Enriched Network-aware Video Services over Internet Overlay Networks

www.envision-project.org

Deliverable D3.1

Initial Specification of the ENVISION Interface, Network Monitoring and Network Optimisation Functions

Public report, Version 2, 19 May 2011

Authors

UCL Raul Landa, Eleni Mykoniati, David Griffin, Miguel Rio

ALUD Nico Schwan, Klaus Satzke

LaBRI Toufik Ahmed, Samir Medjiah, Abbas Bradai, Ubaid Abbasi

FT Bertrand Mathieu, Irène Grosclaude, Selim Ellouze, Yannick Carlinet, Pierre Paris, Valery Bastide, Emile Stephan

TID Oriol Ribera Prats, Armando Garcia, Arcadio Pando, Alvaro Saurín

LIVEU Noam Amram

Reviewers Bezalel Finkelstien, David Griffin

Abstract This deliverable presents the work performed in WP3 during the first year of the ENVISION project, focusing on the results of a survey of existing interfaces, tools and solutions and introducing the initial designs of the project. The first section of this document introduces the network-level ENVISION architecture in order to better understand the relationships between the different functional blocks. The CINA interface (Collaboration Interface between Network and Applications) and a number of scenarios for its discovery by overlay applications are introduced before positioning the CINA interface with respect to related work in this area. The monitoring aspects of the ENVISION system are introduced by presenting metrics that are most relevant for the monitoring aspects in the project, as well as a number of tools that are potentially capable to provide the level of monitoring of those metrics that is required for the project. Finally, the ENVISION network optimisation aspects and a number of network services suitable for the optimisation of content delivery to the end-users on one hand, and for the reduction of the overall network load on the other, are detailed. Among the investigated network services are multicast delivery, caching services, content adaptation nodes, and QoS-based optimising systems.

© Copyright 2011 ENVISION Consortium

University College London, UK (UCL)

Alcatel-Lucent Deutschland AG, Germany (ALUD)

Université Bordeaux 1, France (LaBRI)

France Telecom Orange Labs, France (FT)

Telefónica Investigacion y Desarollo, Spain (TID)

LiveU Ltd., Israel (LIVEU)

Project funded by the European Union under the
Information and Communication Technologies FP7 Cooperation Programme
Grant Agreement number 248565

EXECUTIVE SUMMARY

This deliverable contains the main achievements of work done during the first year in the WP3 workpackage:

- A first description of the network-level functional architecture which enables a better understanding of the relationship between the functional blocks to be designed in this WP has been made. The elaboration of a detailed view for each functional block has been initiated and is ongoing.
- A first analysis and a first proposal for discovering the CINA interface (Collaboration Interface between Network and Applications) allowing overlay applications to discover the CINA server hosted by network operators has been achieved. This first input is part of a draft submitted to the IETF ALTO working group.
- A survey of related work on overlay-ISP interfaces, providing the required level of knowledge on previous work and to helping to identify functions, primitives, or encoding structures that might be reused, has been done and will be helpful for the specification of the CINA interface.
- The definition of the network monitoring architecture has been initiated. Also, a first list of metrics that could be of interest to monitor, as well as an initial survey of existing monitoring tools.
- The survey on Multicast network service (presentation of existing multicast techniques).
- First proposals of new functions for the use of multicast in ENVISION has been made: multicaster box, hybrid multicast, transparent multicast, high capacity node. The first proposals have been discussed and agreed, with design work underway to be completed in the coming months.
- A survey on the Content Adaptation service and relation between the network and the encoding schemes has been completed: adaptation in the network via the use of DPI or not, codec adaptation, bitrate adaptation, FEC at network-level, SVC and multicast.
- Surveys of Caching services and the necessary functions for caching as well as the related work of existing solutions (transparent and explicit caching) have been done and will help to specify the caching node architecture.
- The analysis of models for estimating the hit-ratio of caches has been performed and its first results will enable to better focus and refine the caching work.
- The survey of QoS services: presentation of QoS, existing solutions such as DiffServ and IntServ, as well as a first proposal for a multiple-link approach which will be further developed to specify a solution for ensuring aggregate performance across multiple network interfaces and access networks.

Copyright © ENVISION Consortium, May 2011

TABLE OF CONTENTS

Copyright © ENVISION Consortium, May 2011

Copyright © ENVISION Consortium, May 2011

LIST OF FIGURES

Copyright © ENVISION Consortium, May 2011

Copyright © ENVISION Consortium, May 2011

1.1 Introduction

The goals of WP3 are to define the ENVISION interface, which we call CINA (for Collaboration Interface between Network and Applications), to monitor network metrics and to design/apply solutions for optimising the network. This work is divided in three tasks, one for each activity. All the results achieved within the three tasks are grouped within this deliverable for better clarity and understanding.

This document is organised as follows:

The first section of this document introduces of the ENVISION network-level architecture in order to better understand the relationships between the different functional blocks.

Then, the CINA interface and a number of scenarios for its discovery by overlay applications are introduced. Also, the CINA interface is positioned with respect to related work in this area.

The monitoring aspects of the ENVISION system are introduced by presenting metrics that are most relevant for monitoring in the project, and a number of tools that make feasible the provision of the level of monitoring of those metrics that are required for the project.

Finally, the ENVISION network optimisation aspects and a number of network services suitable for the optimisation of content delivery to the end-users on one hand, and for the reduction of the overall network load on the other, are described. Among the network services investigated are the multicast delivery, caching services, content-adaptation nodes, and QoS-based multi-link optimisation systems.

Copyright © ENVISION Consortium, May 2011

2. ENVISION ARCHITECTURE

2.1 Presentation of ENVISION

Application-layer networks are global overlays running on top of the Internet - today mainly concentrated in the field of file sharing although live media streaming applications are emerging but currently offer only a limited QoE. However, considering the highly demanding nature of future, interactive multi-participatory communications including HD and 3D video, future overlays need to be aware of the underlying networks' capabilities to offer increased QoE to the users. Overlay applications need to be able to influence how their data is transported across the application-layer network making efficient use of the facilities of the underlying ISP networks. Especially considering the inter-domain nature of applications with participation of users around the world this is a challenging task.

In summary the required increase in quality calls for access to specific network resources by application overlays which are today hidden in the walled gardens of network providers. To achieve efficient cross-layer integration, such network capabilities need to be made available by network providers to the service developers and integrators domains. A major research theme of the ENVISION project, therefore, is to expand and enhance the overlay-ISP interaction, by developing a *comprehensive, media-aware open and standardised interface* between the ISPs and the application overlay, which we call CINA (Collaboration Interface between Network and Applications).

The following picture (figure 1) presents the big picture of ENVISION.

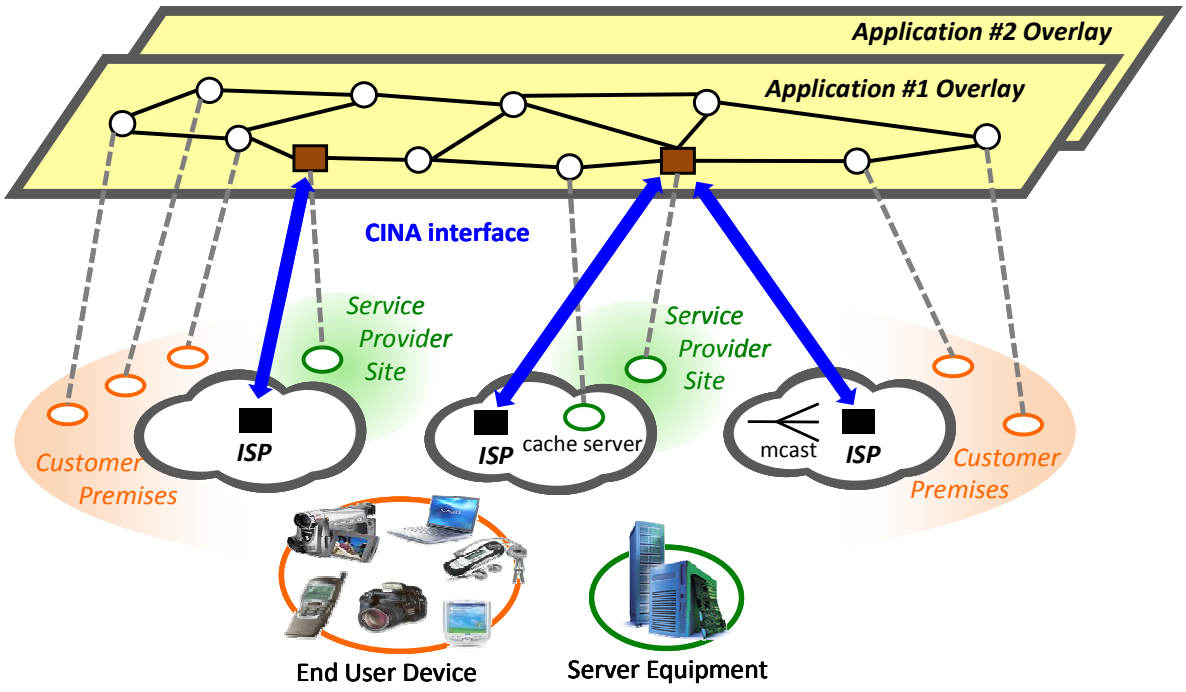

Figure 1: The CINA interface and its relationship with overlay applications, ISPs and service providers

Copyright © ENVISION Consortium, May 2011

2.2 ENVISION network-level architecture

In WP2, the overall functional ENVISION architecture has been designed (figure 2).

WP3 focuses on the network level of the architecture and the interface between the network and the overlay. This section present the first thoughts about the more detailed view of the network-level functional blocks, having in mind it can change during the next months, depending on the technical work, to refine it.

The network-side functional architecture contains 4 functional entities:

- Network Management (9)
- Network Services Control (6)
- Network Data Management (5)
- Network AAA (7)

Those functional entities are hosted at the network side and managed following the network operator policies. They allow having a well-defined cooperation with the overlay application taking into consideration the targeted aspects: exchange of information, mutual aid, security and potentially a second level of cooperation with charging functionalities.

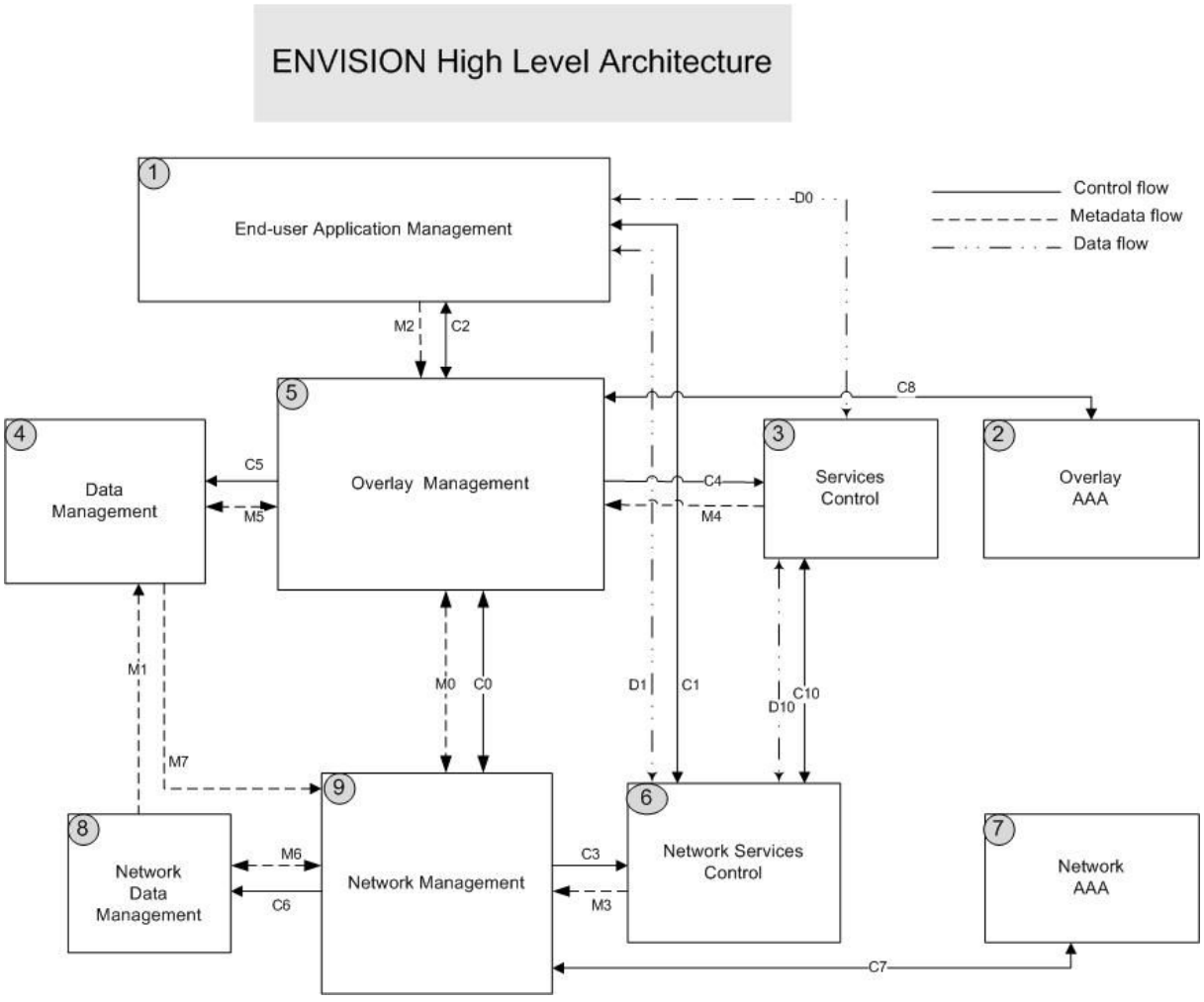

Figure 2: ENVISION overall architecture

Copyright © ENVISION Consortium, May 2011

2.3 Functional Blocks

- Network Management (9)

  The network management (NM)block is the functional entity in charge of the coordination of the whole set of actions provided by the network. It implements the system intelligence at the network level for processing the received data and requests, deciding the actions to take and controlling the network and services equipments. It implements the procedures for supporting the security procedures managed by the Network AAA functional entity such as authentication. It provides information to the Data access level function as well to define the access level and the authorisation procedures for the overlay application. It executes procedures related to overlay optimisation (such as peer ranking procedure) and related to network optimisation (such as enabling transparent caching).

- Network Services Control (6)

  The Network Services Control (NSC) block is the functional entity implementing the mechanisms and procedures executing policy-based network services. It implements the procedures for configuring the network and services parameters relying on the NM commands. It can collect the feedbacks about the current network equipment and services states. It contains the functions responsible for controlling the network and its services (e.g. for setting a multicast tree, network equipments have to exchange signalling messages between them to define the parameters of the tree). It includes the set of functions responsible for the admission and management of the data traffic depending on the rules defined. An example is a differentiated stream with a reserved bandwidth for a specific path is treated in priority by the Resource reservation function at the data plane.

- Network Data Management (5)

  The Network Data Management functional entity is a subsystem responsible of collecting and managing network and overlay related information. It implements procedures for processing and executing the new policies defined by the NM. It is responsible for mapping the available data with an authorisation level and executing any other security mechanisms, e.g. encryption.

- Network AAA (7)

  The Network AAA block is the functional entity that contains the information related to the overlay applications from one side and network services from other side. It should process and answer NM requests, e.g. an authentication request for an overlay application, an authorisation access level for a network service, etc,

  The following sections will focus on the three main features investigated within WP3 and described in this functional view of the network-level ENVISION architecture, i.e., the CINA interface, the monitoring, and the network services that would help to optimise the network.

Copyright © ENVISION Consortium, May 2011

3. CINA INTERFACE

A central aspect of the ENVISION project is the design of a new interface which allows a mutual cooperation between overlay applications and underlying networks. Two protocol specifications will be provided for this interface. The first one is the specification of the CINA interface, detailing the provided methods that allow the exchange of network and application specific information and the activation and invocation of services, as well as the protocol specification which defines how the interface is invoked. The second part of the specification will provide a mechanism and a protocol used by client applications in order to discover the right CINA interface that is responsible for them.

As a first step towards these specifications this chapter is split into two subsections: The first one (section 3.1) presents the motivation for a discovery protocol, an overview on the different discovery scenarios and a discussion on various mechanisms that can be used as discovery protocol. The second one (section 3.2) is dedicated to the CINA interface itself and at this point in time mainly introduces the related work. The actual specification will be based on requirements provided by the Network-Aware Overlay Applications workpackage as well as the Network Monitoring and Optimisation tasks and will be included in the next iteration of this deliverable.

3.1 Discovery of the interface

Before nodes of an overlay application can invoke the CINA interface they first need to discover a contact point. How the discovery is done strongly depends on the deployment scenario and requirements of both the network and the application. While for some use cases a manual configuration could be sufficient, more sophisticated use cases, where nodes of an overlay application are mobile and can potentially join different access networks, need additional mechanisms. Each location and network will have its own CINA interface that is responsible for providing information and services to overlay nodes that are registered in the authoritative domain of the interface. Here a static configuration of the address and access details of the interface of a single access network is insufficient. Instead, a dynamic protocol that allows nodes or their surrogates to discover the appropriate interface for their current location is needed and will be studied. The current Internet architecture does not support this kind of service discovery at the moment, although various IETF working groups are currently in parallel to ENVISION working on a solution [ALTO10][GEO10].

There are numerous service discovery protocols already defined and partly deployed in today's Internet. In order to decide whether one of those already existing approaches can be used for ENVISION the next section first briefly describes the scenarios where nodes of the overlay need to perform the discovery. The following section defines requirements which the protocol needs to fulfil from the ENVISION perspective. Finally the last section introduces existing approaches and shortly discusses how they could be used for ENVISION.

3.1.1 Discovery scenarios

An overlay application that wants to access the CINA interface needs to have a way of finding the right contact point. This contact point is usually the one associated to the access network of the entity that wants to invoke the interface. However in some scenarios this invocation is not done by this entity itself, but by a third party, for example a service node or some logically separate entity. These third parties typically are not part of the same access network and thus have no direct relationship to the CINA interface that is responsible for the user peer. In all cases the contact point of the CINA interface first needs to be discovered by those entities. This section gives an overview on the different possible scenarios and discusses how the discovery can be done depending on the node which initiates the CINA interface request.

Figure 3 gives an overview of the three different possible scenarios. It shows the user peers that are part of the overlay application and registered in an access network. Also part of the overlay

Copyright © ENVISION Consortium, May 2011

application is the SuperPeer, which represents an entity that wants to invoke the CINA interface on behalf of a user peer. This could be for example a hierarchically different entity of the overlay application, such as a Tracker in a BitTorrent system, or it could be a service node of the overlay that optimises or invokes network services to help the user peer. This SuperPeer is usually not located in the same access network (although in some cases it could). The figure further shows the CINA interface of the access network and the ENVISION discovery server. The discovery server represents the not–yet-defined mechanism that helps the peer to discover the right interface.

The figure also depicts three protocols that are utilised in the scenario as well as two logical roles of the entities of the overlay application. The red arrow represents the ENVISION Query protocol, used for invocation of services or for querying network information through the CINA interface. The entity performing this query is thus highlighted in light red. The green arrow represents the protocol that will be used for discovering the contact information of the CINA interface. The entity performing the discovery is framed greenly. The blue arrow illustrates the overlay application protocol that is used by the entities of the overlay application to exchange control information.

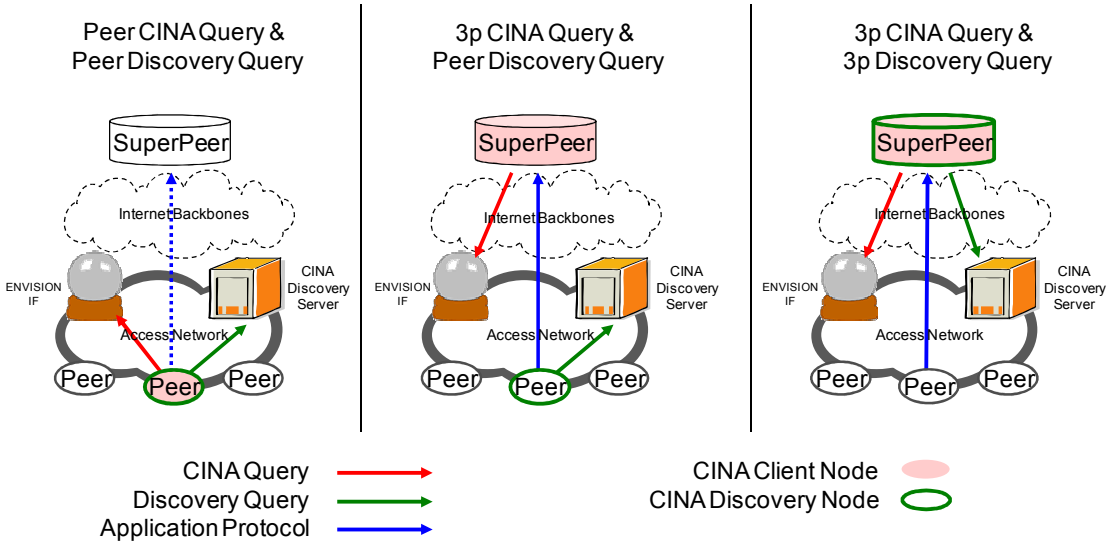

Figure 3: CINA interface discovery scenarios

- The first scenario, illustrated on the left, shows the standard use case where the user peer is performing both the interface discovery as well as the invocation of the CINA interface. This is the simplest case as the user peer doing the discovery already has a relationship to the underlying access network and thus to the CINA interface.
- In the second scenario, illustrated in the middle, a third party – the SuperPeer –invokes the CINA interface. This SuperPeer needs a way to discover the contact point in the network without having a relationship with it. One option to solve this is the user peer performing the discovery by itself and then forwarding this information towards the third party. This typically works only if the overlay application protocol already supports this forwarding of contact information or if a modification or extension of the protocol is feasible.
- If the application protocol does not support forwarding of this contact information, for example in legacy applications or in application where the overlay optimises transparently to the user peer (e.g. in a CDN), the discovery also has to be done by the third party. This is depicted in the third scenario. Here the discovery mechanism must provide a way for the third party to find the CINA interface associated to the access network of the user peer, whereas the third party has only limited information about the access network (typically the IP address of the user peer).

Copyright © ENVISION Consortium, May 2011

3.1.2 Requirements

This section provides requirements that need to be met by the CINA interface discovery protocol. The requirements are derived from the discovery scenarios and reflect the nature of global Internet overlay applications. Specifically the mechanism needs to allow potentially millions of application nodes to discover a specific server which is related to their access network. Thereby the nodes can be registered in different types of access network or they can be located behind residential gateways and NATs.

- ENVISION clients MUST[1] be able to perform the CINA Interface discovery to find one or several CINA interfaces
- ENVISION clients MUST be able to perform server discovery, even if they or the user peer or the discovery node are behind a network address translator (NAT).
- During the evaluation of existing protocols or mechanisms, their availability in various access network architectures and their suitability for third-party queries should be taken into account.
- The interface discovery mechanism SHOULD be able to return the respective contact information for several interfaces.
- The CINA interface discovery mechanism SHOULD be able to indicate preferences for each returned CINA interface.
- The protocol MUST provide an ENVISION client with a response containing one or several suitable CINA interfaces or it MUST inform the ENVISION client if the discovery failed and SHOULD provide the reason for the failure.
- The protocol SHOULD be as simple as possible. The less different mechanisms the protocol has to support the more favourable. The protocol SHOULD require minimal changes only to the overlay application protocol, to the overlay entities and to other network entities.
- The response MUST reach the ENVISION discovery node in a reasonable amount of time.
- The ENVISION discovery SHOULD return a CINA interface that is located in the same domain like the user peer, if such a CINA interface is available. Alternative interfaces can be discovered additionally.
- The protocol MUST work across domain/AS borders.
- The protocol MUST work for all common access network architectures.
- The protocol MUST work in case the ENVISION discovery node is behind a residential gateway.
- Depending on the scenario the discovery node MUST be able to perform many requests, thus a high efficiency of the protocol is favourable. Caching techniques that support the efficiency of the ALTO discovery client SHOULD be taken into account during the evaluation of protocol options.
- The end user SHOULD not need to perform additional configuration steps.

[1] 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 RFC 2119 [RFC2119].

Copyright © ENVISION Consortium, May 2011

3.1.3 State of the art

This section provides an overview of potential service discovery mechanisms that are candidates for the use within the project. The section will further discuss the applicability for the CINA interface discovery use case addressing potential issues and consideration for each.

3.1.3.1 Manual configuration

Manual configuration of the CINA Interface location(s) is one potential option to be used, especially in deployments where the application spans only one single access network. However manual configuration is not well suited when the interfaces of multiple ISPs need to be discovered or if cross-domain services are required. Typically overlay applications, especially p2p networks, are deployed worldwide and the participating nodes often connect to nodes from ISPs that they may not have contacted before. In this case manual configuration cannot be used as neither users nor service operators will be able to efficiently keep the configuration data up to date.

3.1.3.2 Domain Name System (DNS)

The DNS has been used widely in the Internet to discover the server address for applications. DNS provides NAPTR [RFC2915] and SRV [RFC2782] resource records which can potentially be used as service discovery mechanisms. Both resource records are promising candidates. In the end it's a trade-off between flexibility and simplicity whether the use of NAPTR or SRV is preferred. S-NAPTR [RFC3958] and U-NAPTR [RFC4848] mechanisms offer a Dynamic Delegation Discovery System (DDDS) Application to map the domain, service and protocol name to a target URI, or alternatively to a target host and port. SRV records offer the possibility to map the domain, service and transport protocol name to a target host and port only. Both options make DNS a potential candidate to be used as ENVISION discovery mechanism. Which solution better suits is open for discussion and will be decided once the final requirements have been fixed.

A precondition for the use of the described DNS resource records is that the client determines the name of the access network where it is registered in as input for the DNS resolution. There are various ways how a client can do this, for example the use of DHCP, the IANA database, a reverse PTR DNS lookup on the own IP address or manual configuration. Further studies need to be done to determine which of the mechanisms is favourable.

3.1.3.3 Dynamic Host Configuration Protocol (DHCP)

DHCP is a protocol that allows network operators to configure clients within their authoritative domain remotely. Through the specification of a new DHCP option, the CINA interface URI can be pushed to clients directly.

While DHCP is widely deployed in the Internet it still has several limitations. DHCP will not always be adapted, for example in PPPoE based DSL access networks a different mechanism has to be used. Another set of problems with DHCP is based on the existence of many residential gateways or broadband routers with NAT. Typically information such as the CINA interface URI is not automatically forwarded by the residential gateway to the nodes of the overlay application within the local network. To allow this, residential gateways that are already deployed would need a software upgrade that deals with the extension. Also DHCP poorly supports scenarios where the discovery is not done by the client peer but by a third party, which might not be part of the administrative domain of the access network operator running the DHCP server.

3.1.3.4 Provisioning

A configuration file that contains the CINA interface address for its clients provided by either the access network operator or the application provider can be used to keep track of the various contact points of the CINA interfaces. This list can be referenced by the application in order to discover the

Copyright © ENVISION Consortium, May 2011

right interface. However this mechanism is limited due to its low dynamicity and the high maintenance effort of keeping the list up to date. In conclusion this seems to be only realistic to small deployments, spanning only few different access networks.

3.1.3.5 Multicast and broadcast

Also multicast or broadcast can be used for the CINA interface discovery in some scenarios. In general there are two ways of using IP multicast for service discovery:

1) The clients send out discovery requests to a well-known multicast address and port and wait for responses from available servers or service providers, which usually respond in unicast.
2) Servers or service providers send out service announcements via multicast either once when a service is deployed or periodically. Clients listen to these announcements in order to find a service.

Both methods can be used simultaneously by clients and servers. However using multicast or broadcast raises some concerns regarding flooding of networks and clients as well as the limited scope of the discovery area, as multicast forwarding typically is disabled at routers. Examples of multicast-based discovery approaches include [IDChKr10], [IDGoCaLe99], [WSDD05], SLP [RFC2165], and LLMNR [RFC4795].

3.1.4 Discovery protocol

As described in the previous sections there are already various service discovery mechanisms specified in the Internet. Unfortunately none of them supports the interface discovery requirements and scenarios out of the box. However it is assumed that a procedure that leverages one ore more of these already deployed mechanisms can be specified which is suitable for the service discovery.

From the discussed mechanisms the multi- and broadcast based mechanisms (section 3.1.3.5) appear to be the least favourable ones as discovery mechanism. For security reasons multicast generally is disabled by network operators for user nodes which prevents them to employ the mechanism. Even if multicast is enabled it is restricted to only one autonomous system, which makes it a bad candidate for third party discovery scenarios. Also provisioning (section 3.1.3.4) through a file seems to be realistic only for small deployments due to the high maintenance effort and its low dynamicity for global deployments. DHCP (section 3.1.3.3) in turn is a straightforward way for network operators to remotely configure user nodes. A DHCP extension can be standardised for the interface discovery. DHCP however is limited to certain types of access networks only. In conclusion a different mechanism would be needed for access networks where DHCP is not deployed. Also for third party discovery scenarios DHCP is not suitable. The DNS system (section 3.1.3.2) is globally deployed, has been proven to be scalable and can be used for normal and third party service discovery. However at this point in time it is unclear whether the reverse DNS lookup which is needed to map the IP address of a client to the access network domain name is reliable. For example concerns have been raised regarding the use of DNS PTR for IPv6 and also it is unclear whether generally network operators support this type of DNS query in their networks. Manual configuration (section 3.1.3.1) is not desirable due to the dynamicity of nodes and the resulting maintenance effort by the user. However manual configuration can be used as an add-on to allow the skilled user to take influence of the discovery, for example to allow the selection of a third party ENVISION provider.

In summary there is no single mechanism that can be used directly, thus the use of a combination of the discussed mechanisms seems to be the most promising approach to move forward. At this point in time different variations have been discussed already [ALTO3p], without meeting all presented requirements yet. Thus the final specification for the interface discovery protocol will be contained in the next iteration of this document.

Copyright © ENVISION Consortium, May 2011

3.2 Interface description

In ENVISION, one of the main objectives is the definition of an interface between the overlay application provider and the network operator or ISP. This interface is to allow overlay applications to obtain information about the network (i.e. network conditions, available services, etc.) as well as requesting the dynamic activation of some services the ISP can provide. Since many services can be imagined (see section 5), the interface should be generic enough to accommodate many. It should also include a range of network parameters, since the overlay application can use the interface with several network operators but each operator can have its own policies and have a specific agreement with the overlay and thus provide a different set of metrics. The relationship, agreement, business model between overlay providers and network operators is presented in D2.1. Because of agreement between the actors, the interface should also provide security mechanisms and include facilities to allow AAA services. Prior to ENVISION, several interfaces existed, as described in the following sections.

3.2.1 State of the art

3.2.1.1 IP Multimedia Subsystem (IMS)

IMS (IP Multimedia Subsystem) is a system designed by network operators in order to provide a richer set of services, with a standardised platform. Initially proposed by the 3GPP (3rd Generation Partnership Project) [3GPP05] for wireless networks, IMS aims at providing convergence between various access networks (e.g. Fixe-Mobile Convergence) by offering the same services whatever the access network is and the terminal is, using the IP protocol. IMS also should allow mobility of end-users [3GPP06] [IMS06].

IMS advocates the separation of the control plane and the service plane. The following picture depicts the overall IMS architecture, where we can easily see the separation of the three layers (network, IMS and services). The separation of layers resembles to our ENVISION architecture where the overlay/applications could be assimilated to the service layer, the network layer being our network layer and the IMS layer being distributed into the ENVISION overlay and the network layer.

As for ENVISION, we can see that IMS specifies several interfaces between the layers, for instance between the CSCF (Call Session Control Function) and the MRF (Media Resource Function) and IMS GW (GateWay) and between those two modules and the network. There are about 20 interfaces defined between each functional blocks of the architecture. Interfaces are defined for exchanging information between components: user profile information from the HSS database, policies or priority and filers, etc. For those ones, the protocol Diameter is selected. For other interfaces, mainly related to calls and signalling (e.g., session control, session information exchanges, message exchanges, etc.) the SIP protocol is used. For controlling some component such as MCGF (Media Gateway Controller Function) or MRFC (Media Resource Function Controller), which aims at controlling user-plane resources or media stream resources, the H238 protocol is preferred.

