Overview of Resilience Mechanisms Based on Multipath Structures

Michael Menth, Rüdiger Martin
University of Würzburg
Institute of Computer Science
Germany
{menth,martin}@informatik.uni-wuerzburg.de

Arie M.C.A Koster
University of Warwick
Warwick Business School
United Kingdom
Arie.Koster@wbs.ac.uk

Sebastian Orlowski
Zuse Institute Berlin (ZIB)
Takustr. 7, D-14195 Berlin
Germany
orlowski@zib.de

***Abstract*— Multipath structures are the base for many recently developed rerouting and protection switching mechanisms. All of these methods show a similar path layout, rely on traffic distribution, and promise resilience with only little backup capacity. Therefore, it is hard to recognize their commonalities and differences at first sight. This paper provides an overview of these related mechanisms and a comparative analysis regarding their applicability in optical and packet switched technologies, their path layout, their reaction time, their dynamic adaptability, and many other aspects.**

I. Introduction

Rerouting and protection switching deviates traffic around failed network elements in failure cases. To support QoS, sufficient resources must be available on the backup paths. These resources can be shared under some conditions, e.g., if backup paths are signalled only on demand or if different connections can share common resources like in packet-switched networks. Resource sharing is the key to provide resilience with only little backup capacity. This can be achieved as follows. Figure 1 shows two primary paths (solid lines) being protected by two backup paths (dashed lines). The backup paths share a common link. Assuming only one simultaneous failure in the network, it is sufficient to provide only the maximum bandwidth of the two backup paths on the shared link. In case of a failure, any of the two backup paths can use this capacity. This is the principle how backup capacity can be reduced. Multipath structures of resilience mechanisms increase the number of paths sharing a resource and make backup capacity reduction more effective.

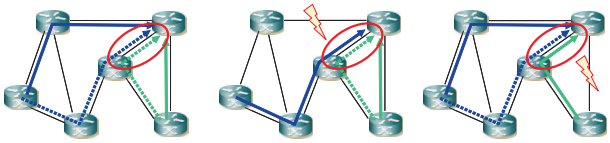

Fig. 1. Shared protection allows the usage of backup capacity by different connections in different failure scenarios.

Two major problems regarding the optimization of resilience mechanisms have been studied in literature which both try to increase backup capacity sharing.

Firstly, there is a *configuration approach*. The topology of a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is described by its set of nodes $\mathcal{V}$ and its set of links $\mathcal{E}$. Such a topology is given together with link bandwidths $c(l)$, a traffic matrix, and a set of protected failure scenarios $\mathcal{S}$ for which a specific routing mechanism should be configured. Each protected failure scenario $s \in \mathcal{S}$ corresponds to a set of failed elements in the network; in partiular, the empty set $\emptyset$ represents failure-free operation. Given a specific configuration of the resilience mechanism, the utilization of a specific link $l \in \mathcal{E}$ in a specific failure scenario $s \in \mathcal{S}$ is given by $\rho_{max}(l,s)$. The objective of the optimization is to find a configuration of the resilience mechanism that minimizes the maximum utilization $\rho_{max}^{\mathcal{E},\mathcal{S}}$ for all links $l \in \mathcal{E}$ and for all protected failure scenarios $s \in \mathcal{S}$.

Secondly, there is a *joint configuration and dimensioning approach*. The topology of a network is given together with its traffic matrix, the set of protected failure scenarios, and the specific resilience mechanism, but the link bandwidths are missing. The objective of the optimization is to find appropriate link bandwidths $c(l)$ and the configuration of the resilience mechanism such that the network costs are minimal.

Recently, several resilience mechanisms have been proposed in literature taking advantage of multipath structures to reduce either the maximum link utilization $\rho_{max}^{\mathcal{E},\mathcal{S}}$ of existing network or the network costs: protection cycles (*p*-cycles) [1], demand-wise shared protection [2], low overhead protection for Ethernet over SONET transport (PESO) [3], optimum backup capacity sharing in packet-switched networks [4], self-protecting multipaths [5], the distributed, responsive, and stable online traffic engineering protocol TeXCP [6], the optimized equal-cost multipath (ECMP) shortest path routing [7]–[9], the adaptive multipath (AMP) [10], and dynamic traffic engineering based on wardrop routing policies (REPLEX) [11]. TeXCP, AMP, and REPLEX are not resilience mechanisms in a narrow sense but dynamic traffic engineering algorithms based on multipath structures. They try to rearrange the traffic distribution within the multipath to minimize the link ultilization. So if reaction speed is not an issue, they can be used to redistribute the traffic in the network after failure.

As these approaches mentioned above allow increased backup capacity sharing, they are economically interesting for network providers. They differ slightly from each other and have, therefore, different constraints that need to be respected by optimization algorithms. Some constraints result from the

©SEE, International Workshop on Design of Reliable Communication Networks (DRCN), La Rochelle, France, October 2007

technology they are intended for, others are needed for quick reaction, for a simple and robust configuration, or for dynamic adaptation. Some multipath-based mechanisms split the traffic over several paths. The contribution of this paper is to give an overview of the presented resilience methods, to point out problems related to traffic distribution, and to analyze their applicability, commonalities, and differences.

The paper is structured as follows. Section II introduces the resilience mechanisms mentioned above. Section III summarizes some performance results regarding the accuracy and dynamics of traffic distribution algorithms. Section IV compares the resilience mechanisms and Section V summarizes this work and gives a conclusion.

II. Multipath-Based Resilience Mechanisms

Resilience mechanisms can be subdivided into protection switching and restoration mechanisms. Protection mechanisms work proactively, i.e., they set up backup paths before failures occur and switch the traffic from the primary to the established backup paths in case of failures. In contrast, restoration mechanisms establish precomputed backup paths after a failure occurred or reroute the traffic over new paths being calculated by distributed routing protocols.

In the following, we group multipath-based resilience mechanisms into three categories. We first introduce protection switching mechanisms that were designed for optical networks. Then, we explain protection switching and restoration techniques for packet-switched network. They are statically configured to work well with a given traffic matrix and a limited set of protected failure scenarios $\mathcal{S}$. Finally, we present traffic engineering (TE) methods that try to dynamically minimize the link utilizations after a failure has occurred and restoration mechanisms like rerouting have led to a new traffic distribution in the network.

