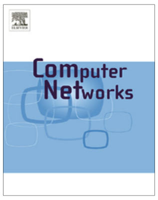
Computer Networks 74 (2014) 1–12

Contents lists available at ScienceDirect

Computer Networks

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



Network protection with multiple availability guarantees [☆,☆☆]

Greg Kuperman [a,*], Eytan Modiano [a], Aradhana Narula-Tam [b]

[a] Massachusetts Institute of Technology, LIDS, Cambridge, MA 02139, USA
[b] MIT Lincoln Laboratory, Lexington, MA 02420, USA

ARTICLE INFO

Article history:
Received 23 August 2013
Received in revised form 6 August 2014
Accepted 19 August 2014
Available online 27 August 2014

Keywords:
Network reliability
Network protection
Network recovery
Network design

ABSTRACT

This paper develops a novel network protection scheme that provides guarantees on both the fraction of time a flow has full connectivity, as well as a quantifiable minimum grade of service during downtimes. In particular, a flow can be below the full demand for at most a maximum fraction of time; if after a network failure the flow is below its full demand, it must still support at least a fraction q of that demand. This is in contrast to current protection schemes that offer either availability-guarantees with no bandwidth guarantees during the downtime, or full protection schemes that offer 100% availability after a single link failure.

We show that the multiple availability guaranteed problem is NP-Hard, and develop an optimal solution in the form of an MILP. If a connection is allowed to drop to 50% of its bandwidth for just 1 out of every 20 failures, then a 24% reduction in spare capacity can be achieved over traditional full protection schemes. Allowing for more frequent drops to partial flow, additional savings can be achieved. Algorithms are developed to provision resources for connections that provide multiple availability guarantees for both the sharing and non-sharing case. For the case of $q = 0$, corresponding to the standard availability constraint, an optimal pseudo-polynomial time algorithm is presented.

1. Introduction

As data rates continue to rise, a network failure can cause catastrophic service disruptions. To protect against such failures, networks typically use full protection schemes, often doubling the cost of resources needed to route a connection. An alternative approach is to provide a guarantee on the maximum time a connection can be disrupted. This is known as an "availability guarantee", and it is a bound on the fraction of time or probability that a connection can be disrupted. However, these disruptions (downtimes) may be unacceptably long; thus, many service providers opt for the more resource intensive full protection. In this paper, we propose a novel protection scheme with multiple availability guarantees. In addition to the traditional availability guaranteed protection, which maintains the full demand for at least a guaranteed fraction of time, we guarantee partial connectivity at all times. Thus, our approach is a hybrid between the traditional availability guarantees and full protection schemes.

Full protection schemes have been studied extensively [1–7]. The most common full protections schemes are $1+1$ or $1:1$ guaranteed path protection [8]. In $1+1$ path protection, two copies of the data are sent over a primary

[☆] Preliminary version of this work appeared in the 2012 IEEE International Conference on Communications (ICC).

[☆☆] This work was supported by NSF Grants CNS-1116209 and CNS-0830961, by DTRA Grants HDTRA1-07-1-0004 and HDTRA-09-1-005, and by the Department of the Air Force under Air Force contract #FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

[*] Corresponding author.

E-mail addresses: gregk@mit.edu (G. Kuperman), modiano@mit.edu (E. Modiano), arad@ll.mit.edu (A. Narula-Tam).

http://dx.doi.org/10.1016/j.comnet.2014.08.013
1389-1286/© 2014 Elsevier B.V. All rights reserved.

path and a failure disjoint protection path. Since two copies of the data are sent, the connection is guaranteed to survive any individual path failure. The downside of this strategy is that the protection resources are always utilized, and cannot be used to protect another connection while the original primary path is functioning. Alternatively, $1:1$ protection reserves resources on a disjoint backup path for protection, but does not utilize that path until a failure has occurred. With proper sharing strategies, protection resources can be used by multiple primary demands as long as they are not needed for more than one connection at any given moment in time. The disadvantage of $1:1$ protection is the additional complexity required for implementation. In this paper, we will refer to both the $1+1$ and $1:1$ protection schemes as disjoint path full protection, and specify the particular form as needed.