To sum up, this set of interfaces enables a specific configuration or activation of some network parameters and functions. Contrarily to what we expect to make within the project, the interfaces are rather fixed, do not allow flexible parameters to be exchanged neither the dynamic activation of a rich set of network services. Finally, the service layer is still under the control of the network operator and a controlled service provider, whereas in ENVISION, the interface is more open to service providers.

Copyright © ENVISION Consortium, May 2011

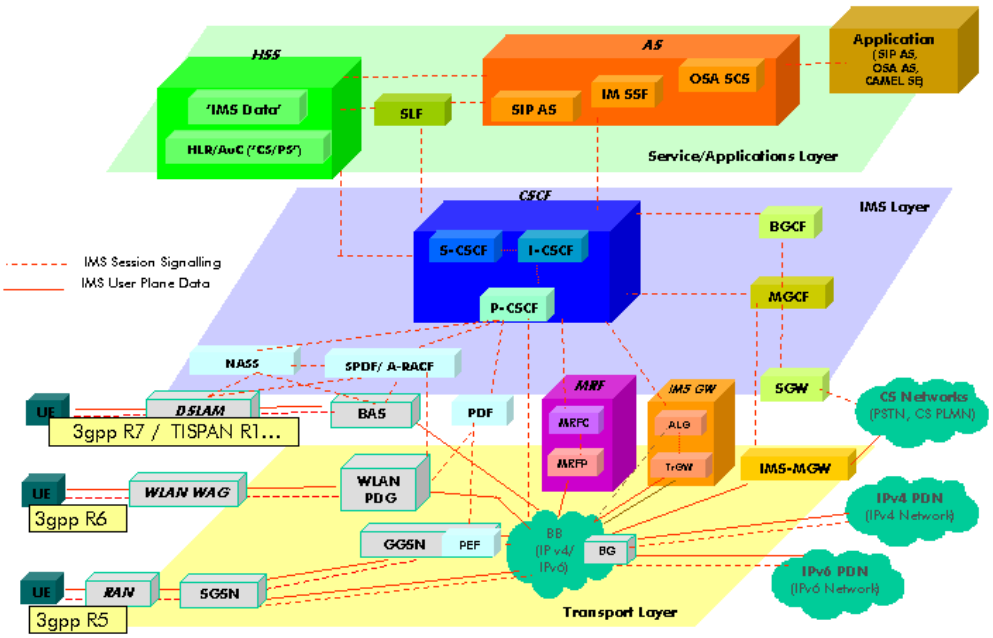

Figure 4: 3GPP/TISPAN IMS Architectural Overview

3.2.1.2 Parlay/X

Traditionally, the telecommunications environment is 'closed', and applications can only be developed internally with specific knowledge of individual network technologies. In the last several years, there has been an enormous increase in efforts to 'open up' these networks for application development [MOER03]. In opening up the network, new business models emerge where applications can be developed and provided by enterprises outside the traditional network operator domain. These applications can utilise the feature-rich service capabilities of the network through standardised Application Programming Interfaces (APIs) with off-the-shelf IT technology and tools such as Java and Web Services.

The Parlay APIs, otherwise known as Open Service Access (OSA), was a set of standardised open APIs that allow applications access to network functionality by packaging and presenting the service capabilities in a manageable fashion. The OSA/Parlay APIs have been jointly developed and published by the Parlay Group [PARL02], Third Generation Partnership Program (3GPP) [3GPP05] and European Telecommunications Standards Institute (ETSI) [ETSI04], and formed the API layer of the 3GPP IP Multimedia Subsystem (IMS). The APIs provided a technology-agnostic abstraction of functions including call control, location and user interaction among others.

Parlay/X also achieved these goals, but further stimulated the development of next generation applications by IT developers who have not necessarily been experts in telecommunications. Parlay/X offered a higher level of abstraction compared to OSA/Parlay APIs, and exposed the interfaces through Web Services technology.

However, uses in both fixed [TURN02, JAIN03] and mobile [GOLD04] networks have been focused on telephony-type applications, particularly those that require call control capabilities, such as:

- "Third Party Call": Creating and managing a call initiated by an application.
- "Geocoding": Get the location address of a subscriber e.g. country, state, district, city, street, house number, additional information, and zip/postal code.

Copyright © ENVISION Consortium, May 2011

- Application-driven Quality of Service (QoS)": Dynamically change the quality of service (e.g. bandwidth) available on end user network connection.

Apart from just standardising interfaces for network services (which Parlay/X did in order to make the access suitable for the web community), there are new initiatives such as GSMA OneAPI that try to set directions in the areas of service-federation, and standardise aggregator-oriented models. GSMA OneAPI is the current valid standard from the GSM association for Telecom third party API.

OneAPI is an official OMA standard profile of Parlay REST, and is an OMA candidate release approved June 2010. There is also an official OMA SOAP OneAPI profile, utilising a subset of Parlay X.

3.2.1.3 OneAPI

The OneAPI [ONEA10] is being standardised via the OMA and will re-use parts of the Parlay/X definitions. OneAPI consists of a set of APIs that expose network capabilities over HTTP, and is developed in public and based on principles so that any network operator or service provider is able to implement OneAPI. The goal is to allow developers of mobile applications to work on a unified API that gives them access to subscribers of different operators without a tight integration of the applications with the operators systems. Network services and capabilities, such as charging messaging, location and user context, will be available in a lightweight and web friendly way through OneAPI. Network operators will be able to support OneAPI without changing their backend architecture. Currently there is one OneAPI Regional Pilot started in Canada and more pilots are planned in Europe and the US [ONEA10].

The currently released OneAPI in the version 1.0 is available as RESTful, HTTP based and JSON encoded version and additionally as WSDL/SOAP, HTTP based and XML encoded version. Four different services are defined so far.

The SMS service allows an application to send and receive text messages. Methods are available that also allow the application to check the delivery status of a message and to subscribe to status notifications for sent text messages. An application further is able to subscribe to notifications for text messages that have been sent to it.

Similar to the SMS service the MMS service provides sending and receiving of multimedia messages. Additionally methods also allow delivery status subscriptions or queries, as well as subscriptions to notifications for MMS messages that have been sent to a web application.

The location service allows an application to get the location of one or many mobile devices. The application specifies either the MSISDN or the Anonymous Customer Reference if supported by the operator as well as the desired accuracy. The response contains the current location of the terminal with parameters for the accuracy, altitude, latitude, longitude and a timestamp.

The fourth service specified by OneAPI 1.0 is a payment service that charging mobile subscribers for the use of an application or content. Direct charging is supported as well as fund reservation for subsequent charging.

The next release OneAPI v2.0 is aimed to include 'Data Connection Profile' (lookup the network name and bearer and other device capabilities) and Click-to-call from a Web page;. Although the release of the second version is scheduled to be available end of 2010, currently there is no public information available about the specification of the second version. OneAPI release 3.0 will add Video Quality (request a Quality of Service to ensure video streams are jitter free and establish that they have been delivered) plus application triggering ('wake up' a device application with SMS/UDH and other technologies), however currently no release date is fixed.

The current version of OneAPI already provides interesting services to third party applications. In addition to traditional telecommunication services such as SMS and MMS especially the location and charging methods seem to be highly interesting for overlay application developers. The potential will

Copyright © ENVISION Consortium, May 2011

even be increased by upcoming methods once they are available, especially the QoS reservation for video streaming services of OneAPI 3.0 as well as the ability to gather network related information of a terminal device seem to be highly interesting. ENVISION, in contrast to OneAPI, is focused on highly distributed overlay applications and thus needs to offer its services for all types of access networks in a dynamic way.

3.2.1.4 Simple Network Management Protocol (SNMP)

SNMP is a protocol defined at IETF, mainly currently used for management purposes. It defines a set of functions, where a management entity can collect information from network equipment such as routers, switches, servers, etc. Each network equipment hosts an SNMP agent and an MIB (Management Information Base), a kind of database containing the configuration and monitoring information that the equipment is able to collect and provide to the management entity when receiving SNMP GET requests. The manager can also change the equipment configuration, via the use of SET requests, which is the operation for modifying values in the MIB.

There are now three versions of the SNMP protocol, the first one being released in 1988 and the last one in 2004.

SNMP could be seen as a possible protocol candidate for the CINA interface at least for overlay applications for collecting network information. However, SNMP is mainly designed and used within an organisation and not designed as an exchange protocol between two different organisations. Indeed, the manager is directly connected to the MIB of the network equipment, which is not a good option for ISPs (they prefer to be the only ones to have control on the network equipment). Even if the newer version of SNMP tries to bring some security features in the protocol, it is not yet enough for ISPs to allow any third-party to access their network equipment. Finally, the primary goal of SNMP is the management of entities and do not really enable the activation of network services such as the ones we envision in the project.

3.2.1.5 Application Layer Traffic Optimisation (ALTO)

The IETF WG “Application Layer Traffic Optimisation” (ALTO, [ALTO10]) is chartered to define protocols for the Discovery of the ALTO Service on one hand, and the information exchange from the ALTO Service to provide network information to applications on the other. An ALTO Server provides its view of a network region to ALTO clients in order to help the nodes of an overlay application to create connections to neighbour peers according to the preferences of the ISP. The goal is to allow a better overall performance of the application while especially for the operator expensive intra-domain traffic can be avoided.

The current ALTO protocol specification [ALPR10] is still under development by the working group and has not reached RFC status yet. In its current version 6 the Internet draft specifies a common transport protocol, messaging structure and a transactional model. The protocol is divided into services of similar functionality. The Server Information Service provides details on the capabilities of a specific ALTO server, such as supported operations and cost metrics or alternative ALTO servers. The ALTO Information Service group contains one basic Map Service which provides the current view of the network region to a client based on Network Maps and Cost Maps. Additionally three services offer additional information: The Map Filtering Service allows the ALTO Server to filter the provided maps according to a specific query of a client. The Endpoint Property Service allows clients to check specific properties of endpoints, such as the Network Location or the connectivity type. Finally the Endpoint Cost Service allows the ALTO Server to rank endpoints directly. Figure 5 illustrates the different available ALTO services.

Copyright © ENVISION Consortium, May 2011

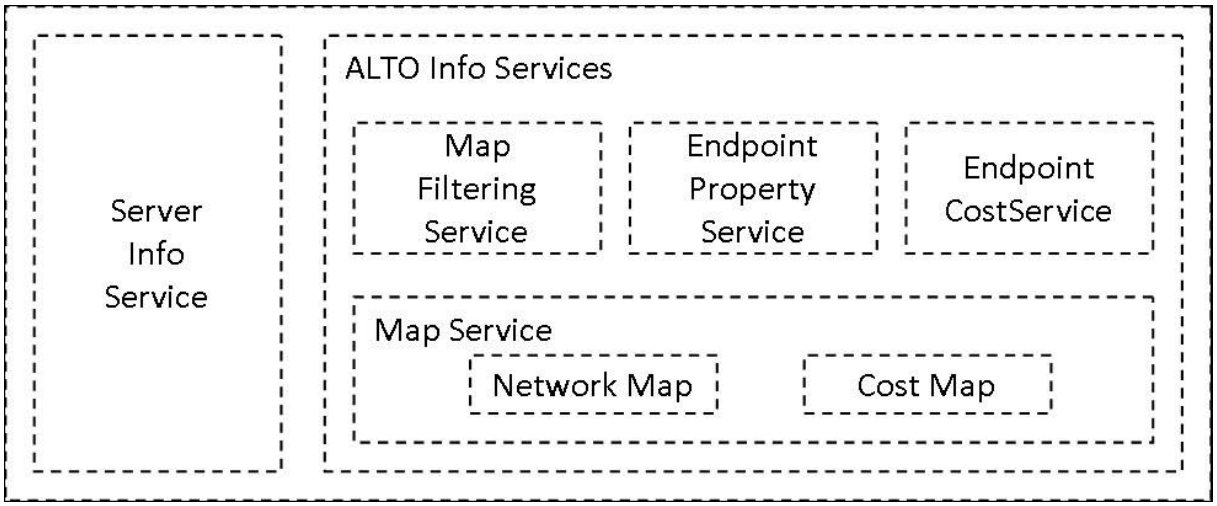

Figure 5: ALTO protocol structure

The ALTO server internally stores an information base which it uses to calculate costs for paths between different endpoints. How the information base is constructed is left open to the ISP that runs the ALTO server, thus it may contain for example routing policies or network measurements. The information base is translated by the ALTO Server into the Network Map and the Cost Map, which an ALTO Client retrieves and uses for path rankings. The Network Map defines the network regions that an ALTO Server considers: It aggregates endpoint addresses, such as IP addresses, together and defines a network location identifier (PID) for each group. The Cost Map then provides pair wise Path amongst sets of source and destination network locations. The costs can represent different metrics, such as air-miles, hop counts or generic routing costs, whereas lower values indicate a higher preference of the network operator for traffic to be sent from a source to a destination network location. The advantage of separating network and cost information types into different maps is that both components can be updates independently from each other, for example in different time scales. While network information is considered to be relatively stable, network conditions and with them the costs may be subject to higher dynamicity. After retrieving both map types an ALTO Client is able to check based on endpoint addresses to which peers it should establish connections preferably. In cases where the end device should not be burdened with the necessary calculations or where the network operator refrains from publishing exact network maps the ALTO Client may use alternative services in order to get pre filtered maps or rankings based on endpoint addresses.

The ALTO protocol encoding approach employs a RESTful interface over HTTP and uses JSON encoding for message bodies. This allows the ALTO protocol to benefit from the already installed base of HTTP infrastructure and the fact that many overlay applications already have an HTTP client embedded in their software. As HTTP, each ALTO transaction consists of one request and one response.

The ALTO approach of providing network guidance for the establishment of application overlay connections is closely related to the ENVISION approach of providing network information to the application layer. ALTO however is limited to provide cost maps or reordering of a set of preferred nodes, mainly as network operators are concerned of exposing the underlying network configuration to third parties. The preferred routing cost however not always reflects the requirements of a dedicated application, for example some applications need to be optimised with respect to latency, and others try to achieve a maximum throughput. ENVISION thus broadens the scope to allow applications to retrieve information according to their very specific requirements.

Copyright © ENVISION Consortium, May 2011

3.2.1.6 NAPA-WINE

The NAPA-WINE project (NAPA10] proposes a common *handle* that consists of a generalised overlay system which has the inherit property of allowing network side feedback for performance control. This generalised overlay system is reusable by multiple applications concurrently and it shares a common control and management interface across all applications. This allows participating nodes to communicate between each other. Combining that control and management interface with the network side feedback allows both network and peer-to-peer applications to run in a cooperative way.

The NAPA-WINE generalised overlay system is based on the Service-Aware Transport Overlay (SATO) system [SATO07] developed in the European Commission's FP6 "Ambient Networks" Research project. The purpose of SATO is to provide a flexible and customisable transport service to the application layer by using overlay networks on top of the transport layer connectivity. Furthermore, it includes auxiliary services such as distributed lookup services (e.g., DHT).

The Napa-Wine EU project proposes an architecture where P2P-TV clients exploit network measures to reduce their network footprint and, if available, exploit ALTO-like services to optimise topology and performance, and to minimise the overall network load.

Contrarily to ENVISION, NAPA-WINE does not address an explicit interface for enhancing the cooperation between overlay and the networking layer, neither aims at enabling the dynamic activation of network services to reduce the network load.

3.2.2 Security aspects for the collaboration

Behind the collaboration scheme between the application and the network promoted by the project stands a fundamental point concerning the security aspects of the interface. It is intended to address these matters carefully as the collaboration is partially based on critical information and operations involving different parties. The security issues of concern for these interactions that require secure communication involve four processes:

1. Authentication: The authentication is the first step of the security process allowing each side to determine the identity of the other side. We identified the following requirements for the authentication process:
    a. The ISP I needs to identify the application A requesting the collaboration
    b. I needs to identify the entity or entities E executing or managing the application and interacting with the ISP to implement the collaboration
    c. E needs to identify I
2. Authorisation: The second step is to determine the privileges accorded to the application A and entities E. Based on its own policies and agreements with A and E, I defines different level of access to its information and resources. The following requirements are under consideration for the interface security aspects:
    a. Classification of the information into classes of privileges
    b. Classification of the ISP resources into classes of privileges
    c. Granting of privileges to each application and entities
3. Communication: the next step is to secure the communication channel between I and E in a manner that no other entity is capable of intercepting the exchanged data or primitives. It is required that encryption with sufficient strength is used for the exchanged data.
4. Data usage: the final step of the process is to ensure that the data securely exchanged between the communicating parties remain confidential and used intentionally for the

Copyright © ENVISION Consortium, May 2011

collaboration process only, preventing any illegal distribution. This protection might be ensured by:

a. Usage of copy protection techniques

b. Usage of watermarks techniques.

3.2.3 Design considerations for the CINA interface

The design of the CINA (Collaboration Interface between Network and Applications) interface depends on the various requirements that are captured in various subsections of the first technical deliverables D3.1, D4.1 and D5.1 of the project. The first step in designing the interface will be the semantic definition of various generic methods that reflect these requirements.

For example the Consolidated Overlay View function that is described in D4.1 requires the retrieval of information through the interface, such as preferences or peer rankings provided by the ISP, which is similar to ALTO. However the same function also benefits from explicit metrics provided by the ISP. This could be proximity metrics (AS path length, intra-domain hop count), the status of the access network (load, loss, failure events), edge-to-edge performance statistics or further predictions, for example on the expected end-to-end performance. Depending on these requirements as well as what kind of metrics finally will be provided by the Network Monitoring function (see section 4), the next step will be to define generic methods that will be used by the application for information retrieval.

In addition the interface will be used to transmit information also from the application to the network. In particular the Consolidated Overlay View function will perform active and passive end-to-end measurements that can be provided to the ISP. Specifically the knowledge of metrics like delay, throughput and loss appears to be beneficial to the ISP, thus the design of the interface also needs to consider this top-down information flow.

Section 5 Network Optimisation discusses different types of network services that will be invoked by overlay applications. The design of the CINA interface needs to comprise methods that also reflect the setup of these services, such as the instantiation of a cache, a multicaster, a high capacity node or an adaption service. By nature each service potentially has different requirements that these methods need to fulfil. At this point in time it is thus not clear how generic the service invocation methods can be designed.

The security mechanisms which will be investigated for implementation within the AAA functions include the widely-used algorithms, used for authentication and confidentiality based on keys, either asymmetric for public/private keys or symmetric with shared secret keys, and cryptographic hashing algorithms such as MD5 and or SHA-1 for integrity checking. There are several authentication techniques depending on the targeted security strength level, starting from the basic access authentication where the password is sent plaintext over the network to the symmetric key AES algorithm. Particularly, three techniques could be interesting for our system. The first one is the Digest Access Authentication, which is more-or-less vulnerable to some attacks such as Man-in-the-Middle but is widely used specially by webservers. The second one is based on Public keys algorithms where one side has a public key for encrypting its credentials and the other side has the private key, the only one capable of decrypting the data. This technique saves the two parties from the problem of securely exchanging a secret key. The last one is based on a secret key which offers a higher level of security than the others. For securing the communication between the different entities after the authentication, we will investigate the SSL/TLS protocol “Secure Socket Layer/Transport Layer Security”. This protocol is above the transport protocol and is compatible with TCP, UDP and SCTP. It is provides security functions for several higher layer protocols such as HTTP, FTP, SMTP, SSH, etc. SSL/TLS offers the following security services; (1) Data confidentiality using cryptographic algorithms such as DES (Data Encryption System), 3DES or RC2 for block ciphering or RC4 for data flow ciphering (2) Data integrity by using Message Authentication Codes (MAC) based on MD5 or SHA-1 algorithms

Copyright © ENVISION Consortium, May 2011

(3) Identification based on X.509 certificates. For providing authorisation functionality support for the system, we will investigate the use of credentials allowing a specific access level to the data or services.

After the semantic methods have been designed further design decisions can be made. These include the transactional model, the encoding and finally the transport protocol. At this point in time one promising approach seems to be the use of a RESTful HTTP based interface that uses JSON encoding, as done in closely related work as ALTO and OneAPI. HTTP provides several advantages for Internet applications, for example it is easy to read and debug, it can leverage a huge installed infrastructure base (e.g. HTTP caches), many P2P clients already speak HTTP and finally it already provides authentication and encryption mechanisms. Also JSON encoding provides a good mix of features. It is ASCII encoded, thus again it is easier to read and debug than binary encoded messages, and compared to XML the overhead is reduced to a large extend. However the final decisions strongly depend on the final requirements of the overlay-network interaction, thus this is only to be seen as preliminary. In order to reflect the progress in the technical tasks the interface will be refined in several interactions until the release of the final specification at the end task 3.1 in month 21.

Copyright © ENVISION Consortium, May 2011

4. MONITORING

4.1 Introduction

A major task within the ISP is to monitor network data for internal network management and other operational and planning purposes. This data could also be used by the ISP for helping in the peer selection of the application as well as for activation of the network services. Peer selection might take part in the swarm creation process where the initial set of serving peers is chosen, this decision will be based on some monitoring history or momentary status of networks availability and may run some real-time tests at swarm creation. In addition peers may be selected dynamically according to ongoing monitoring results. Unlike applicative monitoring where the network decisions are more "routing" oriented in the overlay layers, the network tries to minimise peer traffic footprint on the network by localising the swarms. This is a challenging task, taking into consideration that close peers may not hold the content a peer is requesting with high probability, however for the live scenarios, this may not be the case, as the subject of the live content may refer to a local target audience. Even though there will be many cases where the swarm will be wide spread over remote located peers of different ISPs, Thus the CINA interface could bridge over several ISPs to optimise network resources and cost. The aim of network monitoring within the project context is to analyse which data or information is needed, what are the available sources of information within the existing networks, what are the future monitoring requirements and systems needed to improve the collaboration between ISPs and overlay services, and what are the mechanisms for delivery of the recorded information over the network.

4.2 State of the art

4.2.1 Measurements methods and standard interfaces

4.2.1.1 Monitoring methods

Two main monitoring methods are commonly used by tools and applications: Passive and Active Monitoring, as defined below.

4.2.1.1.1 Passive monitoring

Passive monitoring is defined as a monitoring technique that does not generate new packets for the purpose of monitoring, thus the application traffic/packets are used to carry monitoring information from source to destination and vice versa. This type of monitoring is most suitable to ongoing sessions, the portion of monitoring information in relation to other application traffic is usually minor (i.e. less than 1%). The main advantage of passive monitoring is that it does not introduce large overheads, bandwidth can be measured and throughput adapted based on delay and loss. The disadvantage is that it is not suitable for monitoring where there is insufficient traffic, thus active monitoring probes can complement passive monitoring techniques in areas of the network with low load.

4.2.1.1.2 Active monitoring

Active monitoring techniques refer to techniques which generate dedicated traffic for the purpose of monitoring. Traffic such as "ping" is an example of the active monitoring technique. This monitoring could be used in order to check connectivity, count number of hops between peers, query peers for their resource availability, collect statistics and so on.

4.2.1.2 IP Performance Monitoring (IPPM)

IP Performance Monitoring (IPPM) was instigated by the IETF's Benchmarking Methodology Working Group (BMWG) of the Operational Requirements Area, and was later continued by the IP Performance Metrics Working Group (IPPM) of the Transport. The aim is to define IP metrics to allow

Copyright © ENVISION Consortium, May 2011

interoperability and better understanding of how the metrics are implemented, and to allow similar interpretation of the results. For more information please refer to https://datatracker.ietf.org/wg/ippm/.

The IETF IPPM working group have already defined 83 metrics, including delay (one-way or two-way delay), packet loss, jitter, one-to-group (multicast SSM), division of metric and composition of metrics:

- *Core* metrics
  - Connectivity (5)
  - One-way delay (6)
  - Packet lost (3)
  - Round-trip Delay (6)
  - Lost pattern(6)
  - Ipdv (6)
- *Ancillary* metrics
  - Periodic streams metric (1)
  - Reordering metrics (12)
  - Duplicate packets metrics (6)
- Spatial metrics (7):
- One-to-group metrics (12):
- Composition metrics (13):

The final decision about which metrics are more useful and how to monitor them will be part of the work over the coming months. We also plan to investigate how IPPM work can be useful for the harmonisation of the reporting of the monitoring information (definition, periods and default values) of the project.

4.2.1.3 Spatio-temporal Available Bandwidth (STAB)

Locating the link with the minimum available bandwidth on a path (also called the tight link) has many important uses. It can help understand where the path limitations are and why, and can also help network-aware applications in choosing the best available server. Real time tight link location information will aid network operators in adjusting traffic routes and detecting network anomalies.

Spatio-temporal Available Bandwidth (STAB) is a tool designed to locate the tight link of a path in space and over time. STAB uses probing schemes that combine the powerful concept of self induced congestion and packet tailgating, and has some significant advantages:

1) No topology information required.
2) Robust to multiple bottlenecks.
3) Efficient.

STAB overview:

Available Bandwidth – Numbering the links on a path 1,2,....N starting from the source (see figure 11, the available bandwidth of a link i is defined as its average unused capacity. The available bandwidth of the path is the minimum available bandwidth of all links comprising the path and will be termed as A(1.N).

Copyright © ENVISION Consortium, May 2011

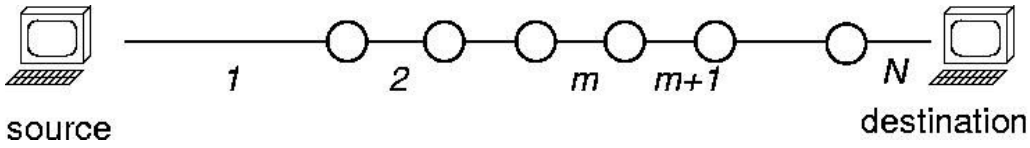

Figure 6: Network path model

Self-induced congestion – The principle of self-induced congestion allows a straightforward technique for estimating A. It relies on the following heuristic: if the probing bit rate R exceeds A then the probe packet will become queued at some router, resulting in an increased transfer time. On the other hand, if R < A, than the packets face no extra delay. A is estimated simply as the probing rate at the onset of congestion.

Chirp trains – In a chirp probing train the inter-arrival time between successive packets decreases exponentially (see figure 12). As a result, chirps rapidly sweep through a wide range of probing bit-rates using few packets. This allows efficient available bandwidth estimations based on the self-induced congestion principle.

Packet Tailgating – Packet tailgating is a powerful technique that provides local information about segments of network paths. It uses special probing trains consisting of large packets interleaved with small tailgating packets. The large packets exit at hop m due to limited TTLs but the small packets travel to the destination while capturing important timing information.

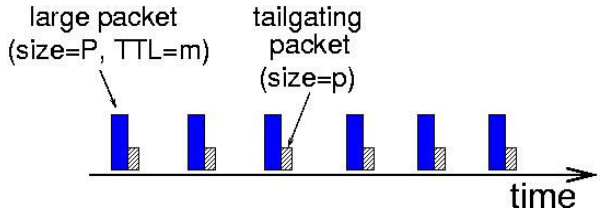

Figure 7: Packet tailgating

Tight link localisation – STAB employs packet-tailgating chirps to locate the tight link of a path. A tailgating chirp is a chirp as described earlier except that each packet is replaced by a large packet closely followed by a small tailgating one. STAB keeps the TTL of the large packets within each chirp fixed. Chirps with large packet TTL *I* provide estimates of A(1,*i*). By varying the large packets TTL from chirp to chirp STAB obtains estimates of A(1,*I*) for *I* = 1,2,…,N. The tight link is the link *I* after which estimates of A(1,*I*) becomes more or less constant.

4.2.1.4 IPFIX/NetFlow

NetFlow is a protocol developed by Cisco which was adopted by many router vendors as well as operating systems other than Cisco IOS. The protocol is specified in RFC 3954- Cisco Systems NetFlow Services Export Version 9 and is widely used in today's networks for collecting IP flows data. The NetFlow records are distributed through either UDP or SCTP to the NetFlow collector. Recently it was suppressed by the Internet Protocol Flow Information eXport (IPFIX) protocol defined in RFCs 5101 and 5102, which is not as common yet as the NetFlow protocol.

Copyright © ENVISION Consortium, May 2011

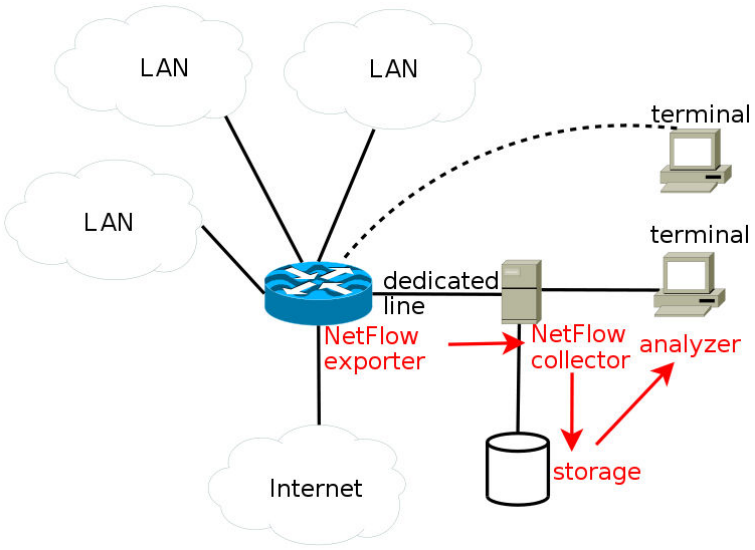

Figure 8: NetFlow architecture

The protocol suggests the following identification for a unidirectional "flow" based on 7-tuples attributes as follow:

- Source IP address
- Destination IP address
- Source port for UDP or TCP, 0 for other protocols
- Destination port for UDP or TCP, type and code for ICMP, or 0 for other protocols
- IP protocol
- Ingress interface (SNMP ifIndex)
- IP Type of Service

By analysing flows data, which contains the number of bytes and packets observed in the flow, one can obtain a good knowledge of the network load. However using the NetFlow at the routers also has some drawbacks related to the performance and security at layer 3, thus alternative solutions such as DPI monitoring over the links, independent from the routers complements the router monitoring approach.

4.2.1.5 Operations, Administration, and Management (OAM)

Operations, Administration, and Management (OAM) tools are deployed in all the networks, it's main use is to monitor network devices, each in it's own area of operation, they can be used to monitor IP traffic and load of devices at the routers, however usually they will not provide application specific or even IP flows specific information.

4.2.1.6 Simple Network Management Protocol (SNMP)

SNMP is a solution widely used by organisations to manage their own networks. Within an organisation, we do not have any longer issues related to security, authentication and access to network equipment since it is an internal management entity. All network equipment (or a lot of them) now support the SNMP protocol and host an SNMP agent and its associated MIB

Copyright © ENVISION Consortium, May 2011

(Management Information Base). Thanks to its wide deployment, SNMP is the less disruptive solution for collecting network information within an ISP.

4.2.1.7 Deep Packet Inspection (DPI)

Unlike OAM and NetFlow, Deep Packet Inspection can provide information regarding the application through deeper analysis of the packets payload and not just at the headers level, there are tools to be deployed externally to the routers on links operator decide to monitor. The monitoring is done in real time and can provide the information such as what is the portion use of link resources used by specific application or service and so on.

4.2.2 ISP ALTO configuration and monitoring

Recently in Oct 2010 the ALTO group has published Internet draft named "draft-alto-deployment-00.txt" that can be viewed at http://tools.ietf.org/html/draft-alto-deployment-00, the document describes ISP ALTO Configuration & Monitoring specifications and metrics.

ISPs like China Telecom as well as application providers have shown interest in deploying commercial ALTO server and services, and the deployment involves both ISPs and network applications. The ALTO WG has identified four major related issues in ALTO deployment to be addressed:

- How does an ISP deploy and configure its ALTO servers? Specifically, an ALTO Server provides the Network Map and the Cost Map. How does an ISP configure these maps? Where does an ISP deploy ALTO servers?
- Which application entities fetch ALTO information?
- How does an application integrate ALTO information into its decision process?
- How does an ISP (potentially with collaboration from applications) monitor the deployment of ALTO, so that the ISP can better understand the status as well as the policy impacts of its ALTO deployment?

The configuration and monitoring draft focuses more on the ISP perspective.

