TSVWG
Internet Draft
draft-briscoe-tsvwg-cl-phb-00.txt
Expires: January 2006

B. Briscoe
G. Corliano
P. Eardley
P. Hovell
A. Jacquet
D. Songhurst
BT

July 11, 2005

The Controlled Load per hop behaviour
draft-briscoe-tsvwg-cl-phb-00.txt

Status of this Memo

Copyright Notice

Abstract

This document defines two per-hop behaviours (PHBs) called Controlled Load step and Controlled Load ramp (CL-step-PHB and CL-ramp-PHB). CL

service is a quality of service (QoS) closely approximating the QoS that same flow would receive from a lightly loaded network element [RFC2211]. CL service is useful for inelastic flows such as those for streaming real-time media.

Explicit Congestion Notification (ECN) marking semantics are defined as part of the PHBs, to provide an "early warning" of potential congestion. The PHBs can be used as a basic building block within an edge-to-edge (CL-ramp-PHB) or end-to-end (CL-step-PHB) QoS architecture, using distributed measurement-based admission control.

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Table of Contents

1. Overview and motivation

Network nodes that implement the differentiated services (DS) enhancements to IP use a codepoint in the IP header to select a per-hop behaviour (PHB) as the specific forwarding treatment for that packet [RFC2474, RFC2475]. This document describes two new PHBs called Controlled Load (CL) ramp and step (CL-ramp-PHB and CL-step-PHB).

The Quality of Service (QoS) that can be achieved with the PHBs corresponds to the Integrated services Controlled Load (CL) class, that is [RFC2211]: "a quality of service closely approximating the QoS that same flow would receive from [a network element that is] lightly loaded".

The CL PHBs are different from PHBs defined so far, in that they define Explicit Congestion Notification (ECN) marking semantics as part of the PHB. [RFC3168] specifies the default ECN marking behaviour of an ECN-capable router for currently standardised PHBs, but allows new PHBs to define different ECN marking behaviour. The default behaviour is to indicate incipient congestion by setting the Congestion Experienced (CE) codepoint in the IP header of packets, with a probability determined by the RED algorithm [RFC2309] based on the moving average queue length - so packets have their CE codepoint set at the same point that a non-ECN-capable router would drop them.

This document specifies a different behaviour: a node sets the CE codepoint in the IP header as an "early warning" of potential congestion, and aims to do so before there is any significant build-up of CL packets in the queue. This information enables appropriate action to be taken to heed the "early warning" so that the router should never become overloaded and forced to queue packets, and hence the service achieves a higher performance. One example of an 'appropriate action' is that an ingress gateway blocks new microflows (Section 3.2 and [CL-arch]). Hence it is possible for the CL packets to suffer minimal queuing delay, jitter and loss - exactly the requirements of real time traffic.
So a node using a CL PHB generates ECN signalling and is also part of a system of other nodes in a closed feedback loop, which allows control of the offered load on the local queue. To summarise: the CL PHBs provide QoS by virtue of their local scheduling behaviour, in combination with admission control.

The basic concepts are:

- Setting the CE codepoint: the router sets the CE codepoint of CL packets before it is actually congested (but when congestion is fairly imminent). Reason: it enables appropriate action to be taken so that actual congestion is not experienced, and hence a very low loss, delay and jitter service is possible.

- Discarding: CL packets should not need to be discarded, but if necessary packets would be discarded on a drop tail basis.

- Priority queuing: at a minimum for the CL-ramp-PHB with adaptive bandwidth (see below), CL packets are prioritised before non-CL packets, ie they are always de-queued first. Reason: This minimises CL packets' delay and jitter.

- Reordering: CL packets within a flow must not be reordered, for example this can be achieved with a first-in-first-out (FIFO) queue. Reason: packet reordering significantly degrades the performance of real time applications.

- Selecting the output link: the router selects the link on which to forward the packet based on normal IP routing.

To indicate that this new behaviour is required, packets are marked with one of two new Differentiated Services Codepoints (DSCPs). It is hoped that (an evolution of) the CL PHBs is standardised, which requires each PHB to have an associated RECOMMENDED DSCP. RECOMMENDED DSCPs, rather than EXP/LU (experimental / local use) ones, enable multi-domain operation and vendor interoperability.

