ANR - IGTMD - 31 juin 2008

Research on High Speed Transport Protocols & Services

Romaric Guillier, Chen Cheng, Pascale Primet
Sebastien Soudan, Paulo Goncalvès
Equipe-Projet INRIA RESO
Ecole Normale Supérieure de Lyon
Pascale.Primet@inria.fr

Summary

- Our goals in IGTMD
  - Explore transport protocols in high speed networks
  - Analyse and characterize limits of existing scheme
  - Study end system issues: disk2disk + 10Gb/s
  - Explore transport services for data intensive applications
  - Understand specificities of LCG traffic
- Current advances & issues
  - Methodology, Metrics, Benchmark
  - Continue protocol analysis (network & traffic model impact)
  - Traffic and flow analysis in the IGTMD environment
  - Explore flow scheduling approaches in LCG context.
- Future work
  - Continue ongoing work
  - Disk to disk and 10Gb/s end system issues
  - Propose and “autonomic transport service configuration” approach
  - Validate innovative transport services with data intensive applications

- Collaborations & animation:
  - GridNet-FJ and Negst & Naregi with AIST japan/ Titech/ University of Osaka
  - EC-GIN (BDTS), Grid5000 & CARRIOCAS
  - OGF (NSI, GHPN) & IRTF (TMRG, ICCRG)
  - Steering committee of PFLDNET & GridNETS

Ongoing work : advances & issues

–Continue protocol analysis:(Romaric Guillier)

- •Network & Traffic models analysis
- •Protocol Characterisation testbed: NXE
  - => Collaboration with ORTF TMRG

–Traffic and flow analysis in the IGTMD environment (P. Goncalvès)

- • Capture & analyse LCG traffic at Lyon (with Gnet1 & on the router (netflow))
- •Metroflux approach
  - ⇒Collaboration with AIST (Japan) & EC-GIN (EU)
- *• IGTMD measurements:*
  - *– pb with captured traffic : only two pairs….difficult to do valuable statistics on the traffic*

–Disk to disk and 10Gb/s end system issues

- • Myrinet cards evaluation: without/with latency at Lyon (with GNET10)

–Explore transport services for data intensive applications

- • RFT, GridFTP (NEGST - Osaka)

–Explore flow scheduling approaches (forward & reverse) in LCG context.

Future work

– Continue ongoing work
– Propose and “autonomic transport service configuration” approach
    • Characterize network path
    • Identify bottlenecks
    • Automatically select the protocol and its parameters
    • Automatically configure the protocol, the network and the application

– Validate innovative transport services with data intensive applications
    •Select realistic workload
    •Run, measure and analyse transfers

Focus

- Metroflux
- BDTS

MetroFlux

A very high performance and programmable system for Network Traffic analysis and modeling

Damien Ancelin, Patrick Loiseau & Paulo Goncalvès
Pascale Vicat-Blanc Primet
Equipe-Projet INRIA RESO
Ecole Normale Supérieure de Lyon
Pascale.Primet@inria.fr

Motivations

- Some questions we have:
  - Does the Grid traffic exhibit same statistical characteristics as the Internet traffic?
  - Can we understand the cause of LRD phenomenon observed in Internet traffic?
  - Has this property a positive or negative impact on QoS or resource utilisation?

- Goal: design & deploy a dedicated network instrument
  - For traffic analysis => to provide consistent models for simulations & expe
  - For network performance measurement => to feed schedulers & middleware
  - For precise experiment analysis => to quantify and understand flow behaviors
  - For very long distance emulation => to enable broad range of latencies investigation

Motivations

- Current (passive measurement) tools limitations
  - Software tools (TCPdump) cannot deal with 1 Gbps and 10 Gbps lines
  - Netflow (Cisco) sflow, MRTG, give coarse grain stats (minute scale)

⇒ Complete time series (packet based) or controlled sampling are required to enable rigourous flow by flow statistical analysis

⇒ 10Gbps links monitoring is a VERY CHALLENGING problem (up to 15Mpk/s)

⇒ Hardware solutions are required for packet capture

- Solution space
  - DAG card : specialized NIC (ASIC) for packet capture & analysis (optical derivation)
  - Programmable card based on Network Processor (ex: Intel IXP card, 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.