4.2.2.1 ALTO deployment monitoring

In current ALTO architecture, there is no method to understand the ALTO service running conditions. For example, it is known that the ALTO service aims to reduce network resource consumption and accelerate download rate, but, how much traffic can be reduced? is it effective... and so on. Thus the group identified several metrics which are also important for ENVISION, as some of them could benefit from enhanced interface, moreover the aim of this project is to show that benefit could be achieved thus adopting ALTO metrics as initial set would be probably a good choice. At a later stage, we will need to see how we can use these or alternative metrics to the mutual benefit of both ISPs and overlays.

4.2.2.1.1 ALTO P2P monitoring metrics from network

4.2.2.1.1.1 Inter-domain IP traffic

To evaluate how ALTO service can reduce network resource consumption, the total traffic information has to be known. For example, what are the resources used before and after deployment. Deploying an ALTO service in one domain of network, we should know the inbound and outbound traffic condition of this domain.

4.2.2.1.1.2 Inter-domain application traffic

ALTO is aimed at optimising application layer traffic. Thus, this metric is very important in ALTO service deployments. The ALTO service mainly reduces inter-traffic between domains through the localisation of the traffic, so the inter-domain application traffic before deploying the ALTO service

Copyright © ENVISION Consortium, May 2011

should be greater. After deploying the ALTO service, this metric can give the main result of ALTO service. This metric is always used with metrics 4.2.2.1.1.1 and 4.2.2.1.1.3.

4.2.2.1.1.3 Inter-domain application traffic

One of the ALTO service primary functions is application traffic localisation. This metric can provide the efficacy result of ALTO service and different policies. This metric is always used with metrics 4.2.2.1.1.1 and 4.2.2.1.1.2.

4.2.2.1.2 ALTO P2P overlay monitoring metrics

4.2.2.1.2.1 Peer user distribution in domain

In P2P overlay, peer distribution information can provide some hints to understanding the overlay architecture. This information can be helpful to understand the P2P overlay. We plan for the overlay to take an active part in this metric derivation, through information exchanged over the CINA interface.

4.2.2.1.2.2 Peer list condition

In P2P application, every P2P client needs to be connected with other P2P clients. The list of these P2P clients is named as peer list for one request per client. The ALTO service aims to achieve a more localised peer list within network segments. Thus, this metric can be helpful to give the efficiency result of the ALTO service and different policies.

4.2.2.1.2.3 Global average download rate

This metric provides the application performance directly. Download means inbound traffic to one user. Global average means the average value of all users download rate in one or more domains. We can compare this metric before and after deployment of the ALTO service in one domain to understand the ALTO service. ENVISION may want to improve the granularity of this metric: for instance to obtain the average download rate per flow.

4.2.2.1.2.4 Global average connection hop

This metric provides the application performance indirectly. One of ALTO protocol's aims is to reduce the network path hops of one data connection. This metric gives the average hops of all connections in one or more domains. We can compare this metric before and after deployment of the ALTO service in one domain to understand the ALTO service. ENVISION may wish to use such a metric when creating swarms, since we aim to minimise the delay in a tree based swarm.

4.2.2.1.2.5 Global average cache time

This metric is for streaming applications and refers to the time between when a client starts to download streaming packets and until the streaming content is first played on the player. Global average means the average value of all cache time in one or more domains. Comparing this metric before and after deployment of the ALTO service in a domain would increase the level of understanding of the ALTO service.

4.2.2.1.2.6 Jitter time

This metric is not well defined beyond its name as it is currently work in progress, however we assume it means the mean jitter across all the flows of a specific service.

4.2.3 Commercial network monitoring tools

Some of the available tools for network monitoring are provided in the list below:

SNMPc is a secure distributed network management system which delivers real-time monitoring for network infrastructure. SNMPc monitors SNMP devices including WAN/LAN links, servers and

Copyright © ENVISION Consortium, May 2011

applications. IT support SNMPV1, V2 and secure SNMP V3. SNMPc has programming and scripting interface and provides automatic reports.

Argus processes packets into detailed network flow audit data for operations, performance and security management.

Cacti allows a user to poll services at predetermined intervals and graph the resulting data.

cFosSpeed performs traffic classification and lets the user display, shape, tag or rate-limit protocols or programs under Windows.

FlowMon by INVEA-TECH is a complete solution for NetFlow monitoring and analysis including probes up to 10 Gbit/s, collectors and other supervision systems.

InterMapper Originally developed for the Macintosh Classic in 1994 by the network manger of Dartmouth College this application uses SNMP, Ping and NetFlow to build a graphical network map similar to HP Openview which shows bandwidth usage by port information and protocol. VLAN aware.

OmniPeek is an end-to-end network monitoring solution, offering support for many packet adapters and remote collectors.

Observium is an auto-discovering network monitoring application focusing on extensive data collection and graphing of network infrastructure.

PRTG runs on Windows, with graphical and web interfaces. It captures packets using Cisco NetFlow or packet sniffing or uses SNMP to monitor bandwidth usages.

PacketTrap Networks - Traffic and Traffic Flow Analyser

Scrutinizer NetFlow and sFlow Analyser provides deep visibility into network traffic behaviour and trends. Leveraging NetFlow, J-Flow, and sFlow data, NetFlow Traffic Analyser identifies which users and applications are consuming the most bandwidth.

Sandvine Intelligent Network Solutions measure and manage network traffic using Policy Traffic Switches

Copyright © ENVISION Consortium, May 2011

4.3 ENVISION monitoring

In this section, we list the requirements and the network information and metrics that might be provided by ISPs to the overlay applications. The information is provided by each ISP to the overlay and it is limited to local information retrieved from the ISP's own domain only. An ISP has then only to monitor its network and does not need to collect and share information from/with other ISPs; it is the role of the COV (Consolidated Overlay View) function at the overlay level to collect information from different ISPs and aggregate them appropriately (Section 2 of D4.1).

4.3.1.1 Requirements

- The ENVISION monitoring system shall collect information and derive metrics from the list provided in section 4.2.2.1.2.
- The ENVISION monitoring system shall explore new sources of information through collaboration with the overlay applications.
- The ENVISION monitoring system shall identify and define new metrics to enrich the list already identified by the ALTO group.
- The ENVISION monitoring system shall monitor metrics per IP flow to support dynamic swarm functions, distribution functions, and content adaptation functions.

4.3.1.2 Metrics and network information

The ENVISION monitoring system will provide overlay applications important and useful network information. In this section we will examine what information ISPs may provide, and applications may need, what real-time network performance should be monitored and what network capabilities should be shared with the overlay application layer.

The ENVISION monitoring system will have two major roles:

1) Provide applications with network capabilities information.
2) Collect metering data and provide applications with real-time (or near real time in some cases) data of network performance (including, but not limited to: available bandwidth, jitter, latency etc.).

4.3.1.2.1 Network capabilities information

Overlay applications may change their behaviour according to network capabilities. For example, a video stream server may switch from multicast to multi-unicast if the number of multicast groups exceeds the maximum multicast streams supported by the network layer. The CINA interface will enable applications to perform network initial discovery during start up and periodic updates during runtime.

Via the CINA interface, ISP may provide the following information to overlay applications:

- Network supported services: Multicast, caching, adaptation, ads insertion etc.
- Network topology information: IP schemes, static and dynamic route information etc.
- Network statistics: average delay, average jitter, average hop count, average network utilisation etc.
- Extended Service information: Specific services may have additional information. For example, multicast service extended information will include: Maximum number of multicast groups one router can manage, Accounting and billing metric from multicast use (per sender/ per number of receivers/per duration) etc. Caching service extended information will include average cache hit ratio, cache TTL limit, cache policy etc.

Copyright © ENVISION Consortium, May 2011

4.3.1.2.2 Metrics

Network performance may vary in time and depends on user's activities, time of day, environmental conditions and more. Real time network performance information is critical to the application layer while selecting new peers or choosing a server to work with.

In order to provide the overlay application layer with metric information, ISP will monitor and collect the following information and may provide it to the application layer:

- ALTO identified metrics as defined in section 4.2.2
- Bandwidth: Available Bandwidth and utilisation.
- Latency: delay between access routers.
- Jitter: variance of latency between routers.
- Packet Loss.

In the next phase we will define metric measurement mythology (active or passive), technology, update frequency etc.

Based on the basic principles of ALTO, ENVISION will follow a similar way and the ISP will help the overlay to select the better than random peers. For this the ISP will return back to the overlay the list of preferred peers. The ranking decision is based on following information, which have to be monitored by the ISP but will not be made public to the overlay or others entities:

- Network topology at granularity the ISP wants: e.g. know bandwidth, delay between access routers of its network, etc
- Number of hops between access routers
- Packet loss ratio
- Peering agreement with others ISPs (AS Number)
- Cost for links between access routers
- List of explicit caches in the network and their IP address

The ISP can return back the list of ranked IP address but depending on the network status, it can also give the indication to the overlay to switch to a multicast tree.

4.3.1.3 Monitoring with COMET

In this section, we outline the plan for collaboration of ENVISION with the COMET project [COMET]. We identify the adaptations needed in the COMET's Content Mediation Plane (CMP) to interface with ENVISION and provide the network metrics based on static information as well as dynamically monitored information from the COMET testbed.

While both projects focus on various aspects of digital data content in the Internet (content access, dissemination, delivery etc), the high-level approaches employed are different. The COMET project aims to solve the problem via an overlay at the network level resulting in a 2-plane approach aiming to mediate the delivery of Internet content via native COMET network entities. On the other hand, the ENVISION project deals with the problem by developing techniques for the content delivery at the application layer and by fostering the collaboration between the applications and the underlying ISP networks to achieve the co-optimisation of the often misaligned application and network performance objectives.

To enable collaboration between the two projects, we envisage the creation of a COMET-ENVISION interface that enables communication between the relevant entities from both sides. Via this interface, we foresee that COMET can supply network performance information monitored within

Copyright © ENVISION Consortium, May 2011

the COMET system to the ENVISION system, thus allowing the related ENVISION functions to perform their optimisation in a more timely and informed manner.

Specifically, the COMET-ENVISION interface is expected to be located between the Server and Network Monitoring Function (SNMF) of COMET, and possibly the Path Management Function (PMF), and the Network Optimisation Function (NOF) of ENVISION – a subcomponent of the Network Management block (block 9 in Figure 2 on page 9).

The candidate information that the SNMF (and PMF) can provide to NOF through this interface includes:

- Long-term inter-domain topological information (e.g. IP addresses of the routers in peering links, peering AS numbers)
- Routing information per destination network, specifically a set of paths per destination network prefix, each path consisting of the following information:
  - COMET Class of Service (CoS) supported along the path
  - Path length expressed in terms of number of domains
  - The list of domains on the path
  - A vector of QoS parameters characterising the path (maximum packet loss probability, maximum delay)
- Load of inter-domain peering links per CoS and the maximum bandwidth per CoS (according to the agreed peering SLAs)
- Long-term intra-domain edge-to-edge QoS parameters (according to the network planning and provisioning). These parameters are used in the routing awareness process in COMET to exchange network reachability information with other domains.

Due to the sensitivity of some of this information, we do not presume that ISPs by default agree to share it with other operators. However, both the SNMF and the NOF are owned by the same ISP. The NOF would use this information directly to improve the network optimisation itself and/or to modify the information exposed to the application, without necessarily revealing the raw information received by SNMF.

Other specifications of the interface (e.g. the messaging format, the interface technology (push/pull)) are dependent on the implementation of both projects and thus will be defined later.

Copyright © ENVISION Consortium, May 2011

5. NETWORK OPTIMISATION

5.1 Network services

This section identified the possible network services that map with the ENVISION objectives and could be useful for both the applications and the network. Several services have been envisioned, some more realistic or feasible than others, some more related to the use-cases defined in WP2. The next section details the most relevant network services for ENVISION and the specified use-cases and will be deeply investigated, specified and designed within the project. Other lower priority services are listed for information in the following section.

5.1.1 Preferred network services being investigated in detail

Five main network services have been identified as most relevant for ENVISION and the use-cases defined in WP2.

5.1.1.1 Multicast-related delivery

This network service concerns multicast data delivery and encompasses several options.

Firstly, one possibility is to dynamically set up a multicast tree (at the necessary scale, might be in a local region or wider) when the number of users in the region exceeds a given threshold. In this case, the overlay application might request the ISP, via the CINA interface, to establish a multicast tree from a given source and towards given destination endpoints. This service will allow improving the QoE delivered to end-users as the number of clients grows but also to reduce the network load, removing many unicast streams for a multicast delivery.

Since some ISPs may be reluctant to allow end-users to directly send multicast traffic in their network, an alternative approach is proposed based on the introduction of a multicast service in the ISP network: this “multicaster” receives unicast streams and forwards them in multicast.

Associated with this multicaster, a unicaster is under specification. This service does the opposite than the multicaster, i.e. receives a multicast stream and forwards the data in unicast toward end-users. The unicaster service is mostly useful for users that could not receive multicast streams, either because of the device capabilities or the network equipments.

An alternative is to integrate in the ISP network some specific nodes, called high capacity nodes in this document, adapted to application level multicast i.e. able to integrate the overlay protocol, and to efficiently replicate a unicast stream towards several destinations.

In ENVISION, we will study how these network service nodes can be used by the overlays, evaluate the impact on the content delivery, look at the quality we can provide, using SVC streams or FEC mechanisms, and also study more specific topics as for instance how to select the best source for multicasting etc.

Section 5.2 details the ongoing activities related to multicast.

5.1.1.2 Content adaptation service

To deliver a good quality to end-users accessing an application using various devices and access networks, content adaptation is required. For this, in the ENVISION project, we design content adaptation service node in the network so that the content can be dynamically adapted to end-users context. The context encompasses the users' profile, devices capabilities, network conditions, etc. All information on the context is grouped within the so-called metadata, which are application metadata or network metadata. This network service is defined in close relationship with WP5.

Several adaptations might be imagined; removing some enhancement layers in SVC streams, adapt the format and container of the data, use FEC mechanisms, etc.

Copyright © ENVISION Consortium, May 2011

Section 5.3 describes this network service.

5.1.1.3 Caching

In WP3, one of the main objectives is the network optimisation. Caching content is then a natural option for saving bandwidth in the network, and delivering content more rapidly to end-users.

Within the project, the caching functions will be investigated in 2 different tasks: one in Task 3.3, where the caching will mainly be related to the network issues and in Task 5.3, which will study it rather at the application level.

In the project, we investigate how caching nodes can be efficient, where to best locate them, which network option to prefer, etc.

Typically, there are currently two main approaches for network caching: *transparent caching*, where the ISP deploys caching nodes in its network that will cache content and deliver them in a transparent way to end-users. In this sense, transparent means that both end-users and content providers are not aware of the caching nodes. Opposite to this approach, the *explicit caching* technique is different since in this case, the caching nodes are clearly known by end-users and the latter connect to the caching node to get the content. In this case, the nodes might be seen a super-peers.

Section 5.4 presents two approaches with description of current solutions.

5.1.1.4 QoS-based services

The project should ensure the quality provided to end-users. Towards this goal, a network service is envisioned which aims at managing the network so that to ensure the QoS to end-users. For this, several options are possible. One is related to the management of users' bandwidth by dynamically changing, reserving it. Another possibility is to have a function which could manage the delivery of content over different links (in case the device is connected to multiple access networks) and thus share the data delivery over the links according to what the networks can do and the expected QoS. A third option for ensuring quality is to have a mapping between application quality requirements and network level quality so that content could be delivered with the expected QoS. This could be done by configuring the network with different priorities (ala DiffServ) or by treating data differently (e.g. prioritise core layers of an SVC stream and have others network paths for enhancement layers).

Section 5.5 presents this network service.

5.1.1.5 Network optimisation logic based on preferences: Preferences announcement

The cooperation between overlay applications and the underlying networks leads to a closer communication between both actors. In this network service, the main idea is that an ISP could announce some preferences to the overlay, based on information gotten via monitoring tools or based on its own policies. For example the ISP can detect that some parts of its network is overloaded by unicast sessions from the same application broadcasting live content, and thus can inform the overlay of the possibility to switch to a multicast-based delivery. One work for this network service is to define the logic that will enable to make decisions with the overall aim to optimise the delivery of data, from the network point of view.

Section 5.6 details this network service.

5.1.2 Additional network services to be specified at a high-level only

In addition to the five services described in the previous section many other network services could be provided by an ISP. Indeed, the availability of an interface between applications and network

Copyright © ENVISION Consortium, May 2011

providers opens a plethora of business and technical possibilities. Applications can make use of these services to drastically improve end-user quality of experience, to improve resilience and security or to implement different business models potentially sharing revenue with network providers.

Here we provide a list of other examples we find interesting, ranging from small incremental evolutions to more innovative paradigms, that will not be investigated in detail (either because they are less challenging or because they are not well suited to the use-cases we have defined), but they are mentioned here since they are candidate extensions ISPs could provide through ENVISION's CINA interface.

5.1.2.1 Traffic prioritisation

This service is about prioritisation of traffic in case of congestion in the network or to ensure a required QoS. This service has relation with the one we will address (QoS-based services) but is more limited and the research work is not so innovative, compared to what currently exist (such as DiffServ for instance).

5.1.2.2 Resource reservation

Although end-to-end flow resource reservation has proven unscalable in the Internet, there are some situations where this might be possible. These include last mile resource reservation where end receivers/transmitters ask for a larger share of the spectrum (ADSL, optical or wireless) to receive/transmit bandwidth rich content. Another scenario is for extremely large data transfers (digital cinema distribution and financial sector data backups are two examples) that can be scheduled in advance and for which the network provider may have to take special actions (for example setting MPLS paths).

5.1.2.3 Content aware policy and security issues

A basic principle of the Internet is that everybody can send to everybody. This has proven to be a double edge sward. The Internet's openness allows for easy communication but it also makes security much harder. The last decade has seen a dramatic increase in attacks (capability exploitation, denial-of-service, etc) that is putting increased pressure on several Internet businesses. Mobile devices are particularly vulnerable since low bandwidth attacks can denial service easily to any application running on them. Again, the CINA interface can be used to actively stop flows being transmitted to the device or, in an extreme scenario, switch to default-off mode of operation where only authorised senders can communicate with the device.

This network service can even be enlarged to wider range of actions an ISP can have regarding content. Typically, such nodes in the network may also take specific actions directed by ISP policies, depending on a range of observed the network conditions and not just when DOS or other attacks are observed. Amongst the possible actions, we can mention redirection of packets, blocking of packets, marking, etc.

5.1.2.4 Geolocation

One main service an ISP can provide is geolocation of end-users. Indeed, for instance, in several wireless scenarios, the network provider is in an advantageous position to determine the geolocation of a given node. For example in closed environments like sports events or shopping malls where there is no line of sight to GPS satellites, the only possible way to determine user location with some accuracy is through cell masts triangulation. This information can be passed to the application through the CINA interface. This service might be of importance for geo-localised services, emergency services, advertisement, etc. This service has been envisioned within the project. Even if possible to include in the CINA interface, it has not been selected because it was not required for the defined use-cases and it has been decided to focus on the network services that would really optimise the network. This service is an additional one that can be proposed by ISP.

Copyright © ENVISION Consortium, May 2011

5.1.2.5 Audience measurement

Audience measurement is a key aspect for content providers. With distributed applications, such as the ones we can have in ENVISION, measuring the audience for content is not straightforward. If the context allows for the network provider to know what content each user is consuming, this can be used, as an aggregate to feedback into the marketing processes of content producers. Even when the network provider does not have access to this information he can use time and data type (port numbers) access information to measure some audiences. It could have a real benefit for applications and might be a service an ISP can monetise. Obviously, all these potential business models need to carefully take into account privacy and legal concerns.

5.1.2.6 Dynamic tunnelling between groups of devices towards destination

In relation with the previously introduced QoS-based services, it has been discussed to have a network service that could establish tunnels between groups of devices towards a destination. This could be complementary to the QoS-based service. But it was not convincing what ENVISION could bring as new in such research topic and we believe that current solutions such as tunnels, or VPN might be applied as is in the ENVISION architecture.

5.1.2.7 Network resilience

Network resilience is another service that network providers can easily make available to applications. Although resilience was a design requirement of the original ARPANET, quality of service in the current Internet is often poor due to the slow convergence of routing protocols when network links fail. Network providers have several ways of achieving extra resilience but often at a cost (management or equipment). If certain applications (e.g. remote medical surgery) need significantly higher levels of resilience they can request this through the CINA interface.

5.1.2.8 Ad/text insertion

Given that the network provider is often the only entity with information about the client, he can make use of this privileged position to insert personalised ads in the data stream. A network service that can dynamically insert ads (e.g., personalised ads for instance) or text (e.g., subtitle or news or weather forecast, etc.) in application data via nodes in the network, has then been discussed. This service might be very interesting for the applications and can nicely illustrate the ENVISION concepts.

Copyright © ENVISION Consortium, May 2011

5.2 Multicast

IP multicast is considered as an efficient mechanism to deliver large-scale content over the Internet, especially for video streaming, to save bandwidth and reduce end-to-end delay.

Application level multicast (ALM) has been proposed as an alternative to IP multicast, to overcome the current limitations on end-to-end deployment of native IP multicast. In ALM, some group members form an overlay network and content is distributed via unicast by relaying packets from one node to another. Peer-to-peer (P2P) networks are one of the examples of such overlay networks. The benefits of ALM however cannot be compared with native IP multicast, as in the former the traffic at the overlay traverses the underlying network many times reducing thus the amount of savings, and the multicast tree topology cannot be optimised without the knowledge of the underlying network topology. Recent research investigations have focused on overlay graph optimisation and efficient routing protocols' design, to minimise the transmission delay for real-time applications [RAT01]. Most of the proposed solutions hide the underlying physical topology and do not take advantage of the IP native multicast capabilities of the network.

In ENVISION, we propose the use of hybrid multicast, to take advantage of native IP multicast capabilities where and when possible and to relay the content via overlay unicast links over these parts of the network where no such capabilities are supported. The objective is to define an interface between ISP networks and overlay applications, such that an application which transmits content to a large number of receivers in an ISP network can benefit from IP multicast, and to study the conditions and the optimisation functions that are related to the use of IP multicast in the network and in the overlay layers.

5.2.1 Multicast terminology

ENVISION considers native IP multicast as well as application level multicast. To avoid ambiguity, the following terminology is used:

- Overlay multicast: a content distribution overlay topology that uses unicast IP streams between the overlay nodes to distribute the content; the nodes create a hierarchical tree topology with the purpose of reducing the load at the content source of uploading to content sinks.
- Hybrid multicast: the combined use of IP multicast (when available) and unicast in a content distribution overlay topology.
- Cooperative hybrid multicast: a hybrid multicast topology that is the result of information and control messages exchanged between the application and the ISP through the CINA interface.
- Transparent hybrid multicast: a hybrid multicast topology that is the result of the ISP replacing some unicast overlay connections with IP multicast connections without the knowledge of the application.
- Multicast source: a node that sends traffic in multicast; this node might be an end-host provisioned by the application outside the core network, or a network node provisioned by the ISP.
- Multicast leader node: an overlay node that is responsible for sending the complete or part of the data associated with a particular content item transmitted over multicast; if the leader node implements multicast then it is also the multicast source, otherwise the multicast leader node sends the traffic in unicast to the multicast source.
- High capacity node: a node that is integrated in the overlay, serving a multitude of peers with unicast IP streams.

Copyright © ENVISION Consortium, May 2011

5.2.2 State of the art – multicast current deployment

Multicast has been the subject of numerous works since the early 1990s, from which has emerged a de facto standard: the PIM-SM protocol (Protocol Independent Multicast – Sparse Mode). Despite the maturity of this protocol, and the fact that multicast is the most efficient way to simultaneously distribute the same content to a large number of receivers - by optimising the network bandwidth usage and saving application server processing resources - there is at the present time no commercial deployment of multicast in the Internet. Multicast is currently used by ISPs for their managed IPTV services, and is generally proposed in business VPN services but to our knowledge none of the ISPs provides a larger, more open, use of multicast to its residential customers.

Analysis of the reasons of the slow deployment of multicast in the Internet can be found in several papers, see [DIO00] for instance. One of the main reasons put forward is that the adopted protocol relies on an open service model, which does not integrate the control functions required for an operational deployment. Indeed, there is no group management mechanism in the PIM-SM protocol: any source can send traffic to any multicast group, any sender can subscribe to any group. Without access control, attacks against the multicast data and control plane are easy to implement. Solutions have been proposed to set up the necessary access control, as described in section 5.2.2.2) but at the present time none have been standardised nor implemented by equipment vendors.

Other reasons are also put forward, for instance the lack of a business model for the ISPs, or the lack of support for network management.

5.2.2.1 The PIM-SM protocol

5.2.2.1.1 PIM-SM ASM

The PIM-SM protocol can be used in two modes: the ASM (Any Source Multicast) mode proposed for many-to-many communications, and the SSM (Specific Source Multicast) mode which can be seen as a simplification of the ASM mode for one-to-many communications.

The ASM mode, which corresponds to the original functioning of PIM-SM, relies on a rendezvous point (RP) in the network. In a first phase, a end user join a shared multicast tree rooted at the RP by subscribing to a multicast group G, typically by sending an IGMP report (*,G) message. All the routers know (by configuration or by using dedicated protocols) the IP address of the RP associated to the multicast group G. Thus PIM join (*,G) messages are relayed towards the RP which, in the reverse direction, builds a multicast tree for the group G. This tree is called “shared tree” or rendezvous-point tree (RPT). End users who have subscribed to the group G receive all the data sent by any source towards G.

When access routers receive some multicast traffic from a source S towards a group G, they encapsulate the traffic in “PIM register” unicast messages and send it towards the RP. When receiving PIM register messages, the RP decapsulates the traffic and sends the multicast flow in the shared multicast tree associated to G.

In a second phase, in order to avoid the unicast encapsulation/decapsulation process which may be CPU consuming, the RP joins the multicast tree rooted at the source (the shortest path tree or SPT) and when it receives the traffic natively in multicast it requests the access router to stop encapsulating the traffic in PIM register messages.

It must be noted that receiving traffic from the SPT tree reduces the risk to receive traffic from a forged source since PIM-SM uses the mechanism of “reverse path forwarding” (RPF): the multicast traffic received on a router interface is forwarded only if this interface corresponds to the shortest path to the source. Thus forged traffic can only be sent by a malicious source on the same sub-network as the original source.

Copyright © ENVISION Consortium, May 2011

In a third phase, access routers which receive a multicast flow (S,G) through a shared tree can decide to switch to the SPT to optimise the path taken by the traffic and avoid any detour by the RP. The switching to the SPT is described as optional in the specification. In practice, the switching to the SPT is triggered by routers when the bit rate of a source passes a configured threshold. Thus it can be noted that the use of a shared tree to carry the traffic depends on the router configuration (typically, if the switching threshold is set to zero, the shared tree is mainly used to learn new active sources).

5.2.2.1.2 PIM-SM SSM

To avoid the complexity of the ASM mode in the case of one-to-many communications, in particular when the source address can be easily known at the application level, the SSM mode has been introduced working on a specific range of multicast group addresses (232/8). The receiver end user directly joins the SPT tree rooted at the source by sending a IGMP report(S, G) message containing the multicast flow source address, there is no more RP to configure.

When an access router receives some multicast traffic from a source, it forwards the traffic only if it has received PIM join messages for this traffic, else the traffic is dropped.

5.2.2.1.3 PIM-SM security issues

As already mentioned PIM-SM (in ASM or SSM mode) is inherently vulnerable to receiver attacks, since without complementary access control mechanism, any end user can create a huge number of multicast states on routers by joining numerous groups. Since the number of multicast states which can be handled by equipments is limited, currently around few thousand states, it is necessary to control the end user subscriptions. This is currently done by statically configuring on the access equipments a list of authorised multicast addresses. Several works exist to allow a dynamic access control, as presented in section 5.2.2.2.

PIM-SM in ASM mode is also particularly vulnerable to source attacks. Without access control, any source can carry on an attack by sending unwanted traffic towards a group, possibly flooding the group and creating congestion. The current solution to solve this problem consists in configuring (statically) a list of authorised sources on RPs.

It must be noted that the multicast control plane can also be threatened by source attacks: the workload of an access router is dependant of the multicast traffic received since this traffic triggers encapsulation in PIM register messages towards the RP. Thus controlling the allowed sources by access lists at the RP is not sufficient. As long as dynamic mechanisms are not available to control the authorised sources at the access, the solution is to rate limit the PIM register messages from access routers to RPs.

The SSM mode is generally considered safer than the ASM mode, in particular the security threats on the multicast control plane which can be caused by a malicious multicast source trying for instance to overburden the RP with PIM register messages is avoided, the control plane is no more dependant of the traffic sent by sources.

A more detailed study of the security issues of PIM-SM can be found in the IETF RFC [SAV06] (only attacks against the multicast routing infrastructures are considered). Solutions exist to mitigate the security issues of PIM-SM and protect the network, but today they only rely on static access control, which requires configuration efforts, and as the ASM mode is concerned do not protect the service itself.

5.2.2.1.4 Advantages and drawbacks of the ASM and SSM mode

The advantages of SSM compared with ASM have been extensively described when SSM has been proposed, see [ALM01] for instance. As underlined in the previous sections, the relative simplicity of the SSM mode and above all the security issues specific to the ASM mode are a strong reason why an ISP could prefer deploying PIM-SM in SSM mode rather than PIM-SM in ASM mode.

Copyright © ENVISION Consortium, May 2011

From an application standpoint, using the SSM mode instead of the ASM mode does not cause any difficulty when the sources can be known at the application level and are relatively stable.

The use of shared trees rooted at rendezvous points nevertheless simplifies the handling of source mobility and dynamicity (at least inside a domain):

- Handling the mobility of a source implying an IP source address change from S1 to S2 is easier: with ASM there is no need to inform the receivers at the application level that they have to subscribe to the flow (S2,G) like in SSM
- In the same way it is easier to replace a source S1 by another source S2 since there is no need to inform the receivers at the application level that they have to subscribe to the new (S2,G) flow.

From the network standpoint, the use of shared trees allows to economise multicast states in equipments when several flows are requested by a same set of end users.

The first work introducing the channel model (i.e. subscriptions to the combination of a source address S and a group address G), called EXPlicitly REquested Single-Source (EXPRESS) multicast and which has inspired the PIM-SM SSM mode, integrated a solution to support multi-source applications. [HOL99] describes how EXPRESS can be extended at the application or middleware level by the introduction of session relays (SRs) that act as the source for the EXPRESS channel (SR,E) to which each participant in a multi-source session can subscribe. The SR can use an application-layer relay protocol or an IP-in-IP-like encapsulation with application level access control on the encapsulated forwarding. The benefits provided by the relay structure cited in the paper are the following:

- An application can select the placement of SRs to minimise communication, contrarily to the PIM-SM ASM mode where applications have no control of the RPs they use since it depends on the network configuration.
- An application can select to use additional backup SRs for fault-tolerance.
- The SR can provide application-specific functionalities beyond simply relaying data and transmitting notifications of new sources, for instance application-specific access and content control, or it can add sequence numbers to relayed packets, as required in some reliable multicast protocols.

Note that the fact that an ISP can provide one or more well-positioned session relay servers as a value-added service for customers is underlined.

Thee idea to build the IP model of ASM on top of SSM by using an SSM proxy service is took up in [ZAP01] with the objective to mitigate the delay from senders to group members due to the detour through the session relay. The approach proposes to allocate several proxies instead of a single one for each group communication session identified by a group address G: each receiver joins the nearest proxy Pi by subscribing to (Pi,G), each sender sends traffic in unicast to the nearest proxy, the traffic being distributed between the proxies in unicast or multicast. The list of authorised senders for a group is coordinated by a primary proxy associated with the group initiator, which advertises other proxies. The interest of the method in terms of delay and bandwidth usage is shown through simulation by comparison with the results given by the approach using (unique) session relays and with the approach based on shortest path trees towards each source. The paper also describes how new proxies can be dynamically allocated to group sessions.

5.2.2.2 Multicast dynamic AAA

A previously mentioned, today only static access control lists describing which multicast group a receiver can access, or which source can send traffic can be configured on equipments.

