



Intelligent Network Infrastructures for Global Grid Computing

Luca Valcarenghi, Piero Castoldi, and Lorenzo Rossi

Scuola Superiore Sant'Anna for University Study and Research

Pisa, Italy

{valcarenghi, castoldi, rossi}@sssup.it

GGF 12

GHPN-RG Session II, 22 October, 2004

National Research Program
Strategic Project on Enabling Technologies for Information Society
FIRB



<http://www.grid.it>

**Enabling Technologies for High-performance Computational
Grids Oriented to Scalable Virtual Organizations**



National Program Coordination



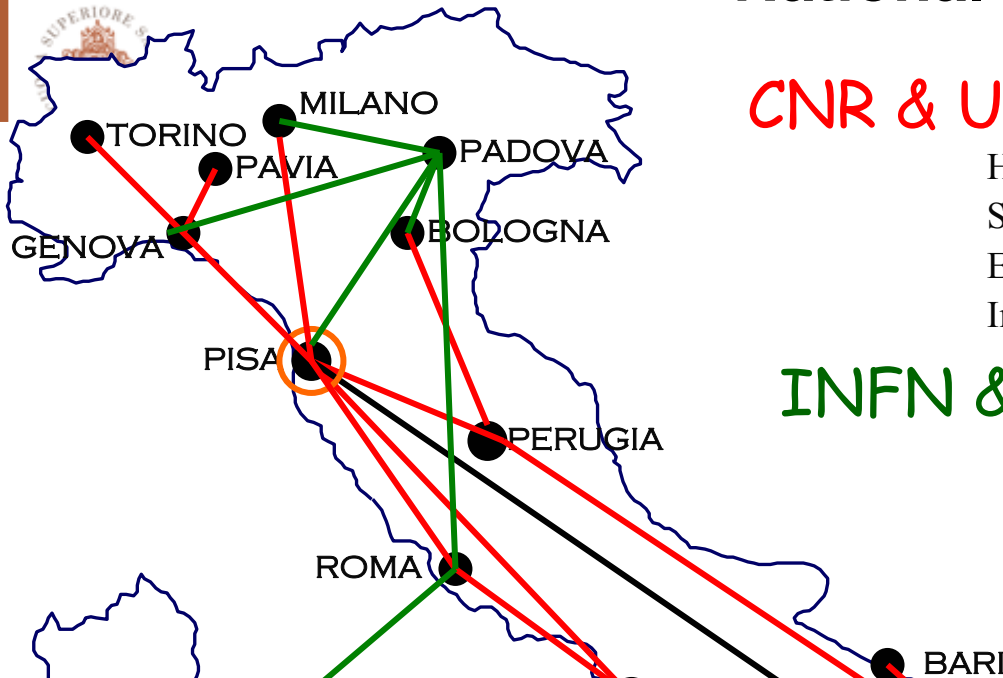
consorzio nazionale
interuniversitario
per le telecomunicazioni

CNR & University

HPC, Parallel Programming, Grid computing,
Scientific libraries, Data base and knowledge discovery,
Earth Observation, Computational chemistry,
Image processing, ...

INFN & University

Grid (INFN-Grid, DataGrid, DataTag),
e-science applications: Astrophysics,
Bioinformatics, Geophysics, ...



GRID.IT 3-Year Project

Cost 11.1 M€

Funding: €8.1 M

(1.1 M€ for young researchers)

Start-up: November 2002

Applications and Demonstrators

(WP10, WP11, WP12, WP13, WP14)

Component-based Programming Environment (WP8),
e-science Components (WP9)

Knowledge Services (WP6), Grid Portals (WP7),
Security (WP4)

Photonic, high-
bandwidth
networks
(WP1, WP2)

Basic Grid Infrastructure
and Data Core Services
(WP3, WP5)

WP 1 - Grid oriented optical switching paradigms (Resp. P. Castoldi)

- Activity 1 – Connections, topologies and network service models

(Resp. R. Battiti/F. Granelli)



- Activity 2 – Grid computing on state-of-the-art optical networks

(Resp. P. Castoldi)



- Activity 3 – Migration scenarios to intelligent flexible optical networks

(Resp. F. Callegati)



- Activity 4 – Control plane and network emulation for optical packet switching networks

(Resp. A. Fumagalli)



- Activity 5 – Enabling technologies for optical switching networks

(Resp. G. Cancellieri)



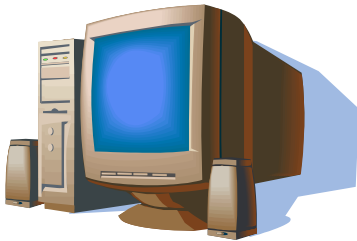


Grid Network Aware Programming Environment

cnit

consorzio nazionale
interuniversitario
per le telecomunicazioni

User

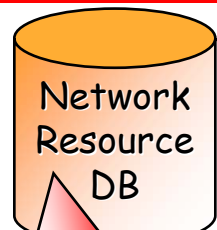


User Interface (UI)

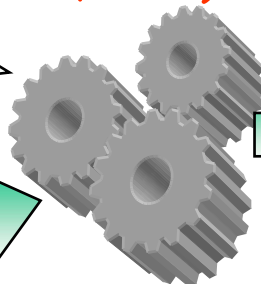
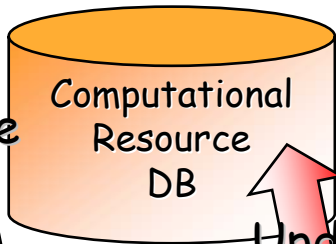
Application

Application requests

Run
(max_exec_time
, reliable, etc.)



