TIE

Traffic Identification Engine

Alberto Dainotti - *alberto@unina.it*
COMICS Research Group
University of Napoli "Federico II"

TRAFFIC CLASSIFICATION

To associate ***flows*** to the ***applications*** that generate them

{*UDP*, IP_{SRC}:*10.0.0.1*, $PORT_{SRC}$:*31215*, IP_{DST}:*212.48.72.19*, $PORT_{DST}$:*80*}

→ ***SKYPE!***

{*TCP*, IP_{SRC}:*10.0.0.1*, $PORT_{SRC}$:*2233*, IP_{DST}:*13.29.10.199*, $PORT_{DST}$:*25*}

→ ***SMTP!***

COMICS Research Group
University of Napoli "Federico II" - Italy

MOTIVATIONS

Why classify traffic?

- To ***understand*** what our links carry
  - -How are people using the Internet?
  - -What's the killer application?
  - -Does it really matter to model this or that?
  - -Is something "strange" happening and we don't know it?

- To ***operate*** networks
  - -Resource allocation and QoS
  - -Enforcement of security policies (e.g. Firewalling)
  - -Billing based on typology of traffic
  - -Network provisioning
  - -Diagnostics: retracing phenomena (e.g. congestion) to specific applications and protocols

APPROACHES

an evolving complex scenario

- **Port-based**
  - ✓Fast and Simple
  - -Unreliable (e.g. TCP:80 ≠ HTTP)

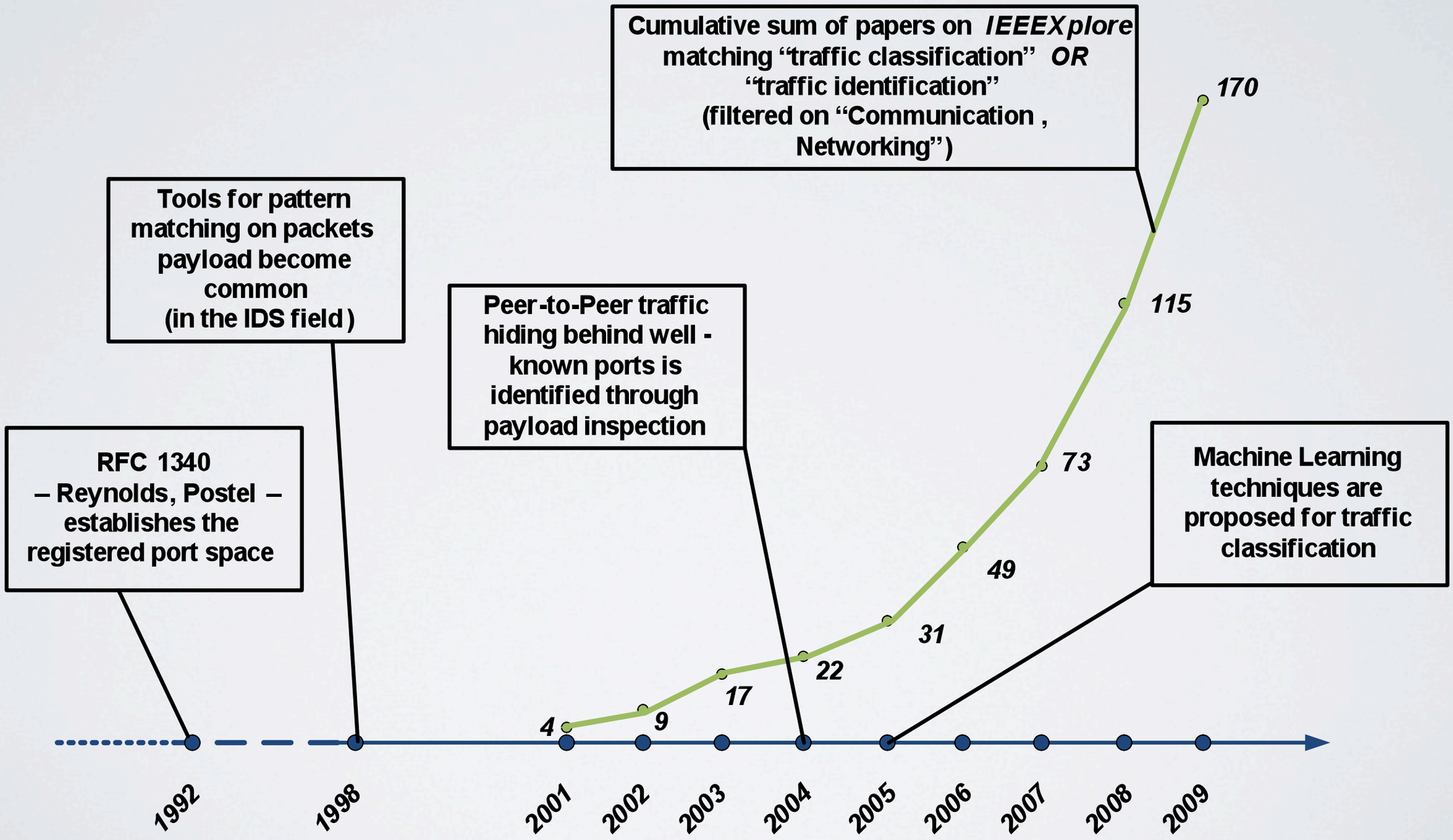
- **Payload inspection**
  - ✓Often reliable
  - -Privacy concerns
  - -Computationally heavy
  - -Can be tricked by protocol encapsulation, encryption, ...

- **Pattern Recognition & Behavioral**
  - ✓Promising with respect to current trends (encryption, obfuscation, novel applications, …)
  - -Experimental
  - -Reliable?

