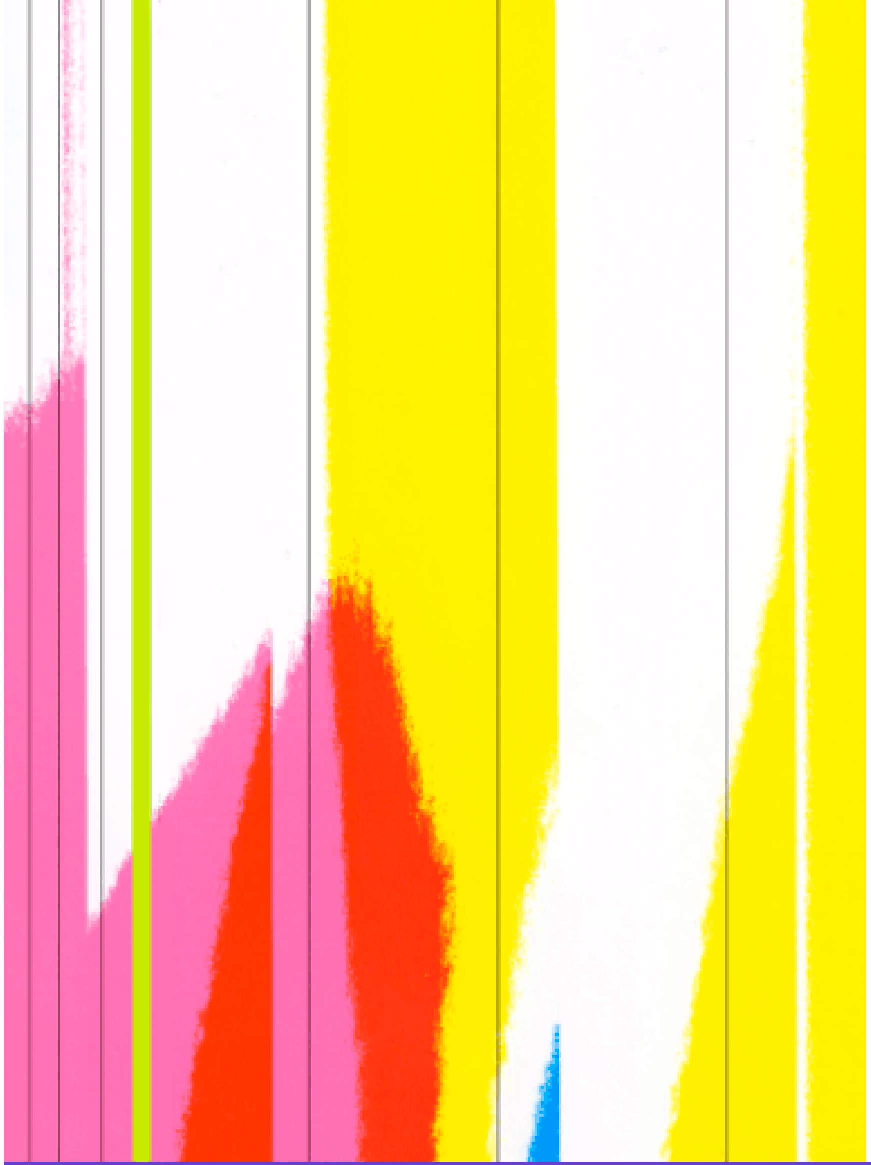


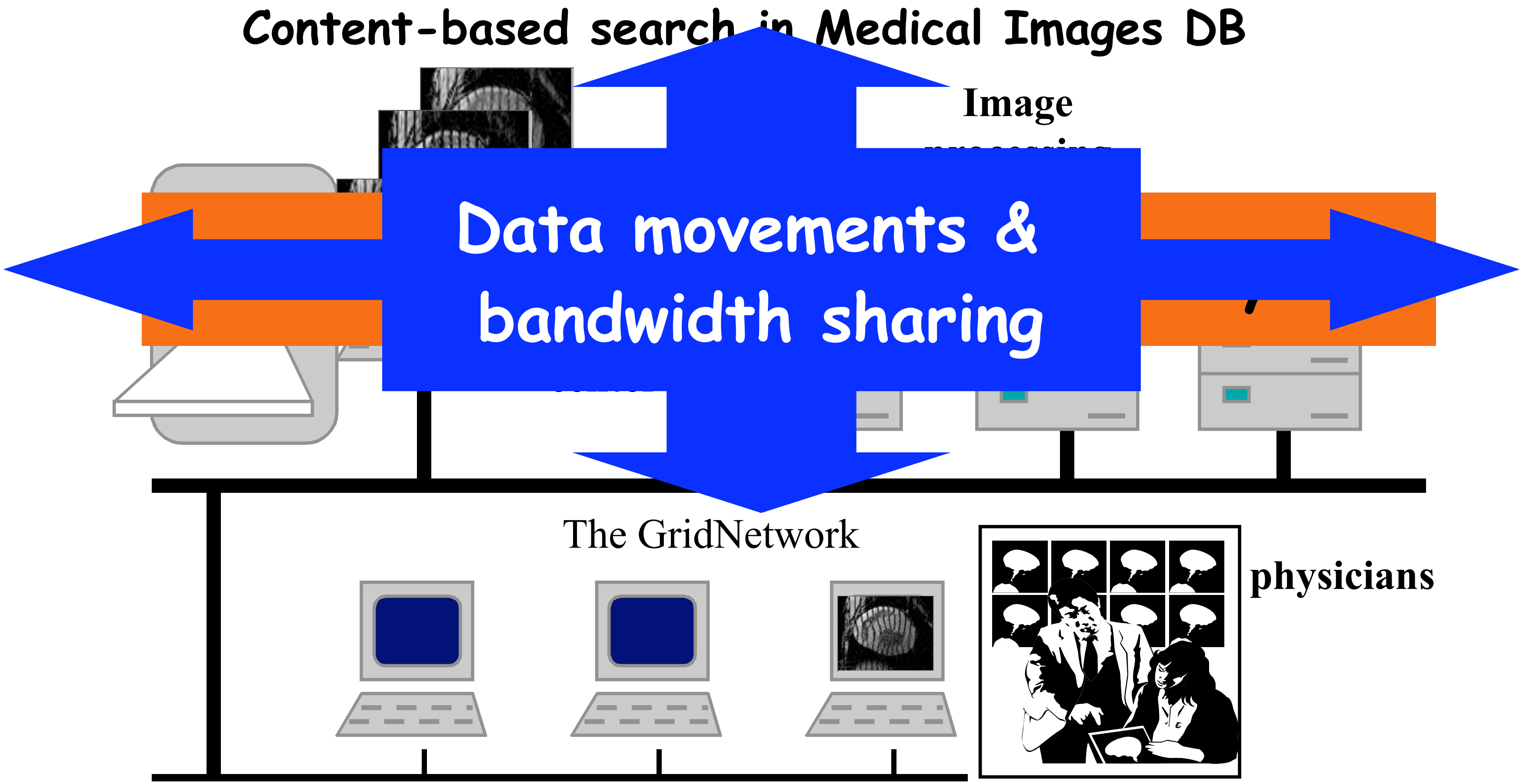
Pfldnet2006
Nara - Japan
2-3 february 2006

Large scale Gigabit Emulated Testbed for Grid Transport Evaluation

P. Primet*, R. Takano**+***, Y. Kodama***, T. Kudoh***
O. Gluck*, C. Otal*

* INRIA RESO - France
***AIST - GTRC - JAPAN; **AXE, inc. JAPAN,

Example of Grid Application

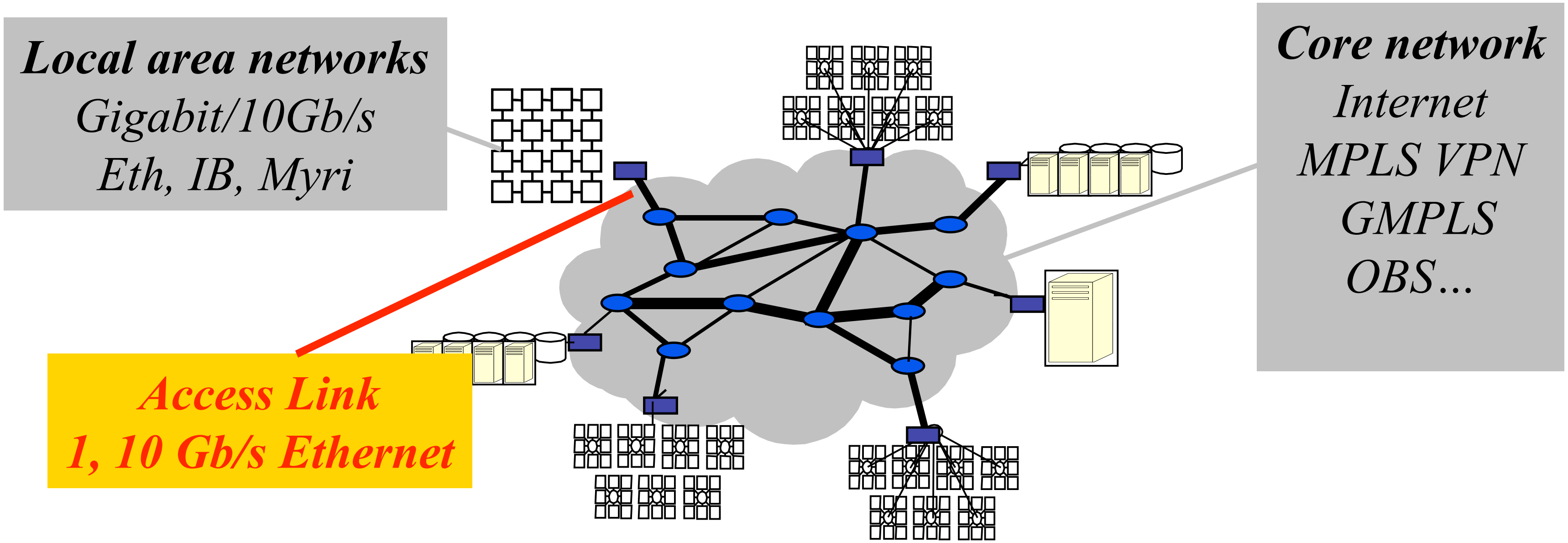


Content-based search in Medical Images DB

Data movements & bandwidth sharing

GridNetwork issues

- The shared resources are interconnected by a complex internetwork
- Applications use Internet protocols: TCP/IP