⇒ Upgrade and integrate packet capture function within GNET10

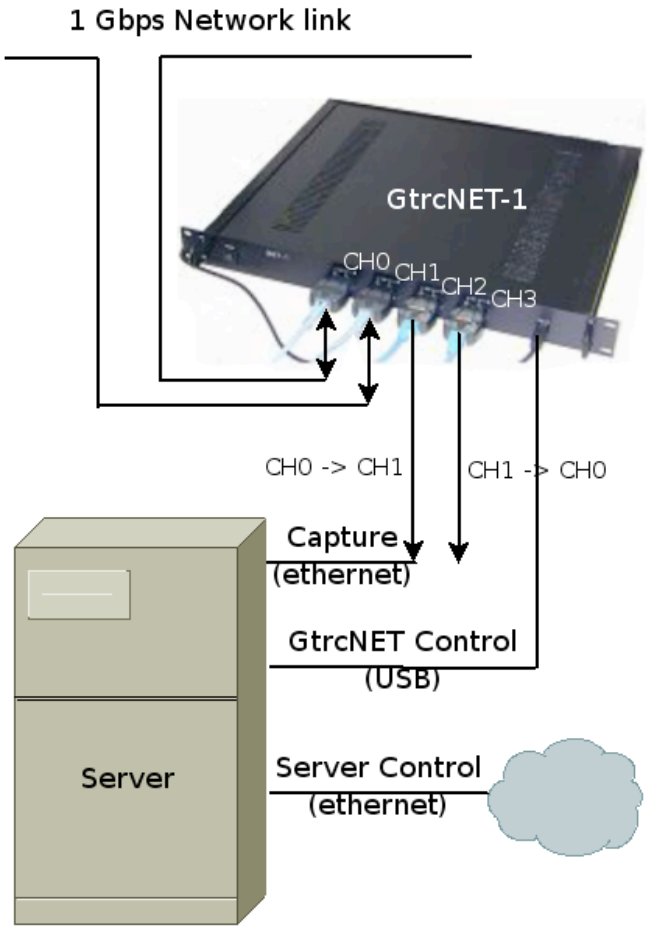
MetroFlux system design

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

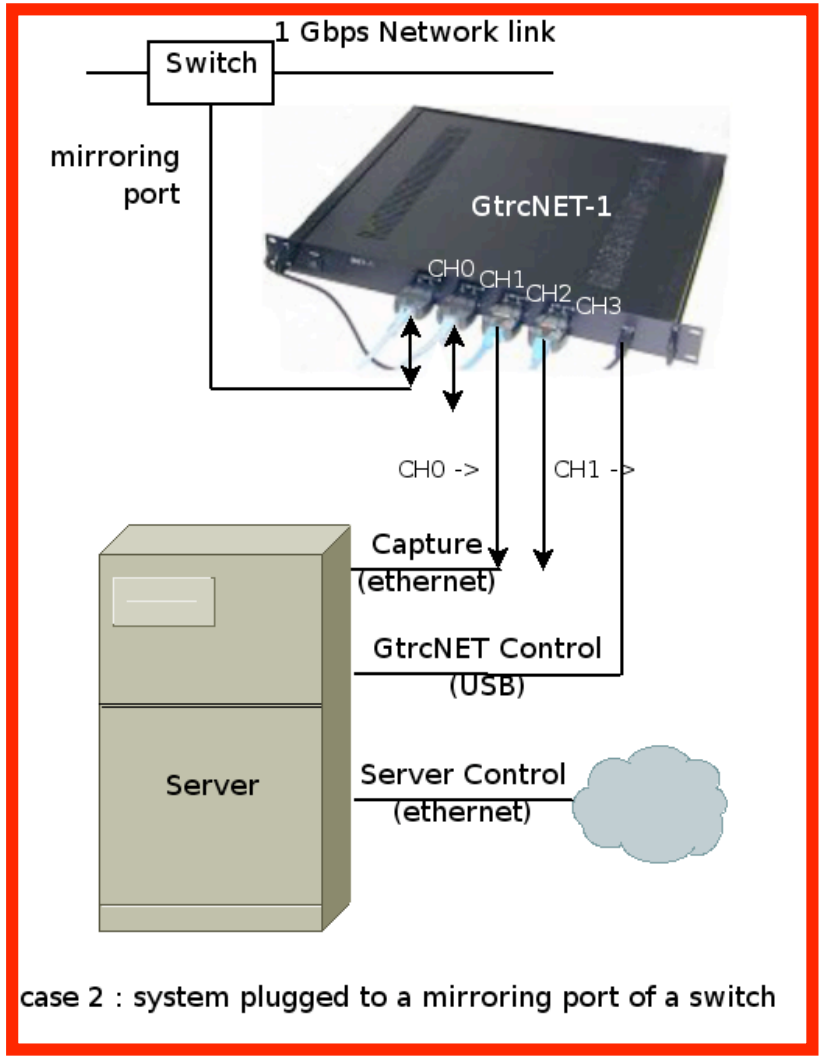
- GtrcNET box for packet capture
  - MetroFlux V1 = 1 Gpbs link (already developed & deployed in G5K-Lyon)
  - MetroFlux V2 = 10 Gpbs link

