2013

Lp-Based Flow-Rate Control And Modeling Of Capacity Collapse Propagation

Wogayehu Y. Gebremariam
North Carolina Agricultural and Technical State University

LP-Based Flow-Rate Control and Modeling of Capacity Collapse Propagation

Wogayehu Y. Gebremariam

North Carolina A&T State University

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Department: Electrical and Computer Engineering

Major: Electrical Engineering

Major Professor: Dr. Marwan U. Bikdash

Greensboro, North Carolina

2013

School of Graduate Studies
North Carolina Agricultural and Technical State University
This is to certify that the Master's Thesis of

Wogayehu Y. Gebremariam

has met the thesis requirements of
North Carolina Agricultural and Technical State University

Greensboro, North Carolina
2013

Approved by:

Dr. Marwan U. Bikdash Major Professor	Dr. Abdollah Homaifar Committee Member
Dr. John C. Kelly Department Chair	Dr. Robert Y. Li Committee Member

Dr. Sanjiv Sarin
Dean, The Graduate School

Biographical Sketch

Wogayehu Gebremariam was born on August 28, 1982, in Addis Ababa, Ethiopia. He received his Bachelor of Science degree in Electrical Engineering from Addis Ababa University in 2006 and a Master of Science degree in Electrical Engineering from North Carolina A&T State University in 2013.

Acknowledgements

This thesis would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

First and foremost, my utmost gratitude goes to my advisor Dr. Marwan U. Bikdash, Professor and Director of Computational Science and Engineering of North Carolina A&T State University, for his unreserved support, sincerity and encouragement in the completion this research work. I am also very thankful to my committee members Dr. Abdollah Homaifar and Dr. Robert Y. Li for their time and invaluable feedback.

I would also like to extend my acknowledgement to Pennsylvania State University and The Defense Threat Reduction Agency for partially sponsoring this work under contract DTRA01-03-D-0010/0020 and Sub-contract S03-34.

I would like to thank my wife Yamerot Gorfu for her care, love, support and patience at all times. My sisters, Wogderess Hailu and his wife Mekdim Eshete have given me their undeniable support throughout for which my mere expression of thanks likewise does not suffice.

Last but not least, I am most grateful to Dr. Abrham Workineh for his kindness and consummate friendship in facilitating my admission to North Carolina A&T State University and for his continuous support throughout my graduate study.

Table of Contents

List of Figures

List of Tables

Abstract

The main focus of this thesis is to understand how congestion that is due to link failure propagates to successive upstream links, and how well the network maintains system flow under abnormal conditions. Alleviating network failures depends on how congestion propagates through the network. In general, units of traffic can move from their origin to their destination quite rapidly, but the change in flow rates tends to propagate slowly. We develop novel capacity collapse propagation models that extends significantly the concept of cell-transmission used to partition links into sections. The sampling is done in such a way that density wave propagates through a section of the link in one time interval.

A general framework to model interaction between merging and diverging flow patterns is developed. The models considered for the nodes take into consideration the different types of intersections that may exist in the network. The capacity collapse propagation models can better represent networks with substantial propagation delay. The speed of the capacity collapse waves will be shown to depend on the magnitude of the failure. We integrate our models within the multicommodity flow framework, in which each commodity (origin-destination pair) uses $k \in \mathbb{N}$ link-disjoint paths to satisfy flow-rate demands. The congestion in the links is used to update the prices of the links, thus affecting the cost of travelling. We solve several minimum-cost linear-programming problems to control path flow-rate routing decisions triggered by the changes in the cost coefficients. We conclude that proposed path flow-rate rerouting in response to the congestion in the links could contribute significantly to network survivability. Numerical simulations of the proposed models are used to illustrate the concepts.

CHAPTER 1

Introduction

1.1 Flow Networks

Networks serve to deliver flow through a system of interconnected nodes and arcs. The purpose of a flow network is the transportation of commodities from specific origins to specific destinations in response to flow-rate demands. Flow usually refers to the amount or rate of traffic associated with a route.

Many infrastructure networks can be modeled using concepts from graph theory. Consider a network represented by a capacitated graph $G(N, L)$, where N is a set of nodes, and L is a set of unidirectional links. Nodes usually model origins, destinations or transshipment of commodities. Links are direct, possibly directed, paths between nodes. A path is formed by a sequential combination of one or more directed links in a network with no repetition of nodes. Every link in a path is directed away from the origin towards the destination, and allows traffic flow in only one direction. A network G is said to be connected if there exists at least one path for any node to any other node in the graph. Basic definitions and elementary properties of graphs are treated in detail in [1] and [2].

A flow is characterized by its source node and destination node. In general, a commodity represents a traffic demand between a pair of one source node $i \in N$ and one destination node $j \in N/\{i\}$. In a practical network, not all nodes participate significantly in meeting a demand pair; there are often transit nodes used solely for the purpose of routing. Let F represent the set of all Origin-Destination (OD) pairs. An OD pair $f \in F$ has a flow-rate demand $d_f(t)$ from source i to destination j. Flow rates are measured in number of traffic units per time interval. Most network flow problems have to be modeled as multicommodity networks where the flow rates associated

with an OD pair compete for the capacities of a typical link [3] and [4].

Disruption of network facilities (links/nodes) can considerably hinder the flow of services through the network. Operational characteristics such as the level of system connectivity, maximum flow capacity, and the cost of network transportation can be affected by facility damage. The performance of the network in the event of link failures depends not only on the physical characteristics of the network but also on the ability of the network to react to failures.

1.2 Flow Survivability

The location and role of network facilities, and the topological relation among them are vital in the operability of network services. The impact of a link failure in network performance is reflected in the concept of network vulnerability and reliability analysis. Several optimization models, also known as interdiction models, have been developed to identify important facilities with regard to impact on system performance [5] and [6]. A general framework for reliability analysis in [3] highlights the importance of routing and rerouting in the reliability of flow networks. Survivability, an emerging principle, extends vulnerability and reliability studies and focuses on maintaining system flow even when the system has encountered undesired events [7] and [8].

The concept of survivability as it applies to different types of network has gained in importance in recent years. A general definition presented in [9] summarizes survivability as the system's ability to continuously deliver services in compliance with the given requirements in the presence of failures and other undesired events. This capability should not depend on the survival of a damaged facility. It is the compromised services, not any particular network component, that must survive.

1.2.1 Failures. A variety of threats, like attacks, accidents, and failures, may cause minor or major service degradations. These undesired events can be broadly categorized as failures and accidents. Accidents describe externally generated events such as natural disasters or targeted attacks. On the other hand, failures represent internally occurring potentially damaging events that are usually caused by deficiencies in the system due to traffic congestion, link/node failure and repair.

Failures and accidents are included as part of survivability. With respect to system survivability, the impact of the event is more important than the type of the event. The definitions of survivability concentrate on the effect of a damaging event without any reference to the events that caused it. In fact, for a network to survive, it must successfully recover from the failure whether the cause is determined or not. A failure in the network can be represented as a specific reduction of link capacity.

1.2.2 Essential services. Essential or critical services are defined as the functions of the system that must be preserved when the network is exposed to undesired events, [7] and [9]. These services have strict requirements for reliability. If an essential service is lost, it must be replaced by another but equivalent service that satisfies the survivability requirements in a different way.

The service in flow networks can be to satisfy the flow demand between specified origin and destination nodes. The availability of paths supporting OD pair flows is a requirement for survivability. In [7], essential services are defined to include alternate set of mutually exclusive essential services that need not be simultaneously available. So, to enhance the survivability of a flow network, the shortest paths are equipped with alternative paths supporting OD pair flow demands. In order to ensure the reliability requirements, a path that delivers flow through a failed link can be replaced with another path that is link-disjoint to the first path but serves the same OD

pair flow.

1.2.3 Reliability measures. A general framework for calculating a reliability measure for several types of flow networks is presented in [3]. The approach emphasizes the importance of routing on top of network connectivity and performability. Performability is defined in [10] as a reliability measure that is commonly used to evaluate how well a flow network reacts to a failure. Unlike connectivity measures, performability considers the flow nature of networks in evaluating network reliability. Connectivity measures are related to the probability of conservation of the graph structural properties in the event of failures, [5] and [9]. Flow rerouting, proposed in [3] as a reliability measure in transportation systems, accounts not only for the probabilities of terminal connectivity or the capacity of the network, but also the ability of the system to adjust its flow after a failure.

1.3 Source Rate Control

Survivability of a system also depends on the routing and the congestion control schemes in place [11]. The need for networks to operate in non-cooperative environments has stimulated work on optimization approaches to rate control algorithms. There are several articles on rate control algorithms. The important papers of Kelly [12] and [13] and Low [14] pinpointed optimization approaches to flow-rate control schemes that have proved to be stable. The basic algorithm requires communication of link prices to sources and source rates to links. The rate control schemes addressed the issue of fairness, as there might be unfair network throughput distribution in situations where a given scheme maximizes network throughput while denying access to some users.

A common approach to flow control is to decompose the problem into a static optimization problem and a dynamic stabilization problem. The former incorporates fairness, capacity

constraints, and utilization. Its solution provides the desired steady-state operating point. The source rate and link price update laws are then designed to guarantee stability and robustness of the equilibrium.

The articles in [12]-[17] motivate the modeling of flow control by an optimization problem and derive their control mechanisms as solutions to the optimization problem. The objective in this approach is basically to maximize the aggregate source utility, and sources with different values of bandwidth should react differently to network congestion. This is accomplished by means of pricing signals transmitted from links to sources. The sources then adjust their transmission rates accordingly. Two types of traffic are renowned in communication networks: elastic traffic and inelastic traffic.

1.3.1 Elastic traffic. Elastic traffic adjusts its throughput between end hosts in response to network condition. It has adaptive transmission rates generated by delay-tolerant traffic such as file transfer or E-mail applications. In the context of data networks, the source flow control models are designed to address flow demands of an elastic traffic. They have the advantage of controlling the packet injection rates depending to the availability of bandwidth.

1.3.2 Inelastic traffic. The other important class of flows is inelastic with fixed flow arrival rates. Inelastic flows usually model delay-sensitive and high-priority applications such as video and audio streaming. The approach in [18] uses an optimization model for heterogenous traffic that consists of both elastic and inelastic flows. The arrival rate of the inelastic flow is assumed to follow a stochastic process that is identically and independently distributed (i.i.d.) over time with a fixed mean rate.

1.4 Traffic Evolution

The network flow control approaches did not consider queuing and propagation delay in the network. They assumed that traffic units at the sources reach their destinations instantly. But in reality, it takes some time to accomplish OD pair flows, i.e. links down in the path sense the flow at the origin at a later time. The propagation delay is significant in most transportation networks.

In situations where there is latency, the concept of cell-transmission is introduced by Daganzo in [19] and [20]. The cell-transmission model predicts the evolution of traffic flows over time based on a simple macroscopic simulation of traffic flow. The cell-transmission model promotes a discrete-time strategy where current conditions are updated every time as the clock advances. The authors in [21] reduce the model into the single-destination dynamic traffic assignment problem.

1.5 Problem Statement

We study the flow control of transport networks in the presence of inelastic traffic requirements. This is the more difficult and general case and apply to other types of networks such as highway networks. Much of the existing work [12]-[17] on flow control approaches concentrates on elastic traffic. They are concerned with maximizing source transmission rates so as to fully utilize the resources of the network while complying with capacity constraints in the links. We will extend that and develop a flow control optimization approach for inelastic flow demand requirement. To that end, we intend to carefully redistribute the OD pair flows into their available routes.

For the sake of utilizing the network flow control approaches, we extend the cell transmission model by assuming that congestion waves travel much more slowly than the traffic to discretize the links of the network. Then we formulate capacity collapse propagation models at

link level. We will also propose possible congestion propagation and conflict resolution models at the merging and diverging nodes of a network.

Flow control algorithms differ in their choice of objective functions or their solution approaches, and result in rather different flow control mechanisms to be implemented at the sources and the network links. In our model, we treat inelastic flows that cannot be controlled using utility functions. This leads systematically to refine the objective of the optimization based flow control (2.2) to load balancing. For this purpose, our approach aims to minimize the cost of network transportation as a measure of network performance.

The effectiveness of flow routing basically depends on the availability of alternative paths. We consider the multicommodity flow problem, in which each commodity uses $k \in \mathbb{N}$ paths to address Origin-Destination (OD) flow-rate demands. The k alternative path are ideally required to be link-disjoint. To that end, we extend the k successively shortest link-disjoint paths generation criterion in [22] to include paths that are first-link disjoint. As a result we have more versatility in availability of substitute paths.

A linear program based controller is then used to assign the flow rates into alternative paths with the objective of minimizing the cost of network transportation. The formulation of the controller satisfies the flow demand requirements in addition to the capacity constraints. The price signals reflect the intensity of traffic congestion in the links and have a hold up in calculating the cost of travelling. The controller reassigns the traffic into relatively inexpensive paths in order to avoid further backlog buildup in the network.

Incorporating rerouting capabilities into the network can substantially reduce the risk of disruptions. Survivability can be further improved through restoration of compromised OD pair flow rates following a damaging event.

1.6 Synopsis

The thesis is organized as follows. In Chapter 2, we present background information that will help to understand the material. The network flow control algorithms and cell-transmission models will be discussed. Chapter 3 focuses on congestion propagation in links and conflict resolution at intersection nodes. The models assumed for diverging and merging nodes are compared and analyzed using numerical example. In Chapter 4 an LP-Based flow control approach will be proposed. The applicability of the network flow models in flow survivability will also be considered. In Chapter 5, numerical examples illustrating capacity collapse propagation and the importance of flow rerouting for single and multiple link failure scenarios will be discussed. Chapter 6 concludes the thesis and points to future research possibilities.

CHAPTER 2

Background Information

2.1 Network Flow Control Schemes

2.1.1 The basic model. Consider a network that consists of a set of L links of finite capacity c_l, $l \in L$. The network is shared by a set P of routes. A route $r \in P$ is a non-empty ordered subset of L and it is associated with an OD pair, also called a source, or a user. The interconnections between the links and the paths are defined through a routing matrix $R = (R_{lr},\ l \in L, r \in P)$. The link-path indicator variable R_{lr} is defined as

$$R_{lr} = \begin{cases} 1 & \text{if } l \in r, \text{ so that resource } l \text{ lies on route } r, \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

A rate x_r is describes to the source r, and its utility is denoted $U_r(x_r)$. The utilization function $U_r(x_r)$ is assumed to be increasing, strictly concave and continuously differentiable in its argument over the range $x_r \geq 0$. The objective of the optimization problem is to maximize the sum of the utilization functions $U_r(x_r)$ over all sources while complying with the capacity constraints of the links:

$$SYSTEM(U, R, c) \quad : \quad \max_{x \geq 0} \sum_{r \in P} U_r(x_r) \tag{2.2}$$

$$\text{s.t.} \quad : \quad y = Rx \leq c,\ x \geq 0 \tag{2.3}$$

where y is the aggregate source rate at the links.

$$y_l = \sum_{r \in P} R_{lr} x_r,\ l \in L. \tag{2.4}$$

The constraint (2.3) enforces that the aggregate source rate at any link should not exceed the capacity of the link. A unique optimizer exists since the objective function is assumed to be

strictly concave and continuous.

By using the Lagrangian multiplier, p, the inequality constraint can be brought into the optimization problem:

$$\begin{aligned} \min_{p\geq 0}\max_{x\geq 0} L(x,p) &= \min_{p\geq 0}\max_{x\geq 0}\sum_{r\in P} U_r(x_r) - \sum_{l\in L} p_l \sum_{r\in P}(R_{lr}x_r - c_l) \\ &= \min_{p\geq 0}\{\max_{x\geq 0}\sum_{r\in P}(U_r(x_r) - x_r\sum_{l\in L} R_{lr}p_l) + \sum_{l\in L} p_l c_l\} \\ &= \min_{p\geq 0}\{\sum_{r\in P}\max_{x\geq 0}(U_r(x_r) - x_r\sum_{l\in L} R_{lr}p_l) + \sum_{l\in L} p_l c_l\} \end{aligned} \tag{2.5}$$

Since the utilities U are unlikely to be known by the network, the approach taken in [12] and [13] decomposes $SYSTEM(U,R,c)$ (2.2-2.3) into two simpler problems: a user subproblem and a network subproblem.

The first term in (2.5), $\sum_{r\in P}\max_{x\geq 0}(U_r(x_r) - x_r\sum_{l\in L} R_{lr}p_l)$, is decomposed into $|P|$ separable subproblems. If p_l represents the price per unit flow at link l, then q_r represents the total price per unit flow for all the links in path r,

$$q_r = \sum_{l\in L} R_{lr}p_l. \tag{2.6}$$

Hence, the user subproblem is to select transmission rates x_r in order to maximize the users total benefit at the given prices q_r. If user r is charged an aggregate price q_r per unit flow, and is allowed to freely vary the flow x_r, then the utility maximization problem for user r is

$$USER(U_r;q_r) : \max_{x_r\geq 0} U_r(x_r) - x_r q_r \tag{2.7}$$

The price vector p takes the role of a coordination signal that lines up the optimal value of $USER$ (2.7) to the optimal value of $SYSTEM$ (2.2).

If the network receives a revenue q_r per unit flow from user r, and is allowed to freely vary

the flows x, then the revenue optimization problem for the network is

$$NETWORK(R, c; q) \quad : \quad \max \sum q_r x_r \tag{2.8}$$
$$\text{s.t.} \quad : \quad Rx \leq c, x \geq 0$$

Theorem 2.1: There exists a price vector $q = (q_r, r \in P)$ such that the vector $x = (x_r, r \in P)$, formed from the unique solution x_r to $USER_r(U_r; q_r)$ for each $r \in P$, solves $NETWORK(R, c; q)$. The vector x then also solves $SYSTEM(U, R, c)$. The proof to the this theorem is given in [12].

The objective function (2.2) is separable in the source rates x_r which are coupled by the constraints (2.3). As a result, solving the optimization problem (2.2-2.3) requires coordination among the users.

