AN APPLICATION DELIVERY ARCHITECTURE
TO LOAD BALANCE IN A
MULTIHOMED L3 VPN SCENARIO

USING SEMANTIC ROUTING PRINCIPLES

Final Report of Internship of end of studies

TELECOM BRETAGNE – IFSIC – IRISA

OSCAR OLVERA IRIGOYEN

TELECOM PARIS TECH

SUPERVISORS: DANIEL KOFMAN AND DARIO ROSSI

Telecom Paris Tech
June 2009

Abstract

Optimal IP load balancing of business applications, onto a multihomed VPN/WAN, is an aim that cannot be accomplish by conventional routing. Selecting the best path for a specific traffic class goes beyond of just using IP reachable prefixes. In fact, this selection requires additional metadata, based on an accurate treatment of application flows, metrics of traffic, metrics of links states and a coherent map that matches business policies with traffic class policies.

In this work, we propose an Application Delivery Architecture for Load Balancing (ADABL) that allows optimizing the use of parent routes. ADABL uses three major tools: application classification, passive and active monitoring and traffic policing enforced by Business Objectives. We argue that these tools integrate three major functionalities of semantic routing and that are required for the Internet of the future. Consequently, ADABL makes possible to optimize the use of redundant links, to warranty that critical applications will be delivered accomplishing business policies and to enhance the quality of experience of end users.
Contents

Preamble ................................................................................................................................................ IV
1. Introduction ........................................................................................................................................... 1
2. Related Work ......................................................................................................................................... 3
3. The Proposed Application Delivery Architecture for Load Balancing (ADALB) ...................... 6
   3.1 Operation ............................................................................................................................................ 7
   3.2 Traffic Profiles and Policies Injection .............................................................................................. 10
      3.2.1 Traffic classes based on prefix .................................................................................................. 10
      3.2.2 Traffic classes based on applications ....................................................................................... 10
      3.2.3 Traffic Class Policies ................................................................................................................ 11
      3.2.4 Link Utilization Policies ........................................................................................................... 11
   3.3. Parent Routes Module .................................................................................................................... 11
   3.4. Passive and Active Monitoring ..................................................................................................... 12
      3.4.1 Passive Monitoring ................................................................................................................... 12
      3.4.2 Active monitoring ..................................................................................................................... 12
   3.5. Classifying and Marking Traffic Flows .......................................................................................... 13
      3.5.1 Traffic Flow Classification ........................................................................................................ 13
      3.5.2 Marking ................................................................................................................................... 14
   3.6. Best Path Selection, CAC, Scheduling and traffic shaping ......................................................... 15
      3.6.1 Best Path Selection .................................................................................................................. 15
      3.6.2 CAC ...................................................................................................................................... 16
      3.6.3 Scheduling .............................................................................................................................. 16
      3.6.4 Traffic Shaping ......................................................................................................................... 17
   3.7. Management of the policies and visibility ..................................................................................... 18
4. Operation scenarios ........................................................................................................................... 20
   4.1 The ADALB modules in the type topology .................................................................................... 20
   4.2. The ADALB operation steps in the type topology ....................................................................... 21
   4.3. The CAC operation in the type topology ..................................................................................... 22
5. Future Work .......................................................................................................................................... 23
6. Conclusions .......................................................................................................................................... 24
References ............................................................................................................................................... 25
Annex A. Gant Chart of the internship ................................................................................................. 29
Annex B. A functional architecture of Internet and Semantic Routing .............................................. 30
SR in the Internet Architecture Re-composition .................................................................................. 31
Challenges to cope the today Internet Architecture Stress ................................................................. 31
Approaches to Analyze the Internet Functional Architecture ............................................................ 31
SR in the Internet Architecture Re-composition .................................................................................. 32
Original Internet functionalities ............................................................................................................ 32
Ongoing Internet functionalities ........................................................................................................... 33
Annex C. Traffic Classes Performance metrics and Link Utilization Performance metrics ................. 36
Preamble

Among others, there are two general approaches to obtain results in the way of research. The first is to take a specific problem and to put in place analytical tools for looking a quantitative solution. The second is to take a framework problem and to apply an extensive (as possible) documental investigation to obtain a new qualitative solution or at least to settle the principles that govern this framework.

Both approaches are necessary. The first allows the progress of specific state of the art solutions. The second allows tracing a systemic view of the origin and possible application trends for those analytical and algorithmic solutions [ClarkMWCDP05]. This approach has views of more long term, but they are required “to abstract and isolate high level goals from low level actions to integrate and act on imperfect and conflicting information, and to learn from past actions to improve future performance…” [ClarkKPI03]

The base of this work is Semantic Routing\(^1\) and as there is a lack of references in the literature of this concept, we took initially the second approach in research, explained before, to intent formalizing it. Semantic Routing differs from conventional IP routing because besides to make use of the IP addresses to route datagrams, it exploits additional packet metadata. This additional metadata serves to take more intelligent routing decisions and includes: extensive traffic classification, traffic policies and link policies (using diverse traffic engineering metrics and policy metrics). This part of the research has been placed at the end of this document as the Annex B.

After formalizing the framework to explain Semantic Routing, we looked for a low level framework to implement its principles. In this way, we proposed an architecture to load balancing in an intelligent VPN environment based on application delivery. This architecture resulted modular and adaptable to solve other application delivery IP issues. This second framework occupies the main place in the work.

Traditional load balancing splits the traffic between different paths packet by packet and depending or not on certain delivery priority levels. However, to really optimize the use of the links in the WAN, additional functionalities are necessary. They are in one hand, the visibility of the state of the network and policies and on the other, the automatic policy enforcement (to deliver integrity of service).

These functionalities are part of the architecture of the Internet of the future. They are based on knowledge of the state of the network and also on the business goals of the organization where this network is working. For this reason, semantic routing in the infrastructure of the networks, taking advance of extra metadata, is well settle as a cornerstone for application delivery.

This research internship was divided in four phases: 1) the evolution of the Internet architecture, its relation with semantic routing and related work, 2) the consolidation of the proposed architecture for path optimization and allow intelligent load balancing, 3) the exploration of operation scenarios and 4) the preparation of the report and revisions of the work.

Annex 1 shows a Gant chart of how the internship was organized.

\(^1\) In this work, the concept of semantic routing is used in the sense of the Internet infrastructure at the layer 3 level, in contrast to the semantic routing used in peer to peer applications, see Annex A.
1. Introduction

Load balancing is essential in a reliable system to optimize the use of the redundant resources. However, the scheme of parallelism required for a system with elements organized in a distributed environment, is a difficult aim. In this way, the range of speeds and complexity of the processes inside of such system make difficult a design where all the redundant elements may be used fairly. This is the case of the today Wide Area Networks (WANs), where the reliability of conflictive distributed applications is put in place using multihoming and where the optimal use of all links is desired.

This paper approaches this problem, limiting its context to a Hub-and-Spoke topology and using Layer 3 Virtual Private Networks (L3-VPNs) see Figure 1.1. The traffic is classified and routed given policies of traffic while links and traffic are monitored passive and actively. This approach differs from classic routing mainly because, the algorithm that selects the best path reasons given: a fine grained classification of the traffic, the active monitoring of IP Service Level Agreements\(^2\) (IP-SLAs) over the paths, and the queuing of flows depending on business policies.

A Hub-and-spoke topology is arranged like a chariot wheel, in which all traffic moves along spokes connected to the hub at the center. In this topology, the spoke consolidates de access to the IT global resources such as: business applications, data storages, Internet services, PSTN network, etc. The existence of a global central point for the access of IT resources permits to simplify the management related processes and to reduce the capital and operational expenditures.

The L3-VPNs are WAN access services offered by Service Providers (SPs) that warrant the transmission of different classes of traffic. They are usually implemented over Multiprotocol Label Switched (MPLS) backbones. These MPLS backbones differ from legacy L2 backbones because instead to offer dedicated point to point links they offer virtual connections over a routed infrastructure. For this reason MPLS backbones can map Diffserv schemes (at the edge) to traffic engineering (TE) techniques (at the backbone). Of course the structure of the MPLS backbone would be transparent to the user of VPNs who only sees an access interface that offers SLAs for some classes of service.

In this context, the first tool to intelligently distribute the load over more than one WAN interface is the correct classification of traffic flows in wire speed. This is possible, implementing port-based and/or payload inspection. This paper assumes that the port or payload inspection must be done before encryption or tunneling.