Local area networks
Gigabit/10Gb/s
Eth, IB, Myri

Core network
Internet
MPLS VPN
GMPLS
OBS…

Access Link
1, 10 Gb/s Ethernet

⇒Issues: Security, **E2E performance prediction and control**

Constraints

Relationships between grid entities (e.g. users, services, resources, virtual organizations, etc.) are **dynamic, and possibly short-lived.**

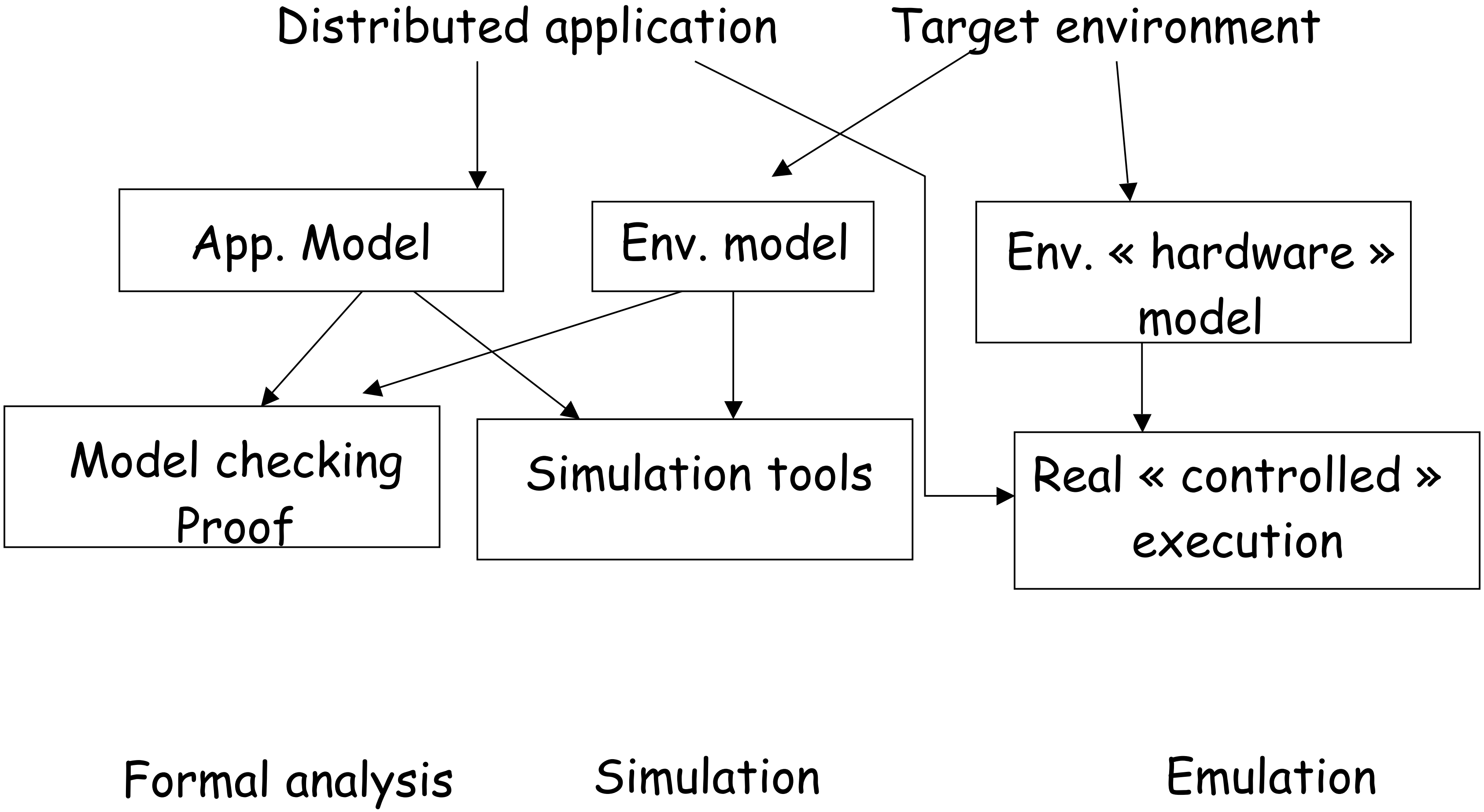
The **network interconnect is dynamic** :

- might grow, diminish, move, etc. It is not a fixed entity w.r.t. location and constitution.

The entities would like to use the communication infrastructure **transparently**:

- end-use, applications, tools and APIs behave like if they were using a regular TCP/IP Internet or intranet.

Distributed application evaluation



GRID5000 french initiative

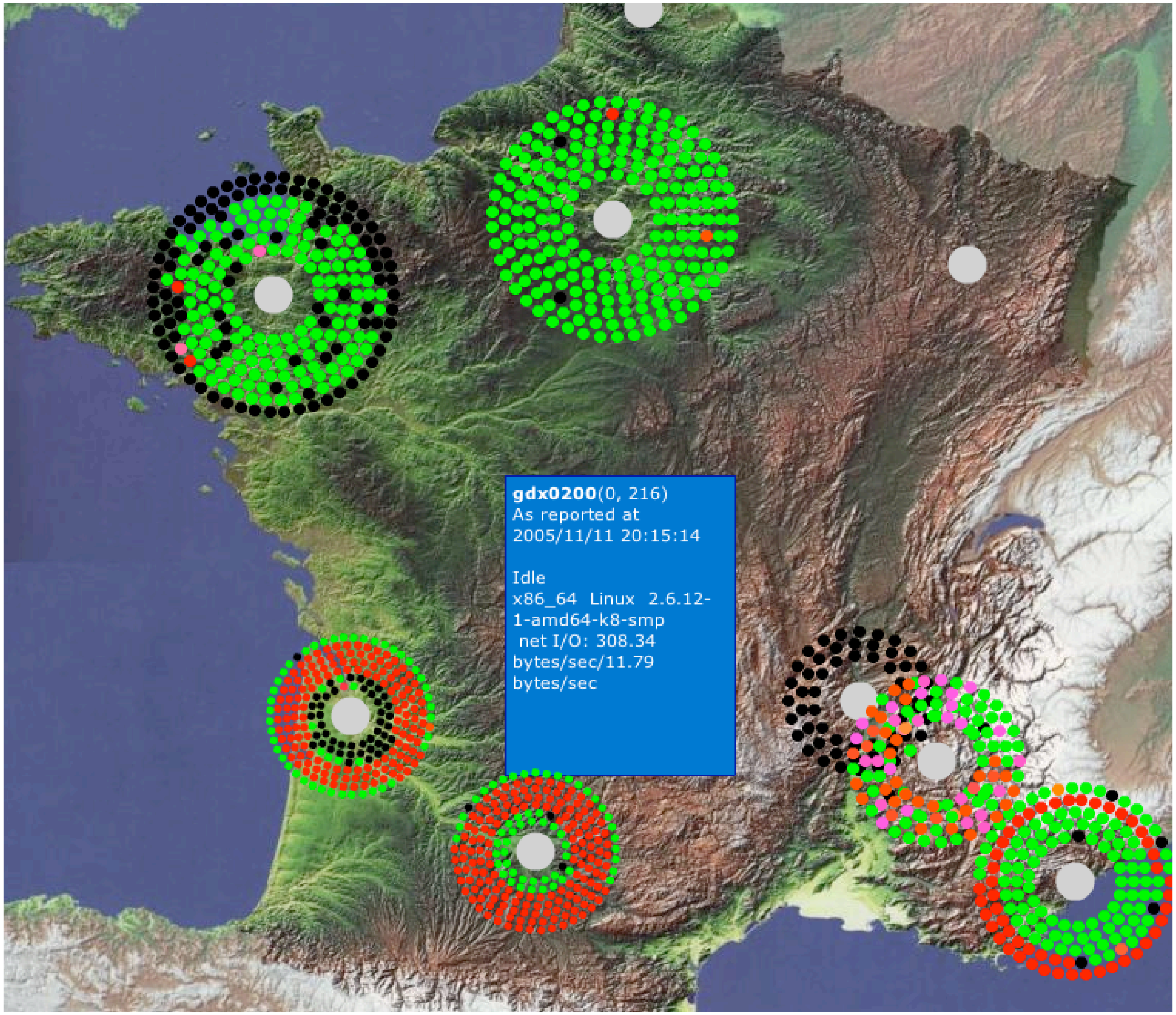
A nation wide experimental platform for Grid researches

- 9 geographically distributed sites
- every site hosts a cluster (from 256 CPUs to 1K CPUs)
- All sites are connected by RENATER (10Gb/s DWDM VPN)
- A system/middleware environment for safe and repeatable experiments

Run Grid experiments in real life conditions

- Address critical issues of Grid system/middleware:
  - Programming, Scalability, Fault Tolerance, Scheduling
- Address critical issues of Grid Networking
  - High performance transport, QoS, measurement, distributed security
- Port and test applications
- Investigate innovative approaches
  - P2P resources discovery, Desktop Grids, active grids

