Congestion Control Mechanism for Traffic Engineering within MPLS Networks

Felicia Holness, Chris Phillips

Department of Electronic Engineering – Queen Mary and Westfield College
University of London – London, UK, E1 4NS
Tel: +44 (0) 20 7882 3755, +44 (0) 20 7882 7989 Fax: +44 (0) 20 7882 7997
Email: f.holness@elec.qmw.ac.uk
Email: c.phillips@elec.qmw.ac.uk

Abstract: The transformation of the Internet into an important and ubiquitous commercial infrastructure has not only created rapidly rising bandwidth demands but also significantly changed consumer expectations in terms of performance, security and services. Consequentially as service providers attempt to encourage business and leisure applications on to the Internet, there has been a requirement for them to develop an improved IP network infrastructure in terms of reliability and performance [1]. Interest in congestion control through traffic engineering has arisen from the knowledge that although sensible provisioning of the network infrastructure is needed together with sufficient underlying capacity, these are not sufficient to deliver the QoS required [2]. This is due to dynamic variations in load. In operational IP networks, it has been difficult to incorporate effective traffic engineering due to the limited capabilities of the IP technology. In principle, Multiprotocol Label Switching (MPLS), a connection-oriented label swapping technology, offers new possibilities in addressing the limitations by allowing the operator to use sophisticated traffic control mechanisms.

However, as yet, the traffic engineering capabilities offered by MPLS have not been fully exploited. Once label switched paths (LSPs) have been provisioned through the service providers' network, there are currently no management facilities for dynamic re-optimisation of traffic flows. The service level agreements (SLAs) between the network operator and the customer are agreed in advance of the commencement of traffic flow, and these are mapped to particular paths throughout the provider's domain and may be maintained for the duration of the contract. During transient periods, the efficiency of resource allocation could be increased by routing traffic away from congested resources to relatively under-utilised links. Some means of restoring the LSPs to their original routes once the transient congestion has subsided is also desirable.

Today's network operators require the flexibility to dynamically renegotiate bandwidth once a connection has been set up [3] preferably using automated solutions to manage an access switch management algorithm and route connections. Although these services are already provided to some extent with provisioning, they tend to occur relatively infrequently (several times in a day) using prior knowledge and manual intervention. There are currently no mechanisms in place within the network to allow the operator to rapidly change the traffic paths in response to transient conditions.

This paper proposes a scheme called Fast Acting Traffic Engineering (FATE) [6][7] that dynamically manages traffic flows through the network by re-balancing streams during periods of congestion. It proposes management-based algorithms

S. Rao and K.I. Sletta (Eds.): INTERWORKING 2000, LNCS 1938, pp. 254-263, 2000.

that will allow label switched routers (LSRs) in the network to utilise mechanisms within MPLS to indicate when flows may be about to experience possible frame/packet loss and to react to it. Based upon knowledge of the customers' SLAs, together with instantaneous flow information, the label edge routers (LERs) can then instigate changes to the LSP route to circumvent congestion that would hitherto violate the customer contracts.

Keywords: MPLS, Traffic Engineering, LDP, LSR, CR-LDP,FATE.

1. Multi-service Provisioning Environment

At present the Internet has a single class of service - "best effort". As a result of this single service all traffic flows are treated identically, there is no priority servicing regardless of the requirements of the traffic. A provisioning scheme that can be applied within an MPLS environment can be described as follows:

Consider **Fig. 1**, which shows a scheduler at each egress port of a LSR. The scheduler has been programmed to visit each class-based buffer at a rate commensurate with the loading of that particular buffer and its identified Quality of Service (QoS) constraint(s) i.e., a long-term guaranteed loss limit.

