Operational decision making for Internet service provision

Mark C. Stewart
Department of Management
University of Canterbury
New Zealand
m.stewart@mang.canterbury.ac.nz

Abstract

Communications services, such as those provided by telecommunications and the Internet, have become an important part in society. The significance of these services means that it is important to investigate how best to meet the ever increasing level of demand. Much research has been conducted on determining appropriate routing systems. However, less attention has been given to determining the best processing capacity to meet the demand, or how to best use the processing capacity available. The problem situation is described for the latter operational decisions in the Internet setting. This includes factors such as processing and routing congestion, demand, and node and arc reliability. A mixed-integer program is presented for the lower level operational processing decisions. An analysis into the four general types of solutions produced by the model is discussed.

1 Introduction

Communication services are an important component in today's electronic world. Telecommunications services, from simple dial up telephony to recent inventions like videoconferencing, are a necessity of society. Internet services, such as search engines and Internet banking, are also rapidly gaining popularity as the growth of the Internet soars [1].

The area of telecommunications has been significantly investigated by researchers in Management Science/Operations Research. One of the main areas of focus for this research has been the routing and transmission of flow over the telecommunications networks [4] [5]. In recent times, however, some attention has also been given to the processing of the information. The main reason for this is because many new services are now available, including multimedia applications such as videoconferencing, and these new services require more information processing than existing services. Therefore, it has become important to examine the networks from a processing capacity point of view. It is unlikely, however, that transportation bottlenecks will be avoided, as many of the new services also require transportation of a large amount of information.

In communication industries there is a move towards a distributed processing environment. In this environment services are platform independent and may be provided by any of the network's computing nodes. To utilise the inherent flexibility of a distributed processing environment, services are made up of smaller entities, called

applications or subservices. The subservices required to provide a particular service can be run independently at different nodes. For a service to be available, the subservices that make up that service must all be available within the network.

An important set of decisions in this environment is how to make best use of the computational resources in the network when meeting demand. As outlined in [3], these processing decisions made are 1) which nodes to run the subservices on, and 2) how to meet demand once the subservices have been installed. Demand for a subservice can only be met at a node if the subservice is installed at that node. In these models there is a single constraining computational resource (processing capacity) that can be used in two ways: as a fixed requirement to make a subservice available on a node, and to meet demand at a node.

There has been some research into analysing, modelling and solving variations of the telecommunications processing decisions [3] [6] [7]. Much of this research has not included transportation restrictions.

Telecommunications and the Internet both facilitate communication and the transfer of information. There is currently a convergence of these two types of communication at the level of technology and the services that they can provide. Therefore, models developed for the telecommunications processing decisions may, with a little effort, be tailored to the Internet. The problem addressed in this paper, following on from [3], is to formulate and analyse a model for the processing decisions in the Internet setting. In this setting it is very important to include transportation factors, such as congestion, in the model.

The remainder of this paper is organised as follows. In Section 2 the problem situation for the Internet processing decisions is described. Section 3 presents a formulation for the Internet processing decisions model. Section 4 discusses the analysis of the mathematical model, and in particular the different solutions the model can produce [3]. Section 5 concludes this paper and discusses some future research opportunities in the area.

2 Problem situation

The situation considered is how to best use the processing resources available to meet demand for Internet services. An influence diagram is presented (Figure 1) which shows the most important factors in the problem situation and how these factors affect each other. For clarity, only the most important factors and interactions are shown in this paper.

The total costs and total revenue depicted in the influence diagram represent the important costs incurred and revenue obtained from meeting demand, respectively. The total costs consist of processing costs and delivery costs, and costs associated with any demand that is rejected. Revenue is obtained from meeting demand.

