Metroflux: a high performance system for very fine-grain flow analysis

Oana Goga, Patrick Loiseau, Paulo Gonçalves, Romaric Guillier, Matthieu Imbert, Yuetsu Kodama, Pascale Vicat-Blanc Primet

INRIA, France – AIST, Japan

Grants: GridNet-FJ, EU EC-GIN, INRIA Bell-Labs common laboratory

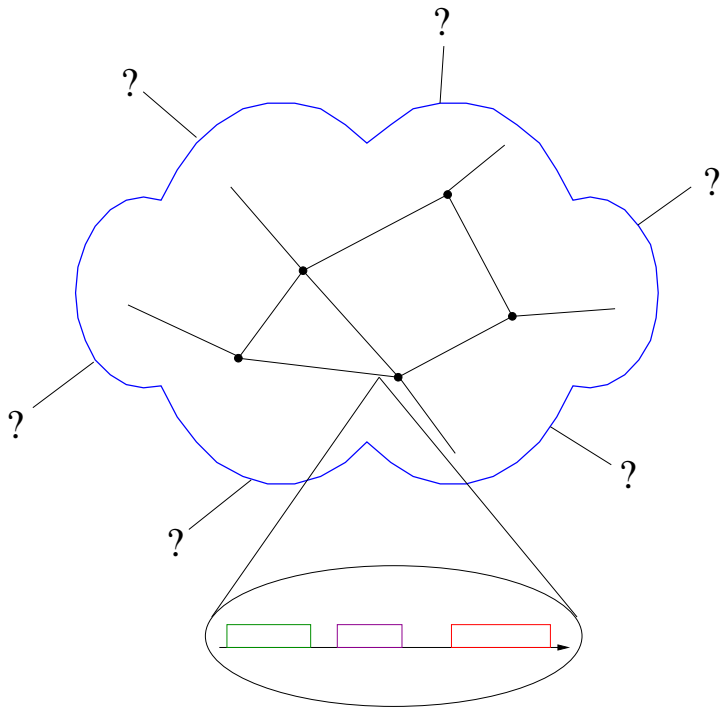
Grid'5000 Spring School, April 9th, 2009

Motivations

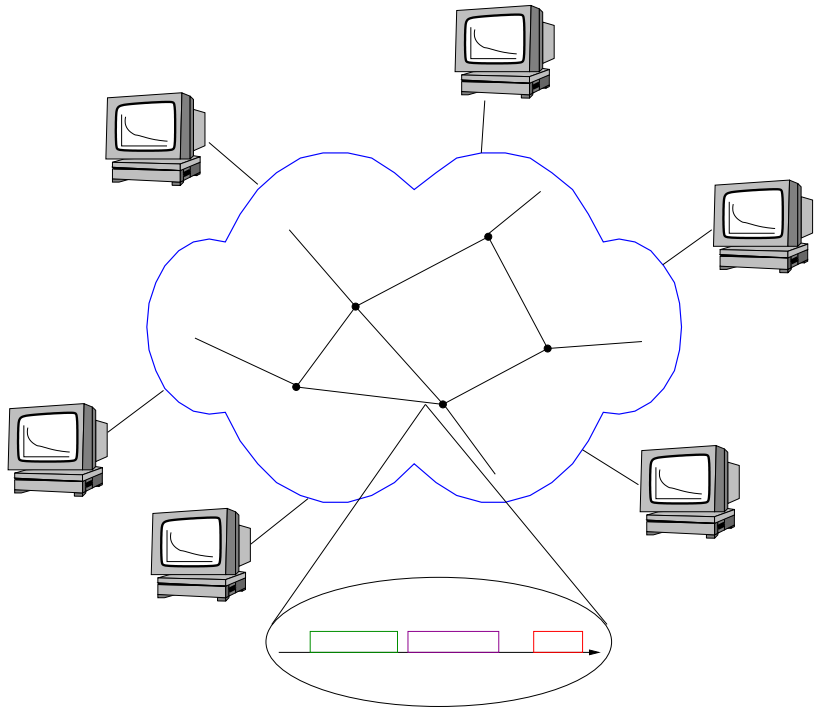
- Main goals
  - Characterize network traffic at very-fine grain (packet grain)
  - Very high speed links
  - Grid'5000 continuous traffic capture
  - On demand (experiment) traffic capture
- Proposed methodology
  1. development of a tool for non-intrusive fine grain capture of traffic traces on high speed links
  2. *off-line* analysis and identification of flows on traffic
     - Uncontrolled experiments (for characterizing network traffic): Collect and analyze traffic traces of uncontrolled sources - for Grid'5000 sources
     - Controlled experiments (for validating theoretical models): Collect and analyze traffic traces of controlled sources (reproducible) - for Grid'5000 experiments

