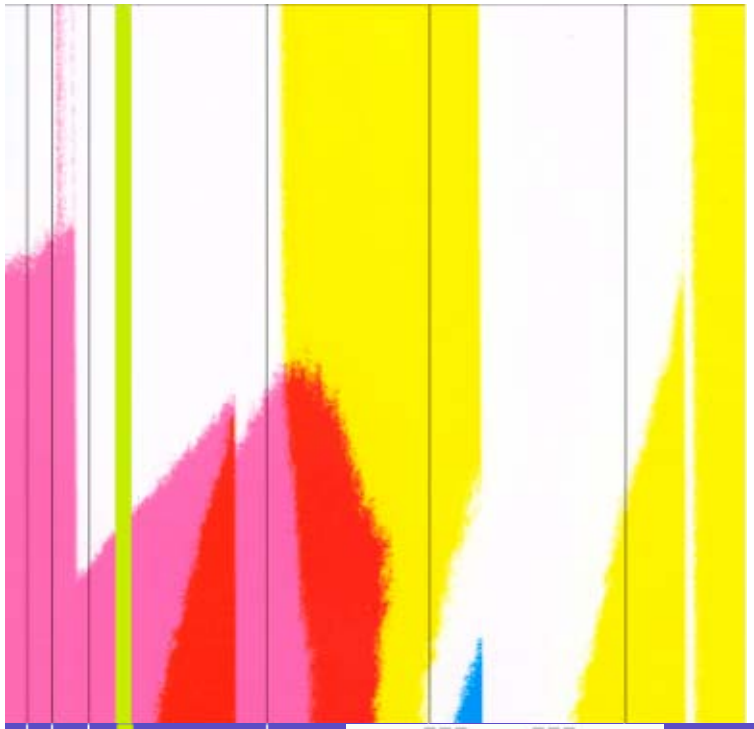


Designing High Performance Autonomic Gateways for Large Scale Grids and Distributed Environments

Laurent Lefèvre

INRIA / LIP

École Normale Supérieure de Lyon, France

laurent.lefevre@inria.fr

CCGSC2006, Flat Rock, North Carolina, Sept. 2006

Outline

Needs and challenges for autonomic gateways in large scale grids

Scenario 1 : Autonomic gateways in industrial context

Scenario 2 : Inter-planetary Grids

Conclusion and future works

Grid applications from the network view

It is difficult to clearly define what is a grid application :

- depends on people you are speaking with

- depends on type of grids (data grid, computing grids, P2P grids, mobile grids)

- depends on protocols/API/environments (MPI, java, corba, Web Services…)

- Need an application grid view and understanding in terms of network

How is used the network ?

Understand more :

- Communications frequency (bursts…)
- Aggregation on shared links/equipments
- Bottleneck effects
- Message patterns

•Network Topology ?

•Sharing of infrastructure with others applications ?

•Impacts of network usage on scalability ?

•How to design network-aware applications ? Usage of network services ?

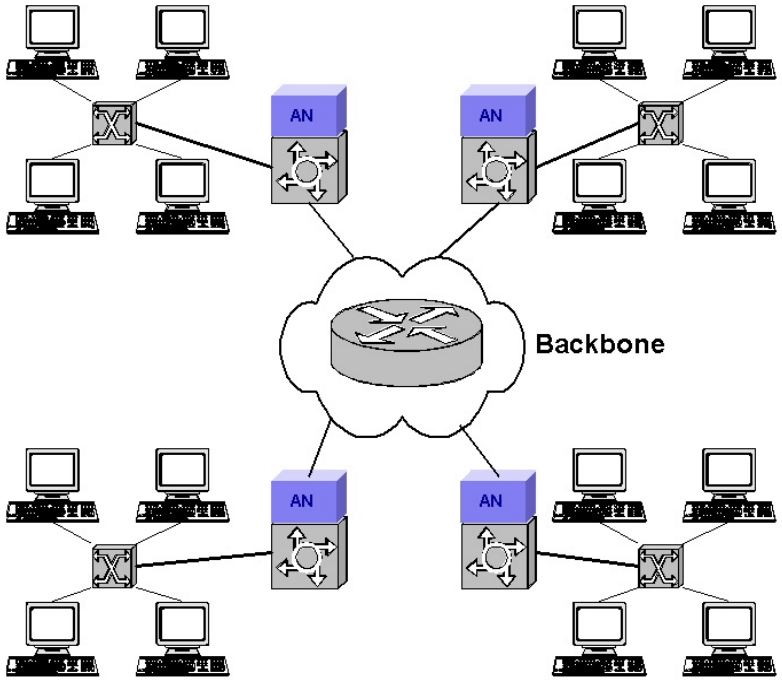
•How my middleware impacts the network ?

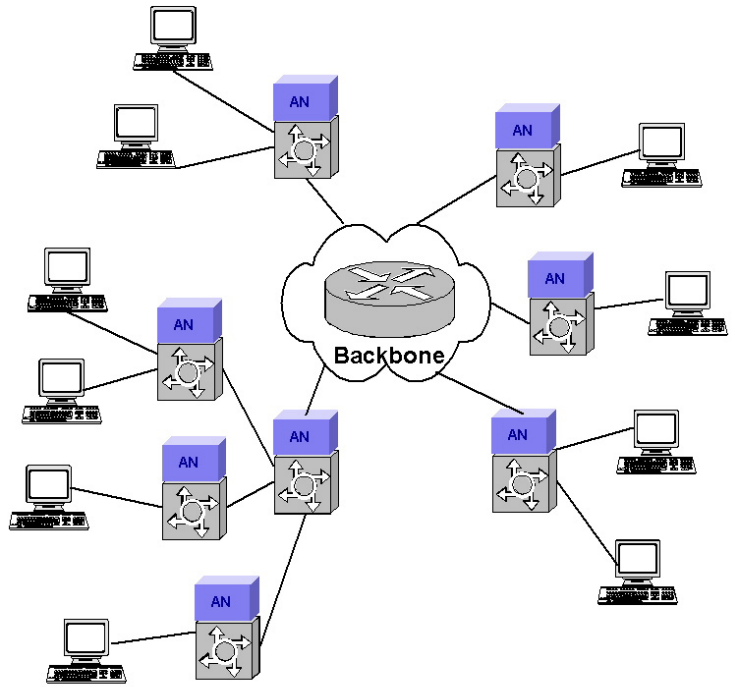
-How to give pertinent information to users ?

Active Grids : improving network usage with new dynamic services

- Exposing network capabilities to Grid middleware
- Support of multi-clusters / P2P Grids with active routers
- Example of services : Reliable Multicast, QoS, service deployment, compression, video adaptation,...
- Services deployed on demand : not enough

- [J.P. Gelas, L.Lefèvre et al. « Designing and evaluating an active grid architecture », FGCS, Feb. 2005]





Need for new services and equipments

Gateway located on strategic locations

Data path

Embedded services :

- Filtering data
- Monitoring / collecting
- Re-injecting

•Context aware equipments

Propositions : Autonomic Networking : “When human intervention is not possible...”

Derived from “Autonomic Computing” (IBM)

Dynamic service deployment

Self-*

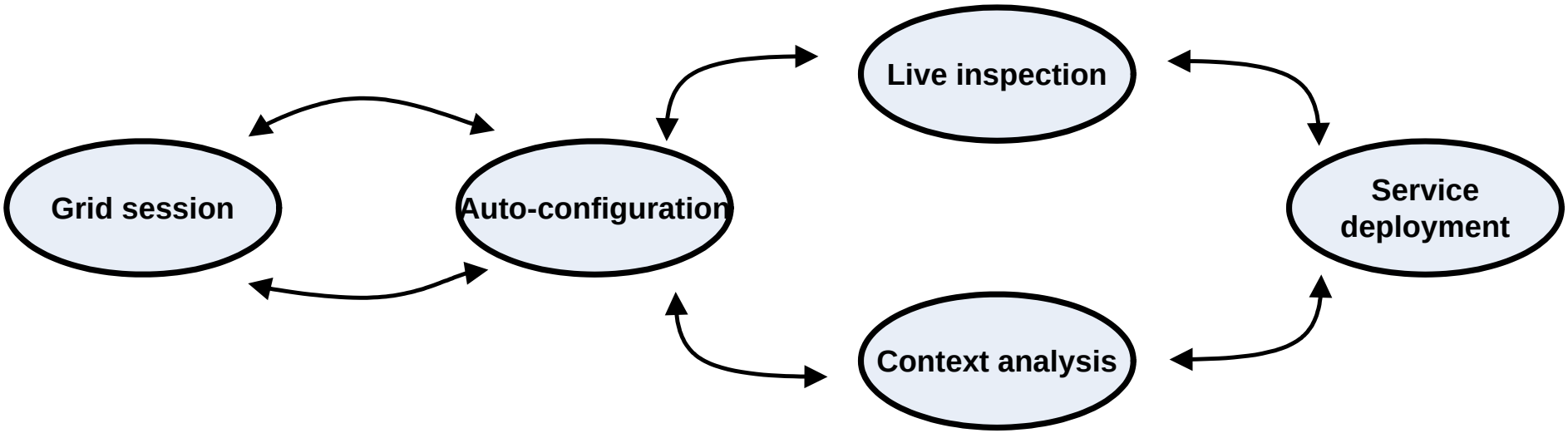
- self-managing
- self-configuring
- self-optimizing
- self-protecting
- self-healing/repairing
- ...