The second tool is the passive and active monitoring of traffic and links. The passive monitoring may consolidate the counters of Management Information Base (MIB) agents of local interfaces. With this counters it is possible to define performance tables of ingress and egress of TCP flows. However, local information is not enough to characterize the end to end state of a path for L4 protocols without feedback, neither to monitor the metrics required by an SLA. For this reason,\(^2\) A service-level agreement (SLA) is a negotiated agreement between two parties where one is the customer and the other is the service provider. The SLA records a common understanding about services, priorities, responsibilities, guarantees and warranties. Each area of service scope should have the 'level of service' defined. The SLA may specify the levels of availability, serviceability, performance, operation, or other attributes of the service such as billing. The 'level of service' can also be specified as 'target' and 'minimum', which allows customers to informed what to expect (the minimum), whilst providing a measurable (average) target value that shows the level of organization performance. In some contracts penalties may be agreed in the case of non compliance of the SLA [WikiSLA].
active monitoring is implemented transmitting IP-SLA probes. These probes are packets of Internet Control Message Protocol (ICMP) that traverse the class traffic paths and test their respective performance conditions. The periodicity of passive and active monitoring depends on the reactivity of the reasoner engine to select the optimal path, assuming that it is necessary to minimize the possible churn effects of transitory instability of links and traffic.

The third and last tool for intelligent load balancing is the definition and enforcement of traffic policies given a business driven policy schema. This schema would be modeled given the business goals of the organization. They match the proportions of the link utilization and traffic engineering metrics with groups of applications. In this way a map between applications and the control plane of the network is defined to enforce the business goals over the network.

The rest of the paper is organized as follows: the second section shows the related work, the third section shows the operation and different modules of the proposed architecture, the fourth section shows the operation scenarios, the fifth section shows the future work, the sixth section shows the conclusions and finally Annex A, B, C, D and E are placed at the end of the paper.

Figure 1.1. Hub-and-Spoke topology using L3-VPNs
2. Related Work

The architecture proposed in this paper to load balance traffic classes is based on several previous networking frameworks. Some of those frameworks go from QoS networking, active networking, policy networking, service oriented networking and application oriented networking among others. All of them cope with performance issues derived from the today nature of Internet.

The actual Internet architecture has at the same time some functionalities that have allowed it to grow rapidly and others that also have contributed to this expansion but that are under stress. The first group includes: robust connectivity, infrastructure simplicity (at routing level), innovation at the edges and a high level of plasticity (at the end user level). While on the other hand the stressed functionalities include: addressing of low level of semantics, stateless of data and control network, and the philosophy of fair share resources.

Many authors have proposed to adapt or increase functionalities of today Internet architecture, see the annex B. When Internet was open to the commercial world, the number of users, applications and access bandwidth was in its first phases of development. For this reason, the routing was focused just in the reachability and the simplicity of connections. However today, new addressing schemes, flow-base traffic treatment, autonomocity (soft-state management), integrity of service, and support of trust are demanded.

Given this context, different architectures that adapt some of these new required functionalities have been proposed to re-dimension the face of Internet. There are propositions that suggest modifying the architecture of Internet from the core or from the edge.

To affect the backbone of Internet imply the agreement of ISPs for the reorientation of the architecture. However, the own interests of each ISP and the debate of the different actors in the Internet arena have been made difficult to achieve deep changes. For example, IPv6 offers stateless address autoconfiguration\(^3\), multicast\(^4\), mandatory network layer security\(^5\), header simplification\(^6\), mobility\(^7\) and so on. Conversely, the global adoption of IPV6 in December of 2008, after 10 years of being designated the successor of IPV4 by the IETF\(^8\), had reached just one percent.

Given the difficulties to make changes in the backbone of the Internet many of the architectural propositions have been inserted at the edge. This is the case of several optimization schemes applied to the enterprise or governmental scale. In these schemes most of the new desired functionalities have been included but the size of the network had been obviously limited. In this way, the problem of scalability for the deployment of a new architecture of Internet persists but some of the new functionalities have begun to be spread.

Active Networking [Smith99], [Tennenhause02], [Calvert06], [Maxemchuk01], QoS networking [CiscoQoS], [Gozdecki03], Policy-Based Networking [Nomura99], [IBM_PBN] [Balaji07], [Westerinen01], [Alb06], Service Oriented Networking and Application Oriented Networking [CiscoSONA], [Chen08], [Tian07] are similar approaches to deliver nearly network

\(^{3}\) hosts can configure themselves automatically
\(^{4}\) unlike IPv4, where it is optional
\(^{5}\) IPsec support is mandatory
\(^{6}\) the packet header in IPv6 is simpler, IPv6 routers do not perform fragmentation, etc.
\(^{7}\) unlike mobile IPv4, Mobile IPv6 (MIPv6) avoids triangular routing
\(^{8}\) Internet Engineering Task Force
functionalities. They all assume to classify the traffic as flows instead of datagrams, to apply traffic policies that match user requirements (services, applications, business objectives), and a degree of autonomicity given the knowledge of the state of policy metrics and flow and network performance metrics. These assumptions match with three of the five desired functionalities of internet: flow-based treatment, autonomicity and integrity of service. Latest versions of Active Networking deal with the problem of addressing/location dichotomy and with the problem of security as well. See annex B to detail the functionalities.

Naturally, the functionalities of these networking approaches lie on techniques of routing that use multiple constraints: QoS Routing [Orda05], [Kant08], Policy-based Routing [Nanda08] or Multi-constraint Routing [Bistarelli08], [Younis03], [Karaman06]. In this way, Semantic Routing can be seen as a synthesis that includes these different types of routing. Thus, Semantic Routing reacts to the emergency of the knowledge, derived from the states of performance flow metrics, performance link metrics and policy metrics.

Given the three networking functionalities identified previously, for semantic routing, it is possible to design a more detailed modular architecture to solve the problem of load balancing. This modular architecture will be detailed in the next section, however before; it is possible to mention some alternative approaches.

In [Chimmanee01], Chimmanee et. al. have proposed in the earlier 2000 a simple mechanism to balance three groups of classes of applications over an edge router running Interior Gateway Routing Protocol (IGRP). This router has a primary VPN interface and a secondary interface which is a modem connection. They assume that the packets of each application must be marked using the Type of Service (IP precedent byte) and that the VPN does not offer classes of services neither SLAs. The cost of each interface is computed given passive performance metrics at router interfaces (access bandwidth capacity (clock rate), packet loss and network delay). Then, each class of application is forwarded through the matched interface. This mechanism implements a rudimentary scheme of policy routing but it does not have an architecture that includes autonomicity for policy management, application classification (low and high levels of semantics), marking flows, scheduling, neither active monitoring.

In [Esaki05], [LeeNOMS08], [Albyrak08], the authors have proposed a load balance mechanisms over mobile routers with access to two different networks (WLAN and Wimax, or WLAN and 3G).

In [Esaki05] no application classification (flow recognition), nor traffic policies, but active monitoring for the available bandwidth is implemented. The traffic is dispatched packet by packet in the different interfaces using preferentially the interface with higher available bandwidth (round robbing is used when the paths are equally used and weighted round robin when the paths are used asymmetrically). In [LeeNOMS08], the previous schema is enhanced using flow policies given user, application and addressing information.

In [Albyrak08] an integral autonomic architecture, based on a middleware, is proposed. This architecture includes, among other details, an application treatment module, a context awareness module and a smart module (with sub-modules of policy, QoS and mobility). This architecture is capable of classifying the applications given low and high level of semantics (fine-grained classification), to translate policies of high level to traffic policies, to monitor the state of policies, performance flow applications and performance paths, to react dynamically to failure conditions and to apply self-protection. The decisions to select the best path and dispatch the applications flows are then reasoning by using the states of policies and flow-path performances.
In [CiscoWANoptSolGu08] and [CiscoOER_12_2sx_book09] two parallel architectures are proposed. The first one is called WAN optimization and the second one is called Optimization Edge Routing (OER). The components of these architectures are part of the global desired Internet functionalities included in semantic routing mentioned before (flow-based treatment, integrity of service and autonomicity).

The WAN optimization includes five components: Classification, Optimization, Monitoring, Control and Network Management. Given a WAN scenario, the application flows are classified and controlled automatically by enforcing traffic policies before being forwarded to the WAN. The traffic services are warranted by the passive and active monitoring of the performance states of policies, flows and paths. The path selection decisions are reasoned by using those states. In this way, it is possible to react dynamically to failure conditions. And finally, the visibility of those states is automatically synthesized by the use of a visibility manager.

The OER architecture proposes a performance loop with five phases: Profile, Measure, Apply Policy, Control and Verify. Given also an enterprise WAN scenario, flow applications are profiled given traffic policies and automatically classified as traffic classes before being routed and forwarded. The path selection decisions are reasoned by using the traffic policies and flow and path performances. In this way, it is possible to react dynamically to failure conditions. The integrity of services is also preserved monitoring passive and actively the states of flows and paths.

However these two architectures do not show the framework to map policies based on business objectives to traffic policies.
3. The Proposed Application Delivery Architecture for Load Balancing (ADALB)

The proposed Application-Delivery-Based for Load Balancing Architecture (ADALB) is the intersection of the defined functionalities of semantic routing (Annex B), the architectures proposed in [Albyrak08], [CiscoOER_12_2sx_book09], [CiscoWANoptSolGu08] and the Business Driven Management Framework proposed in [Alb06].