Motivations

Uncontrolled experiments

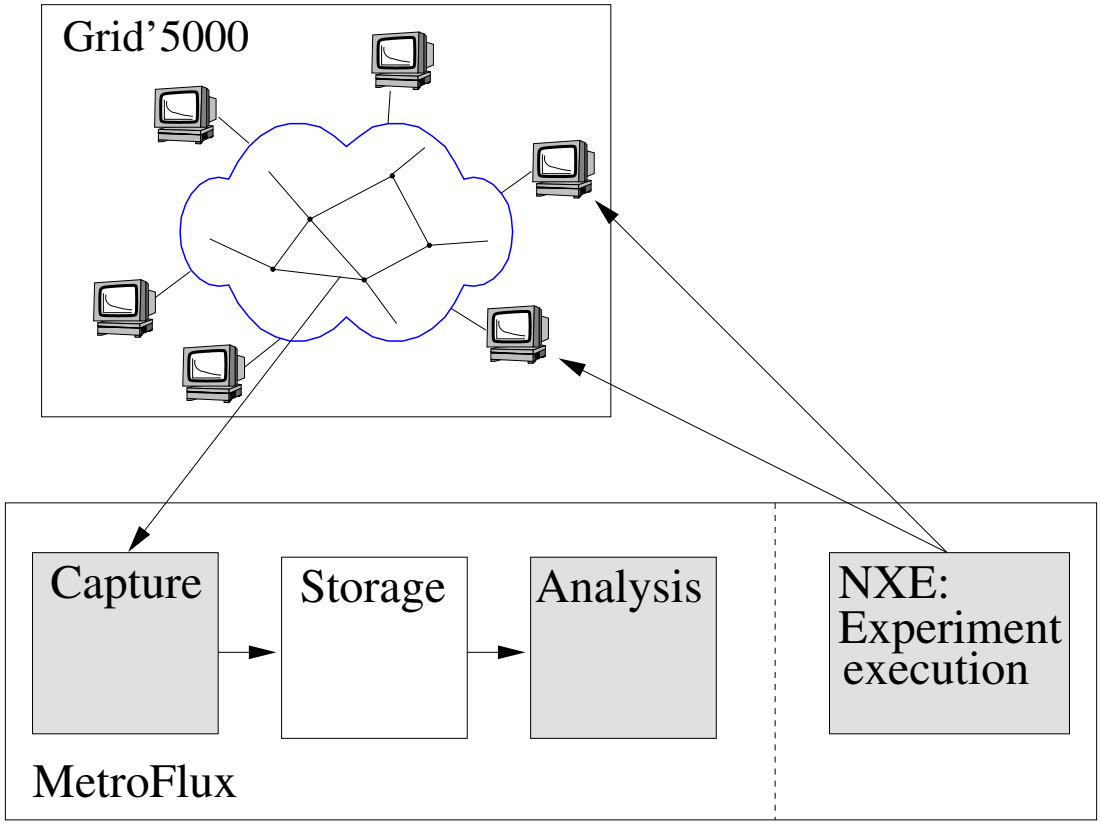


Controlled experiments



- Very-fine grain passive traffic capture at very high speed → *MetroFlux* (with GtrcNet)
- Large scale controlled testbed → Grid'5000 (with *NXE*: Network eXperiment Engine)

General scheme and outline



Outline

1. Capture
2. Analysis – Uncontrolled experiments
3. Controlled experiments
4. Metroflux and Grid'5000

Capture – Fine grain capture of traffic traces on high speed links I

- Rigorous statistical analysis at flow level requires complete packet based time series or controlled sampling processes
- Current (passive measurement) tool limitations
  - Software tools (TCPdump) cannot deal with 1 and 10 Gbps links, end point of measure
  - Netflow (Cisco), sflow, MRTG, give coarse grain stats (minute scale), in core point of measure
  - ▷ Hardware solutions are required for packet capture
  - ▷ 10Gbps links monitoring is a **very challenging** problem (up to 15Mpk/s)

Capture – Fine grain capture of traffic traces on high speed links II

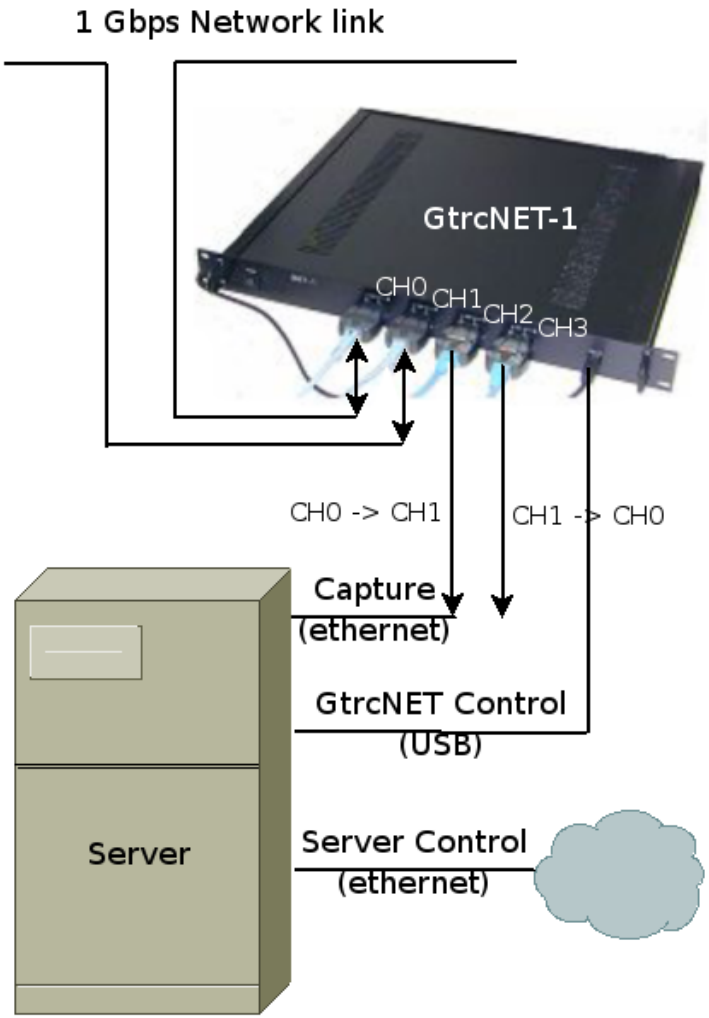
- Hardware solution space
  - DAG card: specialized NIC (ASIC) for packet capture & analysis
  - Programmable Network Processor card (e.g. Intel IXP). No 10Gb/s
  - nGenius sensors : not open, no 10Gb/s solution
  - **FPGA** : flexibility & high performance, two year experience with 10Gb/s
    - AIST GtrcNET : FPGA-based programmable device (1 & 10 Gb/s)
    - Context : EPI RESO & AIST GTRC associated team (GridNet-FJ)
    - 1 GNET10 & 2 GNET1 are installed since 2006 in G5K Lyon.
    - Flexibility: integrate and upgrade packet capture functions within GNET10