The network's optimization problem modeled in primal and dual forms proposed by Kelly [12] and Low [14] lead to two classes of rate control algorithms: the primal algorithm and the dual algorithm. These algorithms provide source and link update laws that are decentralized. The sources do not have information about the utilization functions of other sources, and the links do not have knowledge of the capacities of other links. The flow rates corresponding to a path can only depend on the price of the path, and the price corresponding to a link can only depend on the total flow in the link. The routing information contained in the routing matrix R are unknown to the sources and the links.

2.1.2 The primal algorithm of Kelly [12]**.** In [12], Kelly developed a model in which a user chooses the charge per unit time and the network determines the user's rate. It is shown that a system optimum is achieved when users' choices of charges and the network's choice of allocated rates are in equilibrium. Later in [13], he proposed the primal algorithm using explicit rates based

on link prices, that are shown to provide stability and fairness.

The primal algorithm consists of a first-order source update law and a static link penalty function to keep the aggregate rate below its maximum capacity. Given the utility function for each source, the source update law is given by

$$\frac{d}{dt}x_r = \kappa(U_r'(x_r) - q_r) \tag{2.9}$$

where κ is a constant.

Equation (2.9) corresponds to a response flow-rate by user r to an increase in price by adjusting the flow-rate on route r, x_r. The network attempts to equalize the aggregate price per flow of route r, q_r, to the derivative of the utility of the user, for every $r \in P$.

When link l generates a price signal, it is interpreted as a congestion indicator requiring each user whose route passes through the link to reduce some flow. Suppose that link l generates a continuous stream of feedback signals at rate $f(y_l)$ when the total flow through resource l is y_l, the link update law is given as a penalty function that enforces the link capacity constraint $y_l \leq c_l$,

$$p_l = f(y_l). \tag{2.10}$$

2.1.3 The dual algorithm of Low [14]**.** Equations (2.9-2.10) present a system where rates vary gradually, and prices are given as functions of the aggregate rates. The link rate constraint enforced by using the penalty function fails to take the link queue dynamics into consideration. A dual approach is also proposed in [13], where the links use a first-order dynamics of the price update. Moreover, the source rate update is given as a function of the prices. The continuous-time link update law where link prices vary gradually is given as

$$\frac{d}{dt}p_l(t) = \kappa(y_l(t) - c_l) \tag{2.11}$$

where κ is a constant.

A related approach has been developed in [14] to solve the same optimization problem (2.2-2.3) based on discrete-time models. A gradient projection method is used to solve the dual problem where link prices are adjusted in opposite direction to the gradient. In the special case where $U_r = w_r \log x_r$ the two approaches were shown to provide equivalent results. The discrete-time link update is

$$p_l(t+1) = [p_l(t) + \gamma(y_l(t) - c_l)]^+ \tag{2.12}$$

where $\gamma > 0$ and $[z]^+ = \max\{0, z\}$. The price adjustment rule in (2.12) is consistent with the law of supply and demand: if the total flow at link l exceeds the supply capacity, the price increases; otherwise the price decreases.

The static source rates are given by the primal solution (2.9) as a function of the path price

$$x_r = U_r'^{-1}(q_r) \tag{2.13}$$

Each source solves (2.13) and communicates its rate x_r to links on its path. Given the total source rate y_l through link l, the links then update their prices p_l in accordance with (2.12), and then communicate the new prices to the sources contributing to the aggregate flows in the links.

The first-order link price update law (2.12) indicates that price p_l integrates excess demand, which is exactly what a backlog variable b_l does

$$b_l(t+1) = [b_l(t) + y_l(t) - c_l]^+. \tag{2.14}$$

In other words, prices become proportional to backlogs, and thus an increase in price can only be achieved by increasing the backlog. This deficiency has motivated the work in [15] to couple

the congestion measure p_l with the performance measure b_l. Thus, the price update law (2.12) is replaced by

$$p_l(t+1) = [p_l(t) + \gamma_l(\alpha_l b_l(t) + y_l(t) - c_l)]^+ \tag{2.15}$$

where α_l and γ_l are constants. This second-order price dynamics with an additional term involving the backlog attempts to achieve high utilization while clearing the backlog. The extra integrator, $b_l(t)$, guarantees that any equilibrium will have empty buffers as opposed to large buffers in (2.12). The stability proof for this higher order system in continuous-time is given in [16].

Recent works in [17] extends the primal and dual control schemes to a broader classes of flow control laws using the concept of passivity. The idea of a combined primal/dual flow control with dynamic-source and dynamic-link update laws is also discussed in [17].

2.2 The Cell-Transmission Model

The behavior of multicommodity traffic flows over networks can be predicted over time, based on a simple macroscopic simulation of traffic flow. The cell transmission model introduced in [19], [20] is one such approach for modeling highway traffic flow using the hydrodynamic analogy. The model assumes that every link is divided into small homogeneous sections called cells.

The cell transmission model reduces the hydrodynamic model to simple difference equations by assuming a piecewise linear relationship between flow and density at the cell level. The relationship between traffic flow (y) and density (ρ) is of the form

$$y = \min\{v\rho, y_{max}, v_w(\rho_c - \rho)\}, \text{ for } 0 \leq \rho \leq \rho_c \tag{2.16}$$

where $v, y_{max}, w,$ and ρ_c are constants. v is the free flow speed measured in distance covered per unit time, y_{max} is the maximum flow-rate (or flow-rate capacity), v_w is the backward propagation

speed, and ρ_c is the congestion density. The equation of state of the cell-transmission model can be represented using flow-density graph as shown in Figure 2.1.

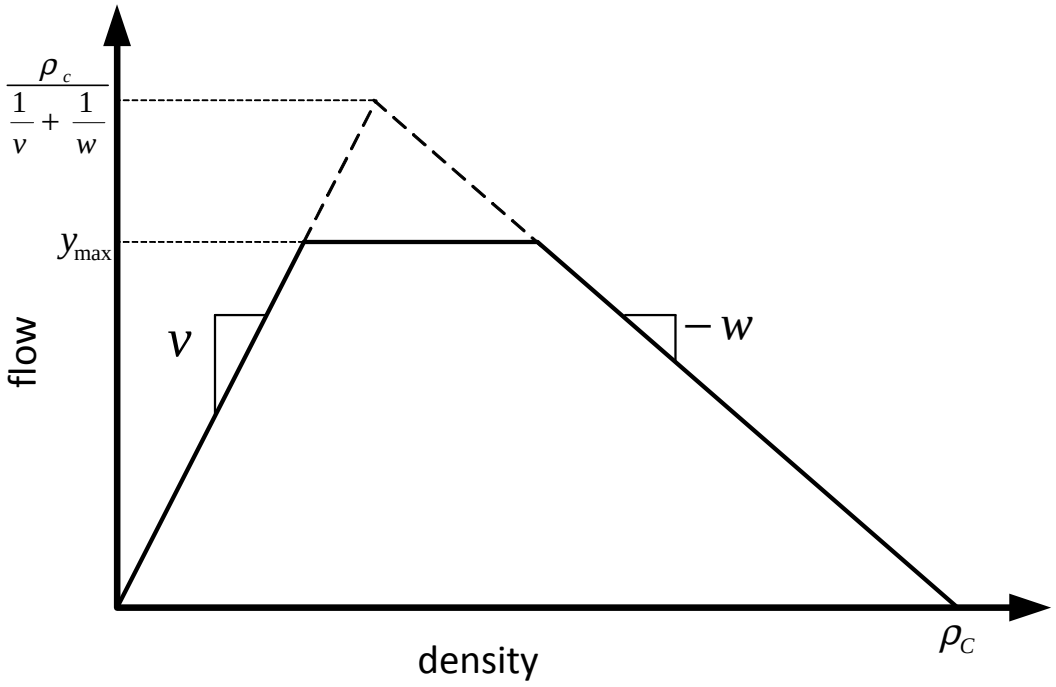

Figure 2.1. The equation of state of the cell-transmission model.

The difference equations can further be reduced to simple linear relationships of flow and occupancy at the cell level. The occupancy level is the product of the cell's length and its density. The length of each cell is chosen in [19] as the distance traveled by free-flowing traffic in one time interval. Free-flowing traffic in a cell advances to the next cell with each clock tick. The time interval is assumed to be 1. Suppose cells are numbered starting with the upstream end of the road. For two consecutive sections s and $s+1$, the system's evolution obeys

$$n_{s+1}(t+1) = n_s(t) \tag{2.17}$$

where $n_s(t)$ is vehicle occupancy in cell s at time instant t. The recursion (2.17) holds unless traffic is slowed down by congestion from a downstream bottleneck where flow exceeds capacity.

The state of the system at time instant t is given by the number of vehicles contained in

each cell, $n_s(t)$. To capture the effect of congestion in each cell, the following parameters are defined: $B_s(t)$, the maximum number of vehicles that can be present in cell s at time t, and $C_s(t)$, the maximum number of vehicles that can flow into cell s when the clock advances from t to $t+1$. These parameters can vary with time to capture time-dependent capacity and flow as per the occurrence of transient traffic incidents. $B_s(t)$ is defined to be the product of the cell's length and its congestion density.

The cell-transmission model is expressed by the following recursive relationship with the state of the system being updated with every tick of a clock,

$$n_s(t+1) = n_s(t) + y_s(t) - y_{s+1}(t), \tag{2.18}$$

where $y_s(t)$ is the inflow to cell s in the time interval $(t, t+1)$. The flows in relation to the current conditions at time t is given by:

$$y_s(t) = \min\{n_{s-1}(t), C_s(t), \sigma[B_s(t) - n_s(t)]\} \tag{2.19}$$

where $\sigma = w/v \leq 1$.

Since the number of vehicles that enter a cell, see (2.19), is only influenced by the current conditions in the cell, the inflow to a cell is unrelated to the number of vehicles that will leave it. The occupancy restriction $y_s(t) \leq \sigma[B_s(t) - n_s(t)]$ in (2.19) is due to the fact that empty slots for vehicles can only travel backwards at a finite speed (the density wave propagation speed) unlikely to be greater than the free flow speed. Therefore, the effects of the outflow should only be noticed upstream after some time. For the cell-transmission model, this lag is one tick of the clock and it is equivalent to assuming that density waves propagate backwards at the free flow speed.

CHAPTER 3

Capacity Collapse Propagation

3.1 Link Discretization

In this section, we extend the cell-transmission model discussed in [19] and [20]. We will refer to cells as sections throughout the remainder of thesis. The equation of state in (2.16) guides the choices of the free flow speed v, the maximum flow-rate $y_{\max}$, and the congestion density ρ. The assumption in the cell-transmission model forces the density wave speed, v_w, to match the free flow speed, v. But, in reality, the waves propagate more slowly than free flowing traffic. This changes the manner in which capacity collapse and density waves propagate in the network.

The backward propagating waves indicate the availability of downstream capacity and occupancy. In [23], we considered the propagation of congestion over long links not including intersection nodes. The model assumes that units of traffic can move from their origin to their destination quite rapidly, but the change in flow rates tends to propagate slowly through the links of the network. Here we will elaborate this model in much more detail. The individual units of traffic move at a speed measured in distance travelled per unit time whereas the flow rates are measured in number of traffic units passing during one unit time interval.

Suppose link L_n in Figure 3.1 is discretized into sections S_1, S_2, $\cdots$ S_7. We designate T as the set of discrete-time instants, i.e. $T = \{0, \tau, 2\tau, 3\tau, \cdots\}$ where τ is the sampling time interval. The discretization is done such that density waves propagate one section of a link within one time interval. The length of each section equals the distance traveled by the wave in one time interval. The free-flowing traffic is assumed to travel so fast that it traverses all the sections of a link in one time step. Therefore, the effects of outflow should only be noticed upstream after some delay.

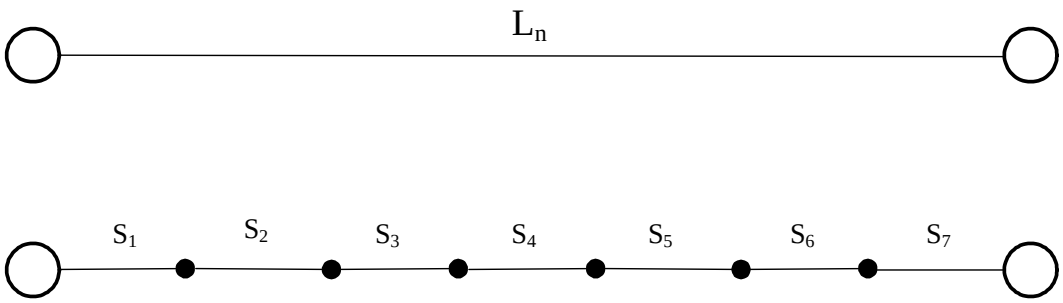

Figure 3.1. Link L_n discretized into sections S_1, S_2, $\cdots$ S_7.

3.2 Congestion Propagation in Links

Let S be the set of sections in a link. To capture the effect of congestion in each section, the following variables are defined. Let $C_s[t]$ denote the maximum capacity of section s during the interval $[t, t+\tau)$ and $B_s[t]$ is the maximum occupancy of section s at time t. These parameters vary with time to model traffic incidents such as link failure.

Let $n_s[t]$ denote occupancy of section s at the beginning of the time interval $[t, t+\tau)$. The available-occupancy in section s, $a_s[t]$, is the amount of empty space for incoming traffic in the section. It is given as

$$a_s[t] = B_s[t] - n_s[t]. \tag{3.1}$$

Let $c_s[t]$ denote the rate capacity of section s; i.e., the maximum flow-rate that can depart the section during the interval $[t, t+\tau)$. A change in rate capacity and available-occupancy can occur when a subsequent section is congested.

For two successive sections s and $s+1$, the capacity update for section s is defined as the smaller of the maximum capacity of the section and the available-occupancy in the next section; i.e.,

$$c_s[t] = \min\{C_s[t], \frac{a_{s+1}[t]}{\tau}\}, s < |S|. \tag{3.2}$$

The rate capacity update at the last section of the link, i.e. $s = |S|$, depends on the type of nodes

and the assumed conflict resolution models at the intersections. This matter is treated in detail in Section 3.4. The instantaneous rate capacity $c_l^s[t]$ and available-occupancy $a_l^s[t]$ capture the time-dependent characteristics, and they depend on the congestion level [19]-[21].

The inclusion of the available-occupancy of the consequent section, $a_{s+1}[t]$ in (3.2) models the capacity collapse propagation to preceding sections at times of congestion. The capacity in a section collapses when the downstream section is clogged. The excess traffic blocks the upstream traffic, and thus the capacity collapse propagates to the beginning of the link. The speed of the capacity collapse wave depends on the rate at which the sections are being filled. The closer the flow-rate to the available-occupancy of a section during the time interval, the faster the section will be occupied, and the faster the collapse will propagate. For very small flow rates it takes a longer time to fill up the section and thus the collapse wave propagates slower.

Example 4.1: Suppose the available-occupancy is 30 units of traffic, and if the flow-rate is 30 units of traffic per time step, it takes only one time step to fully occupy the section. In contrast, if the flow-rate is 5 traffic units per time step, it takes 6 time intervals for the available-occupancy to be used up. In other words, the collapse wave is 6 times slower in the later case. Note that the particle speed is irrelevant in this calculation, unless it is too small.

3.3 Flow-Rate and Occupancy

The flow-rate control algorithms in [12]-[17] essentially assume that traffic injected into the source nodes arrive at their destinations instantaneously. In reality, traffic will reach downstream nodes only after a queuing and propagation delay incurred in the intermediate nodes. The congestion information travels slower than the speed of the actual traffic. The delay in congestion propagation is significant in transportation networks and it is worth considering to understand the buildup of backlogs at bottleneck links. For this reason, we keep track of the flow

rates in the individual sections of a link.

The flow-rate in excess of the capacity of a section will be backlogged and occupy the section at least for one time interval. Let $b_s[t]$ be the traffic backlog in section s and it is defined as

$$b_s[t] = [n_s[t] + y_s[t]\tau - c_s[t]\tau]^+ \tag{3.3}$$

where $[z]^+ \triangleq \max\{0, z\}$ and $y_s[t]$ is the flow-rate entering to cell s within the interval $[t, t+\tau)$.

The occupancy of the section is updated at each clock tick as

$$n_s[t+1] = b_s[t]. \tag{3.4}$$

Equation (3.3) models the backlog build-up process in which traffic in excess of the available-occupancy of a section will spillover to the upstream section.

The links in a network are shared by a set of OD pair flows. The traffic units arriving at the link join the first section and travel through the subsequent sections. Here, we have assumed that units of traffic travel very fast and thus the inflow rate in the sections of the link is given by

$$y_{s+1}[t] = \min\{\frac{n_s[t]}{\tau} + y_s[t], c_s[t]\}. \tag{3.5}$$

As it is indicated in (3.5), the inflow to a section is unrelated to the number of traffic that will leave it. The outflow from a section cannot exceed its capacity which is determined by the availability of empty space at the adjacent downstream section as defined in (3.2). The inclusion of $y_s[t]$ in the total outgoing flow-rate, $\frac{b_s[t-1]}{\tau} + y_s[t]$, models the assumption that the incoming traffic units can leave the section within the same time interval.

Equation (3.3) models the backlog build-up process at the section level. The backlog in

the link is therefore

$$b[t] = \sum_{s \in S} b_s[t]. \tag{3.6}$$

The links then update their price according to the second-order dual algorithm in [15], and the discrete time price update law is

$$p[t+1] = [p[t] + \gamma(\alpha b[t] + y[t] - c[t])]^+ \tag{3.7}$$

where $\gamma > 0$ and $\alpha > 0$ are price sensitivity constants, and $c[t]$ is the rate capacity of the link.

Price is interpreted as a congestion indicator requiring some reaction in the flow controllers. The links feed back the price signals to the flow sources that utilize the information to compute the aggregate prices, $q_r[t]$, $r \in P$,

$$q_r[t] = \sum_{l \in L} R_{lr} p_l[t], \tag{3.8}$$

in order to facilitate the path choice decisions.