Copyright © ENVISION Consortium, May 2011

Several works exist aiming at providing a more dynamic access control. The in-progress IETF draft [HAYdraft10] describes the functions required for a multicast authentication, authorisation and accounting (AAA) unit coordinated between content provider(s) (CP) and network service provider(s) (NSP). Based on these requirements the in-progress IETF draft [SATdraft12] describes a possible architecture (see also [SAT05]). The general idea in this architecture proposal is that when a NSP receives a request from a multicast receiver, the NSP firstly requests authorisation to the content provider, and then the NSP verifies that enough resources are available in the network before accepting the receiver request. Other propositions exist based on similar approaches, see [ISL06a] for instance.

[CASdraft06] proposes the Access Right Distribution Protocol (ARDP) whose finality is to provide full control to content providers (CP) to maintain the access right database integrity on NSP aggregation equipments in real-time by multicasting their service plane and access right policy over the ARDP Backbone (A RSA signature is used to guaranty the protocol datagram authenticity and integrity, as well as datagram sequence numbers). Roughly, the NSP provides a namespace and a point of presence in the ARDP backbone to each trusted CP. CP ARDP servers permanently multicast the CP service plane and multicast the access rights associated to a clientID when requested by the NSP. With the use of clientIDs and multicast, the NSP keeps the information on its topology. The NSP also relies on multicast to distribute information on clientIDs and NSP parameters (IP addresses, MAC addresses, etc).

Alternative approaches propose to modify the multicast access protocol so as to include AAA parameters in the subscription messages (see [ISL06b]). Note that an evolution of IGMP is a long term solution since it requires the evolution of access nodes as well as the evolutions of terminals.

As it can be seen, many works have been proposed to define a dynamic access control mechanism for IP multicast, the necessary building blocks have been described, and even if some protocols for exchanging data between the ISP and the CP must still be described, few technical issues remain.

The availability of such functions in equipments will mostly be conditioned by the interest of ISPs. One point to keep in mind is that access equipments must remain relatively cheap and relatively simple. Thus there is a not inconsiderable risk that the required dynamic access control functions are not available before long.

5.2.2.3 Inter-domain multicast

The model to deploy PIM-SM in ASM mode over several domains has been developed with the assumption that each ISP would prefer managing his own RPs instead of relying on RPs managed by other ISPs and hosting RPs used by other ISP traffic. A new protocol, the Multicast Source Discovery Protocol or MSDP, has been introduced to exchange information on the active sources in each domain. Roughly, the MSDP sessions set up between the different RPs are used to announce the new registered sources, such that other RPs can subscribe to the SPT toward these sources. Periodically the full list of active sources is exchanged. In addition of being hardly scalable, MSDP introduces security issues: an attack initiated by a source towards a RP by sending malicious traffic is easily propagated to other RPs. Thus rate limiting parameters on the number of source announce messages must be finely set up. It must be noted that initially MSDP was proposed as a temporary solution, not dedicated to be largely deployed.

In addition to the potential security issues due to the introduction of MSDP, a large scale use of PIM-SM in ASM mode requires a solution to share the multicast group addresses. Indeed, no mechanisms or global policy has been defined to allocate multicast group addresses such as ensuring that two multicast sessions will not interfere. An experimental static address allocation policy, GLOP, has been proposed and some multicast addresses have been reserved per registered AS in the range assigned to GLOP (233/8), nevertheless only 256 addresses are available per AS.

Copyright © ENVISION Consortium, May 2011

The inter-domain deployment of PIM-SM in SSM mode is easier since no additional protocol is required. Moreover subscription to multicast flows takes into account the source address of the flows, which theoretically solves the issue of having to share the multicast addresses: each ISP can use the multicast addresses of his choice in the SSM range.

It must be noted nevertheless that the presence of L2 equipments at the access makes things a little bit more complicated since these equipments forward the traffic based on the Ethernet multicast group address only.

The inter-domain deployment of IP multicast requires the activation of additional protocols at the peering point:

- Adaptation of the MPGP policy to exchange also multicast routes (allow unicast and multicast routes to be incongruent)
- MSDP for the case of PIM-SM ASM to exchange “Source Active” messages
- PIM-SM to allow the construction of multicast trees across the peering points

Thus enabling IP multicast strongly impacts the peering point engineering: new functions must be tested, new engineering and security rules must be defined by the ISP.

From an operational point of view ISPs are always reluctant to introduce new protocols, all the more at the peering points, and in the case of multicast things are worsened by the fact that PIM-SM is considered as a complex protocol, and associated with security threats. For this reason, the generalisation of multicast peering is very improbable without strong benefit for the ISP in terms of cost.

5.2.2.4 Multicast address dynamic client allocation

An architectural framework and some protocols have been described at the IETF such as providing a way to dynamically allocate multicast addresses to applications.

[THA00] describes the Internet Multicast Address Allocation Architecture. The architecture is modular with three layers, comprising a host-server mechanism, an intra-domain server-server coordination mechanism, and an inter-domain mechanism. The layer 1 corresponds to a mechanism that a multicast client uses to request a multicast address from a multicast address allocation server (MAAS), MADCAP for instance, which is described bellow. The layer 2 corresponds to an intra-domain mechanism that MAASs use to coordinate allocations to ensure they do not allocate duplicate addresses. The layer 3 corresponds to an inter-domain mechanism that allocates multicast address ranges (with lifetimes) to prefix coordinators, for instance the experimental Multicast Address-Set Claim (MASC) protocol or static allocation per AS number.

The Multicast Address Dynamic Client Allocation Protocol (MADCAP), as been defined at the IETF to allow hosts to request multicast addresses from multicast address allocation servers [HAN99] (in a way quite similar as how terminals get unicast addresses with DHCP servers).

The clients can send the following messages:

- DISCOVER: used to discover MADCAP servers, sent using a reserved MADCAP server multicast address. The message may include the same options as the REQUEST message (see below)
- GETINFO: to request information such as the administrative multicast scope list
- REQUEST: the message may include:
    - The Lease Time (by default the maximum available for Lease Time), in units of seconds
    - The Minimum Lease Time (by default no minimum)
    - The Start Time (by default as soon as possible)

Copyright © ENVISION Consortium, May 2011

- The Maximum Start Time (by default no maximum)
- The Number of Addresses Requested (by default one)
- The List of Address Ranges options, describing the addresses it wants to receive (by default any address available)
- The administrative multicast scope
- The Current Time option which is used to detect and handle large clock skew between clients and servers and must be included if the Start Time or Maximum Start Time options are included

- RENEW: to renew a multicast address lease, changing the lease time or start time
- RELEASE: used to deallocate one or more multicast addresses before their lease expires

The servers use the following messages:

- OFFER: used to respond to DISCOVER messages that the server can satisfy
- ACK: if the server can process the client request successfully, with the options corresponding to the message to which it responds (lease time, multicast scope, list of address ranges, start time, current time...)
- NACK

A Lease Identifier is included in each MADCAP message to uniquely identify a lease. A transition is identified by a xid unique identifier. A Shared Lease Identifier can be used, for example for conferencing applications, in this case the server disable any authentication requirements and allow any client that knows the Lease Identifier to modify the lease.

[FIN00] defines An Abstract API for Multicast Address Allocation. The document describes the semantics of the interface that the dynamic multicast address allocation service presents to applications, including the guarantees made to applications. The API is very close to what is done with MADCAP. What can be noted is the impact on the application: “Multicast addresses are allocated for a limited lifetime. An application may attempt to extend this lifetime, but this operation may fail. Therefore, an application must be prepared for the possibility it will not be able to use the same addresses for as long as it desires. In particular, the application must be prepared to either quit early (because its original multicast address assignments have expired), or, alternatively, to occasionally 'renumber' its multicast addresses (in some application or higher-level-protocol dependent way), by making a new allocation. However, if an application needs to consider 'renumbering', it will always know this in advance, at the time it acquired its current address(es) – by checking the lifetime in the returned lease. An application will never need to be notified asynchronously of the need to 'renumber'.”

5.2.3 Assumptions on the IP multicast service

The objective of the ENVISION project, as far as multicast is concerned, is to define an interface between the ISP networks and overlay applications, such that an application which transmits content to a large number of receivers in an ISP network can benefit from IP multicast.

To increase the impact of the ENVISION solution, the proposed IP multicast service must rely on an architecture in the network which could realistically be deployed by an ISP, taking into account operational constraints. The objective is not to link the CINA interface to a specific way of deploying an IP multicast service in the Internet but to make sure that the CINA interface does nor rely on unrealistic assumptions and is also relevant for some IP multicast architectures more constrained than the theoretical open model.

Copyright © ENVISION Consortium, May 2011

The approach taken in the project is pragmatic: the objective is to study how IP multicast could be offered by an ISP, relying as much as possible on existing standard protocols, and without requiring important equipment upgrades or modifications in well-established network engineering practices. Thus the requirements for the IP multicast architecture are the following:

- The multicast service must not introduce security issues in the ISP network.
- The multicast service should not require important equipment upgrades:
    - Opening a multicast service in the Internet requires control of the access to the service. Several works describe how a dynamic access control could be set up. Nevertheless all these approaches assume an evolution of access nodes which is a strong assumption in the short or middle term since this equipment is deployed in large numbers, geographically distributed over the network, and, furthermore, the equipment is often legacy.
- The proposed solution should take into account the existence of legacy equipment in the network which will never implement multicast.
- The ISP must be able to supervise, control and optimise the resources used by the multicast service:
    - The number of multicast states available in the network is limited. The ISP must be able to control the multicast resources and to optimise their use to save bandwidth as far as possible.
    - The multicast relies on UDP, which does not integrate congestion control mechanisms. The ISP should have a mean to control that the UDP traffic does not impact services based on TCP.

From the above requirements:

- Due to the complexity and security vulnerabilities of PIM-SM in ASM mode, the CINA interface must not rely on the assumption that all the ISPs offer the possibility to use PIM-SM in ASM mode, the case where only the SSM mode is offered, which is very probable, must be considered.
- Multicast peering agreements are unlikely at the present time, the CINA interface must consider particularly the case where the multicast is limited to each ISP network.
- The case where an ISP does not have the possibility to use some dynamic access control functions on access nodes for multicast must be considered. In this context, either there is an agreement between the overlay and the ISP such that some nodes of the overlay, with fix IP addresses, can send multicast traffic, with a specified bit rate, towards specified multicast groups, or the ISP uses a multicast proxy in the network, configured with a set of reserved source and group addresses, and which can receive unicast traffic from any node of the overlay and distribute it in multicast. Note that this second scenario only allows to optimise the network multicast resources in real time in function of the application needs.

Consequently, two scenarios will be particularly considered in ENVISION:

- The case where an ISP is able to offer an open IP multicast service in the Internet based on PIM-SM in SSM mode and dynamic access control functions on access nodes.
- The case where an ISP must statically configure the multicast access control. ENVISION proposes the introduction of a function in the ISP network, called multicaster, which translates unicast streams to multicast traffic. The multicaster function is described in section 5.2.4.1. In addition to this function, ENVISION proposes a unicaster function, introduced to translate multicast traffic to unicast streams to allow access equipment that are not multicast capable to benefit from the multicast service. The unicaster function is described in section 5.2.4.2.

Copyright © ENVISION Consortium, May 2011

Different scenarios are possible regarding how the multicast service can be used by overlay applications. The ISP can sell the multicast service. In this context an agreement is concluded with the overlay application which can use a multicast state for traffic with a specified maximum bit rate, during a specified duration. Alternatively, the ISP can use the multicast service as a way to optimise the bandwidth consumption to the common benefit of the network and the application. This case where the network and the application are cooperating to jointly identify the content distribution topologies that would best benefit both the network and the application is called cooperative hybrid multicast. The functionality and the protocols required for cooperative hybrid multicast are briefly presented in section 5.2.5. Finally, in section 5.2.6 we study how multicast could be used in the ISP network transparently to the application, in the case the application cannot cooperate with the ISP (case of legacy applications for instance).

5.2.4 ENVISION IP multicast service enabling mechanisms

5.2.4.1 The multicaster proxy

The multicaster is a function in the network which transforms a content received in unicast to a multicast flow. The multicaster is configured with a set of IP addresses which are used as the sources of the multicast flow, and with a set of reserved multicast group addresses, see Figure 9.

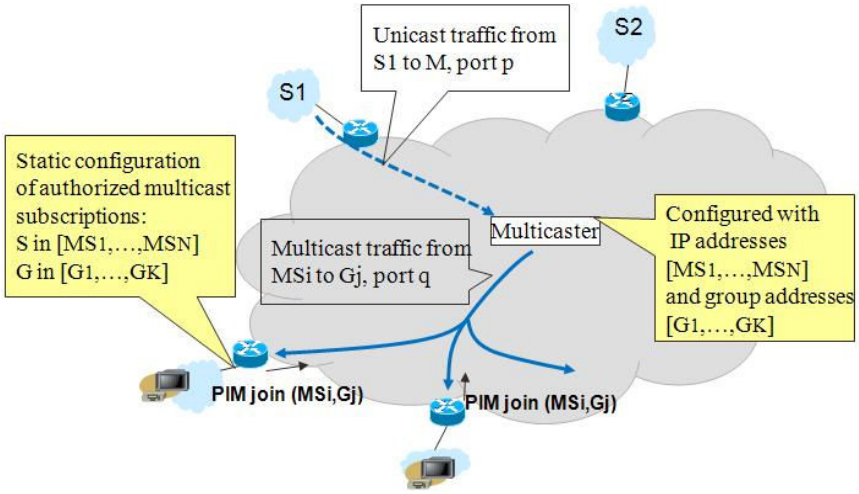

Figure 9: IP multicast configuration

The introduction of the multicaster solves several issues identified as blocking point for the deployment of IP multicast in the Internet:

- The approach does not require dynamic access control mechanism, the access nodes can be statically configured with the multicast source and group addresses dedicated to the multicaster, thus no evolution of the access nodes is required. Note that every end user has access to the multicast flows sent by the multicaster. If the privacy of data is required, the data security must be assured at the application level, through data encryption for instance.
- The approach does not require enabling multicast traffic in the upstream direction.
- The approach enables every source connected to the network to send traffic which is multicasted in the network, even if the source is behind legacy equipments which do not allow multicast.
- The control of multicast traffic is centralised, which simplifies its management.
- The introduction of a multicast proxy service enables the support of multi-source applications with SSM, as already shown in [HOL99] and [ZAP01].

Copyright © ENVISION Consortium, May 2011

On the other hand, the multicaster introduces extra latency in the distribution of multicast traffic, and as it represents a single point of failure, security aspects will have to be studied closely.

5.2.4.1.1 PIM-SM ASM versus PIM-SM SSM

The two modes of PIM-SM, ASM and SSM, could be used to build the tree rooted at the multicaster.

The use of the ASM mode could allow to build shared trees, each tree associated to a multicast group, in which several flows with different sources could be carried. Nevertheless the advantage of using shared tree when the multicaster is used is limited since the multicaster can use the same source for several flows, which can be differentiated by a port number for instance by the end user.

In this context, using ASM introduces unnecessary complexity:

- Every router must know the address of the multicaster (exactly as in the case of pure ASM where the RP address must be known by all the routers). In the case of distributed multicaster, the routers must know which multicaster is used for which group.
- The network must be configured such as assuring that no switching to the SPT tree occurs.

One point that must nevertheless be taken into account is the fact that some popular terminals do not implement IGMPv3 (Apple terminals for instance). Today this issue is solved by SSM-translation functions: the network is configured at the access such as to associate (statically) one source to one multicast group. This point must be considered when associating multicast sources and group addresses to the multicaster, typically the multicaster will always have to use the same (unique) source with a group.

5.2.4.1.2 Unicast content acquisition

Two approaches at least are possible regarding how the multicaster gets the content in unicast from the overlay application:

- The multicaster implements the (P2P) overlay protocol
- Other protocols, typically a L4 protocols are used

These two options correspond to different scenarios. In the first option the multicaster fits into the overlay content distribution protocol, and from the overlay standpoint less adaptation is required. In the second option the multicast service is more generic and can be more easily used by other applications types, and the overlay must be adapted to integrate the multicast service.

The pros and cons of the two options will have to be studied. In particular the efficiency of both approaches: simple packet relaying versus overhead of a transit through the application layer, will be evaluated.

5.2.4.1.3 Multicaster security

The multicaster security is a critical point especially as the multicaster represents a single point of failure in the multicast infrastructure and can be the object of attacks.

The incurred risks will depend on the defined architecture but will have to be studied closely.

5.2.4.2 Handling non multicast equipments

As specified in the list of requirements which should be fulfilled by the multicast architecture, the presence of legacy equipments which are not multicast capable, or on which ISPs does not want to activate multicast should be taken into account.

5.2.4.2.1 Automatic IP Multicast without Explicit Tunnels (AMT)

A draft is currently in-progress at the IETF [THAdraft10] with the objective of normalising a solution such that unicast tunnels can be automatically set up to carry multicast traffic across equipments or

Copyright © ENVISION Consortium, May 2011

network areas where multicast is not available. The solution relies on the introduction of AMT relays in the network and on the implementation of AMT Gateways on client side (as part of the media player or as separate software to which a player can connect).

After discovering a remote AMT Relay in the multicast-aware part of the network, the AMT Gateways encapsulate IGMP messages to subscribe to multicast streams into unicast UDP messages toward this relay. The relays send them the multicast traffic they have subscribed to encapsulated in unicast UDP.

Similarly a server wanting to stream multicast traffic but which would be located behind a non-multicast-aware network can use AMT, which will allow him to encapsulate its multicast traffic in unicast UDP toward an AMT Relay which will decapsulate and forward the multicast traffic as native multicast.

Beyond these basic principles, AMT includes a relay discovery mechanism, and a 3-way handshake for securing IGMP exchanges over UDP.

5.2.4.2.2 The unicaster proxy

AMT is an interesting solution when applications can be adapted to receive traffic in multicast. In the ENVISION project, we also consider the case where existing applications could not be adapted and the case where an ISP wants to optimise its network usage by using multicast inside its network only. In these contexts, it can be necessary to introduce in the network nodes able to transform multicast traffic in unicast.

How these nodes can be integrated in the overlay content distribution mechanism will have to be studied.

Note that the introduction of such nodes could also be interesting to enable the adaptation of multicast traffic to each receiver.

5.2.5 ENVISION IP multicast service control and optimisation functions

The use of the multicast service implies several phases, for which different role sharing can often be considered between the overlay application and the network:

- Multicast service discovery and exchange of information
    - This phase includes the discovery by the application of the support of multicast to all or a subset of the geographical locations covered by a particular ISP, and the corresponding limitations/conditions, e.g. the number of available multicast states. The application needs to retrieve this information from each edge ISP individually.
    - The ISP may need to retrieve information regarding the usage and traffic profiles of a particular application and per content item(s) that can be shared across many endpoints, e.g. the bitrate, the number of endpoints receiving a particular content item, if they support multicast protocols etc.
- Estimation of the benefit of multicast use for a particular content item
    - The estimation is done by the overlay, either based on the current demand for a content item, or based on the expected popularity of the content, and given the restrictions on the locations or the available state imposed by the network.
    - The estimation is done by the network, which implies that the network has the means to correlate high network utilisation at particular links with the distribution of a particular content item to many endpoints.
- Multicast source election

Copyright © ENVISION Consortium, May 2011

- Whether the source is chosen by the overlay or by the network, an important point is to choose a reliable source since a malicious source would impact a large number of receivers.
- To optimise the topology, the elected source needs to be close to the content source, or the ingress point at the domain in case the traffic is generated in a different domain.
- For peer-to-peer networks in particular, where end-user terminals may be used as sources and could therefore drop-out at any point in time, a number of sources may need to be elected, either to enable fast recovery in case the active source leaves the overlay, or for load balancing purposes.

- Multicast service use initiation
  - Using multicast for a particular content item distribution can be initiated by the overlay application. The following information should be provided to the network:
    - Identity of the source(s)
    - Estimated number of receivers
    - Content minimum and maximum bit rate
    - Start time and estimated duration
  - Using multicast for a particular content item distribution can be driven by the network:
    - The network informs the overlay that a multicast address is allocated for this content and that the application could use it for the particular content item
- Supervision of the multicast usage in the network
  - The network needs to have mechanisms in place to monitor how much traffic is injected by the application over multicast and possibly issue warnings or directly rate limit it
  - The application has to have mechanisms in place to monitor the quality experienced over the multicast connections e.g. loss, delay etc. and possibly issue notifications to the network or reduce the rate of the traffic transmitted over multicast and complement it with unicast traffic to increase the overall performance.
- Multicast resources release
  - The multicast resources may be provisioned only for a limited time, after which the overlay is responsible for reverting back to unicast traffic. The overlay may have the possibility to request the extension of the usage time.
  - Alternatively, the application or the network may initiate the release of multicast resources allocated to a certain content item, when certain conditions are met, for example:
    - The content consumption and corresponding drop in network utilisation or increase in application performance are not being satisfactory
    - There are other more demanding content items/applications which would benefit more from the use of the multicast service

As in MADCAP (see section 5.2.2.4), a multicast address is allocated to an application for a specific time interval, that, in most cases, it only covers the time periods when the application is most popular. The impact for the overlay application is that the possibility to use multicast for a specific content has a limited duration, possibly shorter than the time during which the content is distributed, thus the overlay must be able to cope with this situation and integrate handover mechanisms to switch form unicast to multicast and back. These protocols need to ensure that there

Copyright © ENVISION Consortium, May 2011

are no interruptions at the end points and minimise any impact on the user experience, permanent or even temporary, e.g. the significant increase of viewing delay for live video content.

5.2.6 Hybrid multicast scenarios

5.2.6.1 Hybrid multicast driven by the overlay

In ENVISION, we propose the use of hybrid multicast by the overlay to take advantage of IP multicast capabilities when possible and to relay the content via overlay unicast links over the parts of the network where no such capabilities are supported.

Figure 100 shows an example of a hybrid multicast topology, composed of three multicast domains. To establish a hybrid multicast topology, the first step is to group some nodes together into multicast-capable domains, then elect a local leader for each domain, named *multicast leader.* This later will act as a source of multicast into the domain. The multicast leader of each island is connected to the source via overlay unicast links.

Grouping of nodes together to multicast-capable domains follows the restrictions imposed by the ISPs. Without inter-domain multicast, such a domain is restricted to the number of ASs controlled by the same ISP. More fine-grained grouping of nodes to islands depends on their geographic location, the quality of service that they require and other parameters to be further investigated.

Multicast leader nodes should be equidistant from the island members and able to provide the QoS required by the multicast island members. Multicast leader nodes might be multicast enabled, or they may send unicast traffic to the multicaster node (see section 5.2.4.1). A fault tolerant mechanism should be considered to replace the multicast leader in case of failure or graceful leaving, in order to ensure service continuity.

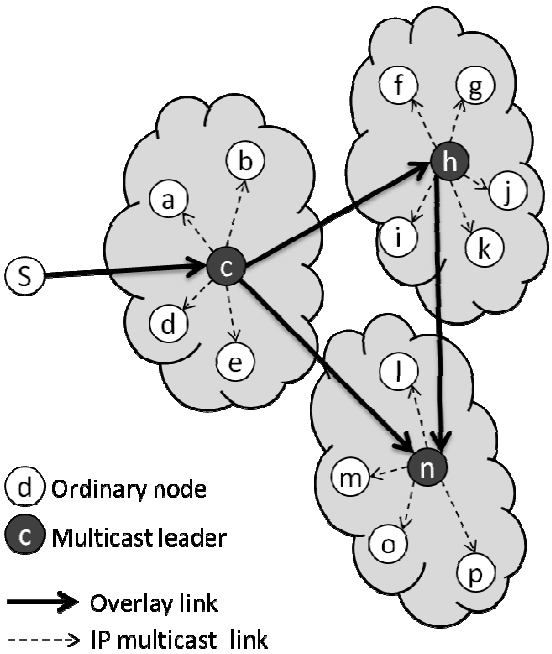

Figure 10: Hybrid multicast overview

5.2.6.2 Transparent hybrid multicast (Hybrid multicast driven by the ISP)

Hybrid multicast over the Internet relies on the assumption that the ISPs accept to open IP multicast in their networks. The multicaster function is aimed at mitigating the technical issues linked to IP multicast. Nevertheless other reasons can drive ISPs not to open IP multicast in their network, like legal impacts, or operational costs for instance.

Copyright © ENVISION Consortium, May 2011

Even if not available for end users, multicast can be used inside the ISP network to optimise the delivery of popular live content, and the multicaster and unicaster nodes are interesting building block to do this.

An ISP can decide to use multicast inside its network completely transparently from the overlay application standpoint, we call it “fully transparent multicast”. This approach is obviously rather complex to set up and will be studied with the idea to identify the difficult points, and how they could be solved if the network gets information or interacts with the targeted overlay application.

5.2.6.2.1 Fully transparent multicast

The objective is to investigate how IP multicast could be used in the ISP network without requiring the adaptation of existing overlay applications. The problem is complex and among other things, the following points will have to be studied:

- Detection/choice of the content to multicast. The ISP requires a mean to know which live content is, or will be, simultaneously watched by numerous users, and to estimate the bandwidth saving provided by the use of multicast for this content. The information can be gotten through deep packet inspection (DPI) or from the overlay if it provides some popularity ranking to end users. This information can be acquired on the basis of usage statistics, or at runtime.
- How the network nodes are integrated in the overlay content distribution mechanism:
  - How the multicaster acquire the content. Since no evolution of the application must be required, the multicaster must implement the overlay protocol.
  - How the end user requests arrive to the unicasters. Since the hypothesis is to not modify the application, the unicaster must implement the overlay protocol. If the overlay content acquisition process is pull-based, the end user request must be directed towards the unicaster. If the overlay content acquisition process is push-based the subscription to the overlay multicast topology must be intercepted.
- Choice of the source: since the traffic will be received by a large set of receivers the network must be able to choose a reliable source.

5.2.6.2.2 Transparent multicast in interaction with the overlay

The objective is to investigate how an interaction with the overlay could simplify or improve the delivery in multicast of popular live content inside the ISP network.

- The choice of the content to multicast can be done in interaction with the overlay.
- The overlay can push the content to the multicaster, the protocol used can be different from the protocol used by the overlay
- The overlay can direct the end users towards the unicasters, the protocol used can be different from the overlay protocol.

5.2.7 High capacity node

Cooperative overlay applications such as peer-to-peer networks leverage the fact that each participating node shares resources with other nodes. Through this cooperation peer-to-peer networks scale to a large number of nodes without the pre-provisioning of expensive server resources. This has led to their success in recent years, especially for non-time-critical applications such as file sharing.

However for the distribution of live or interactive high quality content these mechanisms are limited, as the aggregated upload capacity of the nodes that distribute a particular content determines the upper bandwidth limit that is available to the system. This limit also defines the content quality the

Copyright © ENVISION Consortium, May 2011

system can deliver, without risking jitter. Various strategies have been proposed to cope with this bandwidth limit. Joost [JunLeiXi] uses a pre-provisioned backend server farm that supports the peer-to-peer system by streaming parts of the content. Splitstream [Castro03] distributes the load across all peers of a tree by splitting the content and arranging multiple trees where leave nodes of one tree are intermediate nodes in other trees, optimising the available upload capacity compared to traditional solutions.

We propose a network service that can be offered by network operators to overlay to support them by boosting the overall bandwidth capacity available to the overlay network. Through the CINA interface the overlay will be able to request the instantiation of a high capacity overlay node in a certain area of the network. This node will act as a normal node towards other participating nodes in the overlay, but will support the overlay application remarkably through its high capacity connection.

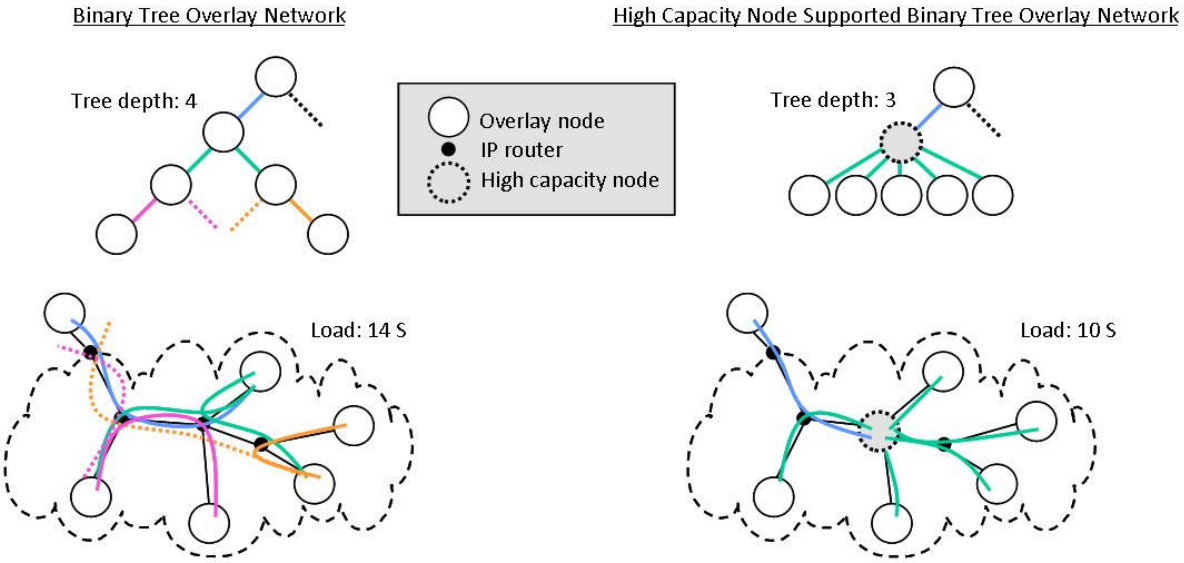

Figure 11: High capacity node overlay integration

Figure 11 illustrates an example, showing how such a node is integrated into a peer-to-peer network and how this network service benefits the network as well as the overlay. The figure depicts the overlay topology on the upper half and a network where the nodes are located with IP links and routers between them in the lower half. The figure shows a peer-to-peer network which maintains a binary tree topology to stream content from the source at the top to the leaves at the bottom. On the left hand side, without a high capacity node, the depth of the tree is 4, whereas on the right side it is only 3. The depth of a tree determines the end-to-end latency that is created by the system. Through the integration of the high capacity node the latency is thus reduced, assuming equal latency between each hop, by 25%. Here also none of the nodes of the local network is burdened by uploading content anymore. In addition there is a gain for the local network, which is shown in the lower half of the picture. Assuming that each IP link is loaded by S bytes per stream a total load of 14 S is created in the network through the system without high capacity node, not counting the potential dotted connection to peers outside of the local network. In contrast through the integration of the high capacity node on the right side the load in the network is reduced to 10 S, leading to an overall reduction of approximately 29% of the overall traffic in the network.

One precondition for such a service is that highly distributed computing resources within the network which will be needed to meet the requirements of distributed and globally scalable overlay applications. One promising approach to realise this is the use of routers that are able to host applications. The Alcatel-Lucent research packet processing platform provides routing and switching functionality with extensible, advanced packet processing. The core switching functions can be

Copyright © ENVISION Consortium, May 2011

complemented with extension cards to perform control and fast path packet processing. Flows are dynamically redirected to these cards through flow classification. The extension cards additionally support Linux with KVM based virtual machines. The platform thus is a natural host for the high capacity node network service.

To allow an intelligent instantiation of this high capacity node network service several research challenges have to be answered:

- How can overlay topologies be constructed according to the underlying network topology to maximise the gain for the nodes in a certain network region?
- How will an overlay detect that a certain area would benefit by the support of the network service?
- How can the network service instantiate nodes in certain distinct regions of the network?
- Where will the node be instantiated for an optimised overlay topology?
- How can the network service support arbitrary overlay algorithms?

The network service and the identified research challenges will be studied, additionally a prototypical implementation might be realised.

Copyright © ENVISION Consortium, May 2011

5.3 Network level adaptation

Content adaptation can be performed at different levels; in this section we will discuss this adaptation at the network level. First, sub-section 5.4.1 discusses the related work regarding Quality of Service management and video awareness mechanisms at the network level. Then, sub-section 5.4.2 proposes the enhancements that may be utilised in ENVISION.

5.3.1 State of the art

5.3.1.1 Video awareness at network level

This section introduces the existing techniques used at the network level in order to identify the traffic types and particularly to distinguish multimedia content and then to apply final adaptation. Toward this end, deep packet inspection (DPI) is needed to identify session request and service invocation at the network level. Network operator can decide later to redirect the request toward a streaming server or an adaptation gateway that are able to deliver the stream with the appropriate quality of service and according to metadata profiles. At the network level, DPI is the technology used to identify and to authenticate protocols and application conveyed by IP.