Capture – MetroFlux: Measurement system

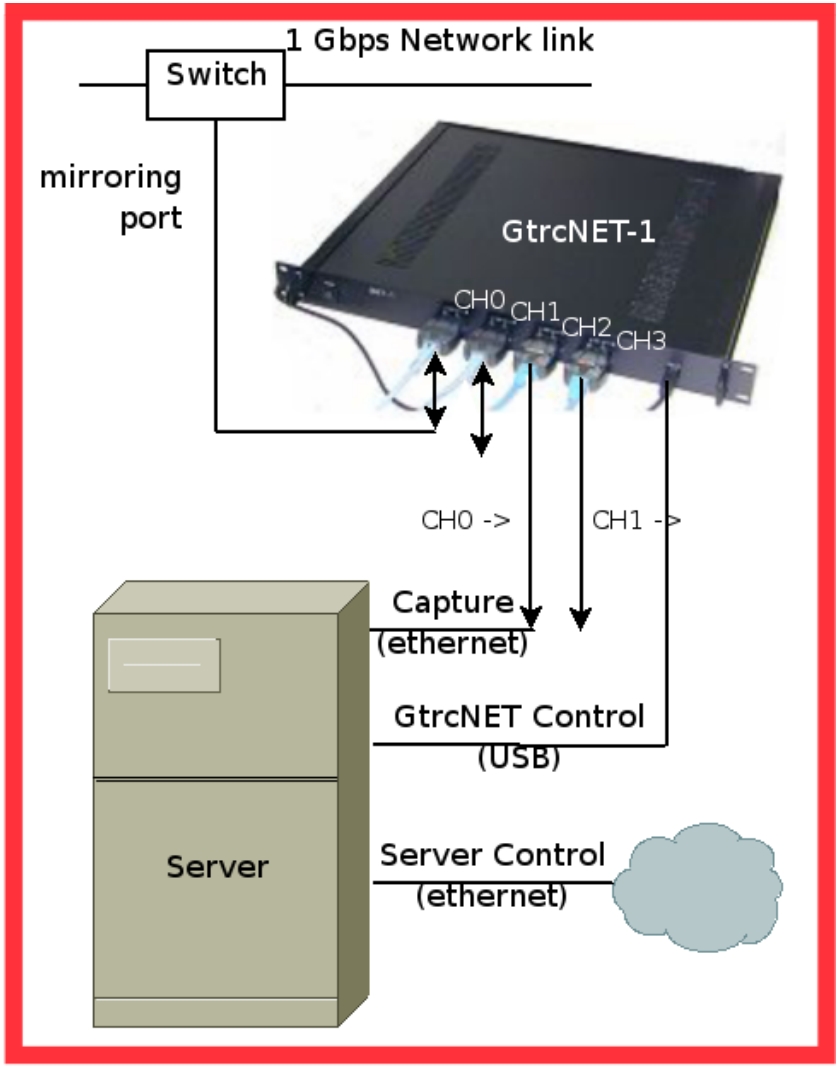
RESO & AIST are jointly designing and developing a system based on:

- GtrcNET box for packet capture
  - MetroFlux V1 = 1 Gpbs link (already developed & deployed in G5K-Lyon)
  - MetroFlux V2 = 10 Gpbs link (under validation)
- Dedicated server for time series storage & processing
  - MetroFlux V1 : 1 quad-core CPU, 5 SAS disks in RAID 0, 4 GB memory, 2 gigabit interfaces
  - MetroFlux V2: min 1 quad-core CPU, 5 300GB SAS 15k rpm disks in RAID0 or 15 1TB SATA 7.5k rpm in RAID6, 4GB memory, 2 x 10GE interface (Myrinet10G)
- A statistical analysis library for off-line processing
- A library for presenting analysis results