3.4 Congestion Propagation at Intersections

Section 3.2 discussed how backlog propagates along links. We are further extending the cell-transmission model to model congestion propagation at intersection nodes. All links in a network have starting and terminating nodes. A node serves as a junction where incoming and outgoing links meet. The behavior of congestion propagation at junctions depends on the types of interactions occurring at the nodes. The nodes in a network can be mainly categorized as merging, diverging, or a combination thereof. A transit node can be considered as a special case with one incoming link and one outgoing link.

3.4.1 Merging nodes. Merging nodes are identified by one outgoing link and one or more incoming links. In Figure 3.2, links L_1, L_2 and L_3 are competing for the resource in L_{out}. Flow disruption occurring in L_{out} affects all the links incident on the node. The scarcity of downstream available-occupancy raises the issue of handling the contention of flow-rate demands amongst the incoming links.

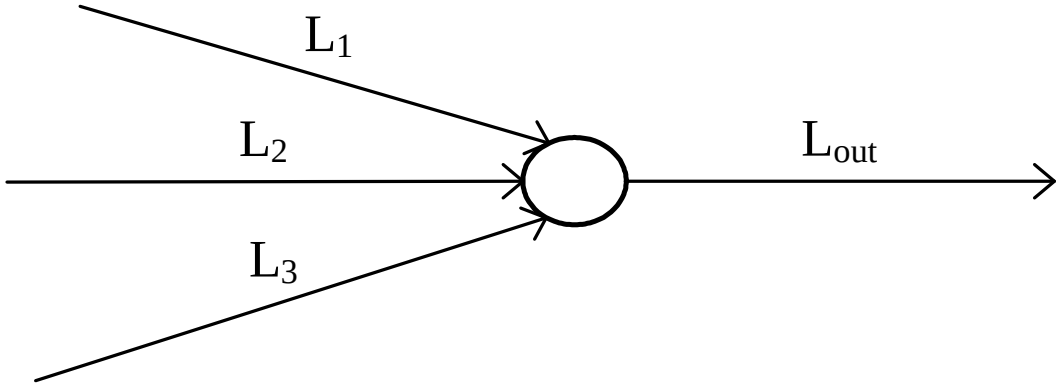

Figure 3.2. A merging node.

Suppose there are m incoming links to a given merging nodes, which are denoted as $L_1, \cdots, L_m$. Let the outgoing link be denoted by L_{out}. The effective capacity of the incoming links combined together is constrained by the availability of the downstream link, L_{out},

$$c_{L_1} + c_{L_2} + \cdots + c_{L_m} \leq \frac{a_{L_{out}}}{\tau}. \tag{3.9}$$

A capacity loss in L_{out} will affect the flow in the incoming links and the failure propagates accordingly. The capacity collapse becomes apparent when the available-occupancy in L_{out} falls short of the total flow entering the node. This conflict can be resolved in many ways.

Following is a list of models that can be applied to determine the capacity distribution among the incoming links at every instant of time $t \in T$.

Model M1: Equal sharing: The available-occupancy at outgoing link is shared equally

among the incoming links at the merging node.

$$c_l = \frac{1}{m} \times \frac{a_{L_{out}}}{\tau},\ l \in \{L_1, \cdots, L_m\}. \tag{3.10}$$

Example 4.2: Suppose $m = 3$ as in Figure 3.2. The capacity of the incoming links is therefore

$$c_l = \frac{1}{3} \times \frac{a_{L_{out}}}{\tau}, l \in \{L_1, L_2, L_3\}.$$

Model M2: Random proportions: The available-occupancy in the outgoing link will be randomly divided among the incoming links. The merging node simply generates a set of random numbers $[\varphi_{L_1}, \cdots, \varphi_{L_m}]$, $\varphi_l \in [0, 1]$, $l \in \{L_1, \cdots, L_m\}$ and

$$\sum_{l \in \{L_1, \cdots, L_n\}} \varphi_l = 1 \tag{3.11}$$

that determine the proportion of the downstream capacity going to each link. In that case,

$$c_l = \varphi_l \times \frac{a_{L_{out}}}{\tau}. \tag{3.12}$$

Example 4.3: Suppose $m = 3$ as in Figure 3.2. Suppose the available-occupancy at L_{out} be 20 units of traffic per time interval. The merging node generates three random numbers $[0.2, 0.7, 0.1]$ respectively for $\{L_1, L_2, L_3\}$. Thus, the capacity apportionment will be

$$\begin{aligned} c_{L_1} &= \varphi_{L_1} \times \frac{a_{L_{out}}}{\tau} = 4 \\ c_{L_2} &= \varphi_{L_2} \times \frac{a_{L_{out}}}{\tau} = 14 \\ c_{L_3} &= \varphi_{L_3} \times \frac{a_{L_{out}}}{\tau} = 2 \end{aligned}$$

Model M3: Based on priority: The merging node can set priorities to the incoming links in several ways. The simplest way could be to set the priority ranking in advance or to do

random ranking in every time interval. The priorities can also be defined using parameters and/or measurements such as the amount of backlog, the number of supported paths, or the waiting time in the incoming links. We concentrate on the case where the priorities are defined based on the amount of backlog in the incoming links. The merging node assigns the highest priority to the link with the largest backlog, and the link with the smallest backlog will have the least priority. The traffic in the lower-ranking links will get a chance to clear out only after the traffic in the higher priority links has flushed out.

Let the priorities be denoted by the permutation $\pi = [\pi_1, \pi_2, \cdots, \pi_m]$ of $1, \cdots, m$. Then the links in descending priority order can be denoted as $L'_1, L'_2, \cdots, L'_m$ with L'_1 having the highest priority. Then, the capacity update laws are

$$\begin{aligned} c_{L'_1} &= \min\left\{C_{L'_1}, \frac{a_{L_{out}}}{\tau}\right\} \\ c_{L'_2} &= \min\left\{C_{L'_2}, [\frac{a_{L_{out}}}{\tau} - (y_{L'_1} + \frac{n_{L'_1}}{\tau})]^+\right\} \\ &\vdots \\ c_{L'_m} &= \min\{C_{L'_m}, [\frac{a_{L_{out}}}{\tau} - \sum_{l \in \{L'_1, L'_2, \cdots, L'_{m-1}\}} (y_l + \frac{n_l}{\tau})]^+\} \end{aligned} \tag{3.13}$$

The backlog on the incoming links can be flushed in duration τ if there is enough available-occupancy in the outgoing link.

Example 4.4: Suppose the set of priorities for the incoming links in Figure 3.2 $\{L_1, L_2, L_3\}$ be $\pi = [1, 2, 3]$ and the sampling time $\tau = 1$. The capacity distribution among the

incoming links using the proposed model is therefore

$$c_{L_1} = \min\{C_{L_1}, a_{L_{out}}\} \tag{3.14}$$

$$c_{L_2} = \min\{C_{L_2}, [a_{L_{out}} - (y_{L_1} + n_{L_1})]^+\} \tag{3.15}$$

$$c_{L_3} = \min\{C_{L_3}, [a_{L_{out}} - (y_{L_1} + n_{L_1} + y_{L_2} + n_{L_2})]^+\} \tag{3.16}$$

Equation (3.14) captures the fact that L_1 has the highest priority to flush out its backlog. L_2 comes 2nd in the rank and will flush off its content only after the traffic in L_1 is cleared out (3.15). Traffic in L_3 has the least priority and it will move forward last.

Example 4.5: Suppose the occupancy in L_1, L_2, and L_3 are 15, 10, and 5 units of traffic respectively at $t \in T$. The priority ranking based on backlog will be $[1, 2, 3]$. Table 3.1 illustrates the possible scenarios depending on the amount of available-occupancy in L_{out}. The link flow and the maximum capacity variables are excluded in the discussion for the sake of simplicity.

Table 3.1

Merging node Rule 3 illustration

$a_{L_{out}}/\tau$	c_{L_1}	c_{L_2}	c_{L_3}	$n_{L_1}[t+\tau]$	$n_{L_2}[t+\tau]$	$n_{L_3}[t+\tau]$
30	15	10	5	0	0	0
20	15	5	0	0	5	5
15	15	0	0	0	10	5
10	10	0	0	5	10	5
0	0	0	0	15	10	5

Model M4: Fixed proportions: Despite the changes in capacity happening in the outgoing link, the merging node statically allocates capacity based on proportions set beforehand. This model is defined as in (3.12) with proportions $[\varphi_{L_1}, \cdots, \varphi_{L_m}]$ known in advance.

Model M5: One link at a time: Among the links terminating at the merging node, one link is selected based on a priority measure, and it will be allowed to use the entire capacity in the outgoing link during the interval $[t, t+\tau)$. Suppose the incoming links are arranged in descending priority order denoted as $L'_1, L'_2, \cdots, L'_m$ with L'_1 having the highest priority. Then

$$c_{L'_1} = \min\{C_{L'_1}, \frac{a_{L_{out}}}{\tau}\}. \tag{3.17}$$

The traffic in $L'_2, \cdots, L'_m$ will be delayed till another selection takes place at the beginning of the next time interval. This model is reminiscent of traffic policeman who prefers to flush a backlogged traffic.

Example 4.6: For the merging node in Figure 3.2, suppose that $L'_1 = L_2$ and $\tau = 1$. The merging node implementing Model M5 assigns the downstream capacity to L_2.

$$c_{L_2} = \min\{C_{L_2}, \frac{a_{L_{out}}}{\tau}\}. \tag{3.18}$$

Model M6: Through rotation: The incoming links to a merging node will be allowed to use the entire downstream available-occupancy in turn at every clock tick on rotation basis. The order of rotation needs to be determined ahead of time. Let the incoming links be arranged based on their rotation order and denoted as $L'_1, L'_2, \cdots, L'_m$.

$$\begin{aligned} c_{L'_1} &= \min\{C_{L'_1}, \frac{a_{L_{out}}}{\tau}\}, \text{ during } [t, t+\tau) \\ c_{L'_2} &= \min\{C_{L'_2}, \frac{a_{L_{out}}}{\tau}\}, \text{ during } [t+\tau, t+2\tau) \\ &\vdots \\ c_{L'_m} &= \min\{C_{L'_m}, \frac{a_{L_{out}}}{\tau}\}, \text{ during } [t+(m-1)\tau, t+m\tau). \end{aligned} \tag{3.19}$$

This is reminiscent of traffic light. For simplicity we have only shown the case where

every competing link gets equal time.

Example 4.7: Suppose in Figure 3.2, the sequence of rotation is L_2, L_1, L_3.. Thus, L_1 and L_3 must be on hold during the interval L_2 is allowed full access to the capacity of L_{out}. During the next time interval, traffic in L_1 will have the privileged to advance to L_4 while L_3 and L_2 wait on hold for their turn. L_3 transmits in the next time slot while traffic in L_2 and L_1 is stalled till a later time.

All the models suggested for a merging node (3.10) - (3.19) can be used to represent transit nodes where $m = 1$. For example, consider the model in (3.13) which reduces to

$$c_{L_1}[t] = \min\left\{C_{L_1}, \frac{a_{L_{out}}}{\tau}\right\}. \tag{3.20}$$

The model in (3.20) represents capacity collapse propagation at transit nodes in harmony with equation (3.2).

3.4.2 Diverging nodes. Diverging nodes are those nodes with one incoming link and more than one outgoing links. Figure 3.3 shows a typical diverging node with incoming link L_{in} and a set of outgoing links $\{L_6, L_7, L_8\}$. Because all the outgoing links share a common node, a failure in one of the links will affect the flow through the other links as well. The multicommodity flow through the incoming link branches off into the different outgoing links towards their destination.

Suppose there are m outgoing links $L_1, \cdots, L_m$ from a given diverging node. The incoming link is denoted by L_{in} and it carries a mix of flows of multiple destinations. The propagation of capacity collapse to the incoming link and the response of a diverging node to changes in occupancy in the branching links can be modelled in various ways. Every model

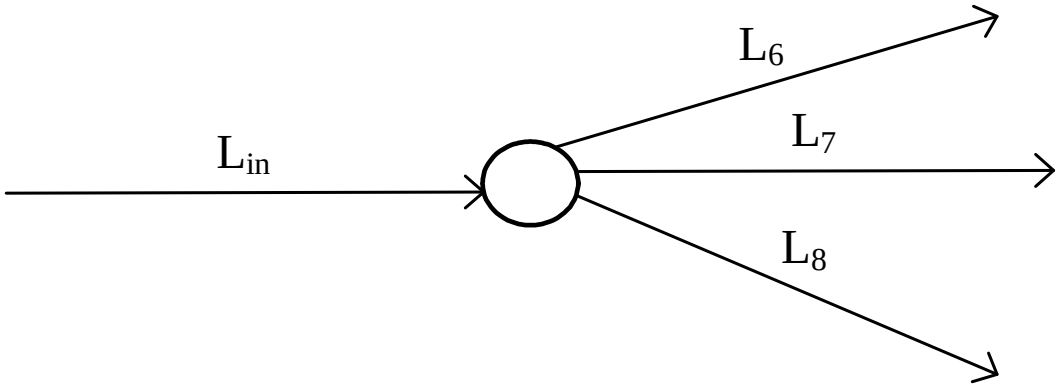

Figure 3.3. A diverging node.

should satisfy the continuity requirement

$$0 \leq c_{L_{in}} \leq \sum_{l \in \{L_1, L_2, \cdots, L_m\}} \frac{a_l}{\tau}. \tag{3.21}$$

If the available-occupancy of the the incoming link goes down to zero, the rate capacity of the link will collapse; i.e.,

$$c_{L_{in}} = 0 \text{ if } a_{L_{in}} = 0. \tag{3.22}$$

In general, the maximum capacity of the incoming link is typically reduced as a function of the backlog occupying the incoming link. Let $C'_{L_{in}}$ denote the net maximum capacity

$$C'_{L_{in}} = [C_{L_{in}} - \frac{\beta \times n_{L_{in}}}{\tau}]^+ \tag{3.23}$$

where $0 \leq \beta \leq 1$ is a constant that can be assumed or determined empirically.

The propagation of capacity collapse to the incoming link and the response of a diverging node to changes in occupancies in the branching links can be modeled in various ways. All of the assumed models implicitly enforce (3.21) and (3.22) in resolving conflicts at the diverging node. Following are some possible models for diverging nodes.

Model D1: The incoming link operates at full capacity.

$$c_{L_{in}} = C'_{L_{in}}, \tag{3.24}$$

where $C_{L_{in}}$ is the maximum capacity of L_{in}.

Model D2: The incoming link operates at a capacity that is the smallest of its maximum capacity and the available-occupancies in the outgoing links,

$$c_{L_{in}} = \min\{C'_{L_{in}}, \frac{a_{L_1}}{\tau}, \frac{a_{L_2}}{\tau} \cdots, \frac{a_{L_m}}{\tau}\}. \tag{3.25}$$

Example 4.8: In Figure 3.3 $m = 4$. The incoming link capacity is therefore,

$$c_{L_{in}} = \min\{C'_{L_{in}}, \frac{a_{L_6}}{\tau}, \frac{a_{L_7}}{\tau}, \frac{a_{L_8}}{\tau}\}. \tag{3.26}$$

Model D3: The incoming link operates at a capacity that is the minimum of the maximum capacity and the total available-occupancy in the outgoing links,

$$c_{L_{in}} = \min\{C'_{L_{in}}, \sum_{l \in \{L_1, L_2, \cdots, L_m\}} \frac{a_l}{\tau}\}. \tag{3.27}$$

Model D4: The incoming link takes up the available-occupancies of the outgoing links on rotation basis. The rotation orders are set in advance. Let the outgoing links be arranged based on their rotation order and denoted as $L'_1, L'_2, \cdots, L'_m$.

$$\begin{aligned}
c_{L_{in}} &= \min\{C'_{L_{in}}, \frac{a_{L'_1}}{\tau}\}, \text{ during } [t, t+\tau), \\
c_{L_{in}} &= \min\{C'_{L_{in}}, \frac{a_{L'_2}}{\tau}\}, \text{ during } [t+\tau, t+2\tau), \\
&\vdots \\
c_{L_{in}} &= \min\{C'_{L_{in}}, \frac{a_{L'_m}}{\tau}\}, \text{ during } [t+(m-1)\tau, t+m\tau).
\end{aligned} \tag{3.28}$$

Model D5: The incoming link operates at a capacity equal to the minimum of its maximum capacity and the available-occupancy of an outgoing link that is chosen randomly.

$$c_{L_{in}} = \min\{C'_{L_{in}}, \frac{a_l}{\tau}\},\ l \in \{L_1, L_2, \cdots, L_m\}. \tag{3.29}$$

3.4.3 Complex nodes. Complex nodes are described by a many to many relationship between sets of incoming and outgoing links. An illustration of a complex node shown in Figure 3.4 has a set of arriving links, $\{L_1,L_2,L_3\}$, and a set of departing links, $\{L_6,L_7,L_8,L_9\}$ sharing a common node. The interaction of multicommodity flows at a complex node is manifold. Complex nodes combine the features in merging and diverging nodes. Hence, it is not straightforward to establish models that could resolve the conflict at the complex nodes when a change in occupancy occurs in the outgoing links.

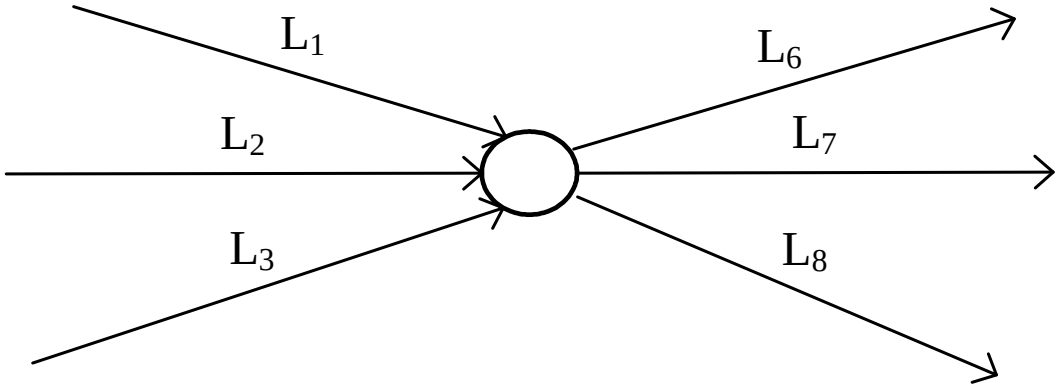

Figure 3.4. A complex node.