Proposing : Autonomic Programmable Network Gateways which measure / monitor network activity, collect and provide network information to schedulers and users (visualization)

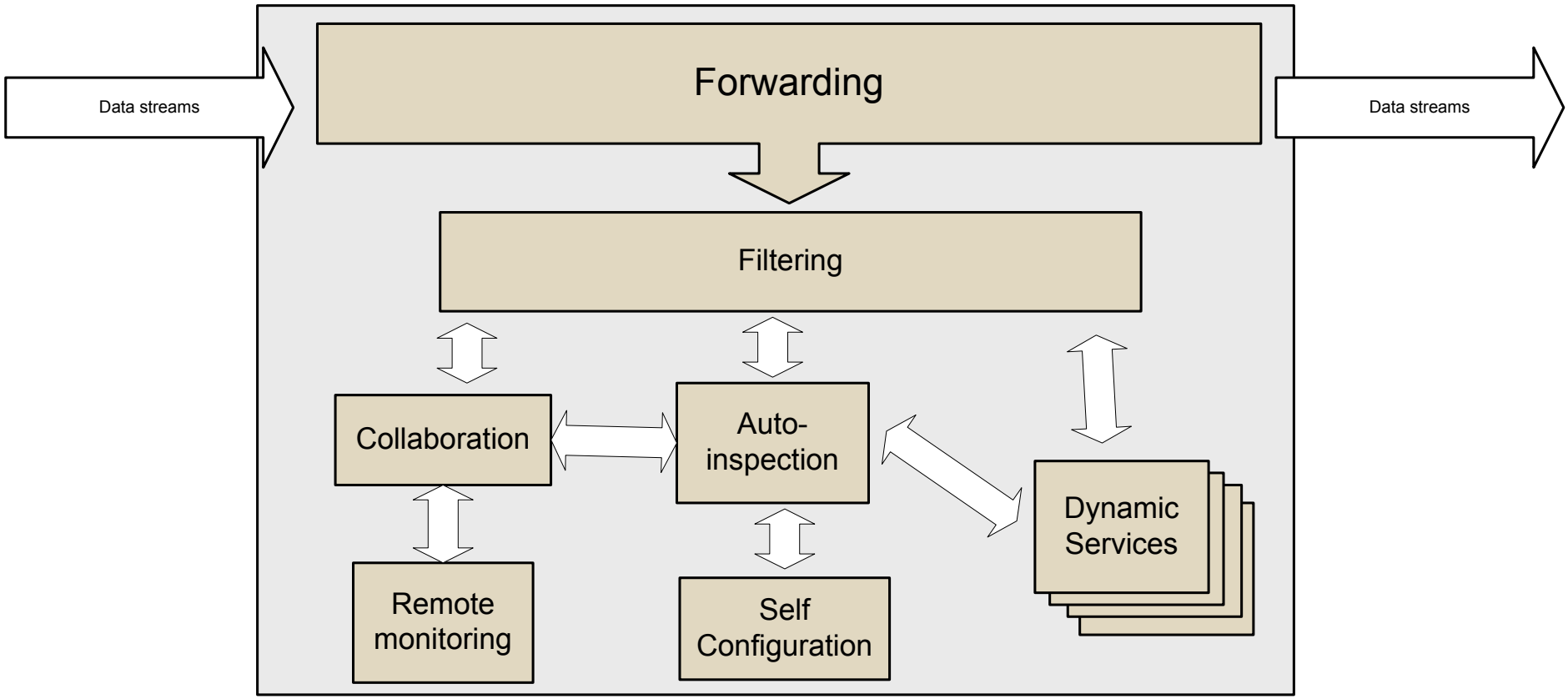
- Without human : not possible (IPG, industrial deployment), not wanted (large scale Grids and environments)

Supporting Grid sessions

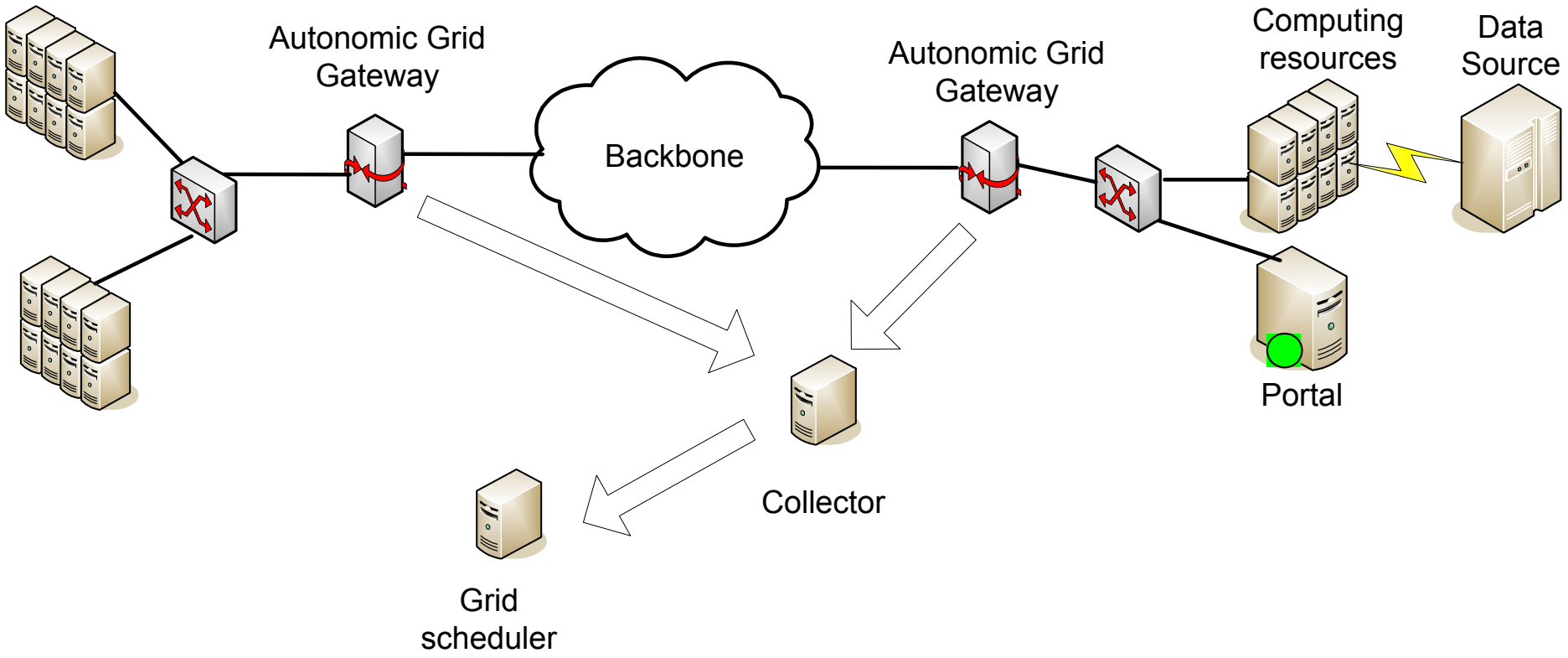
- Focusing on Grid sessions : run multiple times same applications on the Grid
- Monitoring and data collection