- *Mellia et al., "Traffic classification and its applications to modern networks", Elsevier Computer Networks, Dec. 2008*
- *Callado et al., "A survey on internet traffic identification", IEEE Communications Surveys & Tutorials, July 2009.*

SCIENCE EFFORTS

dramatically increased in past years



WHERE WE ARE

difficulties...

- A lot of work is still in experimental stage
- Scarce availability of real implementations
- Sharing traffic data in scientific community
- Lack of benchmarks
- Lack of standard formats

WHERE WE ARE

... and opportunities

- Large interest of different communities
  - -Scientists
  - -Providers
  - -Industry
  - -Society
- Several approaches and code proved to be effective
- Increasing complexity of Internet applications and traffic will continue to keep this topic *hot*!

TIE

Traffic Identification Engine

A software ***platform*** for ***building*** traffic classifiers and for ***experimenting*** with them

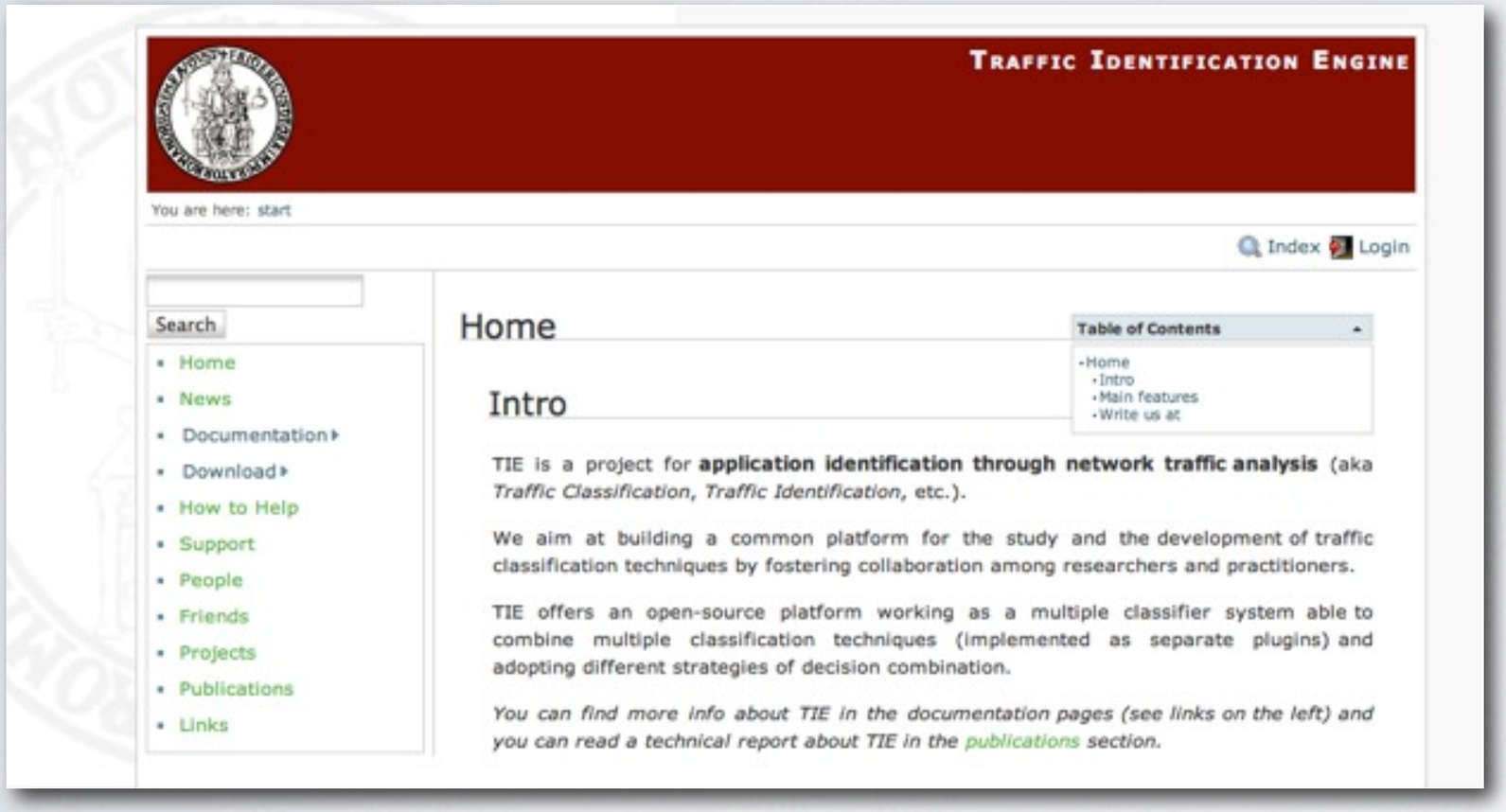
- Multi-approach Framework
- Open source
- Fast (*C language, Libpcap, Endace DAG support, ...*)
- Modular
- Supports *multiclassification*
- Supports *online* traffic classification
- ....

