3.1 Publishable summary

Concept and Project Objectives

Proactive and dynamic QoS management, network intrusion detection and early detection of network congestion problems among other applications in the context of network management and engineering, can benefit from the existence of an accurate and scalable mechanism for online characterization of evolving network traffic patterns.

Nowadays, in a context of continuous exponential growth in network speed and number of devices interacting with each other through the network, current approaches for online network traffic characterization clearly lack scalability and accuracy. Therefore, a new generation of scalable mechanisms and techniques to characterize online network traffic are required. The ONTIC Project proposes to investigate, develop, test and validate such scalable mechanisms and techniques.

To this end, the ONTIC Project proposes to:

(Scientific Objective1) Investigate, implement and test a novel architecture of scalable mechanisms and techniques to be able to (a) characterize online network traffic data streams, identifying traffic pattern evolution, and (b) proactively detect anomalies in real time when hundreds of thousands of packets are processed per second.

(Scientific Objective 2) Investigate, implement and test a completely new set of scalable offline data mining mechanisms and techniques to characterize network traffic, applying a big data analytics approach and using distributed computation paradigms in the cloud on extremely large network traffic summary datasets consisting of trillions of records.

(Technical Objective3) Integrate offline and online mechanisms and techniques into an autonomous supervised or unsupervised network traffic characterization system to be used as cornerstone of a new generation of scalable and proactive network management and engineering applications.

Real testing will be performed by defining three paradigmatic ***Use Cases*** and implementing their corresponding prototypes. Prototypes will be validated with real data obtained from the core network of Interhost-SATEC (a partner of the ONTIC consortium). On average, 1,5Gb of network traffic data will be obtained each second and the capture process will go on for a period of 24 months. These network traffic packets generate 200,000 events per second, which during a whole year means around 10 trillion records; therefore, in order to analyze them, a big data solution is needed. The resulting dataset will be anonymized and made publicly available at the end of the project to foster new research initiatives in the big data analytics area. Final validation of mechanisms and techniques will be deployed to the EMC Greenplum Analytics WorkBench (AWB) and the Google Cloud platform. As a partner of the ONTIC consortium, EMC Spain will provide access to their experimental workbench laboratory for big data solutions. The AWB is a large-scale cluster with more than 1000 nodes and 50 petabytes of storage whose primary purpose is running regular scale validation on Apache Hadoop releases. To test the elastic scalability of the algorithms developed on more efficient platforms like Apache Spark, the Google Cloud Platform will be used.

(Technical Objective 4) Lead the dissemination and adoption of the ONTIC outcomes to other application domains where scalability, accuracy and in some cases real time response are a must. Examples of these domains are bioinformatics, genomic, medicine, physics, social sciences, finances, marketing, etc. To this purpose, ONTIC will generate an open source, highly scalable offline/online analytics framework to be used by developers in other application domains.

ONTIC context

Accurate identification and categorization of network traffic according to application type is an important element of many network management tasks related with QoS such as flow prioritization, traffic shaping/policing, and diagnostic monitoring. For example, a network operator may want to identify and throttle (or block) in traffic from peer-to-peer (P2P) file sharing applications to manage its bandwidth budget and to ensure good performance of business critical applications. Similarly to network management tasks, many network engineering problems such as workload characterization and modeling, capacity planning and route provisioning also benefit from accurate identification of network traffic. Moreover, the detection of network attacks is an important task for network operators in today's Internet. Botnets, Distributed Denial of Service attacks (DDoS) and network-scanning activities are examples of the different threats that compromise the integrity and normal operation of the network every day.

Adaptive/Dynamic QoS implementations (based on online detection of evolving traffic patterns), proactive congestion control mechanisms (based on early problem detection), and Network Intrusion Detection Systems (based on anomaly detection), among others, are applications in the context of network management and engineering that will benefit from the existence of an accurate and scalable mechanism for online identification and characterization of network traffic patterns.

Due to the continuous growth in network speed, terabytes of data may cross the core network of a typical ISP every day. Moreover, in a medium term, the exponential growth of the M2M/IoT scenario is expected to generate more than 50 billion connected devices, each of them producing huge amounts of network traffic. Thus, two major issues hamper network data analysis in the short and medium term: (a) a huge amount of data coming from a huge number of sources can be collected in a very short time, and (b) it is hard to identify correlations and detect anomalies in real-time on network traffic traces that large.

In this context of continuous growth in network speed and number of interconnected devices, current approaches for online network traffic characterization clearly lack scalability and accuracy. Therefore, several **challenges** appear when considering the development of successful solutions.