Architecture : Autonomic Gateway

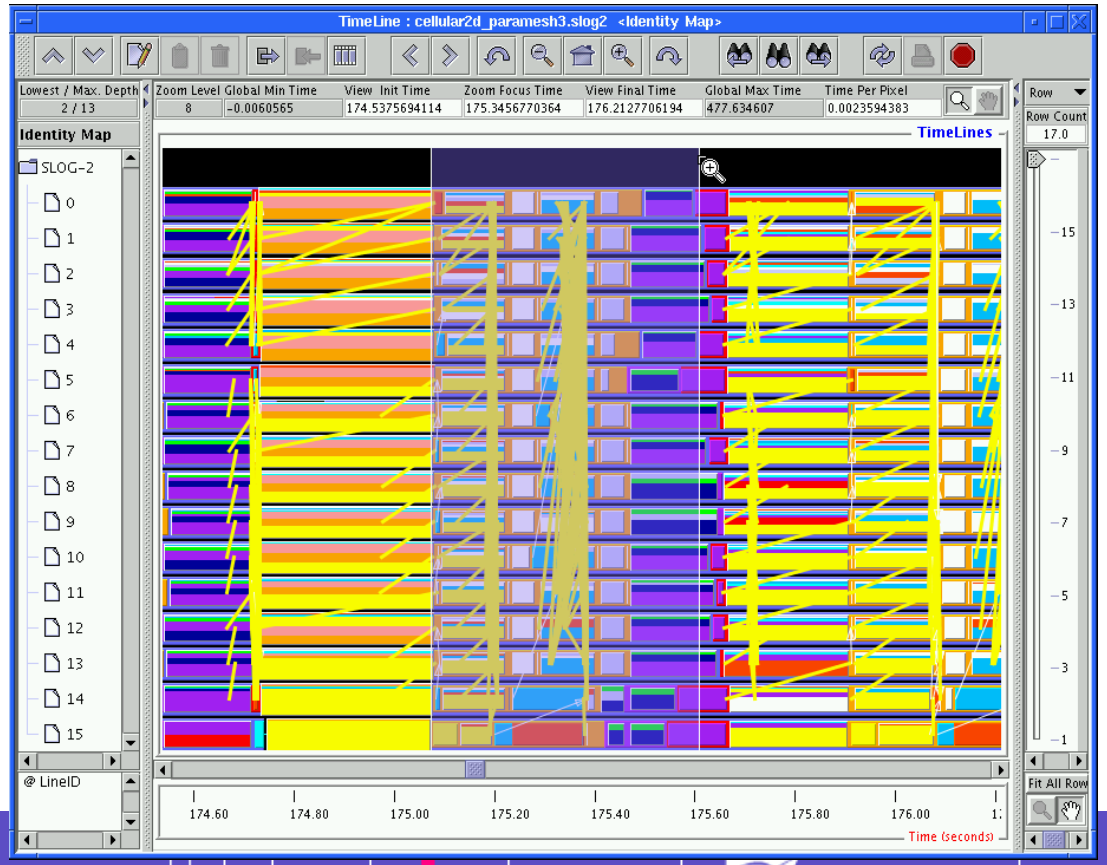


Deployment / infrastructure



Grid visualization

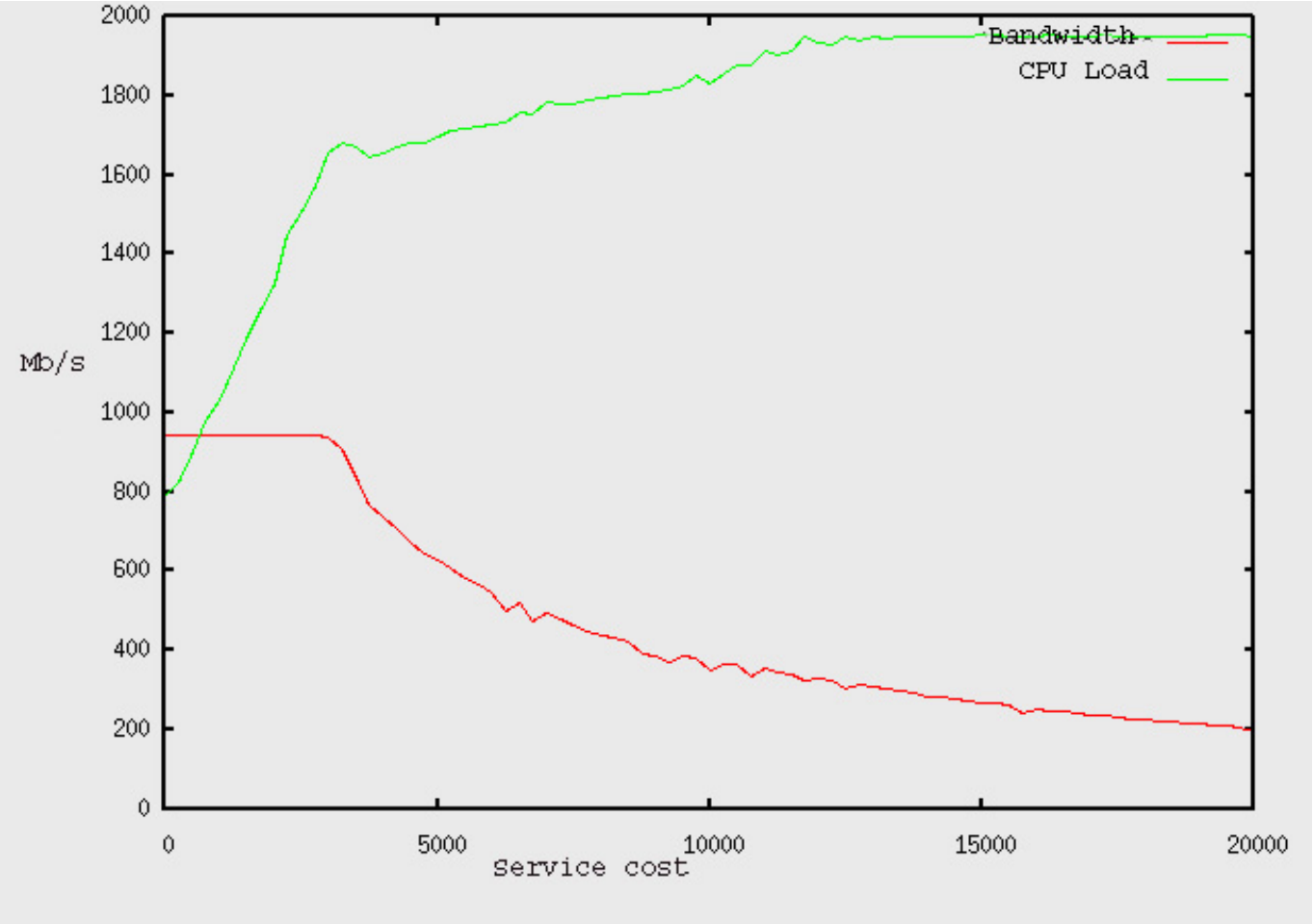
- Understand more and visualize grid sessions in terms of network usage
- Detecting networking problems



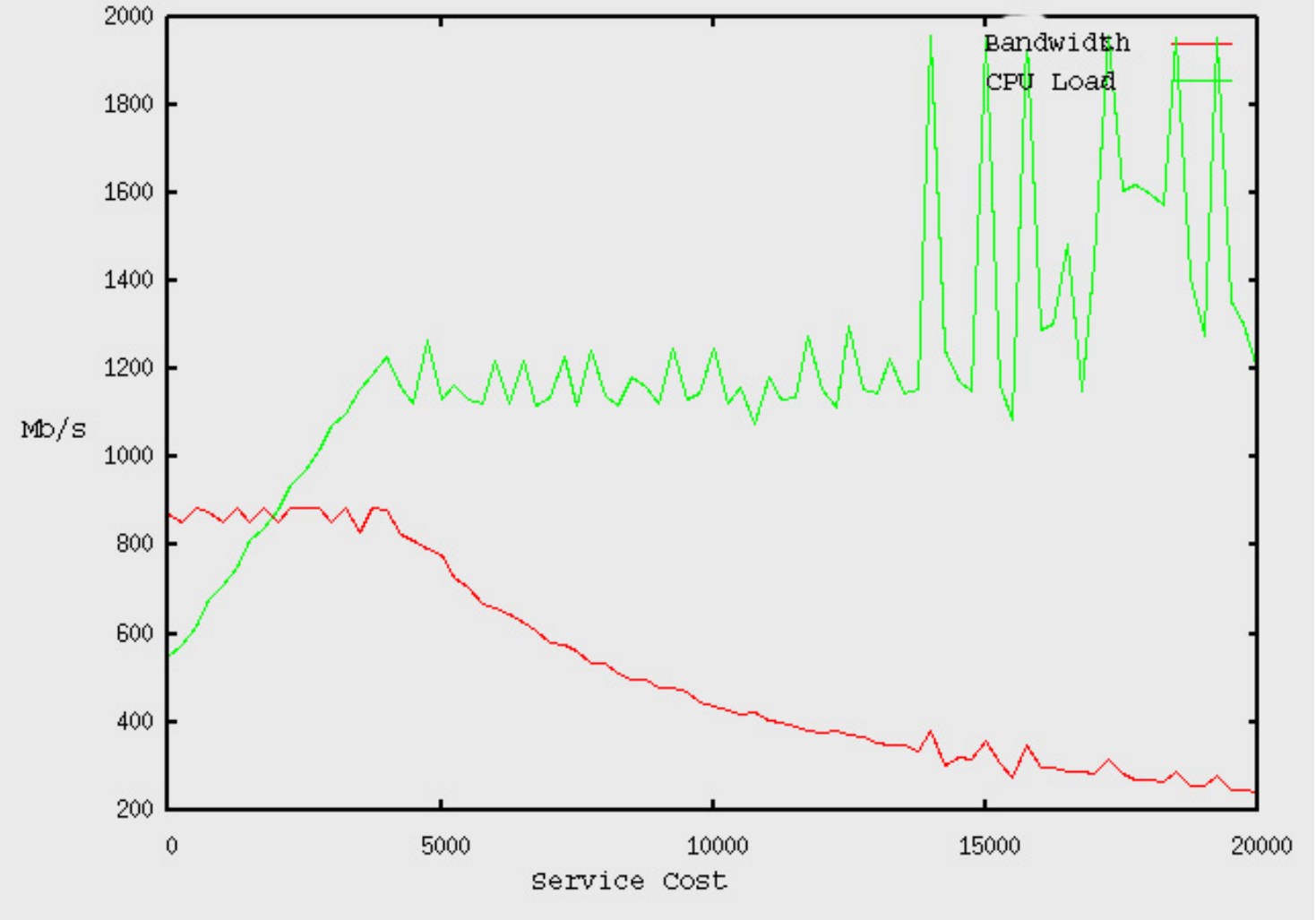
TCP

Bi PIII 1.4 Ghz gateways

GEthernet NICs

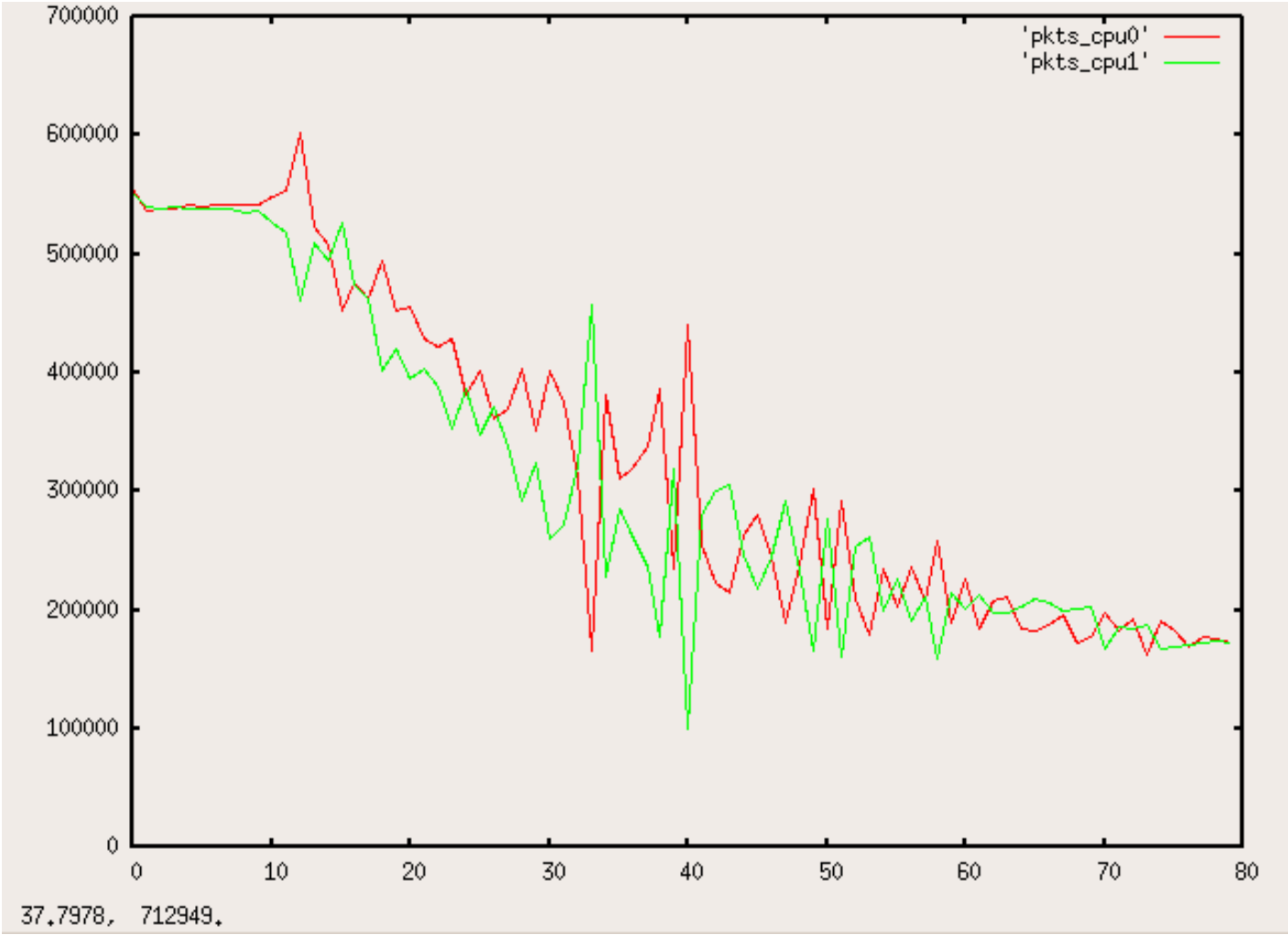


UDP


Bandwidth
CPU Load
Mb/s
2000
1800
1600
1400
1200
1000
800
600
400
200
0
5000
10000
15000
20000
Service Cost

Load balancing between CPUs

TCP



Challenges

- Limit impact/intrusion on data transfers (lightweight services, autonomic adaptive filtering)
- Increase context awareness

Scenario 1 : Industrial autonomic gateway (RNRT TEMIC project)

Scenario requirements

Easily and efficiently deployable hardware in industrial context : Enterprise Grid

Easily removable at the end of the maintenance and monitoring contract.