$f(\text{max_exec_time}, \text{reliable}, \text{etc.})$



Update

Update

Middleware \Rightarrow
Grid Abstract Machine

Notification

Allocation request

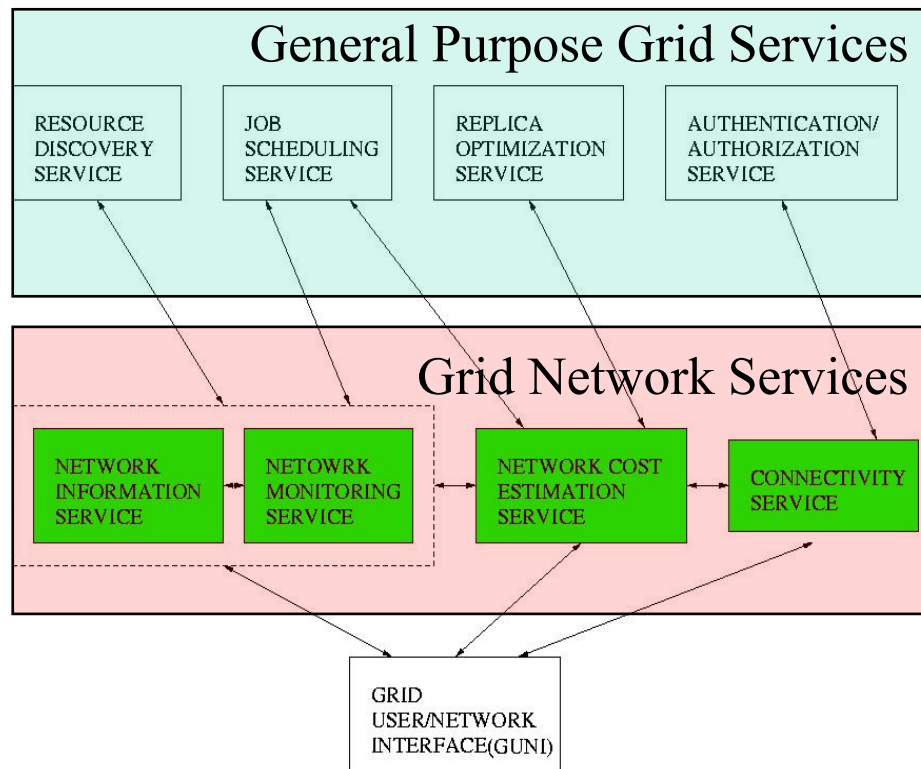
Resource allocation

Notification

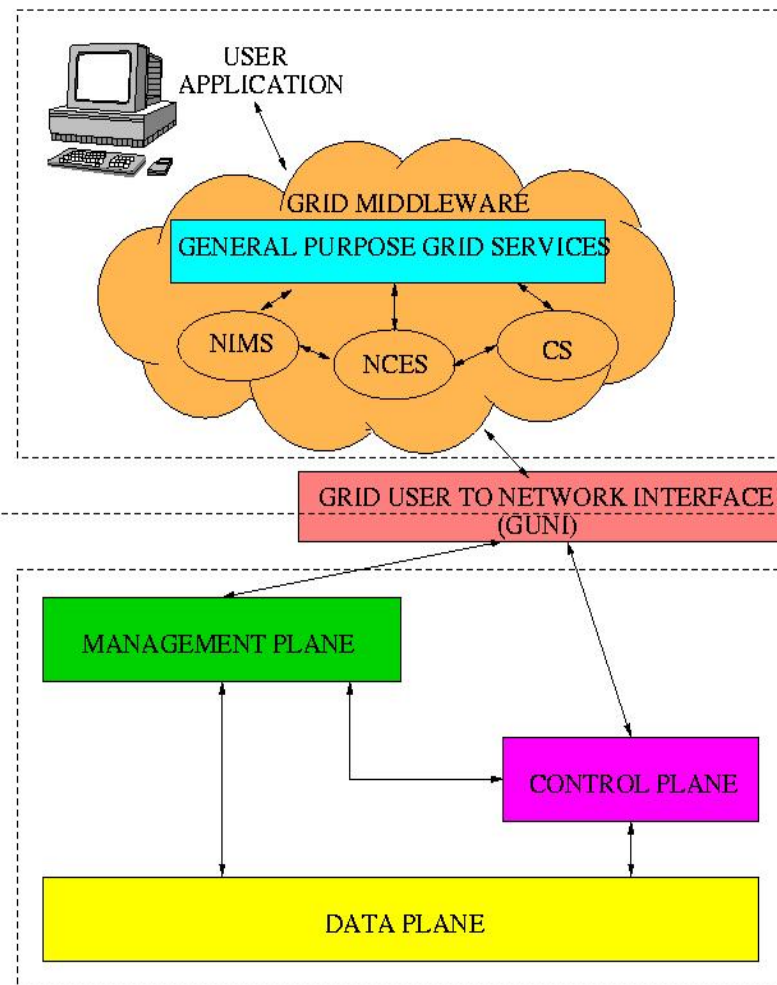
Elaboration

Basic HW+SW

Grid Network Service and Network Management and Control Plane Interaction



Programming
Environment



- A network that provides some ability to recover ongoing connections disrupted by the catastrophic failure of a network component, such as a line interruption or a node failure, is said to be
 - Resilient → resilience (resiliency)
 - Reliable → reliability
 - Survivable → survivability
- Resilience QoS Parameters
 - Restoration Blocking Probability (P_b)
 - Ratio between the number recovered connections and the number of failed connections
 - Recovery Time (RT)
 - Time elapsed between failure notification and transmission restart
 - Restoration Blocking Probability ↔ inter-service communication bandwidth and inter-service connectivity
 - Recovery Time ↔ inter-service communication latency



Approaches for Grid Networking Resilience



consorzio nazionale
interuniversitario
per le telecomunicazioni

Layered Grid Architecture

Failover Schemes

TCP/IP Stack

Application
end-user applications

Application specific fault tolerant
schemes based on middleware
fault detection

Middleware

Collective
collective resource control

Tasks and Data Replicas

Condor-G
checkpointing, migration, DAGMan

Application

Resource
resource management

GT2/GT3
GridFTP
Reliable File Transfer (RFT)
Replica Location Service (RLS)
Fault Tolerant TCP (FT-TCP)