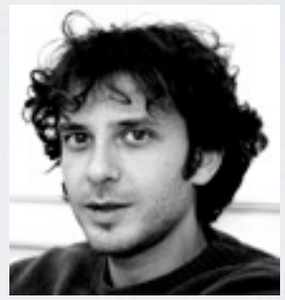
COMICS Research Group
University of Napoli "Federico II" - Italy

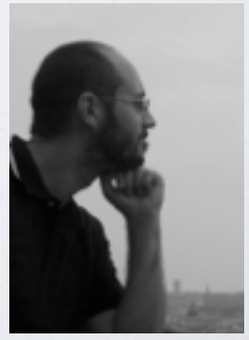
TIE HISTORY

the genesis

Started in **2007** by researchers of the "*TRAFFIC*" project inside COMICS







http://tie.comics.unina.it

COMICS Research Group
University of Napoli "Federico II" - Italy

TIE HISTORY

opening to the world

During these 4 years has been/is the subject of

- Graduate and undergraduate **students** theses
- Collaborations with other **research groups**

- Collaborations with the **Industry** (manufacturing, customer service assurance consultancy, ...)
- National and European **Research Projects**

TIE HISTORY

publications/inventions

Papers

- A. Dainotti, F. Gargiulo, L. Kuncheva, A. Pescapè, C. Sansone, Identification of traffic flows hiding behind TCP port 80, *IEEE ICC 2010* - May 2010, Capetown (South Africa)
- G. Aceto, A. Dainotti, W. de Donato, A. Pescapè, PortLoad: taking the best of two worlds in traffic classification, *IEEE INFOCOM 2010 - WIP Track* - March 2010, San Diego (CA, USA)
- V. Carela-Espanol, P. Barlet-Ros, M. Solé-Simò, A. Dainotti, W. de Donato, A. Pescapè, K-dimensional trees for continuous traffic classification, *International Workshop on Traffic Monitoring and Analysis (TMA'10) @ PAM 2010* - April 2010, Zurich (Switzerland)
- A. Dainotti, W. De Donato, A. Pescapè, "TIE: a Community-Oriented Traffic Classification Platform", International Workshop on Traffic Monitoring and Analysis (TMA'09) @ *IFIP Networking 2009* - May 2009, Aachen (Germany)
- Marco Mellia, Antonio Pescapè, Luca Salgarelli, "Traffic classification and its applications to modern networks", *Computer Networks*, Volume 53, Issue 6, 23 April 2009, Pages 759-760.
- A. Dainotti, W. De Donato, A. Pescapè, P. Salvo Rossi, "Classification of Network Traffic via Packet-Level Hidden Markov Models", *IEEE GLOBECOM 2008* - Dec 2008, New Orleans (LA, USA)

Book Chapters

- G. Aceto, A. Dainotti, W. de Donato, F. Gargiulo, A. Pescapè C. Sansone, "Combining Multiple Traffic Classification Techniques within a Single Platform", *RECIPE Robust and Efficient traffic Classification in IP nEtworks, Fridericiana Editrice Universitaria*, pp.1-16, ISBN: 978-88-833-8081-5, Napoli, Italy, 2009

Technical Reports

- A. Dainotti, W. de Donato, A. Pescape', Giorgio Ventre, "TIE: a community-oriented traffic classification platform", Technical Report TR-DIS-10-2008, Dipartimento di Informatica e Sistemistica, University of Napoli "Federico II", Italy tr-dis-10-2008-tie.pdf