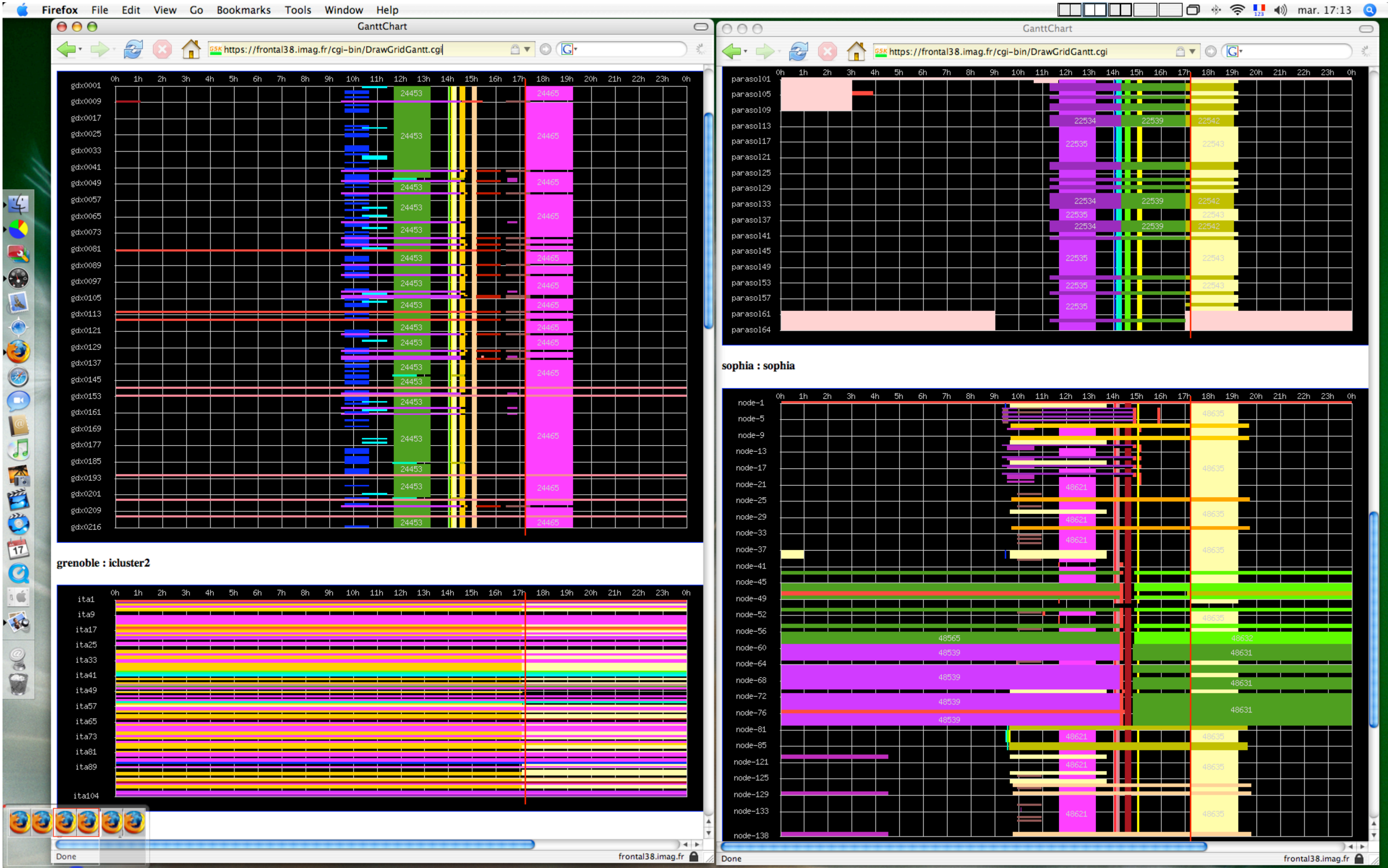
GRID5000 initiative



gdx0200(0, 216)
As reported at
2005/11/11 20:15:14

Idle
x86_64 Linux 2.6.12-
1-amd64-k8-smp
net I/O: 308.34
bytes/sec/11.79
bytes/sec

GRID5000 initiative


Firefox File Edit View Go Bookmarks Tools Window Help
mar. 17:13
GanttChart
https://frontal38.imag.fr/cgi-bin/DrawGridGantt.cgi
0h 1h 2h 3h 4h 5h 6h 7h 8h 9h 10h 11h 12h 13h 14h 15h 16h 17h 18h 19h 20h 21h 22h 23h 0h
gdx0001 gdx0009 gdx0017 gdx0025 gdx0033 gdx0041 gdx0049 gdx0057 gdx0065 gdx0073 gdx0081 gdx0089 gdx0097 gdx0105 gdx0113 gdx0121 gdx0129 gdx0137 gdx0145 gdx0153 gdx0161 gdx0169 gdx0177 gdx0185 gdx0193 gdx0201 gdx0209 gdx0216
24453
24465
grenoble : icluster2
ita1 ita9 ita17 ita25 ita33 ita41 ita49 ita57 ita65 ita73 ita81 ita89 ita104
Done
frontal38.imag.fr
GanttChart
https://frontal38.imag.fr/cgi-bin/DrawGridGantt.cgi
parasol01 parasol05 parasol09 parasol13 parasol17 parasol21 parasol25 parasol29 parasol33 parasol37 parasol41 parasol45 parasol49 parasol53 parasol57 parasol61 parasol64
22534 22535 22539 22542 22543
sophia : sophia
node-1 node-5 node-9 node-13 node-17 node-21 node-25 node-29 node-33 node-37 node-41 node-45 node-49 node-52 node-56 node-60 node-64 node-68 node-72 node-76 node-81 node-85 node-121 node-125 node-129 node-133 node-138
48635 48621 48565 48539 48632 48631
Done
frontal38.imag.fr

High speed transport studies: E2E performance

Study problems related to the transport **protocol & service**

- Compare alternatives for congestion detection and control : BIC, Fast, XCP, TFRC vs Reno,
- Study burstiness, effect of pacing
- Study flow scheduling, load balancing, caching alternatives…

Study problems related to the **protocol and network configuration**

- buffer space allocation, negotiation, bottleneck detection and estimation …

Study problems related to the **end systems & protocol implem.**:

- processor and bus speeds,
- NIC with its associated drivers, protocol implementation;
- memory access, zero-copy sockets, OS by-pass approaches

High performance network emulation

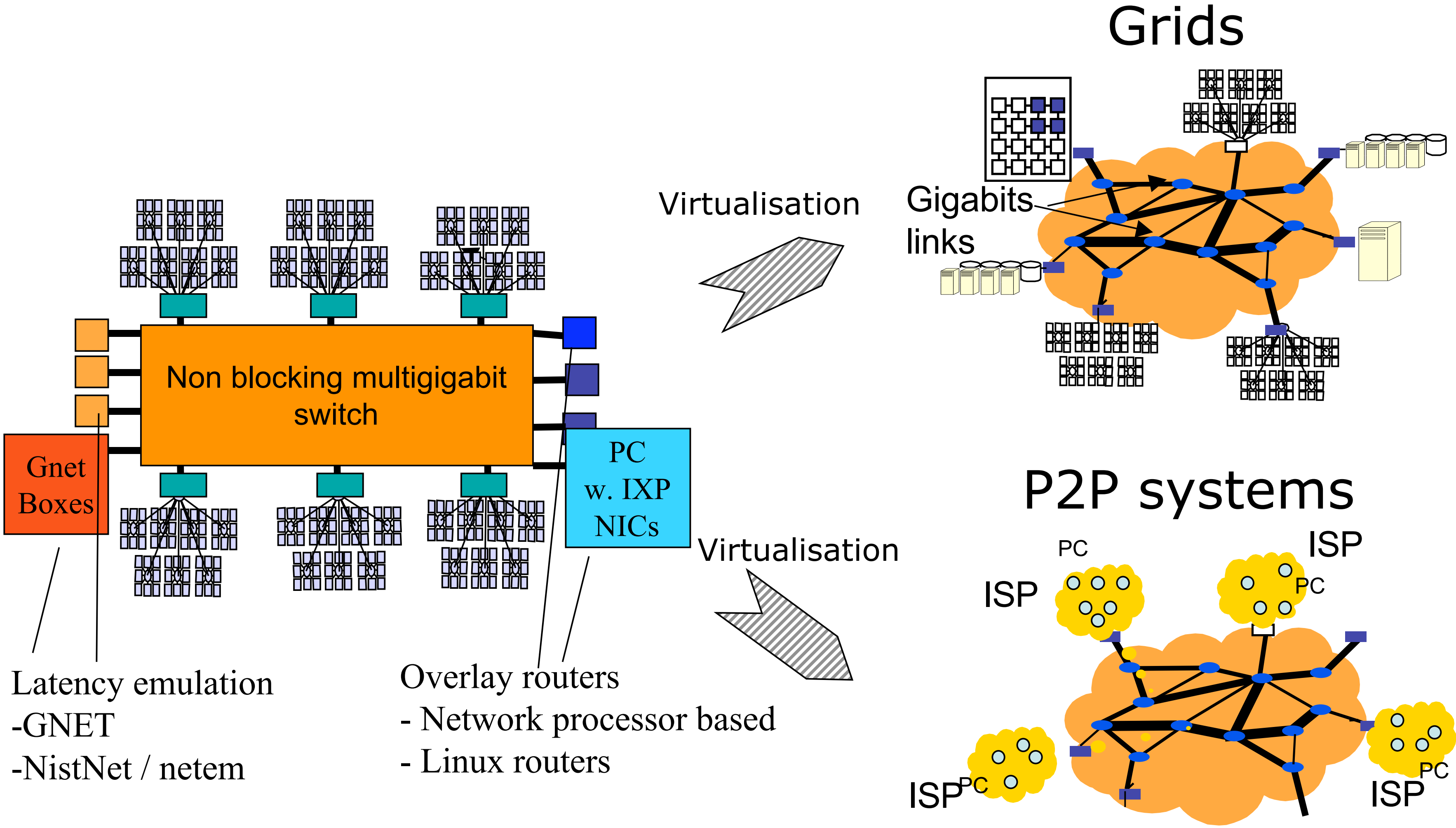
eWAN software has been developed for easily configure and control a wide area emulated network over a high performance cluster

Developed in the framework of the national GRID5000 and GdX projects.