Transport

Connectivity
Inter-process communication,
protection

delegated to WAN, MAN,
and LAN resilience schemes

Internet/Network

Fabric
basic hardware and software

delegated to HW, SW,
and farm failover schemes

Link

Resilience

Protection

Network spare resources are computed and reserved upon connection set up

Restoration

Network spare resources are found and reserved upon failure occurrence

Dedicated

Each connection is assigned dedicated spare resource

Shared

Connection that are not contemporarily involved in the failure event share spare resources

Dynamic

Backup routes are computed and spare resources are reserved upon failure occurrence

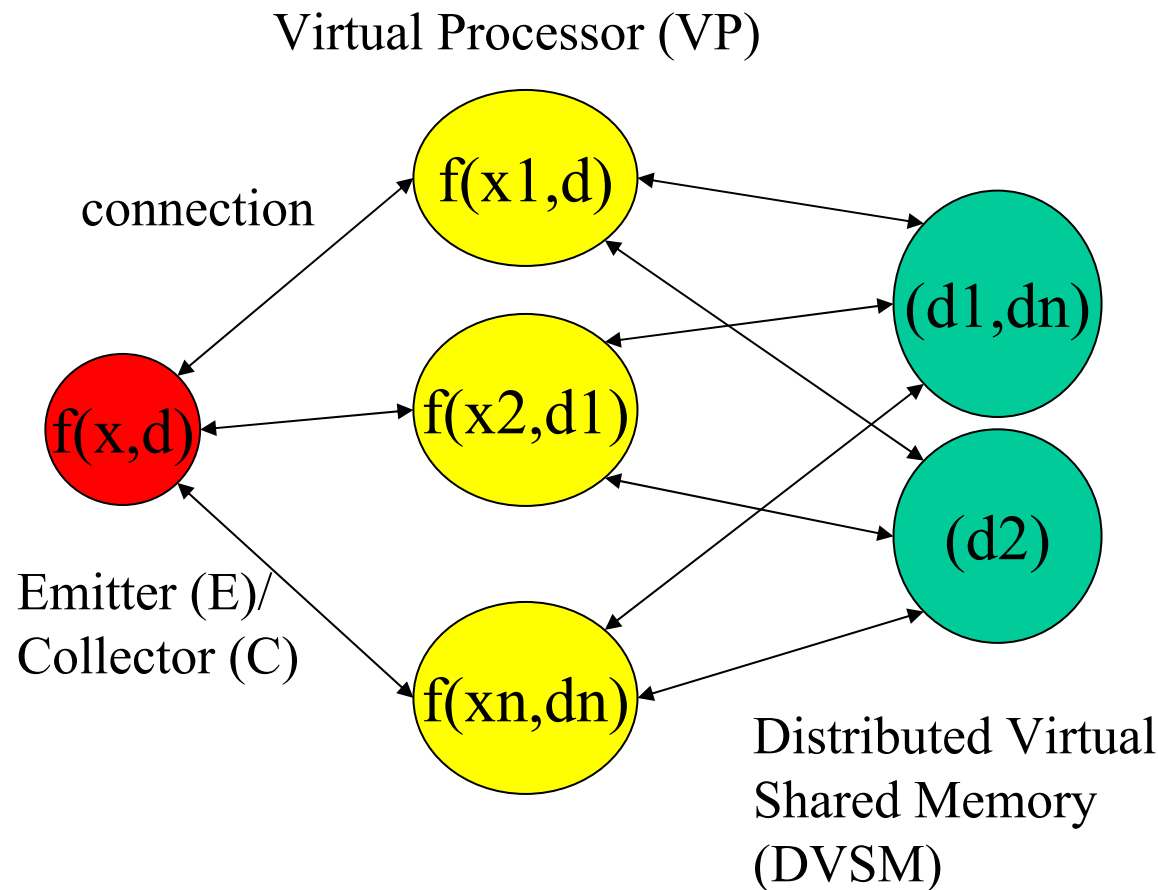
Pre-planned

Backup routes are pre-computed but spare resources are reserved upon failure occurrence

- Protection/Restoration
 - Path level
 - End-to-end connections are independently recovered by finding a new route from source to destination
 - Link level
 - All the connections disrupted by the failure (in this case link failure) are rerouted along the same recovery path by-passing the failure (i.e., the failed link)
- Restoration schemes are commonly available at higher layers (e.g., the IP layer)
- Protection schemes are commonly used at the physical transport layer (e.g., WDM)

Recovery Times

- BGP-4: 15 – 30 minutes
- OSPF: 1s to minutes
- MPLS fast (link) rerouting 50-100ms
- MPLS edge-to-edge rerouting 1-100s
- Spanning Tree 50s
- RSTP 10ms
- FRP-FAST < 2s
- EtheReal ~ 250ms
- SDH / SONET / DWDM: 50 ms
- OCh and OMS restoration ≥ 50 ms
- Dedicated OL protection 10 μ s-10ms
- Shared OL protection 1-100ms