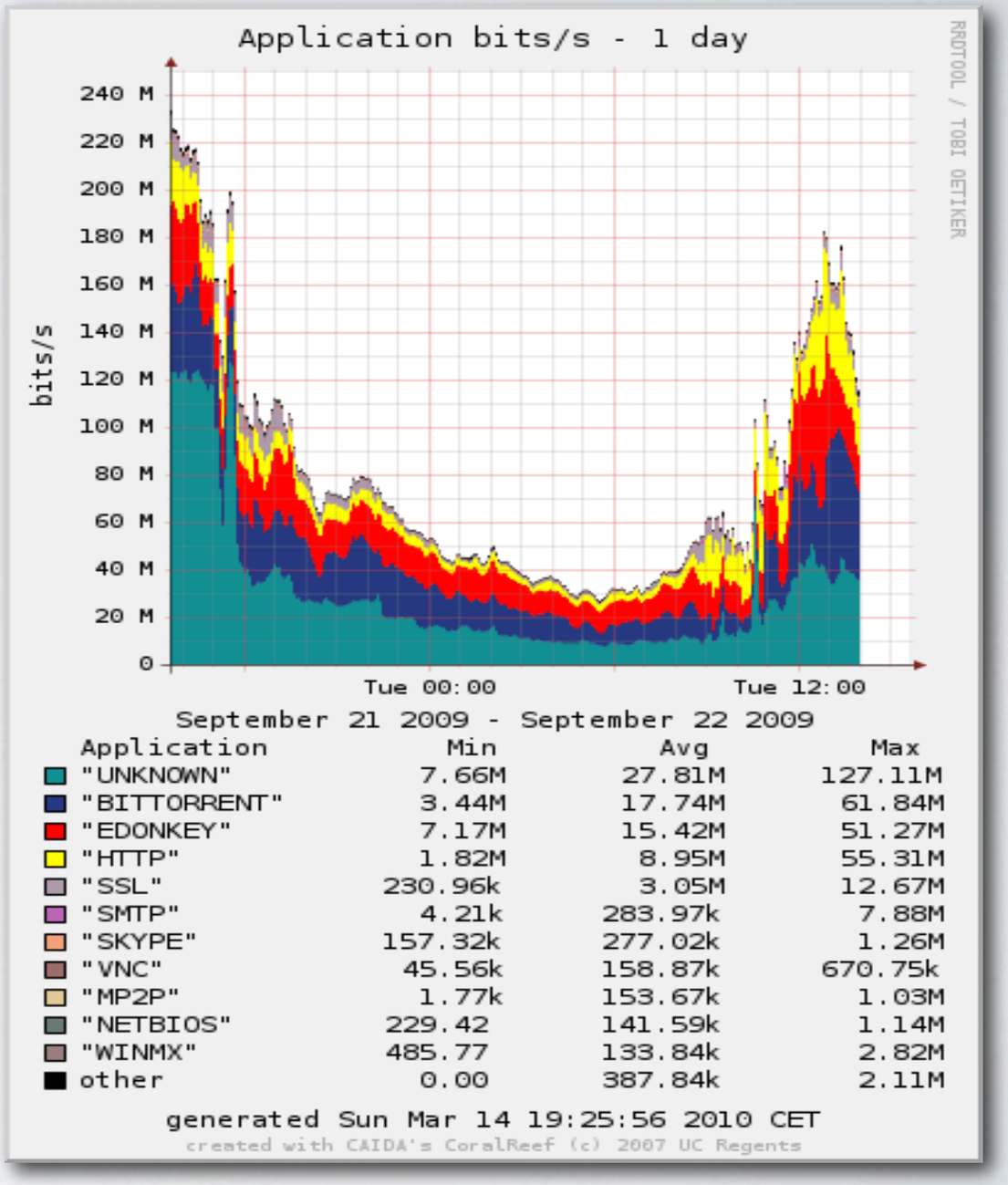
Patents

- A. Dainotti, G. Aceto, W. de Donato, A. Pescapè, "Method and system for traffic classification in communication networks using content-based signatures". 9th March 2010 - code NA2010A000011#

TIE OVERVIEW

operating modes

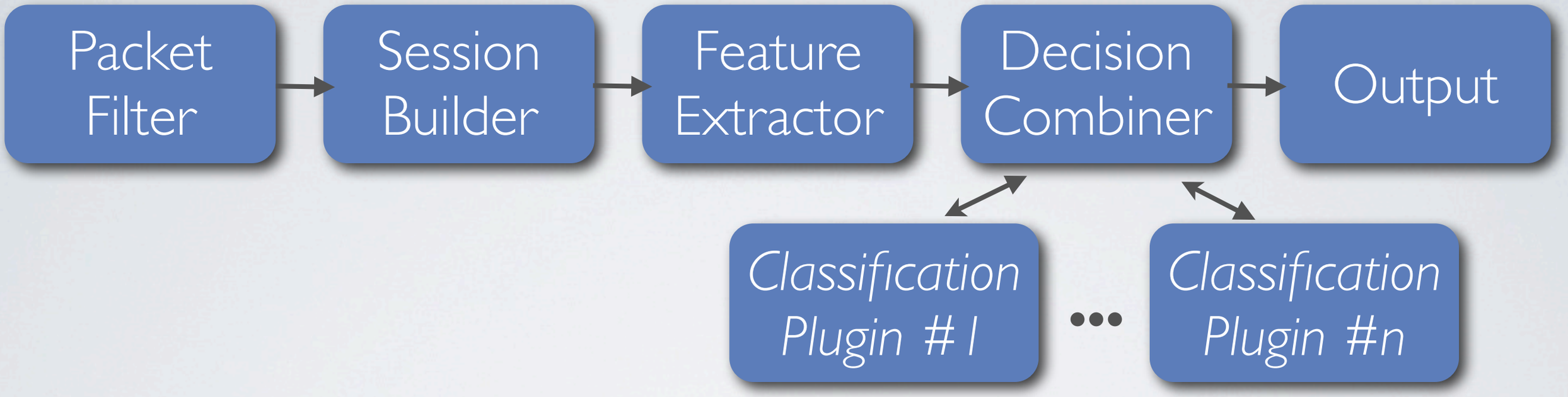
- **Offline**
  - a session is *classified* only when it ends or at the end of TIE execution
- **Realtime**
  - a session is classified as soon as possible and output is immediately available
- **Cyclic**
  -the classification of all live sessions is generated at regular intervals (e.g. each 5 min.)



In Cyclic mode, automated Web Reports can be built using CAIDA's CoralReef tools.

TIE ARCHITECTURE

from packet filtering to classified "flows"



- It can work with configurable definitions of *sessions*
  - **Flows**
    - *<L4Proto, IP$_{src}$, Port$_{src}$, IP$_{dst}$, Port$_{dst}$> + timeout*
  - **Biflows**
    - Same as above but *src* and *dst* are swappable
    - Support for TCP connections through simple heuristics based on TCP flags
  - **Hosts**
    - Under development

CLASSIFICATION PLUGINS

Name	Based on	Status	Contributor
Port	L4 ports	Available	UNINA(signatures from CAIDA)
L7	Deep Payload Inspection	Available	UNINA(signatures/code from Linux L7-filter)
PortLoad	Lightweight Payload Insp.	Licensable	UNINA
GMM-PS	Statistical Approach: PS	Under Test	UNINA
HMM	Statistical Approach: PS, IPT	Under Test	UNINA
FPT	Statistical Approach: PS, IPT	Under Dev	UNIBS
Joint	Machine Learning: PS, IPT	Under Test	UNINA-CENS
GT	Information from hosts	Under Dev	UNINA-UNIBS
OpenDPI	Deep Payload Inspection	Beta	OpenDPI, UNINA, TUM
WEKA	Imports the output of a WEKA classifier	Available	UNINA

OUTPUT

sample ASCII output

```
# tie output version: 1.0 (text format)
# generated by: ./tie -r traffic.pcap -S 2048

# Working Mode: off-line
# 1 plug-ins enabled: l7filter

# begin trace interval: 1222078328

# begin TIE Table
# id    src_ip          dst_ip          proto   sport   dport   dwpkts  uppkts  dwbytes upbytes t_start                t_last                  app_id  sub_id  confidence
844     143.225.229.169 89.96.63.82     6       33837   29867   1       1       4       15      1222078300.965969       1222078300.984039       0       0       0
843     143.225.229.169 213.140.17.96   6       33837   29014   1       1       4       14      1222078300.965951       1222078300.983139       0       0       0
225     143.225.229.169 87.5.180.250    17      33837   13604   1       1       1       1       1222078278.674796       1222078278.674796       163     0       100
503     143.225.229.169 151.8.66.210    6       33837   48781   2       2       2       2       1222078317.672792       1222078317.672792       0       0       0
589     143.225.229.169 87.3.228.234    17      33837   34930   1       1       1       1       1222078290.640406       1222078290.640406       163     0       100
661     143.225.229.169 85.34.207.10    6       33837   16508   1       1       1       1       1222078294.110945       1222078294.110945       0       0       0
134     143.225.229.169 96.20.21.108    17      33837   8056    1       1       1       1       1222078279.994987       1222078279.994987       163     0       100
327     143.225.229.169 74.72.218.29    17      33837   11788   1       1       1       1       1222078281.557751       1222078281.557751       163     0       100
```