A simple but powerful approach is to apply a network transformation to split the complex node in Figure 3.4 into a merging and a diverging nodes as depicted in Figure 3.5. The individual nodes in the transformation are bridged with link L_4 which is assumed to exhibit enough capacity to channel the outbound flow.

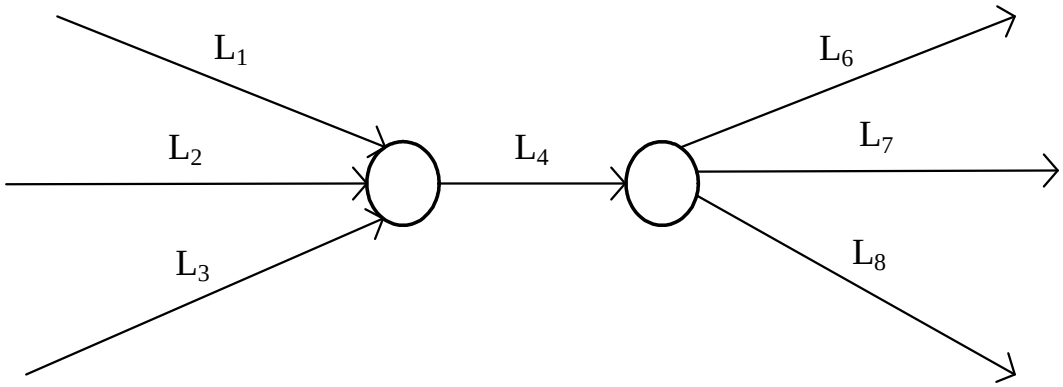

Figure 3.5. A complex node transformed into a merging and a diverging node.

Model C1: Suppose there are m outgoing links at a given complex node, which are denoted as $L_1, \cdots, L_m$. Let L_λ denotes the link that bridges the merging and the diverging nodes in the transformed complex node. For the transformation to keep the properties of the complex node, the maximum capacity of L_λ is set equal to the sum of the maximum occupancies of the diverging links per time interval.

$$C_{L_\lambda} = \sum_{l \in \{L_1, \cdots, L_m\}} \frac{B_l}{\tau} \tag{3.30}$$

The instantaneous capacity of the bridging link, L_λ, is constrained by the capacities of the outgoing links and it follows the models put forward for a diverging node (3.24-3.28). The capacity sharing at the origin of L_λ follows the models posited for a merging node (3.10-3.19). As a direct consequence, we can apply a combination of the merging and diverging models at a given complex node. We have proposed 6 models for merging and 5 models for diverging nodes and thus we can have a total of 30 combinations.

Example 4.9: One model for the complex node can be a combination of M1 of a merging node and D4 of a diverging node.

3.5 A Numerical Experiment As a Demonstration

We consider the network shown in Figure 3.6 which has 13 nodes and 14 links. Nodes N_1, N_2 and N_3 represent source nodes (colored green). N_{12} and N_{13} denote destination nodes (colored red). There are four merging nodes (colored blue): N_5, N_8, N_9, N_{11}, and two diverging nodes (colored yellow): N_4, N_7. Nodes N_6 and N_{10} represent transit nodes (colored black).

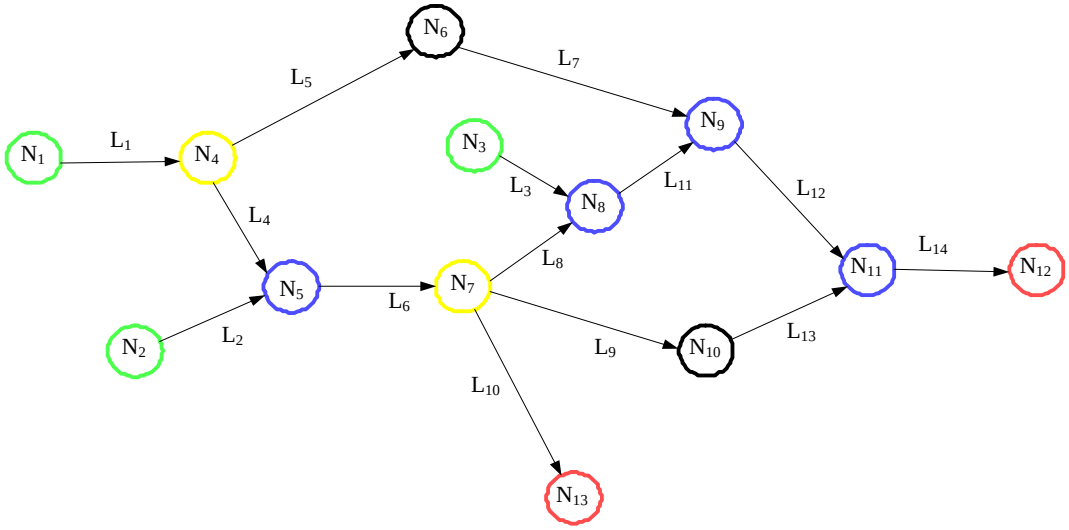

Figure 3.6. An example network to illustrate congestion propagation.

The sections in the links have the same properties: maximum capacity of 30 units of traffic per time interval, and maximum occupancy of 30 units of traffic. Table 3.2 shows the flow demand between respective source and destination node pairs. The shortest path corresponding to each OD pair is also shown.

A simulation of congestion propagation for a 50% capacity reduction in link L_{14} at $t = 5$, and a repair at $t = 80$ is discussed below. The different models proposed for merging and diverging nodes will be analyzed. A combination of 3 merging models (M1 - M3) and 3 diverging models (D1 - D3) gives 9 merging-diverging model pairs whose performance will be compared. The empirical constant β in (3.23) is chosen to be 0.9 at the diverging nodes.

Table 3.2

OD pair flow demand, Example 1

Orig.	Dest.	Flow-rate	Path
N_1	N_{12}	10	P_9 : $N_1 \longrightarrow N_4 \longrightarrow N_6 \longrightarrow N_9 \longrightarrow N_{11} \longrightarrow N_{12}$
N_1	N_{13}	5	P_{10}: $N_1 \longrightarrow N_4 \longrightarrow N_5 \longrightarrow N_7 \longrightarrow N_{13}$
N_2	N_{12}	5	P_{17}: $N_2 \longrightarrow N_5 \longrightarrow N_7 \longrightarrow N_{10} \longrightarrow N_{11} \longrightarrow N_{12}$
N_2	N_{13}	15	P_{18}: $N_2 \longrightarrow N_5 \longrightarrow N_7 \longrightarrow N_{13}$
N_3	N_{12}	5	P_{22}: $N_3 \longrightarrow N_8 \longrightarrow N_9 \longrightarrow N_{12}$

3.5.1 Capacity collapse propagation in the sections of a link. The propagation of a 50% capacity collapse introduced at $t = 5$ in section S_4 of link L_{14} to the preceding sections S_3, S_2 and S_1 is illustrated in Figure 3.7. To simulate the capacity failure, the maximum capacity of S_4 is set to 15 units of traffic per time interval and the maximum occupancy set to15 units of traffic.

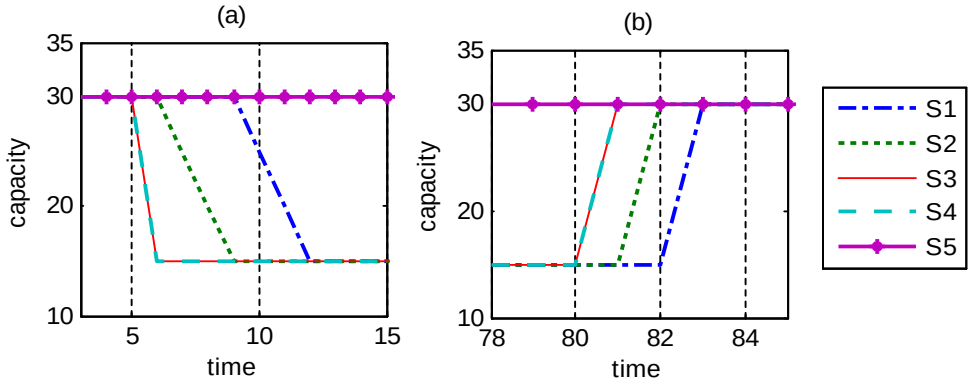

Figure 3.7. Change in capacity of L_{14} propagating through the sections of the link.

Following the introduction of the fault, the capacities of S_4 and S_3 immediately reduced to 15 units of traffic per time interval, as depicted in Figure 3.7(a). The failure propagates upstream and reaches S_1 at $t = 11$. Because the change in capacity propagates slower than the speed of the

traffic, it takes 6 more time steps for the link to feel the effect of the decrease in the capacity of section S_4.

Subsequently a repair is conducted at time $t = 80$. The maximum capacity and the maximum occupancy are set to their initial values. The propagation of the change in the capacity of S_4 towards the first section, S_1, is shown in Figure 3.7(b). It takes 3 time intervals for the link to be affected by the capacity restoration. The difference in propagation time during failure and recovery indicates that the density wave moves faster than the capacity collapse wave.

3.5.2 Comparison of merging models. The capacity diminution at L_{14} affects the flows injected at nodes N_1, N_2 and N_3 through links L_1, L_2 and L_3. We choose L_1 to demonstrate the capacity collapse propagating from L_{14} to the flow sources using assumed capacity collapse models. First every node is assumed to follow M1. Subsequently all nodes follow M2 and so forth. A comparison of the merging models M1 (equal sharing), M2 (Random proportions) and M3 (Based on Priority) is shown in Figure 3.8. The diverging model is set to D3.

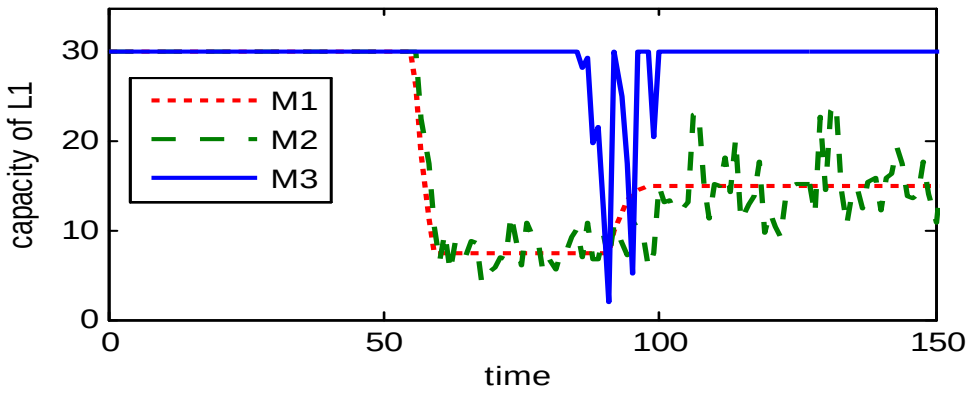

Figure 3.8. Comparison of merging models M1, M2 and M3.

Models M1 and M2 are affected by the failure at $t = 5$ more quickly than M3. They also took longer to respond to the capacity repair at $t = 80$ to the extent that the full capacity is not restored. On contrary, Model M3 is shown to respond very quickly to the repair and the

full capacity is restored. This indicates that model M3 is more efficient and it increases capacity utilization in the network.

3.5.3 Comparison of diverging models. A comparison of the different diverging models is shown in Figure 3.9. All merging nodes are modeled using M3.

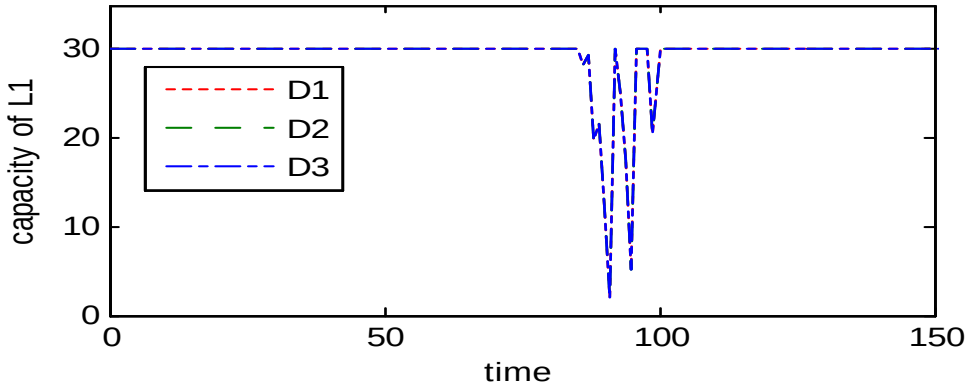

Figure 3.9. Comparison of diverging models D1, D2 and D3.

The diverging models produce similar results shown by the overlap of the lines in Figure 3.9. This indicates that the common factor in the formulation of the models, i.e. $C'_{L_{in}} = [C_{L_{in}} - \frac{\beta \times n_{L_{in}}}{\tau}]^+$, is the limiting factor.

CHAPTER 4

LP-Based Flow-Rate Control and Flow Survivability Using Rerouting

In this chapter, we utilize link congestion information to devise a flow control scheme that carefully assigns traffic flows into alternative paths. The capacity collapse propagation model is used to update the link prices which are factored in the objective of the optimization problem. The ability of the network to reroute its flows as a survivability criterion will be shown to improve the survivability of the network.

4.1 Network Flow Model

Network flow is governed by the interconnection between paths and links through a routing matrix, R, as shown in Figure 4.1. We will follow in general the notation from [12], with some changes that suit our development. The Network $G(N, L)$ is shared by a set P of possible paths serving a set F of OD pairs. Let $L_r \subseteq L$ be a non-empty set of links that path $r \in P$ spans. This defines a link-path indicator binary matrix $R = (R_{lr},\ l \in L, r \in P)$ of dimension $|L| \times |P|$. The l^{th} link and the r^{th} route of this matrix are related as defined in (2.1)

$$R_{lr} = \begin{cases} 1 & \text{if } l \in L_r \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

Suppose that several paths through the network may substitute for one another and serve the same OD pair; i.e., let $P_f \subseteq P$ be a non-empty set of candidate paths that OD pair $f \in F$ uses. This defines an $|F| \times |P|$ OD pair-path indicator binary matrix $H = (H_{fr}, f \in F, r \in P)$. The matrix entries are defined as

$$H_{fr} = \begin{cases} 1 & \text{if } r \in P_f \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

A network can possibly have $|N|(|N| - 1)$ unidirectional flow demands between every pair of nodes in N. In a practical network, not all nodes in the network make up a demand pair;

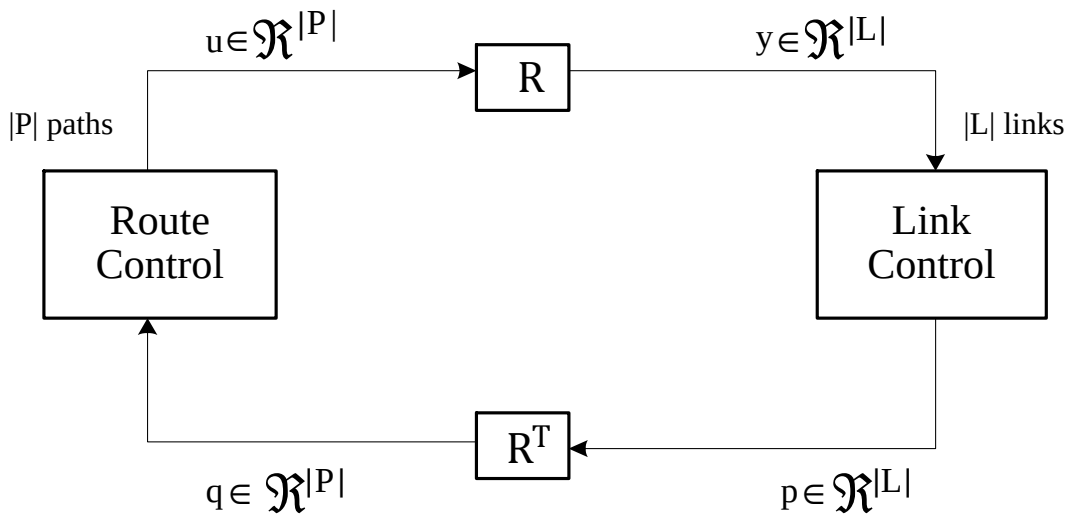

Figure 4.1. Network flow structure.

there are often transit nodes used solely for the purpose of routing.

During the time interval $[t, t+\tau)$, $t \in T$, an OD pair $f \in F$ has a flow demand rate $d_f[t]$ from source i to destination j, where $f \iff (i,j)$, $i \in N$ and $j \in N/\{i\}$. An estimate of the traffic arrival rate $d_f[t]$ for all demand pairs is used during network design in order to guarantee a network with enough capacity and connectivity [2]. We assume this estimate incorporates the additional capacity needed for rerouting at times of disruptions in the network, and external traffic arrives at the beginning of each time slot.

Assuming that the inelastic flows are supported by the network such that there exists a vector of nonnegative path flows, $u[t]$, satisfying

$$\sum_{r \in P} H_{fr} u_r[t] = d_f[t],\ \forall f \in F, \tag{4.3}$$

and

$$y_l[t] \leq c_l[t],\ \forall l \in L \tag{4.4}$$

where $y_l[t]$ and $c_l[t]$ are respectively the aggregate rate and flow capacity at link l. The condition

(4.3) implies that there exists a rate division of the inelastic flow rates over their available routes which can support the arriving traffic. The link capacity constraint (4.4) enforces the requirement that the total flow-rate should not exceed the capacity of the link. The path flow-rate, $u_r[t] \geq 0,$ is measured in terms of the number of traffic assigned to path r at the beginning of the time interval $[t, t+1)$.