ADALB allows application route optimization. It is organized in eight functional modules for best path selection and to improve the user experience, see Figure 3.1 and Figure 3.2.

The manager module is composed by the visibility and policy sub-modules. The visibility sub-module is in charge of showing the topology of the network, the states of policies and performance of traffic classes and links. The policy sub-module permits to configure the traffic and link utilization policies; it must arbitrate the coherence between different policies. The visibility and policy manager together, allow to control and monitor the degree of coincidence among configuration policies, traffic performance and link utilization.

The database module is composed by the TC profile sub-submodule and the TC and LU policy sub-module. The TC profile sub-module contains the identifiers (names and equivalent ids that correspond to the applications running in the network, e. g., interactive, transactional, HTML, Differentiated Service Code Point codes (DSCP), Class of Service codes (CoS), etc.) The TC an LU policy sub-module contains the TE metrics (traffic distribution, delay, jitter, MOS etc.) that must be applied to each class of traffic.

A Configured Traffic Class (CTC) is characterized as an ensemble of the configured metrics:

$$CTC = TC \text{ profiles } \wedge TC \text{ policies } \wedge LU \text{ policies}$$

The parent route module obtains a classic routing table with the routes to the reachable networks. It can be an Interior Gateway Protocol (IGP) as Routing Information Protocol (RIP)
or Open Shortest Path First (OSPF), however it can be used Border Gateway Protocol (BGP) as well. This protocol must be running independently of the best path selection module.

The **classification module** implements static TCP/UDP port recognition, stateful TCP/UDP port recognition, peer to peer protocol recognition and VoIP recognition, on the incoming flows. The recognition of applications can be based on pattern techniques, behavioral techniques, state techniques or heuristic techniques, using Deep Packet Inspection (DPI).

The **remarking module** permits to add a DSCP value that matches each recognized application with the policies defined by the business driven policy scheme. This value will be used to schedule and forward the flows into the WAN.

The **monitor module** is composed by the passive and active sub-modules. The passive sub-module collects header information, volumetry counters, local jitter and local delay of the flows. While the active monitor sub-module collects the delay and jitter of traffic class paths using IPSLA probes. The SLA probes are based on ICMP packets.

A Monitored Traffic Class (MTC) is characterized, in a similar way that a CTC as an ensemble of monitored metrics that include:

\[
\text{MTC} = \text{TC profiles} \land \text{TC monitored policies} \land \text{LU monitored policies}
\]

The **reasoner engine** is composed by the CTC and MTC states sub-module, the Call Admission Control (CAC) sub-module and the best path selection sub-module. The CTC and MTC states sub-module consolidates the states of CTCs and MTCs defining a CTC vs MTC list\(^\text{12}\). This information is obtained from the monitor module. The CAC sub-module reacts to the congestion on the links. In the case of the interactive traffic, it is capable to block a prohibited number of interactive sessions to avoid interactive or voice traffic degradation derived from bottlenecks. The best path selection uses the information provided by the two previous sub-modules to select the optimal path for each traffic class.

The **scheduling and shaping modules** permit to control the priority and maximal throughput assigned to a traffic flow. The scheduling module orders the priority of each packet flow respect to the politics defined in the network. While the traffic shaping module limits the maximal throughput of each traffic class avoiding congestion\(^\text{13}\).

### 3.1 Operation

To explain the general operation ADALB an example will be used. Figure 1.1 shows a typical Hub-and-Spoke scenario. There are a headquarters (HQ) site and N possible branch office (BO) sites.

In the HQ, the global IT resources are consolidated. The Internet connection, the PSTN connection, the application servers and the data center offer different services and applications. The HQ also includes a robust LAN to switch and warranty the optimal transfer of traffic and to facilitate the management of the LAN access. To segregate and prioritize the traffic of users and critical applications the 802.1pq protocol (for virtual LAN tagging and CoS) mapped to DiffServ (for applications) can be implemented.

---

\(^\text{12}\) Monitored traffic classes change constantly given instantaneous variations in traffic and link conditions. Also Configured policies can change, given temporal or definitive modifications in the business requirements.

\(^\text{13}\) In the case of the interactive, not flexible traffic, to shape the traffic is not enough to avoid QoS degradation. If a link is used near of its point of saturation, all the sessions of traffic classes are prone to be queued and the delay and jitter could exceed the acceptable thresholds for the interactive application.
To simplify the example, only three large VLANs with growing priorities are defined. The first VLAN is used for the data users, the second for the users of interactive video and the third one for the VoIP users. To structure the resilient WAN, a master Policy Decision Point (PDP) router and two Policy Enforcement Point (PEP) routers as Costumer Edge routers (CE) are placed at the edge. Each CE is connected to a VPN-MPLS link that offers SLAs for traffic classes paths.

In the LAN of the BOs, the same schema of 802.1q/Diffserv, running in the HQ, is implemented. So, three VLANS with CoS are configured and the critical application packets are marked by a Diffsev code. In the edge of each BO is placed one CE router that includes PDP and PEP functions and that has two WAN interfaces. The interface connected to the ISP1 will be considered the primary interface while, the interface connected to the ISP2 will be considered the secondary one.

Now, the transfer of traffic classes will be explained using a general description of six steps. Thus, BO_A (Figure 1.1) will send three traffic classes to the HQ (the relation of traffic priorities follows that: TC1>TC2>TC3).

In the first step, the manager module injects the TC profiles TC policies and the LU polices to the edge routers. The process of distributing profiles and policies must be arbitrated by the policy submodule. This manager must have a group of coherent policies. For example, if a BO_A is going to be setup to warranty the 40% of TC1 of its WAN link capacity and BO_B will warranty just the 30% of its WAN link capacity for the same TC_1; then the manager would alert the network administrator. This will allow avoiding configuration conflicts. So, the policies must be configured in a coherent manner to warranty correct resource allocation.

In the second step, the Parent Route module, running in router BO_A, acquires a classic IP routing table. This table has all the possible routes to the reachable IP indexes. The traffic of the routing protocol (RIP, OSPF or BGP) must be transferred over the traffic class with the higher priority. In this example all the networks can be reachable by both interfaces of the BO_A router. (Seeing if the path update module will run over each different traffic class path or over)

In the third step, the SLAs correspondent to the traffic class must be monitored. The active monitoring of each traffic class is done by IPSLA probes using ICMP echo packets. This is accomplished using the edge routers of a WAN link as responders to the probes. In figure 1.1, to monitor the SLAs (e. g. voice traffic) that go from the BO_A to the HQ, the BO sends IPSLA probes that the HQ responds and, in the opposite way, the HQ sends IPSLA probes that the BO_A responds. If the performance metrics of CTC match with the performance metrics of the MTC, then the correspondent SLA is in policy.

The forth step consists in classifying and remarking the traffic that comes from the LAN. In the example, the traffic classes are identified using different traffic recognition techniques. After that, the traffic is remarked by the correspondent DSCP to enforce the policies given by the policy manager.

---

14 A virtual LAN, commonly known as a VLAN, is a group of hosts with a common set of requirements that communicate as if they were attached to the Broadcast domain, regardless of their physical location. A VLAN has the same attributes as a physical LAN, but it allows for end stations to be grouped together even if they are not located on the same network switch. Network reconfiguration can be done through software instead of physically relocating devices [WikiVLAN].

15 All policies are stored at the PDP. Whenever the PEP needs to make a decision, it sends all relevant information to the PDP. The PDP analyzes the information, makes the decision, and relays it to the PEP. The PEP then simply enforces the decision [WikiPDP].

16 Active monitoring is mandatory for critical non-best effort flows.
In the **fifth step**, as CTCs could be changed from the policy manager and the MTCs could be changed given the performance conditions of the WAN, then the CTCs MTCs correspondent to different flows are consolidated in a list by the states module. Therefore, the matching between the CTCs and MTCs states must give the information to select the optimal path. At the same time, critical applications that have no elastic traffic require limiting the number of sessions to avoid congestion and QoS degradation. Thus, the CAC module is in charge of limiting the number of flow sessions. Given the states of the MTCs and the admission of the correspondent flow session the flows are routed by the best path selection module.

Finally, in the **sixth step**, the information corresponded to CTCs and MTCs is synthesized to be showed via the visibility manager. Instantaneous measurements, statistics and reports can be obtained.

Next section details each architectural module.
3.2. Traffic Profiles and Policies Injection
The configured traffic classes CTCs can be profiled either automatically or manually.
The traffic class profiles can be divided in two general components: the traffic classes based on
prefix and the traffic classes based on application.

3.2.1. Traffic classes based on prefix
A range of prefixes can be configured by inserting an IP list. The IP prefix list will be matched
with the MTC list. An edge router can monitor and control an exact prefix of any length.\(^{17}\)

3.2.2. Traffic classes based on applications
The traffic profile sub-module allows the configuration of traffic for specific applications based
on metrics in the IP packet header, other than the Layer 3 destination address, see table 3.1.
DSCP values, port numbers, and protocols in addition to prefixes, are stored in the MTC list, see
table 3.2.