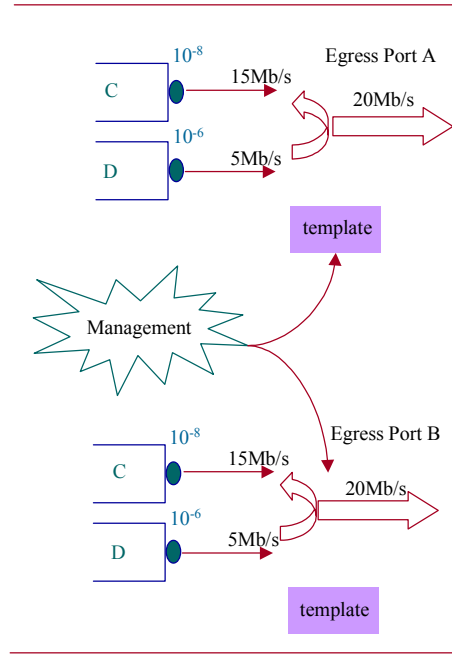

Fig. 1 Multi-Service Provisioning

The order and frequency with which the scheduler services each of the buffers is determined by a port template that may be programmed by the management module and read by the scheduler. In the scenario depicted, where buffer C would be serviced three times more than buffer D, the template could take the format shown below in **Fig. 2**.

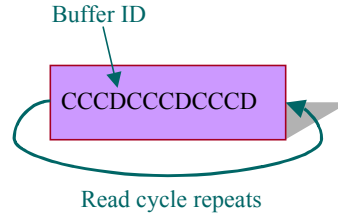

Fig. 2 Example Scheduling Template

A scheduling template is programmed according to a predetermined loss probability threshold for each buffer for a given anticipated load. In a situation where the traffic loading through a buffer stream increases, the management function has the ability to change the template of the scheduler if deemed appropriate. The management function uses its knowledge of the traffic characteristics and the current loading to determine whether a new constraint-based routed label switched path (CR-LSP) can be routed through a particular buffer, whilst attempting to maintain the buffer's traffic engineering constraints below the specified loss probability. Over a specified time period the loss probabilities through each buffer stream are recorded and in the event that the loss exceeds a predefined threshold the management function may decide to alter the template i.e. the scheduler's rate, to accommodate for the additional loading.

Scheduling reallocation is permissible provided the contracted QoS requirements of the LSPs traversing the buffer(s) are not violated.

However, as this method is based on predicted behaviour (estimated over the last time period) it does not cater for transient fluctuations in load. It is a provisioning mechanism. FATE on the other hand provides a means of dynamically redistributing existing CR-LSPs between buffers or via alternative paths in response to short term congestion events.

1.1. Traffic Flow Provisioning

So far basic provisioning has been considered i.e., mapping LSPs to buffers according to their particular QoS requirements and the long term loading situation. This mechanism allows for readjustment of the scheduling templates in response to predicted loading variations. It is relatively slow and operates on buffers – it does not provide granularity to respond to issues associated with individual LSPs.

FATE provides a fast acting mechanism, which augments the above scheme. It allows individual LSPs to be dynamically remapped to higher QoS buffers along a specified path in response to transient congestion situations.

2. Fate

This section describes the mechanisms and procedures that are employed within the FATE (Fast Acting Traffic Engineering) scheme. It proposes a fundamental set of novel mechanisms that can be employed in the effort to either pre-empt congestion or respond to its occurrence in a LSR along which a CR-LSP has already been established.

2.1 Congestion Detected in a CR-LSP

An ingress LER can determine the contribution it makes to the utilisation of the LSRs along each LSP, and can set up CR-LSPs[1] with that limited knowledge. However, it currently has no knowledge of how those same LSRs are being utilised by other LERs. It is this lack of information when deciding which LSP may meet an application's requirement that can lead to congestion occurring within a downstream LSR.

Assume over time that as a result of the increased load through a LSR, it starts to lose packets from a LSP. If the value exceeds a given threshold i.e., the loss probability assigned to that particular buffer, it is taken as a sign of possible congestion in that buffer stream, within that LSR.

Once the packet loss in the best effort LSP has risen above the predetermined threshold value, for an extended time period the LSR creates a LDP notification message containing the proposed *Congestion Indication* TLV. The objective of sending the *Congestion Indication* notification (CIN) message is to indicate to the ingress LER that there are packets being lost from a particular CR-LSP originating from it, allowing the ingress LER to either:

1. 1.Decide that the packet loss it is currently experiencing remains sufficiently low for it to continue to meet its SLA requirements, allowing/permitting no further action to be taken at this time.
2. 2.Renegotiate for new quality requirements along the existing LSP[2].
3. Negotiate for new quality requirements along an alternative LSP.

In order for the ingress LER to act on the received information, it needs to know the following:

1. The identity of the LSP that is experiencing congestion.
2. The current loss in the buffers the LSP is traversing.[3]
3. The LSRs this loss is occurring in.
4. The current loss the LSP is experiencing.