4.2 LP-Based Flow-Rate Control via Pricing

It is useful to treat practical flow control schemes simply as implementations of a certain optimization algorithm. The optimization model then makes possible a systematic method to design and refine these schemes, where modifications to a flow control mechanism are guided by modifications to the optimization algorithm.

In many networks, as in the case of transportation networks and sometimes in communication networks, we cannot control the source flow arrival rates. In the event of a link failure, the price for a unit traffic through that link becomes prohibitively expensive, and that could prevent the sources from transmitting. The solutions based on utility maximization will not be of much help in this situation. If we do not have control over the transmission rates at the sources, congestion can occur at various points in the network and network flow will be interrupted. This justifies the use of for flow routing and rerouting schemes that would enhance the reliability of the network.

Associated with each elastic flow there exists a utility function that determines the equilibrium condition as a function of its transmission rate. On the contrary, the exogenous arrival of each inelastic flow is an uncontrollable process which cannot be described using utility functions. The amount of traffic generated by each inelastic flow is unknown and uncontrollable by the network; however, the number of traffic injected into the network can be controlled by the

network algorithm. Therefore, we propose a network optimization algorithm with an objective of minimizing the total cost of network transportation.

The LP-Based network flow control is designed to control path flow rates by utilizing the congestion information sent back to the controller. Let L_ξ denote a set of links outgoing from the source nodes where OD pair flows originate. Because $L_\xi \subset L$, the routing matrix, R, will be reduced to $Q = (Q_{lr}, l \in L, r \in P)$, where

$$Q_{lr} = \begin{cases} 1 & \text{if } l \in L_r \cap L_\xi \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

The controller decides the amount of traffic routed through the paths based on the capacity of the links in L_ξ in addition to the total cost of travelling. The cost per unit flow of using path $r \in P$ at time $t \in T$, $w_r[t]$, is explained as a function of the aggregate link price

$$w[t] = w_0 + q[t] \tag{4.6}$$

where w_0 is the initial cost when there is no congestion, and $q[t]$ is the implied cost of unit flow as defined in (2.6).

The design of the proposed network algorithm is well suited to be studied using techniques of Operations Research. To use the available routes efficiently in a network of multiple OD pairs, we formulate the problem as a linear programming that intends to minimize the total cost of using the paths in the network. Adapting the path-flow multicommodity flow formulation in [2], the

network flow optimization problem is

$$\text{Objective} \quad : \quad \min_{u[t]} \quad \textstyle\sum_{r \in P} w_r[t] u_r[t] \tag{4.7}$$

$$\text{s.t.} \quad : \quad \sum_{r \in P} Q_{lr} u_r[t] \leq c_l[t], \forall l \in L_\xi \tag{4.8}$$

$$: \quad \sum_{r \in P} H_{fr} u_r[t] = d_f[t], \forall f \in F \tag{4.9}$$

$$: \quad u_r[t] \geq 0, \forall r \in P \tag{4.10}$$

The solution to the optimization problem (4.7-4.10) results in a vector of path flow rates that minimizes the total cost of travel. The problem takes into consideration the capacity constraint (4.8) at the onset of routing. The congestion information in the subsequent links is used to compute the cost associated with the paths (4.6) in the network. Moreover, the problem has a flow-rate demand-constraint (4.9) that reenforces the assumption made (4.3). All flow rates should be satisfied through non-negative path flow rates (4.10).

Solution of this problem can reduce the requirement of complex coordination among sources to only those links in L_ξ. This solution adapts to changing network conditions through flow rerouting, where the rerouting is achieved by means of pricing signals. Each link runs a local algorithm to update its price and communicate its computation result to the sources. The network solves the optimization algorithm (4.7-4.10) and determines the rate distribution. Since the optimization problem has a linear objective (4.7), for any two routes r and r^* serving the same OD pair, if $u_r \geq u_{r^*}$, then necessarily $w_r \leq w_{r^*}$.

The flow control algorithms in [12]-[17] essentially assume that packets injected into the source nodes by the flows arrive at the destination nodes instantaneously. In reality, packets will reach downstream nodes only after a queueing and propagation delay incurred in the intermediate nodes. For this reason, we track the link flow at each section of the links. The sampling interval is

chosen in such a way that a free flow traffic traverses a link in one time interval.

Every link in a network can be shared by more than one set of OD pair flows. The aggregate flow entering a link at $t \in T$, $y[t]$, is given as the sum of the flows from upstream links and the OD pair flows originating at the link.

$$y_l[t] = \sum_{j \in \Lambda_l} y_j^{out}[t - \tau] + \sum_{r \in P} Q_{lr} u_r[t], \ \forall l \in L \tag{4.11}$$

where Λ_l is the set of links incident on link l. y_j^{out} is the amount of traffic leaving link j during the time interval $[t-1, t)$. $Q_{lr}u_r$ is external traffic injected to the link l at the beginning of each time slot.

4.3 Rerouting as a Recovery Technique

A framework in [3] proposes a reliability analysis based on the notion of routing and rerouting after failure. The methodology underlines the importance of the routing in the reliability of flow networks. Each inelastic flow demand is associated with a fixed set of routes. The routes are required to be mutually exclusive if possible.

4.3.1 Span and path restoration. When a link fails, the corresponding flow restoration should take place in a subgraph where the failed link is removed. Span restoration, or backlog rerouting, reroutes flow rates over replacement path segments between the two nodes terminating a span failure. It provides replacement paths originating at the node directly adjacent to the failed link towards the destinations of the disturbed flows.

Given a path r made of consecutive links $l_1, l_2, \cdots l_n$ serving OD pair $O - D$, a span restoration corresponding to a fault in l_j on r reroutes the flow on r to a path r' made of consecutive links $l'_1, l'_2, \cdots l'_m$ such that:

1. r' bypasses the fault and reaches destination D; and

2. the origin of r' is O', a node on r, but where the link l'_1 originating at O' is not part of r.

In contrast, Path restoration can be achieved by rerouting each flow-rate demand affected by the failure individually from its origin to its destination through a replacement path [22], [24].

4.3.2 Successively shortest first-link-disjoint paths. The creation of backup paths in a network is an important network design problem. These paths are needed to restore connectivity in the case of link failure and it is a convenient way to improve the reliability service delivery. The ideal backup path for link failure would have no links in common with the original path for a connection. In this case, a failure anywhere on the path will not disconnect the corresponding flow. In certain topologies it is not possible to find two completely disjoint paths due to the network structure. In such a case, it would be helpful to find the best partially-disjoint backup path.

One of the issues concerning flow survivability is the choice of criterion for rerouting: k successively shortest link-disjoint paths (KSP) or Maximum flow (Max Flow). KSP is faster and easier to implement, but not strictly optimal in terms of finding the maximal number of paths. A theoretically optimal restoration capacity is obtained with a Max Flow criterion. A comparative study of the two criteria shows that they produce extremely close results in span restoration and a similar result could be obtained for path restoration [22].

We have considered the Dijkstra shortest path algorithm in MATLAB to obtain the set of all paths P corresponding to the set of all OD pairs F using a variant of the KSP criterion. The new criterion generates a set of k successively shortest first-link-disjoint paths between two nodes first by finding the shortest path, then the second shortest alternative path that is first-link-disjoint with the first path, and so on. The links used in a path would be assigned more weights so as to discourage the alternative paths using them. First-link-disjoint paths form a set of paths that originate from a given source node through divergent links going towards a given destination node.

The set of paths generated using this criterion contain both partially-disjoint and totally-disjoint alternate paths.

We are more interested in the links used by the paths and the OD pairs served by the paths. The OD pairs and the corresponding paths are contained in H as defined in (4.2). The links corresponding to the paths are held within R as defined in (4.1).

4.3.3 Phased recovery. A phased recovery model in [9] describes the life cycle of failure and recovery in four phases: failure, rerouting, repair, and normal phases. The cycle starts in failure phase and steps through all phases before it returns to failure free mode. The sequences are summarized in Figure 4.2.

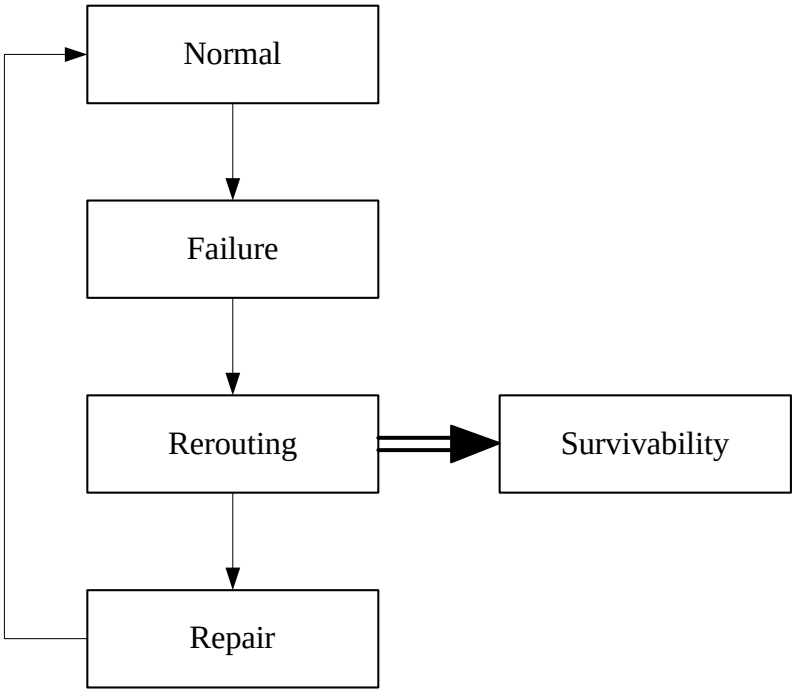

Figure 4.2. Failure recovery model.

Immediately after the failure, congestion information in the form of an increase in link prices will be fed back to the flow sources. In the meantime, the flows are routed according to the original routing scheme. A recovery strategy to restore the essential services in the network will

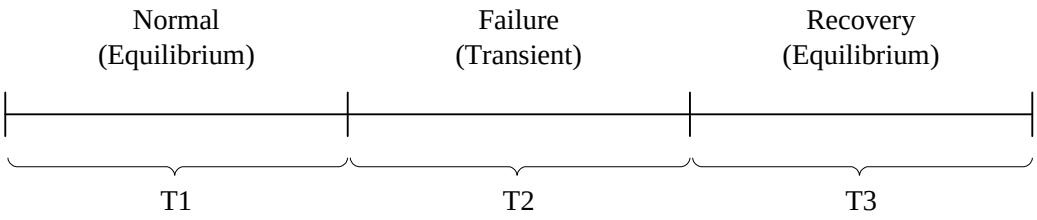

Figure 4.3. Temporal axis partitioning.

be initiated in the next phase. Recovery of services after failure is a key property that survivable systems must exhibit.

In general, rerouting the demand after a failure gives better performance than curtailing the demand when a component fails. A flow affected by the failure is routed to a link-disjoint alternative path that bypasses the failed link. When rerouting is effective reliability enhancer, the flow survives in the presence of the failure. At the end of the repair phase, the system returns to failure free normal state with the original routing scheme put back in place.

4.3.4 A 3-phase model. The dynamics of failure and recovery is portrayed in [3] by dividing the time axis into three phases. Figure 4.3 shows the temporal axis partitioned into three stages that correspond to the normal, failure and recovery phases of the phased recovery approach in Figure 4.2.

During T1 the system is in its normal state and has reached equilibrium. Once a failure takes place at the beginning of T2, a congestion is formed in the failed link due to the reduction of its capacity. During T2, a transient performance unfolds from the instant an undesirable event occurs until steady state where an acceptable performance level is attained. During this phase, the failure propagates and backlog accumulates on the links upstream from the failed link. The increment of congestion in the links degrades the quality of services provided by the network and requires a change in the flow routing scheme. Otherwise, the network will not survive the

failure. In phase T3, a survivable system reroutes each flow demand to paths that are disjoint to the original route, at least at the failed link.

Networks can exhibit large variations in survivability requirements. The time phase T2 in Figure 4.3, also referred to as recovery time, differs from one system to another. In some networks the recovery times can be measured in hours, whereas embedded command and control systems may require recovery times to be in milliseconds. Survivability quantification models in [25] and [9] analyze the transient performance of a network under stress.

The network algorithm we proposed in (4.7-4.10) intends to efficiently assign each flow into the network to traverse each of the available paths. The links individually update their prices (3.7) and the controller is fed back with the aggregate link price (3.8). In the event of link failures, the increase in link prices will be communicated to the controller which in turn reroutes the flow into available alternative routes.

CHAPTER 5

Numerical Simulation of Active Rerouting

5.1 Capacity Collapse propagation

A network of 18 nodes and 21 links is shown in Figure 5.1. Nodes N_1 and N_{10} are flow origins. Nodes N_{17} and N_{18} represent destinations. Nodes N_2, N_7, N_9, and N_{15} are diverging nodes. Nodes N_8, N_{11} and N_{16} are merging nodes. The sections in all the links have a maximum occupancy of 30 units of traffic. The maximum capacity corresponding to each section is 30 units of traffic per time step. Nodes N_8 and N_9 represent a transformed complex node with a bridging link L_{10}. Thus the maximum occupancy and maximum capacity of L_{10} is the sum of the corresponding parameters in the outgoing links.

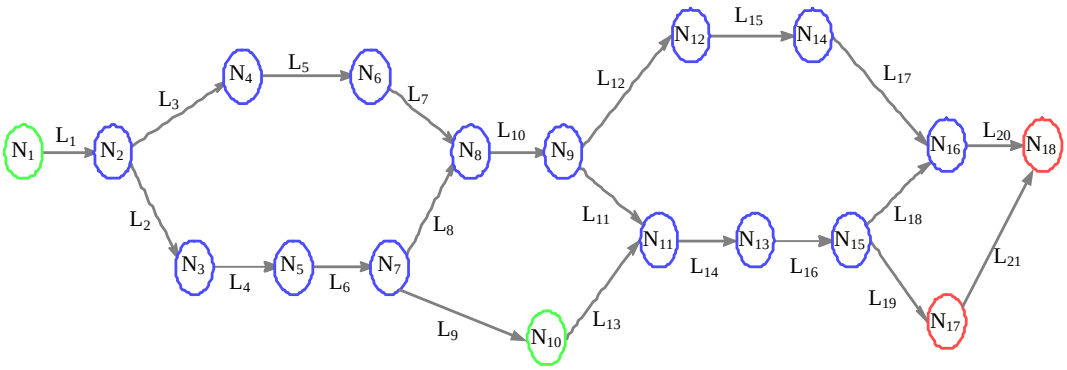

Figure 5.1. A network with 2 origins and 2 destinations.

The OD flow demand is shown in Table 5.1. The demand pairs, the flow demands and the corresponding shortest paths are also shown. For example, the first demand pair has a flow demand of 5 traffic per time from the origin node N_1 to the destination node N_{17} through path P_{16}.

In the following sections different failure scenarios will be discussed. The propagation of capacity collapse is implemented using merging model M3 and diverging model D3, introduced in Chapter 3. A gray scale color map is used to indicate the percentage of capacity collapse in the

Table 5.1

OD pair flow demand, Example 2

Orig.	Dest.	Flow-rate	Path
N_1	N_{17}	$d_{16} = 5$	P_{16}: $N_1 \longrightarrow N_2 \longrightarrow N_3 \longrightarrow N_5 \longrightarrow N_7 \longrightarrow N_{10} \longrightarrow N_{11} \longrightarrow N_{13} \longrightarrow N_{15} \longrightarrow N_{17}$
N_1	N_{18}	$d_{17} = 20$	P_{17}: $N_1 \longrightarrow N_2 \longrightarrow N_4 \longrightarrow N_6 \longrightarrow N_8 \longrightarrow N_9 \longrightarrow N_{12} \longrightarrow N_{14} \longrightarrow N_{16} \longrightarrow N_{18}$
N_{10}	N_{18}	$d_{170} = 5$	P_{131}: $N_{10} \longrightarrow N_{11} \longrightarrow N_{13} \longrightarrow N_{15} \longrightarrow N_{16} \longrightarrow N_{18}$

links. The color map ranges from white representing a 0% capacity to black representing 100% capacity. The empirical constant β at the diverging nodes is chosen to be 0.9.

5.1.1 A 25% capacity reduction in link L_{15}. A 25% capacity reduction is introduced into link L_{15} at $t = 5$. The flow-rate through the sections of link L_{15} is 20 traffic units per time interval. A 25% capacity loss reduces the capacity of the section affected by the failure to 22.5 traffic units per time interval. So, the damage will not affect the flow through the section, and therefore no congestion wave will propagate through the link.

5.1.2 A 50% capacity collapse in L_{15}. Here half of the capacity of L_{15} is lost due to a failure occurring at $t = 5$. The capacity change propagating into the network is shown in Figure 5.2. The capacity reduction in L_{15} propagates through the sections of the link towards L_{12}.

The collapse further propagates through the sections of L_{12} and L_{10} within $t \leq 30$. Between $t = 30$ and $t = 65$, the collapse spreads through the sections of L_7, L_5 and L_3. The collapse wave took 63 time intervals to propagate all the way to L_1. L_{10} has undergone through a 75% capacity reduction from 60 traffic units per time step to 15. All the other links remain unaffected by the failure.

A closer look at L_{15} indicates that the first vertical line represents the capacity of the link. The link has 5 sections whose capacities are represented by the other five lines. The failure happened at the 4th section and propagated backward to preceding sections. The section immediately after that is not affected by the failure.