Firstly, an offline mechanism for traffic characterization should be developed to provide initial network traffic identification. Considering the amounts of summary information that can be generated and stored at a typical ISP per day, and that more information available usually helps achieve better results; two significant needs appear: **i)** a big data solution is required to store and access summarized information of an ISP during a significant period of time (e.g. a one-year summary requires a petabyte-sized database); and **ii)** a true scalable traffic characterization mechanism is needed to efficiently process this huge amount of data.

Secondly, a scalable online mechanism for traffic characterization should be provided to both identify traffic patterns evolution and to detect anomalies in real time, minimizing the false positives rate. Given the high amount of data per second a medium size ISP faces, a tradeoff between the expected accuracy, the computational complexity of the characterization mechanism, how and when to parallelize concrete subsystems, and the amount of data considered at each execution (i.e. temporal window size) have to be paid attention to. In this real time scenario, big data can provide a maximization of the overall system accuracy when identifying traffic patterns and their evolution over time.

Finally, it would be desirable to extrapolate mechanisms, algorithms and techniques applied to a specific domain area (i.e. network traffic) to other domains (bioinformatics, social sciences and statistics among others).

ONTIC is a Small or medium-scale research project (STReP) to fulfill the **Objective ICT-2013.4.2 Scalable data analytics of the EU's FP7 ICT Workprogram** by focusing on the development of tools and skills to deploy and manage robust and high performance data analytics processes over extremely large amounts of data. To this end, ONTIC will investigate, implement and test a novel architecture of scalable online and offline mechanisms and techniques to be able to characterize network traffic data streams. The goal of these mechanisms and techniques is a) to identify network traffic patterns evolution, and b) to proactively detect anomalies, both in real time and when hundreds of thousands of packets are processed per second.

The ONTIC consortium is composed by leading European Industrial, Corporate R&D and Academic partners which will focus their resources and vast know-how in this domain to address the previously mentioned challenges.

WP2 – Big Data Network Traffic Summary Dataset

During the first year of the project the following activities have been carried on:

- **ONTIC Big Data Architecture Design**: a generic and extensible architecture covering form data ingestion to online and offline processing on different platforms has been proposed and designed.
- **ONTS Dataset capture**: the ONTIC data provisioning subsystem has been designed, implemented, tested and deployed in its final location at SATEC. Up to now more than 200 TB of data have been captured as part of the ONTS dataset.
- **Local Big Data testbeds set-up:** in order to test and benchmark different distributed computational platforms and frameworks as well as to validate the algorithms developed so far in small and controlled environments, UPM, POLITO and ADAPTIT have deployed local big data clusters.
- **Hadoop and Spark platforms benchmark:** in the previously mentioned clusters Hadoop and Spark platforms have been extensively tested to clearly identify their strengths and weaknesses.
- **Research and implementation of unsupervised feature selection algorithms:** as a first step to almost any machine learning algorithm a feature selection step has to be performed to lower the dimension of the input data. This provides a number of advantages such as faster processing, reduced probability of overfitting, and in general better classification/clustering results. Two algorithms have been developed so far and currently are being tested.

WP3 - Scalable Offline Network Traffic Characterization System

- **State of the art**: during the first year, WP3 has focused mainly on an extensive study of the state of the art in machine learning and analytics algorithms for network traffic classification. This research has led to the identification of suitable and promising approaches and technologies to be enhanced and applied to the ONTIC use cases. This state of the art compilation and analysis is available in deliverable D3.1.

- **Scientific dissemination**: in the context of this work package, POLITO has produced five scientific publications related to the goals of ONTIC.
- **Big data lab**: POLITO has set up a parallel computing laboratory which has been put at the disposal of project partners.

WP4 - Scalable Online Network Traffic Characterization System

In the context of WP4, the following goals have been attained.

- **Online architecture**: a generic architecture for online network traffic classification has been developed, suitable for both pattern evolution analysis and anomaly detection.
- **Traffic pattern evolution**: an extensive review of the state of the art in this field has been compiled, focusing on the limitations of existing proposals and their applicability to network traffic. This study has revealed a lot of ground to be covered in this domain as well as the advantages of having such a vast data set as the ONTS, collected by the ONTIC project.
- **Network anomaly and intrusion detection subsystem**: an unsupervised learning algorithm based on clustering has been proposed for anomaly and intrusion detection in network traffic. The first results of an evaluation of this method on traffic traces (both normal and honeynet) have been obtained.

Some of these achievements are described in detail in deliverable D4.1.

