SIP Routing Methodologies in 3GPP

Alexander A. Kist and Richard J. Harris
RMIT University, BOX 2476V, Victoria 3001, Australia
Phone: (+) 61 (3) 9925-5230, Fax: (+) 61 (3) 9925-3748
{kist,richard}@catt.rmit.edu.au

July 4, 2003

Abstract

This paper discusses methodologies for efficient Session Initiation Protocol (SIP) message routing on the application layer in the IP Multimedia Subsystem (IMS) of 3rd Generation Partnership Project (3GPP) UMTS networks. The contributions are twofold: Firstly, it introduces a generic multidimensional optimisation metric that can be used in conjunction with existing routing protocols. The metric's sensitivity depends on the network operating conditions. Secondly, it outlines the need for SIP message routing and introduces concepts that allow effective message routing on the SIP layer. It uses the defined optimisation metric which is, in this case, sensitive to delay, reliability and availability of resources. The introduced methodologies are also applicable for routing problems in generic overlay networks.

1 Introduction

The Session Initiation Protocol (SIP) is an IETF protocol that performs user location, session setup and session management. SIP is defined in RFC 3261 [1] that renders RFC 2543 obsolete. The 3rd Generation Partnership Project (3GPP) [2] is a global initiative to develop standards and specifications for next generation *Universal Mobile Telecommunications System* (UMTS) networks. 3GPP has decided to use SIP as the signalling protocol for the IP Multimedia Subsystem (IMS). 3GPP introduces a number of SIP proxy servers called *Call Session Control Function* (CSCF). Commercial service providers need these servers to control session signalling message flows and enable authentication, billing and service provisioning. 3GPP Technical Specification 23.228 [3] (R5) explains these functions in more detail. Logically, SIP nodes are located on the application layer. If these nodes are connected by virtual SIP Links (VSLs) [4], the elements form a *Virtual SIP Overlay Network* (VSON). Messages traversing the VSON can take alternative routes to their destination. Most SIP message routing decisions in 3GPP IMS have to be done during the registration of users. This is required since intermediate nodes record registration state and/or session state information. Once the associations are formed, these nodes have to be traversed for subsequent requests.

Existing routing protocols and methodologies that are used for IP packet routing are well-understood and remain a continuing focus of the research community. Shortest path routing algorithms find paths between an *Origin-Destination Node Pair* (OD-pair) that satisfy a minimal optimisation metric. Traditional routing protocols such as *Open Shortest Path First* (OSPF) [5] use scalar metrics to optimise the paths. Many possible alternative metrics are

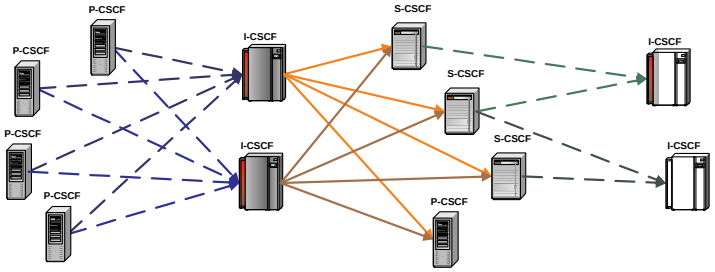

Figure 1: Operator Domain

known, but the inverse of the capacity is most commonly used in IP networks[1]. Factors that have an impact on the message routing on the VSL layer are the hierarchical and logical setup defined by the network structure, the availability of resources in nodes to handle additional requests, resource availability for additional signalling traffic on the transmission media, the close proximity of nodes and the quality on the connection. The routing process should consider these factors.

Current routing models use different schemes to account for multiple metrics. In the first scheme, every metric is multiplied by a factor and then summed into one composite metric. The Interior Gateway Routing Protocol (IGRP) and the more common Enhanced IGRP (EIGRP) [6] both use such a metric. Another approach is to find a minimum cost path where all links have two or more metrics assigned. These problems are known to be intractable (eg. Guérin and Orda [7]).

This paper proposes a multidimensional metric. During defined stages in the network operation one metric within the multidimensional metric will dominate the others. If, for example, two paths have the same number of available users and similar delays of acceptable length, the path with the maximum reliability should be selected. On the other hand, core network connections are likely to have similar reliabilities. In this case, the path selection is mainly based on delay which, in this context, is a measure for distance. If the number of available users on a certain path gets smaller, routing has to take this into account. It is important to note that in the combined metric, one metric dominates the others at a given time, the metric is not a simple weighted average calculation. The scope of this paper is the SIP message routing within one operator domain.