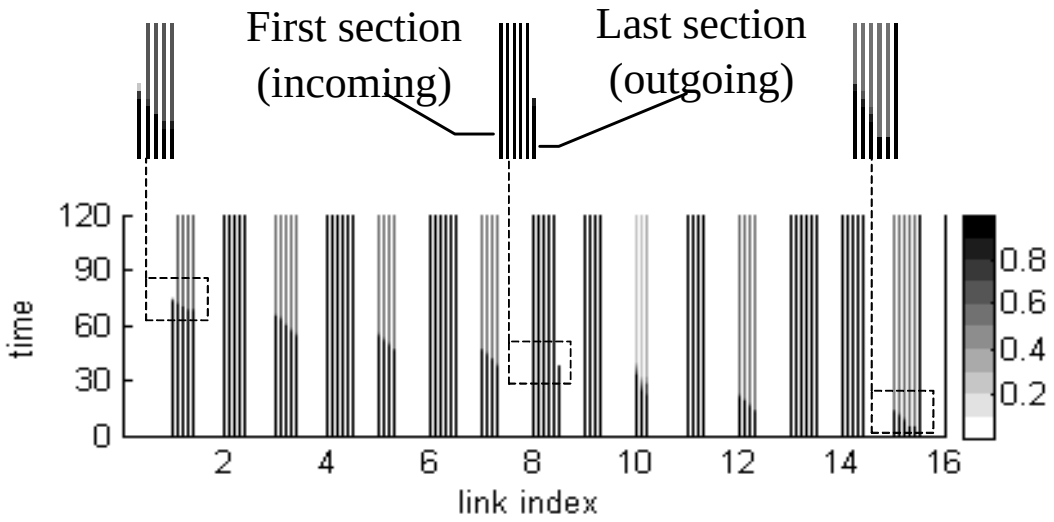

Figure 5.2. Propagation of 50% capacity collapse in link L_{15} section 4 at $t = 5$.

The collapse is sensed by the last section of $L_{8,}$ but it is not propagated further as there was no flow coming through the link. The capacity of L_1 totally collapses as it is fully occupied by the traffic originating at the link.

5.1.3 A 75% capacity collapse in L_{15}. Suppose a 75% capacity reduction of link L_{15} has occurred at $t = 5$. The capacity collapse propagating in the network is shown in Figure 5.3.

A capacity reduction of 70-80% is observed in the sections of links L_3, L_5, L_7, L_{12} and L_{15}. A 80-90% capacity collapse is observed in L_{10}. The sections in L_1 undergo a capacity reduction ranging from 60-80%, and the link has zero capacity to support OD pair flows originating at N_1.The collapse wave took 37 time steps to arrive at the farthest link, L_1. This wave propagated 1.7 times faster than the wave with 50% capacity collapse.

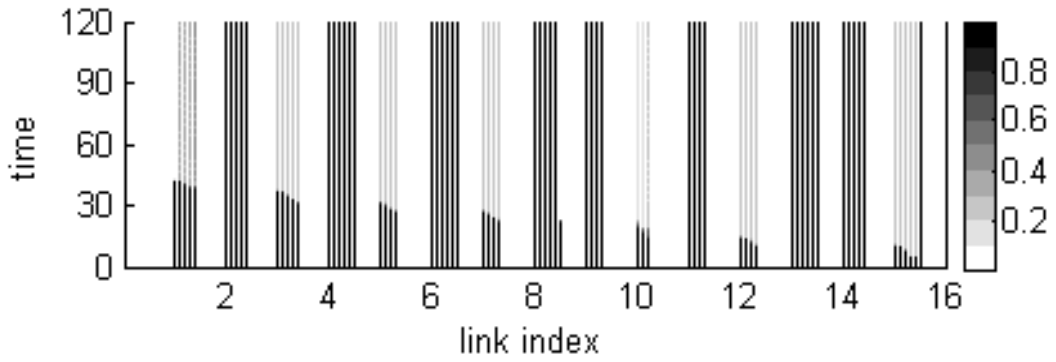

Figure 5.3. Propagation of 75% capacity collapse in link L_{15} section 4 at $t = 5$.

5.1.4 A 100% capacity collapse in L_{15}. A total capacity damage is introduced to link L_{15} at $t = 5$. The collapse in capacity propagating to preceding links in the network is shown in Figure 5.4.

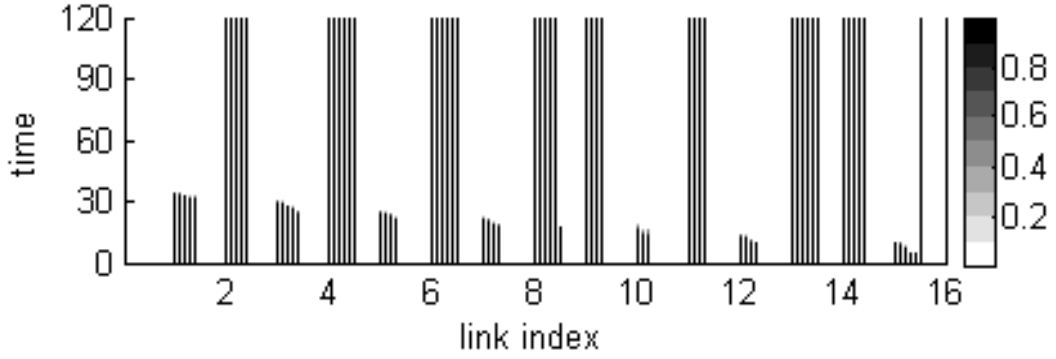

Figure 5.4. Propagation of 100% capacity collapse in link L_{15} section 4 at $t = 5$.

The capacities of L_1, L_3, L_5, L_7, L_{10}, L_{12} and L_{15} are shown to disappear well before $t = 35$. The collapse wave travels much faster in this scenario. It took 29 time steps for the wave to arrive at the first section of L_1. The wave travels 1.28 faster than the scenario with 75% collapse, and it moves 2.17 times faster than the case with 50% failure. The larger the capacity reduction, the faster the capacity collapse propagates in the network.

5.1.5 A repair of the failed link. A repair restores the capacity of the link to its full operation. To demonstrate how the capacity restoration is propagated, we choose a 50% capacity collapse of link L_{15} at $t = 5$ and a repair at $t = 80$. The capacity change propagation due to the collapse and later due to the repair is shown in Figure 5.5.

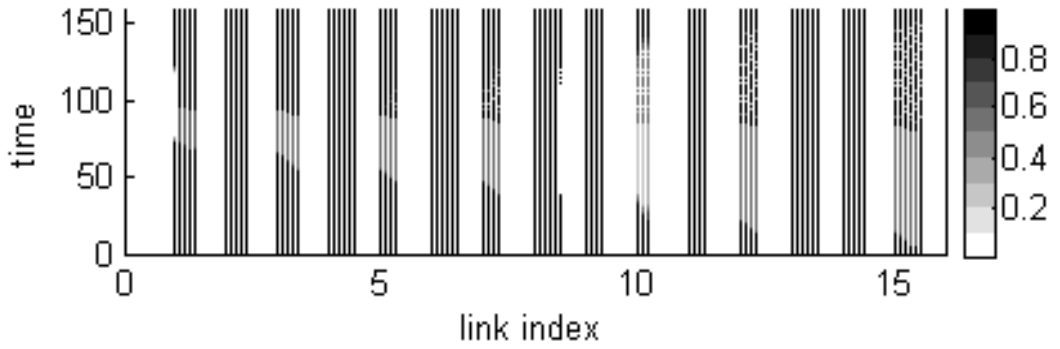

Figure 5.5. 50% capacity collapse in link L_{15} at $t = 5$, and a repair at $t = 80$.

Following the restoration of the capacity of L_{15}, all the links that have been affected by the capacity reduction start operating at their full capacity. The restoration wave takes 14 time intervals to travel all the way to L_1. The restoration wave is shown to be 4.5 times faster than the collapse wave. The capacity swing in L_{15} during the recovery phase is due to the assumed capacity collapse propagation model at the merging node N_{16}. The allocation of capacity to the incoming links, L_{17} and L_{18}, is done based on the occupancy of the links. The link with the larger backlog has a higher priority to claim the resource.

The backlog accumulated at L_1 takes 58 time steps to clear out as shown in Figure 5.6. The capacity of L_1 is fully restored starting from $t = 124$.

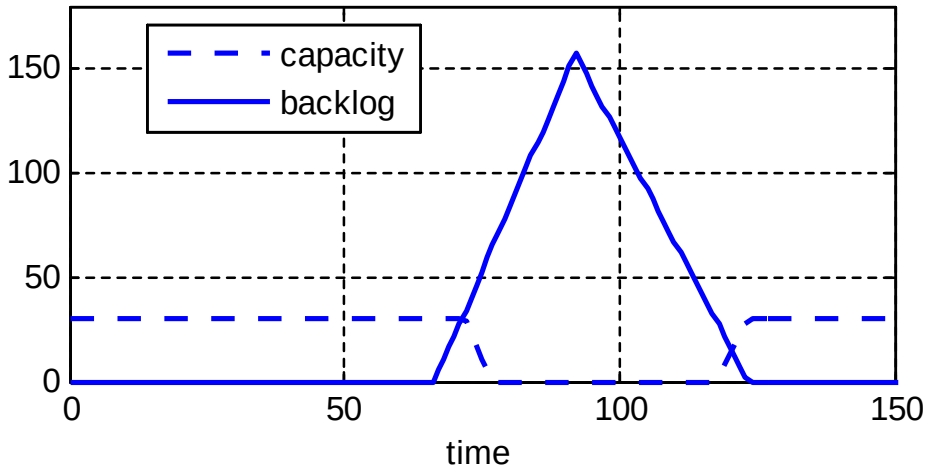

Figure 5.6. Capcity and backlog at link L_1.

5.2 Span Restoration

Most of the node pairs in Figure 5.1 do not have flow demands associated with them at the beginning. But, as the traffic proceeds in the assigned path, a capacity collapse in one of the links would require the traffic to change its route in order to avoid the congested links. The need to reroute the traffic going through a clogged path would reassign the backlog as a flow demand in subsequent nodes. This leads to a span routing approach that reroutes backlogged traffic at a link using alternative path segments starting at the end node of the link.

5.2.1 Capacity collapse in link L_{15}. For the failure scenario where we introduced a 50% capacity loss in L_{15}, the backlog accumulated in the links of the network is shown in Figure 5.7.

The capacity collapse at L_{15} blocks the incoming traffic from N_1 to N_{18} that has been assigned to path P_{17} as shown in Table 5.1. The backlog in L_1 keeps increasing because the backlog in the downstream links is not rerouted.

In contrast, a backlog rerouting scheme at N_9 reassigns the incoming traffic through another path that is link disjoint to the failed link. Figure 5.8 shows that a flow demand between N_9 and N_{18} is being introduced to reroute the flow that was intended to pass through the failed

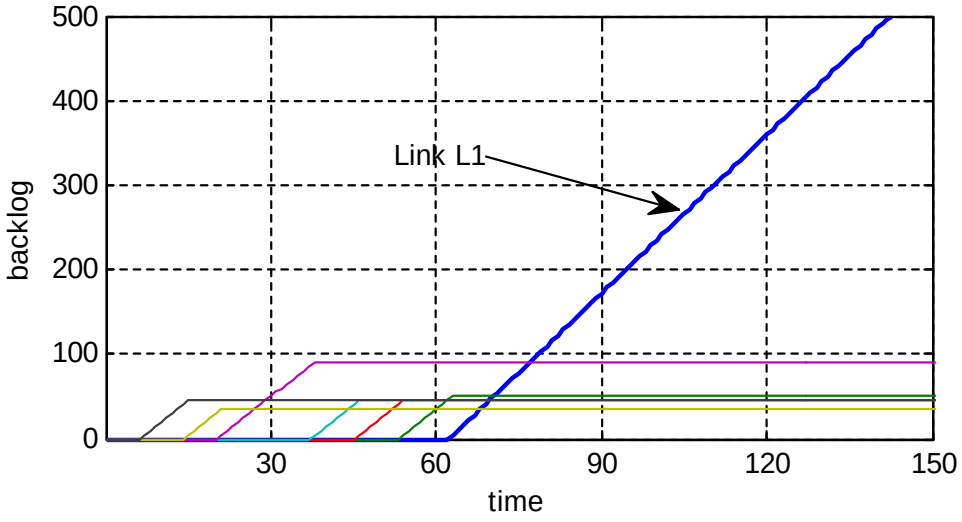

Figure 5.7. Backlog continues to build up unless re-routed.

link, L_{15}.

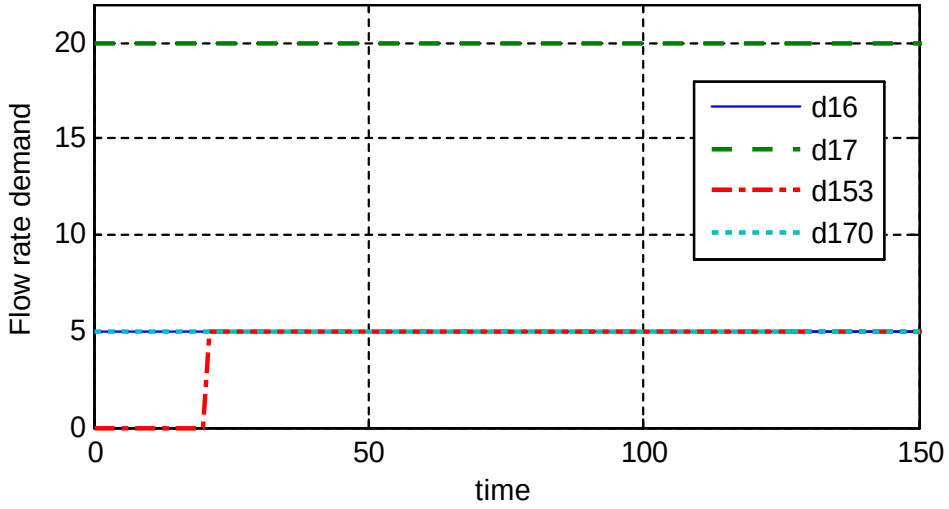

Figure 5.8. Backlog re-routed as demand at intersection node.

The reassignment of the backlog in L_{10} as a new demand at N_9, d_{153}, has reduced the overflow through L_{12} and L_{15} as shown in Figure 5.9. The excess flow that would have resulted in backlog accumulation towards the source of the flow, as in Figure 5.7, is rerouted through P_{125} ($N_9 \rightarrow N_{11} \rightarrow N_{13} \rightarrow N_{15} \rightarrow N_{17} \rightarrow N_{18}$) that serves OD pair N_9 to N_{18}.

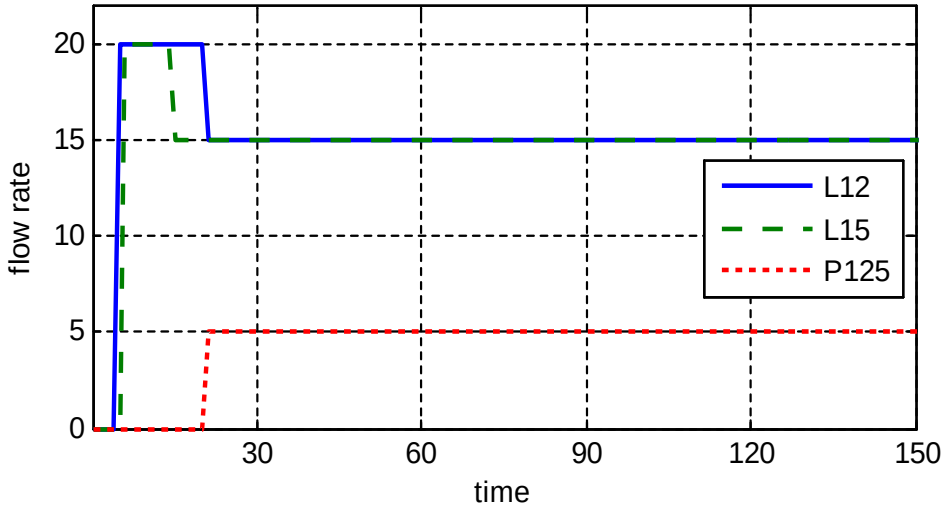

Figure 5.9. Link overflow rerouted through P125.

Due to the backlog rerouting, the continuous increment of backlogs is averted as shown in Figure 5.10. The capacity collapse at L_{15} propagates only through L_{12} and L_{15} and both links continue operating at failed state. In comparison, all the other links that would have been affected by the damage in L_{15} will now operate at full capacity.

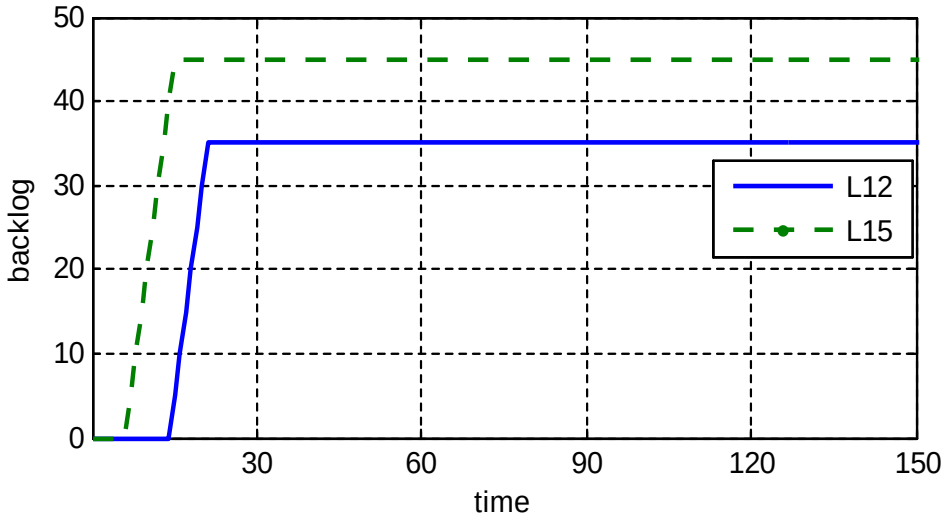

Figure 5.10. Backlog after re-routing.