As a result of this information, the congested LSR generates a *CIN* message. This must contain the identity of the LSR that is experiencing loss, the identity of the CR-LSP along which packet loss was detected, and the packet loss the LSP and buffer are currently experiencing. The congested LSR uses the input port that the packet was received on and the input MPLS label, as an index into the Next Hop Label Forwarding Entry (NHLFE) table to obtain the CR-LSP ID. The CR-LSP ID identifies the *Ingress LSR Router ID* i.e., originating LER, and the local value assigned by that ingress LER to identify the CR-LSP initiated by it. The buffer the LSP traverses through at the congested LSR is obtained using the CR-LSP ID to index a separate *Buffer Table*[4] Each

---

[1] The terms CR-LSP and LSPs are used interchangeable, the main difference is that CR-LSPs are established based on constraints, e.g., explicit route constraints, QoS constraints, etc.

[2] Request that the LSP be promoted to pass through a higher priority buffer along the same path, and within the same LSR.

[3] Although each buffer and its servicing scheduling are dimensioned for a specific CLP, at any time due to traffic loading the current available packet loss within the buffer may have increased or decreased.

[4] The *Buffer Table* is to maintained by each LSR to a record of the availability of each buffer's resources.

LSR experiencing congestion records in its *Congestion Indication Table*[5] the CR-LSP *ID*, and the current LSP and buffer losses. A timer is set. If when the timer expires, the LSR is still suffering from congestion, the LSR will send another *CIN* message with the updated calculated loss values and reset the timer.

The LSR's own IP address is included in the message along with the current packet loss both the CR-LSP and the buffer are experiencing. The *CIN* message is then forwarded to the next hop LSR towards the ingress LER.

Rather than all congested LSRs always generating CIN messages, intermediate LSRs upon receipt of a *CIN* message may append relevant information to it concerning their status if they are also experiencing congestion. If a LSR receives a *CIN* message shortly after sending one, it checks the *Congestion Indication Table* to determine if the timer it has set has expired. If it has not expired, it will simply forward the message without appending its own information, otherwise it will include its information before forwarding.

Timers are used to control the responsiveness of the FATE scheme to traffic loading transients. For example, when a LSR is congested it can issue a *CIN* message. In doing so it sets a retransmission timer. It is not permitted to issue another message until the timer expires, thus avoiding signalling storms whilst improving the robustness of the protocol. Alternatively, if it receives a *CIN* message on route to the ingress LER from another congested LSR, it can simply append its own congestion information and set the timer accordingly. In doing so, avalanches of congestion notification messages towards the ingress LER are prevented. In addition, stability is improved by averaging the observed traffic parameters at each LSR and employing threshold triggers.

When the ingress LER receives a *CIN* message, it may do any of the actions previously outlined.

The motivation behind monitoring individual LSPs through a particular buffer stream stems from the ingress LER's need to ensure the SLAs between the customers and the MPLS network are maintained at all times. To enable it to do this, it needs to have knowledge of the loss encountered by the individual LSPs originating from it. Individual LSPs from a customer site are aggregated into LSPs that share class based buffer resources. As a result of this, the LSP loss rather the individual flow losses is reported back to the ingress LER, who has knowledge of which flows are affected via the flow/LSP binding information.

By monitoring both the losses experienced by individual LSPs and buffer streams, it gives the ingress LER two averages to consider when deciding whether to renegotiate QoS requirements along an existing path or a different path, or whether to accept the current condition.

For example, consider when an ingress LER receives an indication that the loss in a buffer its LSPs are passing through is experiencing a particularly poor loss probability (1 in 10^{-2}). However the loss probability experienced by the buffer it is traversing (1 in 10^{-5}) is acceptable. Or it may decide to set an *Optional Response Timer*; if it receives another CIN message before the timer expires it will take appropriate action. However, if the timer expires and no CIN message is received, it will assume the loss experienced by its flows has fallen within the negotiated value. Some means of averaging the loss statistics provides a useful dampening factor. To prevent an avalanche of CIN messages being sent to a single ingress LER, the congested LSR when it determines that more than one

[5] The *Congestion Indication Table* is maintained by each congested LSR, it contains information about the flow and buffer experiencing loss that has exceeded the predetermine thresholds.