Figure 1 depicts an example VSON. Virtual connections between nodes and servers form the transport independent overlay SIP signalling network on the application layer. Users Equipment (UE) connects to proxy CSCFs (p-CSCFa) that hold registration-state information. During registration of users with a domain, the serving CSCF (S-CSCF) that serves the users is assigned. This process determines the message routing for subsequent messages. The interrogating CSCF (I-CSCF) forwards messages to the S-CSCF [3]. SIP message routing in 3GPP can be divided into two areas: Routing decisions have to be made for routes from I-CSCFs to S-CSCFs and decisions are required for routes from UEs/P-CSCFs/S-CSCFs to I-CSCFs, where the I-CSCF to nodes routing is a one to many routing decision, the routing for P-CSCFs to I-CSCFs is a many to few routing decision. The first require methods that reassemble load balancing behaviour schemes, the latter can use shortest path routing methodologies. These separations are defined by the specific structure of 3GPP SIP overlay

[1]Possibly due to the fact that this is a default setting for commonly used routers.

networks. A methodology for efficient shortest path SIP message routing is the focus of this paper.

The remainder of this paper is organised as follows: Section 2 defines a generic multidimensional metric, Section 3 introduces a routing scheme that can be used in VSONs, Section 4 discusses practical routing parameters and Section 5 summarises comments on the operation of the scheme. Section 6 presents a practical example. Final remarks conclude the paper.

2 Multidimensional Cost Metric

This section introduces a multidimensional metric that allows the comparison of parameters through the use of different scales. A generic notation is used to define the metric. Associations with practical parameters are explained later. Submetrics are components of the multidimensional cost metric and their number is not limited. For simplicity, the definitions are restricted to three: *Metric X* (MX), *Metric Y* (MY) and *Metric Z* (MZ). The metric's range is limited to the interval $[0, b]$. b has to be larger than one: $b > 1$.

The definitions in this section imply an importance ranking between the submetrics. The most important one is metric MX, the second important one is MY and the least important one is MZ. Importance in this context means that if two paths have equal values for metric MX, the size of metric MY matters. Also, in the case when MY reaches its maximum possible value, MY is of the same importance and same order than MX. For paths with equal MX and MY values, MZ size is relevant. As MZ values increase in size, they first reach a similar importance to MY and than MX.

This separation between the various submetrics is achieved by different scales. The *Normal Cost Interval* (NCI) is defined as the range $[0, b]$. It is the same for all metrics. The different scales are realised with orders of b and cost functions. All cost functions map NCI to the interval $[0, b^n]$, where n is the number of different submetrics. In this case, n is equal to three. The definition of the cost function uses the thresholds χ_y, χ_{z1} χ_{z2} that have to be all within the NCI: $\chi_y, \chi_{z1}, \chi_{z2} \in [0, b]$. The first cost function $f_x(m_x)$ is linear for m in the input interval $[0, b]$. It is defined in Equation (1).

$$f_x(m) = m \cdot b^2 \tag{1}$$

This "strongest" cost function is used for MX. Once MY closes in on its upper bound it has to compete with MX. To enforce this behaviour, a piecewise linear cost function is defined. Between 0 and χ_y cost is defined as described above; between χ_y and b it is "lifted" to the same level as the MX cost function f_x. Equation (2) formalises this definition.

$$f_y(m) = \begin{cases} m & if \quad 0 < m \leq \chi_y \\ \chi_y + (m - \chi_y) \cdot \frac{b^2}{b - \chi_y} & if \quad \chi_y < m \leq b \end{cases} \tag{2}$$

The third metric MZ uses a similar definition, this time with three linear segments. The first segment follows the original definition, the second segment competes with the second cost MY and the third segment competes with the first cost MX. The definition is depicted in Equation (3).