Devices must fit industrial requirements:

- reliability
- fault-tolerance

Devices must be *autonomic*!

- auto-configurable
- re-programmable

Our approach

Designing an Industrial Autonomic Network Node (IAN^2):

- Using a reliable and embedded hardware
- Running on a low resource consumption node OS
- Proposing an adapted EE
- Designing a set of services
- Evaluating solution in controled and industrial scenario

bearstech

Hardware / Node OS

A transportable solution.

Reduced risk of failure:

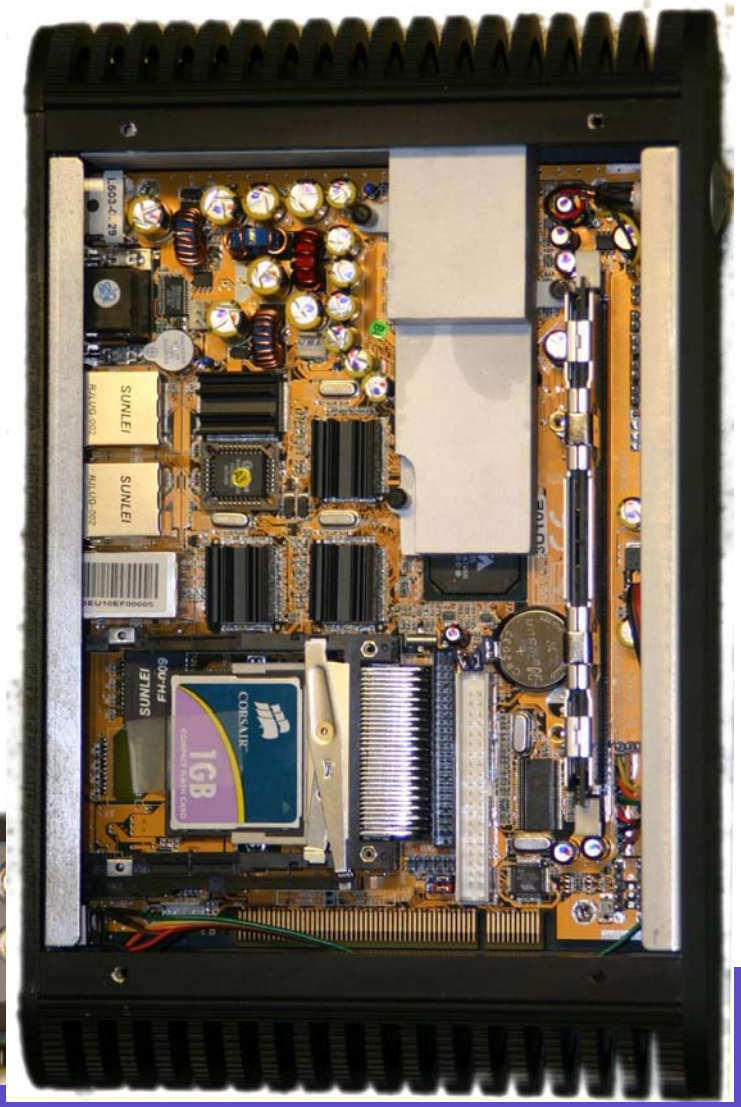
- fanless
- no mechanical hard disk drive

VIA C3 1GHz, 256MB RAM, 3xNIC Gbit Ethernet, 1GB Compact Flash,...

Indutrial Autonomic Network Node (IAN^2) runs over Btux (bearstech.com)

Btux is based on a GNU/Linux OS

- rebuilt from scratch
- small memory footprint
- reduced command set available
- remotely upgradeable

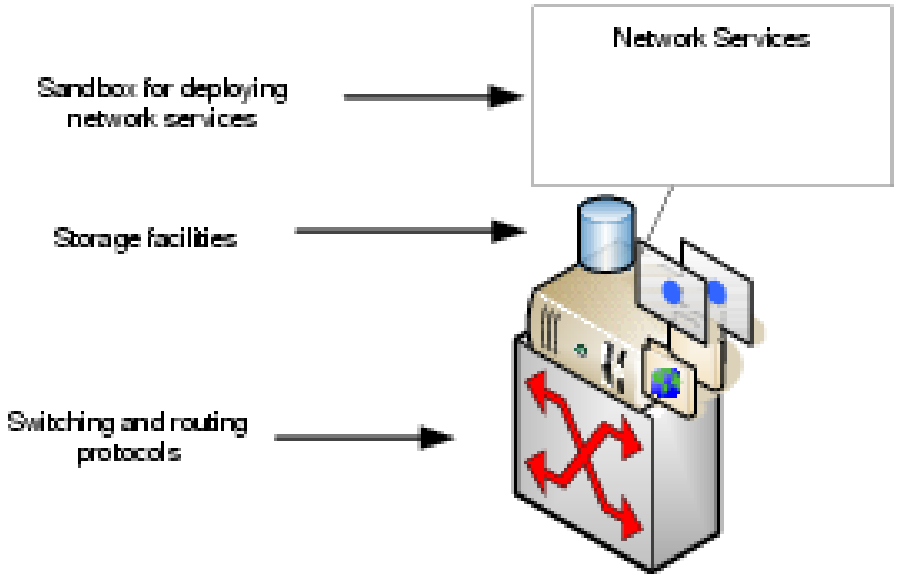


Software Execution Environment:

IAN^2 Software Architecture

Our Industrial Autonomic Nework Node architecture supports:

- wired and wireless connections,
- CPU facility,
- Limited storage capabilities.



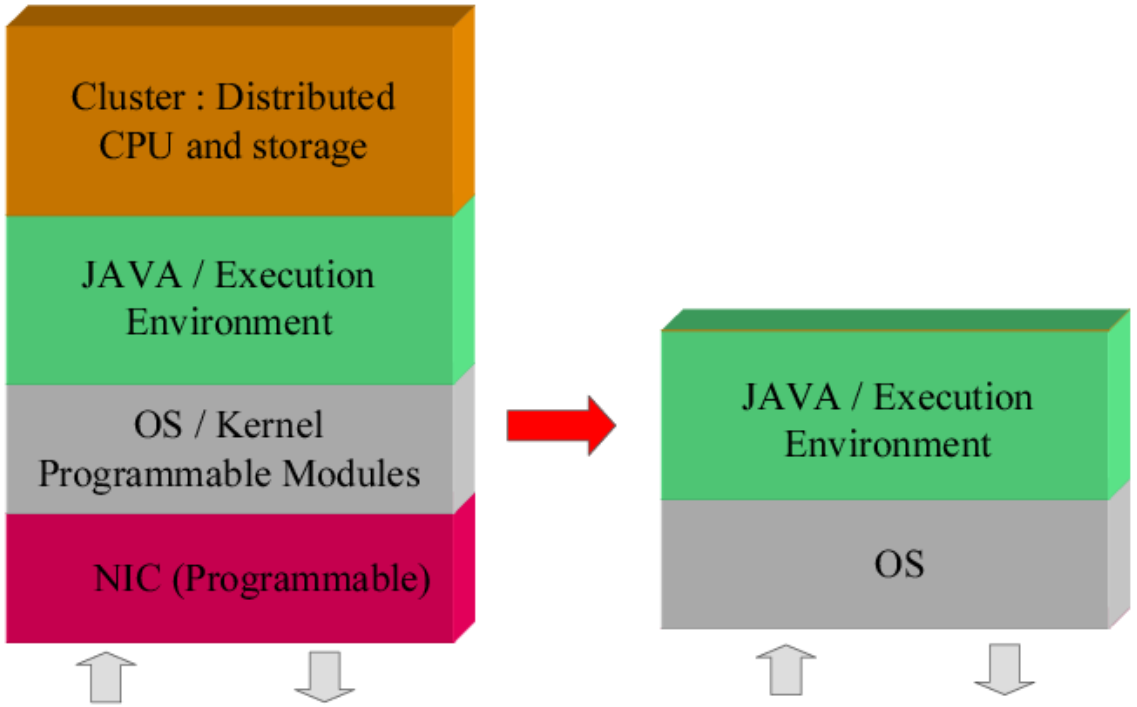
Sofatware Execution Environment

The EE is based on the *Tamanoir (INRIA)* software suite, a high performance execution environment for active networks.

Tamanoir: Too complex for industrial purpose.

Tamanoir^{embedded}:

- reduced code complexity,
- removed unused class and methods,
- simplify service design.