WP5 - Implementation, Integration and Verification

For the most part, WP5 activities have been dedicated to the refinement of the definition of the use cases, as well as the study of the way ONTIC can contribute to tackle related issues.

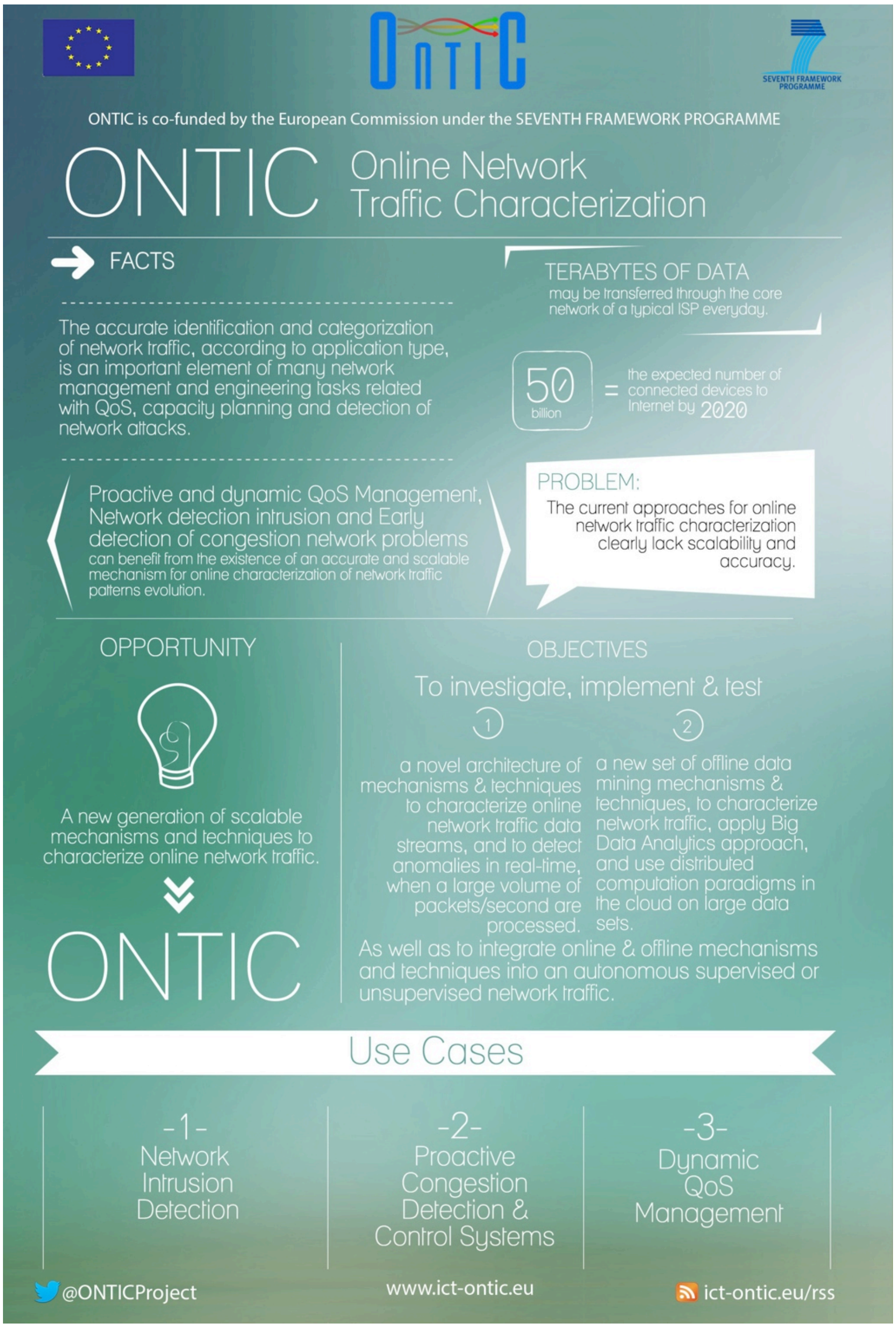
- The use cases have been thoroughly refined, including inputs received from potential stakeholders and the potential market. The expected final results and their potential impact and uses have been determined (including the socio-economic impact and the wider societal implications of the project so far).
- An extensive analysis of the state of the art on techniques for network anomaly detection, congestion control and quality of service policy enforcement has been performed. The shortcomings of existing techniques have been identified and the role of machine learning and big data analysis techniques has been outlined.
- The use cases have been adapted to the user story format of the Agile Scrum methodology.
- Big Data Research and Development activities have focused on the Big Data platform and dataset extraction task, the development of a first Policy Governance Function mock-up, etc. The outputs generated by the R&D activities carried out throughout the first year will be used as a basis to implement the provided User Stories on following years.
- The Quality of Experience concept, including QoS and congestion control, has been developed in several workshops, speeches, conferences, etc.
- Deliverable D5.1 provides a general overview of the ONTIC scenarios and related user stories. In addition, this deliverable offers an introduction to the proposed architecture, a summary of the related SoTA and a first version of the working backlogs.
- The Analytics Virtuous Circle has been defined, enhancing Network optimization in any type of network (IT, Telecom).