$$f_z(m) = \begin{cases} m & if \quad 0 < m \leq \chi_{z1} \\ \chi_{z1} + (m - \chi_{z1}) \cdot \frac{b}{b - \chi_{z1}} & if \quad \chi_{z1} < m \leq \chi_{z2} \\ \chi_{z1} + \chi_{z2} + (m - \chi_{z2}) \cdot \frac{b^2}{b - \chi_{z2}} & if \quad \chi_{z2} < m \leq b \end{cases} \tag{3}$$

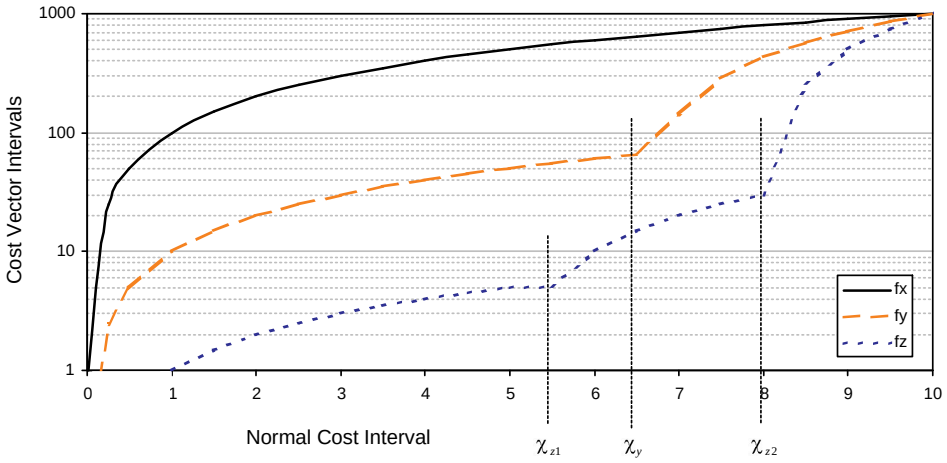

Figure 2: Cost Functions

An example graph of these three cost functions is depicted in Figure 2. The plot uses a decade basis ($b = 10$) and interval boundaries χ_{z1}, χ_{z2} and χ_y. A logarithmic scale is used to visualise the spanning over several decades. The results of the cost functions can be combined into one cost vector C which is depicted in Equation (4).

$$\mathbf{C} = \begin{pmatrix} c_x \\ c_y \\ c_z \end{pmatrix} = \begin{pmatrix} f_x(m_x) \\ f_x(m_y) \\ f_x(m_z) \end{pmatrix} \tag{4}$$

c_x, c_y and c_z are the respective cost values after they are transformed by the cost functions. These optimisation parameters use the same scales and are therefore comparable. As mentioned in the earlier, shortest path problems with multiple optimisation metrics are generally known to be intractable. To be able to use this metric with a conventional routing algorithm, a single scalar value is required. The absolute value of this vector ("Euclidean distance") can be used as an optimisation parameter (Equation (5)).

$$C = \sqrt{c_x^2 + c_y^2 + c_z^2} \tag{5}$$

Since the submetrics are scaled, the absolute value of $\mathbf{C}$, its "length", can be compared. Other common ways of calculating a single value for a composite metric are, to select the minimum, maximum or the sum of all different sub metrics. Practical parameters have to be additive and have to be normalised to the NCI to use the multidimensional cost metric. The next part of this paper introduces practical applications of these metrics and SIP messages routing in VSONs, in detail.

3 Routing in 3GPP VSONs

Four factors are of major concern for message routing: The functional requirements are given by the topology and the proposed SIP node setup. In 3GPP call flows several nodes have to be traversed during the session set up. Most of these associations are formed during registration. Details of these processes can be found in [3]. The second limiting factor is the maximum

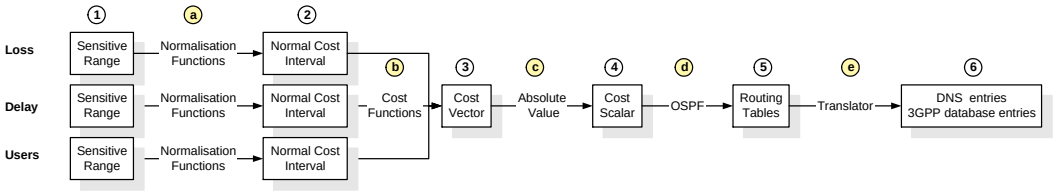

Figure 3: Routing Scheme

number of users that can be served. This applies to nodes as well as connections. Servers can only handle a limited number of user requests and virtual connections are limited by their capacity. The third parameter is the delay encountered by the messages between SIP nodes. It is desirable that the delay is minimal. The message loss probability is the last factor that is considered. Message losses are mostly due to bit errors on the transmission links and overflows in transmission queues.