Goals:

- Enabling large scale gigabit experiments
- For Grid protocol, middleware and application evaluation
- Easy to deploy on standard clusters

eWAN: large scale gigabit WAN emulator



Non blocking multigigabit switch

Gnet Boxes

PC w. IXP NICs

Virtualisation

Grids

Gigabits links

Virtualisation

P2P systems

ISP

PC

ISP

PC

ISP

PC

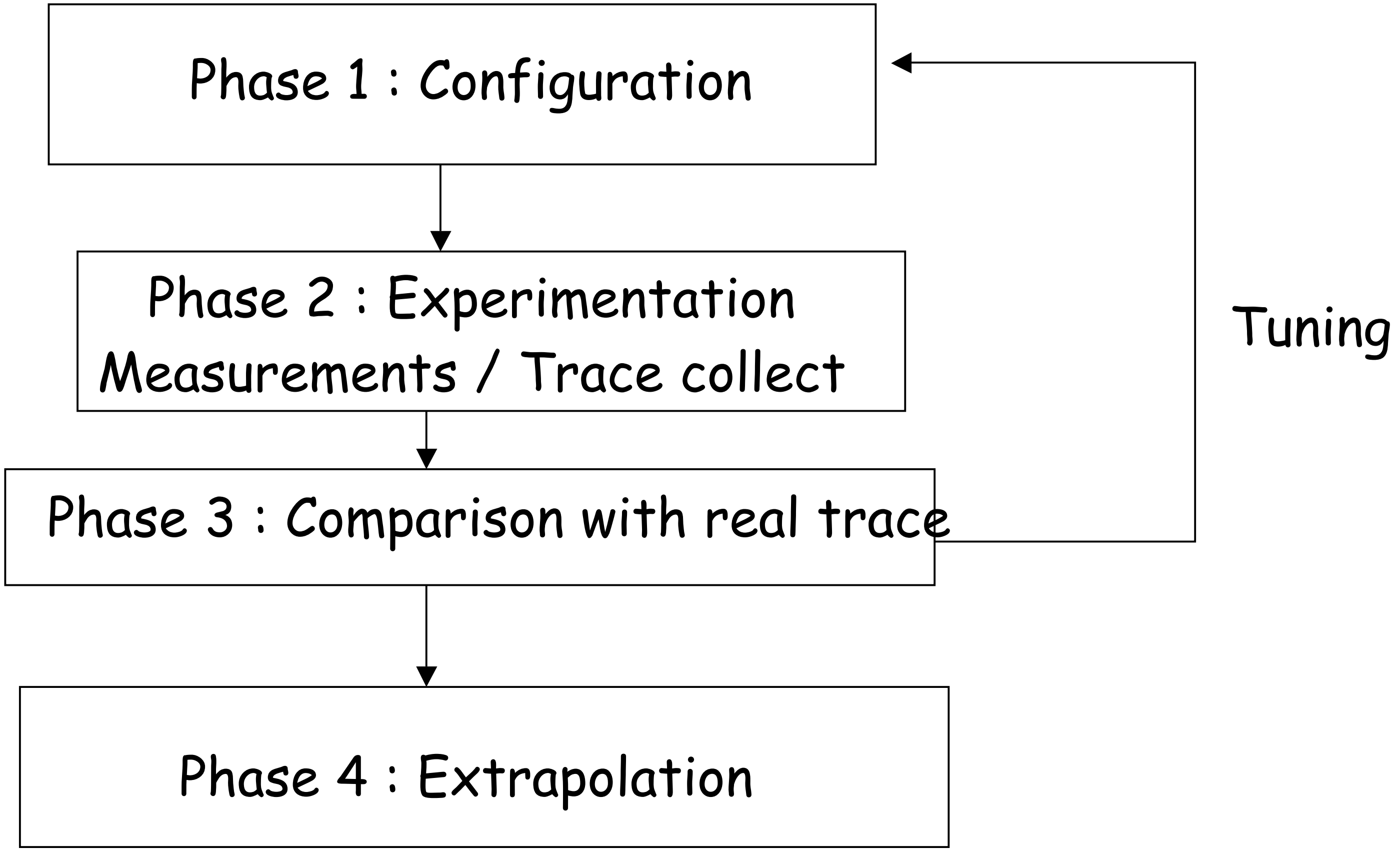
ISP

PC

Latency emulation
-GNET
-NistNet / netem

Overlay routers
- Network processor based
- Linux routers

Emulation loop



Phase 1 : Configuration

Phase 2 : Experimentation

Measurements / Trace collect

Phase 3 : Comparison with real trace

Phase 4 : Extrapolation

Tuning

eWAN Steps

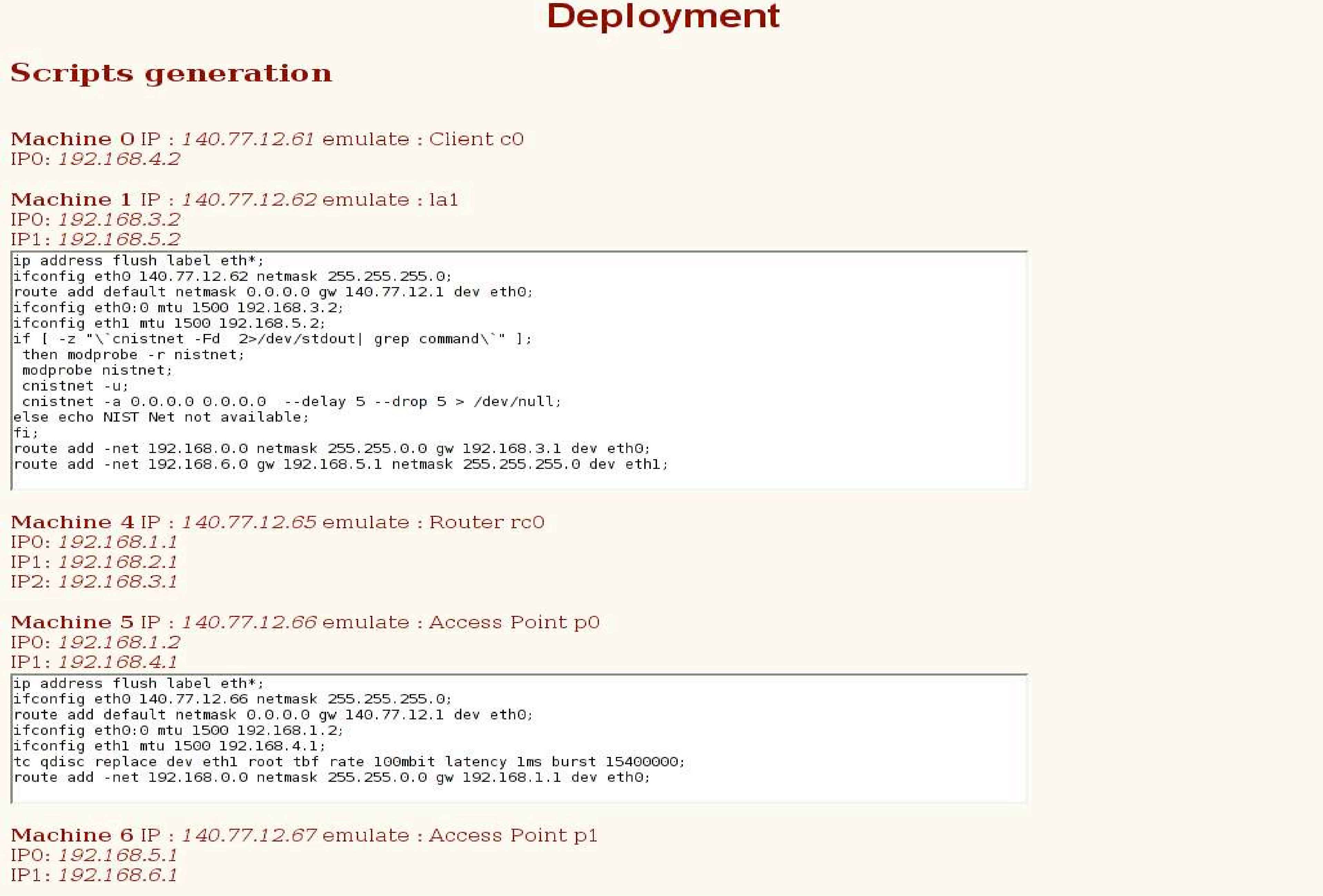
- Load or define a virtual topology
- Choose your emulation options
- eWAN generate scripts for automatic deployment on your own cluster
- Run your experiment on the emulated network environment and collect the logs.

eWAN Functionnalities

- Interface to define topology and save it as rdf/xml file.
    - User-friendly GUI.