A. Resilience Mechanisms for Optical Networks

For the sake of completeness, we start with 1+1 and 1:1 protection although they do not use multipath structures. We discuss several studies regarding $M:N$ protection and Demand-Wise Shared Protection, and finally the p-cycle concept. Note that these mechanisms can also be applied to packet-switched networks.

1) 1+1 Protection: With 1+1 protection, a primary path is protected by a link- and node-disjoint backup path. Traffic is simultaneously transmitted over both paths and if the signal of the primary path cannot be read, the receiver obtains it from the backup path. This approach is rather simple and robust, but also very cost-intensive since backup capacity cannot be reused at all.

2) 1:1 Protection: With 1:1 protection, a primary path is protected by a link- and node-disjoint backup path. Traffic is transmitted over the primary path only, but if the primary path fails, the traffic is switched over to the backup path. As a consequence, the backup capacity can carry additional low-priority traffic in failure-free scenarios. When failures occur on the primary path, the low-priority traffic on the backup path is dropped to accommodate the high-priority traffic from the failed primary path.

3) Dedicated $M:N$ Protection: The concept of 1:1 protection can be extended to $M:N$ protection, where N primary resources are protected by M backup resources [12], usually with $M \leq N$. If more than M primary resources fail, the service cannot be fully restored. In particular, the N primary resources may be split among several paths which are not necessarily disjoint. $1:N$ protection is a special case of $M:N$ protection where N different primary paths are protected by a single backup path.

PESO is a protection scheme for Ethernet over SONET implementing $M:N$ protection [3]. To that end, PESO takes advantage of the virtual concatenation (VC) and the link capacity adjustment scheme (LCAS) in next-generation SONET networks.

4) Demand-Wise Shared Protection: Demand-wise shared protection (DSP) is a survivability concept initially proposed in [2] for optical networks. Bandwidth for a specific demand is reserved on several paths in the network. It is dedicated to a particular demand, and part of it is reserved for backup purposes. The backup bandwidth is shared among different working paths for this demand. DSP is similar to dedicated $M:N$ protection in the sense that a set of not necessarily disjoint paths is protected by another set of paths, and that backup capacity is dedicated to a particular demand. However, there is no strict separation of working and backup paths with DSP. Instead, the same path in the network can carry both working and backup traffic. Thus, traffic is distributed over M+N different paths. If one of them fails, the respective traffic is redistributed to the other working paths. Like dedicated M:N protection, the capacity of the paths can be shared only by the traffic with the same source and destination and backup capacity sharing among different e2e aggregates is not possible. The joint configuration and capacity dimensioning problem has been studied for DSP in [13], [14].

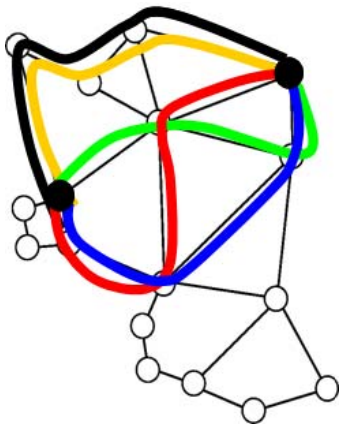

Fig. 2. Example of a bandwidth reservation with DSP; backup capacity can be shared among working paths belonging to the same demand.

An example of a DSP routing is shown in Figure 2. Five traffic units (e.g., wavelength demands) need to be routed

between a given pair of nodes, and three of these traffic units have to be protected against single node or link failures. This can be achieved by reserving altogether five units of bandwidth (e.g., wavelength channels) on the displayed paths. In case of a failure, unprotected traffic is preempted and protected traffic is rerouted over the respective paths. In contrast, 1+1 dedicated protection requires 8 capacity units in this example, namely five working paths and three backup paths. A study [15] found that despite the smaller amount of required backup capacity, the availability of protected connections in DSP is comparable with that of 1+1 protection.

5) Protection Cycles: Protection cycles (*p*-cycles) [1] have originally been proposed for ring-based optical networks where the transmission direction can be reconfigured within milliseconds. Thus, it is a typical physical layer protection scheme for, e.g., WDM or SONET networks.

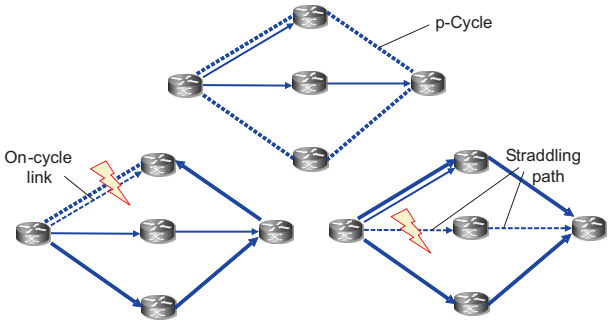

Fig. 3. Protection by *p*-cycles for cycle links and straddling paths.

Figure 3 explains the idea behind p-cycles. If an on-cycle link fails, protection is achieved by operating the cycle in the opposite direction. If a straddling link or path fails, its traffic can be rerouted over both parts of the cycle. Hence, p-cycles provide local protection. This enables fast signalling such that backup resources can be signalled on demand. As backup resources are not dedicated to specific connections in advance, backup capacity sharing among different connections is possible. It allows to achieve protection with only little backup capacity [16]–[19]. *P*-cycles were enhanced to provide protection against link and node failures and they were also discussed for application in packet-switched networks using, e.g., MPLS [20], [21]. For further details, we refer to [22, Chapter 10].

B. Resilience Mechanisms for Packet-Switched Networks

In packet-switched networks, the capacity of a link is not physically bound to any paths or connections. Therefore, it can be shared by different flows without any signalling for resource allocation or release. As a consequence, backup resources can be shared easily among primary and backup paths of different flows. In the following, we present three approaches trying to maximize this effect: optimum backup capacity sharing, self-protecting multipaths (SPMs), and equal-cost multipath IP routing.