CR-LSP traversing its buffers is experiencing a particularly poor loss probability, will aggregate the CIN messages for those individual buffers.

2.2 SCALABILITY

Monitoring losses in individual LSPs is not very scalable, even if those LSPs represent the aggregation of individual connections or flows from a customer site. It is quite possible that at any instance in time, a LSR could be expected to handle a very large number i.e., thousands of these LSPs. As a result of this scalability issue, detecting losses in individual LSPs described previously, may not be a viable option in an MPLS domain expected to maintain a large volume of LSPs[6]. This immediately poses two questions.

How is it possible for an autonomous MPLS network to apply congestion control mechanisms in a situation were it has numerous flows, some of which may be entering the domain just after exiting a customers premises, and others on route from or to another autonomous domain?

How can this service provider ensure the customers SLA is met whilst traversing this network?

In monitoring a single LSP or a number of LSPs that connect between a specific source and destination, connected within a single autonomous system, it is quite easy to identify the ingress and egress LERs, and the exchange of messages can be easily handled under the control of the operator.

Consider the case when the source and destination are not within the same domain, or where the MPLS domain is as an intermediary transport 'pipe'. It is not possible or desirable for the operator to determine the absolute source and destination of each LSP.

The author proposes assigning a ***Virtual Source/Virtual Destination*** (VS/VD) [4] pair for the aggregation of LSPs entering the domain at one point and exiting at another point, using label stacking or tunnelling within the autonomous MPLS domain of interest.

All LSPs arriving at a particular ingress LER and exiting at a particular egress LER are assigned to a FEC. The ingress LER also known as the virtual source, is the entry point to the MPLS domain and it is at this point that an additional label is 'pushed' onto the label stack, and used to 'tunnel' the packet across the network. On arriving at the egress LER, also known as the virtual destination, the label is 'popped' and the remaining label used to forward the packet.

By employing label stacking within the domain and assigning VS/VD pairs, the issue of scalability is removed whilst allowing the operator control of the LSPs traversing its network. It allows for efficient utilisation of the limited network resources and the additional capability of controlling congestion. With the VS/VD paradigm, the congestion control message need only propagate along as far as the virtual source for the ingress LER and to the virtual destination for the egress LER. With the virtual endpoints of the LSP defined, aggregation of many LSPs can be treated as an individual LSP as described previously.

---

[6] However [4] explains how operation and maintenance (OAM) cells are used in ATM for fault management and network performance on a point to point connection basis, thus implying it is possible to monitor a large number i.e., thousands of flows or connections.

2.3 Renegotiation Procedures

On receiving a *CIN* message the ingress LER extracts the following information: CR-LSP ID that encodes the *Ingress LSR Router ID* and a locally assigned value *Local CR-LSP,* from the LSP-ID TLV. The LSR Router ID experiencing loss, and the value of packet loss the LSP and buffer are currently experiencing, from the *Congestion Modificaiton* TLV. The *Ingress LSR Router ID* along with the *Local CR-LSP* identifies that this message has been received by the correct ingress LER. With this information the ingress LER is able to identify the particular LSP and its traffic parameters.

The ingress LER needs to determine whether it should renegotiate along an existing LSP for a higher buffer stream offering improved servicing or whether it should negotiate for a new LSP route. The decision depends on information gathered from Statistical Control messages explained later.

2.3.1 Renegotiation along an Existing LSP within a Higher QOS Buffer

If the ingress LER decides to renegotiate along an existing path for a higher service class, it will carry out the following procedure: The ingress LER formulates a Label Request message with the ActFlag set, to indicate that this is an existing CR-LSP along which the traffic parameters need to be modified. The Label Request message contains the newly requested modified traffic parameters along with the service class it requires. When each LSR receives a Label Request message it uses the globally unique CR-LSP ID as an index into the *Buffer Table* to determine which buffer stream the CR-LSP traverses, the amount of bandwidth initially reserved and the loss probability assigned to that CR-LSP identified by the CR-LSP ID.

The LSR then chooses a higher buffer stream to the one the CR-LSP currently traverses. It then determines whether it can allocate the bandwidth and the minimum loss probability requested within one of the alternative buffer streams. If it can, it temporarily assigns that amount in the new buffer stream, whilst maintaining the original entry. It alters the available bandwidth within the *Buffer Requirement Table* and forwards the Label Request message to the next hop.