<table>
<thead>
<tr>
<th>TC profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Src. IP address</td>
</tr>
<tr>
<td>Dst. IP address</td>
</tr>
<tr>
<td>Src. Port</td>
</tr>
<tr>
<td>Dst. Port</td>
</tr>
<tr>
<td>L3 protocol type</td>
</tr>
<tr>
<td>ToS Byte (DSCP)</td>
</tr>
<tr>
<td>Logical input interface</td>
</tr>
</tbody>
</table>

Table 3.1. TC Profile metrics

<table>
<thead>
<tr>
<th>Name</th>
<th>DSCP</th>
<th>Protocol</th>
<th>Port Src</th>
<th>Port Dst</th>
<th>SrcPrefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>telnet</td>
<td>defa</td>
<td>tcp</td>
<td>23-23</td>
<td>1-65535</td>
<td>1.1.1.0/0</td>
</tr>
<tr>
<td>telnet</td>
<td>defa</td>
<td>tcp</td>
<td>1-65535</td>
<td>23-23</td>
<td>1.1.1.0/0</td>
</tr>
<tr>
<td>ftp</td>
<td>defa</td>
<td>tcp</td>
<td>21-21</td>
<td>1-65535</td>
<td>1.1.1.0/0</td>
</tr>
<tr>
<td>ftp</td>
<td>defa</td>
<td>tcp</td>
<td>1-65535</td>
<td>21-21</td>
<td>1.1.1.0/0</td>
</tr>
<tr>
<td>cuseeme</td>
<td>defa</td>
<td>tcp</td>
<td>7648-7648</td>
<td>1-65535</td>
<td>1.1.5.0/0</td>
</tr>
<tr>
<td>cuseeme</td>
<td>defa</td>
<td>tcp</td>
<td>1-65535</td>
<td>7648-7648</td>
<td>1.1.5.0/0</td>
</tr>
<tr>
<td>dhcp</td>
<td>defa</td>
<td>udp</td>
<td>68-68</td>
<td>67-67</td>
<td>1.1.1.0/0</td>
</tr>
<tr>
<td>dns</td>
<td>defa</td>
<td>tcp</td>
<td>53-53</td>
<td>1-65535</td>
<td>1.1.1.0/0</td>
</tr>
<tr>
<td>dns</td>
<td>defa</td>
<td>tcp</td>
<td>1-65535</td>
<td>53-53</td>
<td>1.1.1.0/0</td>
</tr>
</tbody>
</table>

Table 3.2. TC Profile List example

\(^{17}\) Although an internal public prefix can be manually configured, the edge router does not try to control an
interior prefix unless there is an exact match in the routing information base (RIB) of the topology update module, because the
edge router must not advertise a new prefix to the Internet.
3.2.3. Traffic Class Policies
Traffic class policies are a set of rules with thresholds of performance metrics for each CTC. The performance flow metrics that can be managed are: Reachability, Delay, Packet Loss, Jitter, Mean Opinion Score (MOS). See Annex C for the definition of these metrics.

3.2.4. Link Utilization Policies
Link utilization policies are a set of rules with thresholds of performance metrics of the links connected to the edge router. Link policies are concerned with the performance of the link as a whole. Three metrics can be considered by link policies: Link Utilization Distribution, Range, Cost. See Annex C for the definition of these metrics.

Next lines show a configuration example that includes traffic class profiles, traffic class policies and link utilization policies applied for voice traffic in the BO_A router.

Traffic class profiles

```plaintext
ip prefix-list CONFIG_TRAFFIC_CLASS permit 10.10.10.0/24
permit udp any range 16384 32767 10.10.10.0 dscp ef
```

Traffic class policies

```plaintext
VOICE_MAP 1
match ip address access-list VOICE_TRAFFIC_CLASS
set active-probe jitter 10.10.0.1 target-port 2000 codec g729a
set delay threshold 100 milliseconds
set loss relative 2%
set probe frequency 10 seconds
set jitter threshold 30 milliseconds
set mos threshold 4.0 percent 25
```

Link utilization policies

```plaintext
mode select-exit best
resolve delay priority 1
resolve loss priority 2
resolve range priority 3
resolve utilization priority 4
max range receive percent 35
border 10.10.10.1
maximum utilization receive 75
```

3.3. Parent Routes Module
Prefixes can both be learned dynamically and or configured statically. In whichever of these cases, a parent route is required. Parent routes are redundant routes that can be injected into the routing table by BGP, OSPF or static routes. The parent routes must therefore be of equal cost
and administrative distance so that more than one path for the parent route exists in the routing table of the edge router at the same time\textsuperscript{18}.

The edge router routes TCs using these equal cost parent routes out its external interfaces. DSCP values, protocols, and port numbers are used to optimize the routes in the Route Information Base of the edge router. Then, the routes can be changed depending on the states of links. See the next example.

**BGP enabled**

- Discovered Exit for prefix 10.1.1.0/24 → 10.10.10.1, int LANeth
- Route changed 10.1.1.0/24, 10.10.10.1
- Reason Delay, OOP

### 3.4. Passive and Active Monitoring

#### 3.4.1 Passive Monitoring

Passive monitoring allows measuring TCP traffic flows. These flows can be characterized by: Delay Time (between TCP SYNC and TCP SYNC/ACK in a TCP three-way handshake), Loss TCP sequence numbers, Reachability (Repeated TCP SYNCS without an accompanying TCP SYNC/ACK) and Throughput.

Passive monitoring allows measuring only throughput for non-TCP traffic flows. Table 3.3 shows the metrics that characterize a TC flow to use passive monitoring. One or more of those metrics can be absent. To see the description of TC and LU policies see Annex C.

<table>
<thead>
<tr>
<th>TC Profile</th>
<th>TC Policies</th>
<th>LU Policies</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC name</td>
<td>Reachability</td>
<td>TC required available BW by Link</td>
</tr>
<tr>
<td>Src. IP address</td>
<td>Delay</td>
<td>-</td>
</tr>
<tr>
<td>Dst. IP address</td>
<td>Loss (packets per million)</td>
<td>-</td>
</tr>
<tr>
<td>Src. Port</td>
<td>Delay (For ingress interface)</td>
<td>-</td>
</tr>
<tr>
<td>Det. Port</td>
<td>Jitter (For ingress interface)</td>
<td>-</td>
</tr>
<tr>
<td>L3 protocol type</td>
<td>Throughput</td>
<td>-</td>
</tr>
<tr>
<td>ToS Byte (DSCP)</td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3.3. The metrics that characterize a TC flow to use passive monitoring

#### 3.4.2 Active monitoring

To monitor actively the parent routes for UDP and real time voice or video traffic flows or L4 traffic without ACKs service level agreement (SLA) probes can be generated by the edge routers.

\textsuperscript{18} “Caution must be applied when redistributing static routes into an IGP. The injected routes may be more specific than routes in the IGP, and it will appear as if the edge router is originating these routes. To avoid routing loops, the redistributed static routes should never be advertised over a WAN by an edge router. Route filtering and stub network configuration can be used to prevent advertising the static routes. If the static routes are redistributed to routers terminating the external interfaces, routing loops may occur” [CiscoOER_12_2sx_book09].
and transmitted at a probe frequency value. An active probe is of the type of ICMP echo. For example, if VoIP traffic-class will be monitored:

\[
\text{set active-probe jitter} \\
\text{IP target 10.10.10.1 target-port 33033 codec g729a} \\
\text{probe frequency 2}
\]

In this example, a target IP address 10.10.10.1 is configured. It might be another edge router configured as ip sla responder. Most IP hosts can respond to an ICMP echo, unless administratively disabled or prohibited, nevertheless to resolve delay, jitter and MOS, the capabilities and function of an IP SLA responder must be defined.

The insertion of active probing furthermore has disadvantages. The ICMP ECHO requests that are generated by default constitute supplementary traffic on the network. Activating probing on the Internet may not be desirable. These ICMP packets may be blocked or administratively prohibited and may be considered a threatening or abusive posture to the target hosts. For this reason, active monitoring is best suited for use within the private internal network of the enterprise.

Table 3.4 shows the metrics that characterize a TC flow to use active monitoring. One or more of those metrics can be absent. To see the description of TC and LU policies see Annex C.

<table>
<thead>
<tr>
<th>TC profile</th>
<th>TC policies</th>
<th>LU policies</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC name</td>
<td>Reachability</td>
<td>TC required available BW by Link</td>
</tr>
<tr>
<td>Src. IP address</td>
<td>Delay</td>
<td>-</td>
</tr>
<tr>
<td>Dst. IP address</td>
<td>Loss (packets per million)</td>
<td>-</td>
</tr>
<tr>
<td>Src. Port</td>
<td>Delay (For ingress interface)</td>
<td>-</td>
</tr>
<tr>
<td>Dst. Port</td>
<td>Jitter (For ingress interface)</td>
<td>-</td>
</tr>
<tr>
<td>L3 protocol type</td>
<td>Peak Information Rate</td>
<td>-</td>
</tr>
<tr>
<td>ToS Byte (DSCP)</td>
<td>MOS</td>
<td>-</td>
</tr>
<tr>
<td>Logical input interface</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3.4. The metrics that characterize a TC flow to use active monitoring

### 3.5. Classifying and Marking Traffic Flows

#### 3.5.1. Traffic Flow Classification

Classifying packets enables differentiation for distinctive type of traffic. This classification can be associated with the TC profiles. This work assumes that the application classification is applied before tunneling or encryption.