Capture – MetroFlux: Measurement system

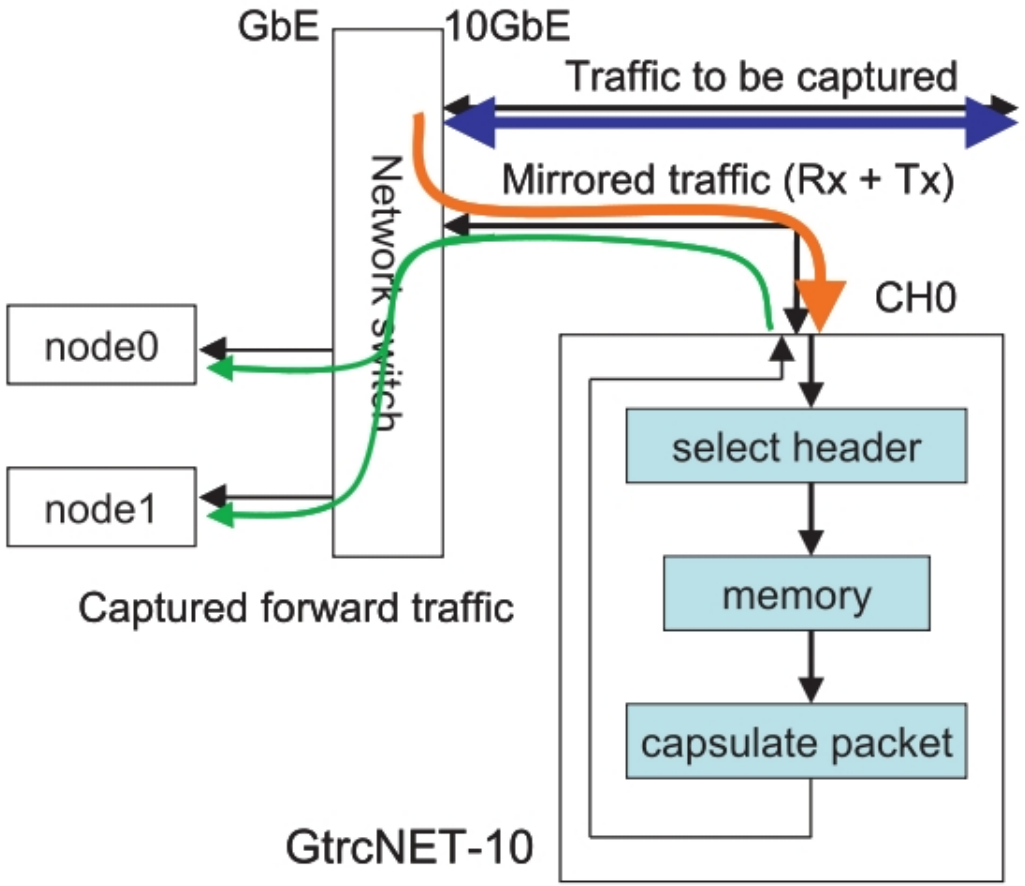


case 1 : system inserted in the link



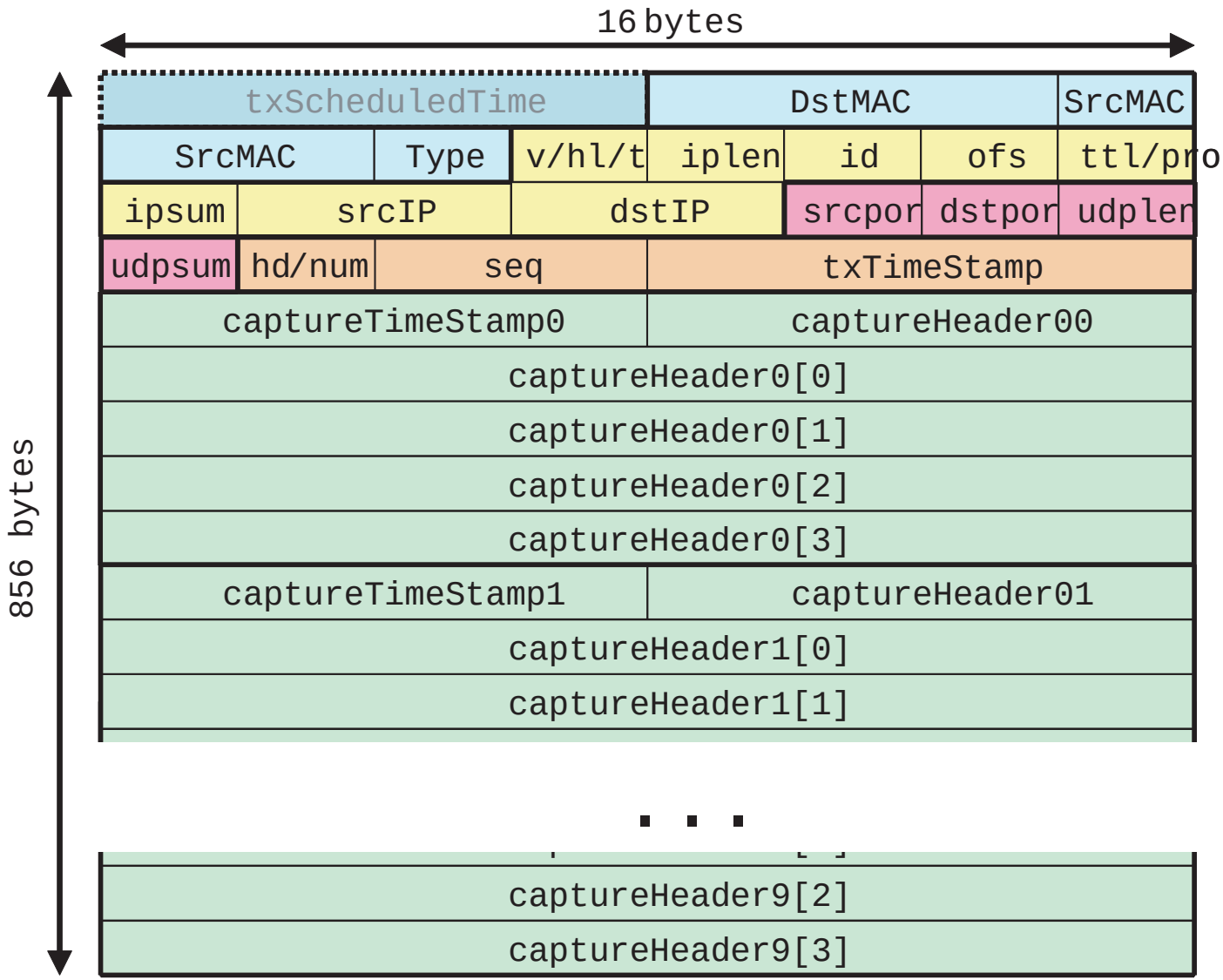
case 2 : system plugged to a mirroring port of a switch

Capture – Example of a packet capture



absorb/header/forward

Capture – Packet format of capture forward



Uncontrolled experiments – Analysis and identification of traffic characteristics - ANR IGTMD project (IN2P3, RENATER, LIP collab)