Demand can be rejected for two reasons: insufficient system resources, or because meeting the demand is not profitable. Demand will be met while the marginal profit associated with meeting an extra unit of demand is more than the rejection costs incurred if it was not met. The unit rejection cost refers to the loss to the decision maker if they fail to meet a unit of demand. This cost should incorporate the potential loss of goodwill which could lead to a loss of future demand and future profits. The rejection cost incorporates a number of complex relationships but these will not be discussed in this paper, and hence the unit rejection cost is assumed to be constant.

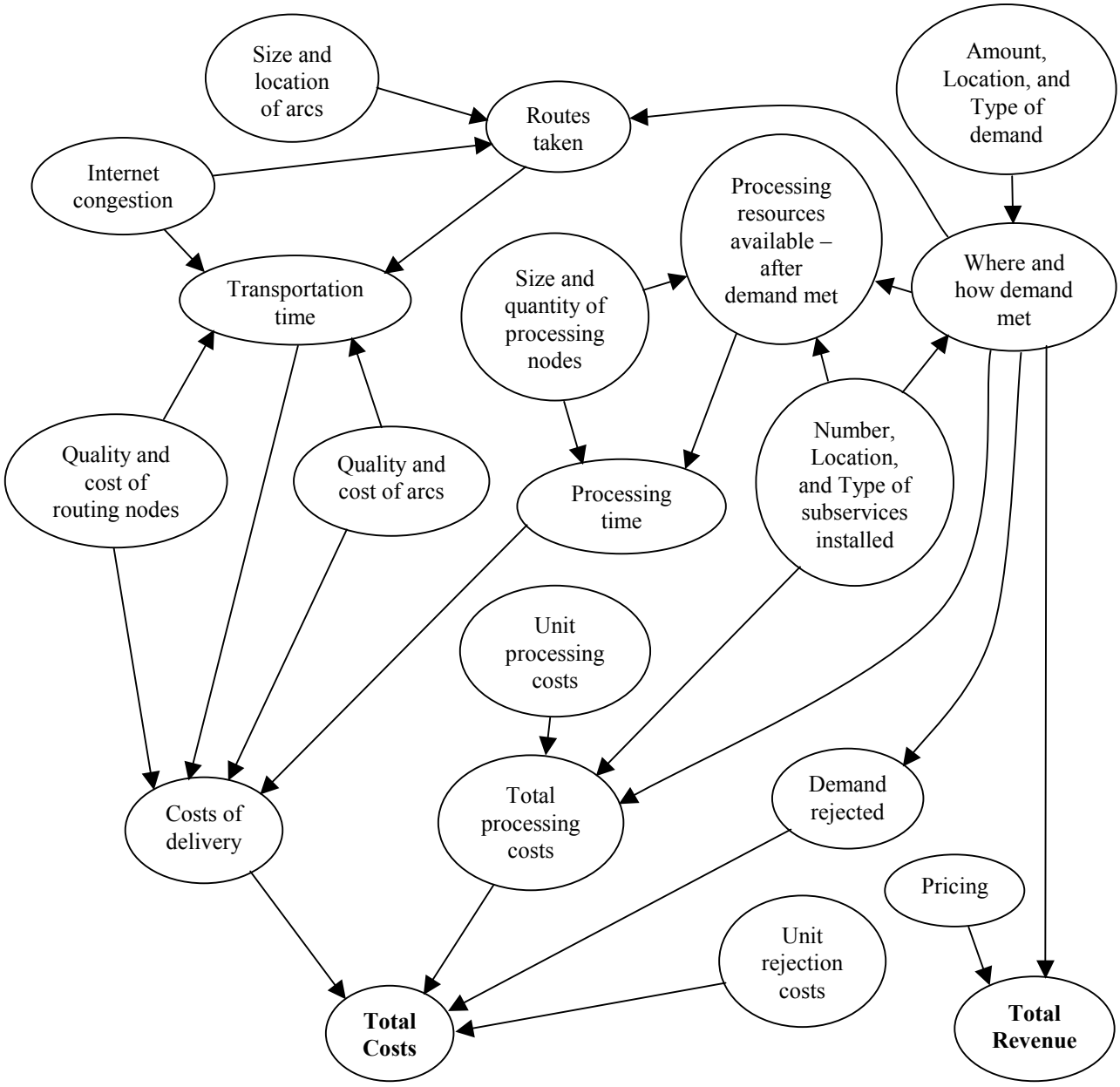

Figure 1. Influence diagram for the Internet processing decisions problem situation.