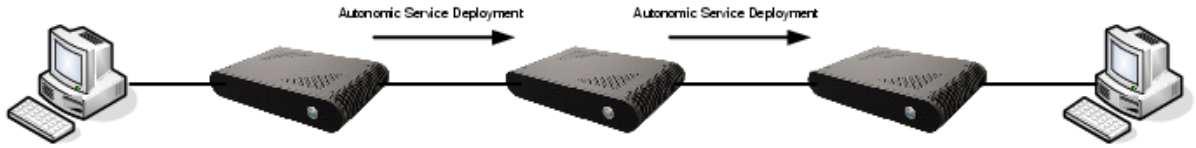
Software Execution Environment:

Autonomic Service Deployment

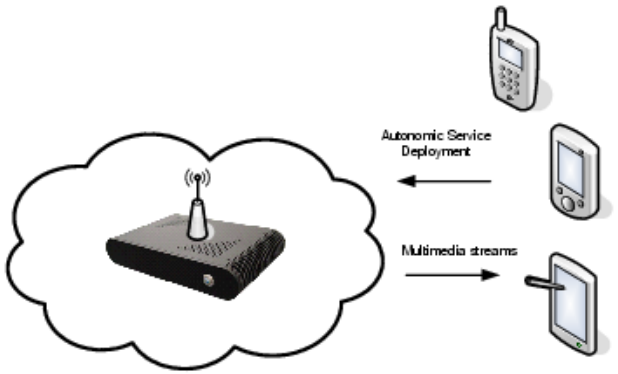
Tamanoir^{embedded} is written in Java and suitable for heterogeneous services.

Provides various methods for dynamic service deployment/update:

- ***from a service repository to a Tamanoir Active Node (TAN),***
- ***from the previous TAN crossed by the active data stream,***



- ***from mobile equipments.***

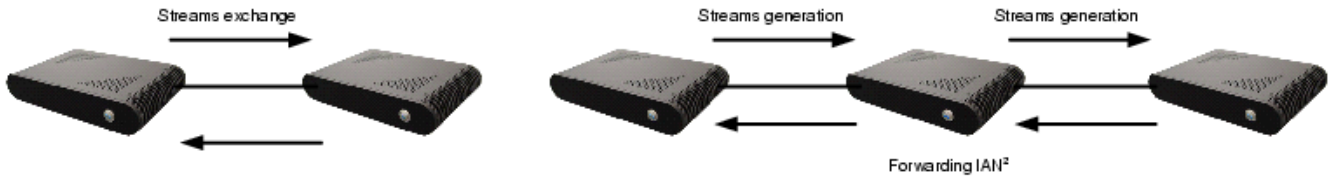


Experimental Evaluation:

Network Performances

Based on *iperf* (bandwidth, jitter, loss) on two topologies.

IAN^2 failed to obtain a full Gbit bandwidth due to the limited embedded CPU and chipset.



Configuration	Throughput	cpu send	cpu recv	cpu gateway
back-2-back	488 Mbps	90%	95%	N/A
gateway (1 stream)	195 Mbps	29%	28%	50%
gateway (8 streams)	278 Mbps	99%	65%	70%

Experimental Evaluation:

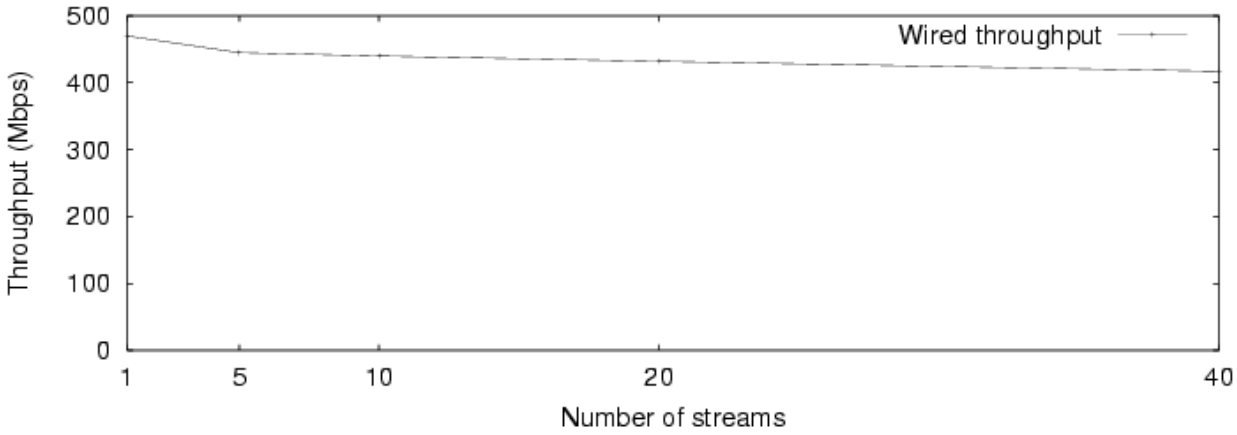
Network Performances

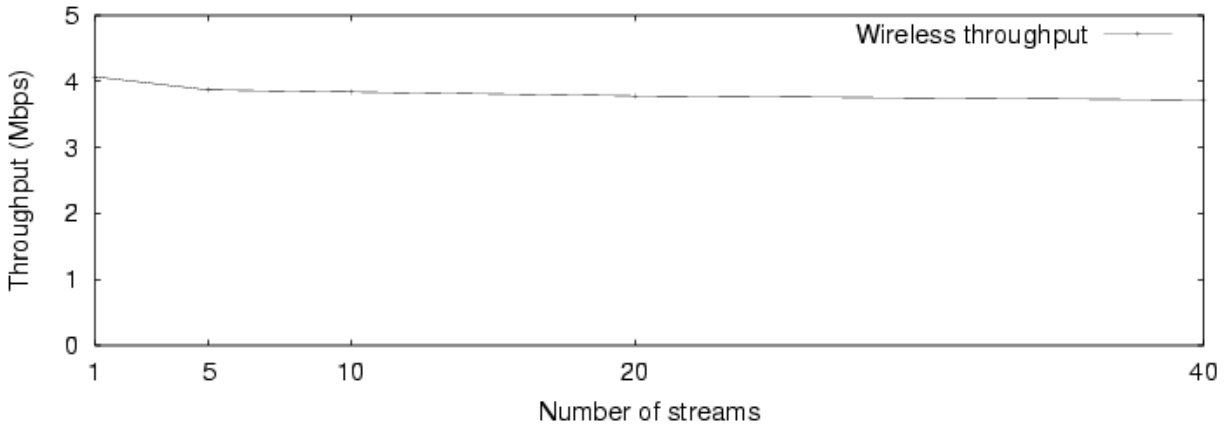
GigaEthernet:

480 Mbps

Wireless (802.11b):

4 Mbps





Experimental Evaluation:

Autonomic Performances

We ran two different active services:

- A lightweight service (MarkS)
- A heavyweight service (GzipS)

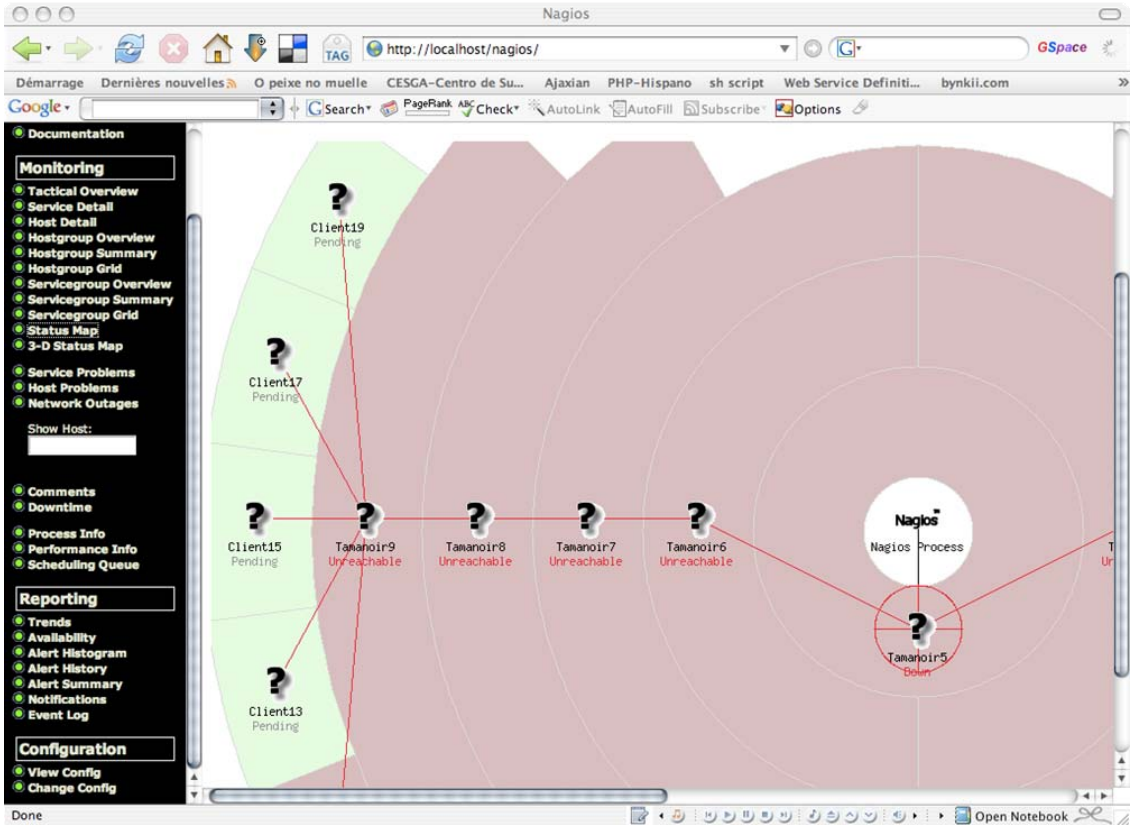
EE and services run in a SUN JVM 1.4.2

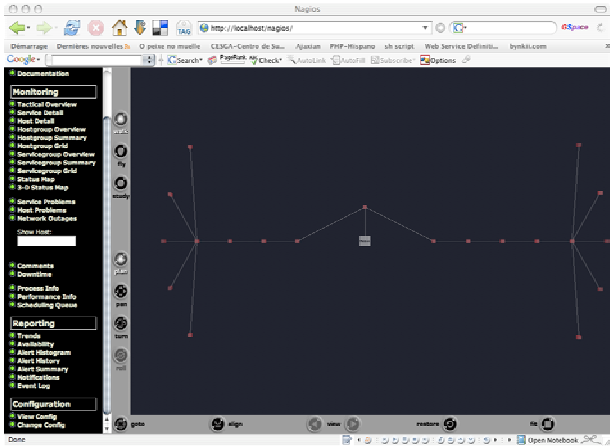
	4kB	16kB	32kB	56kB
MarkS	96	144	112	80
GzipS	9.8	14.5	15.9	16.6