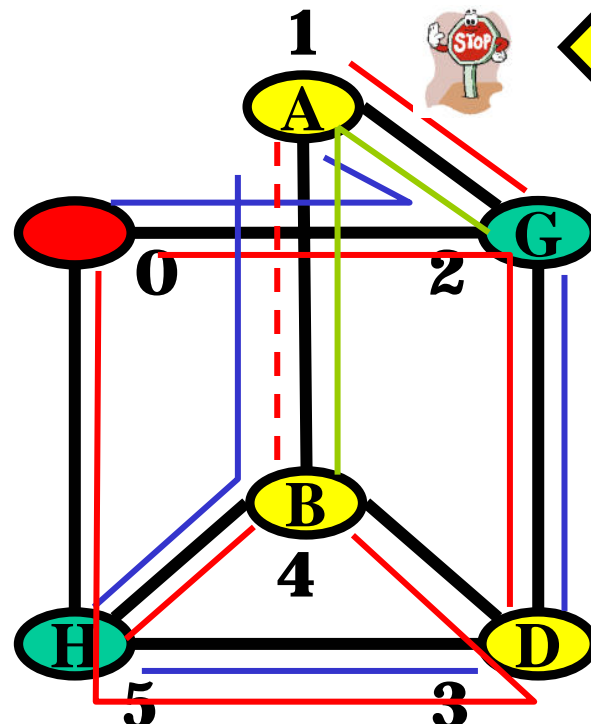
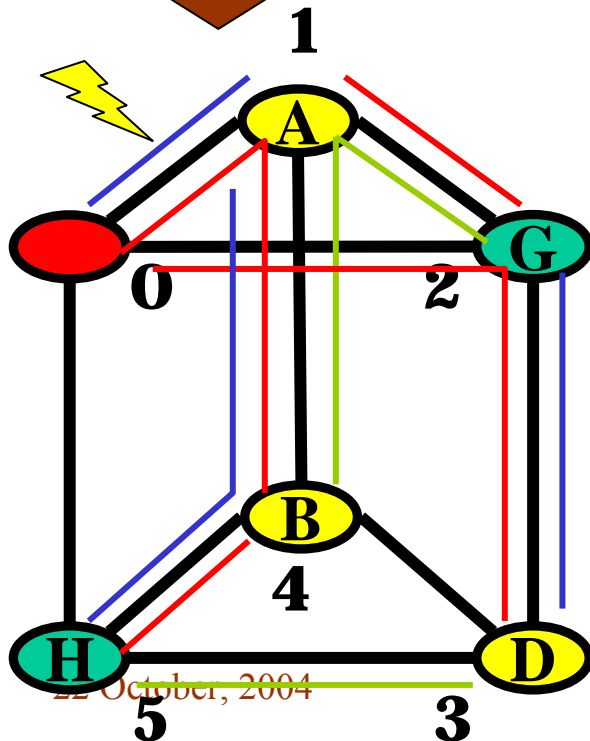
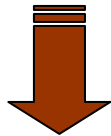
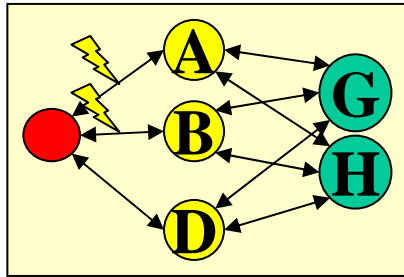






- Motivations
 - Guaranteeing the successful calculation of function $f(x,d)$ in spite of grid infrastructure failures
 - Nowadays grid failure guaranteed by middleware and network resilient schemes independently
- ↓
- Grid network failover improved by Grid network infrastructure resilient scheme integration

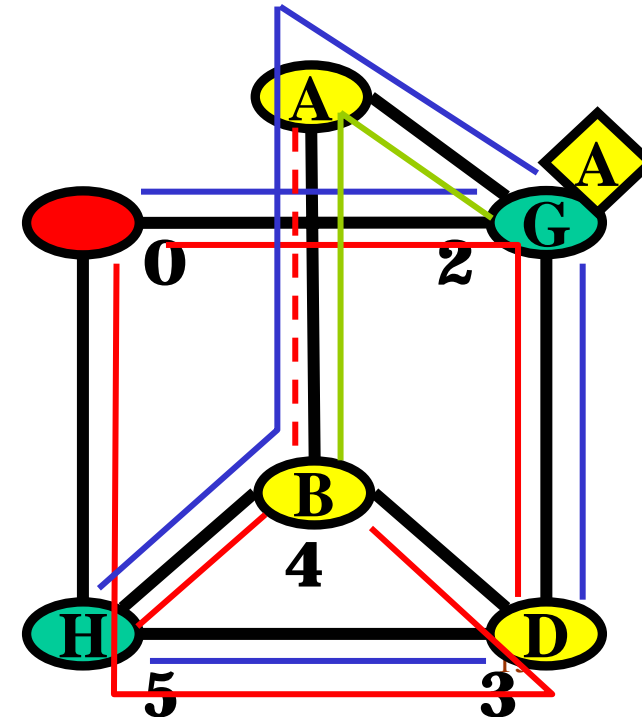
- Fabric fault tolerant scheme already implemented in the LAN
- Utilize fault tolerant scheme in the global internet to implement a reliable fabric
- Integrate with higher layer (e.g., collective, job scheduling) fault tolerance schemes
- Try to move the most of the resilience burden to the fabric making a lighter service for collective and higher layers