The number of nodes and the size of each node determines the initial processing capacity available, the maximum capacity. Processing capacity is used by either installing subservices or meeting demand. The processing capacity available and the processing capacity used give the final processing resources available. This provides an implicit feasibility test in the influence diagram. Also, the processing capacity used incurs direct costs. The unit processing cost allows the possibility of paying for processing capacity to be included, for example, if processing was done on another service providers node. Therefore, it is important to distinguish between the processing capacity used at different locations.

An Internet service request involves routing flow over arcs and through routing nodes. It is assumed that the decision maker has full knowledge of the route the flow is sent on. The first type of delivery costs included are direct costs of routing over arcs and routing through nodes. Including these costs allows different arcs and nodes to have different owners, who all have their own pricing structures.

As with rejected demand, long transportation times can cause a loss of goodwill. For example, if transportation times were high then customers could become frustrated at not receiving their service promptly. This could harm their possible future requests as well. The transportation times associated with meeting demand are influenced by a number of factors. The quality of the arcs and routing nodes are an important factor. For example, good quality routing nodes route information faster than nodes of a lesser quality (see [2] for an explanation of how flow is routed). Another factor is Internet congestion. If the level of flow through a node exceeds the routing nodes capacity the excess flow joins a queue, and is hence delayed. Therefore, the transportation time per unit through a node increases as the amount of flow through the node increases. Figure 2 shows an example of the total transportation time through a node as flow through the node increases. Note that as the total level of Internet traffic cannot be known with certainty the transportation times also can not be known with certainty. As they are a key factor in the delivery costs, and hence the decision making process, it may be important, in later models, to include this uncertainty explicitly.

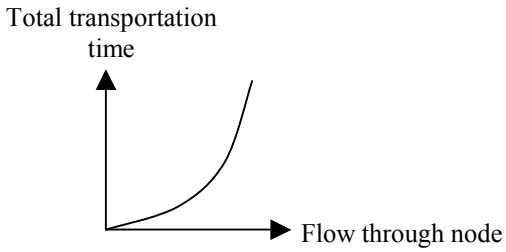

Figure 2. Function for transportation times through nodes.

Long processing times also cause costs associated with a loss of goodwill. Similarly to transportation times, processing times are influenced by the quality of the processing node and also by the level of processing at the node. The more processing capacity being used the longer the processing time per unit. This is because the processing resources have to be shared, and cannot be devoted entirely to meeting one lot of demand.

Therefore, in summary, the delivery costs function includes the following factors:

- Arc costs
- Route through node costs
- Delivery time costs:
  - Processing times (= f (subservices installed, demand met))
  - Transportation times:
    - Time through nodes (= f (flow through node))
    - Time over arcs

The final part of the influence diagram are the factors set by the decision maker. The decisions made are where to install subservices and meet demand, and the routes taken. The first two decisions define the processing capacity used and therefore affect the available processing resources. These decisions are constrained by the amount, location, and type of demand. Demand is influenced by many factors, such as price and convenience. However, for reasons of simplicity these factors are left out. The routing

requirements are determined by where and how demand is met. Also, arc capacities restrict the choice of routes available, and congestion will influence the routes taken, as routing through highly congested areas will incur significant costs.

An important point to note for the next section is that there are actually only two constraining factors – the node capacity and the arc capacity – and the rest of the problem situation determines costs and revenues.

Note that the major difference between the Internet problem situation described above and the telecommunications problem situation [6] is that the Internet model includes transportation and delivery factors in the decision making.