Two usage scenarios are suggested in Section 3 (others are possible):

- A host-to-host service, potentially through several domains, using the CL-step-PHB

- An edge-to-edge service across a single domain or across cooperating domains, using the CL-ramp-PHB.

How the CL PHBs can co-exist with existing DiffServ (DS) PHBs and with the default [RFC3168] ECN behaviour is discussed in Sections 2.6 and 2.7.

The CL PHBs - unlike the PHBs for Expedited Forwarding [RFC3246] and Assured Forwarding [RFC2597] - do not require routers to be configured with a minimum departure rate, although it is not precluded.

There are two CL PHB groups each of which consists of a single individual PHB. They are intended for general or local use.

One motivation for this specification is that some network operators are planning to carry all their traffic on a single unified network, because this is expected to reduce operating costs dramatically. Therefore such a network must carry real time inelastic flows like voice and video calls, which are very delay sensitive. Per microflow reservations are too unwieldy in the core and backbone of the network. It is possible to achieve low delays in the core and backbone with DS today, by triggering flow admission control when traffic entering a DS domain exceeds its contracted rate. But for longer topologies, the chances increase that traffic will focus on a resource near the egress, even though it is within contract at the ingress [Reid]. This is because the ingress contract must allow any destination address, if it is to remain scalable. Even though networks can be engineered to make such failures rare, when they occur all inelastic flows through the congested resource fail catastrophically.

2. Detail

The CL PHBs define forwarding behaviour for packets in the CL behaviour aggregate.

2.1. Queuing

CL and non-CL packets are put into logically separate queues; if required, a CL packet can pre-empt non-CL packet(s) in the total buffer (see below).

2.2. Setting the Congestion Experienced (CE) codepoint

A CL implementation in a DS node MUST detect and respond to potential congestion within the CL traffic aggregate by setting the CE codepoint of CL packets, with a probability determined by the algorithms described below, when it forwards them. 'Potential congestion' means a little before the CL packets start suffering a significant delay due to queuing in the node.
There are two PHBs with slightly different behaviour:
1. CL-step-PHB: The probability that a node sets the CE codepoint of a packet is either 0 or 1 (ie 'on' or 'off').

2. CL-ramp-PHB: The probability that a node sets the CE codepoint of a packet can be any value.

Implementation details include the precise algorithm and the value of its parameters.

2.2.1. Possible algorithms for setting the CE codepoint

Each node in the CL-region runs an algorithm to determine whether to set the CE codepoint of a particular CL packet. In our description we assume that a bulk token bucket is used (note that there is not a token bucket per microflow). Tokens are added when packets are queued and are consumed at a fixed rate. The idea is that an excess of tokens is seen before the queue of CL packets has got long enough to cause the CL packets to suffer a significant delay - the algorithms are explained more fully below and are slightly different for the CL-step-PHB and CL-ramp-PHB. Other implementations are possible.

The two CL PHBs use different algorithms for determining how the amount of tokens whether a packet has its CE codepoint set:

1. CL-step: The CE codepoint setting is either 'on' or 'off'. As soon as the amount of tokens is greater than a threshold value, then all CL-step packets have their CE codepoint set. To prevent instability, the metering function has built-in hysterisis, ie this continues until the amount of tokens has decreased below a second threshold value.

2. CL-ramp: There are two alternatives described in detail in [CL-arch]:

o Configured bandwidth: the node has a fixed maximum bandwidth allocated to CL-ramp traffic, under the control of management configuration. Tokens are consumed slightly slower than this rate. The probability that a node sets the CE codepoint of a CL-ramp packet depends on the number of tokens in the bucket. Below one threshold value of the number of tokens no packets have their CE codepoint set and above the second they all do; in between, the probability increases linearly.

o Flexible bandwidth: the node allows an adaptive trade-off between CL-ramp and non-CL-ramp traffic. Tokens are consumed at slightly less than the (total) outgoing service rate. The probability that a node sets the CE codepoint of a CL packet depends on the number of tokens in the bucket *plus* the number of queued non-CL packets. Below one threshold value of this sum no packets have their CE codepoint set and above the second they all do; in between, the probability increases linearly.