- A set of utilities is distributed with TIE for the post-processing of the output
- In ***realtime*** mode, the output can also be sent through network sockets to another application

A CASE STUDY

*PortLoad**

- TIE's modular framework allows to easily **implement a new classification technique** and run it on real traffic

- By using a unified framework and standard definitions and formats it is easy to **compare and benchmark** three different classification techniques

G. Aceto, A. Dainotti, W. de Donato, A. Pescapè, PortLoad: taking the best of two worlds in traffic classification, IEEE INFOCOM 2010 - WIP Track - March 2010, San Diego (CA, USA)

Patent pending "Method and system for traffic classification in communication networks using content-based signatures". 9th March 2010 - code NA2010A000011#

COMICS Research Group
University of Napoli "Federico II" - Italy

PORTLOAD

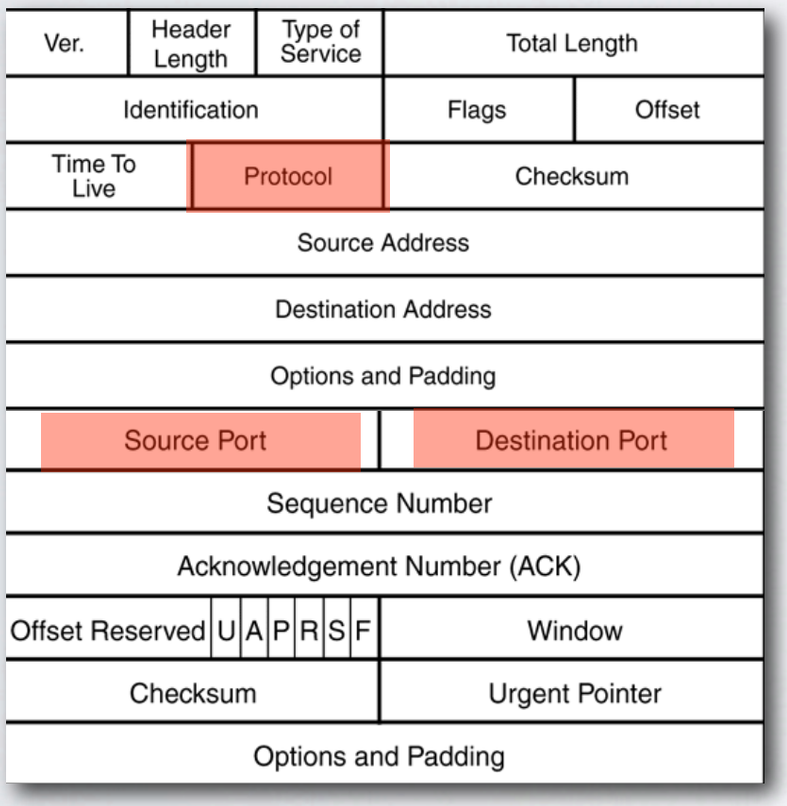
merging two "worlds" in traffic classification

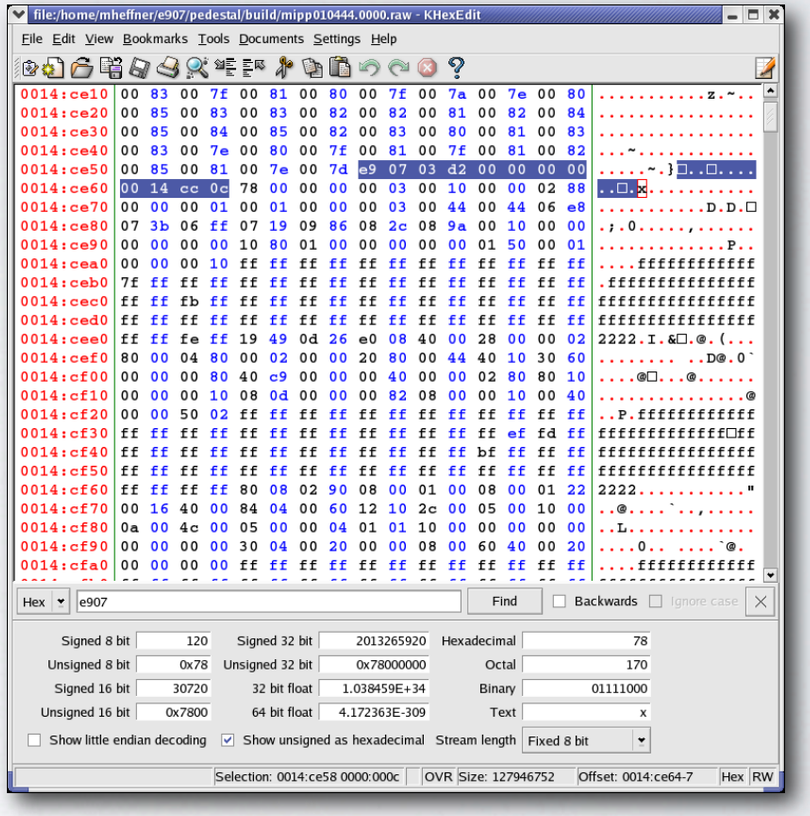
Port-based approach

- Very inaccurate
+ Simple & Fast
+ Privacy-friendly

Deep Packet Inspection

+ Accurate
- CPU intensive
- Doesn't care about Privacy





COMICS Research Group
University of Napoli "Federico II" - Italy

PORTLOAD

do we need all that payload?