3 Formulation

The Internet processing decisions model has two main purposes. They are:

- To provide the operational decision makers with solutions to the shorter term problem of how to provide services. When making these decisions the resources available have been set, the demand is known with more certainty, and hence the model will be used to make the best decisions possible for meeting demand.
- To help the strategic decision makers in determining solutions to the longer term problem of what resources to provide. The shorter term processing decisions model would be used as feedback for the longer term capacity investment decisions, in order to show how the resources were actually used and how the profit was affected. For example, if the decision maker wanted to increase the capacity of a certain node the processing decisions model could be used to evaluate how different the solution would be given these changes to the network, and to what extent the profit would be affected.

The mathematical model formulated for the Internet processing decisions situation is discussed and presented below. Constraints referred to in the discussion relate to the formulation at the end of this section. The formulation uses the following indices:

a – arc.
dr – node where the demand is received.
k – subservice.
n – node where the demand is processed.
s – service.

Constraint (2):

$$\overline{X}_{s,dr} + R_{s,dr} = D_{s,dr} \qquad \forall\, s,dr,$$

Demand for service *s* received at node *dr* ($D_{s,dr}$) is either met ($\overline{X}_{s,dr}$) or rejected ($R_{s,dr}$).

Constraint (3):

$$\sum_{n} x_{k,dr,n} = \sum_{I_k} e_{k,s}\, \overline{X}_{s,dr} \qquad \forall\, k,dr,$$

Let $e_{k,s}$ represent the processing units of subservice *k* used by service *s*, and I_k the set of all services that use subservice *k*. The right hand side is the total processing requirement for subservice *k* and the left hand side is the total processing for subservice *k* performed in the network. Here, $x_{k,dr,n}$ is the demand for subservice *k* received at node *dr* that is processed at node *n*.

Constraints (4) & (5):

$$\sum_{k}\sum_{dr} x_{k,dr,n} + \sum_{k} z_{k,n}\, u_k \;\le\; S_n \qquad \forall\, n,$$

$$x_{k,dr,n} \;\le\; z_{k,n}\, M \qquad \forall\, k,dr,n,$$

Let $z_{k,n}$ be a binary decision variable with a value of 1 if subservice k is installed at node n and 0 otherwise. Constraint (5) ensures that demand for a subservice can only be met at a node if the necessary subservice is installed there (M is a large number). Constraint (4) ensures that the total processing capacity used at a node does not exceed the node capacity available at that node (S_n), where processing capacity can be used by meeting demand or installing the necessary subservices, respectively. The processing capacity used to have subservice k available is u_k.

Constraints (6), (7), & (8):

$$arc\_usage_a \;\le\; C_a \qquad \forall\, a,$$

$$arc\_usage_a \;=\; f_{au}(x_{k,dr,n}) \qquad \forall\, a,$$

$$delivery\_costs \;=\; f_d(.....)$$

Constraint (6) ensures that the level of flow on an arc (arc_usage_a) does not exceed the capacity of that arc (C_a). For a particular set of nodes between which flow must be sent the cheapest route is chosen, and this in turn defines the arc usage (Constraint (7)).

For simplicity, in this formulation of the Internet processing decisions model the complicated delivery costs function (discussed in the previous section) is not presented explicitly (Constraint (8)).

Objective function (1):

$$\text{Max} \quad \sum_{s}\sum_{dr} (\overline{X}_{s,dr}\, p_{s,dr} - R_{s,dr}\, h_{s,dr})$$

$$- \Big(\sum_{k}\sum_{dr}\sum_{n} b_n\, x_{k,dr,n} + \sum_{k}\sum_{n} u_k\, b_n\, z_{k,n}\Big)$$

$$- \; delivery\_costs$$

The five terms in the objective function calculate profit. The first term is the revenue obtained from meeting demand ($p_{s,dr}$ is the price obtained for service s received at node dr). The second term is the cost associated with rejecting demand ($h_{s,dr}$ is the cost of rejecting service s received at node dr). The third and fourth terms are the costs of using processing capacity (b_n is the cost of using processing capacity at node n). Finally, the fifth term represents the delivery costs associated with the chosen method of meeting demand.