1) Optimal Shared Protection: Murakami and Kim [4], [23] proposed a joint optimization for resilient path layout and link capacity assignment (JOA) in the context of ATM networks. Thus, there is a different routing for every failure scenario $s \in \mathcal{S}$ to minimize the overall required bandwidths. They consider both line restoration (LINE-JOA), i.e., if a link fails, the traffic rerouted to the next hop, and end-to-end (e2e) restoration (ETE-JOA), i.e., if a link fails, the traffic is rerouted directly towards its destination.

The authors propose a linear programming approach to compute cost-optimal multipath structures with respect to a continuous capacities model. However, implementing such a routing in practice may lead to technical problems. Each e2e flow is possibly routed over a different path in each failure scenario $s \in \mathcal{S}$. Thus, up to $|\mathcal{S}|$[1] different primary and backup paths need to be established. Consequently, the number of backup paths per e2e aggregate is potentially too large to be administered by routers. The exact failure need to be broadcasted through the network such that ingress routers can choose the appropriate backup paths. This is difficult because broadcasting is required during unstable network conditions. In particular, an aggregate may also be rerouted if its primary path is not affected by a failure. This raises timing issues because primary paths need to be relocated first to get free space for other backup paths. The resulting optimal path layout possibly consists of arbitrary multipaths requiring traffic splits also at interior routers. This is in contrast to the approach of choosing parallel disjoint paths, which requires traffic splits only at their ingress routers.

2) Self-Protecting Multipath (SPM): The self-protecting multipath is an end-to-end (e2e) protection switching mechanism that has been proposed first in [5], [24]. Each e2e aggregate demand d requires a SPM structure. Its path layout is depicted in Figure 4. It consists of multiple preestablished parallel paths between source and destination. The multipath should be node- and link-disjoint and can be implemented, e.g., with MPLS. The traffic of the demand d can be distributed over all parallel paths of the multipath according to a traffic distribution function $\mathbf{l}_\mathbf{d}^\mathbf{f}$. It depends on the pattern $\mathbf{f}$ of working and non-working paths. Thus, to protect the failure of any path, the SPM in Figure 4 requires four different traffic distribution functions: one for failure-free operation and three for the failure of any of the paths. If a path fails, the SPM redistributes the traffic over the working paths according to the respective distribution function. In contrast to the conventional primary and backup paths concept, the SPM does not distinguish between dedicated primary and backup paths.

To configure the SPM the following heuristic can be used. First, a k-disjoint-shortest paths algorithm finds a suitable path layout with up to k parallel paths for which algorithms from [25] can be used. Then, the traffic distribution functions $\mathbf{l}_\mathbf{d}^\mathbf{f}$ are optimized. Linear programs (LPs) have been presented in [26] to optimize the configuration of the SPM in already provisioned networks, i.e., the maximum link utilization is

[1] $|\mathcal{X}|$ denotes the size of set $\mathcal{X}$.

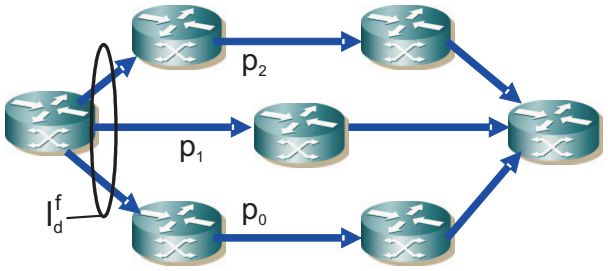

Fig. 4. The SPM distributes the traffic of a demand d over disjoint paths according to a traffic distribution function $\mathbf{l}_\mathbf{d}^\mathbf{f}$ which depends on the pattern $\mathbf{f}$ of working and non-working paths.

minimized. The LPs in [27] optimize it jointly with capacity dimensioning, i.e., the required backup capacity is minimized. The so-computed traffic distribution functions are optimal with respect to these objective functions and we call this approach the optimized SPM (oSPM). However, the oSPM potentially needs traffic splitting over several paths, which makes the implementation unnecessarily complex, and the LPs for the optimization become computationally infeasible for large networks. To bypass these problems, a heuristic integer assignment is presented in [26]. The traffic distribution functions of this integer SPM (iSPM) take only values of 0 and 1, thus, they effectively become path selection functions. The iSPM is only little less efficient than the oSPM and due to its decreased complexity and fast computability it is suitable for application in practice. The failure-specific SPM (FSPM) is presented in [28]. Its traffic distribution functions depend on the exact failure s within the failed path and not just on the pattern $\mathbf{f}$ of working and non-working paths. Therefore, the FSPM requires more traffic distribution functions. This makes the optimizing LPs and the implementation in routers more complex because the location of the exact failures within the paths need to be signalled back to the ingress router. However, the FSPM leads only to marginal improvements compared to the oSPM such that it is not recommended for application in practice.

So far the multipath structures are found using a k-shortest path algorithm and is completely decoupled from the assignment of the traffic distribution functions. A joint optimization of path structure and distribution functions still remains for future research.

P-cycles have also been proposed for application in packet-switched networks. However, it is possible to emulate them by SPMs if their requirement for disjoint paths is dropped: straddling links are emulated by 3-SPMs and on-path links are emulated by 2-SPMs and appropriate traffic distributions.

3) Single Shortest Path (SSP) and Equal-Cost Multipath (ECMP) Rerouting: In case of a failure, conventional IP routing communicates the failure of network elements to all routers in the network using a distributed routing protocol. Based on this information, all routers can recompute consistent forwarding tables. However, the signaling of the failures requires some seconds because of the standard settings of expiring timers. Sub-second reconvergence is possible, however, the timers cannot be reduced to arbitrarily small values [29]. As this mechanism is slow, it is used only to restore low priority traffic. Usually, IP routing uses a single shortest path. However, if several shortest paths exist towards the destination, the equal-cost multipath (ECMP) may be used to forward the traffic equally over the respective interfaces. ECMP is a standard option of the OSPF [30] and IS-IS [31] routing protocols. Some proprietary implementations also allow ECMP with RIP and other routing protocols [32].

ECMP is also used for IP fast reroute (IP FRR), i.e., for sub-second rerouting in IP networks. If a router detects the failure of a link, it can quickly redistribute the traffic towards a specific destination to the remaining least-cost paths if such alternatives exist. While the global routing reconvergence takes in the order of seconds, the router detecting the link failure can react much faster. However, this solution does not cover a large set of failure scenarios [33] since there are not always alternate path of equal length available. Nevertheless, the use of ECMP is one option in the IP FRR framework for fast failure reaction currently under development by the IETF [34].