(Throughput in Mbps)

Current / future experiments

- Evaluating large scale deployment with the Grid5000 platform



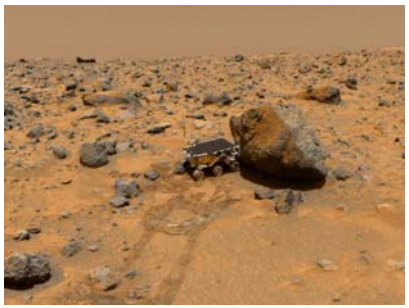


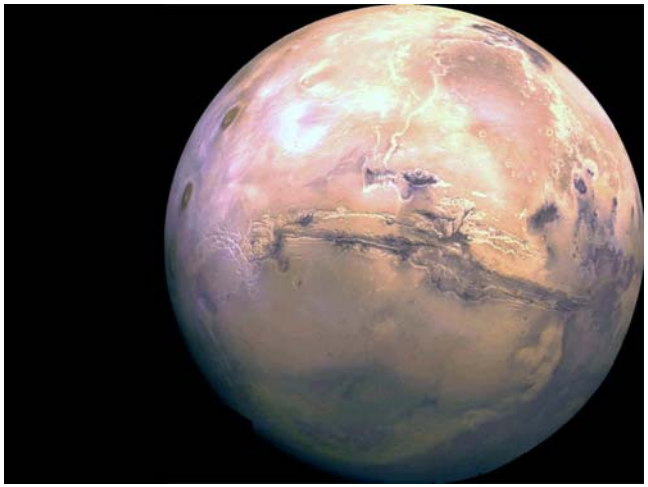
- Autonomic gateways around DSL infrastructure (DSLLAb project)

Scenario 2 : Inter-planetary Grid

Challenges

- Space missions will/already require computing/storage ressources to process collected data (from robots, cameras, sensors...)
- Sending large computing equipments on remote planets : too expensive!
- Need for a computing Interplanetary Grid which can support space challenges and provide an unified framework for computing collected data.





Pictures from : mars55.atomic-pigeon.net

Delay Tolerant Networking : “An approach to interplanetary internet”

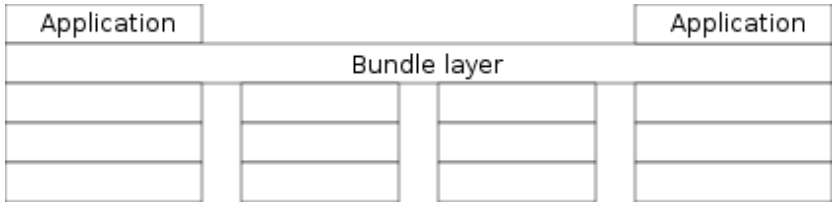
DTN community works on networks which must deal with:

- high latencies
- frequent disconnections
- no end-to-end path
- power saving constraints
- ...

Bundle layer	
TCP	Transport
IP	Network
Ethernet	Link

Based on a additional protocol layer.
The *bundle layer*, which provides:

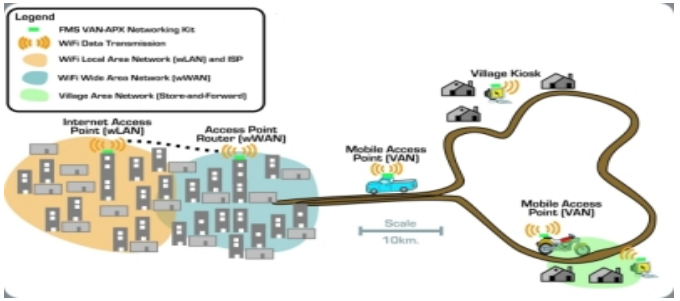
- intermediate storage
- **adaptation** to all kind of networks
- high latencies and long disconnections support



[S.Burleigh, A.Hooke, L.Torgerson, K.Fall, V.Cerf, B.Durst, K.Scott and H.Weiss, IEEE Communications, June 2003]

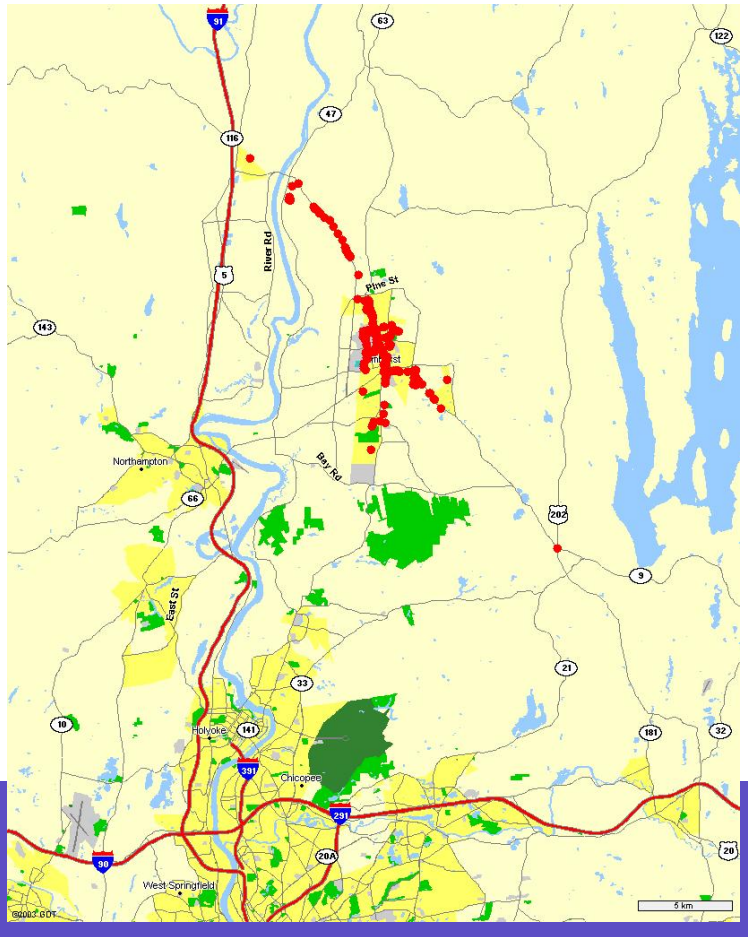
Some (terrestrial/marine) DTN projects: “When connection is not always available...”

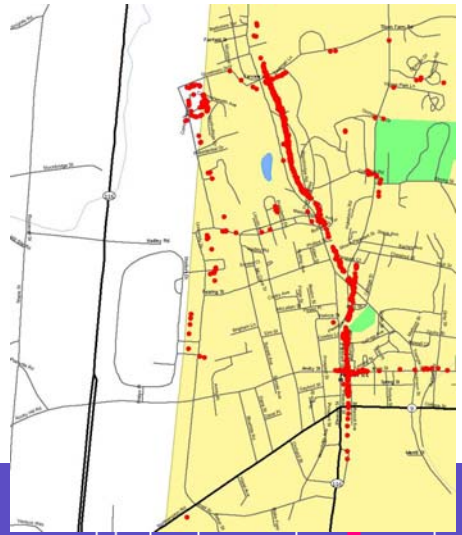
- UMassDieselNet http://prisms.cs.umass.edu/diesel
- ZebraNet http://www.princeton.edu/~mrm/zebranet.html
- DakNet http://firstmilesolutions.com
- SaamiNetworks
- DTN train demo
- ...