Figure 3 depicts an overview of the routing scheme that is proposed for the VSON in this paper. The components are numbered and the description begins with the results and explains the steps that are necessary. Domain Name System (DNS) entries and the entries in the various 3GPP databases (6) are the final output. These are derived from the routing tables (5). SIP in the IMS uses these databases for its message forwarding. This information is generated by simple translation (e). Routing tables are the final output of the shortest path routing protocol (d), for example OSPF. Issues concerning the routing protocol are for further study. For the discussions in this paper, it is assumed that the information is propagated through the SIP overlay network and it is possible to calculate the shortest path. The basis for the shortest path calculation is a scalar metric (4). The remaining boxes at the left show how this metric is calculated: In the first step, the sensitive ranges (1) of relevant submetrics are normalised (a). This yields the normal cost intervals (2). In the second step, the costs are transformed using the cost functions (b) and combined in a cost vector (3) as explained in Section 2.

4 Practical Routing Parameters

This section defines practical parameters for SIP session routing and their mapping to the normalised range. Since the cost functions uses an implicit ranking of parameters, this has to be reflected in the parameter assignment. Initially, the parameters have to be ordered by their importance. The message loss probability is the most important parameter under the assumption that all parameters have similar values. Message loss instantly results in additional delays caused by the resending of these lost messages. Delay is the second important parameter for paths with similar loss probabilities. Shorter delays are desirable since they improve the interoperability with legacy systems, the media bearer utilisation and the user's satisfaction [8]. If paths have similar costs while considering both the loss probability and the delay, the user count is measured. It is desirable to facilitate the available resources in a way that the load is balanced. The effect of unavailable resources is discussed in later sections. To achieve the above-described behaviour of the metric, the original values have to be normalised, scaled and mapped to the cost functions.

The metrics have a range of values. In this range, changes have an impact on the routing decisions. This interval is defined as the *sensitive range* (SR). This section discusses the

mappings between SRs of practical values and the NCI $[0, b]$. Two definitions additionally use infinity to mark parameters that are out of the sensitive range. Message loss probability, delay and user count are discussed in this section. Other parameters can be used accordingly.

4.1 Message Loss Probability

The probability $\overline{P}$ that no messages are lost has a multiplicative composition rule (e.g. [9]). This probability can be calculated with the message loss probability P by $\overline{P} = 1 - P$. To incorporate this metric in the multidimensional framework an additive composition-rule metric is required. The logarithm function can transform the metric into an additive-rule metric $log(1 - p)$. If the sensitive range of the message loss probability is between p_{max} and p_{min}, Equation (6) defines the mapping of this interval to the cost range of $[0, b]$ where p is the message loss probability.

$$m_x(p) = \begin{cases} 0 & if \quad 0 \leq p < p_{min} \\ b \cdot \frac{log(1-p)-log(1-p_{min})}{log(1-p_{max})-log(1-p_{min})} & if \quad p_{min} \leq p \leq p_{max} \\ \infty & if \quad p_{max} < p \leq 1 \end{cases} \tag{6}$$

These parameters are rather difficult to measure and complicated to calculate. A counter for the number of resends within a specific time interval is a simpler alternative. Every time a SIP proxy server resends a message it increases a link specific drop counter. For example, a bit error of 10^{-9}(respectively 10^{-5}) yields a message loss probability of $4 \cdot 10^{-6}$ (respectively 0.039) for a message size of 500 bytes. For a session arrival rate of 10 sessions per second and 14 messages per session this yields about 1 (respectively 9828) lost message in 30 minutes. Since these message drops are proportional to loss, the count can be used as a measure of reliability. Equation (7) depicts the definition using the count n.

$$m_x(n) = \begin{cases} 0 & if \quad 0 \leq n < n_{min} \\ b \cdot \frac{n-n_{min}}{n_{max}-n_{min}} & if \quad n_{min} \leq n \leq n_{max} \\ \infty & if \quad n_{max} < n < \infty \end{cases} \tag{7}$$

n_{min} is the minimum number of dropped messages and n_{max} is the maximum number of dropped messages. These drop counts follow additive composition rules. For a sensitive range of $[1, 10000]$, a basis of 10 and a message drop count of 2 (respectively 5000) Equation (7) yields a cost m_x of 0.001 (respectively 4.999).