IP routing can be optimized by setting link costs which are used to calculate the least-cost paths. Heuristics and exact algorithms for this have been proposed in [7]–[9], [35]. An acceleration of the heuristics was suggested in [36], and a performance study regarding the optimization quality and speed has been presented in [37]. The analysis in [38] shows that ECMP routing has a larger optimization potential than SSP routing.

C. Dynamic Traffic Engineering for Packet-Switched Networks

The resilience mechanisms presented above are optimized for a given traffic matrix and a set of protected failure scenarios. Deviations of the current traffic pattern from the expected traffic matrix or unprotected failures may lead to congestion and thus to packet loss and delay, as there is no precomputed action. This shortcoming calls for adaptive mechanisms.

Adaptive routing implemented by dynamic load-dependent link cost settings modify the shortest paths. It has been investigated in the early ARPAnet [39], but can lead to heavy oscillations.

Furthermore, adaptive traffic distribution methods have been proposed. They rely on multipath structures and in case of overload on a path, they shift traffic to alternative paths. They operate at much finer granularity than adaptive routing and, therefore, prevent oscillations.

The following dynamic TE methods are not resilience mechanisms in a narrow sense, but they cooperate with them to improve QoS. They rely on other restoration schemes like multiple parallel paths or ECMP rerouting. If the restoration has finished, dynamic TE methods try to rearrange the traffic distribution within the multipath to minimize the link utilizations. We now briefly review the dynamic TE algorithms TeXCP, AMP, and REPLEX.

1) TeXCP: TeXCP [6] distributes the traffic over a multipath structure consisting of single paths between network ingress and egress. The algorithm adjusts the traffic distribution over the paths according to feedback from probes sent along the paths. The objective of this method is to minimize the maximum link utilization in the network.

A k-shortest paths algorithm finds the required, not necessarily link-disjoint multipath structure between any ingress-egress-pair in the network. The resulting explicit path structure can be implemented by MPLS. To achieve small path delays, the length of a path is set to its propagation delay. The TeXCP agent in the ingress router periodically sends probe messages along the available paths towards the egress after the expiration of a probe timer T_p. The nodes on the path update the congestion information contained in the probe with the congestion seen on their output link and the TeXCP agent in the egress router returns this information. The ingress TeXCP agent adjusts the traffic fraction sent over the individual paths according to the probed congestion states. This is done after the expiration of a decision timer T_d that should be at least $5 \cdot T_p$ to achieve stability [6]. Two TeXCP ingress agents whose paths share one or more common links still may both increase the traffic load for the paths containing a shared link virtually at the same time. This may result in overload for this link leading to oscillations. Since the core routers know all ingress-egress-pairs leading over their links, they attach explicit feedback to the respective probes indicating the maximum amount of utilization increase allowed for each agent. Additionally, to keep the propagation delay low, the TeXCP agents prefer shorter paths.

2) Adaptive Multipath Routing (AMP): The Adaptive Multipath (AMP) [10], [40] is a load balancing mechanism operating on arbitrary multipaths. To generate large multipaths, the relaxed best paths criterion is used for routing decisions. Any neighboring node which is closer in terms of link costs to the destination than the current node is considered a viable next hop for multipath routing. However, AMP can be integrated into any routing protocol that discovers multiple paths, provided that routers can exchange information messages with their upstream neighbors.

Given that a router knows several paths to the destination, a node running AMP distributes the load over those paths. In case that one of the paths is congested, it reduces its share of the load and redistributes it to other less congested paths. However, load balancing based on only local information is not sufficient to minimize network-wide link utilizations. Therefore, backpressure messages (BMs) from neighboring routers inform about the congestion state further downstream (see Figure 5). These BMs contain information about the degree to which the node is responsible for the congestion on the outgoing links of the neighbors. For instance, in Figure 5, router Y_0 integrates both local load information about its congested link (Y_0, X) and its link (Y_0, Z_3) and feedback from other BMs into one BM to router Z_1. The BM to router Z_3 in turn contains load information about the links (Y_0, X) and (Y_0, Z_1) and feedback from other BMs. Further details on the exact creation of the BMs, considerations on stability issues, and a simulative evaluation of the concept can be found in [10], [40].

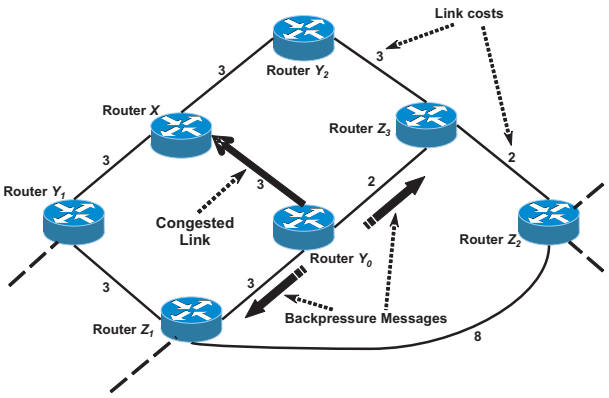

Fig. 5. An AMP capable router recognizes congestion and uses backpressure messages to inform its neighbors about the degree to which they are responsible for this congestion.

3) REPLEX: Similarly to AMP, REPLEX [11] relies on local knowledge and aggregated feedback from other routers. However, both the local knowledge and the feedback are given not only per interface, but per destination and interface. In addition, aggregated feedback from other routers is desirable, but not vital for its operation. REPLEX can be deployed on top of virtually any routing protocol.

The algorithm can be configured for different objective functions like, e.g., minimum path delay or minimum link utilization. The analysis in [11] is based on minimum link utilization which is also the appropriate measure to rebalance the traffic distribution after rerouting due to failures. REPLEX distributes the traffic at router r to a destination t over the possible outgoing interfaces (r, v_i) to next hops v_i according to a destination-dependent weight function $w(r,t,v_i)$. Any router r periodically performs measurements $l(r, v_i)$ concerning the target optimization function for every destination t in its routing table. In addition, the next hops signal back aggregated feedback about the condition of their destination-specific paths. A combination of the local measurements and the feedback leads to an adjustment of the current traffic distribution weights $w(r,t,v_i)$.