Proposed Approach

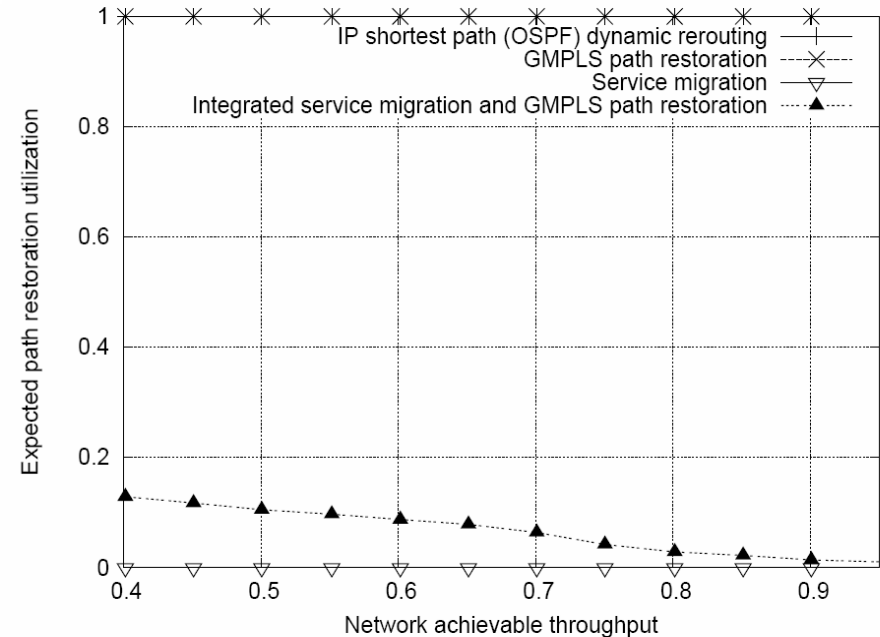
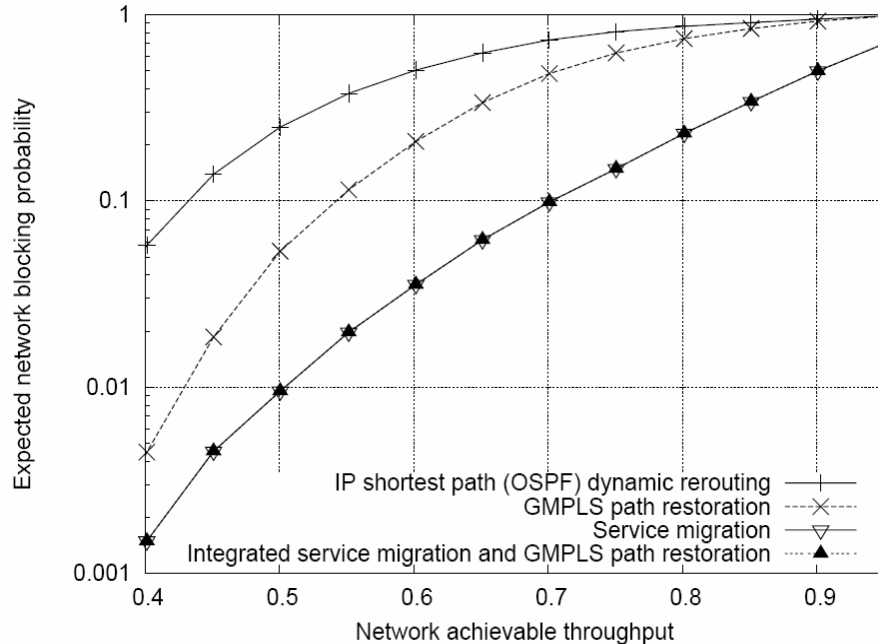
- Integrating network layer connection rerouting with task/data replication/migration
- Integrated scheme model by MILP problem formulation
 - Objective: maximizing the number of connections restored after failure



  original task/data
  task/data replica



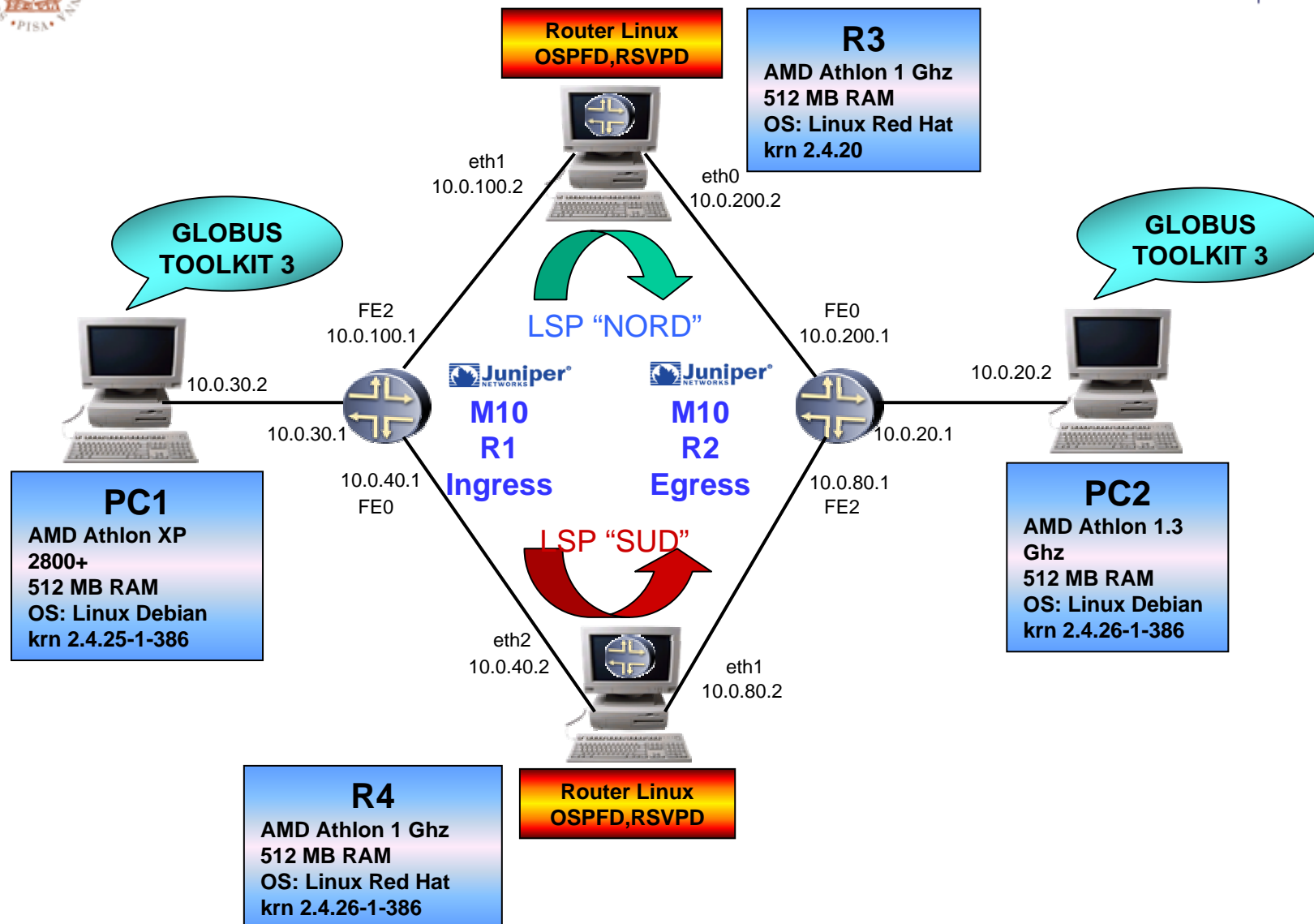
Integrated Restoration Performance



- Integrated restoration outperforms OSPF dynamic rerouting resilience
- Integrated restoration performs as well as service migration resilience but by utilizing path restoration decreases the need for service synchronization and restart