- Experiments on sample traces with TIE-L7 (L7-Filter DPI based on regular expressions)
  - -Evaluated where the matches happen
  - -Packet position inside flow
  - -Bytes in payload

- E.g.

Site	Date	Size	Pkts	biflows
Univ. Napoli	Oct 3rd 2009	59 GB	80M	1M

  - -87% of the matches start at the first packet
  - -Almost all matching strings start (99.98%) and finish (90.77%) in the first 32 bytes of payload of a packet

PORTLOAD

taking the benefits of both approaches

- **Port-based** is *fast* and *privacy-friendly* because:
  - -It needs the 1st packet only
  - -It uses fixed fields (protocol and port)
  - -It uses few data
  - -It can be considered as a special case of packet-classification techniques developed for routers, flow-monitors, etc.

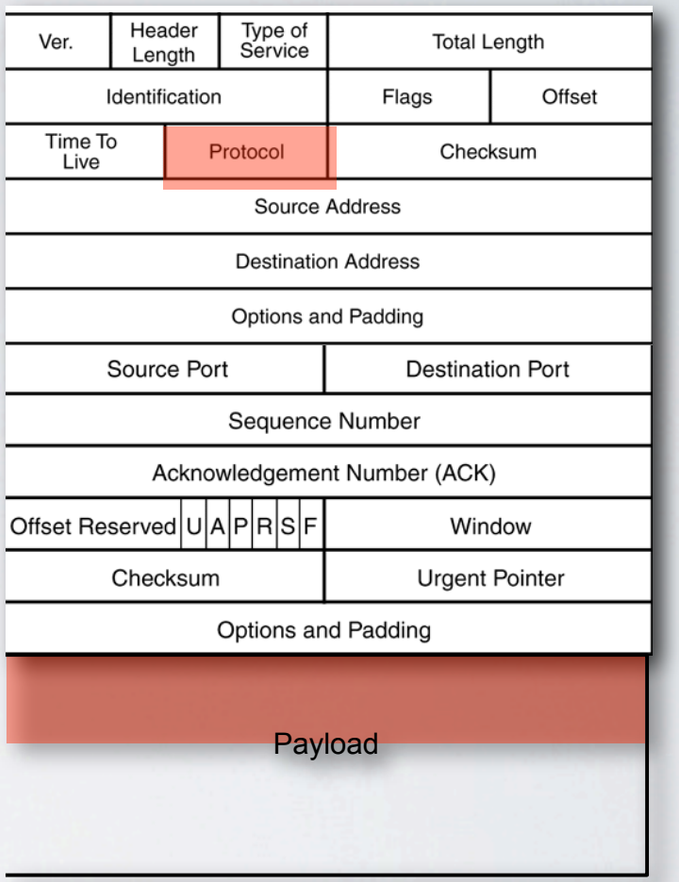
- **Payload-based** is *accurate* because it relies on application-level headers and other information from the payload
  - -Payload-based signatures

COMICS Research Group
University of Napoli "Federico II" - Italy

PORTLOAD

Port + Payload = PortLoad

- A single packet (1st one with payload), fixed fields, limited data (e.g. 32B of payload)



- Payload-based signatures

App_ID	TCP/UDP	direction UP/DW/BOTH	offset	fields 1	2	3	4	5	6	7	8
34	UDP	BOTH	0	I	C	Y	\x20	⊙	⊙	⊙	\x20

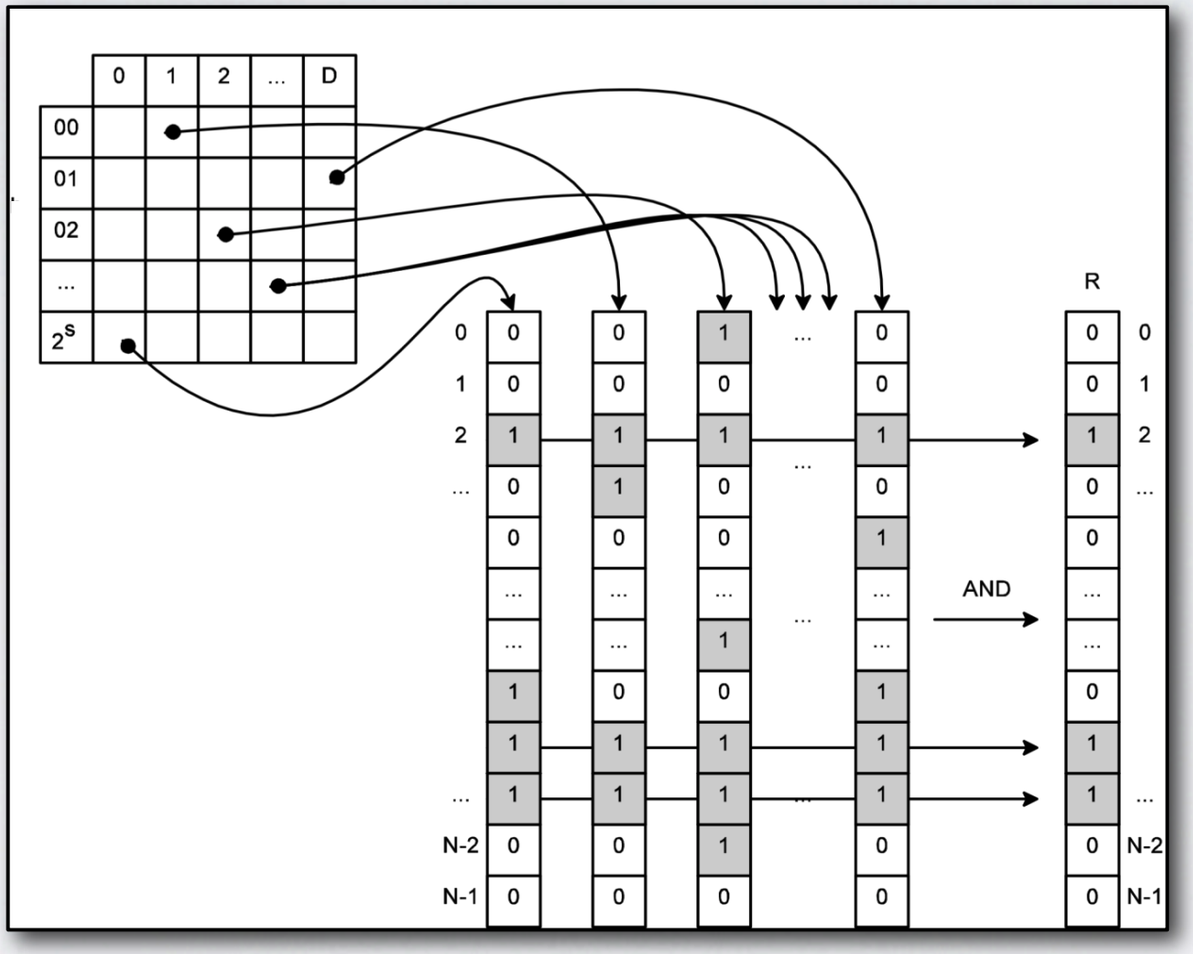
Example of signature for the Shoutcast MP3 streaming application