The adjustment of the weights $w(r,t,v_i)$ follows a so-called $(\alpha - \beta)$-exploration-replication policy leading to the name REPLEX. This policy is inspired by game theory. Exploration means that all paths are equally likely to be examined whether a traffic shift towards them leads to a reduction of the target function. This leads to the exploration of new, possibly unused paths. Replication prefers paths that are already popular assuming that this is a quality indication. The algorithm is controlled by a set of parameters that influence its stability. The stability is backed by theoretical considerations and simulative parameter studies in [11].

III. Accuracy and Dynamics of Load Balancing Algorithms

In packet switched networks, traffic splitting onto individual paths requires load balancing algorithms on the packet or flow level. Due to stochastic effects of the traffic, the load balancing accuracy may differ fundamentally from the specified target values. This aspect impairs the capacity savings potential of the multipath concept and is important for the resource management to dimension the required backup capacity. Thus, we comment on the accuracy and dynamics of load balancing algorithms in this section.

Packet level versus flow level: The most intuitive approach to load balancing in packet switched networks is the round robin distribution of consecutive packets. Due to different delays on the paths, this can lead to packet reordering if two packets of the same flow are sent over different paths and has detrimental effects on, e.g., the TCP throughput [41]. Thus, load balancing on the flow-level is required. For this purpose, hash-based schemes [42] have been proposed since it is not feasible to store the mapping of the flows to the paths for each flow explicitly. The hash is computed on the flow ID and a lookup operation on the hash value from a much smaller domain yields the corresponding outgoing path.

Static versus dynamic hashing: The typical data structure of load balancing algorithms, table-based hashing, is shown in Figure 6. The hash values are mapped to intermediate bins and the number of bins connected to the paths determines the intended traffic distribution. If the assignment of hash values to the paths is fixed, these methods are referred to as static. Dynamic algorithms re-adjust the assignment of bins to the paths periodically.

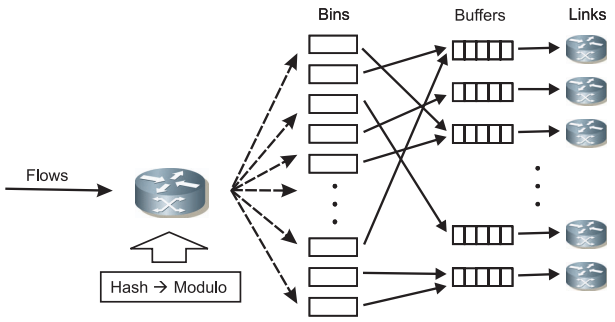

Fig. 6. Data structure of table-based load balancing algorithms.

Accuracy: The basic accuracy of static and dynamic hash-based load balancing techniques has been studied in [43]. In a very simple simulation scenario where one node splits traffic onto two paths only with target distribution $50\% : 50\%$, a difference of up to 30% from the target value was observed with non-negligible probability. In a more complex simulation scenario where a node splits traffic over four paths with target distribution $40\%, 30\%, 20\%, 10\%$, dynamic mechanisms increase the accuracy by factor 10. This demonstrates the need for dynamic operation of load balancing algorithms, but it also emphasizes that the accuracy of load balancing algorithms is relevant for resource planning.

Dynamics: Dynamic load balancing is the only option to achieve an acceptable load balancing accuracy, however, the reassignment of the mapping between bins and paths may lead to packet reordering. Furthermore, dynamic decisions made at one node influence the decisions made at other nodes further downstream. The hash operation polarizes the traffic, i.e., only flows with certain properties of their flow ID reach the downstream routers, and reassignments shift part of the input traffic away from one node to another one, changing the statistical properties there. In [44] the authors evaluated methods to cope with these interdependencies and dynamics. There are effective anti-polarization mechanisms and with the most efficient algorithms the dynamics in terms of flow reassignments increases only slightly along the paths. But still, when configuring multipath mechanisms, flows should not undergo balancing steps too often.

The results concerning the accuracy and dynamics of load balancing algorithms have different effects on the described multipath mechanims. The SPM and ECMP rely on pre-configured traffic distribution targets and, therefore, the observed inaccuracies must be taken into account for capacity planning. TeXCP, AMP, and REPLEX react to inaccuracies inherently, however, the dynamics must be considered. We further comment on the implications of the observed accuracy and dynamics of load balancing algorithms in Section IV-C.

IV. Comparison

In this section, we compare the approaches presented above with regard to applicability, ability for capacity sharing, reaction speed, and flexibility and we point out open research issues. The main points of our comparison are summarized in Table IV and explained in the following.

A. Applicability and Potential for Capacity Sharing

The 1+1, 1:1, and M:N protection mechanisms and DSP do not require backup capacity sharing among different demands and can therefore easily be implemented in optical technologies. Other mechanisms taking advantage of flexible backup capacity sharing can be implemented only in packet-switched networks since fast claim and release of optical resources is too challenging. *p*-Cycles are an exception since they provide local protection and, thus offer fast signalling on demand.

a) Backup capacity reuse: As 1+1 protection carries traffic simultaneously both on the primary and on the backup path, backup capacity cannot be reused for other purposes. With 1:1 and M:N protection as well as with DSP, the backup capacity is only used in failure scenarios and can therefore be used to carry low priority traffic under failure-free conditions.

b) Backup capacity sharing within a single demand: With M:N protection, several primary paths can share a set of commonly used backup paths. The backup paths are established in advance, but the traffic assignment is done only in case of a failure. This constraint makes M:N protection mechanism well applicable for optical networks. Less than 100% backup capacity is required to protect primary paths due