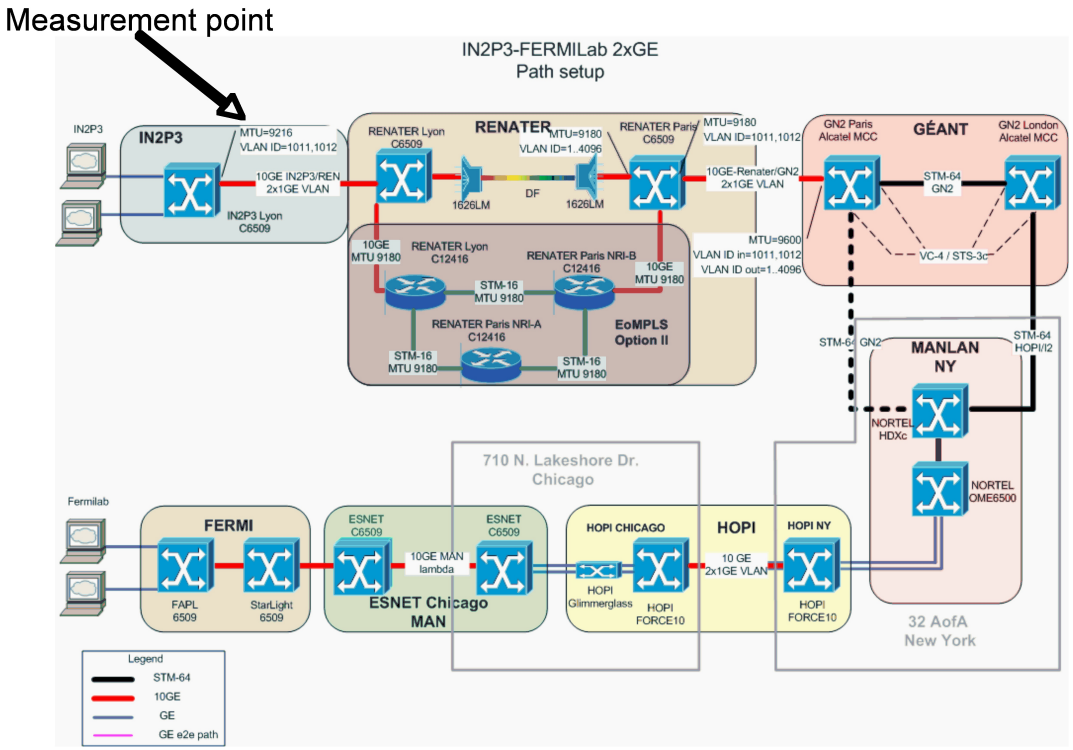
IN2P3 (Lyon) ↔ FermiLab (Chicago)

- Replication of LHC measurements (particule collisions)
- "EGEE" traffic type (cosmology, astrophysics, meteorology. . .)

▷ 60ns time stamp precision

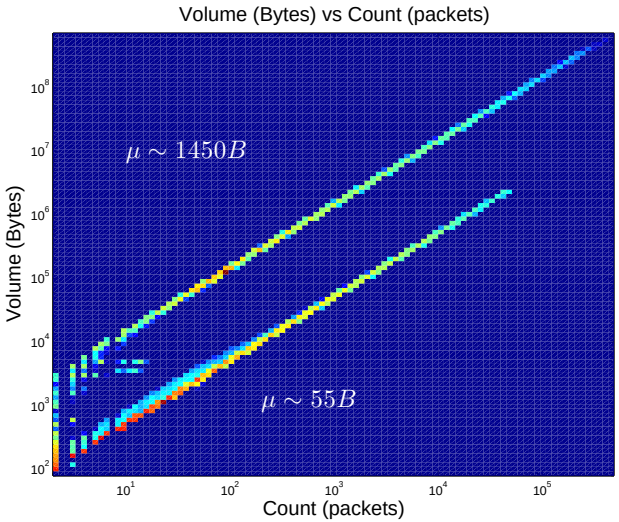
▷ 1Gbps aggregated link monitoring : bidirectional traffic

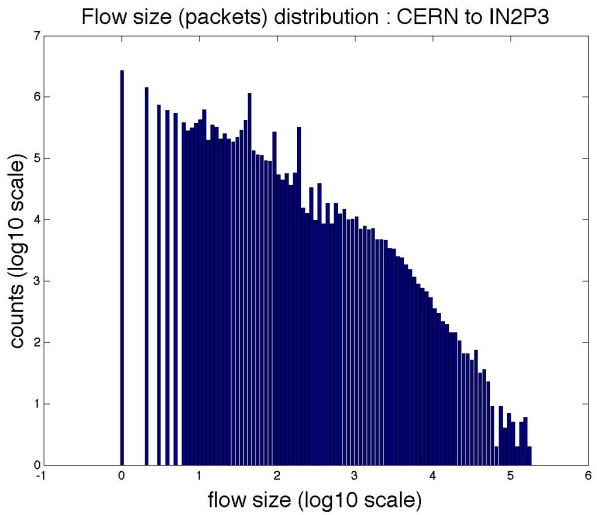
▷ 50 days continuous capture : Feb. - Mar. 2008