In the standard packets inspection process (shallow packet inspection), basic information (e.g. IP addresses (L3), Port numbers (L4)) were extracted from IP packets' headers and thus reveals the principal communication intent.

However, this technique dwells insufficient in order to identify any application-related information. For example, analysing the source and the destination addresses from packets header cannot inform us about the fact that the application is trying to set up additional connections.

In the other hand DPI provides information about the application by inspecting the content of the packets headers and the packets payload as well. Thus, DPI can enable the network operator to analyse any service invocation and to classify network traffic in order to optimise the service delivery and to enhance the network performance.

The DPI process can also rely on other aspect of protocol and application signature. The signature is very similar to fingerprint as it is used to uniquely and completely identify a protocol or an application (multimedia session, user agent used …).

As a new protocol/application is encountered in the network, a signature will be created and will be stored in a signature database. This signature has to be regularly checked and maintained as several application/protocols change their behaviour with new updates. It is known that Skype and BitTorrent protocol signatures change frequently in order to escape from being captured by some network providers.

The signature database maintenance is very crucial for an efficient DPI. Indeed, classification problems may arise if the signature does not uniquely identify a protocol/application. For example, an application may be identified as something it is not (false positive) or not identified as it should be (false negative). A common situation where false negatives occur is where the application behaves differently depending on the client situation (with/without the use of proxy/firewall).

The signature creation process is based on application behaviour analysis; here we introduce the different techniques used to classify an application/protocol in order to create a signature.

1) ***Analysis by port number:*** this may be the easiest and the well known technique for signature analysis. This technique is based in the fact a many applications use default ports or ports that are chosen using a specific pattern. (Example applications: POP3 110/995, SMTP 25, FTP 21 …). However, port analysis is very weak since some application use random ports or ports that allow them to be identified as other application (for example the port 80 syndrome, where the

Copyright © ENVISION Consortium, May 2011

application traffic may be seen as HTTP traffic). For these reasons, port analysis cannot be used alone and should be used along other analysis technique.

2) ***Analysis by string match***: this technique tries to search for one or several specific character sequences within the packet. This technique relies on the fact that most applications announce their names in the transactional messages. For example, the field `User-Agent` in a typical `HTTP GET` request.

L7-Filtering: Layer-7 filter is a software package for traffic classification in Linux. The main motivation behind its development is the identification of P2P traffic. There are two versions of the L7-Filter package. The first one is implemented as a kernel module, while the second one (which is still experimental phase) runs as a user-space program. L7-Filter relies on the use of regular expression to identify the application/protocol and thus fall in the category of signature analysis by string match as mentioned earlier. Figure 12 presents some L7-Filter patterns related to instant messaging protocols/applications and P2P applications.

```
skypetoskype
^..\x02.............

jabber
<stream:stream[\x09-\x0d ][ -~]*[\x09-\x0d ]xmlns=['"]jabber

msnmessenger
ver [0-9]+ msnp[1-9][0-9]? [\x09-\x0d -~]*cvr0\x0d\x0a$|usr 1
[!-~]+[0-9. ]+\x0d\x0a$|ans 1 [!-~]+ [0-9. ]+\x0d\x0a$

msn-filetransfer
^(ver [ -~]*msnftp\x0d\x0aver msnftp\x0d\x0ausr|method msnmsgr:)

bittorrent
^(\x13bittorrent protocol|azver\x01$|get /scrape\?info_hash=get
/announce\?info_hash=|get /client/bitcomet/|GET/data\?fid=)|
d1:ad2:id20:|\x08'7P\)[RP]
```

Figure 12: Some L7-Filter Patterns.

In the case of the `BitTorrent` filter, the following packets will be marked as BitTorrent protocol packets:

[header]	\rbittorrent protocol	...
[header]	get /scrape?info_hash=	...
[header]	get /announce?info_hash=	...
[header]	get /client/bitcomet/	...
[header]	GET /data?fid=	...

3) ***Analysis by numerical properties***: this technique tries to make use of the application/protocol numerical characteristics by analysing one or several packets. Examples include payload length, fields' offsets within packets, and number of packets as a response to a specific transaction.

4) ***Analysis by behaviour and heuristics***: the behaviour of an application/protocol is the way in which that application/protocol acts and operates in terms of transactions and/or sent/received packets. Heuristic analysis is expressed by the mean of statistical parameters of examined

Copyright © ENVISION Consortium, May 2011

properties. An example of behavioural pattern that can be traced is an action leading to another action (a UDP connection transforms into a TCP connection using the same port/address).

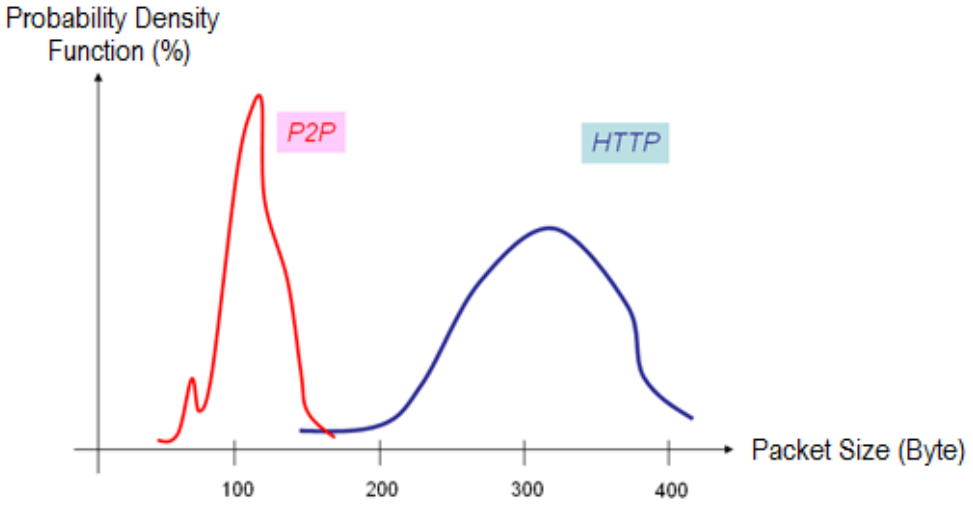

Figure 13: Packet size distribution in case of HTTP and P2P file sharing applications [Allot Communications 2007]

An example of heuristic analysis is the file sharing application (see Figure 13: Packet size distribution in case of HTTP and P2P file sharing applications [Allot Communications 2007]). In a HTTP file sharing application, packets are relatively big in term of packet size (packet size is dense around 300 bytes) while in a P2P file sharing application packets may be smaller (packet size is dense around 100 bytes). This statistical information may be used in order to know whether a port 80 connection is used to carry pure HTTP traffic or P2P traffic.

While DPI may refer to several packet inspection techniques, we can go further with the classification and we can have the following sub classes of DPI:

1) ***Cross Packet Inspection (XPI):*** Rather than analysing the data packets separately, XPI technique (also known as Deep-DPI) examines a set of data packets. For example, an HTTP GET request may be split into three IP packets.

2) ***Cross Session Inspection (XSI):*** Cross Session Inspection technology (XSI), based on XPI technology, tries to reassemble a complete session by performing an XPI. For example, to get a full overview of a streaming session, many control and data sessions (e.g. RTSP + RTP/RTCP) have to be reassembled.

3) ***Deep Session Inspection (DSI):*** Essentially used for the calculation of KPI (Key Performance Indicator) and metrics, Deep Session Inspection (DSI) parses the sessions' payload looking for a relationship between the different sessions. For example, in order to compute the average access time to a web page including its pictures, DSI can be used to analyse the different sessions (HTML/XML session and Pictures' sessions) and compute the average access time.

5.3.1.2 DPI Architectures

DPI architectures have been standardised over the past few years by several important international telecommunications standards organisations (3GPP, IETF, ETSI TISPAN, etc.). These architectures mainly consist of two elements: Policy Decision Point (PDP) and Policy Enforcement Point (PEP). The PDP is an intelligent and computation-intensive device. The purpose of PDP is to make policy decision on behalf less intelligent devices. The PEP acquires deep packet inspection (DPI) policy from the PDP and processes data stream from a terminal in accordance with the DPI policy. PEP is essentially any

Copyright © ENVISION Consortium, May 2011

piece of subscriber equipment that is capable of enforcing a policy decision. Some vendors also provide DPI solutions with integrating these two elements (PDP and PEP) in a single module.

5.3.1.3 DPI-based content adaptation

The content adaptation at the network level can be achieved by introducing content adaptation gateways. Thus, after inspecting IP packets, adaptation service may be invoked in order to adapt the requested/transmitted content to meet the network conditions and/or the user context in a seamless fashion.

Figure 14 shows a typical situation where two users (User A and User B), with different profiles, are requesting the same multimedia content. In this example, the user A is supposed to have a profile which enables the direct consumption of the original content (no specific requirement for content adaptation), while the user B is supposed to have a profile where an adaptation of the content must be performed before its consumption. In this scenario, a dedicated network node performing deep packet inspection (DPI) on the metadata flows, may decide whether to re-route the data flow through an adaptation gateway (e.g. serving the user B) or not (e.g. serving the user A).

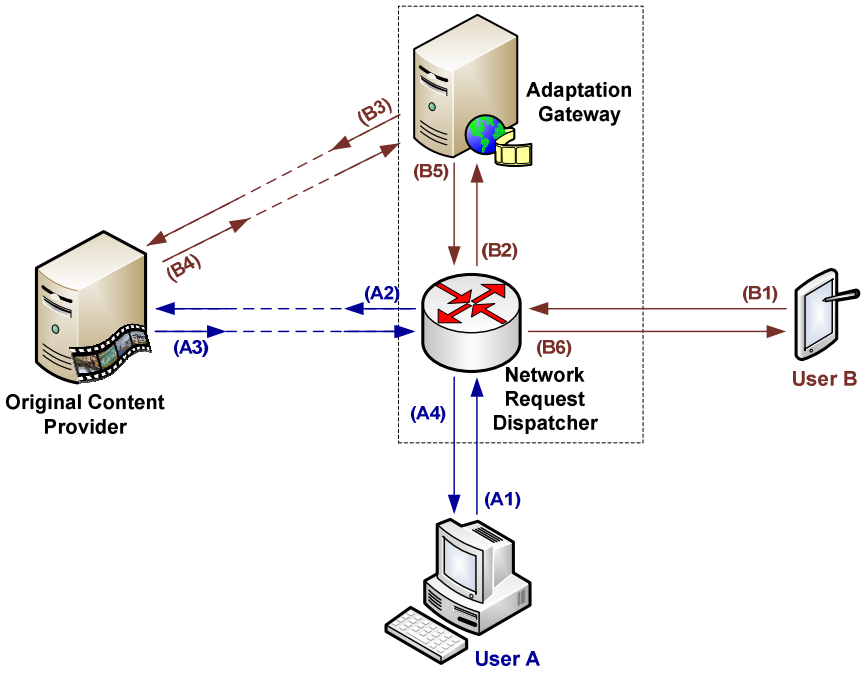

Figure 14: Content adaptation inside the network

Content adaptation may include the techniques described in the following subsections.

5.3.1.4 CODEC adaptation

The adaptation/content encoding service will allow the transcoding from a certain format to another one. For example, two end users (User A and User B in Figure 14) requesting SVC content with different terminal capabilities (different codec requirements). User A has the capability to support H264/SVC format, and thus it can consume the content directly. On the other hand, user B can only support MPEG2 format. After analysing metadata flows, the Network Request Dispatcher can re-route the SVC content through the Adaptation Gateway in order to be transcoded to MPEG2 which is supported by the User B terminal, while this content is routed directly to the User A. Figure 15 shows the typical example of codec adaptation.

Copyright © ENVISION Consortium, May 2011

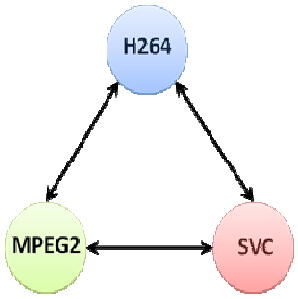

Figure 15: Codec adaptation

5.3.1.5 Bitrate adaptation

The bitrate adaptation in video streaming context is the modification of the stream bitrate in order to meet the network bandwidth requirements or the receiver device capability, while keeping a QoS/QoE above a certain threshold. This adaptation can be performed by reducing or enhancing the fine-grained video quality (Figure 16). It is also known as the SNR (Signal to Noise Ratio) adaptation. The bitrate adaptation can also be performed by reducing/increasing the spatial resolution of the video. This is what we call spatial adaptation (Figure 17). Finally, the bitrate adaptation can be performed by deleting some pictures from the video, and consequently reducing the number of frames played per time unit (Figure 18). This is called temporal adaptation.

Figure 16: Quality adaptation

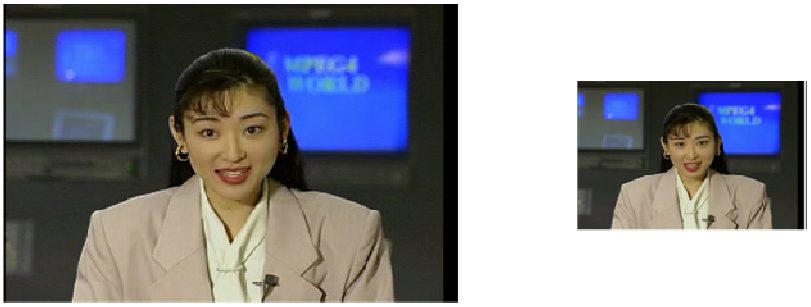

Figure 17: Spatial adaptation

Copyright © ENVISION Consortium, May 2011

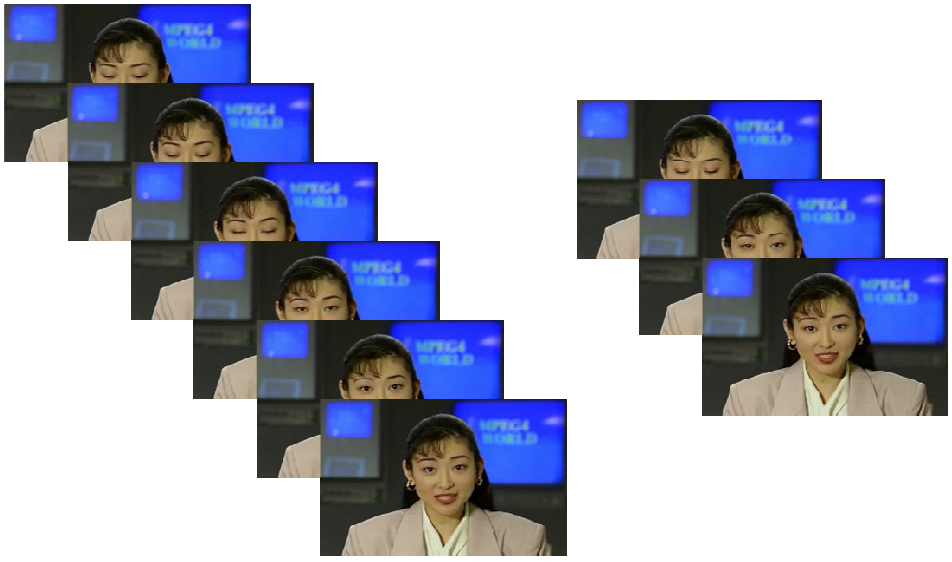

Figure 18: Temporal adaptation

For example, a user with a full capable terminal (User A) and a user with a limited capacity terminal (User B) are requesting the same multimedia content from the Original Content Provider (for example HD video with 1280x720 image resolution). After analysing the metadata flow (using DPI), the Network Request Dispatcher will decide whether the content can be directly consumed by the user or not. In the case where the user cannot consume the content in its original form (a mobile device cannot take full advantage of a HD video as its screen resolution does not allow him to view a 1280x720 resolution), the network operator may invoke a content adaptation service in order to provide the same quality of the requested video but with the appropriate spatial resolution.

5.3.1.6 Protocol adaptation

Different content is transmitted using different protocols. The adaptation service will include a protocol adaptation (see Figure 19). For example, an end user behind a firewall would not be able to receive a content using RTSP/RTP as the necessary ports may be blocked by the firewall or NAT limitations. Thus, the adaptation gateway needs to be invoked to deliver the content using HTTP rather than RTP in order to serve this user which is behind a firewall/NAT.

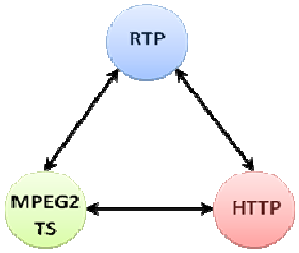

Figure 19: Protocol adaptation

5.3.2 Smart packet dropping

Real time video adaptation at the network level may be in the form of dropping packets at the network to react on a given congestion, smart dropping could be achieved in several ways, for instance packets priority could be higher due to several aspects, their payload or content importance to QoE of the users, time to display (maybe correction could be applied by means of retransmission), level of implied error correction (also, dropping packets from different FEC blocks), and the number

Copyright © ENVISION Consortium, May 2011

of users receiving that stream. SVC is the simplest example where higher layers that are less significant can be dropped, in addition, in a given layer there are different level of importance, like B frames and P frames are less important than I frames. Moreover it is also assumed that dropping of frames is more suitable than dropping packets, thus the network may need to associate different packets belonging to a single frame.

Packet dropping though should be the last option, it will be better to first adopt the content prior to drop packets at the network, thus, the option to either communicate between the network and overlay may results in QoE gain over dropping packets, smart dropping though aims to minimise the negative affect of such severe action.

5.3.3 FEC service at network level

In many cases, the network may be able to reproduce parity packets; this could be at caching serving nodes for instance where multi-level caching is done at different layers of the network, rather than on the peers themselves, or at the routers connected to the peers. The main idea is to be able to take active role of protecting the packets in links where it is needed based on network level monitoring. The FEC at the network level, reduces the need to implement it any overlay application, and to control the overhead of different services dynamically at the network. Although we present the FEC at the network, we still have doubts regarding its importance, and further study will clarify it.

5.3.4 SVC in P2P swarms

The weaknesses of many P2P streaming systems come from static selection of streaming parameters that are based on average peer resources. This selection might work if all systems in the network would have equal resources, which is not true due to the heterogeneity of the Internet. Internet devices are heterogeneous not only in their resources, but also in the type of connections they have. Therefore, bandwidth, delay and reliability vary drastically, rendering current P2P video streaming techniques best effort, i.e. they either work or not. A possible solution to the problem of supporting streams with different qualities is achieved by creating a different video file for each quality level and therefore different overlays or swarms. However, this solution is not only inefficient due to data duplication across overlays, but also limited with respect to the level of possible collaboration between strong and weak peers across different overlays. To overcome these problems SVC based P2P streaming is introduced.

The scalable video coding (H.264-SVC) is new extension format for H.264 standard that is considered as a promising video format for media streaming over heterogeneous networks. In SVC encoding scheme, each video stream is encoded in multiple quality layers. The first layer which provides the basic quality of the video is called the "base layer", while others layers which are used to enhance the video quality of the base layer, are called "enhancement layers". The different dimensions of scalability offered by SVC are as follows.

- Temporal scalability is based on providing different frame rates for a video stream. This is achieved through structuring picture and motion estimation dependencies such that complete pictures can be dropped from the bitstream while still providing the possibility of decoding the video stream.
- Spatial scalability is based on providing different resolutions for a video stream. This is achieved through the usage of lower resolution pictures to predict data of higher resolutions pictures.
- SNR scalability is based on providing different quality levels for a video stream. This is achieved through hierarchical construction of quantisation coefficients for each picture.

For streaming, an SVC stream is divided into chunks. Each chunk contains layers in the three dimensional quality spaces. The smallest quality unit is called a NAL Data Unit (NAL: Network

Copyright © ENVISION Consortium, May 2011

Abstraction Layer). A NAL unit will be used as basic unit for fetching and distributing video data across the network.

The following approaches can be used in order to perform layered content adaptation (e.g. SVC) at the network level:

5.3.4.1 Adaptation using DiffServ

The transmission of SVC content over the Internet faces many challenges. Different layers of the SVC stream contribute differently to the stream quality. Consequently these layers can be transmitted with different priorities. In DiffServ architecture, the packets of different layers can be mapped to different classes of service. For example, packets of the base layer are marked to the higher class of service compared to enhancement layers. For example, the base layer is mapped to the Expedited Forwarding (EF) class of service while the enhancement layers are mapped to the different (AF) classes according to their priority.

5.3.4.2 Adaptation using packet dropping

Another scenario of SVC content adaptation at the network level is the layer dropping. Due to limited bandwidth or network congestion, a network element may drop packets. After analysing SVC data flows using DPI, and in case of network congestion or limited bandwidth, a network element may decide to suppress packets that belong to higher layers (i.e. less important layers) in order to ensure the service continuity. However, the packets of the base layer must not be dropped as they are required to decode any other layers.

5.3.4.3 Simultaneous multicast delivery of SVC layers

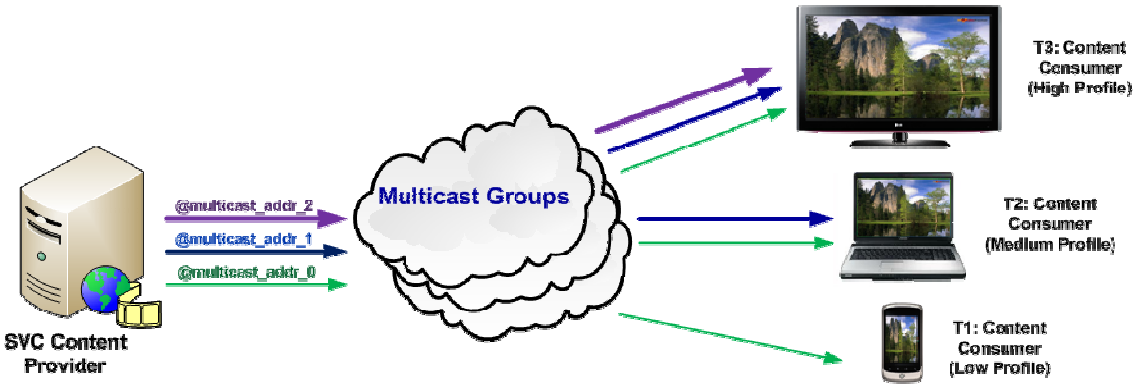

Figure 20: Simultaneous SVC layers multicasting

The distribution of SVC stream in P2P streaming environments is done through several possible ways. A smooth approach to distribute SVC content is to send each layer in a separate multicast session. Indeed, each SVC layer is transported in its own IP multicast group identified by its own IP multicast address, and terminals subscribe to layers utilising IP multicast mechanisms, namely IGMP. This implies that a terminal should subscribe to several layers (several IP multicast groups) to have better quality.

Figure 20 illustrates an example of this approach. The SVC stream considered in this example is composed of three layers: The base layer and two enhancement layers: enhancement layer 1 and enhancement layer 2. Three types of terminal are connected to a SVC content provider, through Internet, over links with different bitrates capacities. The terminal T1 with low bitrate capacity link subscribes to the multicast session distributing the base layer. Consequently, it receives a low quality related to its capacity. T2, with higher bitrate capacity link than T1, subscribes, in addition to the multicast session providing the base layer, to the multicast session which distributes the enhancement layer 1. The received content quality in this case is medium. Finally the terminal T3

Copyright © ENVISION Consortium, May 2011

joins the three multicast groups to obtain the three layers in which the SVC stream is coded, in order to obtain the best offered quality.

5.4 Caching

Having in mind network optimisation, caching content looks a natural option for saving bandwidth in the network, and delivering content more rapidly to end-users.

Within the project and according to the defined use-case, caching could be seen at two levels: long-term caching content, e.g., for 3D Web conference, for caching background, materials and others fixed things that can be shared and short-term caching for Live content for instance when content is stored for only few seconds.

For both use-cases, efficient caching solutions should provide several properties, such as introduced in the next sections. Amongst others, the location of the caches in the network and the hit ratio (for efficiency of the caches) are the most important ones that we have a look at.

In this section, after a presentation of the functions a cache system should have, we present the existing network caching solutions, both transparent and explicit, with some results of cache efficiency. Then an analysis of current models for estimating cache performance we have conducted is presented. Finally, we present current thoughts of caches utilisation for use-cases we have in ENVISION such as the Live content.

5.4.1 State of the art

During these last years, OTT (Over The Top) traffic and more precisely video traffic grew in a significant manner. The consequence of the proliferation of Internet-connected devices, 10 Mbps access and more, the increase of media consumption per viewer, the availability of high quality and long-form content, involves expensive network costs for NSP (network service providers) with weak incomes for this type of traffic inside its own network as well as on peering links. To make nothing would bring the risk of seeing customers churning due to the decreasing of QoE.

Before this period, ISPs were faced to P2P traffic with the same issues in term of non revenue traffic and high bandwidth occupation. To limit the impact of these traffics, network service providers can use different ways like IP QoS DiffServ (NSP traffic in premium, OTT in best effort), DPI (OTT traffic constraint by shaping or limiting) and caching solutions. The last one is an improvement of the first Web caches used to accelerate Web pages viewing on end-user screen, getting better QoE. Now, caching solutions are deployed to store and deliver long-form content to end-users instead of origin servers, saving bandwidth on peering links and network and providing faster download times to end-users i.e. QoE.

Copyright © ENVISION Consortium, May 2011

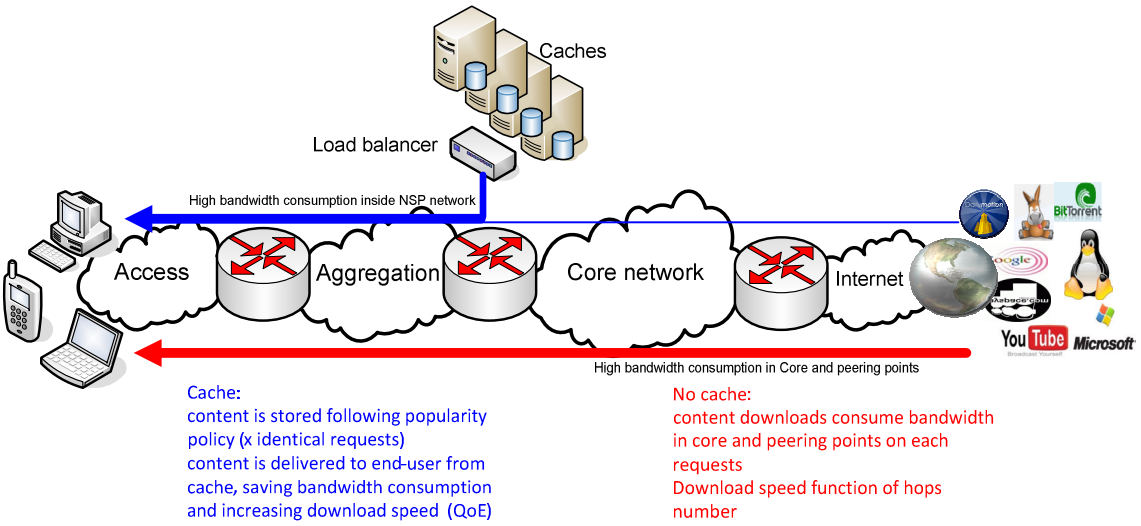

Figure 21: Caching solution located in NSP

Multiple solutions are proposed to NSP based on different models and principles.

5.4.1.1 Cache mechanism: main functions

Cache mechanism needs to provide multiple functions to manage content. Main functions are the ingestion, the storage, the delivery and the managing.

To ingest content, a cache must be aware of end-user requests sent to a content provider on Internet cloud. So, it is necessary to divert the end-users requests to the cache.

1) Divert function

The divert function, which is external to the cache, can be supported by different mechanisms:

- Pure network mechanism present in routers: PBR (policy based routing), WCCP (Web Cache Communication Protocol), BGP (Border Gateway Protocol)
- Additional network device: DPI (deep packet inspection) able to recognise L7 traffic and divert it to the cache
- DNS: NSPs need to configure its own DNS to divert URL to the cache

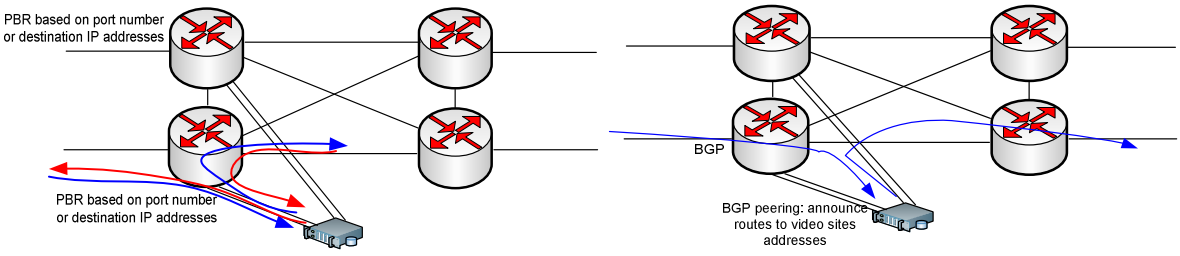

Figure 22: Divert function based on PBR and BGP

With PBR, routers are able to divert from one or a set of interfaces to another one interface the traffic following the port number or destination IP addresses. For example, all packets with TCP port number equal to 80 (main http traffic) and greater than 1024 (P2P traffic) will be diverted from interface A to interface B with an ACL (access control list) like:

```
"ip access-list extended Internet.2.Local
 permit tcp any eq www any
```

Copyright © ENVISION Consortium, May 2011

```
 permit tcp any gt 1024 any gt 1024
…
…
ip access-list extended Local.2.Internet
 permit tcp any any eq www
 permit tcp any gt 1024 any gt 1024”
```

With BGP, one of the mechanisms is based on the cache capacity to announce for the proximity router the route to the content provider site. The end-user requests are diverted to the cache then analysed. To avoid BGP loops (in case of redirection), added engineering is necessary.

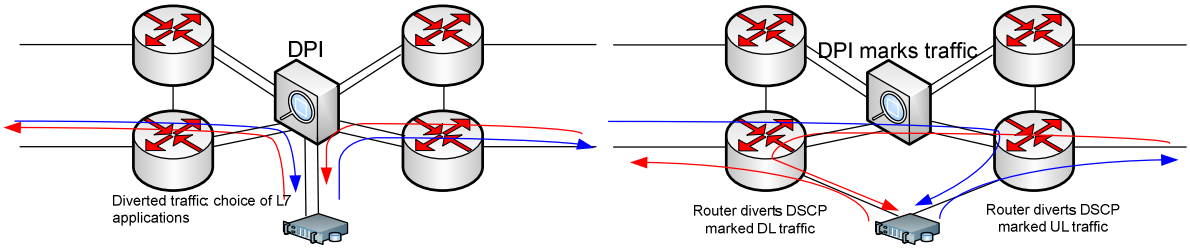

Figure 23: Divert function based on DPI with and without integrated divert functions

DPIs are devices cutting the links and capable to identify layer 7 applications then mark the different traffic, apply rate limiting, quotas and shaping policies. These equipments are currently standalone boxes with or without divert functions but the new approach is to integrate these equipments inside network devices like routers.

DPIs permit to divert traffic targeting specific applications or P2P protocols instead of large amount of traffic which is not interesting to cache. When DPI is not equipped with divert function, it is possible to use DPI to mark the targeted traffic (DSCP field) which will be treated by classical network equipments (routers and ACL based on DCSP field values)

2) Ingest function

Once requests are diverted to the cache, ingest function must analyse these requests:

- The request is not for targeted traffic => no analyse, request is sent to origin server
- The request is in relation with targeted traffic but not with content download (e.g. signalisation between end-users and origin servers for statistics or BW monitoring) => request is sent to origin server
- The request is in relation with targeted traffic => analyse the “GET” content:
  - Content is not yet stored in the cache:
    - The number of similar request is equal of configurable popularity threshold => the “GET” must be treated and content must be stored in the cache, the end-user request is sent to origin server
    - If not, request counter is incremented and end-user request is sent to origin server
    - Ingestion is made in parallel with end-user download (synchronous mode) or independently (asynchronous mode)
  - Content format coding: ingestion can be for
    - Entire content: the content is stored as a unique object. To get the content, it is necessary to avoid a cut request in synchronous mode

Copyright © ENVISION Consortium, May 2011

- Content divided in pieces: the content is distributed in multiple pieces of same size. The mechanism could be like P2P chunking or based on http byte-range. In that case, content can be stored partially.