	Technology	Capacity sharing	Path selection	Signalling	Reaction time	Traffic distribution	Failure coverage
1+1	opt & ps	no	multiple explicit e2e paths	egress router decision	< 10ms	no	single failures
1:1	opt & ps	only in ps	multiple explicit e2e paths	e2e monitoring & ingress router decision	< 100ms	no	single failures
M:N/DSP	opt & ps	partly	multiple explicit e2e paths	e2e monitoring & ingress router decision	50-150ms	preplanned	preplanned failures
p-cycles	opt & ps	partly	multiple explicit local paths	on-path local decision	< 100ms	static	preplanned failures
Optimal shared protection	only for ps	yes	multiple explicit e2e paths	network-wide failure notification & decision	?	preplanned	preplanned failures
SPM	only for ps	yes	multiple explicit e2e paths	e2e monitoring & ingress router decision	< 100ms	preplanned	preplanned failures
TeXCP	only for ps	yes	multiple explicit e2e paths	e2e measurements & ingress router decision	several seconds	adaptive	arbitrary failures
ECMP	only for ps	yes	shortest paths	local or network-wide failure notification & local or all routers decision	< 100ms (IP FRR), several seconds	static	arbitrary failures, no full coverage for fast reaction
AMP	only for ps	yes	arbitrary paths	local measurement & local decision	several seconds	adaptive	arbitrary failures
REPLEX	only for ps	yes	arbitrary paths	local measurement & local decision	several seconds	adaptive	arbitrary failures

TABLE I

MAIN CHARACTERISTICS OF MULTIPATH RESILIENCE STRUCTURES IN OPTICAL (OPT) AND PACKET-SWITCHED (PS) NETWORKS.

to backup capacity sharing. However, the backup capacity can be shared only among primary paths with the same source and destination. This is due to the application in optical networks where connection setup comes with an exclusive claim of physical resources. With 1:1 protection, backup capacity sharing is not possible within a single demand. Therefore, if it is applied in optical networks, $\geq 100\%$ backup capacity is required if primary and backup paths are preestablished.

c) Backup capacity sharing among different demands: To share capacity among different demands, the capacity must not be bound in advance to any specific e2e demand. In optical networks this could be achieved by setting up and releasing connections on demand. In general, this is complex and increases the failover time because the backup paths still need to be signalled. With the local protection mechanism *p*-cycles on the other hand, only local signalling is required to set up the backup paths, leading to quite short recovery times. In packet-switched networks, backup capacity sharing is trivial because capacity is not physically bound to any demand. Most of the discussed resilience mechanisms require backup capacity sharing capabilities: optimal shared protection, SPM, TeXCP, ECMP, AMP, and REPLEX. These mechanisms need to increase the traffic share of any demand on any link depending on the current network-wide failure situation, and, therefore, implementing these mechanisms in optical technologies with a priori claim of physical resources would be very complex.

B. Path Selection, Signalling, and Reaction Time

The path layout of 1+1, 1:1, and M:N protection, as well as the one of SPM and DSP, follows explicit e2e paths. Note that these paths are preferentially but not necessarily disjoint.

With 1+1 protection, a deteriorated signal is recognized at the tail end router and the backup signal is used on demand. This can be done within a few milliseconds. The other mechanisms require e2e link management protocols monitoring the availability of the paths, e.g., by hello messages and explicitly reporting a path failure to the head end router which then takes countermeasures. This kind of signalling requires time and, therefore, the reaction time of these mechanisms is in the order of 100 ms.

With p-cycles, paths for e2e demands consist of several subpaths that are protected by p-cycles. End nodes of a failed link – on-path or straddling link – locally recognize the failure and change the operation of the protecting p-cycles to provide a bridge over the failed network element. The objective is to achieve this goal within 50–150 ms.

Joint optimal capacity and flow control allows for arbitrary multipath structures. As this approach stems only from a theoretical study, no signalling mechanisms are proposed, but the practical implementation will be difficult due to the reasons mentioned in Section II-B.1. Therefore, it is hard to give estimates for reaction times.

TeXCP also sets up explicit paths and monitors their utilization. Based on this feedback, traffic is slowly redistributed onto other paths. TeXCP takes 10-15 iterations to converge to within 10% of an optimal (i.e. minimal) link utilization. Considering the fact that one iteration is recommended to take place every $T_d = 500$ms, the traffic redistribution process of TeXCP is slow. Thus, the reaction time of TeXCP lies in the order of several seconds. Therefore, TeXCP can be used as a resilience mechanism only if this relatively long reconvergence time can be afforded.

In contrast, ECMP, AMP, and REPLEX have a rather general multipath structure. The one of ECMP is constrained by IP routing while AMP and REPLEX can be combined

with any routing. As a result of the more complex path layout, the notion of explicit e2e paths vanishes. In contrast to TeXCP, congestion is measured on links, and to indicate congestion, link utilization and feedback from downstream links is recursively signalled to local upstream neighbors taking into account the contribution of the local neighbor to the observed congestion. Based on this information, traffic distributions are changed and a balanced resource utilization is restored when link failures lead to overload due to redirected traffic. Similarly to TeXCP, the concepts AMP and REPLEX require several seconds until a balanced resource utilization is achieved. With ECMP a fast reaction can take place as proposed with IP Fast Reroute. TeXCP follows a connection-oriented concept, whereas ECMP, AMP, and REPLEX are connection-less approaches operating on local information.

C. Traffic Distribution and Failure Coverage

1+1 and 1:1 protection do not allow traffic distribution. M:N protection and DSP require preplanned traffic distribution for specific failures of primary paths.

With *p*-cycles, arbitrary distribution of backup traffic over several paths is not possible, but overlapping *p*-cycles can be arranged in such a way that a certain traffic distribution is achieved. The distribution of the backup traffic to different *p*-cycles must be preplanned. Example: Consider a link with a capacity of 60 MBit/s. It may be a straddling link for a first p-cycle with 40 MBit/s and an on-path link for a second p-cycle with 20 MBit/s whose other half coincides with one half of the first p-cycle. Now, 20 MBit/s of the traffic is deviated over both sides of the first *p*-cycle if the considered link fails, and 20 MBit/s of the traffic is deviated over the backup of the second *p*-cycle. Finally, a preplanned traffic distribution of 1:2 is achieved.

Optimal shared protection and the SPM use explicit traffic distribution, i.e., the traffic should be distributed onto different backup paths according to a preplanned distribution. The optimum failure-specific traffic distribution is calculated offline and applied if this failure occurs. However, it is not always possible to realize the required traffic distribution in practice due to inaccuracies of load balancing algorithms (cf. Section III. This deteriorates the optimality of the preplanned solution. Similarly, the preplanned traffic distribution becomes suboptimal or even counterproductive if the traffic matrix changes substantially. If unplanned failures occur, traffic distributions must be adapted online, e.g. by interpolating appropriate distributions, but also they are suboptimal in the end.