- Packet-classification matching approach
  - -Indipendent field searches
  - -E.g. bitmap intersection (Lakshman and Stiliadis, SIGCOMM Computer Communication Review, 1998)

PORTLOAD

Bitmap Intersection

- A **bitmap** is assigned to each Field-Value pair
- 1's in a bitmap indicate signatures compatible with that pair
- AND-ing the bitmaps corresponding to packet content will return the matching signatures



COMICS Research Group
University of Napoli "Federico II" - Italy

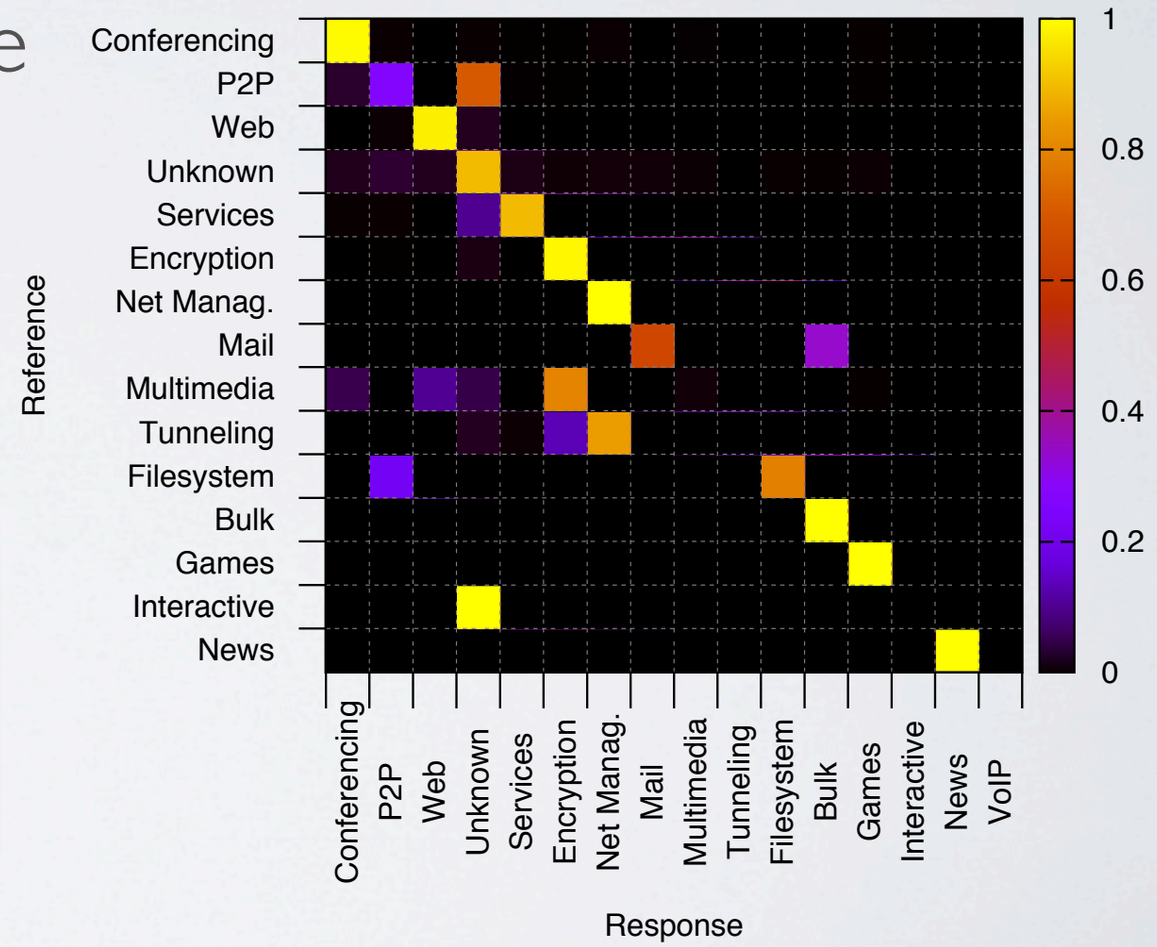
PORTLOAD

evaluation of classification accuracy

- Evaluation (accuracy against TIE-L7) on UNINA trace from Oct. 2009, with a preliminary set of signatures