4.2 Delay

The observations in this paper assume that the transmission delay is much smaller than the propagation delay. This is true if the capacity of the connection is sufficiently high. Queuing delays are considered in a different context [4] and are also assumed to be smaller than the propagation delay. The distance between two nodes therefore approximates the delay between nodes. If the *Round Trip Time* (RTT) is known it can be used for this metric. SIP nodes can measure the round trip time for INVITE messages. This has two reasons. Firstly, every session initiation requires an INVITE message, secondly INVITE messages are acknowledged on a hop-by-hop basis. Delay is an additive metric. In this case the mapping to the cost range is straightforward. The definition is depicted in Equation (8) for a sensitive range of $[d_{min}, d_{max}]$.

$$m_y(d) = \begin{cases} 0 & if \quad 0 \leq d < d_{min} \\ b \cdot \frac{d-d_{min}}{d_{max}-d_{min}} & if \quad d_{min} \leq p \leq d_{max} \\ b & if \quad d_{max} < d < \infty \end{cases} \tag{8}$$

For a sensitive range of $[10, 1000]$ ms, a basis of 10 and a delay of 20 (respectively 200) ms Equation (8) yields a cost C_P of 0.1 (respectively 1.9).

4.3 User/Message Count

This metric is a measure of the availability of resources. Servers can handle a limited number of registered users at a time; connections are limited by the number of messages they can accommodate per second. This metric can be used for two parameters: The first is the average number of messages transmitted via a connection. For example, VSLs are defined for a maximum number of messages per second. The second is the number of registered users in a server. In both cases it is possible to calculate the maximum number of available users/messages. The metric takes this number into account. It is similar to the metric "available capacity". Note that VSL traffic only compromises signalling traffic and not general network traffic. Equation (9) depicts the definition, where n is the number of available user spaces and SR is defined as the interval $[n_{min}, n_{max}]$.

$$m_z(n) = \begin{cases} 0 & if \quad n_{max} < n \\ b \cdot \frac{n_{min}}{n} & if \quad n_{min} \leq n \leq n_{max} \\ \infty & if \quad n < n_{min} \end{cases} \tag{9}$$

The upper bound is not considered in this definition. The metric is mainly sensitive between n_{min} and $b \cdot n_{min}$. Larger numbers of available resources yield very small cost values. If no space is available for additional users the path in question has to be "blocked". This is achieved by an infinite cost. Under other circumstances the maximum cost is limited to b. For the interval $[100, 10000]$, a basis of 10 and a count of 200 (respectively 1000) Equation (9) yields 5 (respectively 1).

The maximum number of servable users will differ considerably between various node and connection types, for example, I-CSCFs serve a large number of users whereas P-CSCFs only serve a fraction of the I-CSCF load. To make these servers comparable for routing a bin count can be used instead of the original count. The basis for the routing decision is then the number of "full" bins of size a. The adopted metric can then operate on different scales for different node types. In this case n in Equation (9) is substituted by $n_a = n/b$ in all calculations. If a bin size is chosen that is proportional to the maximum number of users per node, large servers with many connected nodes will be comparable to smaller servers with only a few connections. The same applies for connections of various capacities. Note that in both cases the functional hierarchy of nodes and links has to be considered.

5 Remarks

This section discusses the operation of the routing scheme. It shows a simple transformation to map node restrictions into link limits. The routing update interval and the complexity of the required calculation are also discussed.