MPLS Grid (MGRID) Testbed

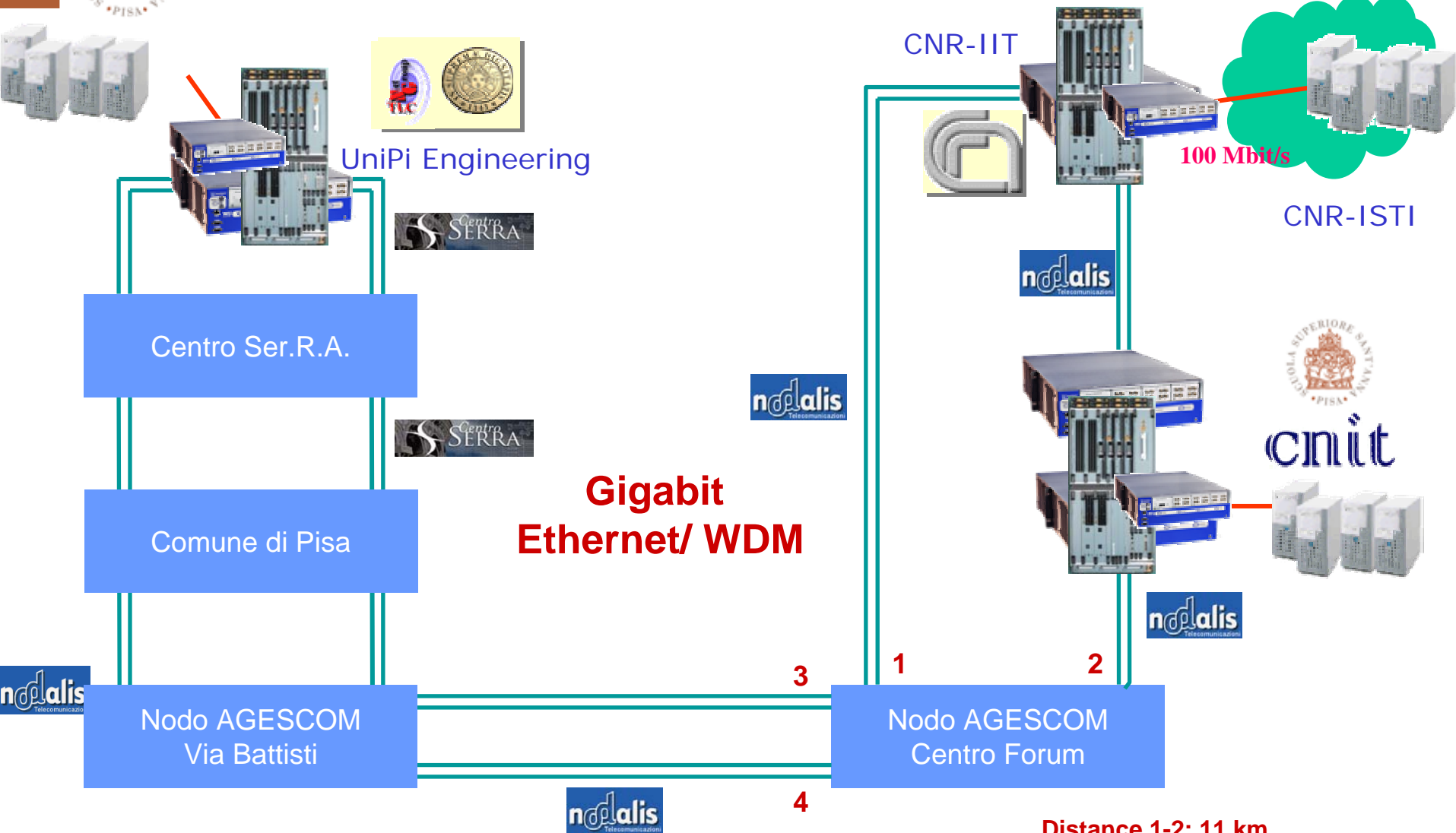




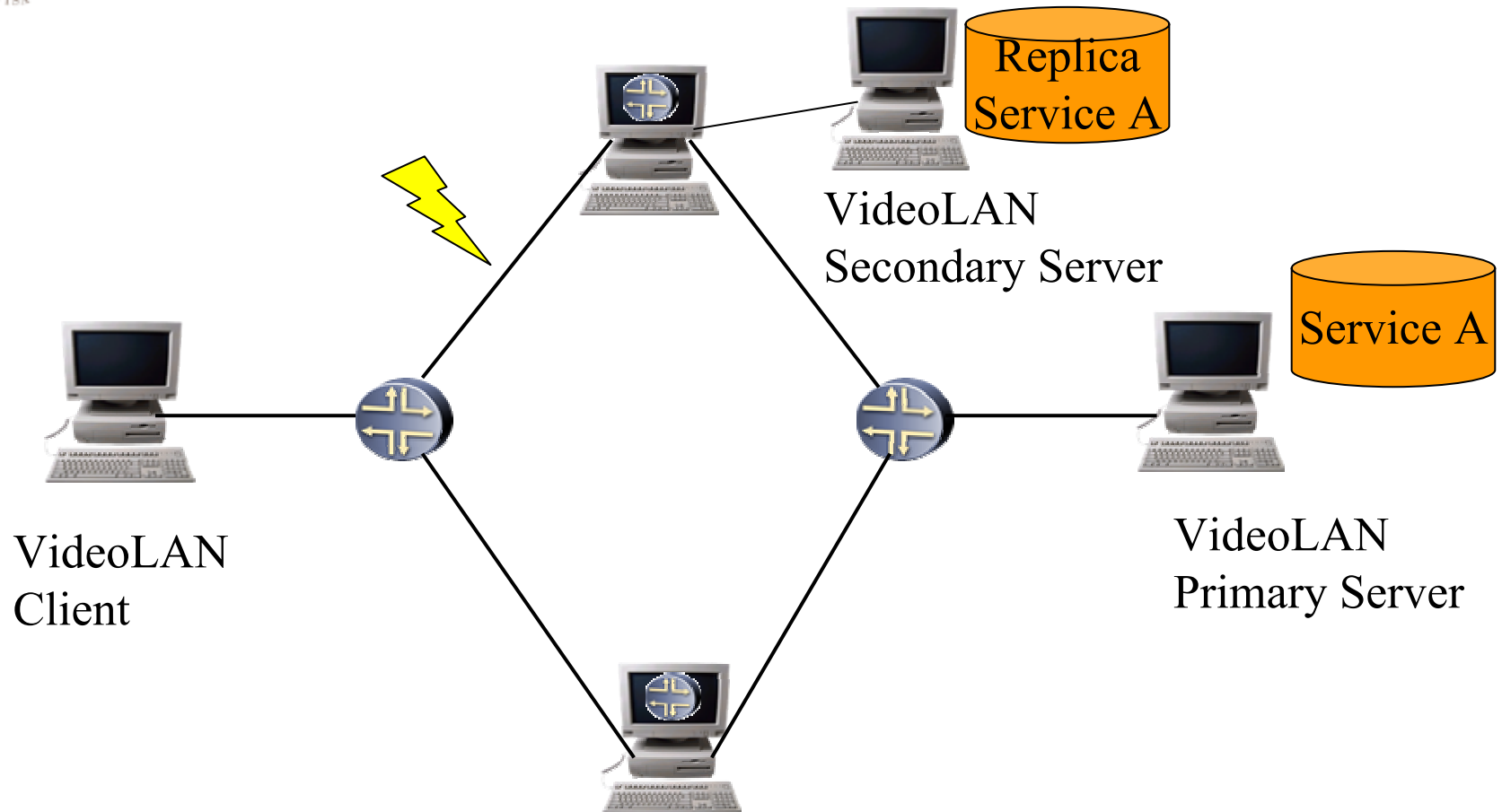
The Metrocore/VESPER testbed



consorzio nazionale
interuniversitario
per le telecomunicazioni



Distance 1-2: 11 km
Distance 3-4: 12,5 km



Connection Between Client and Primary Server Working

Shell - Konsole <2>


Sessione Modifica Visualizza Segnalibri Impostazioni Aiuto

```

[00000343] main video output warning: late picture skipped (-990)
[00000343] main video output warning: late picture skipped (274176)
[00000343] main video output warning: late picture skipped (224240)
[00000343] main video output warning: late picture skipped (190885)
[00000343] main video output warning: late picture skipped (140846)
[00000343] main video output warning: late picture skipped (107510)
[00000368] main private debug: decoded 106/108 pictures
[00000343] main video output warning: late picture skipped (106366)
[00000343] main video output warning: late picture skipped (73060)
[00000343] main video output warning: late picture skipped (76979)
[00000343] main video output warning: late picture skipped (43680)
[00000343] main video output warning: late picture skipped (-6359)
[00000343] main video output warning: late picture skipped (58077)
[00000343] main video output warning: late picture skipped (24798)
[00000343] main video output warning: late picture skipped (61333)
[00000343] main video output warning: late picture skipped (11344)
[00000368] main private debug: stream periodicity changed from P[3] to P[1]
[00000343] main video output warning: late picture skipped (380165)
[00000343] main video output warning: late picture skipped (346875)
[00000343] main video output warning: late picture skipped (296859)
[00000343] main video output warning: late picture skipped (263494)
[00000343] main video output warning: late picture skipped (213454)
[00000343] main video output warning: late picture skipped (218782)
[00000368] main private debug: stream periodicity changed from P[1] to P[3]
[00000343] main video output warning: late picture skipped (115325)
[00000343] main video output warning: late picture skipped (-1362)
[00000343] main video output warning: late picture skipped (55419)
[00000343] main video output warning: late picture skipped (22124)
[00000343] main video output warning: late picture skipped (101448)
[00000343] main video output warning: late picture skipped (51517)
[00000343] main video output warning: late picture skipped (18163)

```

VLC (X11 output)