The entire formulation is as follows. Note that the decision variables representing the subservices being installed are binary, while the rest of the decision variables are non-negative.

Max

$$\sum_{s}\sum_{dr}(\overline{X}_{s,dr}\ p_{s,dr} - R_{s,dr}\ h_{s,dr})$$

$$- (\sum_{k}\sum_{dr}\sum_{n} b_n\ x_{k,dr,n} + \sum_{k}\sum_{n} u_k\ b_n\ z_{k,n})$$

$$- \ delivery\_costs \qquad (1)$$

s.t. Constraints

$$\overline{X}_{s,dr} + R_{s,dr} = D_{s,dr} \qquad \forall\ s,dr, \qquad (2)$$

$$\sum_{n} x_{k,dr,n} = \sum_{I_k} e_{k,s}\ \overline{X}_{s,dr} \qquad \forall\ k,dr, \qquad (3)$$

$$\sum_{k}\sum_{dr} x_{k,dr,n} + \sum_{k} z_{k,n}\ u_k \le S_n \qquad \forall\ n, \qquad (4)$$

$$x_{k,dr,n} \le z_{k,n}\ M \qquad \forall\ k,dr,n, \qquad (5)$$

$$arc\_usage_a \le C_a \qquad \forall\ a, \qquad (6)$$

$$arc\_usage_a = f_{au}(x_{k,dr,n}) \qquad \forall\ a, \qquad (7)$$

$$delivery\_costs = f_d(.....) \qquad (8)$$

$$z_{k,n}\ binary,\ x_{k,dr,n},\ \overline{X}_{s,dr},\ R_{s,dr} \ge 0.$$

4 Analysis of Model

This section discusses one important part of the analysis, that being the types of solutions the Internet processing decisions model can produce. It was found that solutions to this model could be grouped into four general solution types. These solution types will now be discussed. To aid in this explanation a simple example of the four solution types, where three processing nodes are being used to meet demand received at those three nodes, is shown in Figure 3.

'Distributed subservices' solution

		Demand processed		
		1	2	3
Demand received	1	40%	40%	20%
	2	30%	30%	40%
	3	30%	20%	50%

'Cheapest node first' solution

		Demand processed		
		1	2	3
Demand received	1	100%	0	0
	2	100%	0	0
	3	100%	0	0

'No transportation' solution

		Demand processed		
		1	2	3
Demand received	1	100%	0	0
	2	0	100%	0
	3	0	0	100%

'Middle ground' solution

		Demand processed		
		1	2	3
Demand received	1	40%	30%	30%
	2	0	100%	0
	3	0	0	100%

Figure 3. Fraction of demand received met at node *n*, for each solution type.

- A ‘distributed subservices’ solution has the installation of the subservices and the meeting of demand shared, or distributed, among all the nodes. This type of solution leads to an even usage of processing capacity at all nodes in the network. The important point about this type of solution is that, from the perspective of the node where demand was received, the cheapest node for meeting a unit of demand changes as the amount of demand already met changes. This results in the meeting of demand for that node being split between many processing nodes in the network.

  This type of solution occurs when the processing time function increases relatively steeply, meaning high congestion occurs as the level of processing performed increases. In these circumstances spreading processing around nodes incurs the least processing time costs.

- For a ‘cheapest node first’ solution all demand received at a node is met at the node which is cheapest initially, taking processing and transportation costs into consideration. The important point is that all demand received at a node is met at the one node only, there is no distributing processing between nodes.

  This type of solution occurs when the processing time and transportation time functions are close to linear. This is because the processing time cost per unit does not change much as more and more demand is met, meaning congestion does not significantly influence the costs.