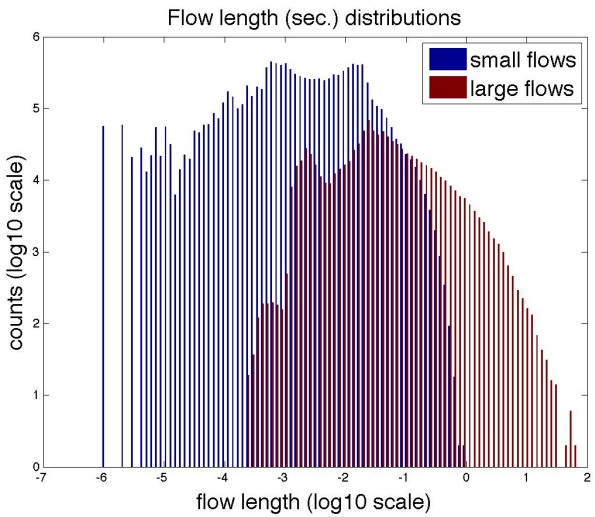


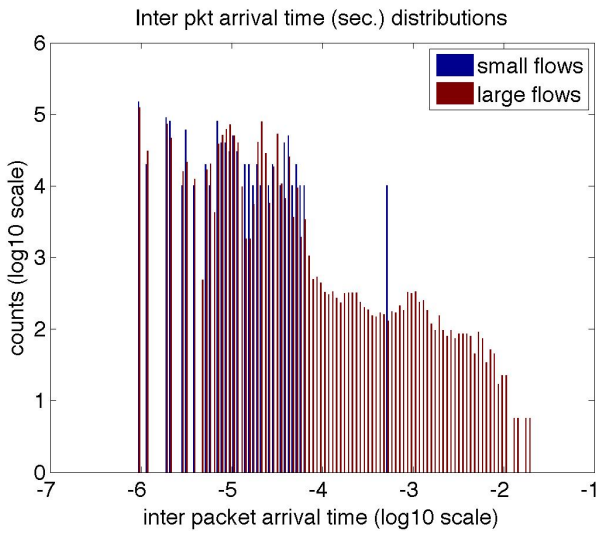
Uncontrolled experiments – Behavioral Analysis

▷ Flow level statistical analysis : traffic model

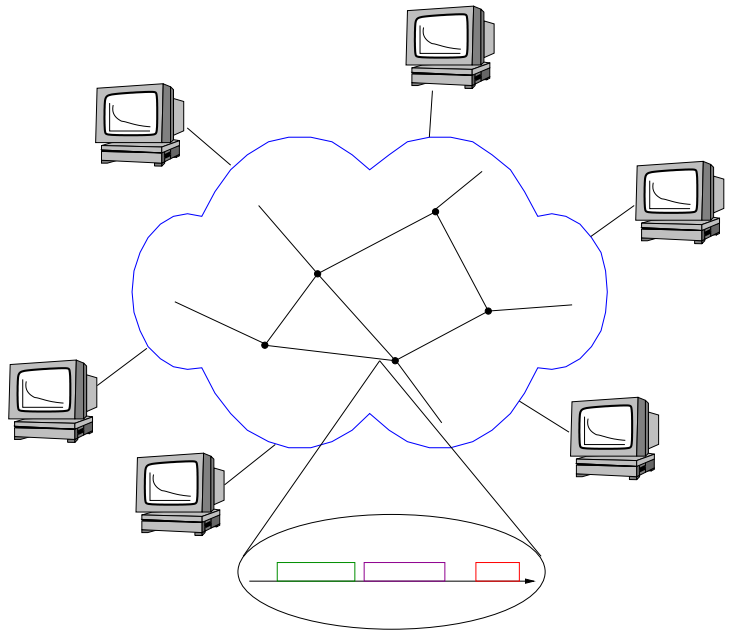






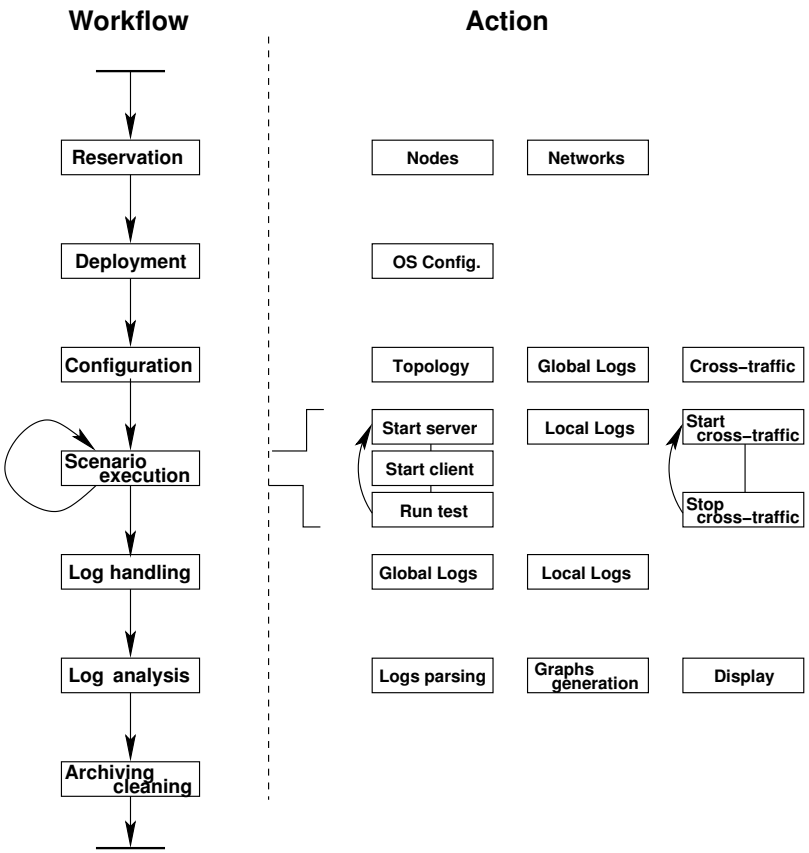


Controlled experiments – Scenario elaboration



- Scenario elaboration
  - Topology definition
  - Traffic characteristics definition (flow size distribution, protocol, etc.)
  - Measurement point choice and traffic capture
  - Off-line results analysis