WP6 - Exploitation and Dissemination

The work of WP6 has been mostly aligned with the goals established by the related deliverables. All of these have been satisfactorily submitted according to schedule and are briefly described below.

- D6.1 - Public Project Presentation according to the EU specifications.
    - This deliverable featured the ONTIC logo and the project fact sheet.

Figure 1. The ONTIC logo



ONTIC is co-funded by the European Commission under the SEVENTH FRAMEWORK PROGRAMME

ONTIC Online Network Traffic Characterization

FACTS

The accurate identification and categorization of network traffic, according to application type, is an important element of many network management and engineering tasks related with QoS, capacity planning and detection of network attacks.

Proactive and dynamic QoS Management, Network detection intrusion and Early detection of congestion network problems can benefit from the existence of an accurate and scalable mechanism for online characterization of network traffic patterns evolution.

TERABYTES OF DATA may be transferred through the core network of a typical ISP everyday.

50 billion = the expected number of connected devices to Internet by 2020

PROBLEM: The current approaches for online network traffic characterization clearly lack scalability and accuracy.

OPPORTUNITY

A new generation of scalable mechanisms and techniques to characterize online network traffic.

ONTIC

OBJECTIVES

To investigate, implement & test

1 a novel architecture of mechanisms & techniques to characterize online network traffic data streams, and to detect anomalies in real-time, when a large volume of packets/second are processed.

2 a new set of offline data mining mechanisms & techniques, to characterize network traffic, apply Big Data Analytics approach, and use distributed computation paradigms in the cloud on large data sets.

As well as to integrate online & offline mechanisms and techniques into an autonomous supervised or unsupervised network traffic.

Use Cases

-1- Network Intrusion Detection

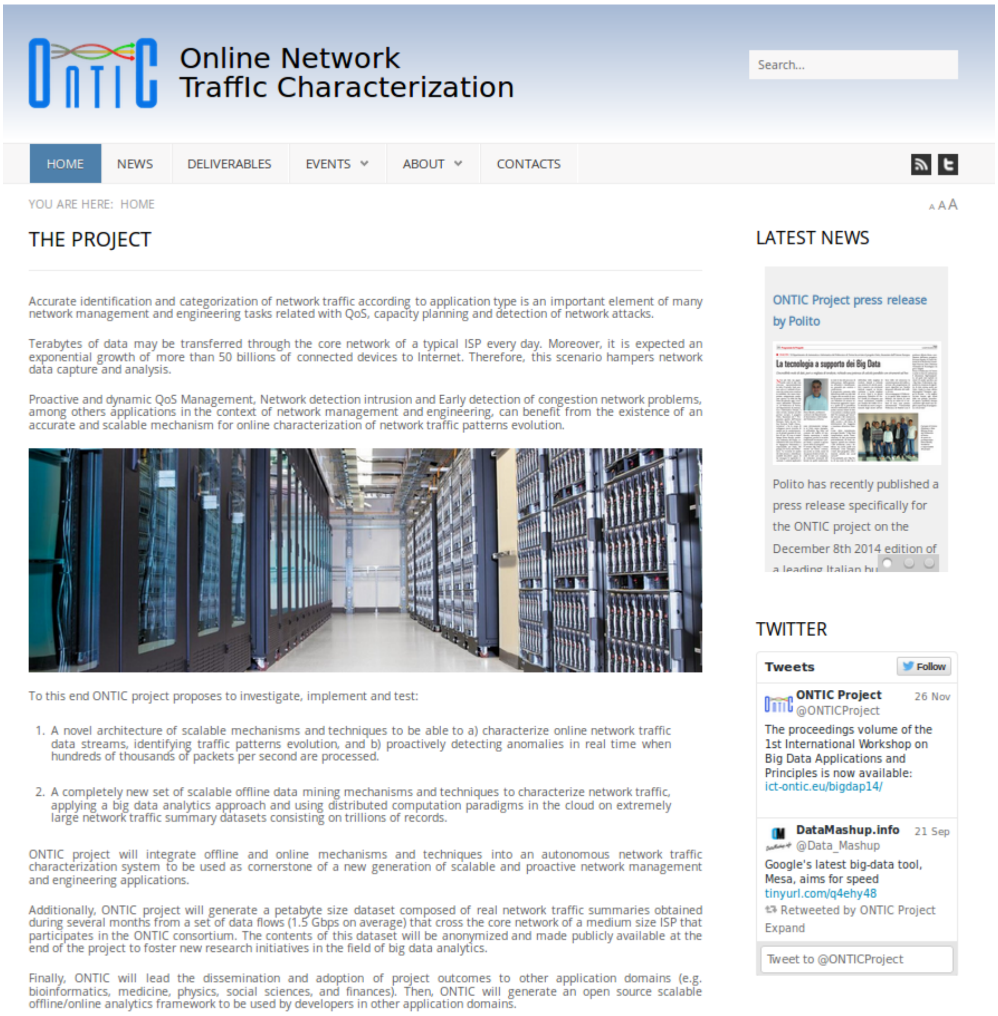
-2- Proactive Congestion Detection & Control Systems