In the flexible bandwidth case, the probability reflects the load of both CL and non-CL traffic. The reason is to ensure a 'fair balance' between the two classes, by rejecting CL session requests if non-CL demand is very high. Alternatively, if the number of queued non-CL packets is not included, then the admission of a CL microflow is independent of the amount of non-CL traffic.

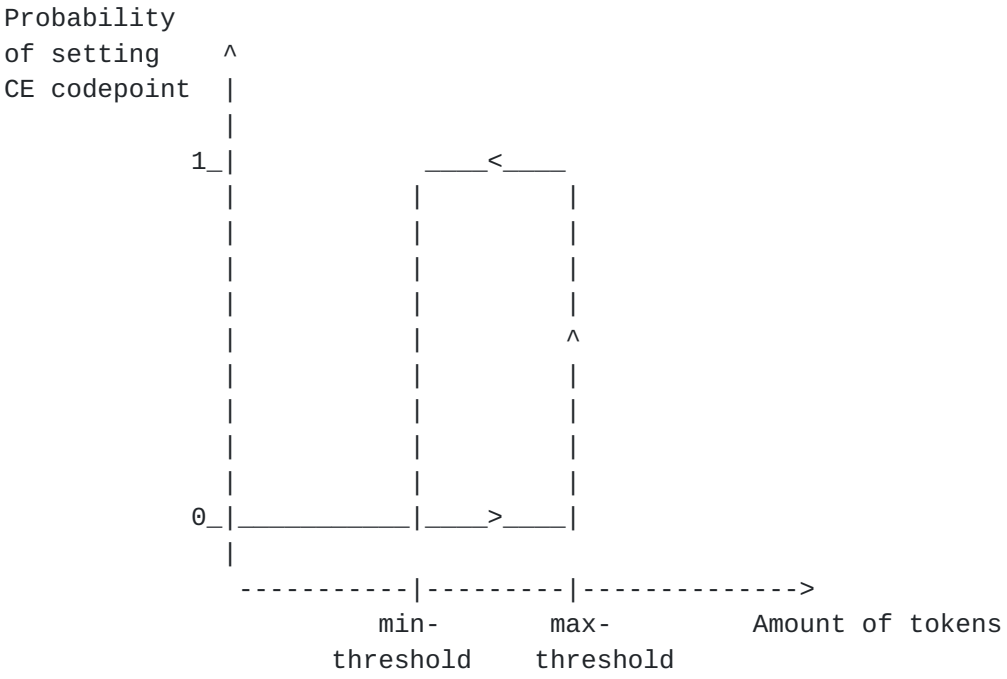

Figure 1: Setting Congestion Experienced Codepoint for CL-step-PHB

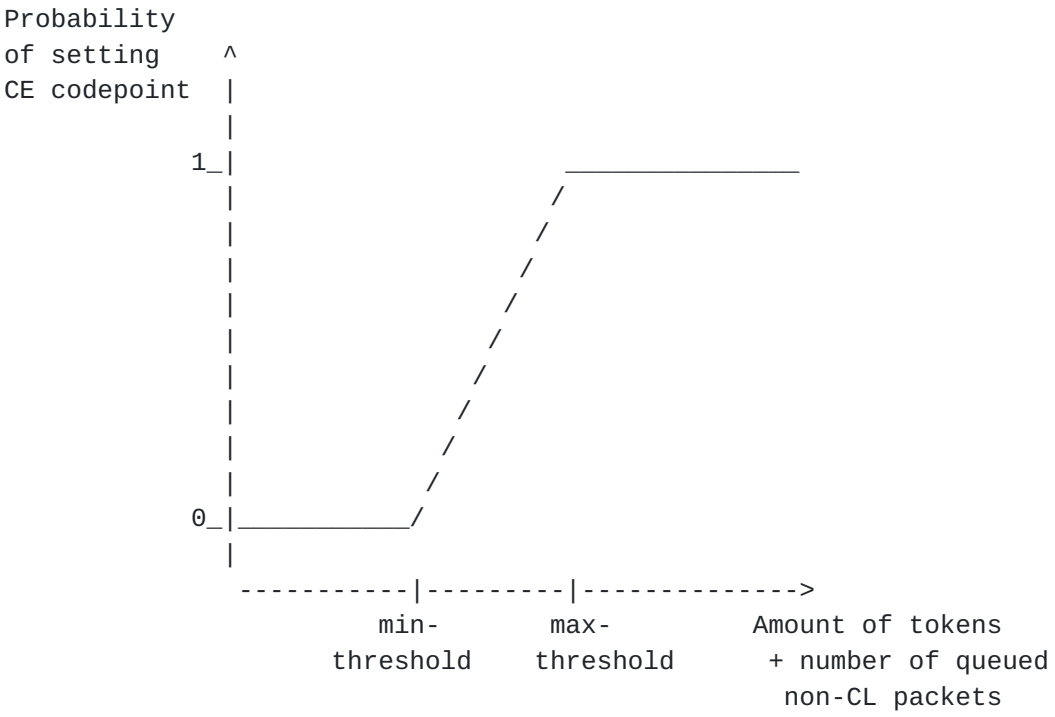

Figure 2: Setting Congestion Experienced Codepoint for CL-ramp-PHB

2.3. Scheduling

In the CL-ramp case with flexible bandwidth, CL packets are always scheduled ahead of non-CL ones, in order to minimise their delay and jitter, and FIFO (First In First Out) scheduling is used to prevent reordering within a CL microflow. This is needed because the arrival rate of CL packets is unknown. The fixed bandwidth case has more options, for example the CL-ramp and non-CL traffic could be serviced by a Weighted Round Robin scheduler.

2.4. Damage limitation

It is expected that setting the CE codepoint will enable appropriate action to be taken early enough to prevent significant congestion. However, in some circumstances (for instance if a node fails) it will not be sufficient. Therefore, if a CL PHB is implemented by a mechanism that allows unlimited pre-emption of other traffic (eg a

priority queue), the implementation SHOULD include some means to limit the damage CL traffic could inflict on other traffic.

2.5. Applicability of CL PHBs

It is suggested that the CL-step-PHB is defined for probing. One advantage of the CL-step-PHB is that it sets the CE codepoint immediately it notices there is potential congestion, so the admission control can react to a single CE packet.

It is suggested that the CL-ramp-PHB is defined for either data or probing. Advantages of the CL-ramp-PHB are that it can discriminate between different levels of congestion and accumulate the congestion signal across the network (so it can detect when several nodes are experiencing a low level of congestion such that a step algorithm would be 'off' in all nodes).

Use cases for the CL PHBs are briefly described in Section 3.

2.6. Interaction with other PHBs