Controlled experiments – Scenario execution: NXE workflow and interface [Guillier, Vicat-Blanc Primet 08, 09]



Reservation available resource allocator services contacted to get the resources needed by the experiment

Deployment reboot of the nodes with adequate kernel image, or simple setting of the OS internal variables

Configuration available hardware (e.g. hardware latency emulator, routers) contacted to alter the topology, or to activate statistics collect

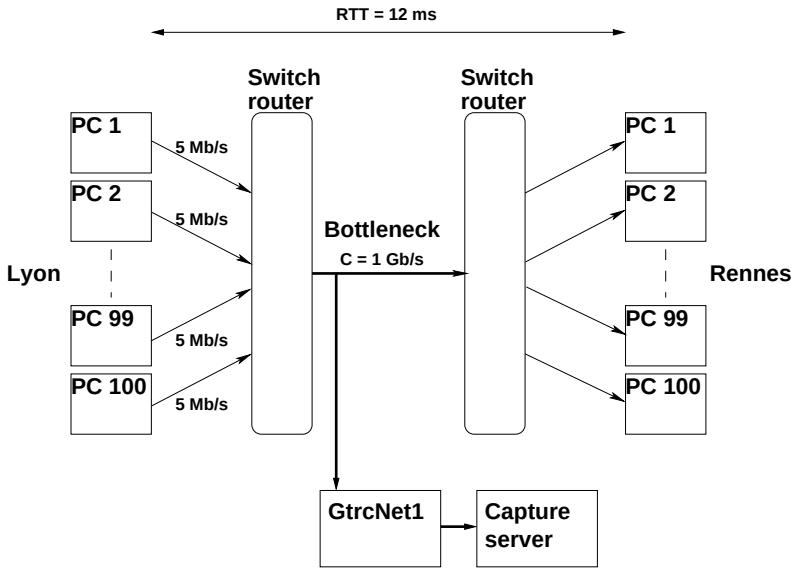
Scenario execution start actual scenario execution (scenario can repeatedly run for consistency purpose)

Log handling nodes' logs and global logging facility merged for analysis

Log analysis logs' parsing and metrics computation to generate graphs that are easy to interpret.

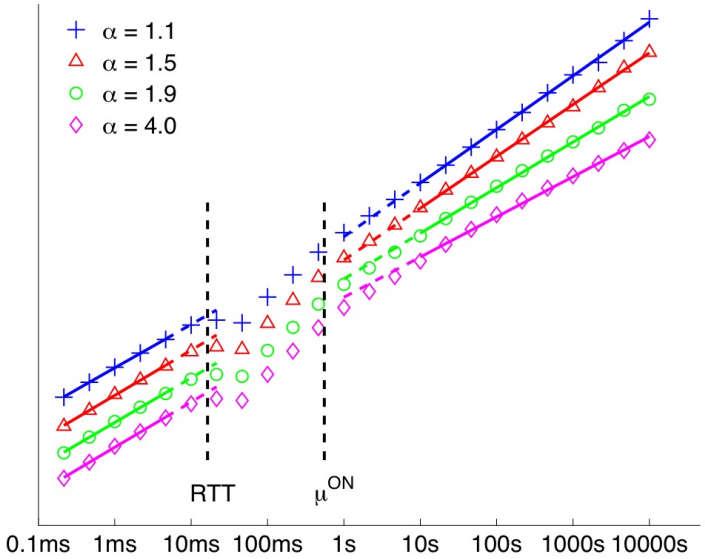
Archiving and cleaning resources reset and release. XML formating of experiment general conditions.

Controlled experiments – Example of a Grid'5000 and NXE based experimental setup



- 100 sources on each side
- 8 hours of TCP / UDP independent random connections
- flow sizes drawn at random from a heavy-tailed distribution
- aggregated traffic captured at the LYON access router
- *Metroflux* connected to mirror the output port

Controlled experiments – Example of results [Loiseau & al. 08, 09]