-3- Dynamic QoS Management

@ONTICProject www.ict-ontic.eu ict-ontic.eu/rss

Figure 2. Project factsheet

- D6.2 - Project Website.
  - The website is up to date and fully operational, featuring information on the project, news, events, access to public deliverables and more. It is available at http://ict-ontic.eu/.



THE PROJECT

Accurate identification and categorization of network traffic according to application type is an important element of many network management and engineering tasks related with QoS, capacity planning and detection of network attacks.

Terabytes of data may be transferred through the core network of a typical ISP every day. Moreover, it is expected an exponential growth of more than 50 billions of connected devices to Internet. Therefore, this scenario hampers network data capture and analysis.

Proactive and dynamic QoS Management, Network detection intrusion and Early detection of congestion network problems, among others applications in the context of network management and engineering, can benefit from the existence of an accurate and scalable mechanism for online characterization of network traffic patterns evolution.

To this end ONTIC project proposes to investigate, implement and test:

1. A novel architecture of scalable mechanisms and techniques to be able to a) characterize online network traffic data streams, identifying traffic patterns evolution, and b) proactively detecting anomalies in real time when hundreds of thousands of packets per second are processed.
2. A completely new set of scalable offline data mining mechanisms and techniques to characterize network traffic, applying a big data analytics approach and using distributed computation paradigms in the cloud on extremely large network traffic summary datasets consisting on trillions of records.

ONTIC project will integrate offline and online mechanisms and techniques into an autonomous network traffic characterization system to be used as cornerstone of a new generation of scalable and proactive network management and engineering applications.

Additionally, ONTIC project will generate a petabyte size dataset composed of real network traffic summaries obtained during several months from a set of data flows (1.5 Gbps on average) that cross the core network of a medium size ISP that participates in the ONTIC consortium. The contents of this dataset will be anonymized and made publicly available at the end of the project to foster new research initiatives in the field of big data analytics.

Finally, ONTIC will lead the dissemination and adoption of project outcomes to other application domains (e.g. bioinformatics, medicine, physics, social sciences, and finances). Then, ONTIC will generate an open source scalable offline/online analytics framework to be used by developers in other application domains.

LATEST NEWS

ONTIC Project press release by Polito

La tecnologia a supporto dei Big Data

Polito has recently published a press release specifically for the ONTIC project on the December 8th 2014 edition of a leading Italian bu

TWITTER

ONTIC Project @ONTICProject 26 Nov
The proceedings volume of the 1st International Workshop on Big Data Applications and Principles is now available: ict-ontic.eu/bigdap14/

DataMashup.info @Data_Mashup 21 Sep
Google's latest big-data tool, Mesa, aims for speed tinyurl.com/q4ehy48
Retweeted by ONTIC Project

Figure 3. Website home page

- D6.3 - Exploitation Strategy
- D6.4 - Exploitation and Dissemination Plans
  - 12 screening talks
  - 4 Business ideas
  - 1 Patent application
  - 1 Co-operation/license agreement
  - Outreach to 29 new R&D partners
  - Outreach to 23 new industrial partners

- D6.8 - Big Data analytics Workshop #1
  - The First International Workshop on Big Data Applications and Principles (BIGDAP 2014) was held at the facilities of the UPM in September 2014 and featured:
    - 7 Papers (Proceedings Volume)
    - 3 invited talks.
    - 6 industrial talks
    - 76 attendees.