- A ‘no transportation’ solution has all the demand met at the node where it was received, which clearly leads to no transportation. This means that all subservices have to be installed at all nodes.

  This type of solution occurs when the costs associated with transporting flow between nodes are relatively very high. In this case it is worth incurring more processing costs so that the high transportation costs are not incurred. Also, because of the large amount of processing capacity used to install subservices, this type of solution only occurs when the processing time costs per unit do not increase significantly as more processing capacity is used.

- A ‘middle ground’ solution is essentially a combination of the other three ‘extreme’ solutions. It has distribution of processing, but some nodes will have all their demand met at only one node. Also, transportation may occur, but some arcs or nodes may be avoided in the routing.

  This type of solution occurs when circumstances are such that the extreme solutions did not quite occur. For example, a ‘middle ground’ solution would be produced if some areas of the network had high congestion while other areas did not.

The remainder of this section shows mathematically how the Internet model can produce the four solution types. The network and assumptions necessary for this are shown in Figure 4 and below.

Assumptions:

(1) One service is available and can be demanded from node 1 only. This service requires only one subservice.

(2) Installing a subservice does not use any processing capacity, i.e., $u_1 = 0$.

(3) It costs more to process at node 1 than it does at node 2, i.e., $b_2 < b_1$.

(4) It is profitable to meet demand only if up to m units of processing capacity at a node are used meeting that demand, i.e., $p - b_n - d_1 > -h$, but $p - b_n - d_2 < -h$.

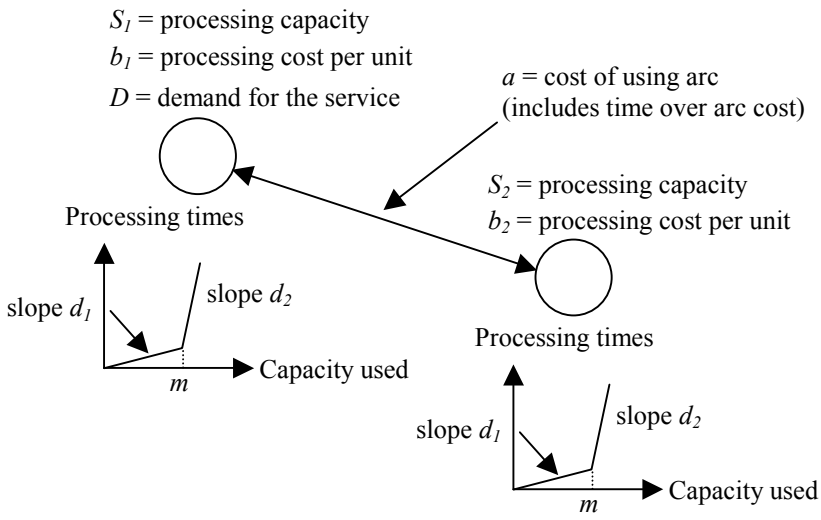

Figure 4. Internet network.

The formulation for the Internet model, where $x_{n,j}$ is the demand met at node n where the processing time per unit is d_j, is:

Model:

$$\text{Maximise } (p - b_1 - d_1)\, x_{11} + (p - b_1 - d_2)\, x_{12} + (p - b_2 - d_1 - a)\, x_{21} + (p - b_2 - d_2 - a)\, x_{22} - h\, r$$

$$\text{s.t.} \quad x_{11} + x_{12} + x_{21} + x_{22} + r = D$$

$$x_{n,1} + x_{n,2} \le S_n \qquad n = 1,2$$

Optimal solution:

$$x_{11}^* = \min\{m, D - m\},\ x_{12}^* = 0,\ x_{21}^* = m,\ x_{22}^* = 0,\ r^* = D - x_{11}^* - x_{21}^*.$$

Therefore, given the respective values of a, b_n, d_j, h, m, and D, the solution given from the Internet model can be found from Figure 5.