Node Restrictions The shortest path algorithms mentioned in this paper considers only link metrics. To account for node restrictions a simple transformation is necessary. Every node is transformed into a node-link-pair. The node's parameters are attributed to the link. Ingress links stay connected to the original node, egress links are connected to the newly added

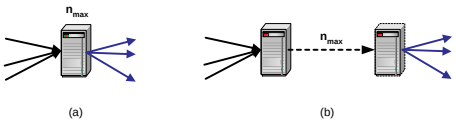

Figure 4: Transformation

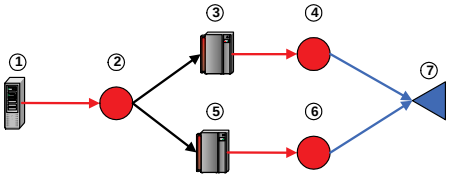

Figure 5: Example

node. Figure 4 shows an example of such a transformation. The server with the assigned cost of n_{max} is shown in part (a) and the transformed node is depicted in part (b). The cost is attributed to the dummy link that connects the original node with the new dummy node. The egress links are connected to the added dummy node.

Routing Update Interval Most of the metrics discussed in this paper change under normal operating conditions very slowly. Network faults are an exception. The number of registered users is the only metric that is changing more frequently. The methodologies are sensitive to fluctuations above the threshold of the bin size in user numbers. The routing protocol updates have to take this into consideration. Once the threshold is reached, updates are sent to other nodes within the routing domain. The maximum errors in the user number information under no error conditions is the number of additional register requests/messages that are allowed in the time interval between the last update and a new update. This is just the minimum number of users or the bin size. As for most routing schemes that use online traffic data for their calculations, this scheme has a tendency to oscillate between different paths ("route flapping"). This behaviour is not crucial in the discussed context, since once a routing decision is made for a user, it is persistent for a long time, usually the registration period. In practice, only new registering arrivals will oscillate between two competing nodes. This is acceptable behaviour if one node provides a better service than the other.

OSPF Protocol The discussion in this paper assumes that a routing protocol is available for use. It proposes the use of the OSPF protocol for this purpose. Where it is believed that no major changes to the protocol are required, adjustments are necessary in the case where the protocol is in use on the VSON layer. The OSPF *Autonomous System* (AS) where the nodes have complete knowledge of the topology has to reflect the operator domain. In particular, the server discovery and the routing message delivery are for further study as they depend principally on further standardisation and implementation decisions. In general, any shortest path routing protocol can be used.

Complexity of Calculations The major differences between existing routing schemes and this approach is the complexity of the metric. Every server has to calculate the metric. The complexities of the calculations are not significant since the changes are not frequent and the calculations required are of $O(1)$. The node measures its parameters and triggers routing protocol updates. The propagation delay between fixed network nodes is unlikely to change over long time periods. The mobility of users is handled beyond the P-CSCF in the *UMTS Terrestrial Radio Access Network* (UTRAN) and is not in the scope of this paper. Some general simplifications are possible in the calculation of the metric, for example, Equation (5)

Table 1: Example

		l_{12}	l_{23}	l_{34}	l_{47}	l_{25}	$l_{56}(1)$	l_{67}	$l_{56}(2)$
(a)	Loss	0	100	0	0	100	0	0	0
	Delay	0	50	0	0	40	0	0	0
	Users	1000	5000	2000	∞	2000	2000	∞	120
(b)	m_x	0	0.1	0	0	0.1	0	0	0
	m_y	0	0.8	0	0	0.6	0	0	0
	m_z	1	0.2	0.5	0	0.5	0.5	0	8.3
(c)	$\mathbf{C}$	$\begin{pmatrix} 0 \\ 0 \\ 10 \end{pmatrix}$	$\begin{pmatrix} 99 \\ 80 \\ 2 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 5 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 99 \\ 60 \\ 5 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 5 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 167 \end{pmatrix}$
(d)	C	10	127	5	0	116	5	0	167

can use C^2 instead of C and save the square route calculation. This is especially important since square route computations would take significant amounts of processor time on a server.

6 Example

This section introduces a very basic example to show the operation of this routing scheme. A example network is depicted in Figure 5. For simplicity, this example uses three network nodes (1), (3) and (5). Node (1) represents a P-CSCF node and node (3) and (5) are I-CSCFs. Every network node in Figure 5 is depicted with its associated dummy node (2), (4) and (6) respectively as a result of the transformation described in Section 5. Node (7) is the dummy destination node that enables the optimised selection of one of the network nodes (3) or (5). The example uses the basis $b = 10$. The cost boundaries are chosen to be $\chi_{x1} = 6$, $\chi_{x2} = 8$ and $\chi_y = 8$ respectively. The sensitive ranges are: $10 \ldots 1000$ for the user count, $1 \ldots 10000$ for the number of lost messages in 30 minutes and $10 \ldots 500$ for the delay. Table 1 shows the first example.