- 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 SAS disks in RAID 0, 4 GB memory, 2 x 10GE interface (Myrinet10G)

- A Statistical analysis library for online and offline processing
- A library for results presentation



case 1 : system inserted in the link

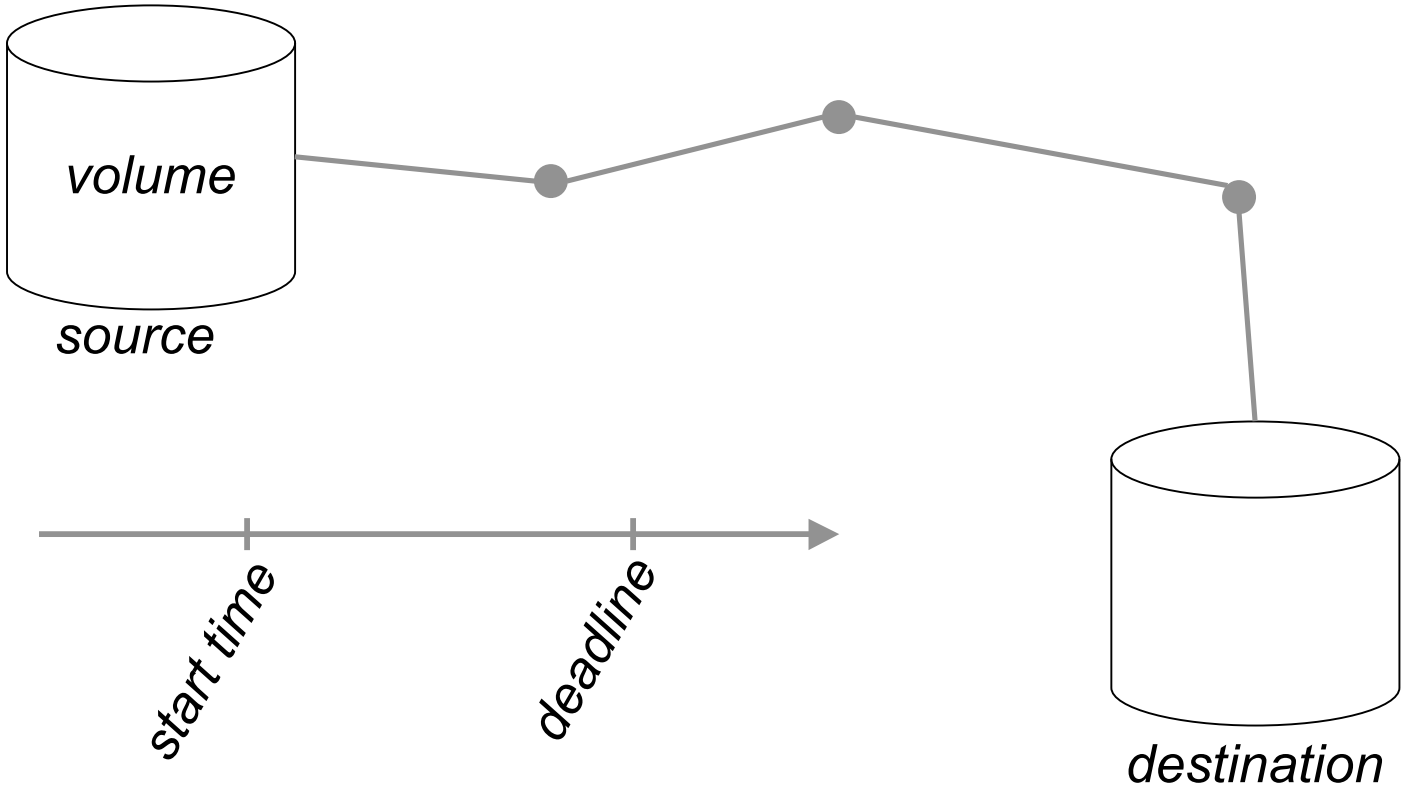


case 2 : system plugged to a mirroring port of a switch

BDTS: Bulk Data Transfer Scheduling

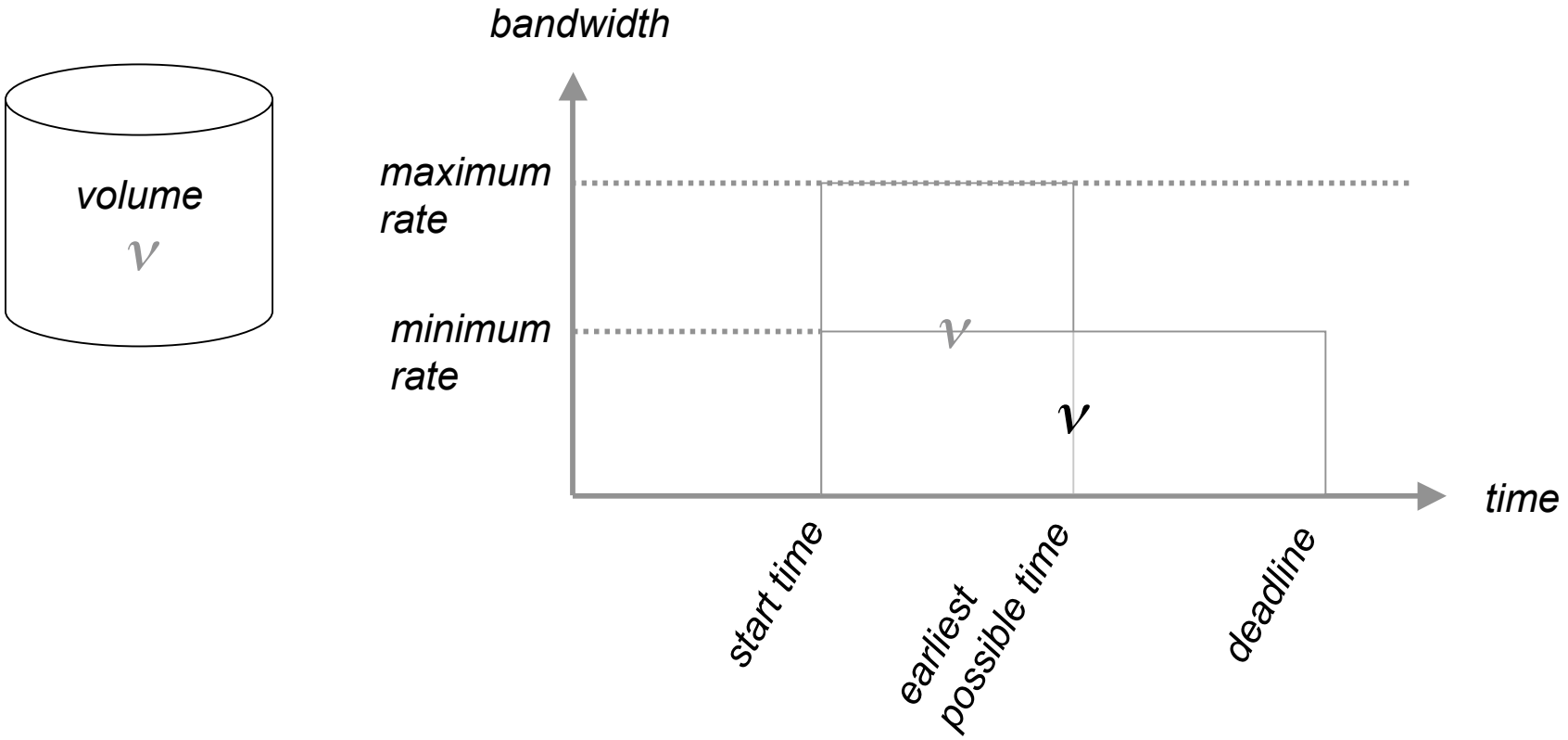