5.2.2 Capacity collapse in links L_9 and L_{15}. We introduce a 100% failure in L_9 and a 50% capacity reduction in L_{15} at $t = 5$. The capacity collapse propagating to other links of the network is shown in Figure 5.11.

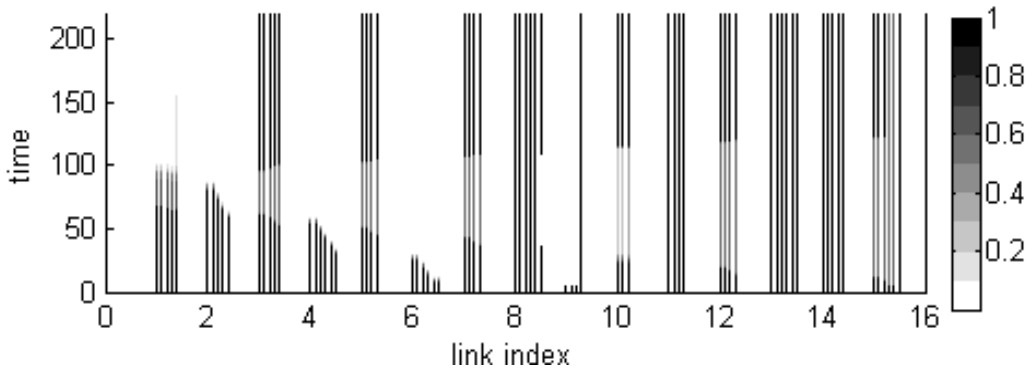

Figure 5.11. 100% capacity collapse in L_9 and 50% collpase in L_{15} at $t = 5$.

Next we compare Figure 5.11 with the situation in Figure 5.2. There we had a single link failure, which was a 50% capacity reduction in L_{15}. The comparison reveals that in Figure 5.11 the capacities of links L_1, L_2, L_4 and L_6 were reduced to a 90-100% failure before $t = 100$. The traffic units in L_1 intending to go through the failed links totally block the incoming traffic. This is reflected in L_3, L_5, L_7, L_{10}, L_{12} and L_{15} as the capacities are restored.

The backlog accumulated in the links due to the failure in links L_9 and L_{15} is depicted in Figure 5.12(a). It has spread into many links and the backlog in link L_1 is shown to be increasing indefinitely. Figure 5.12(b) illustrates the backlog under backlog rerouting.

The rerouting of the backlogs into other paths to avoid the failure puts up new flow demand requests at other junctions as shown in Figure 5.12(d). d_{118} and d_{153} denote OD pair flow-rate demands from N_7 to N_{17} and from N_9 to N_{18} respectively. The traffic units in L_9 will have to wait until the failure in the link is repaired. Links L_{12} and L_{15} operate at reduced capacity with some backlog till recovery. The rerouting has prevented backlog build up in all the other links and improves network capacity utilization.

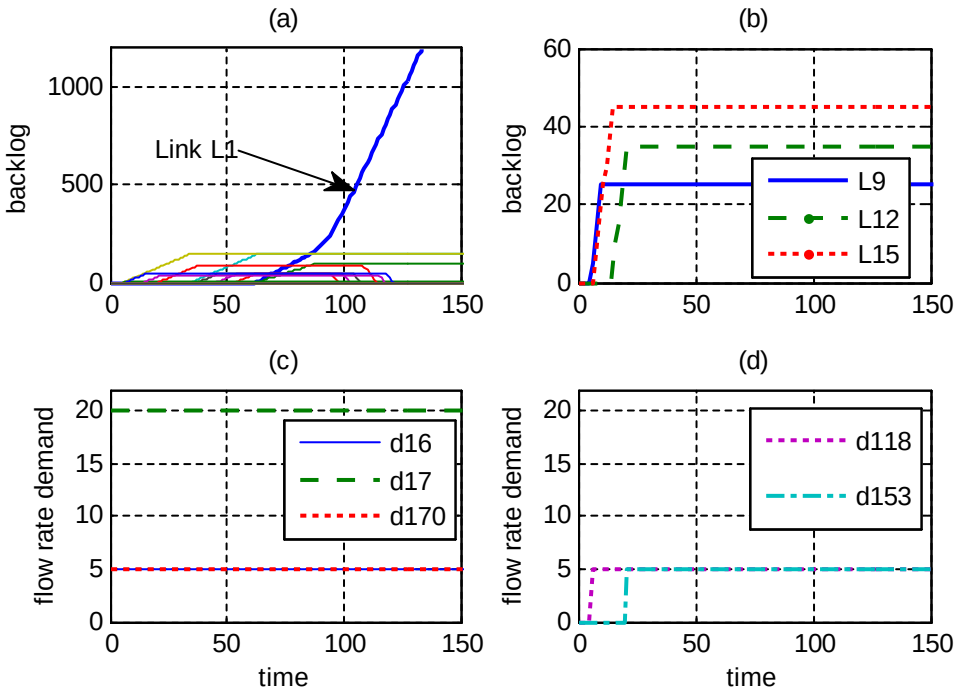

Figure 5.12. Backlog with and with out re-routing.

5.3 Improving Flow Survivability via Routing

To illustrate further the importance of routing to reliability and survivability of flow networks, we simulate the network of 21 nodes and 28 links shown in Figure 5.13. Note that although this network has no cycles, our methodology does not require this assumption.

The nodes are categorized as diverging, merging and transit in Table 5.2. Nodes N_1 and N_{13} are flow origins. Nodes N_{20} and N_{21} represent destinations.

Table 5.2

The nodes grouped by their type

Type	Nodes
Diverging	N_1, N_2, N_9, N_{10}, N_{17}, N_{18}
Merging	N_5, N_6, N_{15}, N_{11}, N_{16}, N_{19}, N_{20}, N_{21}
Transit	N_3, N_4, N_7, N_8, N_{12}, N_{13}, N_{14}

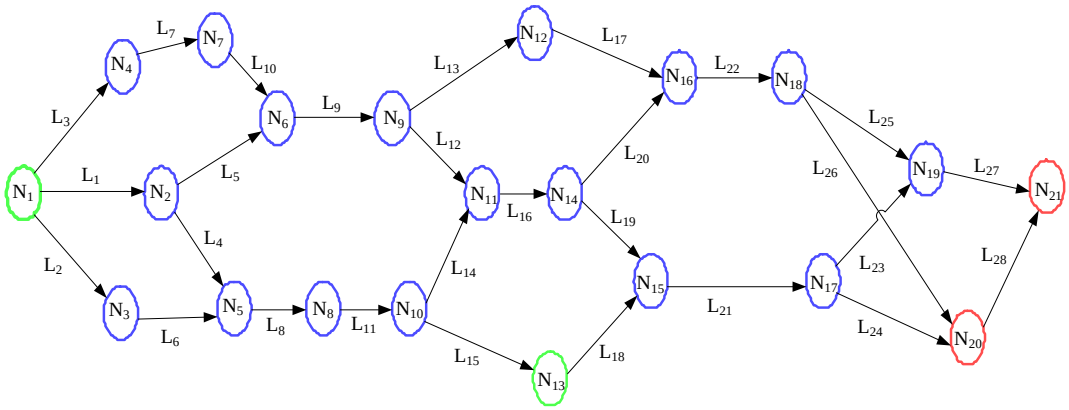

Figure 5.13. A network to demonstrate path and span rerouting.

The sections in all the links are assumed to have identical properties: maximum occupancy, $B_s = 30$ units of traffic, and maximum capacity, $C_s = 30$ traffic units per time step. Nodes N_{11} and N_{14} represent a transformed complex node with a bridging link L_{16}. Thus the maximum occupancy and maximum capacity of L_{16} is the sum of the corresponding parameters in the outgoing links, L_{19} and L_{20}.

The demand pairs, the flow demands and the corresponding first-link-disjoint shortest paths are shown in Table 5.3. OD pair flow demand d_{19} has a flow-rate of 5 traffic units per time step from origin node N_1 to destination node N_{20} through paths P_{40}, P_{41} and P_{42}. P_{40} is the shortest, and P_{41} and P_{42} are alternatives to P_{40}.

There are also three paths serving OD pair demand d_{20} and only one path along demand pair d_{260}. The routing control algorithm takes link congestion information into account, in addition to the shortness of a path, when selecting a path.

5.3.1 Capacity collapse in link L_{13}. Suppose a 100% failure is introduced to link L_{13} at time $t = 5$. The link has 5 sections, and the failure occurred at section 4. The failure affects OD pair flow demands d_{19} from node N_1 to N_{20} and d_{20} from node N_1 to N_{21} whose flow has been assigned to paths P_{40} and P_{43} respectively. The propagation of capacity collapse wave to upstream

Table 5.3

OD pair flow demand, Example 3

OD pair demand	Flow-rate	OD pair Paths
d_{19}	5	P_{40}: $N_1 \rightarrow N_2 \rightarrow N_6 \rightarrow N_9 \rightarrow N_{12} \rightarrow N_{16} \rightarrow N_{18} \rightarrow N_{20}$ P_{41}: $N_1 \rightarrow N_3 \rightarrow N_5 \rightarrow N_8 \rightarrow N_{10} \rightarrow N_{13} \rightarrow N_{15} \rightarrow N_{17} \rightarrow N_{20}$ P_{42}: $N_1 \rightarrow N_4 \rightarrow N_7 \rightarrow N_6 \rightarrow N_9 \rightarrow N_{11} \rightarrow N_{14} \rightarrow N_{16} \rightarrow N_{18} \rightarrow N_{20}$
d_{20}	20	P_{43}: $N_1 \rightarrow N_2 \rightarrow N_6 \rightarrow N_9 \rightarrow N_{12} \rightarrow N_{16} \rightarrow N_{18} \rightarrow N_{20} \rightarrow N_{21}$ P_{44}: $N_1 \rightarrow N_3 \rightarrow N_5 \rightarrow N_8 \rightarrow N_{10} \rightarrow N_{13} \rightarrow N_{15} \rightarrow N_{17} \rightarrow N_{19} \rightarrow N_{21}$ P_{45}: $N_1 \rightarrow N_4 \rightarrow N_7 \rightarrow N_6 \rightarrow N_9 \rightarrow N_{11} \rightarrow N_{14} \rightarrow N_{15} \rightarrow N_{17} \rightarrow N_{20} \rightarrow N_{21}$
d_{260}	5	P_{190}: $N_{13} \rightarrow N_{15} \rightarrow N_{17} \rightarrow N_{19} \rightarrow N_{21}$

links is shown in Figure 5.14.

The backward travelling wave moves through link L_{13} in the interval between $t = 5$ and $t = 10$. The link failure spreads to links L_9 and L_5 in the interval $11 \leq t \leq 15$. During the interval $t = 16$ to $t = 20$, the collapse further diffuses through L_1 towards the flow origin node N_1. All the other links remain immune to the failure at L_{13}. The flow-rate assignment and the backlog in the

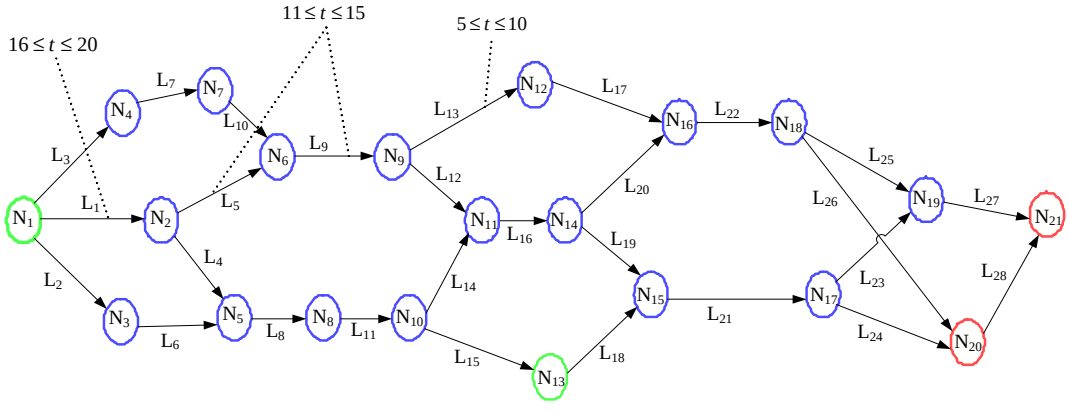

Figure 5.14. Propagation of capacity collapse wave.

links is shown in Figure 5.15.

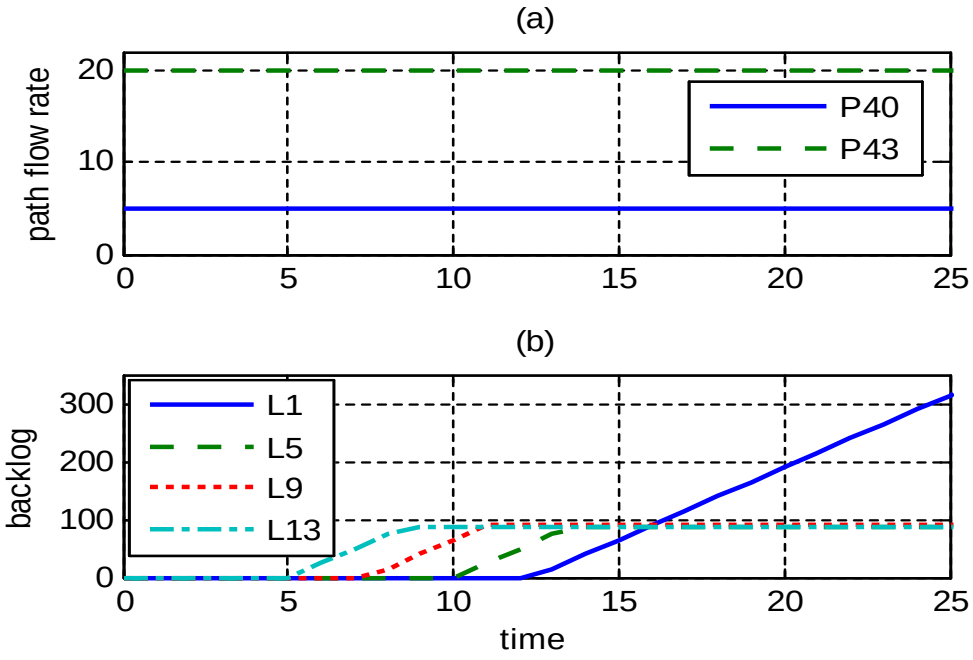

Figure 5.15. Path flow and backlog in the links of the paths.

The flow-rate demands of node pairs (N_1,N_{20}) and (N_1,N_{21}) will be continually loaded to P_{40} and P_{43}, respectively, unless rerouting is considered, as shown in Figure 5.15(a). The decline in capacity has disrupted the flow through the links in the failed paths, P_{40} and P_{43}. This results in a backlog build up as depicted in Figure 5.15(b). The backlog in L_{13}, L_9, and L_5 saturate but the backlog in L_1 keeps increasing because the incoming flow is not rerouted.

The uncontrolled increment of traffic backlog at L_1 suggests rerouting the flow in paths P_{40} and P_{43} to alternative paths. The alternate paths serve the same OD pair but pass through links that are least affected by the damage in L_{13}.

5.3.2 Flow restoration. The accruement of traffic backlog in the links affirms that the network fails to survive the link failure. A simulation of flow restoration through path flow rerouting is illustrated in Figure 5.16.

The failure in link L_{13} at $t = 5$ is communicated to the flow assignment controller at node

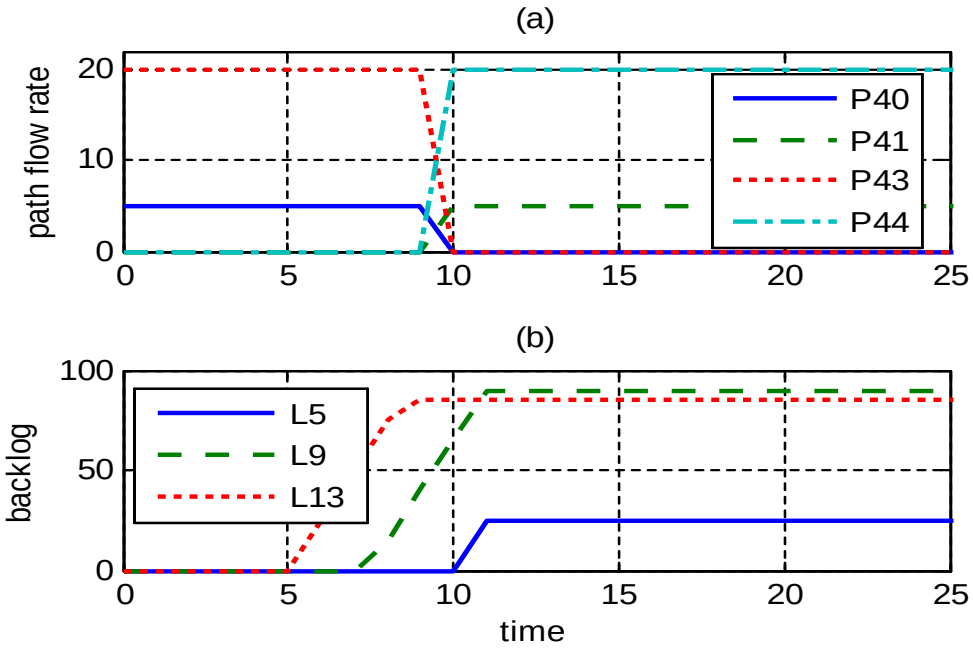

Figure 5.16. Path flow rate and backlog in the links of the paths with path re-routing.

N_1 through an increase in the prices of the links affected by the failure. The links update their price at each time instant, and the controller computes aggregate prices every tenth time instant.