In addition to full protection schemes, there has also been a growing body of literature for backup provisioning to meet availability guarantees [9–15]. In all of these, primary and backup flows are allocated such that the connection is disrupted for at most a specified fraction of time or probability. During these down-states, the service is completely disrupted. A version of availability guarantees is considered in [16], where an end-to-end flow having a certain expected capacity, based on link availabilities, is found; multi-path routing is used to distribute risk, but no guarantees on flow are provided. In our paper, a flow is guaranteed to be at least a fraction q of the full demand at all times, which is known as "partial protection". Our novel approach is the first to combine the traditional availability guarantee and partial protection guarantee to allow the user to specify flows with different availability guarantees. Moreover, it is particularly applicable to IP-over-WDM networks where MPLS tunnels are used to provision resources.

The partial protection framework was first introduced in [17]. More recently, [18,19] developed a "theory" of partial protection such that after *any* single link failure, the flow can drop to the partial protection requirement. In [18,19], a fraction q of the demand is guaranteed to remain available between the source and destination after any single link failure, where q is between 0 and 1. When q is equal to 1, the service will have no disruptions after any single failure, and when q is 0, there will be no flow between the two nodes during the down state. In our work, flows can drop below the full demand for at most a specified fraction of time, and maintain at least q of that demand at all times.

The novel contributions of this paper include a framework for Multiple Availability Guaranteed Protection (MAGP) and providing associated algorithms to provision resources to meet these guarantees for both the cases when protection resources can and cannot be shared. Moreover, in the $q=0$ case, corresponding to the previously studied scenario where full availability is guaranteed for a fraction of time, we develop an optimal pseudo-polynomial algorithm. A preliminary version of this work was published in [20].

This paper is outlined as follows. In Section 2, the model for MAGP is described. In Section 3, MAGP is shown to be NP-Hard, and the minimum-cost solution to MAGP is formulated as an MILP. In Section 4, optimal solutions and algorithms for MAGP are developed when protection resources cannot be shared, and in Section 5, an algorithm is developed for when protection resources can be shared.

2. Multiple availability guaranteed protection

In this paper, routing strategies are developed and analyzed to minimize the total cost and capacity allocation required to satisfy each demand's protection and availability requirements. A demand needs to be routed from its source s to destination t such that the flow must be fully available for some given percentage of time. In other words, a flow can drop below the full demand for at most some specified downtime for any given time period, and must maintain at least a fraction q of that full demand at all times. Primary and protection resources are provisioned at the time of routing for a connection, which guarantees that sufficient capacity exists after a failure for that flow to meet its availability requirements. Similar to [11–15], the probability of simultaneous failures is assumed to be negligible, and we only consider single-link failures. To simplify the analysis, a "snapshot" model is used: The network state is considered after a failure has occurred. Let p_{ij} be the conditional probability that edge $\{i,j\}$ failed given that a network failure has occurred. For ease of exposition, instead of availability or maximum downtime, the Maximum Failure Probability (MFP) is considered, and its value is denoted by P. After some network failure occurs, a flow can be below the full demand, but at least a fraction q of the demand, with at most probability P. When a flow is below its full demand (but always at least q), that connection will be considered in a "downstate". The maximum failure probability is the conditional probability that a connection is in a downstate given some link disruption has occurred in the network.

This maximum failure probability can be related to the metric of availability by accounting for the expected time between failures and mean time to repair. Assuming that both the time between failures and the length of repair of any failure as exponential random variables with parameters $\frac{1}{\lambda}$ and $\frac{1}{\mu}$, respectively. The expected proportion of time there will be some failure is $\frac{\mu}{\lambda+\mu}$. With MAGP, after some failure in the network, a connection can fall below its full demand with probability less than P. In other words, $(1-P)$ percent of failures must have no effect (i.e., zero repair time). With a maximum failure probability P, the expected value for repair time becomes μP, and the proportion of time a connection is down is $\frac{\mu P}{\lambda+\mu P}$. We note that with MAGP, when a connection is "down", it still maintains a fraction q of the original connection's demand.

We assume that the graph G, with a set of vertices V, edges E, and edge failure probabilities $\mathcal{P}$, is at least two-connected. Since only single-link failures are considered, edge failures are disjoint events; hence, the sum of all the link failure probabilities is equal to one (i.e., $\sum_{\{i,j\}\in E} p_{ij} = 1$). Similar to previous works, the primary flow is restricted to a single path. After the failure of a link, a network management algorithm reroutes the traffic along the allocated protection paths.