Other PHBs and PHB groups may be deployed in the same DS node or domain with a CL PHB. CL traffic is prioritised over all other PHBs, but the admission control procedure prevents the CL traffic affecting their service.

It is believed that the CL-ramp-PHB configured and flexible bandwidth cases belong to the same PHB because they can both be used in the same domain, since the difference between them doesn't affect the way the ingress and egress nodes do admission control and metering.

It is believed that the CL-step-PHB and CL-ramp-PHB are different PHBs, because they cannot be used within the same domain, since the CE codepoint is interpreted differently. With the CL-step-PHB a single CE packet can cause the admission controller to block a new microflow, whereas with the CL-ramp-PHB sufficient packets are needed to estimate the congestion level. However, it would be possible to use CL-step-PHB for probe traffic and CL-ramp-PHB for data traffic within the same domain.

2.7. Interaction with default ECN behaviour

The CL PHBs define behaviour for setting the CE codepoint that is different from the default ECN marking behaviour [RFC3168]. To address the concerns of [Floyd] it therefore must be ensured that the packets only travel through nodes with a CL PHB. This could be done by signalling (to verify that the nodes on the path all understand

the CL PHB) or by configuration (eg in usage scenario 2 the whole domain runs the CL-ramp-PHB).

2.8. Mutability

CL packets with the ECN field cleared (`Not ECT') should be re-marked to Best Effort. In most scenarios, all CL packets should be ECN-capable, so it MAY be appropriate to raise a suitable management alarm the first time this is not the case.

In the edge-to-edge usage scenario, which is described below in Section 3.2, if some packets of a CL microflow are over the reservation at the ingress edge, then they SHOULD be marked to Best Effort.

2.9. Microflow misordering

Packets that belong to a single microflow within the CL behaviour aggregate passing through a device SHOULD NOT be re-ordered in normal operation of the device. In a microflow where packets are marked to Best Effort by the ingress node (see Mutability sub-section) they may become misordered, but note that these packets are never part of the CL behaviour aggregate.

