FRR : $n(n-1)$ working paths, and as many protection tunnels as links.

[2] MPLS-ECMP, link protection: as many tunnels to build as edges; the average degree of our topologies is lesser than 5, so we build $O(n)$ tunnels. Node protection: depends on the number of paths following 2 hops, hence $O(numberofnodes * edgedegree^2)$. All in all, $O(n)$.

We will just analyze the case of the optimized POEM algorithm and compare it to the classic OSPF and ECMP schemes.

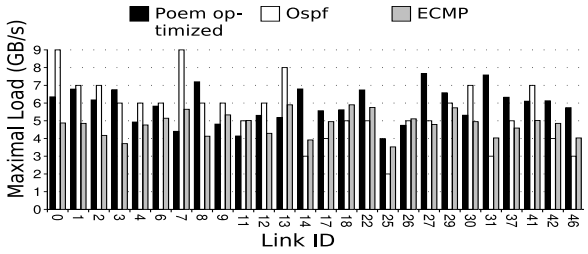

Fig. 6. Maximum bandwidth usage per link with single failure.

For the COST network, the results are the best for ECMP (fig.6), like in the normal state case. ECMP needs a maximal link capacity of just under 6 Gbps. For OSPF the maximal capacity used can go up to 9 Gbps, because no load balancing at all is used. With the POEM approach, the iterative optimization allows the maximal capacity to be of just under 8 Gbps.

Some tests were made in the single failure case to optimize the layouts of the MPLS tunnels in order to ensure better overall bandwidth usage. But the influence of modifying the link costs in advance seen by the algorithm are very limited: Since all possible link and node failures are taken in account when computing the maximum bandwidth usage on a link and because of the multiple paths used by the protocol no further optimization could be obtained.

The results we obtained with the optimized iterative algorithm are very promising given the fact that POEM enhances the resilience of the network compared to the classic OSPF and ECMP protocols. The more imperfect load distribution we observed in the normal state (compared to both OSPF protocols) has to be regarded together with the relative small load increase by POEM at the occurrence of single failures in the network.

VI. Implementation Options

There are two options for the introduction of the MPLS-enhanced OSPF-ECMP approach in networks. The algorithm could be implemented for offline calculation of the required MPLS LSPs, like a planning tool (fig.8), or online, like a configuration server (fig.7).

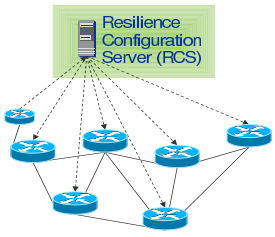

Fig. 7. POEM Resilience Configuration Server

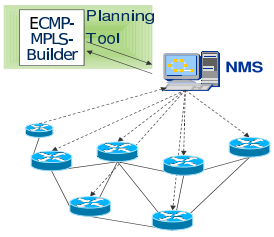

Fig. 8. Network Management System

Offline Tools

In an offline tool the topology is an input to the algorithm, which computes the LSPs the operator has then to set up. Integration in the NMS environment or least electronic transfer of this configuration information is highly desirable but not trivial.

Online Integration

A more automated approach is to integrate the algorithm in the network management and control functions. In recent publications instances like a network control server [2] have been proposed. Such instances are well suited to include the algorithms described in this paper for the calculation of MPLS LSPs. Moreover these servers can directly initiate the setup of these LSPs in the network.

VII. Conclusion

The idea described in this paper allows to extend the fast recovery of OSPF-ECMP to a complete network domain. Although it needs the use of MPLS it requires only very few and very static MPLS configurations. Therefore it does not suffer from the administrative overhead as most other MPLS resilience approaches do. The first results we obtained, for simple topologies with equal link capacities are very promising and trigger the continuation of the research. Topics to be investigated further include in particular the adaptation to heterogeneous link capacities and quality of service considerations.

Obviously the approach bears the potential to introduce fast protection switching in routed networks in the near future. More complex approaches like MPLS protection switching or fast reroute can then be replaced by simple IP routing technology.

References

[1] A. Shaikh and A. Greenberg, "Experience in black-box ospf measurement," in *ACM SIGCOMM Internet Measurement Workshop (IMW)*, nov 2001.

[2] G. Schollmeier, J. Charzinski, A. Kirstädter, C. Reichert, K. J. Schrodi, Y. Glickman, and C. Winkler, "Improving the resilience in IP networks," in *2003 Workshop on High Performance Switching and Routing (HPSR 2003)*, jun 2003.

[3] J. Moy, "OSPF version 2," http://www.ietf.org/rfc/rfc2328.txt, Network Working Group, RFC 2328, Apr. 1998.

[4] C. Harrer, "Verhalten von ip-routingprotokollen bei ausfall von netzelementen," Master's thesis, Lehrstuhl für Kommunikationsnetze,Technische Universität München, april 2001.

[5] P. Batchelor et al., "Ultra high capacity optical transmission networks," COST 239," Final report of Action, jan 1999. [Online]. Available: http://www.cnlab.ch/cost239

[6] A. Frot, S. Pasqualini, A. Iselt, and A. Kirstaedter, "Simulative comparison of the performance of mpls protection switching and ospf routing," in *International Conference on Communications (ICC)*. IEEE, sep 2003.

[7] A. Autenrieth and A. Kirstaedter, "Engineering end-to-end ip resilience using resilience-differentiated qos," *IEEE Communications Magazine*, vol. 40, no. 1, pp. 50–57, jan 2002.

[8] Cisco homepage. http://www.cisco.com. Cisco Corp. [Online]. Available: http://www.cisco.com