The Criteria of Classification can include:
- Static Ports TCP and UDP (Low Level Syntax Complexity)
- Sub-port classification based on DPI (High Level Syntax Complexity)
- HTTP Traffic by URL, Host, or MIME for instance:
  - match protocol http url /latest/whatsnew.html
  - match protocol http c-header-field "somebody@telecom-paristech.fr"
- Real-time audio and video traffic complete identification
  - Citrix ICA Traffic by Published Application or Tag numbers
  - RTP Payload Type Classification: Audio and Video codecs
  - RTP Payload Type Classification
    - match protocol fasttrack file-transfer "*.mpeg"

3.5.2. Marking

To marking TCs, a general DiffServ principle is to mark or trust traffic as close to the source as administratively and technically possible. However, certain traffic types might need to be re-marked before handoff to the service provider to gain admission to the correct class. If such re-marking is required, it is recommended that the re-marking be performed at the CE’s egress edge. This is because of the internal complexity of the ISPs, it will be easier to manage if re-marking is performed only at the CE egress edge.

Additionally, in some cases, multiple types of traffic are required to be marked to the same DiffServ code point value to gain admission to the appropriate queue. For example, on high-speed links, it might be desired to send Voice, Interactive-Video, and Call-Signaling to the service provider’s Real-Time class. If this service-provider class admits only DSCP EF and CS5, two of these three applications would be required to share a common code point. See the next example; both Interactive-Video and Call-Signaling are re-marked to share DSCP CS5.

```
profile-map CE-INGRESS-EDGE
  class-map match-any VOICE
  match ip dscp ef
  class-map match-all INTERACTIVE-VIDEO
  match ip dscp af41
  class-map match-any CALL-SIGNALING
  match ip dscp af31
  match ip dscp cs3

profile-map CE-EGRESS-EDGE
  class VOICE
  priority percent 18
  class INTERACTIVE-VIDEO
  priority percent 15
  set ip dscp cs5 (Interactive-Video is remarked to CS5)
  class CALL-SIGNALING
  priority percent 2
  set ip dscp cs5 (Call-Signaling is also remarked to CS5)
```
3.6. Best Path Selection, CAC, Scheduling and traffic shaping

3.6.1. Best Path Selection

Edge routers select the best path based on MTC states see Figure 3.3:
Excluding links currently overloaded
Best MTC performance (the MTC state includes the failure conditions)

3.6.1.1. CTC-MTC State Sub-module

Figure 3.3. Flow chart of the processes to obtain the MTC states

Processes to obtaining the MTC states:

1. CTC Injection and activate the **periodic timer**
2. Passive monitoring → if MTC=CTC → in Policy → Forwarding
3. If MTC ≠ CTC and if alternate path is Reachable → send IPSLA probes and active **Hold-down timer**
   3.1. If MTC = CTC → In Policy → Forwarding
   3.2. If MTC ≠ CTC → send IPSLA probes and activate the **back-off timer**
      3.2.1. If the TC is correctly configured → OOP (out of policy) → Forwarding
      3.2.2. If the TC is not in control → Default → Forwarding
4. After back-off timer expires repeat from 3
5. After periodic timer expires repeat from 1

The three timers required to obtaining the MTC states varying depending on the degree of monitoring overhead desired into the network and on the possible failure scenarios. The periodic
timer determines the mandatory period of repetition to send IPSLA probes if the MTCs are in policy. If the MTCs change to the state of OOP and there is a parent route (redundant route by a secondary interface), the hold-down timer is activated until the parent route is confirmed in policy. If the parent route is not in policy or there is an unconfiguration of configuration, the back-off timer is activated. After the back-off timer expires, the IPSLA probes are resend. Each time that the back-off timer is activated consecutively the back-off timer increases to reduce the monitoring overhead in case of a permanent damage in the links. An example of the states of links is showed in Table 3.5.

<table>
<thead>
<tr>
<th>External Interface</th>
<th>Interface Capacity (kbps)</th>
<th>Max BW (kbps)</th>
<th>BW Used (kbps)</th>
<th>Load (%)</th>
<th>Status</th>
<th>Exit Id</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>256</td>
<td>204</td>
<td>74</td>
<td>28</td>
<td>UP</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>256</td>
<td>192</td>
<td>67</td>
<td>25</td>
<td>UP</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.5. Link States

3.6.2. CAC

If multiple applications try to reserve bandwidth for a specific queue type and the reservation cannot be met, the network denies the reservation, which prevents the degradation of the application experience and protects existing reservations.

To admit a new TC, the Peak Information Rate (PIR) of the TCT (PIR_TCT) must be less than the available bandwidth of the MTC (Av-Bw_MTC). If this condition is accomplished the edge router will take the optimal path for the TC and will schedule and forward the packets of this flow. If the condition is not accomplished the request will be rejected and a warning message will be sent to the visibility manager. As the request for the connection will not be responded, the application terminal will send to the user a message of unreachability. See the flow chart of the CAC process in Figure 3.4.

Figure 3.4. Flow chart of the CAC process

3.6.3. Scheduling

Queuing algorithms are used to sort and prioritize traffic overflow before packets are transmitted onto the network. Packets are scheduled for transmission according to their CTC and queuing mechanism applied to the WAN interface. There are different types of queuing mechanisms such as First-In-First-Out (FIFO) or Weighted-Fair-Queuing (WFQ). Traffic flows classified as expedite forwarding will be placed in the high priority queue and transmitted first. Traffic flows requiring a certain amount of bandwidth will be sorted and placed on different queues before being transmitted. In times of congestion, some of these queues will overflow and excess traffic will be dropped.
Selective packet drop can be implemented with Weighted Random Early Detection (WRED). By assigning a drop probability to a traffic class, it is possible to select which traffic class will drop packets at the time of congestion. In general, a higher drop probability is given to traffic classes considered to be of lower priority.

Tail drop is used for WFQ traffic classes unless a WRED configuration would be used to drop packets as a means of avoiding congestion.

The following example configures WRED to use the DSCP value af11. The minimum threshold for the DSCP value af11 is 24 packets (for the queue) and the maximum threshold is 40 packets (for the queue). Table 3.5 shows the matching between WRED and DSCP values. This table also shows the probability to drop packets and the current counters for transmitted packets, and random drops and tail drops.

```
set wred dscp-based
  wred dscp af11 min pkts 24 max pkts 40
  wred dscp af12 min pkts 28 max pkts 40
  wred dscp af13 min pkts 24 max pkts 40
```

<table>
<thead>
<tr>
<th>dscp</th>
<th>Transmitted pkts/bytes</th>
<th>Random drop pkts/bytes</th>
<th>Tail drop thresh</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mark WRED prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>af11</td>
<td>0/0</td>
<td>2/100</td>
<td>1/100</td>
<td>24</td>
<td>40</td>
<td>1/10</td>
</tr>
<tr>
<td>af12</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>28</td>
<td>40</td>
<td>1/10</td>
</tr>
<tr>
<td>af13</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>24</td>
<td>40</td>
<td>1/10</td>
</tr>
</tbody>
</table>

Table 3.5. WRED-DSCP queue configuration example

### 3.6.4. Traffic Shaping

Traffic shaping limits the rate of transmission of data. The data transfer can be limited to one of the following:

- A specific rate
- A derived rate based on the level of congestion

The rate of transfer depends on three components that constitute a token bucket: burst size (also called the Committed Burst (Bcz)), mean rate, time (measurement) interval (Tc). The mean rate is equal to the burst size divided by the time interval.

When traffic shaping is active, the bit rate of the interface will not exceed the mean rate over any integral multiple of the time interval. It means, for the period of every time interval, a maximum of burst size can be sent. Inside the interval, nevertheless, the bit rate may be faster than the mean rate at any certain time. One extra variable acts in traffic shaping: excess burst size (Bez). The Bez corresponds to the number of noncommitted bits that are still accepted.

Packets corresponding the specified criteria are placed in the token bucket. The maximum size of the token bucket is the Bcz plus the Bez. The token bucket is filled at a constant rate of Bcz worth of tokens at every Tc. This is the configured traffic shaping rate.
If the traffic shaping mechanism is active\textsuperscript{19} at every $T_c$, the traffic shaper checks to see if the transmission queue contains enough packets to send\textsuperscript{20}.

- If there are enough tokens in the token bucket, the packet is sent (transmitted).
- If there are not enough tokens in the token bucket, the packet is placed in a shaping queue for transmission at a later time.