- We compared results on the same traffic trace obtained with
  - **TIE-L7**
  - **TIE-PortLoad**
  - **TIE-Port**

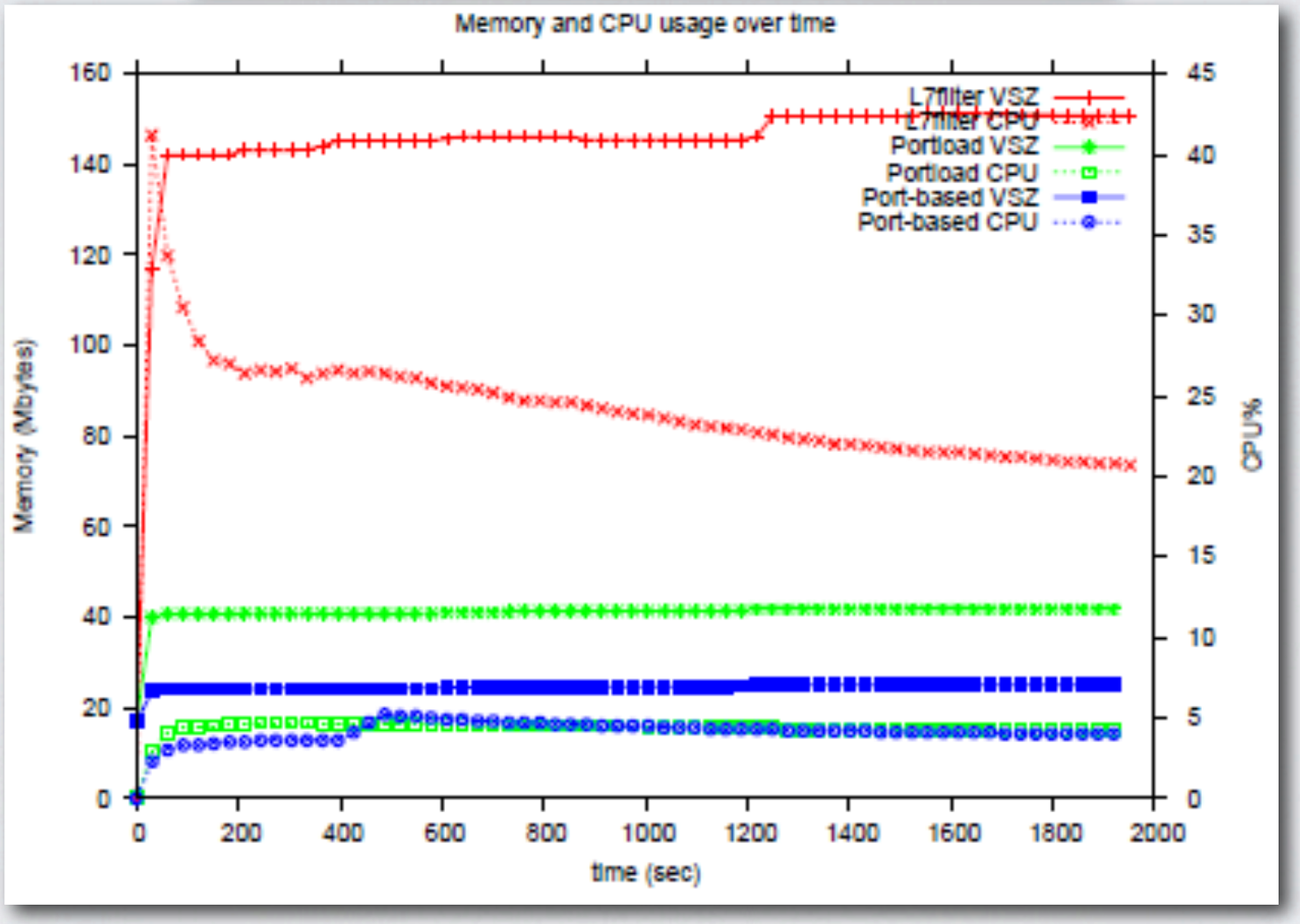
Classifier	Accuracy on applications	
	sessions	bytes
PortLoad	74.24%	97.83%
Port-based	19.57%	25.12%



PORTLOAD

evaluation of performance

Classifier	Mean Time (μsec)	Mean Time (*vs* Port-based)	Variance (μsec^2)
Port-based	2.48	1.0	0.88
PortLoad	6.99	2.8	11.15
L7-Filter	211.4	85.2	47057.88



TIE DEPLOYMENT

what do you need at least

- A **Linux/FreeBSD** box
- An **optical splitter** or switch/router doing **port mirroring**
- A spare **network adapter** or an ENDACE **DAG card**
- The ***pcap*** library
- The CAIDA's **CoralReef** library for live web reports

E.g. we live monitor a 200Mbps link with a Xeon box / FreeBSD 6.3 and a ~$800 DAG card.

RIPE MEETING

TIE and Internet Service Providers

- We are always seeking for **collaborations**

TIE can be used by ISPs for:

- Deploying traffic classification with ***low costs***
- Developing traffic classifiers targeted to specific needs and ***operating problems*** (novel/custom network protocols and encapsulations, specific classes of traffic and applications, etc.)
- Helping in ***monitoring and diagnosing*** network events
- Deploy differentiated ***QoS*** or ***security*** policies
- ***Forecasting*** users-traffic trends
- ...

- We are particularly interested in **listening to ISPs needs** and unsolved technical problems and to discuss them

COMICS Research Group
University of Napoli "Federico II" - Italy

THANKS

feedback is very welcome

http://www.grid.unina.it/Traffic

This work has been funded by the European Project "OneLab2" (ICT FP7 IP 224263).

COMICS Research Group
University of Napoli "Federico II" - Italy