3) Store and manage functions

The cache needs to store content inside specific space. Possibilities are:

- DAS: internal hard disk
- SAN: Storage area network (linked by SCSI, Fibre channel, iSCSI, etc.)
- RAM: internal ram memory

Depending on cache solution, the store space can be one or all these possibilities. For example, a cache will store high popular content inside the RAM, popular content inside DAS and long tail inside SAN.

If the cache is based on hierarchy architecture (multiple level of cache), storing can also based on this hierarchy.

Content management needs to take in account the storage. In this document, manual deletion or manual ingestion is not studied. The main management functions are deletion and positioning. Cache solutions must support these two functions without any constraint on human management, so caching must integrate automatic functions. Deletion can be provided by algorithm like LRU or LFU systems which delete content when storage reaches a high percentage of use and content positioning could be manage between RAM, DAS, NAS following the request popularity or a content deployment policy.

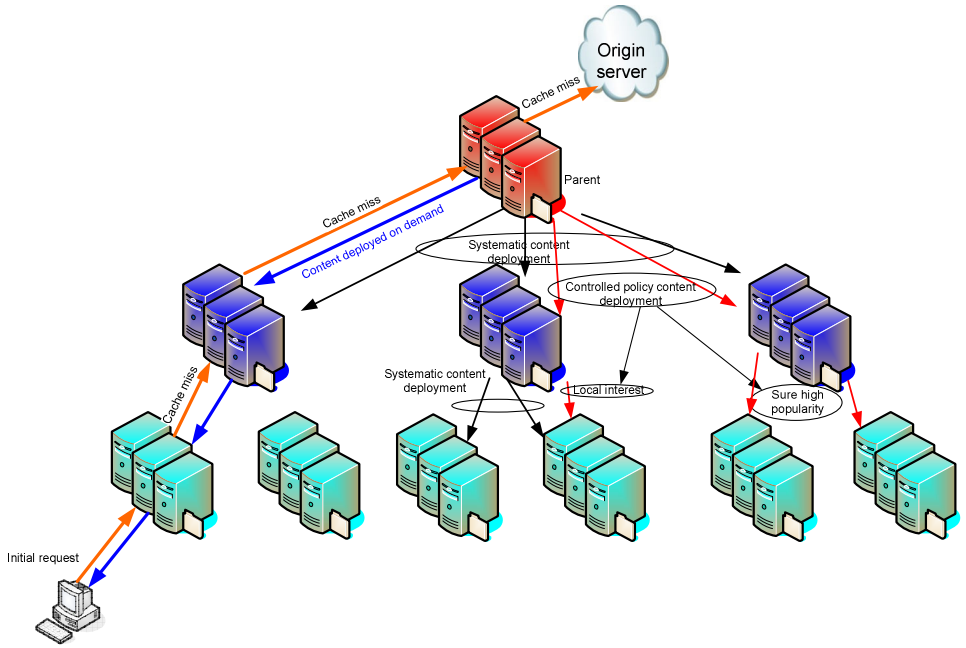

Figure 24: Hierarchic cache solution

A hierarchical cache can provide a solution based on content popularity, permitting to manage storage size and location in a network. The lowest level of caches is dedicated to high popularity content, either local or global interest.

4) Delivery function

Content is stored in the cache:

Copyright © ENVISION Consortium, May 2011

- The “GET” request is treated by the cache depending cache systems:
  - HTTP redirect sent to end-user to redirect request to delivery cache function
  - Cache stops the delivery from origin server and delivers content keeping with origin server one “keep-alive” connection
- Content in cache:
  - Complete file: delivery without cutting
  - Partial file: the content is delivered one part by the cache, other part by the origin server => cache must be able to switch the delivery from cache to origin server

5) Transparency or otherwise

Depending on cache method, 2 approaches can be designed: transparent and not transparent caching solutions. Both are in common to the respect of end-users and origin servers with no integration of additional software in their equipments.

Transparent caching can be defined as end-users and origin servers are not aware of cache presence. This constraint needs for the cache to be able to spoof the IP addresses of each side: cache spoofs end-user IP address for origin server and origin server IP address for end-user.

Non transparent caching solution introduces that end-users and origin servers are aware of cache presence, cache IP address can be seen by both parts. IP spoofing is not used in that case.

6) Cache efficiency

The efficiency of a cache is described as Hit Ratio and Byte Hit Ratio.

- Hit Ratio: Requests made by end-users are analysed by cache and some of them reach its content in the cache. Hit Ratio is the percentage of requests that hits in the cache. Hit ratio is the consequence of requests popularity. The most popular files are requested more than one time so the probability to get these files in the cache is strong.

  Content popularity is an external element of cache system which cannot be controlled by the cache system itself.

  What are the possibilities to improve the Hit ratio? Cache system can improve it with different solutions but mainly with targeting the long tail of less-popular content. To make this, the cache storage size needs to be increased to permit to store less-popular content, first content deleted by automatic LRU algorithm which manages storage. This LRU algorithm can also be configurable to increase the hit for a maximum number of content. The popularity policy in charge of requests analyse to determine the most popular of them (so caching the most popular content) can also be modified to take in account a low number of similar requests to store the content (e.g. store content at each different requests). Depending of cache ingestion mechanism, another way to improve the hit ratio is to store the content (partial or complete) even if end-user stops its request before the end of downloading.
- Byte Hit Ratio (BHR): Byte Hit Ratio is the percentage of volume delivered by the cache. Problematic is about the question: how to calculate this percentage?

  Byte Hit Ratio is often given by different definitions representing different point of view (and one of them is often “marketing” point of view showing very high scores).

  A realistic approach of BHR is the number of bytes delivered by the cache (bytes served) against total number of bytes requested.

Copyright © ENVISION Consortium, May 2011

The number of bytes delivered by the cache against total number of bytes for requests delivered by cache often presented by manufacturers and closed to 100% can be named as “marketing BHR”.

Byte Hit Ratio is based on cached files size. That involves possible BHR improvement with caching large size objects and, like HR, storing and delivering partial content.

7) Impact on the network

Cache integration in network involves some constraints based on cache architecture and model.

Full transparent cache (e.g. PeerApp, Ankeena) needs to get both traffic directions (upstream and downstream of a same session). This point implies to have a bi-directional traffic on the links where the traffic should be cached. If it is not the case, network engineering must be modified to obtain this bi-directionality. To solve this issue, caches could be located in the access network where traffic is in majority bi-directional. But, the caches are working on statistics rules based on request popularity, so depending on high requests number, so on high end-users number. On the contrary, location at the high level in network permits to get high number of requests but the probability to get unidirectional traffic on links is important.

These issues involve the Caching dilemma for NSP: how to get the best hit ratio with the best bandwidth saving and the lowest investment?

- Caching at the core level involves a high hit ratio, the interconnection bandwidth saving and low number of cache equipment. However core, aggregation and access networks are not impacted by the bandwidth saving and cache equipment is necessarily expensive having to manage high capacity line rates (10 Gbps or more).
- Caching at the access level involves a high bandwidth saving (interconnection and core level), on less-sophisticated equipment (e.g. with link capacity equal to or less than 1 Gbps). However the hit ratio is a function of the number of end-users and might be low unless the quantity of equipment is large.

5.4.1.2 Solutions including several type of traffic

In OTT traffic, NSPs need to target a cacheable traffic as large as possible:

- P2P (e.g. Bittorrent, eDonkey, encrypted, unencrypted)
- HTTP Progressive download (e.g. Youtube, Dailymotion videos)
- HTTP Downloading (e.g. file sharing as Megaupload, RapidShare; software update, all type of files downloaded by http transfer)
- HTTP Adaptive streaming (e.g. Smooth Streaming, Apple Adaptive Bitrate streaming, Adobe HTTP Dynamic streaming)

Currently, no market solution provides a cache with all these features.

In addition, as P2P and associated cache solutions become a major type of traffic, some specific features of this type of traffic can complicate the caching solutions. In fact, the cache needs to support a large number of P2P application protocols, every P2P protocol is continuously evolving with new features and P2P applications use proprietary protocols or support end to end encryption. The objective of the IETF working group DECADE (DECoupled Application Data Enroute) is to design an in-network storage protocol to address the problems listed before.

DECADE approach is to access to in-network storage through standardised interfaces. This open standard protocol will be a solution to decouple P2P data transport from P2P application control and signalling. The Main idea is to specify a protocol to use between caches themselves or between caches and peers.

Copyright © ENVISION Consortium, May 2011

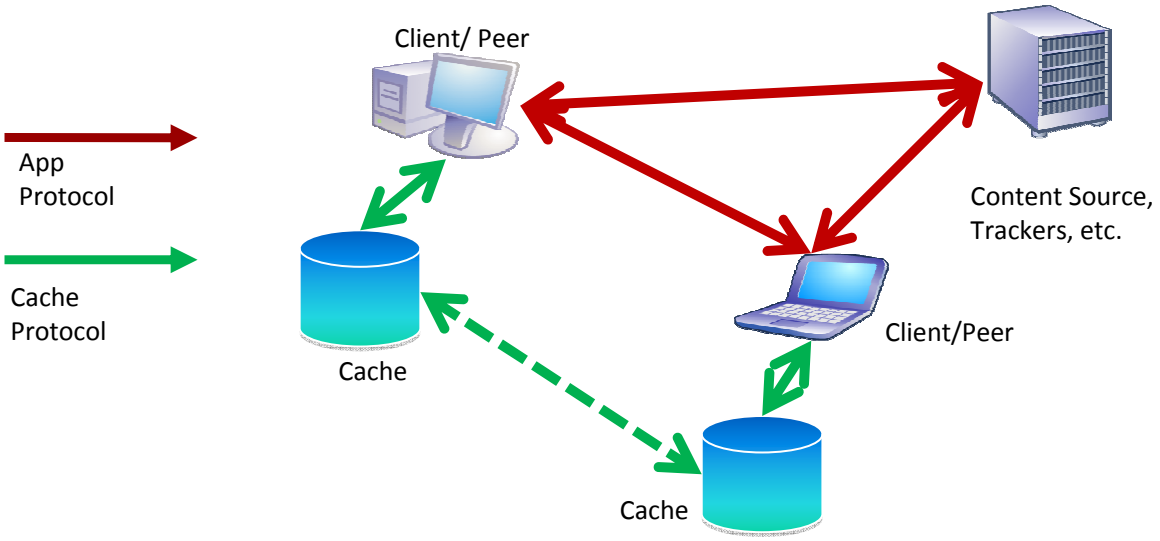

Figure 25: IETF DECADE: in-network storage

As we can see, with this approach, there is no need for caches to implement all possible existing P2P protocols. On the contrary P2P applications need to implement this standardised protocol to be able to use caches.

The in-network storage service and the protocol which support it will include storing, retrieving and managing data as well as specifying both access control and resource control policies in the in-network storage pertaining to that data

In the next section, the list will present different market solutions including most of the targets traffic presented before. This list is not exhaustive and will probably improve following deeper research.

5.4.1.3 Oversi

Target: P2P, HTTP Download, HTTP Progressive Download

This solution is based on integration of P2P overlay network elements like peers, trackers, super peers. It is not a transparent solution but it supports encrypted Bittorrent and eDonkey protocols.

The solution is named “out of band”, because it needs only to intercept and analyse the upstream traffic (from end-user to origin server). This system is based on 2 different devices, first one (HM or http manager) which analyses the requests from end-users, second one (HCS or http cache server) in charge of storage and delivery. Both communicate under http. End-users are redirected to HCS, so cache is not transparent for end-user. HCS is in charge to ingest content, so cache is not transparent for origin server. Even if the solution is not transparent, the interest is to avoid diverting all traffic through the cache (like “in-band” solution)

Copyright © ENVISION Consortium, May 2011

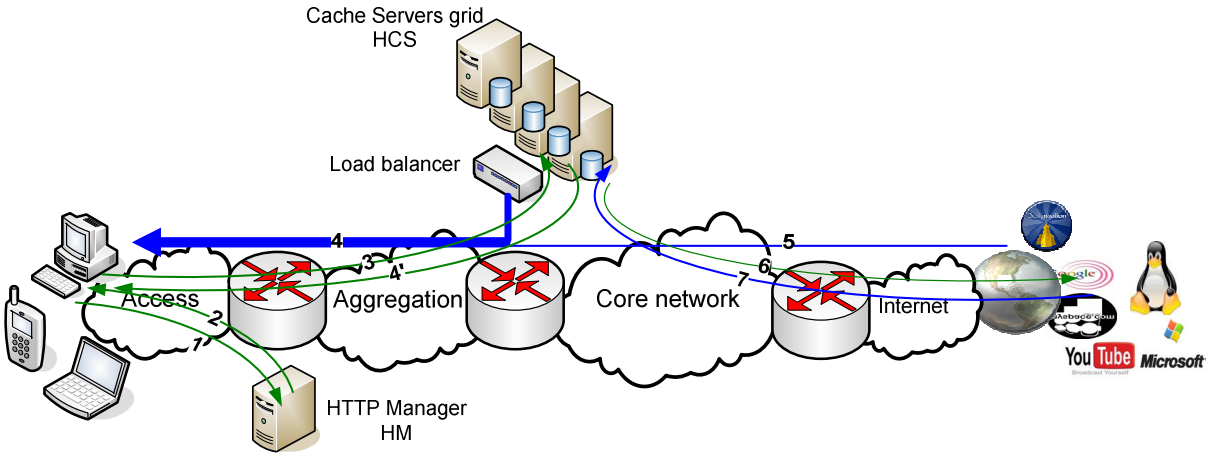

Figure 26: "Out of band" cache mechanism (as Oversi)

For http traffic, HM is in charge to analyse http GET messages. These messages are diverted to HM by BGP announcement (HM announces to ISP routers, it is the next hop for content sources to be cached) or by PBR. HM answer to end-user with a redirect to HCS (http 303). In case of cache miss, HCS send to end-user a redirect to the origin server (http 303). End-user receives content from origin server. Asynchronously and following popularity policy, HCS send its own request for the same content and store it. If content is already in cache, the HCS serves the end-user. Many content providers use http GET messages to calculate statistics on content demands, available bandwidth, etc…These specific messages are treated by HM as empty GET and redirects to origin server.

P2P caching mechanism: Oversi provides P2P cache based on using overlay P2P network elements as peers, super peers, trackers. This solution can be used to Bittorrent and eDonkey protocols, in unencrypted/encrypted mode. As for HTTP, Oversi solution uses 2 main appliances, PM (peer manager) which is in charge to manage information about cached content and promotes the cache (P2P peer or super peer) in the peer lists, and PCS (peer cache server) which is in charge of ingestion and delivery of the chunks. PCS is in fact a concentration of peers which appears as elements of P2P overlay network.

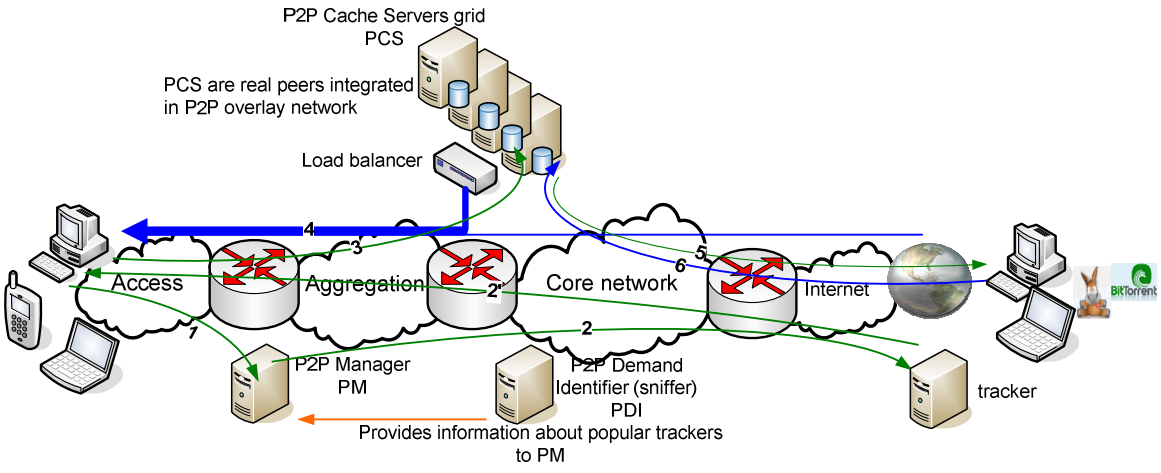

Figure 27: Oversi's solution for P2P caching

Present in multiple countries in the world, Oversi provides its solution for P2P and HTTP to ISP or Cable communication companies [OVERSI07]

Copyright © ENVISION Consortium, May 2011

5.4.1.4 PeerApp

Target: P2P, HTTP Download, HTTP Progressive Download

This solution is based on real transparent cache mechanism. The solution spoofs IP addresses: end-user "sees" Origin server IP address, origin server "sees" end-user IP address.

As mentioned above, this method needs to have UP and DN traffic direction analysed together, so unidirectional traffic (2 different network ways) have to be joined in a same path.

Traffic is diverted to the cache by PBR or DPI method. The cache has to be able to manage all the diverted traffic and also its own traffic generated by the content which is inside the cache.

The cache is able to recognise http requests (GET based) or P2P control messages for main P2P protocols and associated P2P clients (Bittorrent, eDonkey, Gnutella, FastTrack). Currently, it is not able to recognise encrypted P2P protocols.

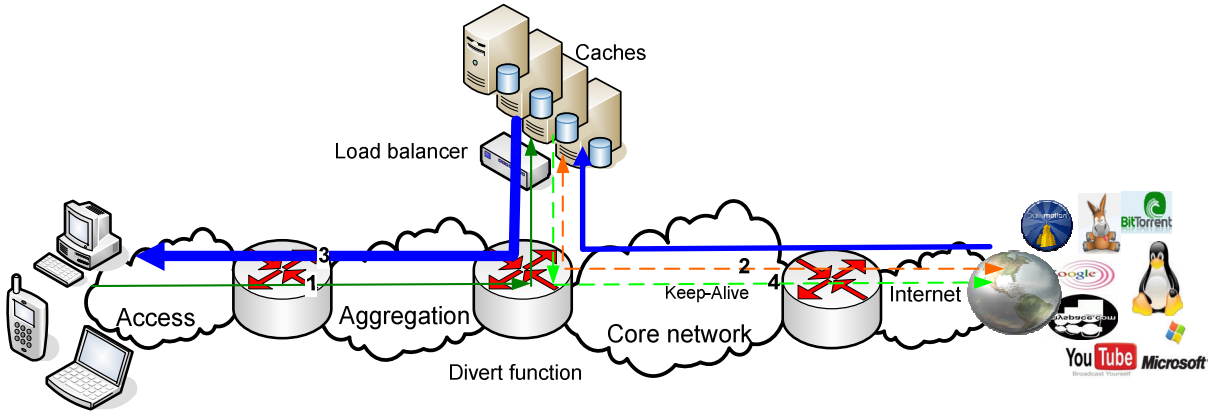

Figure 28: Full transparent caching solution (from PeerApp)

This solution is full transparent and the mechanisms are similar for equivalent market products.

Session between end-user and Internet content source (http or P2P) is diverted to the cache equipment thanks to a "divert function". This function, as described above, is based on L4 on L7 classification criteria. When end-user sends a request (e.g. http GET), the cache receives it and propagates it as it is to the content source, including original source IP address, port number and application request. Exchanges between content source and end-user are made to control business logic (access control, authorisation and accounting) and in final, content source approves the transaction. In case of transaction denial, the cache doesn't take any action.

If content is not already stored in cache (cache miss), the cache forwards traffic as it is between source and destination. Following popularity policy (store only most popular content) and configurable content size, cache catches content (video files, P2P chunks) during the transfer inside the cache.

In case of "cache hit", the cache takes over the session cutting the download from content source after few packets, maintains the session (keep-alive) to transparency respect and delivers the data from storage using content source IP address.

When content is not present in the cache, cache ingests the content simultaneously with the end-users, thanks to the divert mechanism which redirects up and down stream to the cache.

Copyright © ENVISION Consortium, May 2011

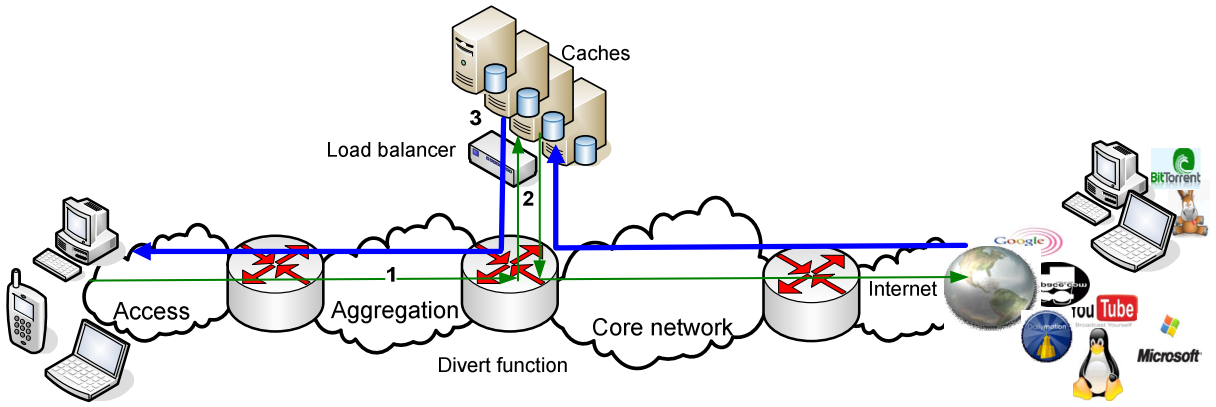

Figure 29: Ingestion in cache network (from PeerApp)

When ISP subscriber queries P2P network and finds a file to download from P2P users on the Internet, the divert function transparently redirects P2P traffic to the cache. If file is already stored in cache, cache stops the download from remote peers, serves ISP subscriber's request directly hence saving bandwidth on the Internet transit link and maintains the connection with the Internet user to respect transparency.

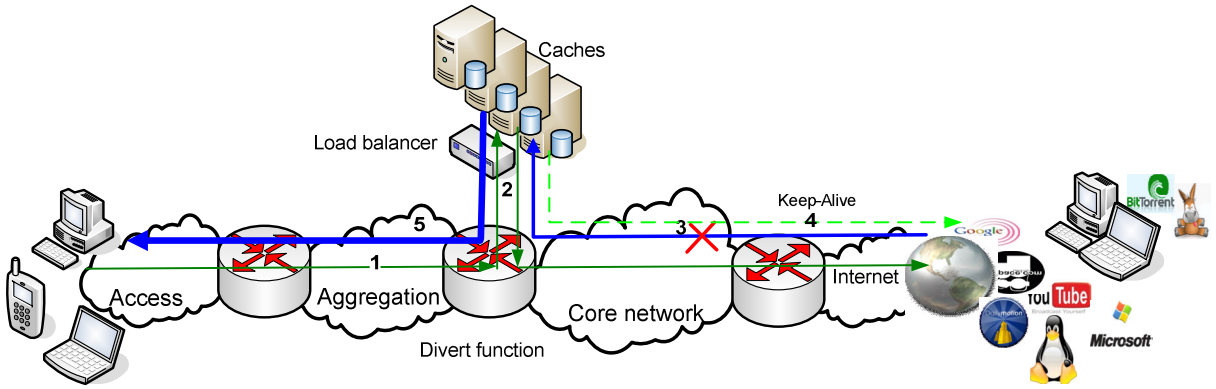

Figure 30: Download with cache hit

In case of request from remote peer, cache is able to serve the content instead of local peer, saving bandwidth consumption in the last-mile. In that case, the mechanism is the same as above, divert function transparently redirects P2P traffic to cache. Cache stops the upload traffic from local peer, serves content to remote peer and maintains the connection with local peer, respecting transparency.

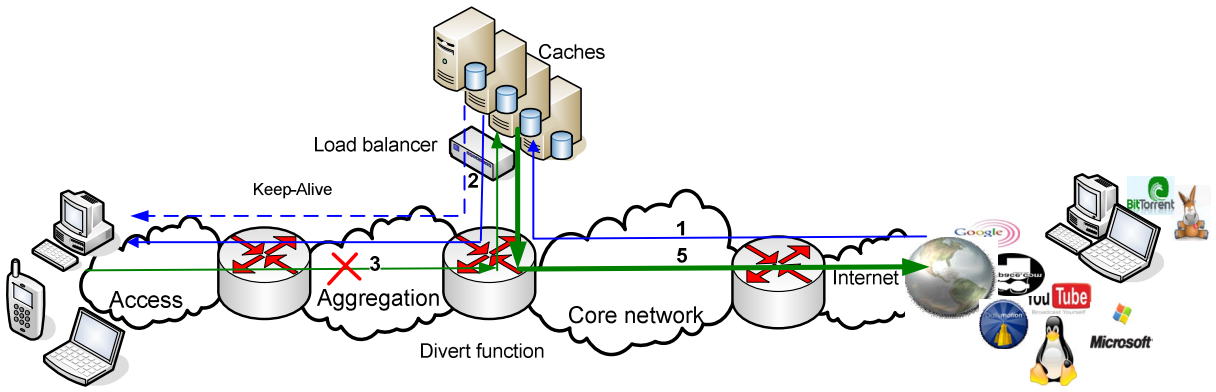

Figure 31: Upload with cache hit

Copyright © ENVISION Consortium, May 2011

PeerApp is present in many countries in the world. [PeerApp09]

5.4.1.5 Coblitz

Target: HTTP Download, Progressive download, adaptive streaming

Based on Codeen solution (Planet Labs), it is a hierarchical solution using a "P2P" model to store pieces of content, but it is no fully transparent. Each node is a proxy and redirector. Cache nodes store pieces of content and cache miss involves a redirect to the upper level of cache node to retrieve the missing pieces. In case of total cache miss, the request is sent to the origin server. Divert method is based on WCCP (Web Cache Communication Protocol), a Cisco protocol close to PBR with additional features like load-balancing. Specific equipment named CoTTC is in charge to analyse the request and redirects it to the cache nodes.

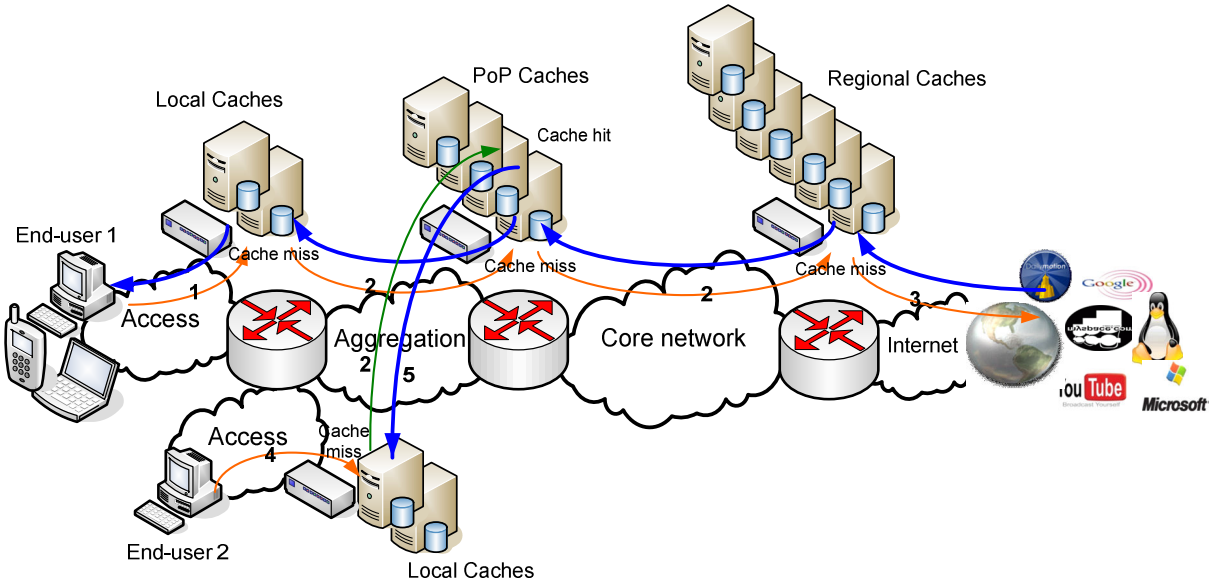

Figure 32: Coblitz architecture. Peering and parenting

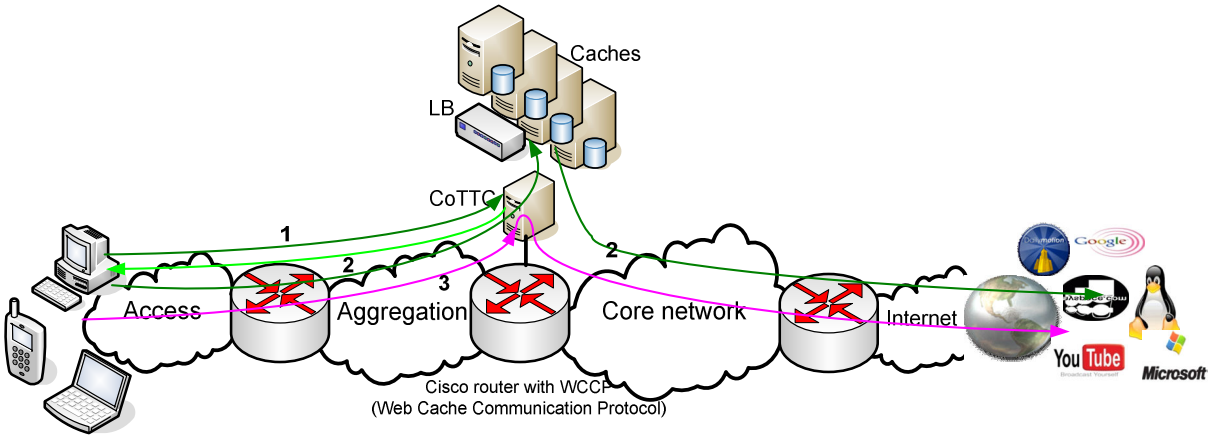

Figure 33: Coblitz divert mechanism (CoTTC)

Each cache contains an agent which is charge to manage requests for large files, dividing it in multiple requests for small pieces of the content i.e. chunks. These new requests are not seen by the client or origin server because it is an internal URL name. The agent of this node sends the different

Copyright © ENVISION Consortium, May 2011

requests to retrieve all the chunks in the different nodes following the byte range. The chunks are reassembled to provide content to the end-user.

In case of cache miss, the requests are sent to the origin server (request for content with the byte range). The origin server sends a set of partial content corresponding to the requests sent by the different nodes. The chunks are stored inside the different nodes and are now available when an identical request is sent by an end-user [Coblitz06]

5.4.1.6 Other solutions

For P2P caching, solutions are not so many than http caching.

In this last family, the software solution (open source) named PCache ([PCache]) is a full transparent caching solution based on a transparent proxy which intercepts the P2P requests, identifies P2P traffic (protocol) and serves content if it is already inside the cache. For all possible cache misses (complete or partial content), the request is forwarded to P2P network. This solution is faced to the main issue which is to recognise P2P protocols.

For http traffic caching, many solutions are on the market. These solutions aim to cache web objects from the smallest one to large items of content such as video. The new improvements in these solutions are to provide capacities to cache new video transport protocols as SmoothStreaming, Adobe HTTP Dynamic Streaming or Adaptive Bitrate Stream.

Classical solutions like BlueCoat, Appliansys, Squid (open source) are often based on transparent proxy and target all type of files downloaded thanks to HTTP protocols. A solution like Ankeena provides a full transparent solution closed to PeerApp but supporting also adaptive streaming.

Some products like Steelhead from Riverbed [Riverbed] are considered as WAN optimizers and in this way develop caching solutions to accelerate business applications including file sharing for example. In the same category of products and also used to improve application performance on branch offices, Branchcache [Branchcache] is a new peer-to-peer caching solution included in Windows® 7 and Windows Server® 2008 R2. Branchcache can operate in Hosted Cache mode or Distributed Cache mode depending on where the cache is located. The Hosted Cache mode operates by deploying a computer that is running Windows Server 2008 R2 as a host from which users can retrieve cached content when available. With fewer than 50 users, BranchCache can be configured as a peer-to-peer architecture in Distributed Cache mode with local Windows 7 clients keeping a copy of the content and making it available to other authorized clients that request the same data.