If all the LSRs along the CR-LSP are able to meet the requirement on receipt the egress LER will create a LDP notification message containing a *RenegSuccess* TLV indicating the resources have been reserved and send it to the upstream LSR towards the ingress LER of the CR-LSP.

On receiving a *RenegSuccess* notification message each LSR will permanently assign the resources to the path. The *RenegSuccess* notification message is then passed upstream. On receipt of a *RenegSuccess* message the ingress LER updates the FEC/label binding to reflect the higher buffer stream through which the CR-LSP will now be routed.

The *Reneg Success* notification message includes the CR-LSP ID, along with the parameters agreed[7] on, in terms of bandwidth required and minimum LSP loss probability

If a LSR cannot allocate the additional resource it will send a proposed *RenegFailure* TLV within a notification message to the message source and not propagate the Label Request message any further. The LSR will append to the *RenegFailure* notification message the maximum current available bandwidth it can allocate within each of its buffer streams that are also capable of meeting the minimum loss probability requested.

---

[7] This document assumes the bandwidth is controlled by the operator by possibly using policing and shaping mechanisms, but these mechanisms are beyond the scope of this document.

On receipt of a *RenegFailure* notification message, the LSR will deduce that another LSR further upstream has been unable to allocate resources for a LSP which traverses one of its own buffers.

The ingress LER on receiving a *RenegFailure* notification message will have enclosed a single value representing the lowest currently available bandwidth that can be offered along that CR-LSP, whilst realising that renegotiation along the existing path has failed for that CR-LSP and decides on remedial action. The protocol supports a "crank back" mechanism. For instance, when the ingress LER receives a *RenegFailure* notification message it can select an alternative path either by referring to a topological link cost database maintained by a separate routing protocol or the decision is made by the network management module. It then sends a Label Request message along the revised path. When it receives a Label Mapping confirming a new path has been set up, it replaces the old Label Mapping with the newly received Label Mapping, it can then delete the original label or keep it to send other data along the path it represents. If the decision is to delete the original label, the ingress LER will send a Label Release message [5] including the newly replaced Label along the LSP to the egress LER. This procedure results in the label being removed from the pool of "in use" labels. This Label Release message should be sent a few seconds after the last packet is forwarded along that path to ensure the egress LER receives the last packet before it removes the label from forwarding use[8].

2.4 Monitoring Procedures

Proposed *Statistical Control* TLVs contained within LDP notification messages, known as *Status Requests*, are sent into the network periodically from the ingress LER or when the ingress LER receives a *CIN* message.

When the ingress LER chooses to issue a Status Request, it uses the CR-LSP ID to determine which CR-LSP it refers to. It then formulates the Status Request message with the explicit route and CR-LSP ID included and transmits it to the next hop in the ER.

As each LSR receives it, it appends its own statistical information to the message. This includes the loss probability experienced by this CR-LSP. It also includes the current losses of all the alternative class-based buffers the CR-LSP could pass through at this LSR along the specified path[9]. It then forwards the *Status Request* to the next LSR. When the message reaches the egress LER, it is sent back to the ingress LER. Upon receipt of a *Status Request* message that it earlier issued, the ingress LER extracts the CR-LSP ID, and records for each LSR along that path the loss experienced both by this CR-LSP and the loss currently being experienced by all the relevant buffers at each LSR. This information is recorded in a *Statistical Buffer Table* for monitoring purposes.

The *Status Request* messages provide an overall view of the status of the links and LSRs along a particular CR-LSP. It includes the available bandwidth and loss probabilities within every buffer stream within a LSR, as well as the loss experienced by a CR-LSP.

---

8 Alternatively a 'flushing' mechanism could be used to ensure all data sent along the former path has reached its destination prior to forwarding more data along the new path [4].

9 In this thesis loss is used as an example statistical parameter, however, this could be easily generalised to a variety of traffic engineering performance metrics.

The CIN messages only return status information about the CR-LSP suffering unacceptable loss and the particular buffer it traverses in the congested LSRs between the ingress LER and the initiator of the message i.e., not the entire CR-LSP.

Subsequently, if the ingress LER receives a *CIN* message, it examines the information held in its *Statistical Buffer Table* to help determine whether it should renegotiate along the existing path, as the higher buffer streams seem capable of meeting its QoS requirements. Alternatively, it can choose to negotiate for an alternative path.