- Deploy rdf topology
    - Configure all stations automatically.
    - Dry-run or deploy

Deployment



Scripts generation

Machine 0 IP : *140.77.12.61* emulate : Client c0
IP0: *192.168.4.2*

Machine 1 IP : *140.77.12.62* emulate : la1
IP0: *192.168.3.2*
IP1: *192.168.5.2*

```
ip address flush label eth*;
ifconfig eth0 140.77.12.62 netmask 255.255.255.0;
route add default netmask 0.0.0.0 gw 140.77.12.1 dev eth0;
ifconfig eth0:0 mtu 1500 192.168.3.2;
ifconfig eth1 mtu 1500 192.168.5.2;
if [ -z "\`cnistnet -Fd  2>/dev/stdout| grep command\`" ];
 then modprobe -r nistnet;
 modprobe nistnet;
 cnistnet -u;
 cnistnet -a 0.0.0.0 0.0.0.0  --delay 5 --drop 5 > /dev/null;
else echo NIST Net not available;
fi;
route add -net 192.168.0.0 netmask 255.255.0.0 gw 192.168.3.1 dev eth0;
route add -net 192.168.6.0 gw 192.168.5.1 netmask 255.255.255.0 dev eth1;
```

Machine 4 IP : *140.77.12.65* emulate : Router rc0
IP0: *192.168.1.1*
IP1: *192.168.2.1*
IP2: *192.168.3.1*

Machine 5 IP : *140.77.12.66* emulate : Access Point p0
IP0: *192.168.1.2*
IP1: *192.168.4.1*

```
ip address flush label eth*;
ifconfig eth0 140.77.12.66 netmask 255.255.255.0;
route add default netmask 0.0.0.0 gw 140.77.12.1 dev eth0;
ifconfig eth0:0 mtu 1500 192.168.1.2;
ifconfig eth1 mtu 1500 192.168.4.1;
tc qdisc replace dev eth1 root tbf rate 100mbit latency 1ms burst 15400000;
route add -net 192.168.0.0 netmask 255.255.0.0 gw 192.168.1.1 dev eth0;
```

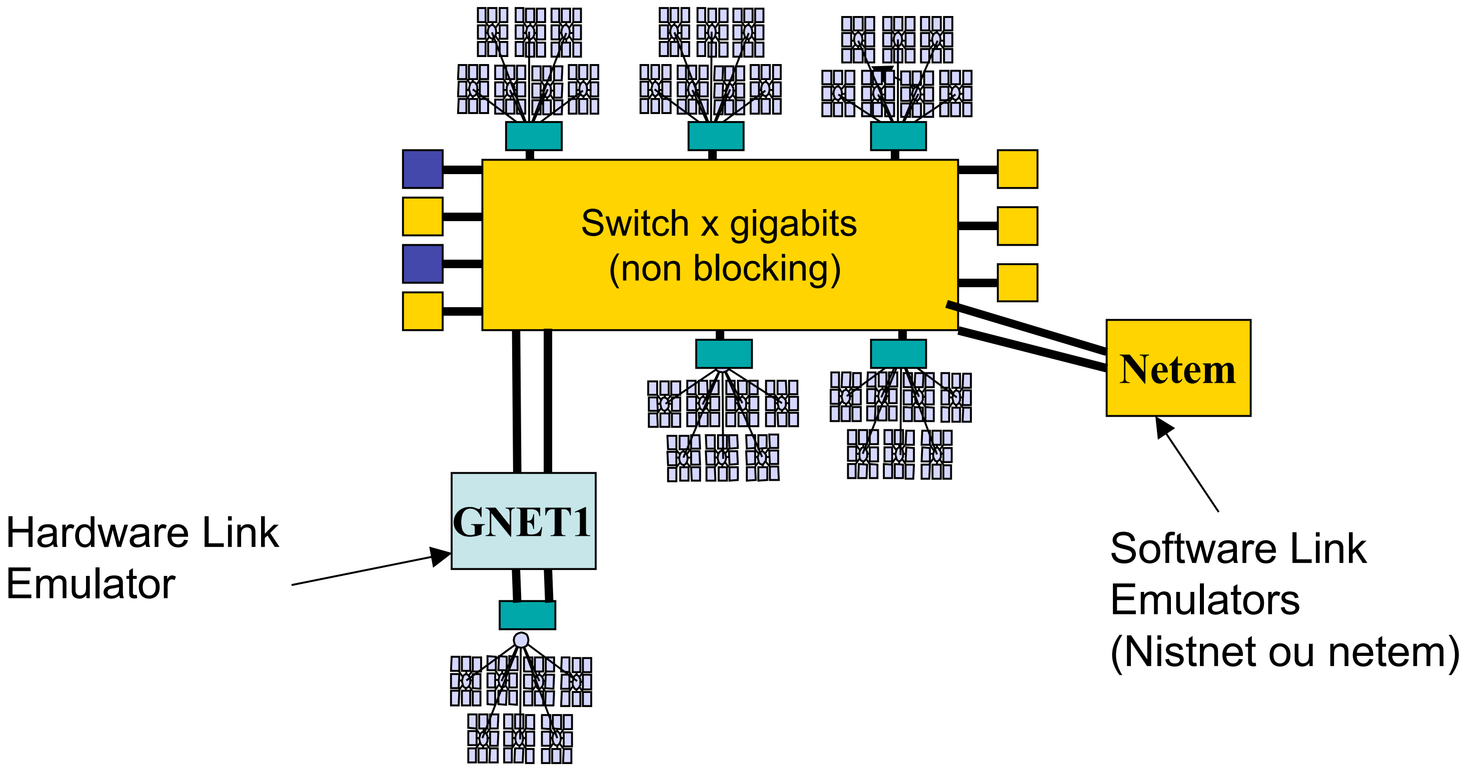
Machine 6 IP : *140.77.12.67* emulate : Access Point p1
IP0: *192.168.5.1*
IP1: *192.168.6.1*

High performance link emulation

- AIST GTRC group has developed a Gtrc-NET1 box that allows 1Gb/s and 10Gb/s long distance links emulation.
- based on a large scale FPGA, 4 high speed memory blocks and 4GbE ports.
- By configuring the circuits on the FPGA, various functions such as network emulation, traffic measurement, packet capturing and traffic generation can be achieved.
- GNET1 description and configuration modules have been integrated in the eWAN software.

- Several experiments have been conducted in collaboration with the AIST GTRC group.

eWAN + GNET1


Switch x gigabits
(non blocking)
Netem
GNET1
Hardware Link
Emulator
Software Link
Emulators
(Nistnet ou netem)

eWAN Calibration

Evaluate Hardware parameters:

- PCI bus, CPU speed, NICs, memory size, switch characteristics

Evaluate Software parameters:

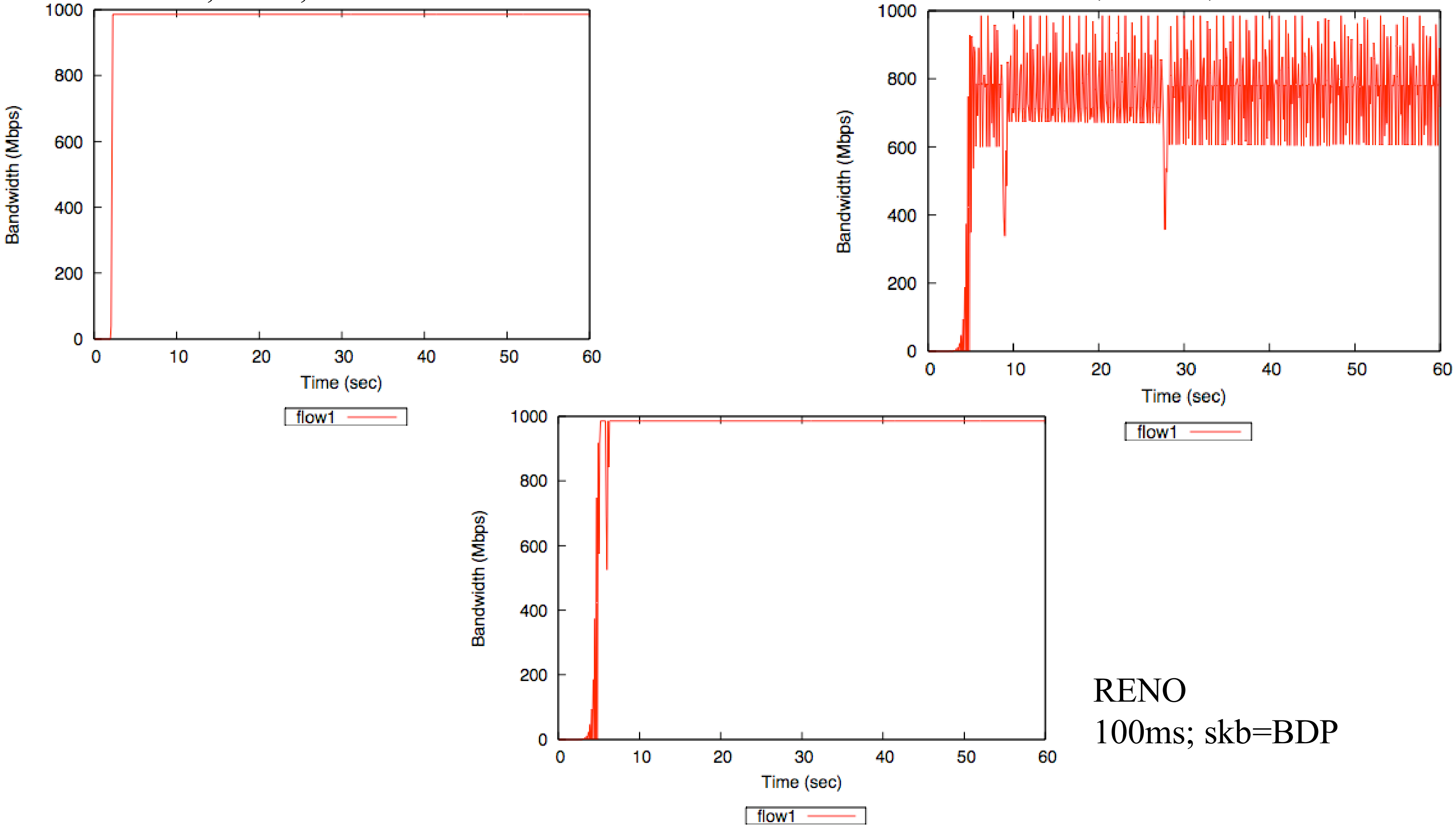
- Link emulation, traffic control, software routing, virtual interfaces