	$b_1 > a + b_2$			$a > p - b_2 - d_1 + h$
	$D < m$	$m < D < 2m$	$D > 2m$	
Solution	$x_{11}^* = 0$, $x_{12}^* = 0$, $x_{21}^* = D$, $x_{22}^* = 0$, $r^* = 0$.	$x_{11}^* = D - m$, $x_{12}^* = 0$, $x_{21}^* = m$, $x_{22}^* = 0$, $r^* = 0$.	$x_{11}^* = m$, $x_{12}^* = 0$, $x_{21}^* = m$, $x_{22}^* = 0$, $r^* = D - 2m$.	$x_{11}^* = \min\{m, D\}$, $x_{12}^* = 0$, $x_{21}^* = 0$, $x_{22}^* = 0$, $r^* = \max\{0, D - m\}$.
Solution type	‘cheapest node first’	‘middle ground’	‘distributed subservices’	‘no transportation’

Figure 5. Solutions for the Internet model.

Therefore, it has been shown that the Internet model can produce a solution of any type, where the type of solution actually produced depends on the relative sizes of the parameters. Upon analysis of the model it was found that the delivery costs, and in particular the processing time costs and to a lessor extent the transportation time costs, led to the possible distributed processing solutions found from the Internet model. Note that the first three assumptions could be relaxed and the same solution process would occur. The only difference is that the model would be more complicated, and hence the insight as to exactly how demand would be met would be more difficult to determine.

5 Conclusion and Future work

This paper presents research done on how to make best use of the available computational resources when meeting demand for Internet services. These decisions are especially important given the rapid growth of the Internet in recent times. The research presented in this paper extends work done in a distributed processing environment for the telecommunications processing decisions [3]. The reason this has been possible is because of the current convergence of telecommunications and the Internet at both the level of technology and the services that they can provide.

An influence diagram is presented to help discuss the Internet processing decisions problem situation, and a mathematical model, useful for both operational and strategic decision makers, is developed. This model includes transportation factors, such as congestion and network reliability, in more detail than the previous telecommunications research [3]. Finally, an analysis of the four general solution types for the Internet processing decisions is discussed, and is shown mathematically.

There are a number of possible extensions to the work presented in this paper. These include extending the 1-stage formulation to a 2-stage formulation, where the subservices are installed in the first stage, and demand is met and routed in the second stage. Alternatively, a stochastic version of the Internet processing decisions model could be investigated, meaning the stochastic parameters, such as demand and congestion, would be modelled differently. These possible extensions would enable the model to more closely represent reality. This proposed research should follow from the explanation of the problem situation and development of the 1-stage deterministic model which are presented in this paper.

Acknowledgements

I would like to acknowledge the assistance of Dr Shane Dye for his invaluable input into this paper.

References

[1] Chapman, A., Kung, H.T., *Enhancing Transport Networks with Internet Protocols*, IEEE Communications Magazine, 36 (1998), pp 100-104.

[2] Dowd, K., *The best routes for Net links*, InformationWeek, n617 (1997), pp 53-59.

[3] Dye, S., Tomasgard, A., Wallace, S.W., *New Aspects of Service Provision and Technology Strategies in Telecommunications*, Telektronikk, 3/4.98 (1998), pp 67-73.

[4] Kelly, F. P., *Routing and Capacity Allocation in Networks with Trunk Reservation*, Mathematics of Operations Research, 15n4 (1990), pp 771-793.

[5] Neuman, I. , *Class Dependent Routing in Backbone Computer Networks*, INFOR, 28n3 (1990), pp 247-265.

[6] Tomasgard, A., *Aspects of service provision and distributed processing in a telecommunication network*, Dr. ing. Thesis, Norwegian University of Science and Technology (1998).

[7] Tomasgard, A., Dye, S., Wallace, S.W., Audestad, J.A., van der Vlerk, M.H., *Stochastic Optimization Models for Distributed Communications Networks*, Working Paper #3/97 (1997).