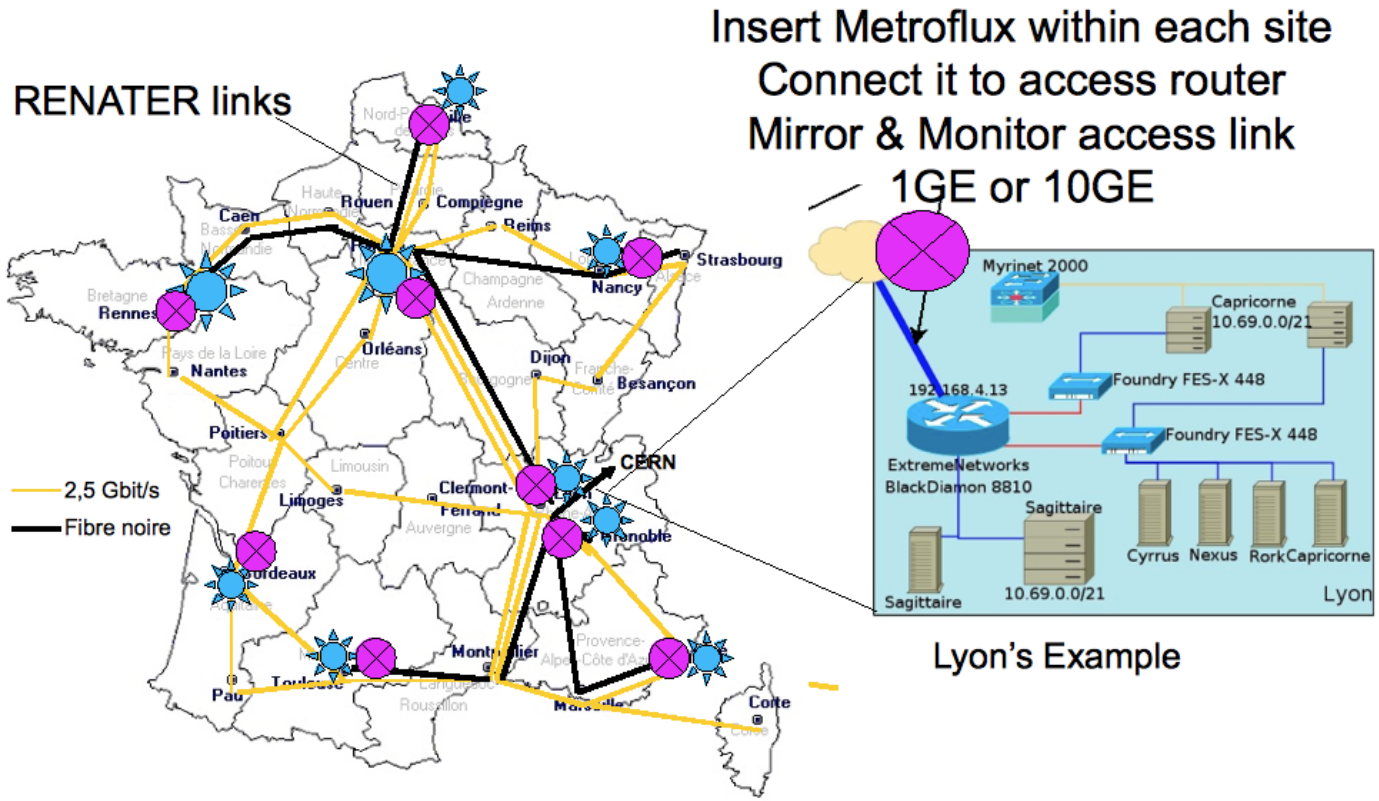


- Possibility to analyze the scaling behavior of the traffic at every scale:
  - small scale ($<$0.1 ms) thanks to the very fine grain capture
  - intermediate scale (protocol characteristic scale)
  - large scale (validation of Taqqu Theorem)

▷ Allows for experimentally verifying theoretical results on a realistic and fully controllable facility

▷ Provides with a flexible testbed for the development of protocols and control policies

Metroflux and Grid'5000 – Architecture Grid Level



- support capture from different types of probes (GNET, sFlow, NetFlow, PCAP, DAG) and provide an unified interface for traffic analysis
- integration with OAR
- support interaction with other programs
- management of captures and analysis, giving the possibility to later access the results
- real time statistics and user request traffic analysis

Metroflux and Grid'5000 – Summary of required metrics

Metric	Description and Notes
RTT	Time between a packet being sent and a reply packet being received. Useful for general performance troubleshooting.
OWD	The time between a packet entering the network and it being received at the destination. Useful for monitoring performance of QoS networks.
IPDV	Difference of two OWDs, sometimes known as jitter.
Packet Loss	The fraction of packets sent but not received. One-way measurements useful for QoS monitoring, two-way for real life application performance.
Capacity	The maximum amount of data per time unit available for a particular path when there is no competing traffic.
Utilisation	The capacity currently being consumed on a given path.
Available Bandwidth	The maximum amount of data per time unit that a path can provide given the current utilisation. Often derived directly from the capacity and utilisation rather than measured directly.

Conclusions – Perspectives

- Conclusions
  - MetroFlux permits very fine grain traffic capture on very high speed link
    $\longrightarrow$ with GtrcNet
  - MetroFlux is a portable and non intrusive device to analyze unknown real network traffic
  - MetroFlux combined with Grid5000 offers a large scale testbed for fully controlled experiments
    $\longrightarrow$ with NXE
- Perspective
  - design an equivalent system suitable for on-line analysis
  - multiple measurement points on Grid5000
  - study of possible integration of MetroFlux with COMO (OneLab)
  - integration into the future internet for traffic awareness

Contacts – References

- Contacts
  - GtrcNet: Yuetsu Kodama, AIST (y-kodama@aist.go.jp)
  - NXE: Romaric Guillier, INRIA (Romaric.Guillier@ens-lyon.fr)
  - MetroFlux – Grid'5000: Oana Goga, INRIA/ENS Lyon, Patrick Loiseau, Université de Lyon/ENS Lyon, Paulo Gonçalves and Pascale Vicat-Blanc Primet INRIA/ENS Lyon ({Oana.Goga, Patrick.Loiseau, Paulo.Goncalves, Pascale.Primet}@ens-lyon.fr)
- References
  - Romaric Guillier and Pascale Vicat-Blanc Primet, *Methodologies and Tools for Exploring Transport Protocols in the Context of High-Speed Networks*, IEEE TCSC Doctoral Symposium, 2008
  - Romaric Guillier and Pascale Vicat-Blanc Primet, *A User-Oriented Test Suite for Transport Protocols Comparison in DataGrid Context*, ICOIN, 2009
  - Patrick Loiseau, Paulo Gonçalves, Guillaume Dewaele, Pierre Borgnat, Patrice Abry and and Pascale Vicat-Blanc Primet, *Investigating self-similarity and heavy-tailed distributions on a large scale experimental facility*, INRIA RR, 2008
  - Patrick Loiseau, Paulo Gonçalves, Stéphane Girard, Florence Forbes and Pascale Vicat-Blanc Primet, *Maximum likelihood estimation of the flow size distribution tail index from sampled data*, Sigmetrics, 2009

Questions?

Thank You!!