Verify you can get the line rate or the specified rate with one flow

If not,

- all your results will have to be scaled,
- may be also distored

Impact of the latency



RENO; 10ms; skb=BDP

RENO; 100ms; skb<BDP

RENO
100ms; skb=BDP

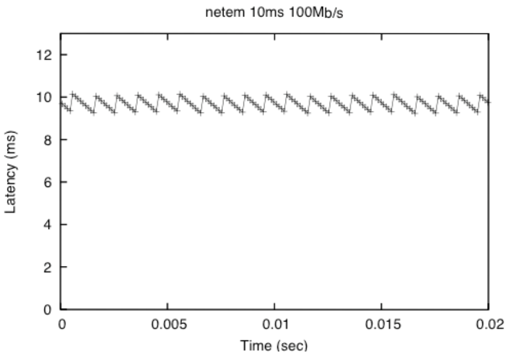
Trade-off between accuracy & scalability

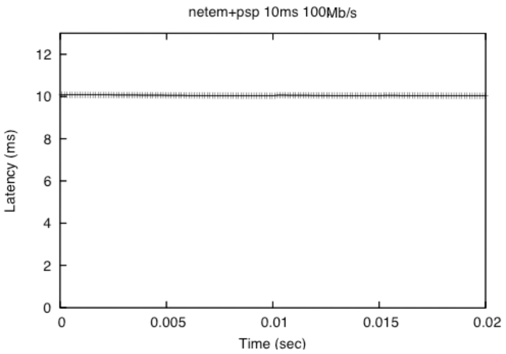
Software emulation: netem: easy to deploy but kernel timers are limited by the system time tick rate of 1000Hz (1ms) on Linux 2.6.

Hardware emulation: GNET: very accurate (1us)

eWAN uses netem linux module for large scale deployment in any cluster

- buffer management for delay emulation makes the traffic highly bursty.
- AIST pacing software PSP has been integrated to control the rate within the network emulator





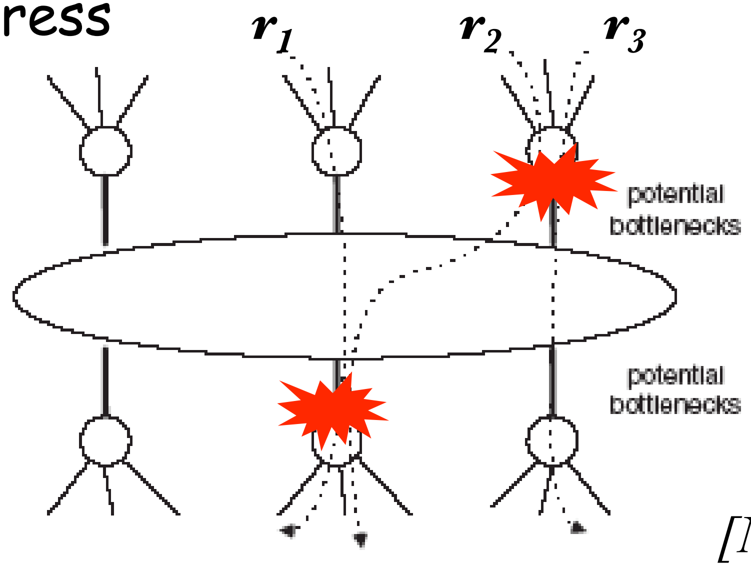
Application example: Bandwidth sharing in grids [globecom05]

Assumption: core is well /over provisionned

Access links & nodes links capacities are in same order (Gb/s)

Access links are potential bottlenecks

at Ingress & Egress



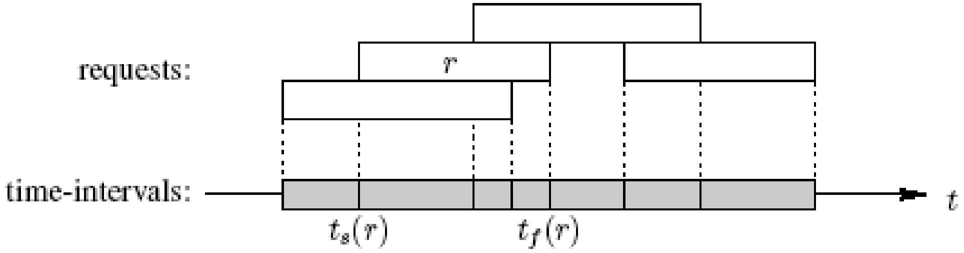
[NBP Albuquerque TON 04]

Bandwidth sharing in LFN

- INRIA RESO is proposing a Grid Network Service for controlling Bandwidth Sharing *(Globecom05).*
- The goal of the experiment on the grid emulmator is to verify the initial assumptions and to study the interaction between the session level algorithm and transport level protocol.
- The objectif is to compare transfer delays (average or minmax) when utilizing different protocol stacks and rate control mechanisms.

Flow scheduling heuristics

Notations



⇒ *Time window of request r : [ts (r), tf (r)]*

⇒ *Volume of data to send of request r : vol(r)*

⇒ *Schedule window of request r: [σ (r), τ (r)]*

⇒ *ts (r) < σ (r) < τ (r) < tf (r)*

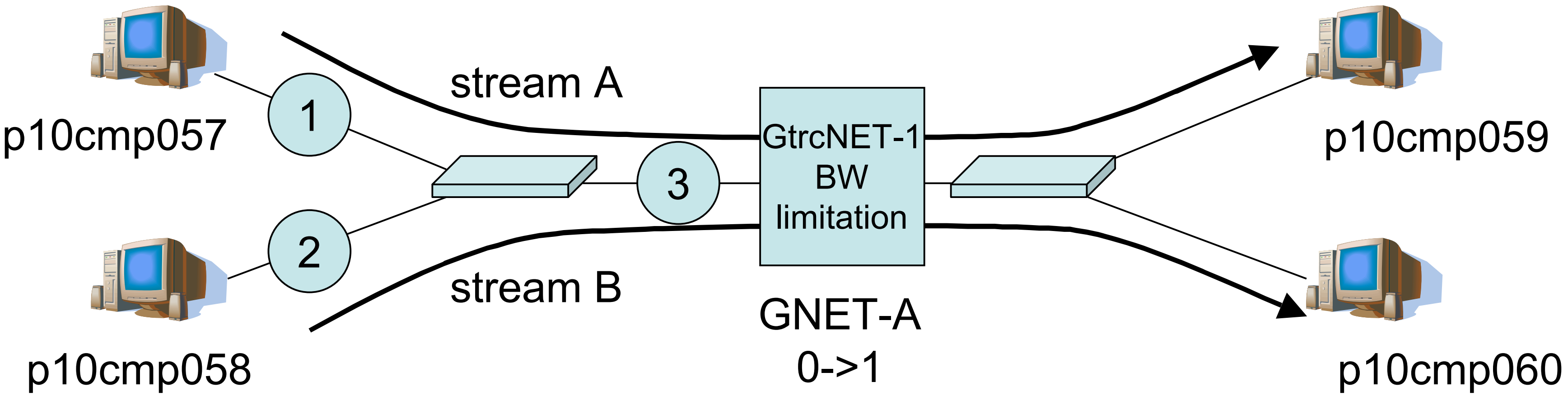
⇒ $\tau(t) = t + \frac{vol(r)}{bw(r)}$

Constraints

$$\forall t,\ \forall i \in \mathcal{I}, \sum_{\substack{r\in\mathcal{R},\ ingress(r)=i,\\ \sigma(r)\leqslant t<\tau(r)}} bw(r) \leqslant B_{in}(i)$$

$$\forall t,\ \forall e \in \mathcal{E}, \sum_{\substack{r\in\mathcal{R},\ egress(r)=e,\\ \sigma(r)\leqslant t<\tau(r)}} bw(r) \leqslant B_{out}(e) \quad (2)$$