See a graphical description of the token bucket in figure 3.5.

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{token_bucket.png}
\caption{Token bucket}
\end{figure}

**TC INGRESS PER CLASS**
- class REAL\_TIME
  - set ip dscp cs5
- class CALL\_SETUP
  - set ip dscp cs5
- class STREAMING\_VIDEO
  - set ip dscp cs5
- class TRANSACTIONAL\_DATA
  - set ip dscp cs3
- class NETWORK\_MANAGEMENT
  - set ip dscp cs3
- class BULK\_DATA
  - set ip dscp af21

**SHAPING map EGRESS PER CLASS**
- class GOLD
  - shape average 460800
- class SILVER
  - shape average 768000
- class class\_default
  - shape average 1536000

3.7. Management of the policies and visibility

Simple and intuitive configuration interfaces, satisfactory monitoring, reporting, and troubleshooting are crucial elements of successful network optimization. The value of these features rise above functional teams, implemented technology and business managers with the information needed to enhance productivity. In this way, it is possible enabling IT and network

\textsuperscript{19} It means that, packets exceeding the configured traffic shaping rate already exist in a transmission queue

\textsuperscript{20} It means that, up to either $B_c$ (or $B_c$ plus $B_e$) worth of traffic
architecture groups to project the business goals; and reduce time spent solving network management problems.

ADALB uses the policy and visibility manager in order to:

1. Profiling traffic classes
2. Configuring traffic policies and link policies
3. Monitoring the states of TCs
4. Analyzing the performance of the network
5. Simplifying configuration processes

Annex D shows the Business Driven Management Framework referred in this work to map business goals to networking policies.
4. Operation scenarios

4.1 The ADALB modules in the type topology

Figure 4.1 shows the typical scenario for deploying ADALB. A hub-and-spoke topology is organized with the headquarters as hub and the branch offices as spokes. In the respective LANs, the segregation of traffic classes is accomplish given a Diffserv scheme for critical applications that are mapped over the protocols 802.1pq. The edge routers have functions of PDP/PEP and two interfaces to two independent SPs.

See in Figure 4.1 how the eight functional modules are disposed at the edge router interfaces.

Figure 4.1. Eight functional modules at the edge router interfaces
4.2. The ADALB operation steps in the type topology

Six main steps to load balance are defined given ADALB. These steps are:

1) the injection of CTC, 2) the acquiring of parent routes, 3) the passive and active monitoring of TCs in order to obtain the MTC states, 4) the classification and remarking of application flows, 5) the path selection, scheduling and traffic shaping and finally 6) management of policies and the monitoring of states in the network.

Figure 4.2 shows the ADALB processes in the type topology, for a description of the process see section 3.1.

![Figure 4.2. The ADALB processes in the type topology](image-url)
4.3. The CAC operation in the type topology

One of the particular processes implemented in ADALB is the Call Admission Control. It avoids the degradation of real time interactive applications before congestion. In figure 4.3 four interactive video sessions are being transmitted. Only 10% of bandwidth remains in each ingress parent route of BO_A. As for each TC1 session is required 30% of the total link bandwidth, the edge routers in the HQ site block a fifth session. In this case the Peak Information Rate (PIR) of the TCT (PIR_TCT1) is greater than the available bandwidth of the MTC (Av-Bw_MTC1) and the required condition to accept the call is not accomplished. As the request for the connection will not be responded, the application terminal will send to the user a message of unreachability.

Figure 4.3. The Call Admission Control in ADALB (in the fifth TC1 session PIR_TCT1 > Av-Bw_MTC1)
5. Future Work

Even if related architectures can be implemented using current technologies, a lack of performance studies is evident. Some of the global variables that can strongly influence the performance of ADALB to optimize the use of parent routes to load balancing can be:

- The algorithms (using timers and signalization) and measuring techniques that constitutes the monitoring module
- The assumptions to defining policies of high level (aligned with business goals)
- The systems to automate the translation of high level policies to low level policies
- The best path selection algorithms
- The Call Admission Control algorithms

By other side, as ADALB is a modular architecture, it can be adapted to solve. The extension of ADALB for fulmesh topologies (scalability of ADALB).
6. Conclusions

ADALB is an architecture based on Semantic Routing functionalities. It permits the fine-grained regulation of application flows. This regulation enforces automatically business policies to traffic policies. It is situated in the edge of the ISPs and optimizes the use of L3 VPN/WAN links in hub-and-spoke topologies. In this manner, it balances application flows depending on the traffic class states. The traffic class states are characterized by a group of metrics of rich semantics such as: IP prefixes, L4 ports, sub-port meanings, flow traffic metrics and link states metrics. These metrics are obtained using traffic classification tools, monitoring tools and traffic policy enforcement tools.

Given that traffic policies are enforced by the business goals, (SLAs are supervised and proactive decisions are taken to select optimal WAN exits) the quality of experience of end users would be enhanced. However no quantitative evidence of the performance of architectures like this has been reported, for this reason a number of analytical, simulation and experimental studies are required. As this modular architecture coincides with major functionalities demanded for emergent IP services and technologies, we argue that these further studies have a fertile field to be pursued.
References


Annex A. Gant chart of the timing in the internship
Annex B. A functional architecture of Internet and Semantic Routing

The pitfalls of the current Internet architecture impede its own natural evolution [Ratnasamy05]. Location/identification dichotomy, mobility addressing restrictions, overhead middle boxes, QoS/Policy end-to-end issues, overhead manual configuration in network devices, etc. are some of the deficiencies that demand a functional re-composition or the redirection (redefinition of the IP protocol) of the internet architecture and the correspondent attention to routing functions as catalytic elements.

There exist a number of propositions to put in place these re-compositions or redirections and to cope with one or many of these issues [ClarkAR03] [Kempf04] [Ratnasamy05]. They go from a new version of the IP protocol (e.g. IPv6) or to the protocolary stack redefinition (e.g. Autonomic Network Architecture project) until underlay and overlay mechanisms to get QoS, addressing versatility, policy management or security functionalities (e.g. Diffserv, Inserv, IP-sec, NAT, SOA, MPLS). However, due to technological and market issues, most of these propositions have been not extensively deployed and most of the pitfalls remain intrinsic to the Internet Infrastructure. In the first part of this annex, a number of references is explored to explain the today Internet architecture and its contrast with the new functional requirements. In this way, the Semantic Routing functions are derived to recompose these architectural discrepancies without initially changing the IP protocol.

From a global view, we defined Semantic Routing (SR) as a soft-state component for the regulation of flows in Internet. It is the result of the constant evolution of Internet and where it is possible, to distinguish four main motivators for its development.

The initial two motivators are the requirements to optimize the resources at network layer and to alleviate the today IP architectural stress. Meanwhile, the last two motivators are the reduction of the price in the access for the transmission per bit, on the client side, and the reduction of the CAPEX-OPEX in the infrastructure on the ISP side, which will drive the emergency of add value and new pervasive services in a competitive market [ClarkMWCDP05].

The initial two motivators are the base of this work and will be used to explain a) the roll of SR in the architectural re-composition of Internet, b) to show some of the SR applications, c) the SR main architectural functions and d) the challenges for its implementation. The last two motivators will not be analyzed but are mentioned as natural prerequisites of the market environment that set in motion the growing and distribution of services over the Internet.

Even, if there are contrary opinions in regards to where the functional re-composition or redirection should be initially placed [Ratnasamy05] (from the core or from the edge) we argue that SR can be progressively adapted on the edge layer, on the distribution layer or core layer routers, extending new architectural functionalities, to finally upgrade a more homogeneous and may be standard architecture. It is evident that years of debate in regards to the evolution of the Internet architecture [Clark88] [Carpenter96] [Kempf04] cannot be easily simplified, however we

---

21 Soft-state means the ability to have enough information (traffic metrics, topology, etc.) to characterize the state of the data and control plane of the network.

22 as the capacity to infer paths given implicit and/or explicit context predicates in a networking environment
think that it is possible to define the main functions that the SR should be accomplish in order to be a versatile change agent at the network level.

**SR in the Internet Architecture Re-composition**

**Challenges to cope the today Internet Architecture Stress**

- Addressing (identification vs. location)
- Mobility
- Enabling flow treatment in contrast with datagram treatment
- Empowering the management systems by the application of policies over a knowledge coherent plane
- Reduction of human error configuration over the ISP infrastructures by the implementation self-configuration, self discovery and self-repairing platforms
- Empowering the pricing by bit
- Explosion of new competitive ISP business models
- Simplification of the QoS mechanisms to get global performance functionality
- Simplifying the anycast and multicast mechanisms over inter-domine ASs
- Simplification of the security management, defining trust zones by using flow, application and service aware capabilities

**Approaches to Analyze the Internet Functional Architecture**

The analysis of the Internet architecture is not a simple aim. Its history shows an organic evolution, where the functions and structure of this complex network have been changed progressively [Postel81] [Clark88] [Carpenter96] [Kempf04] [Ratnasamy05]. Two of the key points of the success of Internet are the local implementation flexibility together with the scalability of the IP protocol, where the fast expansion of nodes and services result natural consequences. In spite of this desired growing capacity, the local implementation flexibility of IP is also the origin of the absence of a strict global architecture. This fact makes Internet not a static well-known artifact but an evolutionary complex network. It has a structure and functions that tend to have a complex behavior [Strogatz01] [Braha03].