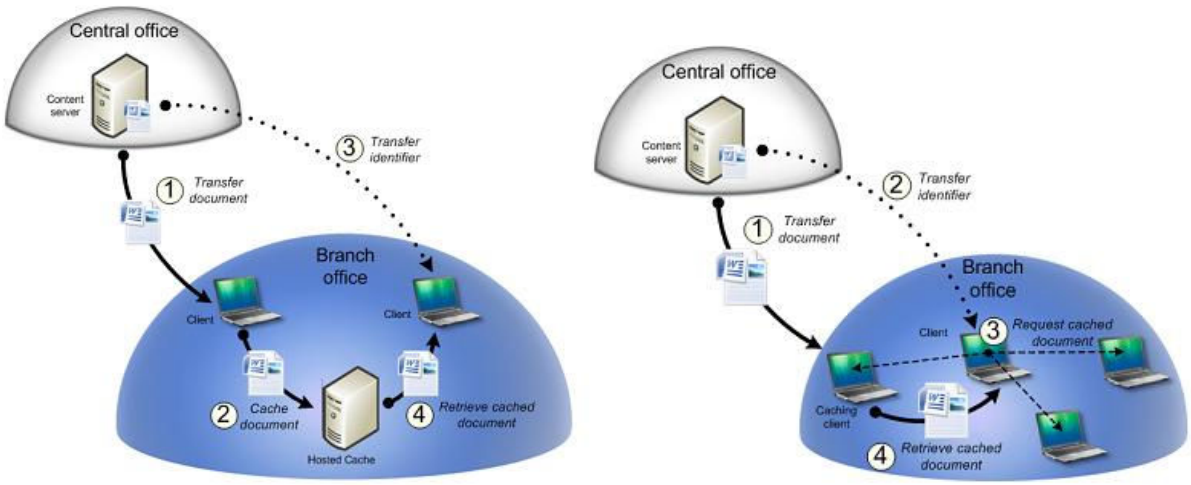

Figure 34: Hosted Cache mode & Distributed Cache mode

Copyright © ENVISION Consortium, May 2011

The mechanism used by Branchcache for reducing bandwidth is to send content metadata to clients, which retrieve the content from within the branch. The content is broken into a collection of blocks and a hash is computed for blocks providing a unit of download and segment of blocks providing a unit of discovery.

5.4.1.7 Evaluation of some deployed cache solutions

This paragraph presents an example of the deployment of a cache solution onto an actual network with less than 20K IP addresses**,** a 400 Mbps Cache and a storage capacity of 12 TB. ISP researches both improvements: bandwidth saving and QoE for its end-users. In term of QoE, the goal was to improve the download speed.

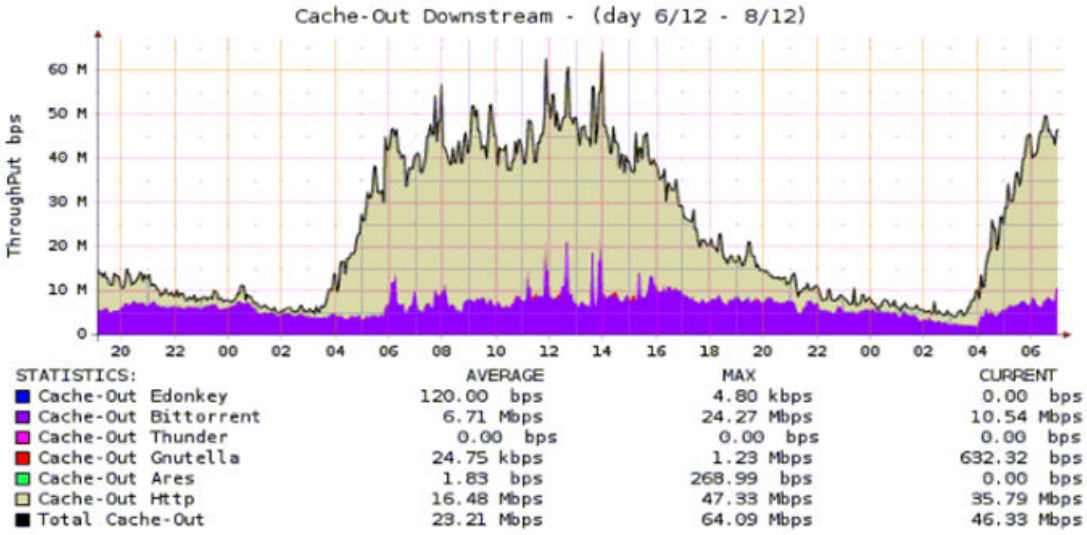

Figure 35: Traffic generated by cache network for http and P2P (mainly Bittorrent)

Max/Average data	Total download (Mbps)	Cache Out (Mbps)	% (cache out vs. total DL)
HTTP traffic	156/62	47/16	30/26
P2P traffic	28/15	25/6	90/40
Total	189/84	64/23	33/27

Figure 36: Byte hit ratio of a cache network

With such a solution the network operator can save between 25% and 35% of its bandwidth needs. For P2P the gain is rather between 40 and 90%.

The figures below show the QoE improvement with HTTP and P2P Cache in term of download speed

Copyright © ENVISION Consortium, May 2011

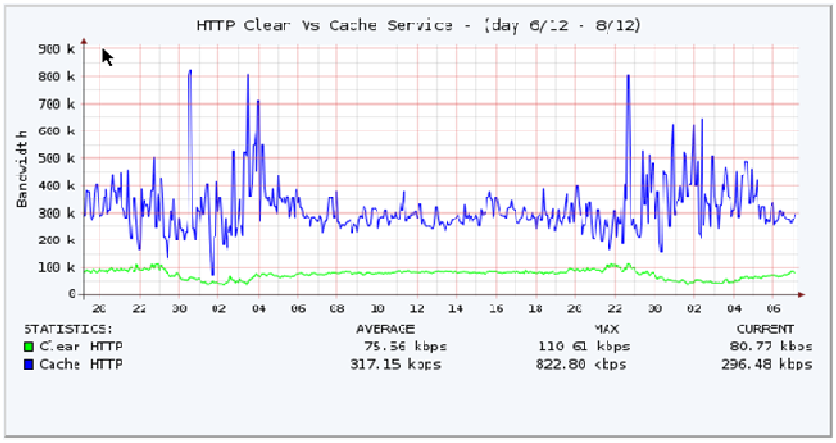

Figure 37: QoE based on download speed. Gain = 400% for HTTP traffic

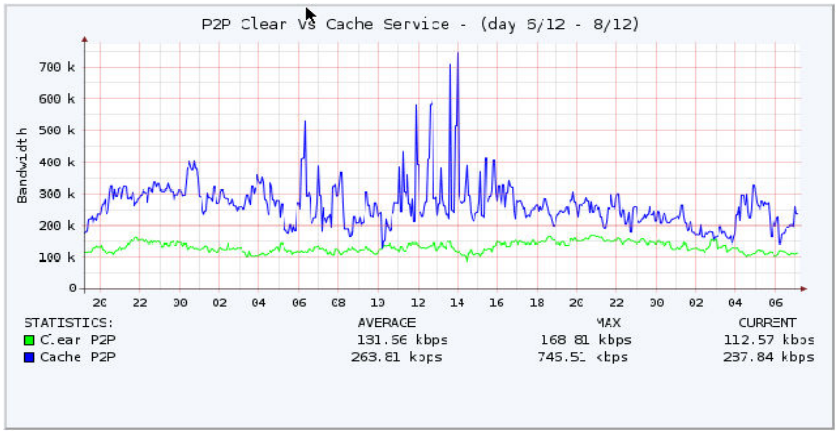

Figure 38: QoE based on download speed. Gain = 200% for P2P traffic

For this ISP, gain in term of bandwidth is not negligible when international links are based on satellite and submarine cables links. The QoE is also improved: the download speed is improved 4 times in comparison with no cache deployment.

The following graph represents the incoming/outgoing traffic on a cache and its added value where the transit link fails. Incoming traffic fall to “0” but a part of P2P traffic is generated by the cache and sent to ISP customers.

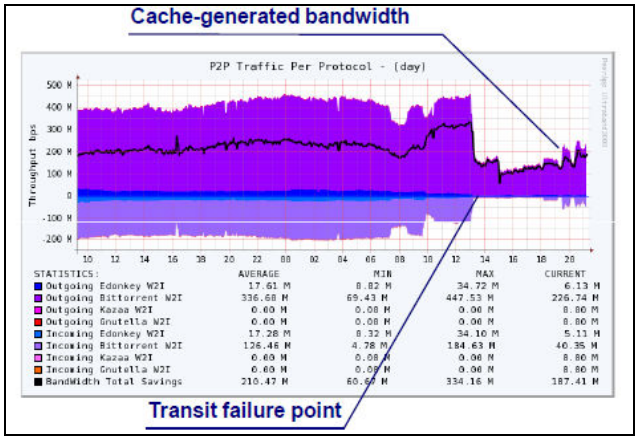

Figure 39: P2P generated traffic by cache. Transit link failure is partially covered by cache delivery

Copyright © ENVISION Consortium, May 2011

5.4.2 Cache performance evaluation

In this section, we study the evaluation of cache performance in relation with the underlying network. Caches can benefit greatly the network by alleviating the load on particular links. They are especially useful when the links are overloaded because they diminish redundancy in the over-the - top traffic. In consequence, caches can prevent congestion, even if the traffic is greater than the link capacity. Evaluating cache performance is an important topic because it can greatly impact architecture choices made when designing the network.

In this section, we will give an overview of the mathematical models of caches and apply them on the particular case of real traffic observed in the Orange network.

The most widely used metric for evaluating cache performance is the Hit Ratio (HR). The HR is the ratio of requests fulfilled by the cache and the total number of requests seen. However, depending on the context in which the cache is used, other metrics can be more relevant. For instance, in order to determine how to provision a particular link in the network, it is useful to know the maximum throughput of the traffic. In that case, it is useful to know for instance how many sessions can be served by a cache when the traffic is at a peak. Knowing this allows determining how to provision links so as to have a good quality of service for users, while avoiding costly over-provisioning of network resources. In that particular case, a metric that is interesting to look at is the Peak Hit Ratio (PHR), i.e. the hit ratio at traffic peaks.

In the following, section 5.4.2.1 gives a state-of-art on analytical studies on cache performance. Then in section 5.4.2.2, we study the property of the traffic of a particular service, namely Video on Demand (VoD). On this particular case, we apply the analytical studies detailed in the section before. Next, in section 5.4.2.3 we give the results of a cache simulation based on the traffic trace used in the section before. We then compare the results of the simulation with the results of the analytical studies and we conclude that the mathematical models developed so far are not valid in practice for a commercial VoD service. Finally, section 5.4.2.4 concludes this part on cache performance evaluation and describes the future work in this area within the project ENVISION.

5.4.2.1 Analytical studies on cache performance evaluation

In 1992, Flajolet and al. have determined the hit ratio of a caching algorithm [FLA92]. To do so, they have formulated the following assumptions:

1) The requests are independent of each others
2) The caching algorithm is LRU (Least Recently Used), i.e. when the cache is full and a new object must be stored, the object that was the least recently used is removed.
3) The popularity of objects follow a Zipf law, i.e. the probability that an object is requested is given by equation 1, with r the rank of popularity of an object, c a constant, and α the Zipf parameter ($0 < \alpha$).

$$p(r) = \frac{c}{r^{\alpha}} \qquad \textbf{Equation 1}$$

Given these assumptions, Flajolet and al. obtain a formula for the hit ratio but unfortunately it is not usable for practical purpose because of its combinatory complexity. As an illustration, with a cache of size 20 and 1000 objects, the formula requires 10^{40} operations to be computed.

Breslau et al. [2] have simplified the formula of Flajolet and al. with the following additional assumptions:

1) The cache size is very large
2) The Zipf parameter α is strictly greater than 1

Copyright © ENVISION Consortium, May 2011

In this case, the hit ratio is given by equation 2, with H the hit ratio, C the size of the cache, N the total number of objects, Γ the Euler function, and q_i the probability of request for object i.

$$H(C) \approx 1 - \left[\left(1 - \frac{1}{\alpha}\right) \left(\Gamma(1 - \frac{1}{\alpha} \right)^{\alpha} \sum_{i=C+1}^{N} q_i \right] \quad \textbf{Equation 2}$$

In the case when the LFU (Least Frequently Used) algorithm is used, the hit ratio is simply expressed by Breslau and al. in [BRE99] by equation 3, with the same notations as before.

$$H(C) = \sum_{i=1}^{C} q_i \quad \textbf{Equation 3}$$

Note that equation 3 does not assume a large cache size and a particular value for α. Actually when C is very large and α = 1, we have $H(C) \approx \ln(C)$. Similarly when C is very large and 0 < α < 1, we have $H(C) \approx C^{(1-\alpha)}$.

The popularity distribution may not follow a Zipf law (as we will see in the next section), but rather the Weibull law, as expressed in equation 4, with c a constant, and β, λ the parameters of the law.

$$p(r) = c.e^{-\lambda r^{\beta}} \quad \textbf{Equation 4}$$

In the case of Weibull popularity and with assumptions 1, 2 4, Jelenkovic [JEL99] has given the following approximated value of the hit ratio, cf. equation 5, with the same notations as before and e^{γ} a constant (≈ 1.78107…).

$$H(C) \approx 1 - \left(e^{\gamma} \sum_{i=C+1}^{N} q_i \right) \quad \textbf{Equation 5}$$

This equation can also be approximated by equation 6.

$$H(C) \approx 1 - \left(\frac{e^{\gamma}}{\beta\lambda} C^{1-\beta} q_C \right) \quad \textbf{Equation 6}$$

Panagakis and al. [PAN08] have investigated the case when the requests are not independent of each others, in the context of web caching and when the popularity of objects follows a Zipf law.

However, it was shown that for other services than web pages the popularity of objects is better modelled with other laws than a Zipf law. Indeed Saleh and Hefeeda [SAL06] and Carlinet and al. [CAR10] have shown that peer-to-peer (P2P) exchanges follow a Mandelbrot-Zipf law. In the same manner we will see in the next section that video on demand (VoD) traffic is better modelled with a Weibull law, due to the very nature of the service. In consequence, the mathematical models developed for predicting the hit ratio of a cache with requests following Zipf are not valid when requests follow a Weibull law for instance. Indeed there are fundamental differences between these two laws that change dramatically the performance of caches, one of these differences is that Zipf is heavy-tailed, while Weibull is light-tailed (popularity decreases exponentially).

5.4.2.2 The case of the Video on Demand service

In ENVISION, caching can be used for a catch-up TV service during an event, as explained in the micro-journalism use-case for instance. This service is very different in nature from web browsing;

Copyright © ENVISION Consortium, May 2011

hence the distribution of popularity of objects is likely to be very different as well. The profile of the demand is closer to the one of a web streaming or a VoD service. Like in VoD, videos of previous parts of the show are made available to the clients of the service to download and watch.

In this section we investigate the properties of a VoD traffic trace in order to determine the cache theoretical performance.

To this end, we have measured the traffic of the Orange VoD service inside the operational network. The traffic trace obtained this way is a record of the requests of real users for a commercial service in operation. The trace was rendered anonymous before any other analysis. The trace is 2-year-long (from June 2008 up to July 2010) and contains around 1,840,000 requests, performed by 42,700 distinct clients, on 44,700 movies. Around one third of the requests are trailers, while the remaining two thirds are full-length movies.

The Figure 393 shows the distribution of the popularity of the objects in the VoD service, i.e. the movies. The X-axis represents the rank of the object by order of decreasing popularity, i.e. the object with rank 1 is the most requested. The Y-axis represents the number of requests.

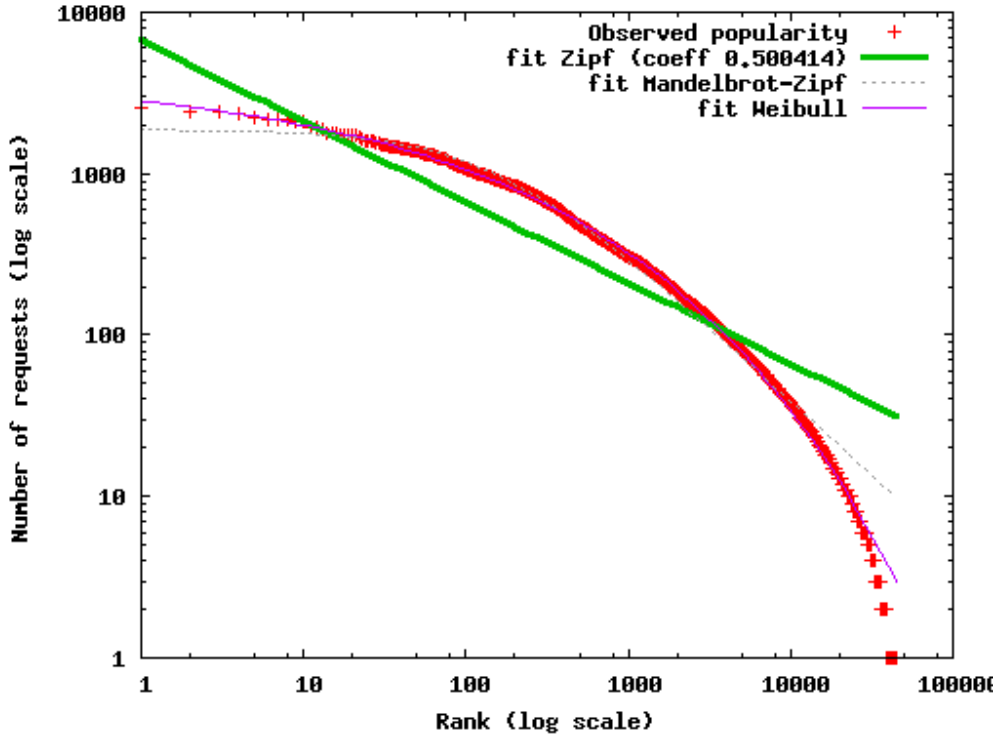

Figure 40: Popularity of VoD movies (log scale)

The figure also shows several fits, with a Zipf law, a Mandelbrot-Zipf law and a Weibull law. The Mandelbrot-Zipf law is given by the function of equation 7, with c a constant and q and α the parameters of the law.

$$p(r) = \frac{c}{(r+q)^{\alpha}}$$ **Equation 7**

The fits are performed with the least-square algorithm. We can clearly see on the figure that the observed popularity is not Zipf, since Zipf on a log-log scale is represented by a line. Indeed, the Weibull law is the one fitting best the popularity distribution.

Copyright © ENVISION Consortium, May 2011

Next we are using these values of popularity, measured in our trace, in order to compute the equations given in the previous section.

Equation 2 cannot apply on this data trace because the Zipf parameter is not greater than 1. However we can apply equations 3, 5 and 6 that we will name respectively 'LFU Breslau', 'LRU Jelenkovic' and 'LRU Jelenkovic 2' for more clarity. Figure 394 shows the results of the computation of the analytical studies detailed in the previous section.

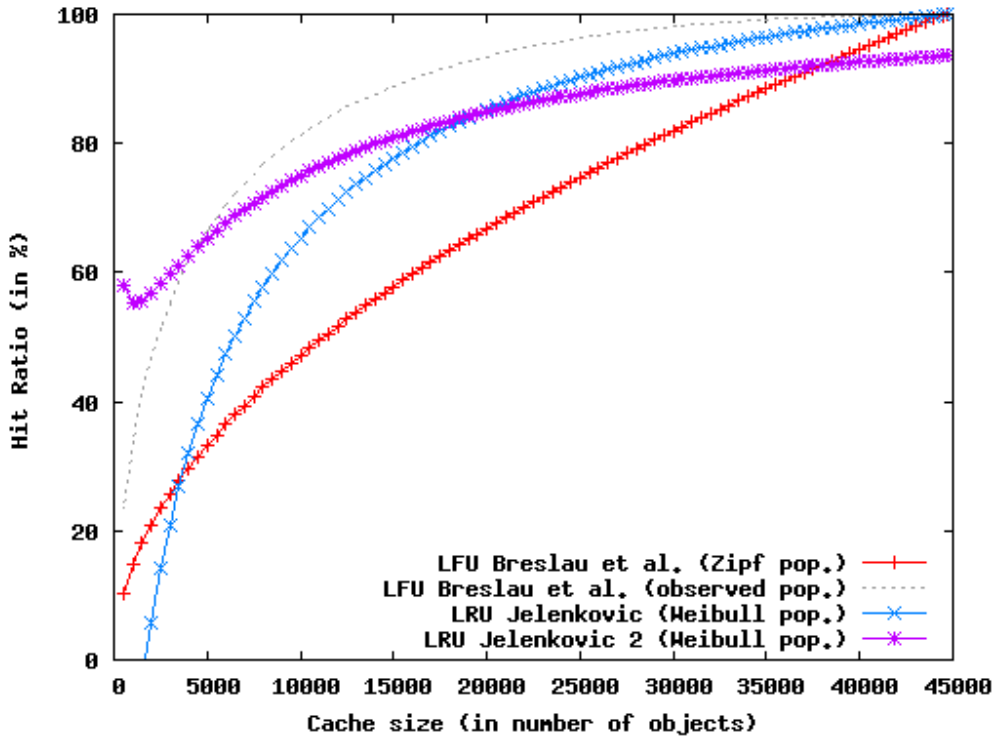

Figure 41: Numerical computation of analytical studies on hit ratio

In the figure, the difference between the two instances of 'LFU Breslau' is that in the equation, we can use either the popularity values that we got when fitting the popularity, or alternatively we can use the observed values.

We can observe on the figure that the most efficient algorithm is LFU (in theory). We can also see that the performance of 'LFU Breslau' is very low with the values of the Zipf fit. This is of course because the Zipf fit is very poor; therefore the computed theoretical hit ratio is off.

Of course these results were obtained with assumptions that are not necessarily verified in the real traffic trace; therefore those theoretical results should not be taken for granted. For instance, the assumption that all the objects are of the same size is not verified in the trace. More importantly, requests are very likely to be correlated with each others, while all the models assume they are independent.

In order to check if the theoretical results give useful results, we will perform a cache simulation based on the real traffic trace and compare with the results of this section.

5.4.2.3 Cache simulation results

The cache simulation is very simple. It assumes the cache sees all the requests in the trace. When the request is on a movie already stored in the cache, we assume the cache serves it and we count a hit. When the movie is not in the cache, we assume it stores the content when the movie is transmitted

Copyright © ENVISION Consortium, May 2011

from the VoD server to the client. When a movie is stored, it is then available for the requests afterwards. When the cache is full and it has to store a new movie, it removes a movie according to its replacement policy (LRU or LFU). For comparison purpose, the simulation also implements a third replacement policy called "random" where the movie to be removed is selected randomly among the movies stored in the cache.

Figure 395 shows the results of the cache simulation.

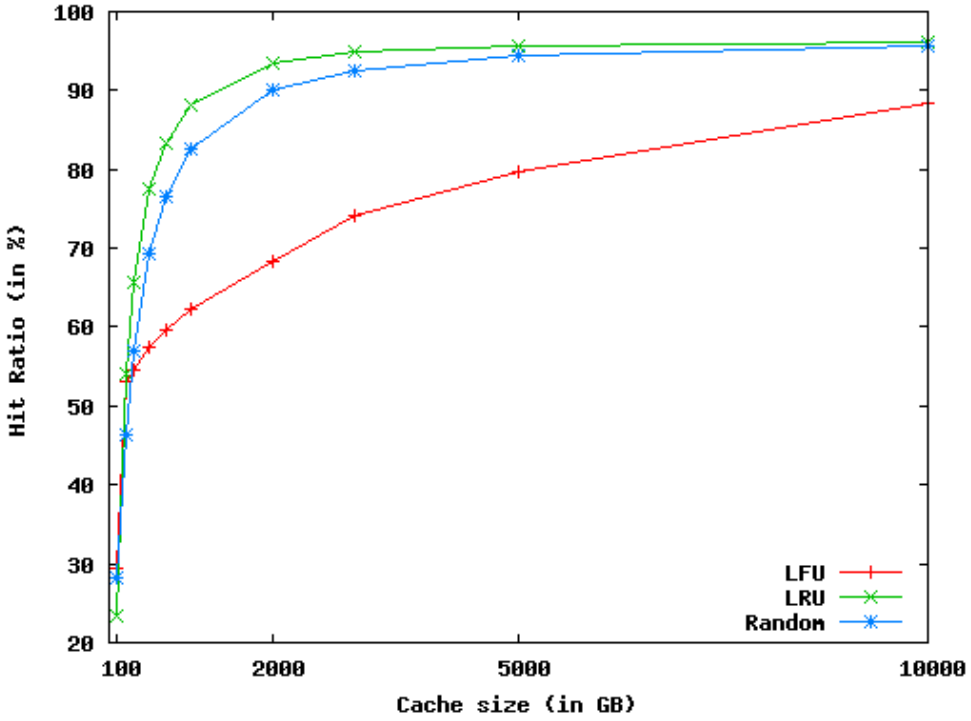

Figure 42: Hit Ratio of simulated cache

We observe that LFU is much worse than the two other replacement policies, even "Random". This is unexpected because the theoretical performance of LFU, computed in the previous section, is the best of all the replacement policies. The second point to notice is that the "real" performance (in the sense of the performance measured on real traffic) is higher than expected by the mathematical models.

The low performance of LFU can be explained by the fact that popular movies are requested a lot during a relatively small amount of time. Therefore a popular movie (with a lot of requests) will be stored in the cache and when it is not requested anymore, it will stay in the cache forever, or at least for a very long time, until it is replaced by more popular movies. In summary, the LFU cache is quickly filled with objects that were requested a lot but are not requested afterwards.

The high performance of LRU and Random can be explained by the fact that requests are temporally correlated. This factor is not taken into account by the mathematical models, but it has a great positive impact on performance. This is because when a movie is stored in the cache; it is very likely to be requested again soon after, thereby increasing the hit count.

So as to have a better idea of the difference between the theoretical performance and the trace-based performance, Figure 396 shows both curves on the same plot. The figure also shows the performance of each algorithm based on a randomised trace (labelled "Rand.-tr."). This randomised trace is the traffic trace in which the order of requests was changed in a random fashion, thereby destroying the correlations between requests.

Copyright © ENVISION Consortium, May 2011

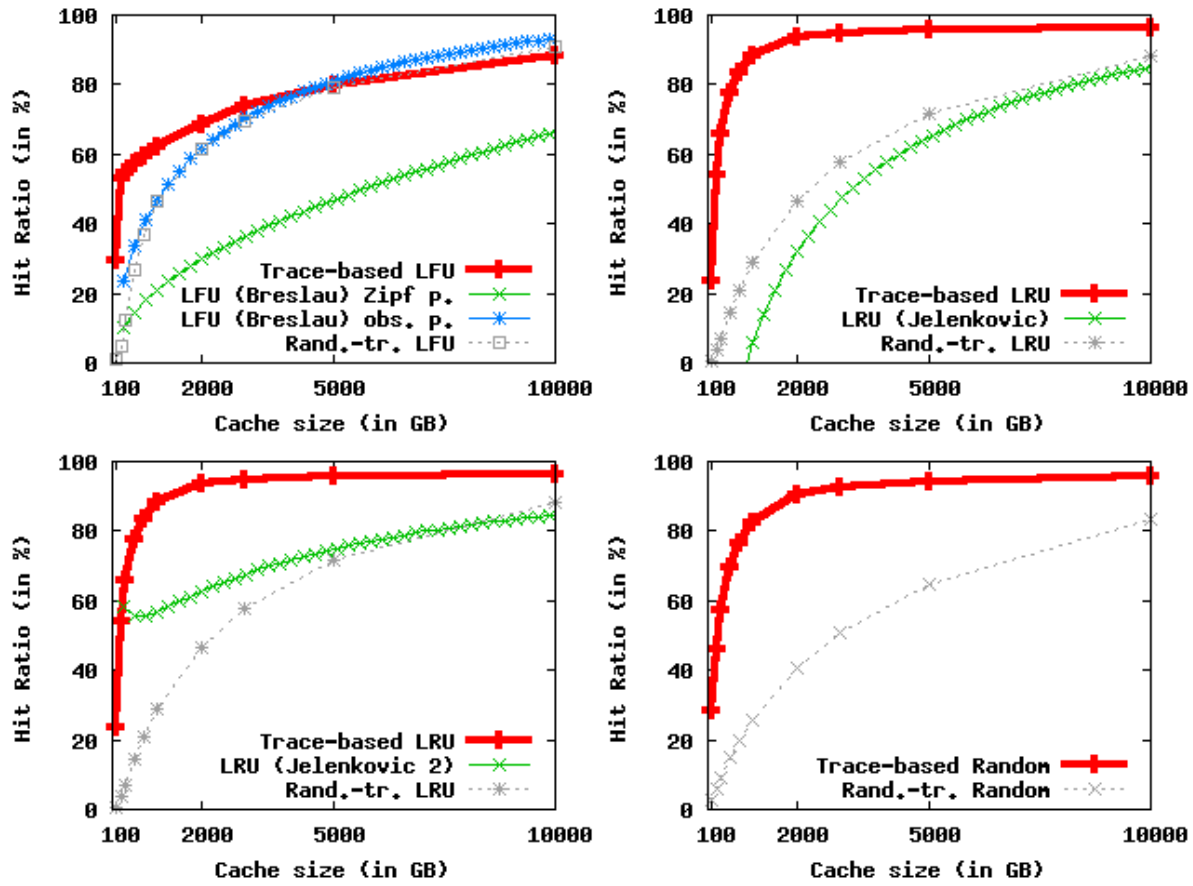

Figure 43: Comparison between trace-based HR and theoretical HR

The plot at the top on the left hand side shows the HR of an LFU cache, respectively based on the traffic trace, with the Breslau formula (with fit and observed values of popularity), and on the randomised trace. The plot at the top on the right hand side shows the HR of LRU, respectively on the traffic trace, with the Jelenkovic formula and on the randomised traffic trace. The plot at the bottom at the left hand side is the same but with the second Jelenkovic formula. Finally, the plot at the bottom on the right hand side shows the HR of the "Random" replacement algorithm on the traffic trace, and then on the randomised trace.

This figure clearly uncovers the impact of correlations between requests, since when the correlations are destroyed; the performance takes a big cut (as one can see by comparing the red line with the grey dotted line, on each plot).

5.4.2.4 Conclusion on cache performance evaluation

We have presented a state-of–the-art on mathematical modelling of a cache. Next we have applied the theoretical formulae with a real traffic trace from a VoD service in operation. We have performed cache simulations in order to compare the results with the theoretical results. We could then conclude that several factors cannot be neglected when trying to predict the cache Hit Ratio, such as requests correlations, for instance.

The traffic of every network service has different properties that must be taken into account when designing a network cache architecture. For instance, web traffic is directed from a lot of users to a lot of objects, with long-term patterns. On the contrary, VoD traffic is coming from a limited number of clients to a limited number of objects and after a while, objects disappear from the traffic. The web video streaming service is probably in between these two.

Copyright © ENVISION Consortium, May 2011

In the next deliverable D3.2 (Refined specification of the CINA interface, network monitoring and network optimisation functions), we will study the optimal placement of caches in the network. Indeed, caches can be located close to the source of the content, or close to the users. On one hand, if the cache is close to the source it sees the requests of a lot of clients, therefore it has a high Hit Ratio, but it only saves bandwidth on a short path (between the source and the cache). On the other hand, if the cache is close to the users, it saves more bandwidth, but the Hit Ratio is lower due to the lower redundancy in the traffic. There is a compromise to find between the Hit ratio achieved and the bandwidth saved. In order to find the optimal point, costs must be taken into account.

We will also study another service such as web video streaming over mobile terminals, that might be more closely related to the bicycle race use case. We will perform another monitoring experiment on network equipments in the mobile networks, in which we log information about web streaming. Finally, we will develop a model that takes into account the correlations between requests, so as to predict the hit ratio of a cache with great reliability.

5.4.3 Live streaming distributed caching

While overlay applications are today able to download stored content, it is highly demanding to deliver fair quality live streams shared amongst users because the underlying network does not enable a rich service capable of delivering live streams in a cost effective way taking benefit of shared buffers amongst peers.