Experimentation setup & parameters



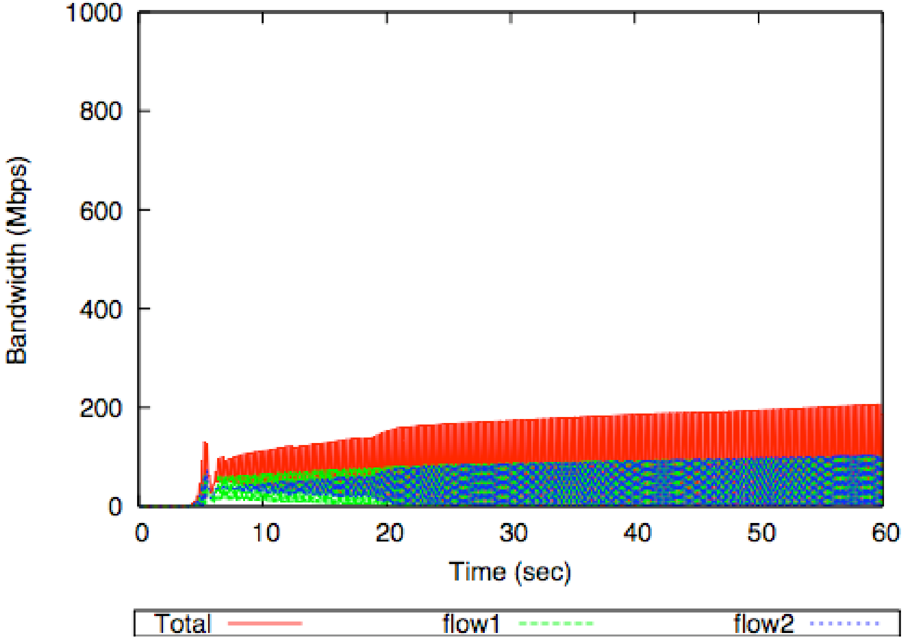
Hardware setups

Place	Setup A	Setup B	Setup C
1	GNET-B 0->1	netem p10cmp061	none
2	GNET-B 2->3	netem p10cmp062	none
3	none	none	netem p10cmp061

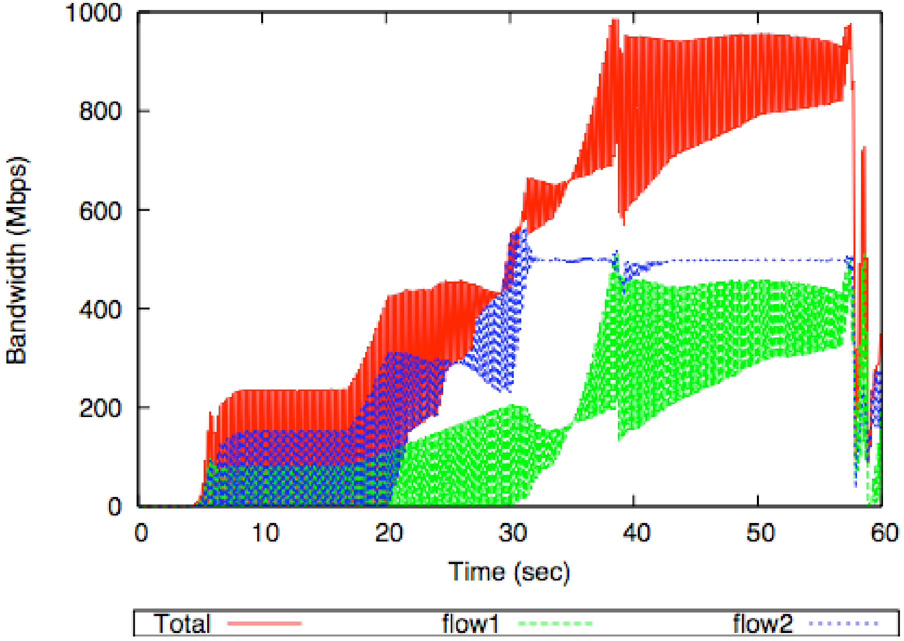
Software parameters

Tx. BW Ctrl.	Delay	TCP stack	BW limit. rate.
none	1ms	BIC	1G
TB	10ms	HS	100M
PSPacer	100ms	Scalable	10M

Flow interaction: Impact of the protocol



Reno



Bic

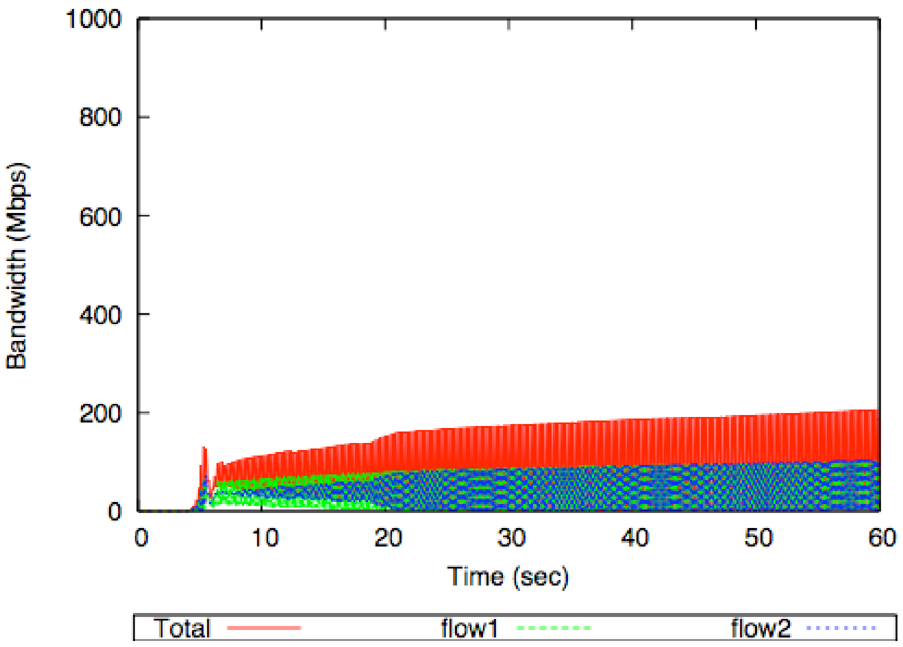
Schedule two TCP sources from independent machines

Rate limited at 490Mb/s - Bottleneck of 1Gb/s

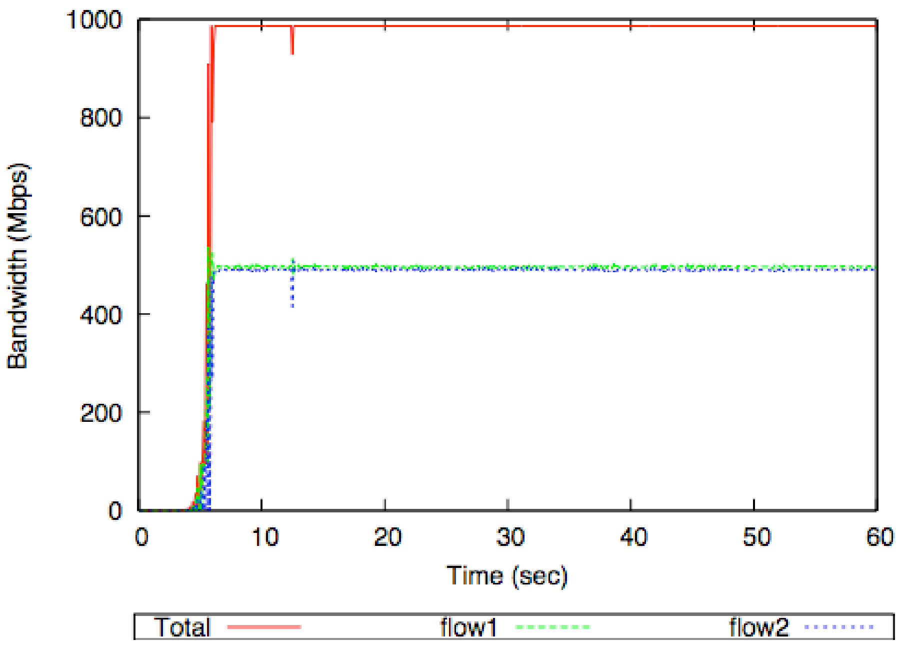
Flow interaction: Impact of the sources

Two TCP Reno sources limited by tbf to 490Mb/s

Bottleneck of 1Gb/s



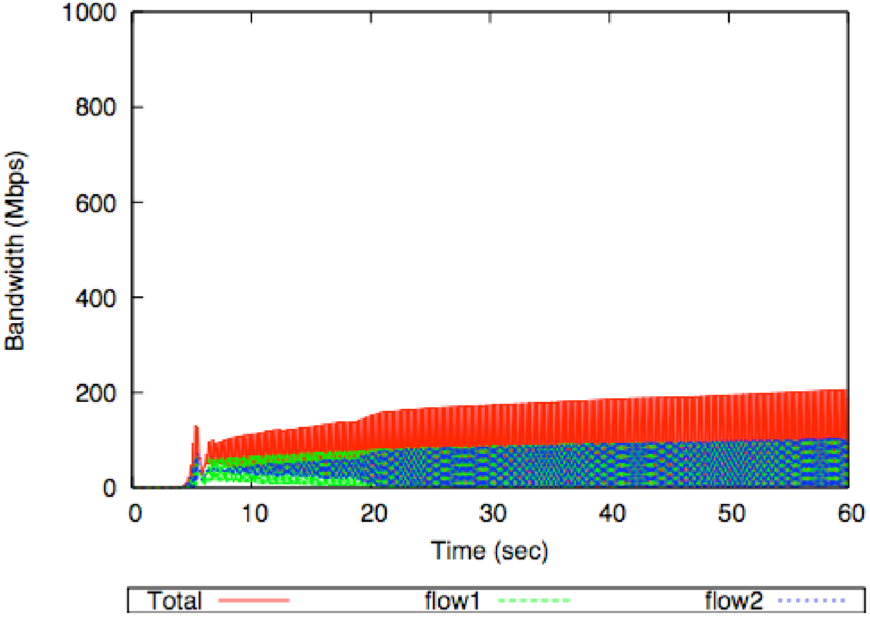
Independent machines



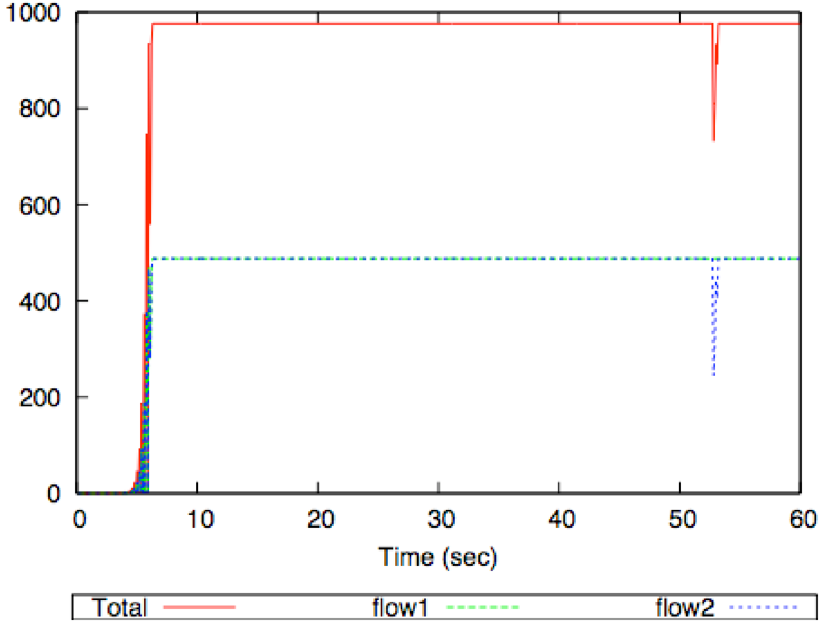
Same machine

Impact of rate limitation tool

Two TCP Reno sources from independent machines
limited 490Mb/s - Bottleneck of 1Gb/s