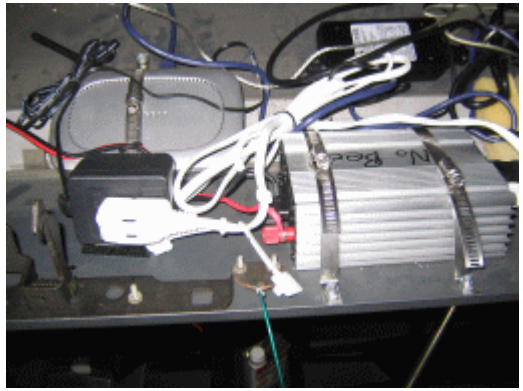


Connection / services in transport : DieselNet

- UMASS / Amherst
- 40 buses
- Bus to bus throughput : 2 Mbits



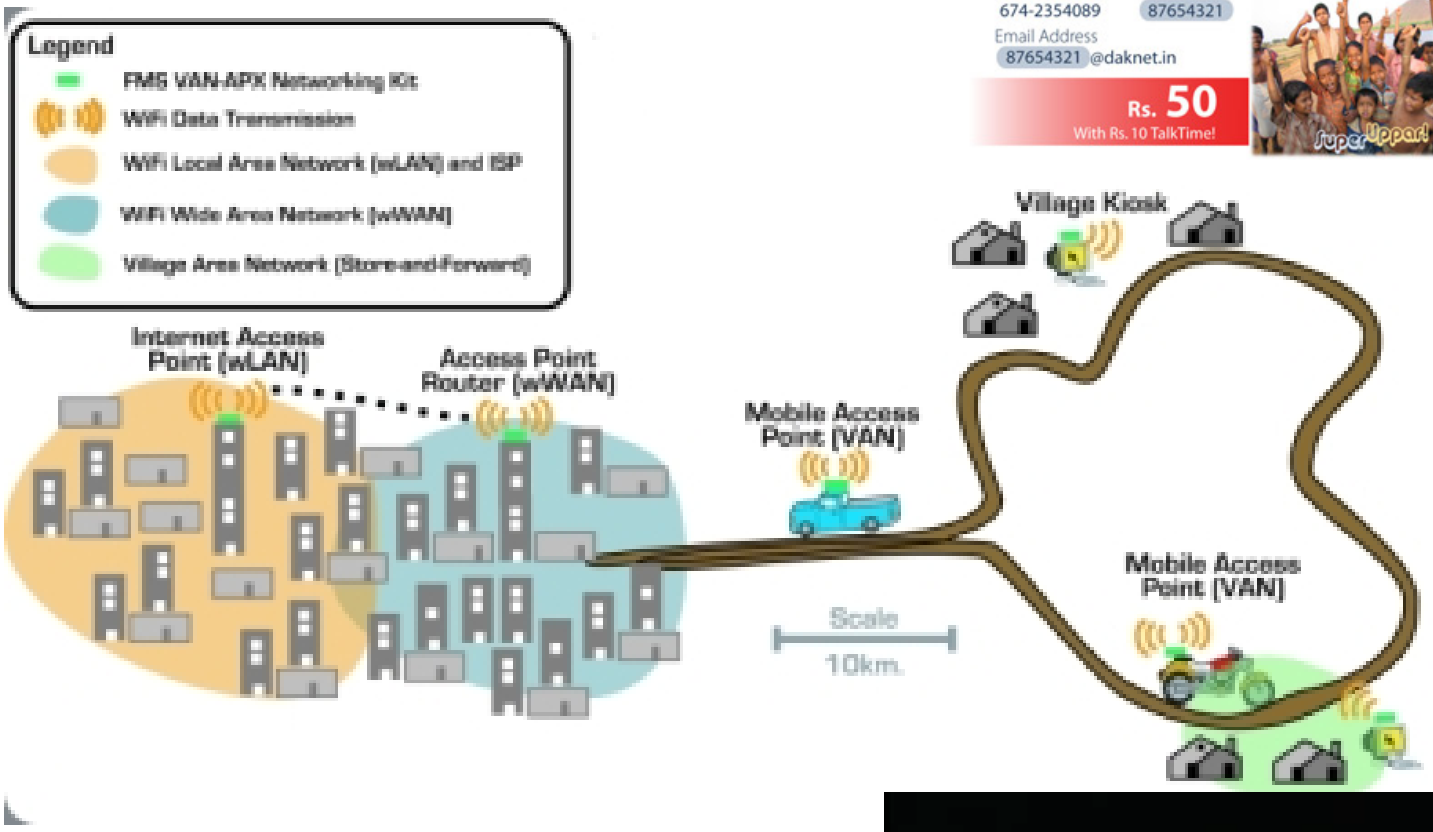




INRIA

Rural connections

- Ex company making money and providing services with DTN : (First Mille Solution)
- Services :
  - Offline web search
  - Emails
  - Voicemails/video mails/ SMS



Multiple Definitions of an Interplanetary-Grid ?

- Infrastructure definition :
  - Derived from Interplanetary networks
  - Heavy computing resources on Earth
  - Lightweight computing remote resources
- Services definition :
  - Remote intervention without human
  - Ultra long latencies networks
  - Disruptive connections
- Applications definitions :
  - Supporting space missions applications with local and remote ressources

- IPG = Grid + Autonomic Gateways + DTN

New services required but problems already exist...

- If the network is out of reach equivalent to a very large network congestion
- Needs to introduce equipments with new services
- In a large scale context, man can not really intervene
- Autonomic services are required...

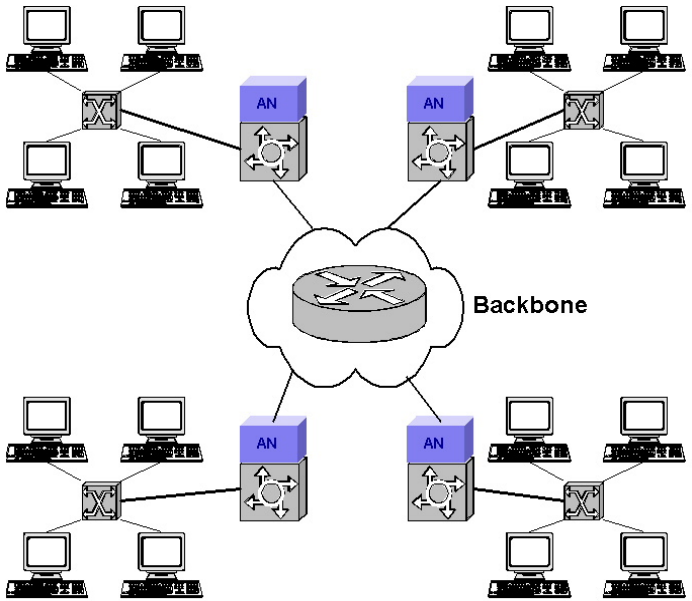
Why? (1)

- Today, applications must be adapted to support (very) high latency.
- Can not use end-to-end protocols. "*Store-and-forward*" technics required.
- Can not use negociation protocols. Protocols must take decisions locally and autonomously.

Why? (2)

- Grids' clusters connections can be through unreliable public links, providing no guaranty.
- Clusters owner may decide to disconnect their cluster from public access (own usage, management, upgrades,...)

Other clusters running the application **should not stop** because a cluster disappear for maybe just few hours!



Constraints

- *Transport protocols*, *routing*, *name space*... must be changed to fit new requirements.
- To build our architecture we need to take into account :

Classical Grid constraints

- Processing power
- Bandwidth
- Latency

IPG constraints

- Power consumption
- Volume (size)
- Ultra high latency
- Fault tolerance (no human intervention)

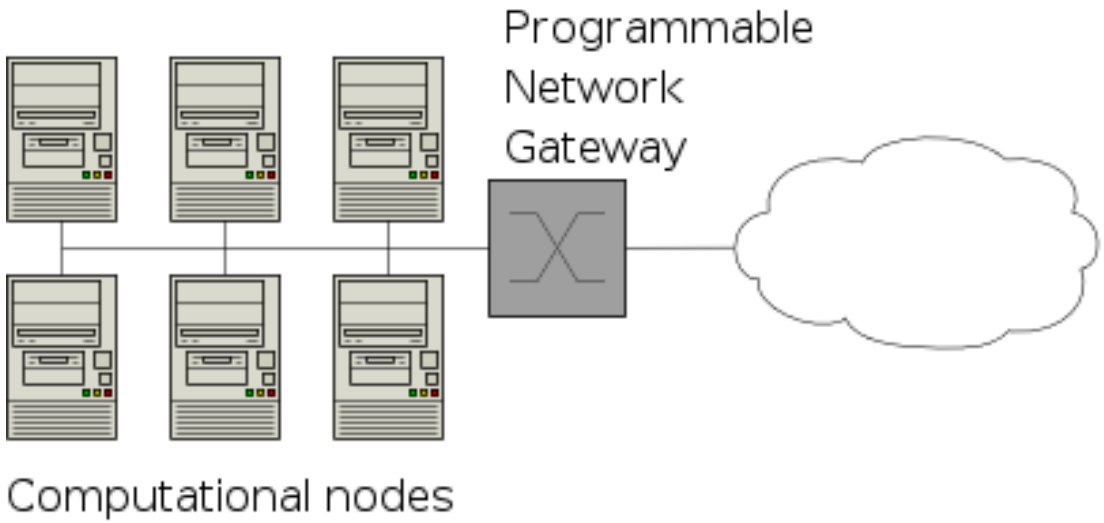
Our approach : designing services for gateway for IPG

- Considering disrupted infrastructure as ultra high latencies (or null bandwidth)
- Remaining as transparent as possible for *users*, *applications* and *Grid middleware*
- Designing an Autonomic *Programmable Network Gateway (APNG)*
- *Proposing adapted services for IPG*
- *Deploying APNG on strategic locations (between clusters and the external networks)*

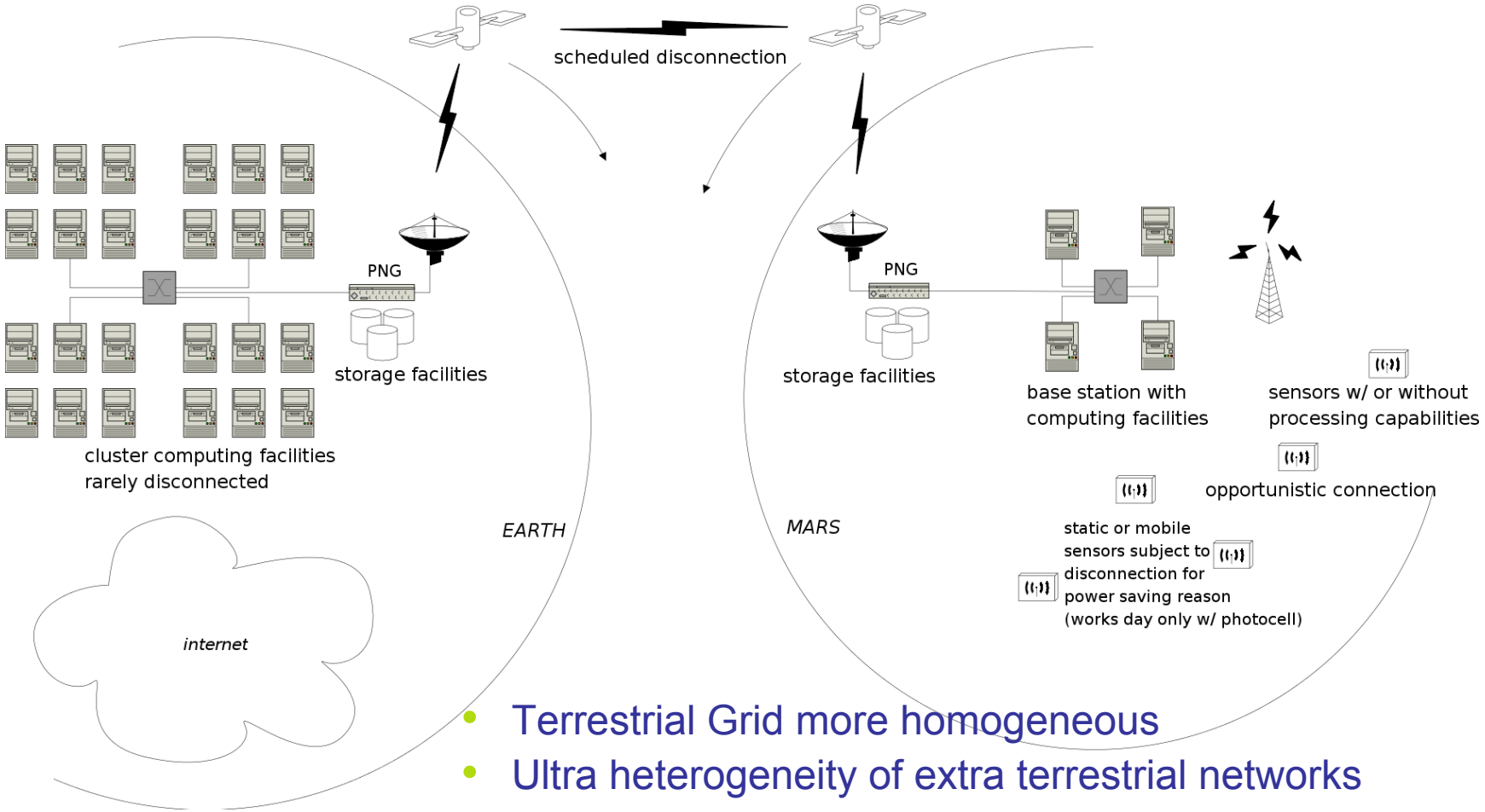
Autonomic Programmable Network Gateway (APNG)