Based on the price information, starting from $t = 10$, the controller reroutes the flow demands originating at N_1 to paths P_{41} and P_{44} as depicted in Figure 5.16(a). Paths P_{41} and P_{44} are substitutes to P_{40} and P_{43} serving OD pair demands d_{19} and d_{20}.

The backlog is prevented from further buildup towards L_1 as portrayed in Figure 5.16(b). But, the traffic accumulated in links L_5, L_9 and L_{13} will be backlogged until the failed link is restored. Otherwise span restoration needs to be in place.

The LP-based controller responds to the link failure by rerouting the OD pair flow at departure through path restoration. There will be some delay between the event of a failure and the reaction of the controller in response to the failure. Meanwhile, the traffic that were assigned through the failed links will not have a chance to avoid the failure unless we implement backlog

rerouting.

Figure 5.17 demonstrates span routing that assigns the backlog into paths that bypass the failed link. The traffic units backlogged in L_9 reset the origin to N_9 so that it will be routed along with the traffic originating at N_9.

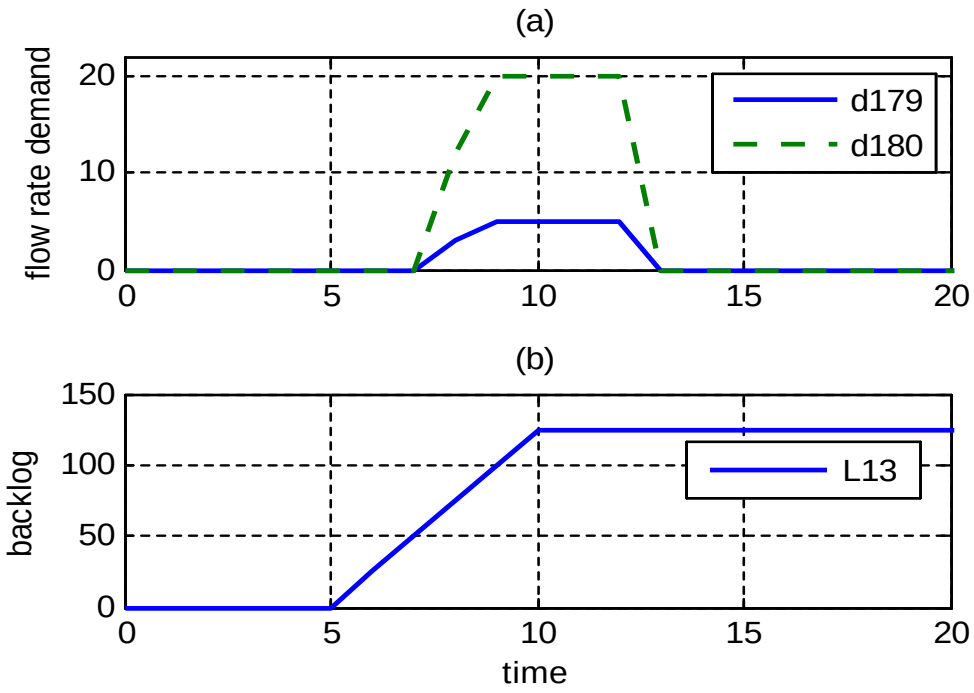

Figure 5.17. Flow demand and backlog with path and backlog re-routing.

The new flow demands, d_{179} from N_9 to N_{20}, and d_{180} from N_9 to N_{21} are shown in Figure 5.17(a). They represent the traffic that would have been backlogged in L_9 and L_5. The span restoration supplements the path flow restoration in diverting the traffic and clearing the backlog in the network.

The introduction of flow rerouting into the network has restored the interrupted OD pair flows in the presence of the failure. This feature contributes to flow survivability. The backlog accumulated in link L_{13}, Figure 5.17(b), will clear when the failed link is recovered.

5.3.3 Failed link recovery. The capacity collapse in link L_{13} has compromised the OD pair flows through paths P_{40} and P_{43}. The LP-based flow controller has restored the flow through alternative paths that carry out the tasks of P_{40} and P_{43}. If the controller were not in place, the blocked traffic would have waited until the failed link is rescued. Suppose that the capacity of L_{13} has fully recovered starting from $t = 80$. The backlog accumulated in the links during the failure phase starts to dissipate following the flow restoration through L_{13} as depicted in Figure 5.18.

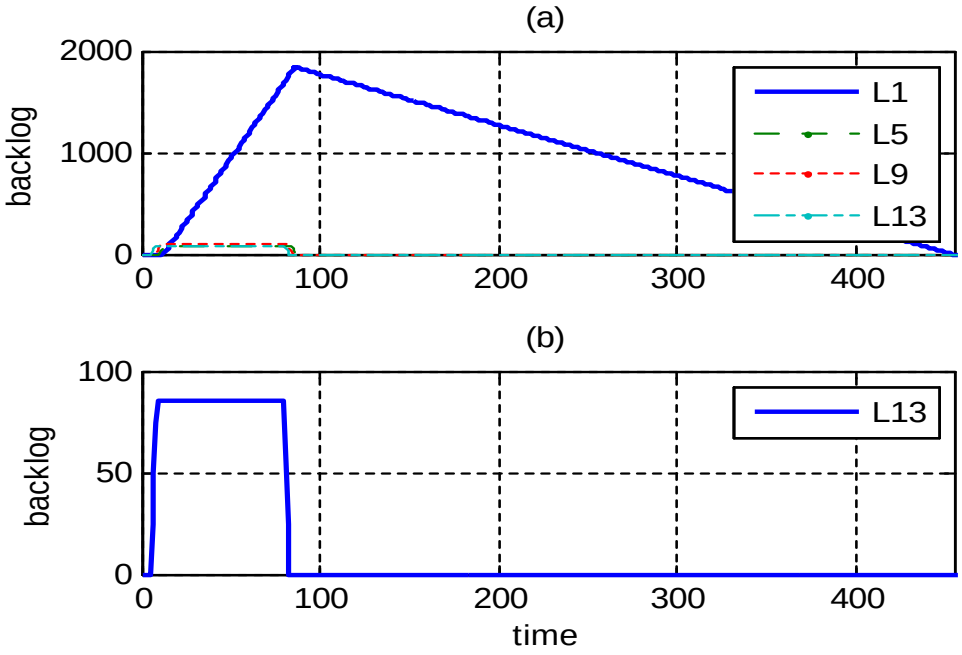

Figure 5.18. Flow restoration across failed link, without and with re-routing.

The amount of time it took to clear the backlog with no rerouting scheme is shown in Figure 5.18(a). It took 371 less time intervals for the backlog to clear out using rerouting as depicted in Figure 5.18(b). This asserts the importance of flow rerouting in flow survivability. After completing the backlog clearance, the controller restores the OD pair flows back to their initial routes.

The rerouting controller responded more quickly to the changes in the network. The network is shown to have better survivability with flow rerouting.

5.3.4 Two link failure. In addition to the 100% capacity collapse in L_{13}, a 50% capacity reduction is now introduced to L_{21} at $t = 5$. Link L_{21} lies on the alternative paths P_{41} and P_{44}. The remaining capacity in L_{21} will accommodate the flow demand d_{260} originating at N_{13}, but not the rerouted traffic.

The total backlog build up in the network due to the link failures is shown in Figure 5.19. The total backlog in the network with no rerouting, Figure 5.19(a), is observed to increase at a higher rate when compared with path rerouting, as in Figure 5.19(b).

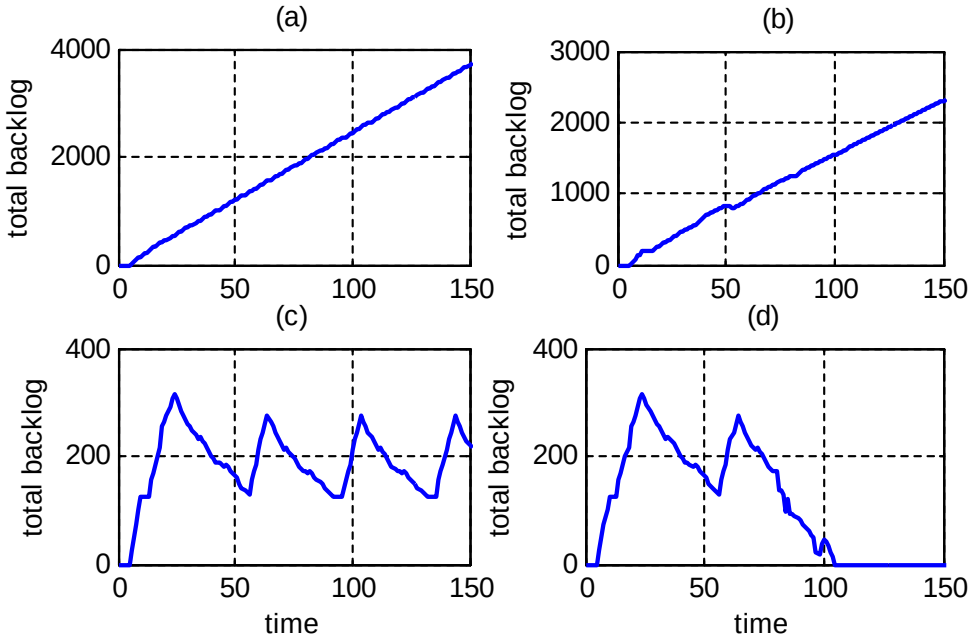

Figure 5.19. Two link failure backlog build up.

The span restoration on top of the path restoration in Figure 5.19(c) further reduces the backlog, thus keeping it within the occupancy limits of the links. The backlog is localized around N_{15} with the traffic overflowing L_{21} has been confined in L_{15}, L_{18} and L_{19} well before it spreads towards the origin of the flows. Similarly, The backlog in L_{13} is well controlled before it spills over to L_9.This ensures that the disturbed flow survives with routing control employed in the network in the presence of the failure.

The failed links are restored at $t = 80$ to their initial capacities. The backlog clears out quickly and the flow will be restored to their original routes if rerouting is in place as shown in Figure 5.19(d). The network controller using no rerouting would have taken 17 times as much time intervals for the compromised OD pair flows to fully recover.

CHAPTER 6

Conclusions And Recommendations

6.1 Conclusions

We assumed that a change in the capacity of a link travels slower than the traffic. The links of a network were discretized into a number of sections so that the capacity collapse in a given link travels a section of the link in one time interval. The nodes of the network are categorized as either transit, merging, diverging, or complex depending on the number of links coming in and going out of the node. Depending on the type of the node, specific capacity collapse propagation models at intersections were proposed. Different models were assumed for merging and diverging node types and evaluated using numerical simulation. Complex node types with more than one incoming and more than one outgoing links were treated using merging-diverging models combination. The propagation of capacity collapse for a range of capacity failures was discussed.

The network flow was formulated as path-based multicommodity flow problem with the objective of minimizing the total cost of travelling. The LP-based optimization controlled the flow assignment-decisions into k first-link-disjoint alternative paths. The aggregate prices of the paths were factored into the decision making process. Recovery of disrupted OD pair flows using path and span restorations was addressed.

The following observations and conclusions were made:

- A difference in propagation times during failure and restoration was observed. The time for the upstream sections of a link to be affected by the fault in the downstream section of the link was more than the time for the upstream sections to feel the effect of capacity restoration.
- The comparison of the merging models M1 - equal sharing, M2 - Random proportions and

M3 - Based on Priority revealed that

  - Models M1 and M2 were affected by a failure in the network more quickly than M3. They also took longer to respond to the capacity restoration.
  - Model M3 was shown to respond very quickly to the capacity restoration. It was superior to M1 and M2 that it improved the capacity utilization in the network.

- The different diverging models produced similar results for the simulation setup we considered.
- The collapse wave was shown to increase faster with increase in magnitude of the fault. The restoration wave was affirmed to be much faster than the collapse wave. For the scenario of a 50% capacity failure, the restoration wave was 4.5 times faster than the collapse wave.
- Multiple numerical simulations affirmed that the proposed controller with rerouting efficiently rerouted the compromised OD pair flows satisfying flow-rate demand requirements in addition to link capacity constraints. The proposed controller was also shown to be applicable in flow network survivability in the presence of failures.
- The steady-state performance in the recovery phase was attained quickly with short intervals of the recovery time.
- The network was shown to survive flow interruptions better with flow restoration schemes in place. Otherwise, the traffic had been delayed until the failed link was recovered and the accumulated backlog was cleared.
- Span restoration on top of path restoration further reduced the backlog in the network and improved network capacity utilization. The rerouting techniques averted the buildup of backlogs in many links. The failure was localized to the links whose flow could not be

rerouted. The stacked traffic had to wait until the failure was removed.

- The time for the backlog to clear out using rerouting was much smaller than the case with no rerouting. This asserted the importance of rerouting in flow survivability. The controller restored the OD pair flows back to their initial routes once the backlogged traffic was cleared.

6.2 Recommendations

There is a huge number of possible combination of options in the proposed control algorithm and detailed numerical and experiment comparisons are needed. We believe the outcome of this thesis will encourage research in adaptive real-time rerouting in transportation networks. We recommend that the proposed controller and congestion propagation models shall be further tested using existing networks and real-time data.

References

[1] R. Ahuja, T. Magnanti and J. Orlin, Network Flows: Theory, Algorithms, and Applications, Prentice Hall, edition 1, February 1993.

[2] D. Medhi, and K. Ramasamy, Network Routing : Algorithms, Protocols, and Architectures, Morgan Kaufmann Publishers, March 2007.

[3] B. Sanso and F. Soumis, "Communication and transportation network reliability using routing models", IEEE Transactions on Reliability, vol. 40, no. 1, April 1991, pp. 29-38.

[4] A. Girard and B. Sanso, "Multicommodity flow models, failure propagation, and reliable loss network design", IEEE/ACM Transactions on Networking, vol. 6, no. 1, February 1998, pp. 82-93.

[5] Y-S Myung and H. Kim, "A cutting plane algorithm for computing k-edge survivability of a network", European Journal of Operational Research, vol. 156, August 2004, pp. 579–589.

[6] T. Matisziw, A. Murray, "Modeling s-t path availability to support disaster vulnerability assessment of network infrastructure", Computers & Operations Research, Vol. 36, January 2009, pp. 16-26.

[7] R. Ellison, D. Fisher, R. Linger, H. Lipson, T. Longstaff and N. Mead, ”Survivable Network Systems: An Emerging Discipline”, Carnegie Mellon/Software Engineering Institute, Technical Report No. CMU/SEI-97-TR-013,1997.

[8] P. Heegaard and K. Trivedi, "Network survivability modeling", Computer Networks, Vol. 53, June 2009, pp. 1215-1234.

[9] A. Murray, T. Matisziw and T. Grubesic "Critical network infrastructure analysis: interdiction and system flow", Journal of Geographical Systems, vol. 9, no. 2, 2007, pp. 103–117.

[10] B. Sanso, L. Milot, "Performability of a Congested Urban Transportation Network When Accident Information is Available", Transportation Science, vol. 33, no. 1, 1999, pp. 68-79.

[11] D. Tipper, J. Hammond, S. Sharma, A. Khetan, K. Balakrishnan and S. Menon, "An analysis of the congestion effects of link failures in wide area networks", IEEE Journal on Selected Areas in Communications, vol.12, no.1, Jan 1994, pp.179-192.

[12] F. Kelly, "Charging and rate control for elastic traffic", European Transactions on Telecommunications, vol. 8, 1997, pp. 33–37.

[13] F. Kelly, A. Maulloo and D. Tan, "Rate Control for Communication Networks: Shadow Prices, Proportional Fairness and Stability", The Journal of the Operational Research Society, vol. 49, no. 3, March 1998, pp. 237-252.

[14] S. Low and D. Lapsley, "Optimization flow control - I: Basic algorithm and convergence", IEEE/ACM Transactions on Networking, vol. 7, no 6, Dec 1999.

[15] S. Athuraliya, S. Low, V. Li and Q. Yin, "REM: active queue management", IEEE Network , vol.15, no.3, May 2001, pp.48-53.

[16] F. Paganini, "A global stability result in network flow control", Systems & Control Letters, vol 46, no. 5, July 2002, pp. 165-172.

[17] J. Wen, and M. Arcak, "A Unifying Passivity Framework for Network Flow Control", IEEE Transactions on Automatic Control, vol. 49, no 2, Feb. 2004, pp. 162-174.

[18] L. Ruogu, E. Atilla, Y. Lei and B. Ness, "A unified approach to optimizing performance in networks serving heterogeneous flows", IEEE/ACM Transactions, vol. 19, no. 1, Feb 2011, pp. 223-236.

[19] F. Daganzo, "The cell transmission model: A dynamic representation of highway traffic consistent

with the hydrodynamic theory", Transp. Res. Part B, Methodological, vol. 28, no. 1, August 1994, pp. 269-287.

[20] F. Daganzo, "The cell transmission model, part II: Network traffic", Trans. Res. Part B, Methodological, vol. 29, no. 2, April 1995, pp. 79-93.

[21] A. Ziliaskopoulos, "A Linear Programming Model for the Single Destination System Optimum Dynamic Traffic Assignment Problem", Trans. Sci., vol. 34, February 2000, pp. 37-49.

[22] D. Dunn, W. Grover and M. MacGregor, "Comparison of k-shortest paths and maximum flow routing for network facility restoration", IEEE Journal on Selected Areas in Communications, vol. 12, no. 1, January 1994, pp. 88-99.

[23] W. Gebremariam and M. Bikdash, "LP-based flow-rate control and modeling of capacity collapse propagation over long links", Southeastcon 2012 Proceedings of IEEE, March 2012, pp.1.

[24] R. Iraschko, M. MacGregor; W. Grover, "Optimal capacity placement for path restoration in STM or ATM mesh-survivable networks", IEEE/ACM Transactions on Networking, vol.6, no.3, June 1998, pp.325-336.

[25] Y. Liu, K. and Trivedi, "Survivability quantification: the analytical modeling approach", International Journal of Performability Engineering, vol. 2, no. 1, January 2006, pp. 29–44.