The traffic distribution of ECMP is static since the traffic is distributed equally to all outgoing interfaces for the corresponding destination. However, like above, equal traffic splits are rather the objective and the traffic distribution result depends on stochastic effects.

TeXCP, AMP, and REPLEX use adaptive traffic distribution that is controlled by online feedback. Thus, they are self-regulating systems. Rather than prescribing a certain traffic distribution result, the existing traffic distribution is modified by changing the partitioning of the hash space towards a certain objective. Another control loop allows for additional corrections if the effective change of the traffic distribution does not lead to the desired effect. The drawback of adaptive traffic distribution is clearly that it leads to a slow reaction time when link utilizations are unbalanced due to likely failures. For such scenarios, preplanned traffic distributions can provide fast reaction times. However, adaptive traffic distribution can well cope with changed traffic matrices and unprotected failure scenarios.

V. Summary and Conclusion

The contribution of this paper is an overview of recently studied protection switching, rerouting, and dynamic traffic engineering (TE) mechansims based on multipath structures. So far, their relation to each other was unclear in the research community which has led to frequent misunderstandings. The common objective of the multipath-based resilience mechanisms is the reduction of backup capacity requirements by increased backup capacity sharing or – in other words – a reduction of link utilization in working and failure scenarios by an improved traffic distribution.

1+1,1:1, DSP, M:N protection, and p-cycles can be applied in optical networks while all other mechanisms are suitable rather for packet-switched networks only. Self-protecting multipaths (SPMs) implement protection switching, are relatively simple, and can be efficiently optimized for preplanned failure scenarios. Other TE mechanisms like TeXCP, AMP, or REPLEX adaptively restore a balanced link utilization after failures occurred and redirected traffic lead to congestion.

Our study compared the applicability and the potential for backup capacity sharing of these mechanisms, their basic path layout, signalling and their time to react to failures, and, finally, their traffic distribution approach and failure coverage. We also gave an overview of the implications of the accuracy and dynamics of load balancing algorithms required to distribute the traffic over multipaths. Our discussion aimed at classifying the different resilience mechanisms and to improve their understanding by contrasting them to similar approaches. We believe that this overview is useful for the improvement of existing and the development of new resilience mechanisms.

Acknowledgements

The authors would like to thank Dieter Stoll (Alcatel-Lucent) for valuable pointers and Ivan Gojmerac (Forschungszentrum Telekommunikation Wien, FTW) for his input on AMP.

References

[1] W. D. Grover and D. Stamatelakis, "Cycle-Oriented Distributed Preconfiguration: Ring-Like Speed with Mesh-Like Capacity for Self-Planning Network Restoration," in *IEEE International Conference on Communications (ICC)*, June 1998, pp. 537–543.

[2] A. M. C. A. Koster, A. Zymolka, M. Jäger, and R. Hülsermann, "Demand-wise shared protection for meshed optical networks," *Journal of Network and Systems Management*, vol. 13, no. 1, pp. 35–55, 2005.

[3] S. Acharya, B. Gupta, P. Risbood, and A. Srivastava, "PESO: Low Overhead Protection for Ethernet over SONET Transport," in *IEEE Infocom*, Hong Kong, 2004.

[4] K. Murakami and H. S. Kim, "Comparative Study on Restoration Schemes of Survivable ATM Networks," in *IEEE Infocom*, Kobe City, Japan, Apr. 1997, pp. 345 – 352.

[5] M. Menth, A. Reifert, and J. Milbrandt, "Self-Protecting Multipaths - A Simple and Resource-Efficient Protection Switching Mechanism for MPLS Networks," in *3rd IFIP-TC6 Networking Conference (Networking)*, Athens, Greece, May 2004, pp. 526 – 537.

[6] S. Kandula, D. Katabi, B. Davie, and A. Charny, "Walking the Tightrope: Responsive Yet Stable Traffic Engineering," in *ACM SIGCOMM*, Portland, OR, Aug. 2005.

[7] A. Nucci, B. Schroeder, S. Bhattacharyya, N. Taft, and C. Diot, "IGP Link Weight Assignment for Transient Link Failures," in *18th International Teletraffic Congress (ITC)*, Berlin, Sept. 2003.

[8] D. Yuan, "A Bi-Criteria Optimization Approach for Robust OSPF Routing," in *3rd IEEE Workshop on IP Operations and Management (IPOM)*, Kansas City, MO, Oct. 2003, pp. 91 – 98.

[9] B. Fortz and M. Thorup, "Robust Optimization of OSPF/IS-IS Weights," in *International Network Optimization Conference (INOC)*, Paris, France, Oct. 2003, pp. 225–230.

[10] I. Gojmerac, T. Ziegler, F. Ricciato, and P. Reichl, "Adaptive Multipath Routing for Dynamic Traffic Engineering," in *IEEE Globecom*, San Francisco, CA, Nov. 2003.

[11] S. Fischer, N. Kammenhuber, and A. Feldmann, "REPLEX – Dynamic Traffic Engineering Based on Wardrop Routing Policy," in *CoNEXT (formerly QoFIS, NGC, MIPS)*, Dec. 2006.

[12] D. W. Griffith and S. Lee, "Dynamic Expansion of M:N Protection Groups in GMPLS Optical Networks," in *Workshop on Optical Networks*, Aug. 2002.

[13] R. Wessäly, S. Orlowski, A. Zymolka, A. M. C. A. Koster, and C. Gruber, "Demand-wise Shared Protection Revisited: A new Model for Survivable Network Design," in *International Network Optimization Conference (INOC)*, Lisbon, Mar. 2005, pp. 100–105.

[14] C. Gruber, R. Wessäly, S. Orlowski, A. Zymolka, and A. M. C. A. Koster, "A Computational Study for Demand-wise Shared Protection," in *International Workshop on Design of Reliable Communication Networks (DRCN)*, October 2005, pp. 421–428.

[15] R. Hülsermann, M. Jäger, A. M. C. A. Koster, S. Orlowski, R. Wessäly, and A. Zymolka, "Availability and Cost Based Evaluation of Demand-wise Shared Protection," in *ITG Workshop on Photonic Networks*. Leipzig, Germany: VDE Verlag GmbH, 2006, pp. 161–168.