2.10. Recommended codepoint for this PHB

Codepoint xxxxxx is RECOMMENDED for the CL-ramp-PHB and codepoint xxxxxx is RECOMMENDED for the CL-step-PHB. RECOMMENDED codepoints are needed, rather than EXP/LU ones, to enable host-to-host and inter-operator usage.

2.11. Tunnelling

When CL packets are tunnelled, the tunnelling packets SHOULD be marked as CL and the ECN field copied between headers as in RFC3168.

2.12. Security Considerations

TBD

2.13. IANA considerations

TBD.

3. Use cases

Two possible usage scenarios for the CL PHBs are discussed in this section:

- A host-to-host service, potentially through several domains

- An edge-to-edge service across a single domain or across cooperating domains.

3.1. Host-to-host service

This usage scenario is based on [RTECN-usage]. A source wanting to establish a real time inelastic microflow sends probe packets, which have their DSCP set to the value for the CL-step-PHB and the ECN field set to the ECT codepoint. Nodes en route set the CE codepoint if necessary as an "early warning" of potential congestion. The receiver informs the source about the value(s) of the ECN field of the probe packets, enabling the source to decide whether or not the new microflow is acceptable. To ease migration, [RTECN] suggests that to start with the mechanism is deployed in a limited number of nodes - those known to be points where the bandwidth is constrained.

3.2. Edge-to-edge service

This usage scenario is described more fully in [CL-arch]. The CL-ramp-PHB is used within a single domain. The ingress node to the domain sets the DSCP to the value for the CL-ramp-PHB and the ECN field set to the ECT codepoint. Nodes en route set the CE codepoint if necessary as an "early warning" of potential congestion. The egress node of the domain meters CL-ramp packets that have their CE codepoint set. It calculates the fraction of the total (CL-ramp) bits that are in CE packets. The calculation is done as an exponentially weighted moving average ('Congestion-Level-Estimate'). Congestion-Level-Estimate provides an estimate of how near the domain is getting to a load where the CL-ramp traffic will start suffering significant delays. Note that the metering and calculation are done separately for CL-ramp packets from each ingress router, because (as discussed in Section 1) there may be sufficient capacity on all the nodes on the path between one ingress node and a particular egress, but not from a second ingress.

In order to decide whether to admit a new real time inelastic microflow, the current value for Congestion-Level-Estimate is compared with a threshold value; if it is greater then the request is refused, otherwise it is accepted. It would be possible to have different service priorities (eg for emergency calls or important users) by the ingress having different thresholds for Congestion-Level-Estimate.

The details depend on the end-to-end signalling protocol, but for example with RSVP, Congestion-Level-Estimate is included as an opaque object within the RESV message. If the current value for Congestion-Level-Estimate is unknown (eg if there are no microflows at present between the relevant ingress and egress nodes), then probe packets are sent, from which the egress node can initialise its meter. These probe packets could use either the CL-ramp-PHB or the CL-step-PHB. It is also possible for several adjacent domains to cooperate. Then only the ingress and egress nodes of the combined region take part in the admission control procedure; border nodes within the combined region do not take part in signal processing or hold path state. The domains can even be run by different operators; in this case the border routers between operators only have to do bulk accounting - per microflow metering and policing is not needed [Briscoe].

4. Acknowledgements

We thank Joe Babiarz for very helpful discussion about this document and [RTECN].

This work evolved from the Guaranteed Stream Provider developed in the M3I project [GSPa, GSP-TR], which in turn was based on the theoretical work of Gibbens and Kelly [DCAC].

5. Comments solicited

Comments and questions are encouraged and very welcome. They can be sent to the Transport Area Working Group's mailing list, tsvwg@ietf.org, and/or to the authors (either individually or collectively at gqs@jungle.bt.co.uk).

6. References

A later version will distinguish normative and informative references.

[Briscoe] Bob Briscoe and Steve Rudkin, "Commercial Models for IP Quality of Service Interconnect", BT Technology Journal, Vol 23 No 2, April 2005.