The Bicycle Race use case, for instance, would be very costly to achieve nowadays for a software development company of overlay applications because:

- The mobile bandwidth is limited, and only few video streams can be supported in the cell
- The Internet traffic between mobiles of different ISPs is not guaranteed

A way to override those for the overlay application would be make use of a live streaming cache service provided by the network via the CINA interface. It is unfeasible to expect that overlay applications by themselves will each of them construct their own implementation since the distributed caching of live streaming is complex to research and develop.

To deliver unique user experience minimising costs to live streamed content amongst ENVISION applications, the goals are:

- Reduce the ISP traffic of live streams by designing live caches in the edges of the ISP network
- Provide a simple ingestion point for the live stream via the CINA interface, and offloading to the ISP the responsibility for distributing the live stream.
- Enable overlay applications to provide end users with “rewind capabilities” on live streams. Therefore, caching buffer replicas across the network.
- Minimal investment costs in equipments designing a mechanism potentially implementable over any plain 1K€ servers.

In order to reach these goals, the research relays in P2P amongst caches hosted in the edge of the network for efficient cache synchronisation.

Copyright © ENVISION Consortium, May 2011

5.5 Network-aware multilink distribution

5.5.1 Introduction

We refer to the term *multilink* in the context of an end-host with multiple layer 2 interfaces to different access networks. The interfaces may be to wired or wireless networks of a mix of technologies (2G, 3G, Wi-Fi, WiMax, ADSL, optical, etc.). It is possible for a host to have more than one interface to the same provider but, in the general case, each interface will access the Internet through different access providers/ISPs. Multilink nodes are therefore multihomed and are able to communicate with other nodes/servers/peers through any, some or all of their interfaces simultaneously. There are three basic scenarios to be considered as depicted in Figure 44: a multilink-enabled node receiving content simultaneously from several senders (Figure 44(a)); a multilink-enabled node sending simultaneously to multiple destinations over several networks (Figure 44(b)); or two multilink-enabled peers communicating across several networks exploiting the aggregate capacity across all access networks (Figure 44(c)).

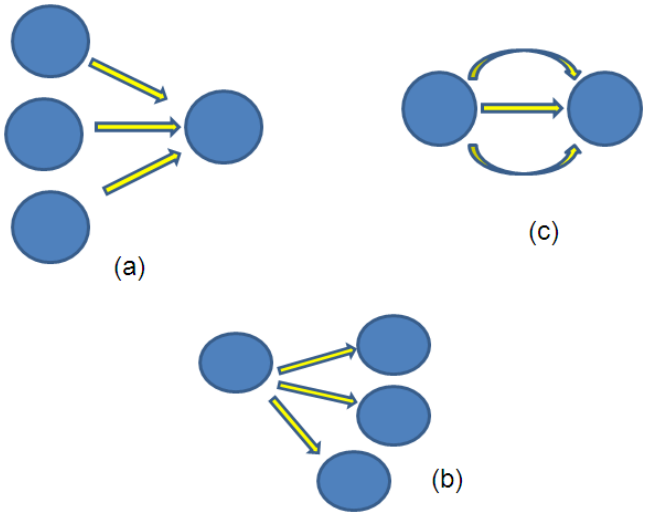

Figure 44: Multilink illustrations

The *links* in this definition consist of physical layer 2 network capabilities. IP flows may be established between peers that are originated or terminated over these links and more than one IP flow may exist simultaneously on a single access link and each may be terminated at different destinations. The arrows in Figure 44 illustrate the topological connectivity between nodes from the perspective of layer 2 communications as seen by each end-host only: they are not intended to represent the full complexity of the end-to-end communication paths for flows established between those nodes. In general, an IP flow will be routed over multiple links interconnected by layer 3 routers and the end-to-end path may span more than one administrative domain/AS.

Network aware multilink refers to network services which take into consideration the multilink capabilities and provide services over those links. Aggregated services may refer to bandwidth aggregation, load balancing, priorities and more.

Multilink-enabled nodes participating in P2P swarms as sources, sinks or relays will be at an advantage from two perspectives: they may exploit the additional network resources available from multiple access links to send or receive more data simultaneously; and, when selecting a remote peer as either a source or a destination, they may optimise the selection according to the expected network performance and/or cost to the remote peer across the range of access networks to which either node is connected. However, in this deliverable we are focusing on the ISP network. The ALTO work is aiming to localise connectivity across swarms by advising the P2P overlay applications of their preferences for peer selection with the target of reducing cross-ISP traffic. An ISP may have several

Copyright © ENVISION Consortium, May 2011

access networks so an issue for an ISP supporting multilink peers would be to extend the ALTO specification to provide preferences (or other parameters considered in ENVISION, see sections 3.2 and 4.3) per peer and per access network to capture the fact that an ISP may have a high preference for a specific peer when accessed through one interface/access network but a medium or low preference from another.

5.5.2 Challenges of real-time video services

The use of multilink fits both wired and wireless networks, the challenge of broadband mobile services, especially for high quality video, however, may speed up demand for multilink. The micro journalism scenario as defined in D2.1 for the bicycle race use case brings live video from anywhere to be further distributed to millions of peers. Both professional photographers and amateur could generate the content, which could be distributed to residential users via several ways, through P2P and TV channels. Thus the importance of achieving high quality at the source is critical and the use of multiple wireless-links is an attractive solution.

5.5.3 State of the art

5.5.3.1 Priority handling in network nodes

The Internet Protocol (IP) was designed to provide best-effort service for delivery of data packets and to run across virtually any network transmission media and system platform. The increasing popularity of IP has shifted the paradigm from "IP over everything," to "everything over IP." In order to manage the multitude of applications such as streaming video, Voice over IP (VoIP), and others, a network requires Quality of Service (QoS) in addition to best-effort service. Different applications have varying needs for delay, delay variation (jitter), bandwidth, packet loss, and availability. These parameters form the basis of QoS. The IP network should be designed to provide the requisite QoS to applications. For example, VoIP requires very low jitter, a one-way delay in the order of 150 milliseconds and guaranteed bandwidth in the range of 8Kbps -> 64Kbps, dependent on the codec used.

When packets are classified at the edge of the network, specific forwarding treatments, formally called Per-Hop Behaviour (PHB), are applied on each network element, providing the packet the appropriate delay-bound, jitter-bound, bandwidth, etc. This combination of packet marking and well-defined PHBs results in a scalable QoS solution for any given packet, and any application.

The IETF defined models, IntServ and DiffServ, are two ways of considering the fundamental problem of providing QoS for a given IP packet.

5.5.3.2 Integrated Services architecture (IntServ)

The integrated Services (IntServ) model relies on the Resource Reservation Protocol (RSVP) to signal and reserve the desired QoS for each flow in the network. A flow is defined as an individual, unidirectional data stream between two applications. Two types of service can be requested via RSVP (assuming all network devices support RSVP along the path from the source to the destination). The first type is a very strict guaranteed service that provides for firm bounds on end-to-end delay and assured bandwidth for traffic that conforms to the reserved specifications. The second type is a controlled load service that provides for a better than best effort and low delay service under light to moderate network loads. It is possible (at least theoretically) to provide the requisite QoS for every flow in the network, provided it is signalled using RSVP and the resources are available.

However, there are several drawbacks to this approach:

- Every device along the path of a packet, including the end systems such as servers and PCs, needs to be fully aware of RSVP and capable of signalling the required QoS.

Copyright © ENVISION Consortium, May 2011

- Reservations in each device along the path are "soft," which means they need to be refreshed periodically, thereby adding to the traffic on the network and increasing the chance that the reservation may time out if refresh packets are lost.
- Maintaining soft-states in each router, combined with admission control at each hop and increased memory requirements to support a large number of reservations, adds to the complexity of each network node along the path.
- Since state information for each reservation needs to be maintained at every router along the path, scalability with hundreds of thousands of flows through a network core becomes an issue.

5.5.3.3 Differentiated Services architecture (DiffServ)

The Differentiated Services (DiffServ) approach to provide quality of service in networks employs a small, well-defined set of building blocks from which a variety of aggregate behaviours may be built. A small bit-pattern in each packet, in the IPv4 ToS (type of service) octet or the IPv6 traffic class octet, is used to mark a packet to receive a particular forwarding treatment, or per-hop behaviour, at each network node. A common understanding about the use and interpretation of this bit-pattern is required for inter-domain use, multi-vendor interoperability, and consistent reasoning about expected aggregate behaviour in a network. The typical architecture of a DiffServ node is shown in Figure 45.

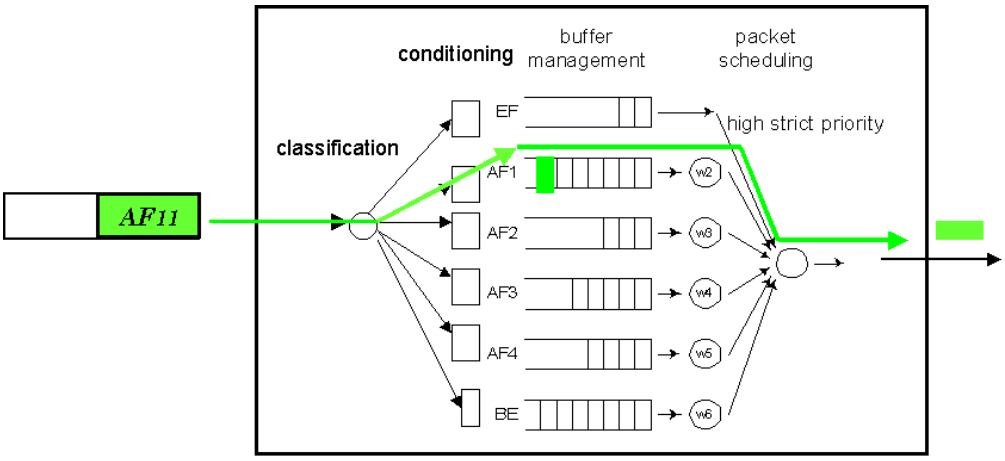

Figure 45: Typical architecture for DiffServ nodes [MH 01]

In order to deliver end-to-end QoS, DiffServ architecture has several major components that are packet marking using the IPv4 ToS byte and behaviours (per hop, per domain and end-to-end).

- **Packet Marking:** Unlike the IP-precedence solution, the ToS byte is completely redefined. Six bits are now used to classify packets. The field is now called the Differentiated Services (DS) field, with two of the bits unused (RFC-2474). The six bits replace the three IP-precedence bits, and is called the Differentiated Services Codepoint (DSCP). With DSCP, in any given node, up to 64 different aggregates/classes can be supported. All classification and QoS revolves around the DSCP in the DiffServ model.
- **Per Hop Behaviours (PHB):** Now that packets can be marked using the DSCP, how do we provide the QoS that is needed? First, the collection of packets that have the same DSCP value (also called a codepoint) in them, and crossing in a particular direction is called a **Behaviour** Aggregate (BA). Packets from multiple applications/sources could belong to the same BA. The PHB refers to the packet scheduling, queuing, policing, or shaping behaviour of a node on any given packet belonging to a BA, and as configured by a Service Level Agreement (SLA) or policy.

Copyright © ENVISION Consortium, May 2011

- **Per Domain Behaviour (PDB):** A PDB consists of computable attributes that define the treatment that each particular BA may experience from edge-to-edge in a particular DiffServ domain. For example the PDB may specify the edge-to-edge delay that the traffic belonging to Assured Forwarding (AF) class may experience in the domain. PDB is constructed based on aggregating a set of PHB from ingress to egress nodes. The attributes that can be part of the PDB are like delay, packet loss and throughput. For the measurement of these attributes, network specific parameters need to be specified.
- **End-to-End Behaviour (E2EB):** An E2EB is a measure of an attribute that defines the treatment of each particular BA from end-to-end. It consists of the sum of the same attribute within all the traversed domains (the sum of the same PDBs).

5.5.4 The ENVISION approach

The network aware multilink concept through the collaboration via the CINA interface stands within the focus of this section, while several communication paths toward a receiving device may cross the ISP network through different entry points/edges, there are yet not enough network transport services to bundle the different flows to provide aggregated services. Aggregated network services refer to network policies which are employed for a group of IP flows which may also be arriving from distinct access networks. The concept of a group of IP flows may be extended in various ways, including the cases where IP flows belong to a single device or multiple devices, IP flows of single overlay application or multiple and so on.

There are many technological directions in which novel techniques for a Group of IP Flows (GOIF) could be based upon techniques like QoS and priority handling. Through the collaboration with the overlay application, the network could discover and provide aggregate services to a GOIF. This approach may be beneficial to both sides: the network may guarantee a certain aggregate of service level agreements to a peer's GOIFs, while signalling to the overlay the costs and/or preferences associated with accessing destinations from each of the multilink-enabled node's interfaces. In this way the ISP can influence the overlay nodes to select ingress points of traffic from multilink nodes according to the destination so as to minimise cost, congestion, delay or to maximise throughput while satisfying the aggregate agreements in a flexible manner.

To achieve this the following network services are identified:

- **Link and flow aggregation:** while aggregating the traffic over the different links, the network could employ QoS mechanisms to achieve certain SLA such as maximal information rate, committed information rate, etc.
- **Discovery of interfaces:** through the ENVISION protocol, the CINA interface could provide the overlay with preferences, costs and other metrics for destinations reachable through each of the multilink nodes interfaces. In such cases, the network provider may balance the traffic through different access networks in its domain provided that the overlay nodes comply with the announced network information and cost metrics.
- **IP flow Priority:** the network could treat different IP flows with different priorities. By means of information exchanged via the CINA interface, the network could provide information to the overlay regarding the priority mechanism a flow may receive when established through a specific interface/access network. In this way, an ISP could reward the overlay nodes for using less congested access networks and associated core network resources by providing higher service levels on lower cost/less congested access networks. Moreover, the overlay may send more important data through interfaces with higher priority and the overlay may also exploit prioritised network level treatment to send traffic to higher priority peers as selected by the overlay optimisation algorithms: e.g. peers that are high in the hierarchy of a tree based swarm or to other critical nodes with a high outdegree of subordinate peers.

Copyright © ENVISION Consortium, May 2011

In detailed specification and design work we will study how the above mentioned services could be integrated with the overlay algorithms being designed for live and interactive ENVISION applications. This optionally includes the high capacity node approach in section 5.2.7, implications on the discovery as defined in section 3.1 and the overlay resource optimisation for interactive and live services as defined in D4.1 sections 4 and 5 respectively.

While this section has concentrated on multilink aggregate network services provided through multiple access networks to the same ISP a topic of further study is on mechanisms for aggregating services across multiple providers. One option would be through an intermediary network aggregation mediator that has access to resources from several providers. The multilink enabled nodes would access the resources of the aggregation mediator rather than those of the ISP/access provider directly, possibly through established business models for Mobile Virtual Network Operators (MVNOs). The mediation layer would provide preferences, costs and load balancing information to multilink nodes in the overlay, through its own CINA interface, according to the usage of the resources it has contracted with the underlying ISPs.

Copyright © ENVISION Consortium, May 2011

5.6 Network optimisation logic based on preferences: Preferences announcement

5.6.1 ISP preferences description

To allow the co-optimisation between the ISP and the application, the CINA interface should allow the ISPs to express their preferences in a way that can be used by the applications to assess the impact on their performance, while at the same time allowing the ISP to conceal sensitive information.

Although it is clear that the semantics of the preferences expressed by the ISP need to be commonly understood by the ISP and the application, in this deliverable, we do not make any assumptions regarding how explicitly the ISP preferences need to reflect its policies. Specifically, we consider the following set of preference expressions that can be communicated to the application:

- Overlay node ranking: The ISP may produce a sorted list of overlay nodes, ranked according to its preferences for creating connections from or to any or specific nodes in its domain.
- Overlay node weights: The ISP may produce a list of overlay nodes, each associated with a vector of weights, set according to the ISP preferences for creating connections from or to nodes in its domain. Note that in the case of a vector of weights, each dimension may represent a different criterion and the semantics need to be clearly communicated to the application.
- Network performance predictions: In addition to pure ISP preferences, the ISP may be revealing information regarding objective network metrics, like delay and loss. This information may be limited to historical observations, or it may be enhanced with predictions, taking into account particular routing and forwarding policies that the ISP has in place.
- ISP policies: The ISP may specify a set of explicit policy rules, capturing explicitly how overlay nodes or particular traffic generated by overlay nodes is preferred over other, or how it may be routed or forwarded differently by the network.

Please note that at this point, we make no assumptions regarding the granularity and the completeness of the information exposed by the ISP. One could think of a scenario where the application provides the ISP with a set of nodes or queries the ISP for a given service or stream and the ISP sends a list of possible nodes tagged with extra information on its preferences. In another scenario, the ISP might provide information for a set of local and remove IP prefixes without requiring any input from the application.

While the above list describes the possible preference expressions, it is still open to define the different factors that may shape these preferences. These may include but are not restricted to:

- ISP cost model: In some cases possibly the most important criterion for an ISP is the cost incurred by customer-provider agreements with other ISPs, for access and for transit inter-domain traffic.
- Network optimisation objectives: To facilitate operations, minimise congestion and recover from failures, the traffic engineering functions within an ISP may have optimum operation points which are translated to particular preferences for the formulation of the traffic matrices.
- Customer satisfaction objectives: An ISP may receive different value from different types of customers or particular SLAs, or it may want to differentiate the traffic treatment depending on the QoS requirements of the corresponding applications, e.g. avoiding long paths for delay-sensitive applications, heavily policing greedy applications in times of congestion etc.

The formulation of an optimum way of setting preferences to be announced to the applications is considered to be specific to each ISP and is further discussed in the following section.

Copyright © ENVISION Consortium, May 2011

5.6.2 Preference announcement optimisation

If one considers the traffic matrix of greatest benefit to a given content distribution overlay, it is clear that it will depend on its preferences regarding cost, QoS, resource availability, data caching and replication. On the other hand, if one considers the traffic matrix of greatest benefit to the ISP providing the overlay with network connectivity, it will depend on the infrastructure and transmission costs of the ISP, the background traffic that it carries and its traffic engineering policies. Thus, a tension arises between the preferences of the overlay and the ISP [ABEA+06], which can lead to complex interactions and potentially unpredictable overlay behaviour. As content distribution overlays can have a significant impact on IP traffic matrices in very short timescales, this can make traffic engineering difficult [DJ09]. In addition, traffic demand unpredictability can lead to suboptimal infrastructure purchase decisions, and thus to lower ISP profitability.

To resolve this tension, ENVISION allows application overlays and their underlying ISPs to exchange network and policy-related information. The overlay then can use this ISP-provided information to optimise its own traffic distribution. By strategically revealing this information, it may be possible for ISPs to have an effect on the behaviour of application overlays.

Starting from a set of basic assumptions regarding the preferences of overlay applications, cost models and network optimisation objectives for the underlying ISPs, WP3 will study the performance or financial gains achieved by the strategic revealing of preferences by the ISP.

Game-theoretical solutions have long been considered for problems that entail independent set of actions and objectives. Mathematically, a game is a model of the interaction of autonomous agents called players. At any given point of the game, each player has a set of actions available to it. However, the benefit that a player can eventually obtain from any given action is dependent on the actions of the other players of the game. The game eventually ends with an outcome that is a result of the decisions of the players, which can then rank, according to their preferences, all the possible outcomes that can result form all possible combinations of player actions.

To explore the problem of setting preferences to influence the application behaviour, we will focus at first on a simple situation with a single application overlay deployed over a single ISP. In this case, these two entities constitute the players of the game. Each player has at its disposal a distinct set of actions. In the case of the ISP, these actions will include routing and traffic management policies; in the case of the overlay, these actions include end-to-end traffic allocations and the invocation of network services. We assume that each application will attempt to maximise the utility it gets from the connectivity provided by its underlying ISPs as a function of its own preferences, the network quality it experiences, and the prices of the offered network services. Each ISP, on the other hand, will attempt to maximise its profit as a function of the aggregated traffic matrix of all the overlays it serves and its infrastructure costs.

By considering a convenient solution concept over this strategy space (such as Nash or dominant strategy equilibria) it is possible to find player strategies that are robust under strategic interaction. This formulation, however, is very computationally demanding. Even by using advanced algorithms, the identification of explicit Nash equilibria may be prohibitively expensive [NRTV07]. Thus, simplifications may be considered.

Our research work has recently been started on this topic and the next deliverable will present the models and results.

Copyright © ENVISION Consortium, May 2011

6. CONCLUSIONS

This deliverable presents the first results of work achieved within WP3 during the first year of the ENVISION project, including a survey of related work and first proposals on the ENVISION solutions.

A first version of the network-level functional architecture has been designed but is subject to refinements after more investigations of interactions between components and further design of the specific algorithms and protocols being developed.

The design and specification of the CINA interface, which is one of the major results expected by ENVISION, has begun and a survey of existing work has resulted in an understanding of the requirements, the missing aspects of current solutions for interactions between ISPs and overlay applications and an outline of the design of the interface. The studies on network monitoring and network metrics part as well as those defining the network services, provide a good basis for the detailed interface specification. An initial design of discovery mechanisms for finding the ENVISION servers and associated CINA interfaces has been initiated and presented in this deliverable. This has been proposed and is currently being discussed in the IETF ALTO working group.

The network monitoring is still under discussion, mainly the monitoring network architecture, the metrics to measure and the ways to collect the values. Analysis of work done by the IETF IPPM working group has been made in order to identify possible definition of metrics that could be of interest for ENVISION.

Finally, regarding the network services to be supported by ISPs and made available to overlay applications through the CINA interface, four main network services have been identified that aim at optimising the network while improving the applications' QoE. The network services are: multicast, caching, content adaptation and QoS. A good survey of existing work has been described in this deliverable and first proposals of improvements and new research work have been presented.

Copyright © ENVISION Consortium, May 2011

7. REFERENCES

[3GPP05] 3GPP: TSG Core Network and Terminals WG5. Technical report, http://www.3gpp.org/TB/CT/CT5/CT5.htm (2005)

[3GPP06] Technical Specification Group Services and System Aspects (2006), IP Multimedia Subsystem (IMS), Stage 2, TS 23.228, 3rd Generation Partnership Project

[AA09] Abbasi, U., Ahmed, T.: COOCHING: cooperative prefetching strategy for P2P video- on-demand system. In: Lecture Notes in Computer Science; Wired-Wireless Multimedia Networks and Services Management, vol. 5842, pp. 195–200. Springer, Berlin (2009).

[ABEA+06E] Altman, T. Boulogne, R. El-Azouzi, T. Jiménez, and L. Wynter. A survey on networking games in telecommunications. Comput. Oper. Res., 33(2):286-311, 2006.

[Allot Communications 2007] Digging Deeper into Deep Packet Inspection (DPI). White Paper, Allot Communication 2007. https://www.dpacket.org/articles/digging-deeper-deep-packet-inspection-dpi

[ALM01] Challenges of Integrating ASM and SSM IP Multicast Protocol Architecture, K. C. Almeroth, S. Bhattacharyya, C. Diot, International Workshop on Digital Communications: Evolutionary Trends of the Internet, 2001

[ALTO3p] Kiesel et. Al, ALTO Server Discovery Protocol, draft-kiesel-alto-3pdisc

[ALTO10] https://datatracker.ietf.org/wg/alto/

[BranchCache] http://technet.microsoft.com/en-us/library/dd637832%28WS.10%29.aspx

[BRE99] Lee Breslau, Pei Cao, Li Fan, Graham Phillips, and Scott Shenker, "Web caching and zipf-like distributions: Evidence and implications", In INFOCOM, pages 126-134, 1999.

[CAR10] Y. Carlinet, L. Mé, Y. Gourhant, H. Debar, "Caching P2P Traffic: What are the Benefits for an ISP?", ICN (International Conference on Networks) 2010.

[CASdraft06] Access Right Distribution Protocol (ARDP), A. Cassen, IETF draft-cassen-access-right-distribution-protocol-06, 2009

[Castro03] M. Castro, P. Druschel, A-M. Kermarrec, A. Nandi, A. Rowstron and A. Singh, "SplitStream: High-bandwidth multicast in a cooperative environment", SOSP'03,Lake Bolton, New York, October, 2003.

[Coblitz06] KyoungSoo Park and Vivek S. Pai: Scale and Performance in the CoBlitz Large-File Distribution Service. http://nsg.cs.princeton.edu/publication/coblitz_nsdi_06.pdf

[COMET] http://www.comet-project.org/

[DIO00] Deployment Issues for the IP Multicast Service and Architecture. C. Diot, B. N. Levine, B. Lyles, H. Kassem, D. Balensiefen, IEEE Network, 2000, Vol. 14, No. 1.

[DJ09] D. DiPalantino and R. Johari. Traffic engineering versus content distribution: A game theoretic perspective. In Proc. of INFOCOM, 2009.

[ETSI04] ETSI TISPAN: Webpage. Technical report, http://portal.etsi.org/tispan/ (2004)

[FIN00] An Abstract API for Multicast Address Allocation, R. Finlayson, IETF RFC 2771, 2000.

[FLA92] Flajolet, Gardy, and Thimonier, "Birthday paradox, coupon collectors, caching algorithms and self-organizing search", journal of Discrete Appl. Math, vol. 39, pp. 207-229, Elsevier 1992.

Copyright © ENVISION Consortium, May 2011

[GEO10] Geographic Location/Privacy Working Group, Charter, http://datatracker.ietf.org/wg/geopriv/charter/

[GOLD04] Golding, P.: Next Generation Wireless Applications. Wiley, John & Sons, Incorporated (May 2004)

[HAN99] Multicast Address Dynamic Client Allocation Protocol (MADCAP), S. Hanna, B. Patel, M. Shah, IETF RFC 2730, 1999.

[HAYdraft10] Requirements for Multicast AAA coordinated between Content Provider(s) and Network Service Provider(s), T. Hayashi, H. Satou, H. Ohta, H.He, S. Vaidya, IETF draft-ietf-mboned-maccnt-req-10, 2010.

[HOL99] IP Multicast Channels: EXPRESS Support for Large-scale Single-source Applications, H. W. Holbrook, D. R. Cheriton, SIGCOMM 1999.

[IDChKr10] Cheshire, S. and M. Krochmal, "Multicast DNS", draft-cheshire-dnsext-multicastdns-11 (work in progress), March 2010.

[IDGoCaLe99] Goland, Y., Cai, T., Leach, P., Gu, Y., and S. Albright, "Simple Service Discovery Protocol/1.0 Operating without an Arbiter", October 1999, <draft-cai-ssdp-v1-03>.

[IMS06] Miikka Poikselka, Aki Niemi, Hisham Khartabil, Georg Mayer, "The IMS: IP Multimedia Concepts and Services", John Wiley & Sons, 2006

[ISL06a] A Framework to Add AAA Functionalities in IP Multicast. S. Islam, J.W. Atwood, AICT/ICIW, 2006.

[ISL06b] The Internet Group Management Protocol with Access Control (IGMP-AC), S. Islam and J. W. Atwood, IEEE Conference on Local Computer Networks, 2006.

[JAIN03] Jain, R., Anjum, F., Bakker, J.L.: Programming Converged Networks: Call Control in JTAPI, JAIN, and Parlay/OSA. Wiley, John & Sons, Incorporated (November 2003)

[JEL99] P. R. Jelenkovic, "Asymptotic Approximation of the Move-To-Front Search Cost Distribution and Least-Recently-Used Caching Fault Probabilities", Annals of Applied Probability, Vol. 9, No. 2, pp. 430-464, 1999.

[JunLeiXi] Jun Lei, Lei Shi and Xiaoming Fu, “An Experimental Analysis of Joost Peer-to-Peer VoD Service”, Computer Networks Group, University of Göttingen, Germany

[MH 01] A. Markopoulou, and S. Han, “Transmitting scalable video over a DiffServ network,” Final Project, Stanford Univ., 2001.

[MOER03] Moerdijk, A.J., Klostermann, L.: Opening the Networks with Parlay/OSA: Standards and Aspects Behind the APIs. IEEE Network (2003) 58–64

[NAPA10] Luca Abeni, Arpad Bakay, Marco Biazzini, Robert Birke, Emilio Leonardi, Renato Lo Cigno, Csaba Kiraly, Marco Mellia, Saverio Niccolini, Jan Seedorf, Tivadar Szemethy, Giuseppe Tropea, Network Friendly P2P-TV: The Napa-Wine Approach, IEEE P2P 2010, Delft, August 2010

[NRTV07] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani. Algorithmic Game Theory. Cambridge University Press, New York, NY, USA, 2007.

[ONEA10] http://www.gsmworld.com/oneapi

[OVERSI07] http://www.oversi.com/images/stories/white_paper_july.pdf

[PAN08] A. Panagakis, A. Vaios, I. Stavrakakis, "Approximate analysis of LRU in the case of short term correlations", Elsevier Computer Networks journal, issue 52 (2008), pp. 1142-1152.

[PARL02] The Parlay Group: Parlay APIs 4.0 and Parlay X Web Services. Whitepaper (2002)

Copyright © ENVISION Consortium, May 2011

[PCache] SFU, “Modelling and Caching of P2P Traffic”, 2008, http://nsl.cs.sfu.ca/wiki/index.php/Modeling_and_Caching_of_P2P_Traffic

[PeerAp09] http://www.peerapp.com/App_FCK/file/Accelerating%20the%20Video%20Internet%20PeerApp%20October%202009.pdf

[RAT01] Sylvia Ratnasamy, Mark Handley, Richard Karp, and Scott Shenker. Application-level Multicast using Content-Addressable Networks. In Proceedings of NGC, 2001.

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

[RFC2165] Veizades, J., Guttman, E., Perkins, C., and S. Kaplan, "Service Location Protocol", RFC 2165, June 1997.

[RFC2782] Gulbrandsen, A., Vixie, P., and L. Esibov, "A DNS RR for specifying the location of services (DNS SRV)", RFC 2782, February 2000.

[RFC2915] Mealling, M. and R. Daniel, "The Naming Authority Pointer (NAPTR) DNS Resource Record", RFC 2915, September 2000.

[RFC3958] Daigle, L. and A. Newton, "Domain-Based Application Service Location Using SRV RRs and the Dynamic Delegation Discovery Service (DDDS)", RFC 3958, January 2005.

[RFC4795] Aboba, B., Thaler, D., and L. Esibov, "Link-local Multicast Name Resolution (LLMNR)", RFC 4795, January 2007.

[Riverbed] http://www.riverbed.com/us/products/steelhead_appliance/steelhead_appliance.php

[RFC4848] Daigle, L., "Domain-Based Application Service Location Using URIs and the Dynamic Delegation Discovery Service (DDDS)", RFC 4848, April 2007.

[SAL06] Saleh and Hefeeda, "Modelling and Caching of Peer-to-Peer Traffic", IEEE ICNP 2006

[SAT05] Authentication, Authorization and Accounting Framework for Multicast Content Delivery, H. Satou, H. Ohta, J. Nishikido, T. Hayashi, Asia-Pacific Conference on Communications, 2005

[SATdraft12]Admission Control Framework for Multicasting, H. Satou, H. Ohta, T. Hayashi, C. Jacquenet, H. He, IETF draft-ietf-mboned-multiaaa-framework-12, 2010.

[SATO07] Ambient Networks Deliverable: “System Design of SATO & ASI”, website http://www.ambientnetworks.org/Files/deliverables/D12-F.1_PU.pdf, December 2007.

[SAV06] Protocol Independent Multicast - Sparse Mode (PIM-SM) Multicast Routing Security Issues and Enhancements, P. Savola, R. Lehtonen, D. Mey, IETF RFC 4609, 2006.

[THA00] The Internet Multicast Address Allocation Architecture, D. Thaler, M. Handley, D. Estrin, IETF RFC 2908, 2000.

[THAdraft10] Automatic IP Multicast Without Explicit Tunnels (AMT), D. Thaler, M. Talwar, A. Aggarwal, L. Vicisano, T. Pusateri, IETF draft-ietf-mboned-auto-multicast-10, 2010.

[TURN02] Turner, K., Magill, E.H., Marples, D.J.: Communications Services: The Technology of Call Control. Wiley, John & Sons, Incorporated (January 2002)

[WSDD05] Beatty, J., "Web Services Dynamic Discovery (WS-Discovery)", April 2005, <http://specs.xmlsoap.org/ws/2005/04/discovery/ws-discovery.pdf>.

[ZAP01] Using SSM Proxies to Provide Efficient Multiple-Source Multicast Delivery, D. Zappala and A. Fabbri, IEEE GLOBECOM,2001

Copyright © ENVISION Consortium, May 2011