[16] W. D. Grover and D. Stamatelakis, "Bridging the Ring-Mesh Dichotomy with *p*-Cycles," in *International Workshop on Design of Reliable Communication Networks (DRCN)*, Apr. 2000.

[17] W. D. Grover, J. Doucette, and M. Cloqueur, "New Options and Insights for Survivable Transport Networks," *IEEE Communications Magazine*, no. 1, 2002.

[18] D. A. Schupke, C. G. Gruber, and A. Autenrieth, "Optimal Configuration of *p*-Cycles in WDM Networks," in *IEEE International Conference on Communications (ICC)*, New York, 2002.

[19] D. Rajan and A. Atamtürk, "Survivable network design: routing of flows and slacks," in *Telecommunications Network Design and Management*, G. Anandalingam and S. Raghavan, Eds., 2003, pp. 65–81.

[20] D. Stamatelakis and W. D. Grover, "IP Layer Restoration and Network Planning Based on Virtual Protection Cycles," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 10, Oct. 2000.

[21] J. Kang and M. J. Reed, "Bandwidth protection in MPLS networks using *p*-cycle structure," in *International Workshop on Design of Reliable Communication Networks (DRCN)*, Oct. 2003, pp. 356–362.

[22] W. D. Grover, *Mesh-Based Survivable Networks: Options and Strategies for Optical, MPLS, SONET, and ATM Networking*. Prentice-Hall, Inc., 2003.

[23] K. Murakami and H. S. Kim, "Optimal Capacity and Flow Assignment for Self–Healing ATM Networks Based on Line and End-to-End Restoration," *IEEE/ACM Transactions on Networking*, vol. 6, no. 2, pp. 207–221, Apr. 1998.

[24] M. Menth, "Efficient Admission Control and Routing in Resilient Communication Networks," PhD thesis, University of Würzburg, Faculty of Computer Science, Am Hubland, July 2004.

[25] R. Bhandari, *Survivable Networks: Algorithms for Diverse Routing*. Norwell, MA, USA: Kluwer Academic Publishers, 1999.

[26] R. Martin, M. Menth, and U. Spoerlein, "Integer SPM: Intelligent Path Selection for Resilient Networks," in *IFIP-TC6 Networking Conference (Networking)*, Atlanta, GA, USA, May 2007.

[27] M. Menth, R. Martin, and U. Spoerlein, "Network Dimensioning for the Self-Protecting Multipath: A Performance Study," in *IEEE International Conference on Communications (ICC)*, Istanbul, Turkey, June 2006.

[28] ——, "Failure-Specific Self-Protecting Multipaths – Increased Capacity Savings or Overengineering?" in *International Workshop on Design of Reliable Communication Networks (DRCN)*, La Rochelle, France, Oct. 2007.

[29] G. Iannaccone, C.-N. Chuah, S. Bhattacharyya, and C. Diot, "Feasibility of IP Restoration in a Tier-1 Backbone," *IEEE Network Magazine (Special Issue on Protection, Restoration and Disaster Recovery)*, vol. 18, no. 2, pp. 13–19, Mar. 2004.

[30] J. Moy, "RFC2328: OSPF Version 2," April 1998.

[31] ISO, "ISO 10589: Intermediate System to Intermediate System Routing Exchange Protocol for Use in Conjunction with the Protocol for Providing the Connectionless-Mode Network Service," 1992.

[32] D. Thaler and C. Hopps, "RFC2991: Multipath Issues in Unicast and Multicast Next-Hop Selection," http://www.ietf.org/rfc/rfc2991.txt, Nov. 2000.

[33] A. F. Hansen, T. Cicic, and S. Gjessing, "Alternative Schemes for Proactive IP Recovery," in *2nd Conference on Next Generation Internet Design and Engineering (NGI)*, Valencia, Spain, Apr. 2006.

[34] M. Shant and S. Bryant, "IP Fast Reroute Framework," http://www.ietf.org/internet-drafts/draft-ietf-rtgwg-ipfrr-ip-mib-00.txt, Oct. 2006.

[35] A. Bley, "Routing and capacity optimization for IP networks," PhD thesis, Technical University of Berlin, February 2007.

[36] A. Sridharan and R. Guerin, "Making IGP Routing Robust to Link Failures," in *IFIP-TC6 Networking Conference (Networking)*, Ontario, Canada, May 2005.

[37] M. Menth and M. Hartmann, "Robust IP Link Costs for Multilayer Resilience," in *IFIP-TC6 Networking Conference (Networking)*, Atlanta, GA, USA, May 2007.

[38] M. Menth, R. Martin, M. Hartmann, and U. Spoerlein, "Efficiency of Routing and Resilience Mechanisms," http://www3.informatik.uni-wuerzburg.de/TR/425.pdf, University of Würzburg, Würzburg, Germany, Tech. Rep. 425, 2007.

[39] A. Khanna and J. Zinky, "The Revised ARPANET Routing Metric," in *ACM SIGCOMM*, 1989.

[40] I. Gojmerac, T. Ziegler, and P. Reichl, "Adaptive Multipath Routing Based on Localized Distribution of Link Load Information," in *International Workshop on Quality of future Internet Services (QofIS)*, Stockholm, Sweden, Oct. 2003.

[41] M. Laor and L. Gendel, "The Effect of Packet Reordering in a Backbone Link on Application Throughput," *IEEE Network Magazine*, vol. 16, no. 5, pp. 28–36, Sept. 2002.

[42] Z. Cao, Z. Wang, and E. Zegura, "Performance of Hashing-Based Schemes for Internet Load Balancing," in *IEEE Infocom*, Tel Aviv, Israel, 2000.

[43] R. Martin, M. Menth, and M. Hemmkeppler, "Accuracy and Dynamics of Hash-Based Load Balancing Algorithms for Multipath Internet Routing," in *IEEE International Conference on Broadband Communication, Networks, and Systems (BROADNETS)*, San Jose, CA, USA, Oct. 2006.

[44] ——, "Accuracy and Dynamics of Multi-Stage Load Balancing for Multipath Internet Routing," in *IEEE International Conference on Communications (ICC)*, Glasgow, Scotland, June 2007.