To approximate the functional architecture of internet we have considered three approaches. The initial approach points to the Internet router topology as a static ensemble of random networks in the space, which has a high degree of redundant edges between the nodes [Newman08]. This approach comes from the graph theory and explains the robustness and reliability of Internet as main properties and as functions of the average node degree in the graph [Xu01].

Secondly, the networking approach, that points to the technical functionalities of Internet. It comes from the technological, engineering and economic optimized tradeoffs. It is a synthesis of the views of experts involved in the historical conception and development of the internet standards [Carpenter96] [Clark88]. It gives the general functions of the IP protocol [Postel81] and the goals of Internet as a technological product.

And the third approach, that points to the statistical distribution of the nodes, showing Internet as an evolutionary network with properties of scale-free, high clustering degree and short average
distance between every two nodes (small world phenomena) [Newman08]. This approach comes from the Complex Networks Theory [Newman03] and permits to explain the degree distribution of the Internet as a result of the relationships between optimization principles, growth rules and topology [Strogatz01] [Braha03]. In this way, it is possible to show that real complex networks do not follow random degree distributions and may be show a Robust Yet Fragile (RYF) characteristic, given its growing speed and its preferential attachment property [Barabási03]. However, Internet, as a technological network, can be also characterized as a High Optimized/Organized Tolerance Tradeoffs (HOT) network, considering the optimization tradeoffs of its technological structure [Doyle05].

In Figure 1 it is showed the interaction of the three approaches to analyze the Internet architecture.

SR in the Internet Architecture Re-composition
Contrasting the three approaches for analyzing the Internet Architecture, it is possible to differentiate the original functionalities of Internet respect to the new required ones and to introduce the SR as a change indirection agent, see figure 2.

Original Internet functionalities
In this annex, we had divided the original Internet functionalities in three groups, see the left column of figure 2. The first is a group of four most general functions, which is stable and consequence of the “end-to-end principle” [Kempf04]. From this group, the initial two functionalities (explained from the graph theory) are connectivity robustness, conferred by the average redundancy path degree between two nodes (routers) in the Internet graph [Xu01]; and the cost-effective nature, conferred by its simple and effective topological expansibility. The last two

---

23 As these functionalities are on the very base of the adaptability and the expansion dynamics of Internet, they must remain, even if the initial phases of the architecture re-composition represent investments of not immediate reward.

24 Essentially the end-to-end principle suggests that specific application functions preferentially do not be built into the lower levels of the system, [Blumental01] or in other words that it is better do not to mixture protocol layer functions [Saltzer84].
functionalities of this group are the requirement to let the innovation at the edges and the Plasticity to accept a large variety of protocols and applications. Where, they are essentially conferred by following “proper layering” [Saltzer84] and by the IP capacity to be adapted to different datagram sizes (fragmentation and reassembly).

In the second group, it is feasible to distinguish another four functionalities that are essentially explained by the networking approach. Firstly, the addressed host to host datagram transmission functionality, which is conferred by using the classical IP addressing and the store and forward discipline. Secondly, the datagram unit base transmission functionality, which is conferred by the transmission of individual datagram as separated entities. Thirdly, the state-less functionality, which is a consequence of taking routing local decisions to route the datagrams, or in other words, because the nodal autonomy of the distributed routing algorithms. (Each router infers paths given local or partial global information). And finally, the fair resource sharing functionality is result of the absence of QoS and Policy constraints in the initial deployment of Internet. we identify this group as the functionalities that suffer stress in the Internet architecture.

The third and last group includes an intuitive property of node aggregation: a random aggregation property at router level. It means that intuitively it is possible to believe that the extensibility of the Internet topology follows a random pattern (routers are aggregated randomly over the entire network).

**Ongoing Internet functionalities**

In this work, we had matched the ongoing functionalities of Internet (result of its evolution) with these original ones that suffer stress. For this reason, the ongoing functionalities should partially or completely substitute these stressed ones. we have divided the ongoing Internet functionalities in two groups; see the right column of figure 2. The first group is integrated by five networking functionalities as follows: location/identification resolution, flow treatment, soft-state, integrity of service, and support of trust. And the second group contents only a statistical approach Internet property that we classified as: the evolutionary global functionality.

The location/identification resolution is a large issue, where its solution requires the definition of an addressing schema with a number of not easy tradeoffs to compromise. These tradeoffs may include a) simple implementation, b) adaptability to different potential contexts, c) address range sufficiency, to cope with an explosive introduction of new nodes and d) possible to standardize. In spite of the complexity of a coherent addressing schema, the SR can be introduced to manage packets that transport supplementary information and that differ with traditional IPv4 or IPv6 headers. The SR routers could be implemented with alternative optimized address resolution protocols given its capacity to manage explicit context predicates, carried into a specific section of the payload packets.

The flow treatment is part of one of the main features of SR and allows identify flows instead of just datagrams. So, in SR it is feasible to implement different schemas of flow recognition that include application, service or statistical means. Therefore, SR can be explained as a flow regulation component of Internet.

The soft-state functionality of SR supposes a knowledge plane, which is extended over the data and control planes of the current Internet infrastructure. The knowledge plane allows a global view of the flow activity over the network; it includes traffic engineering metrics, service and security policies, monitoring means etc. Thus, the soft-state functionality adds a deterministic but flexible
and granular regulation flow capacity, as an intrinsic property on the ongoing Internet architecture.

The integrity of services that come from heterogeneous applications is a functionality that comes from the QoS and policy management. So, contrasting the original transparency of the internet for all datagram transfers, today it is mandatory the correct discrimination and control of individual information flows. Thus, implementing flow treatment over the SR routers, in concert with the Knowledge plane, will simplify and optimize the today QoS underlying and overlaying mechanisms. The inference capacity of the SR routers, given multiple constraints, will allow the simple and gradual export of the end to end QoS warranties from the edges to the core of the Internet ASs but as well between ASs themselves.

The support of trust is a functionality that can be achieved also on the base of the control of the flows over specific zones. If a group of routers, inside of a zone, can follow the behavior of the flows that circulate through them, then these routers can privilege the well behavior of trust users and to limit the possible abnormal ones.

The evolutionary global functionality is the result of at least three main phenomena into the Internet statistical structure: a) the growing speed of the node aggregation, b) the “preferential attachment tendency” to the concentration of nodes on specific zones and c) their constant internal economic, technological and engineering optimization tradeoffs. The recognition of this supra-functionality, in the intricacies of internet, allows highlighting the fact that the re-composition of the Internet architecture is not just a trend but an unavoidable requirement (associated to the principle of change of complex networks).
Figure 2. Original and Ongoing Internet functionalities
Annex C. Traffic Classes Performance metrics and Link Utilization Performance metrics

The major traffic class metrics that can be managed are: Reachability, Delay, Packet Loss, Jitter, Mean Opinion Score (MOS). While the Link Utilization metrics that can be managed are link utilization distribution, range and cost [CiscoOER_12_2sx_book09].

With the exception of reachability, none of these performance characteristics can be managed within the constructs of conventional routing protocol metrics. The reachability can be extended (beyond ensuring that a particular route exists in the routing table) by automatically verifying that the destination can be reached through the indicated path.

Reachability
Reachability can be specified as the relative percentage or the absolute maximum number of unreachable hosts, based on flows per million (fpm). If the absolute number or relative percentage of unreachable hosts is greater than the user-defined or the default value, the edge router determines that the traffic class entry is out-of-policy and searches for an alternate exit link. The relative unreachable host percentage is based on a comparison of short-term and long-term measurements. The short term measurements have an order of magnitude of some minutes while the long term measurements have an order of magnitude of some tens of minutes.

The edge router measures the difference between these two values as a percentage. If the percentage exceeds the user-defined or default value, the traffic class entry is determined to be out-of-policy. For example, if 5 hosts are unreachable during the long-term measurement and 6 hosts are unreachable during short-term measurement, the relative percentage of unreachable hosts will be 20 percent.

Delay
Delay (also referred as latency) is defined as the delay between when the packet was sent from the source device and when it arrived at a destination device. Delay can be measured as one-way delay or round-trip delay. The largest contributor to latency is caused by network transmission delay.

Round-trip delay affects the dynamics of conversation and is used in Mean Opinion Score (MOS) calculations. One-way delay is used for diagnosing network problems. A caller may notice a delay of 200 milliseconds and try to speak just as the other person is replying because of packet delay. The telephone industry standard specified in ITU-T G.114 recommends the maximum desired one-way delay be no more than 150 milliseconds. Beyond a one-way delay of 150 milliseconds, voice quality is affected. With a round-trip delay of 300 milliseconds or more, users may experience annoying talk-over effects.