Bulk Data Transfer Scheduling Service

Transfer Job: t-job

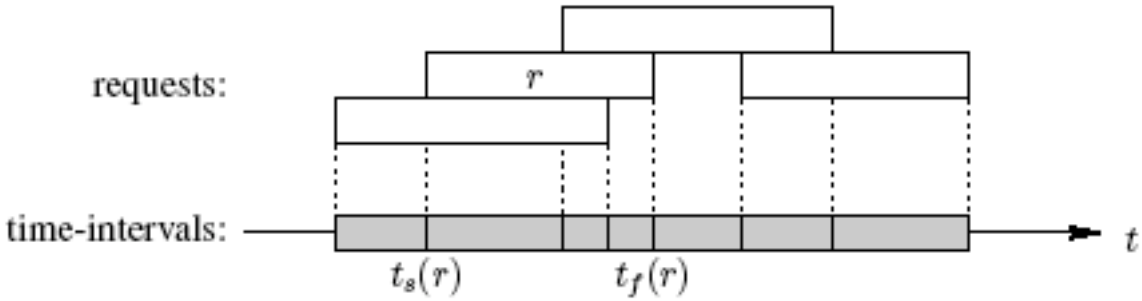


Bulk Data Transfer Scheduling Service

Flexibility: rate and time allocation



Bulk Data Transfer Scheduling Service



Input:

- *Time window of request r :* $[\eta_r , \psi_r]$
- *Volume of data to send of request r :* ν_r

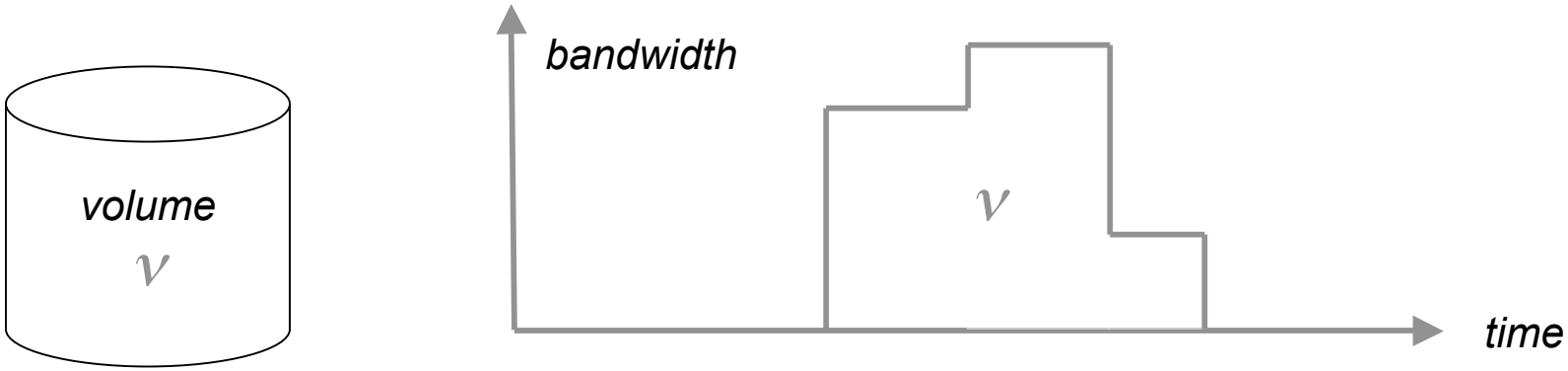
Output:

- *Schedule window of request r:* $[\sigma(r), \tau(r)]$
- $\eta_r \leq \sigma(r) < \tau(r) \leq \psi_r$

$\Rightarrow \tau(r) = \sigma(r) + \nu_r / \lambda$

Bulk Data Transfer Scheduling Service

Step function: profile



$$\lambda(t) = a_0 + a_1 h_\omega^{b_1}(t) + a_2 h_\omega^{b_2}(t) + \cdots + a_{n-1} h_\omega^{b_{n-1}}(t) \quad (1)$$

where:

$$h_\omega^b(t) = \begin{cases} 0 & t \in [\eta, b) \\ 1 & t \in [b, \psi] \end{cases} \quad (2)$$

Bulk Data Transfer Scheduling Service

Architecture