2.5 Simulation Results

One scenario considered for the simulation involved a network of two LSRs and two LERs as shown in Figure 3. A CR-LSP between the ingress and egress LERs is established, with flows passing through all the buffer streams within each LSR/LER ①. After a specified simulation time the number of flows are increased to cause the loss probability within LSR 1 to rise above a predetermined threshold. The moment at which that point is detected the FATE congestion mechanism responds. In this particular scenario, the flows are switched onto a higher buffer stream along the same LSP ②.

The associated graph shows the point at which congestion is detected in a buffer and within a LSP. It also shows the operation of congestion indication mechanisms as witnessed by the subsequent reduction in packet loss once the flows are transferred to a higher buffer stream.

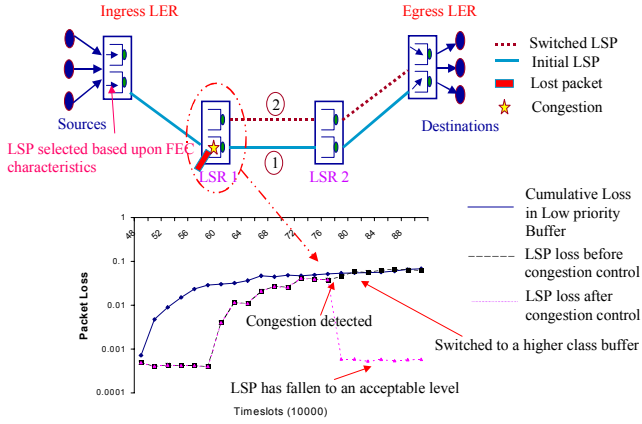

Fig. 3 Loss detected before and after FATE Mechanism

2.6 Discussion

Preliminary results from the FATE scheme illustrate the benefit of dynamic renegotiation of LSPs to control congestion within an MPLS domain. Without this control, network operators have no viable means of redistributing traffic flows at short notice onto under-utilised links / LSRs. The issue of scalability has been addressed by exploiting LSP aggregation to provide tunnels between VS/VD pairs. This mechanism also allows the

operator to exercise full traffic engineering control within their domain without affecting the content of the received flows.

Timers are used to control the responsiveness of the FATE scheme to traffic loading transients. For example, when a LSR is congested it can issue a Congestion Indication notification message. In doing so it sets a retransmission timer. It is not permitted to issue another message until the timer expires, thus avoiding signalling storms whilst improving the robustness of the protocol. Alternatively, if it receives a Congestion Indication notification message on route to the ingress LER from another congested LSR, it can simply append its own congestion information and set the timer accordingly. In doing so, avalanches of congestion notification messages towards the ingress LER are prevented. In addition, stability is improved by averaging the observed traffic parameters at each LSR and employing threshold triggers. Although the averaging window has so far been set to 5 seconds, current research is examining the stability of the system as this parameter is adjusted.

A further area of ongoing research is the extension of FATE, called FATE+ where the decision in situations of congestion is the responsibility of the congested LSR and is not passed to the ingress LER. FATE+ is suitably employed along 'loosely' routed CR-LSPs.

References

1. Coombs. S, Nortel Networks Initiates Major Step in MPLS Multivendor Interoperability, http://www.newswire.ca/releases/March1999/16/c4402.html, March 1999.
2. Borthick.S, '*Router Startups Betting on MPLS',* http://www.bcr.com/bcrmag/11/nov98p14.htm.
3. Semeria.C, '*Traffic Engineering for the New Public Network'*, http://www.juniper.net/techcenter/techpapers/TE_NPN.html, 22/02/00, 25/02/00.
4. McDyson, Spohn, '*ATM Theory and Applications'* ISBN 0-07-645356-2
5. Andersson,L Doolan,P et al, LDP Specification, http://search.ietf.org/internet-drafts/draft-ietf-mpls-ldp-08.txt.
6. Holness,F Phillips,C '*Dynamic QoS for MPLS Networks',* accepted at the 16th UK Teletraffic Symposium, (UKTS), Nortel Networks, Harlow, England, May 22-24, 2000.
7. Holness,F Phillips,C '*Dynamic Traffic Engineering within MPLS Networks',*accepted at the 17th World Telecommunications Congress (WTC/ISS2000) Birmingham, England, May 7-12,2000.