Packet Loss
Packet loss can occur due an interface failing, a packet being routed to the wrong destination, or congestion in the network. Packet loss for voice traffic leads to the degradation of service in which a caller hears the voice sound with breaks. Although average packet loss is low, voice quality may be affected by a short series of lost packets.
Jitter
Jitter means interpacket delay variance. When packets are sent consecutively from source to destination, for instance, 20 ms apart, and if the network is behaving ideally, the destination should be receiving them 20 ms apart. But if there are changes in the delay over the network (like queuing, arriving through alternate routes, etc) the arrival delay between packets might be greater than or less than 20 ms.

A positive jitter value indicates that the packets arrived more than 20 ms apart. If the packets arrive 24 ms apart, then positive jitter is 4 ms; if the packets arrive 16 ms apart, then negative jitter is 4 ms. For delay-sensitive networks like VoIP, both positive and negative jitter values are undesirable; a jitter value of 0 is ideal.

Mean Opinion Score (MOS)
With all the factors affecting voice quality, many people ask how voice quality can be measured. Standards bodies like the ITU have derived two important recommendations: P.800 (MOS) and P.861 (Perceptual Speech Quality Measurement [PSQM]). P.800 is concerned with defining a method to derive a Mean Opinion Score of voice quality. MOS scores range between 1 representing the worst voice quality, and 5 representing the best voice quality. A MOS of 4 is considered “toll-quality” voice.

Traffic Distribution for Link Utilization Policies
A link utilization distribution policy consists of an upper threshold on the amount of each traffic class that a specific link can carry. Every $t$ seconds, edge router reports the link utilization, link utilization distribution thresholds can be configured. If the exit or entrance link utilization is above the configured threshold\(^\text{25}\), the exit or entrance link is in an Out of Policy (OOP) state and the monitoring process starts to look for an alternative exit for the traffic class.

Link Range Policies
A link range policy is defined to maintain all links within a certain utilization range, relative to each other in order to ensure that the traffic load is distributed. If the difference between the links becomes too great, the edge router will attempt to bring the link back to an in-policy state by distributing traffic classes among the available links.

Cost Range Policies
Cost-based optimization allows you to configure policies based on the monetary cost (ISP service level agreements [SLAs]) of each exit link in your network. To implement OER cost-based optimization the OER master controller is configured to send traffic over exit links that provide the most cost-effective bandwidth utilization, while still maintaining the desired performance characteristics.

Peak Information Rate
The Peak Information Rate (PIR) is the maximum bit emission rate of a CTC. PIR is specified in bytes per second. The PIR measures the IP packet transmission rate and so in counting the number of bytes of an IP packet, the entire packet including the IP header is considered.

\[^{25}\] Cisco recommends for a similar case a threshold default of 75%.
Annex D. Classification and marking model

Similar to Layer 2 headers, the IP header has fields that can be used to classify traffic into treatment groups. The most widely used L3 marking techniques are Type of Service (ToS) and DSCP.

**ToS**

ToS was originally defined in RFC 791 and 795 and was further modified/updated by other RFCs like RFC 1122, RFC 1123, and RFC1349. Although the field has been there for quite sometime, it has not been widely used. Its use has been superseded by DSCP today. The ToS fields are shown in Figure 1. Figure 2 shows the precedence bits of ToS field.

<table>
<thead>
<tr>
<th>Bit</th>
<th>2</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2</td>
<td>P1</td>
<td>P0</td>
<td>T2</td>
<td>T1</td>
<td>T0</td>
<td>ECN</td>
</tr>
</tbody>
</table>

See the code assignation for the ToS Byte in the next layout.

<table>
<thead>
<tr>
<th>Binary</th>
<th>Decimal</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>0</td>
<td>Routine</td>
</tr>
<tr>
<td>001</td>
<td>1</td>
<td>Priority</td>
</tr>
<tr>
<td>010</td>
<td>2</td>
<td>Immediate</td>
</tr>
<tr>
<td>011</td>
<td>3</td>
<td>Flash</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>Flash Override</td>
</tr>
<tr>
<td>101</td>
<td>5</td>
<td>Critical</td>
</tr>
<tr>
<td>110</td>
<td>6</td>
<td>Internetwork Control</td>
</tr>
<tr>
<td>111</td>
<td>7</td>
<td>Network Control</td>
</tr>
</tbody>
</table>

- IP precedence—three bits (P2 to P0)
- Delay, Throughput and Reliability—three bits (T2 to T0)
- Explicit Congestion Notification (ECN) — two bits

- Delay - when set to 1, the packet requests low delay.
- Throughput - when set to 1, the packet requests high throughput.
- Reliability - when set to '1,' the packet requests high reliability.
DSCP
Differentiated Service Code Point (DSCP) was defined in RFC 2474 and RFC 2475. DiffServ (DS) has more priority levels than that of ToS because DS has more priority bits. DiffServ uses the same three most significant like ToS to define priority, but uses the next three bits to further refine them. DS fields are used to determine the per-hop behavior (PHB) of the packet. The following byte layouts illustrate the difference between ToS and DiffServ.

<table>
<thead>
<tr>
<th>DS5</th>
<th>DS4</th>
<th>DS3</th>
<th>DS2</th>
<th>DS1</th>
<th>T0</th>
<th>ECN</th>
<th>ECN</th>
</tr>
</thead>
</table>

- **DSCP**—six bits (DS5-DS0)
- **ECN**—two bits

The ECN bits were not in the original DSCP RFCs. They were later added later by RFC 3168 to allow for congestion notification in the path.

RFC 2597 for DiffServ defines Assured Forwarding (AF) PHB that can be used by a service provider to provide different forwarding assurances based on different AF classes. There are four different AF classes each with three different drop probabilities.

RFC 2598 for DiffServ defines Expedited Forwarding (EF) PHB. "The EF PHB can be used to build a low loss, low latency, low jitter, assured bandwidth, and end-to-end service through DS (DiffServ) domains. Such a service appears to the endpoints like a point-to-point connection or a "virtual leased line." This service has also been described as “Premium service.” Codepoint 101110 is recommended for the EF PHB.

Three-Class Provider-Edge Model: CE Design

In this model, the service provider offers three classes of service: Real-Time (strict priority, available in 5-percent increments), Critical Data (guaranteed bandwidth), and Best-Effort. The admission criterion for the Real-Time class is either DSCP EF or CS5; the admission criterion for Critical Data is DSCP CS6, AF31, or CS3. All other code points are re-marked to 0. Additionally, out-of-contract AF31 traffic can be marked down within the service provider’s MPLS VPN cloud to AF32.

Under such a model, there is no recommended provision for protecting Streaming-Video (following the “Don’t mix TCP with UDP” guideline), nor is there a service-provider class suitable for bulk data, which consists of large, nonbursty TCP sessions that could drown out smaller data transactions. Figure 3 shows a re-marking diagram for a three-class service-provider model.

![Figure 3. Three-class service-provider model](CiscoEnterpriseQoS)
Annex E. Business Driven Management Framework

In this work we have use the Business Driven Management Framework proposed in [Alb06] which enables the:

- Management by Business Objectives (minimization of service violation, account strategy business)
- SLA management
- Resource Control

This framework is divided into major layers: the business management layer and the resource control layer. To have a graphical view of this framework see Figure 1 in [Alb06].

The Business Management Layer

“The Business Management Layer is responsible for optimizing the alignment of utility computing (UC) resources usage with the objectives of the utility provider based on a set of business objectives defined and audited over relatively long periods of time (monthly, quarterly, etc.).”

Business objectives are the reflection of the utility provider’s business strategy and range over diverse key performance indicators, as: service operations, service level agreements, etc.

Business relationships contracted by the utility provider are formalized by SLAs and modeled using the Generalized SLA (GSLA) information model.

Each role in the SLA is associated with a set of Service Level Objectives (SLOs) to be achieved at a set of intrinsic policies related to the role behavior.

A special engine, we call the Role-to-policies mapping engine, translates Roles, SLOs and rules into a set of enabling policies that are applied in the resource control layer.

The resource control layer

The resource control layer refine these enabling policies to lower level policies (LLPs) that enclose all the low level logic required to correctly drive the utility resources.

LLPs are dealt with by the Policy Decision Point (PDP) module of the resource control layer.

Part of the PDP’s task is to monitor and respond to system events and notifications by selecting, activating, and scheduling the enforcement of the appropriate policies at the appropriate utility resources.

The PDP contains also sub-components for policy run-time conflict detection, root cause analysis, generation of the set of options available in the presence of some incident or problem, as well as the generate of appropriate configuration flows in order to enforce active policies.

If a change on a policy is required by a PDP, the PDP passes up the control to the BDM. BDM will select the one that will maximize the value to the utility provider. That is, the option that will result in the closest alignment to the business objectives. Such interactions offer also the opportunity for the architecture to learn and refine the policy repository.