A convenient way to support:

- network disruptions
- no access to the recipient nodes
- Processing/adaptation on the fly of data streams



An interplanetary Grid scenario: Interplanetary Grid between Earth an Mars

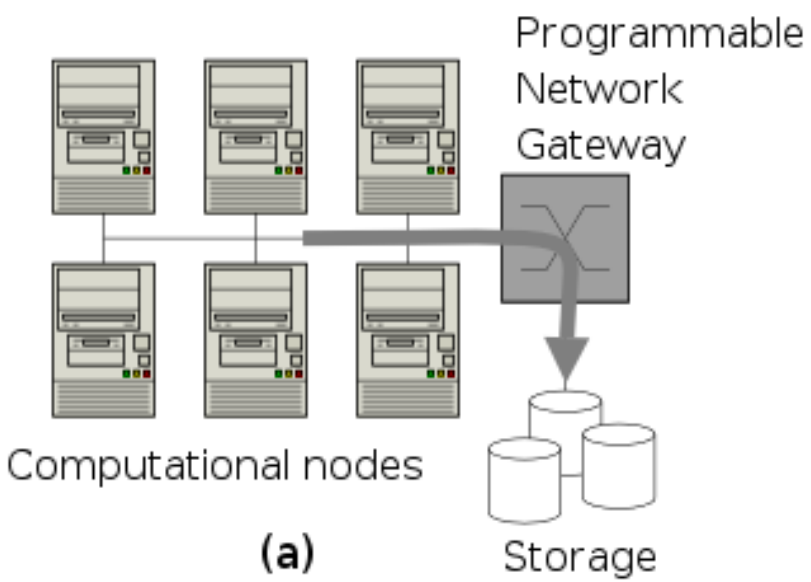


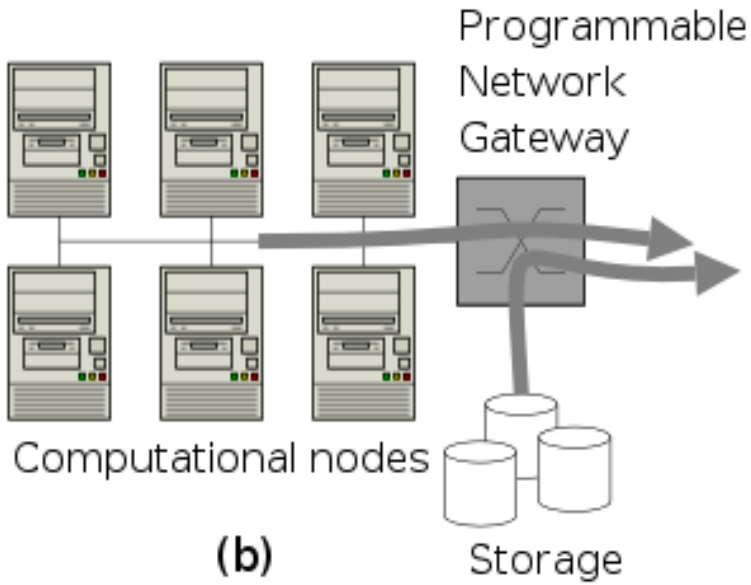
- Terrestrial Grid more homogeneous
- Ultra heterogeneity of extra terrestrial networks

Autonomic Programmable Network Gateway (APNG)

When a cluster is disconnected from the network the APNG should be able to:

- temporarily store data sent by the cluster's node in a local storage
- send a special acknowledgement (*TACK*) to the application





IPG : constraints and heterogeneity

3 levels of disruptions :

- Local (on earth) disruptions : between cluster/sites
- Long distance network disruptions (between earth and distant planet)
- Remote disruptions : between remote sensors and remote APNG

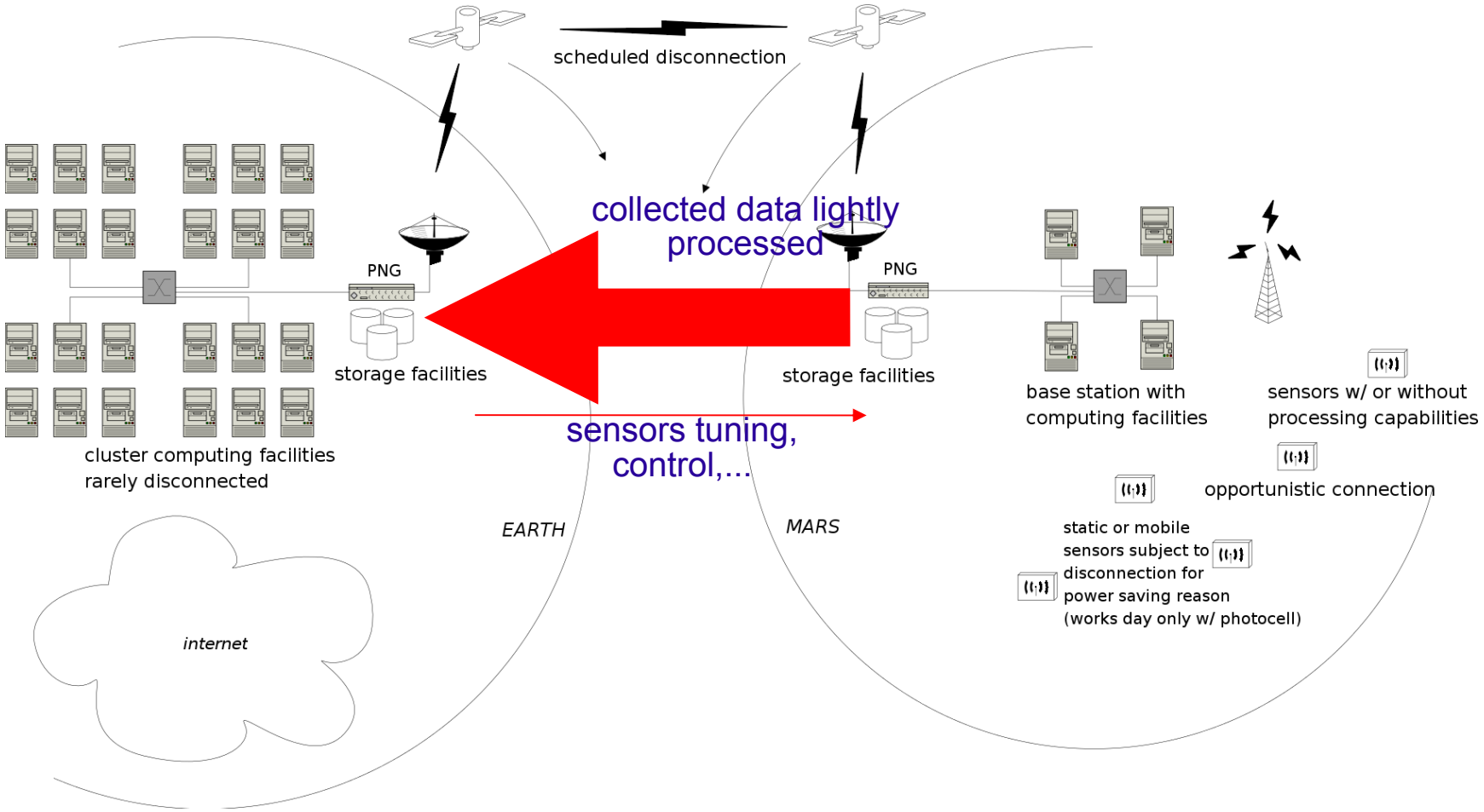
2 computing levels :

- Heavy computing on Earth
- Lightweight computing / filtering / storage on remote planet/space station

3 Networking levels :

- High speed networking : between clusters on earth
- High latency networking : satellite link between eath and remote planet
- Low power networking : between sensors and light processing capabilities

Heterogeneity in communications



Conclusions

- Given the available technologies, the concept of InterPlanetary Grid (IPG!) is far from Sci Fi
- The proposed architecture can also be applied to Grid infrastructure dealing with unreliable long distance network connections
- Disruption == long latency (minutes, hours, days)
- Our approach : first step to DTG : Disruption Tolerant Grids

Performance challenges

- Industrial embedded gateways enough efficient for low performance infrastructure (DSL…)
- Classical PC architecture : OK for Gbit infrastructure

•What about 10G ? -> Looking for gateways with 2 X 10G NIC with enough PCI/CPU

To show limitations, need for network processor (hardware) support, experiment with 10G networks

Current / Future works

- ***First experiment is on going work : inclusion of DTN in autonomic network platform***
- ***Currently designing an DTG/IPG emulator***
- ***Evaluation on a large scale with Grid'5000 project***

- **Combine and interface APNG with SBLOMARS: SNMP-Based Load Balancing Monitoring Agents for Resource Scheduling in Grids (Univ. Politechnica Catalunya, Barcelona)**

Acknowledgments

Jean-Patrick Gelas

Damien Nicolet

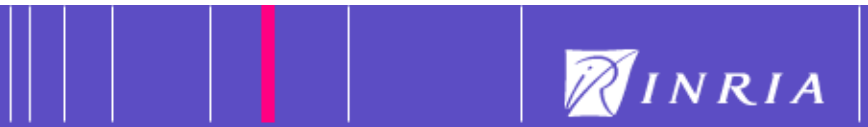
Pierre Bozonnet

Martine Chaudier

Edgar Magana (UPC, Barcelona)

Questions?

laurent.lefevre@inria.fr




INRIA


INRIA