Row (a) shows the original cost for each link l_{ab}. This first example uses $l_{56}(1)$. Row (b) gives the normalised costs for loss m_x, delay m_y and the number of users m_z. Row (c) depicts the cost vectors $\mathbf{C}$ after the cost functions are applied. A cost calculation for the two paths $P_1 = \{1, 2, 3, 4, 7\}$ and $P_2 = \{1, 2, 5, 6, 7\}$ yields the costs 142 and 131 respectively. P_2 is therefore the shortest path. Comparing the values in row (a) shows that both paths have similar reliabilities and that both paths can accommodate additional users. The difference in the number of available users spaces between link l_{23} and link l_{25} 5000 to 2000 respectively, has no significant impact and the path with the smallest delay is selected.

A second example uses all settings of Table 1 with the link $l_{56}(2)$. The difference in this example is that the number of available user spaces in node five is very limited (120). This value is close to the minimum number of available user spaces in this node. This changes the cost situation considerably. In this example, the path cost calculations still yield a value for path P_1 of 142, but for path P_2 a cost of 293. Thus, path P_1 is now selected as the shortest path in this case.

The selection of P_1 indicates that node (3) was chosen over node (5). In the first example, when enough resources were available in node (5) it was selected since the loss on link $l_{2,5}$ was less than the loss on link $l_{2,3}$. In the second case, the resources in node (5) were limited, so node (3) was chosen, this seems to be a reasonable routing decision. Cases for the dominance of reliability can easily be derived from this example.

7 Conclusions

This paper introduced routing methodologies for signalling messages sent between SIP nodes on the application layer. It defined a multidimensional metric with various sensitivities for different network operating conditions. The principal routing scheme as well as the multidimensional metric can be applied to other generic networks and routing problems. The question of where the routing scheme can be used in an operational 3GPP IMS, depends on the availability of information, for example, is load-information accessible for routing decisions or not? The proposed scheme is flexible; it can operate in selected network areas or server clusters as well as on a large scale.

The use of OSPF as the routing protocol on the application layer was suggested, but the detailed implementation is for further study. In particular, the discovery of the nodes, the layering, update interval, information flooding etc. require further investigation. Further study is also necessary for the exact practical implementation and performance evaluation in operational networks, as SIP systems are deployed and further information becomes available.

Acknowledgements

The authors would like to thank Ericsson AsiaPacificLab Australia and the Australian Telecommunications Cooperative Research Centre (ATcrc) for their financial support for this work. The helpful comments by Bill Lloyd-Smith in the CATT Centre, RMIT University, are gratefully acknowledged.

References

[1] J.Rosenberg, H.Schulzrinne, G.Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. *SIP: Session Initiation Protocol*. IETF, June 2002. RFC 3261 (Obsoletes RFC 2543).

[2] 3rd Generation Partnership Project. *About 3GPP*, October 2002. http://www.3gpp.org.

[3] 3rd Generation Partnership Project. *IP Multimedia (IM) Subsystem - Stage 2 (Release 5)*, July 2001. 3GPP TS 23.228 V5.1.0.

[4] A.A. Kist and R.J. Harris. *Using Virtual SIP Links to Enable QoS for Signalling*. In ICON 2003, Sydney, Australia, September 2003. To appear.

[5] J. Moy. *OSPF Version 2*. IETF, April 1998. RFC 2328.

[6] Cisco Systems, Inc. *Enhanced Interior Gateway Routing Protocol*, October 2002. http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/en_igrp.htm.

[7] R.A. Guérin and A. Orda. Qos Routing in Network with Inaccurate Information: Theory and Algorithms. *IEEE /ACM Transactions on Networking*, 7(3):350–364, June 1999.

[8] A.A. Kist and R.J. Harris. SIP Signalling Delay in 3GPP. *In Proceedings of Sixth International Symposium on Communications Interworking of IFIP - Interworking 2002, Perth Australia, October 13-16*, October 2002.

[9] Z. Wang and J. Crowcroft. Quality-of-service routing for supporting multimedia applications. *IEEE Journal of Selected Areas in Communications*, 14(7):1228–1234, 1996.