[CL-arch] B. Briscoe, G. Corliano, P. Eardley, P. Hovell, A. Jacquet, D. Songhurst, 'An architecture for edge-to-edge controlled load service using distributed measurement-based admission control', draft-briscoe-tsvwg-cl-architecture-00.txt", (work in progress), July 2005

[DCAC] Richard J. Gibbens and Frank P. Kelly "Distributed connection acceptance control for a connectionless network", In: Proc. International Teletraffic Congress (ITC16), Edinburgh, pp. 941ù952 (1999).

[Floyd] S. Floyd, 'Specifying Alternate Semantics for the Explicit Congestion Notification (ECN) Field', draft-floyd-ecn-alternates-00.txt (work in progress), April 2005

[GSPa] Karsten (Ed.), Martin "GSP/ECN Technology \& Experiments", Deliverable: 15.3 PtIII, M3I Eu Vth Framework Project IST-1999-11429, URL: http://www.m3i.org/ (February, 2002) (superseded by [GSP- TR])

[GSP-TR] Martin Karsten and Jens Schmitt, "Admission Control Based on Packet Marking and Feedback Signalling ¡-- Mechanisms, Implementation and Experiments", TU-Darmstadt Technical Report TR-KOM-2002-03, URL: http://www.kom.e-technik.tu-darmstadt.de/publications/abstracts/KS02-5.html (May, 2002)

[Reid] ABD Reid, 'Economics and scalability of QoS solutions', BT Technology Journal, Vol 23 No 2, April 2005

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2211] J. Wroclawski, Specification of the Controlled-Load Network Element Service, September 1997

[RFC2309] Braden, B., et al., "Recommendations on Queue Management and Congestion Avoidance in the Internet", RFC 2309, April 1998.

[RFC2474] Nichols, K., Blake, S., Baker, F. and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998

[RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z. and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

[RFC2597] Heinanen, J., Baker, F., Weiss, W. and J. Wrocklawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.

[RFC3168] Ramakrishnan, K., Floyd, S. and D. Black "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

[RFC3246] B. Davie, A. Charny, J.C.R. Bennet, K. Benson, J.Y. Le Boudec, W. Courtney, S. Davari, V. Firoiu, D. Stiliadis, 'An Expedited Forwarding PHB (Per-Hop Behavior)', RFC 3246, March 2002.

[RTECN] Babiarz, J., Chan, K. and V. Firoiu, 'Congestion Notification Process for Real-Time Traffic', draft-babiarz-tsvwg-rtecn-03" Work in Progress, February 2005.

[RTECN-usage] Alexander, C., Ed., Babiarz, J. and J. Matthews, 'Admission Control Use Case for Real-time ECN, draft-alexander-rtecn-admission-control-use-case-00', Work in Progress, February 2005.

[vq] Costas Courcoubetis and Richard Weber "Buffer Overflow Asymptotics for a Switch Handling Many Traffic Sources" In: Journal Applied Probability 33 pp. 886--903 (1996).

Authors' Addresses

Bob Briscoe
BT Research
B54/77, Sirius House
Adastral Park
Martlesham Heath
Ipswich, Suffolk
IP5 3RE
United Kingdom
Email: bob.briscoe@bt.com

Dave Songhurst
BT Research
B54/69, Sirius House
Adastral Park
Martlesham Heath
Ipswich, Suffolk
IP5 3RE
United Kingdom
Email: dsonghurst@jungle.bt.co.uk

Philip Eardley
BT Research
B54/77, Sirius House
Adastral Park
Martlesham Heath
Ipswich, Suffolk
IP5 3RE
United Kingdom
Email: philip.eardley@bt.com

Peter Hovell
BT Research
B54/69, Sirius House
Adastral Park
Martlesham Heath
Ipswich, Suffolk
IP5 3RE
United Kingdom
Email: peter.hovell@bt.com

Gabriele Corliano
BT Research
B54/70, Sirius House
Adastral Park
Martlesham Heath
Ipswich, Suffolk
IP5 3RE
United Kingdom
Email: gabriele.2.corliano@bt.com

Arnaud Jacquet
BT Research
B54/70, Sirius House
Adastral Park
Martlesham Heath
Ipswich, Suffolk
IP5 3RE
United Kingdom
Email: arnaud.jacquet@bt.com

Intellectual Property Statement

Disclaimer of Validity

Copyright Statement