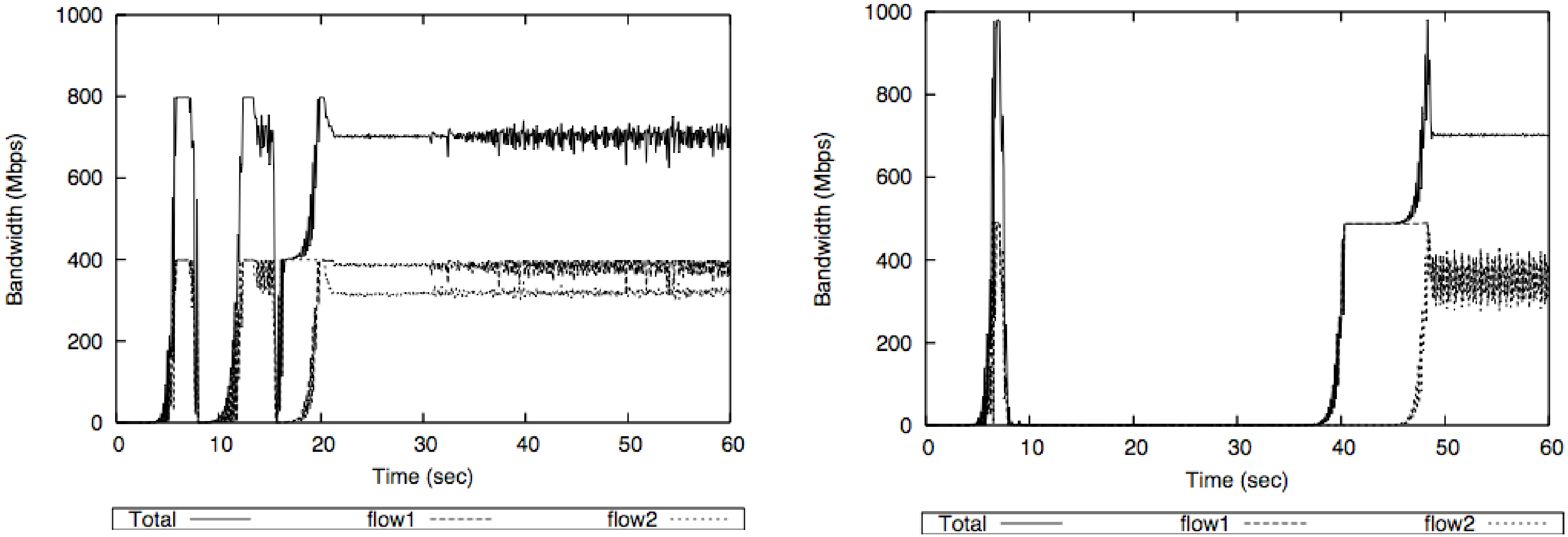


Token bucket filter



Precise Software Pacer

Impact of the congestion level



Two TCP BIC sources limited by PSPacer sharing a bottleneck of 700Mb/s
Congestion level = 10% (left) = 20% (right)
Sources rate = 400Mb/s (left) ; Sources rate = 490 MB/s (right)

Conclusion

We have presented an easy to use software to configure and run large scale experiments on an emulated grid.

To achieve high rates, high performance computers must be used

Hardware solution like GNET1 are more accurate and give better performance.

The experiments show that:

- Calibration and tuning are of great importance for diagnosis and debugging.
- token bucket filter is not an ideal tool for rate limitation at very high speed as it creates burstiness and interferes with TCP. PSPacer works much better
- Interaction of independent large TCP flows can be disastrous even if the congestion level is small
- Statistical bandwidth reservation is not sufficient, traffic has to be shaped
- A good combination of L5, L4 and L2 mechanisms seems to be required.

Perspectives

We continue to develop eWAN and add new functionalities

- monitoring tools, jumbo frames support, realistic topologies generator…

New experiment plan:

- Continue systematic exploration of flow interaction problem
- Study other protocols (westwood, fast…) and services (MPI, Gridftp…)
- Study heterogeneous protocols interaction
- Impact of reverse traffic
- 10Gb/s emulation

Contacts: Pascale.Primet@inria.fr & t.kudoh@aist.go.jp

Related works

Modelnet

- Network emulation / node virtualization

- Emulab
  - Network emulation / nodes simulation
- PlanetLab
  - Large scale distributed architecture (1000 nodes)
  - Real condition
  - Reproductability ?

2 types of emulation

distance (latency, bandwidth, failure rate …)

⇒ dummynet based solution

⇒ Difficulty : high number of configurations (High speed network, VTHD, Internet)

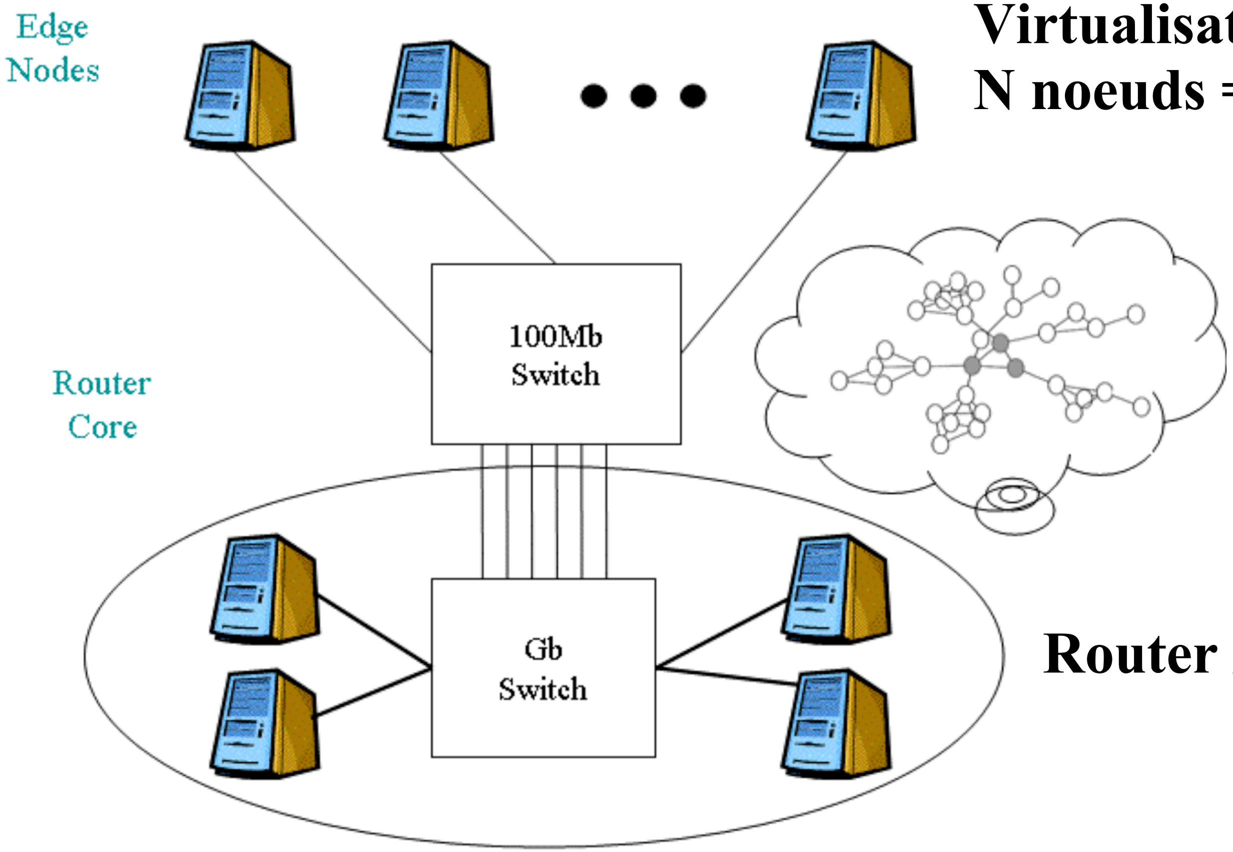
Number of nodes

⇒ Partitioning resources of a node

⇒ Difficulty: side effect

Modelnet

Http://issg.cs.duke.edu/modelnet.html



Virtualisation
N noeuds => M noeuds
N>>M

Router / Dummynet