Primary Server and Secondary Server Link Failure

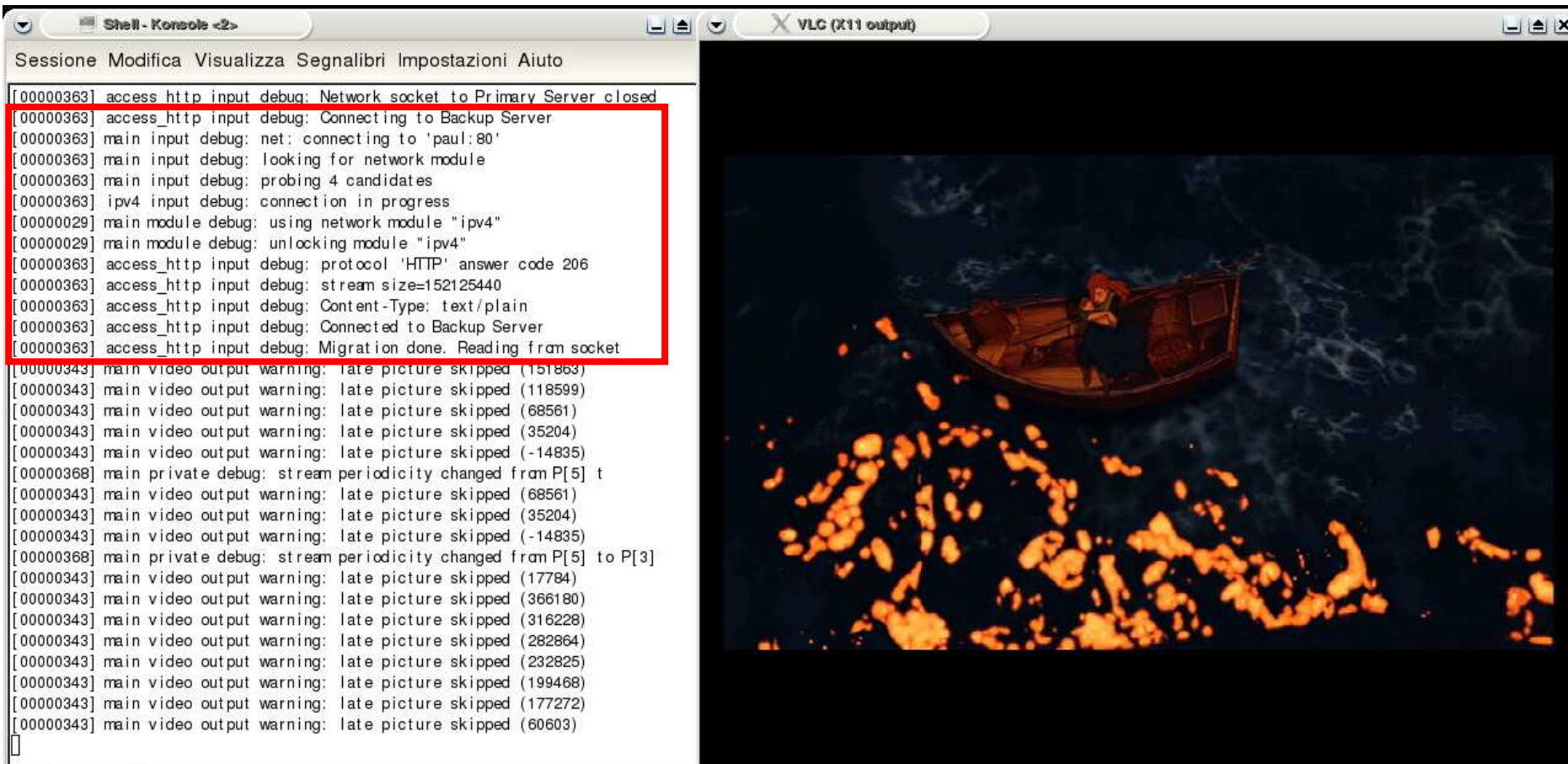
Shell - Konsole <2>

Sessione Modifica Visualizza Segnalibri Impostazioni Aiuto

```
[00000343] main video output warning: late picture skipped (274980)
[00000343] main video output warning: late picture skipped (241626)
[00000343] main video output warning: late picture skipped (191587)
[00000368] main private debug: stream periodicity changed from P[3] to P[5]
[00000343] main video output warning: late picture skipped (189528)
[00000343] main video output warning: late picture skipped (56117)
[00000343] main video output warning: late picture skipped (91725)
[00000343] main video output warning: late picture skipped (58438)
[00000343] main video output warning: late picture skipped (8400)
[00000363] access_http input debug: No data read from network socket
[00000363] access_http input debug: Starting server migration
[00000363] access_http input debug: Closing network socket to Primary Server
[00000363] access_http input debug: Network socket to Primary Server closed
[00000363] access_http input debug: Connecting to Backup Server
[00000363] main input debug: net: connecting to 'paul:80'
[00000363] main input debug: looking for network module
[00000363] main input debug: probing 4 candidates
[00000363] ipv4 input debug: connection in progress
[00000029] main module debug: using network module "ipv4"
[00000029] main module debug: unlocking module "ipv4"
[00000363] access_http input debug: prot
[00000363] main input debug: looking for network module
[00000363] main input debug: probing 4 candidates
[00000363] ipv4 input debug: connection in progress
[00000029] main module debug: using network module "ipv4"
[00000029] main module debug: unlocking module "ipv4"
[00000363] access_http input debug: protocol 'HTTP' answer code 206
[00000363] access_http input debug: stream size=152125440
[00000363] access_http input debug: Content-Type: text/plain
[00000363] access_http input debug: Connected to Backup Server
[00000363] access_http input debug: Migration done. Reading from socket
```

VLC (X11 output)





The screenshot shows a terminal window titled "Shell - Konsole <2>" and a video player window titled "VLC (X11 output)".

The terminal window displays the following log messages:

```

[00000363] access_http input debug: Network socket to Primary Server closed
[00000363] access_http input debug: Connecting to Backup Server
[00000363] main input debug: net: connecting to 'paul:80'
[00000363] main input debug: looking for network module
[00000363] main input debug: probing 4 candidates
[00000363] ipv4 input debug: connection in progress
[00000029] main module debug: using network module "ipv4"
[00000029] main module debug: unlocking module "ipv4"
[00000363] access_http input debug: protocol 'HTTP' answer code 206
[00000363] access_http input debug: stream size=152125440
[00000363] access_http input debug: Content-Type: text/plain
[00000363] access_http input debug: Connected to Backup Server
[00000363] access_http input debug: Migration done. Reading from socket
[00000343] main video output warning: late picture skipped (151863)
[00000343] main video output warning: late picture skipped (118599)
[00000343] main video output warning: late picture skipped (68561)
[00000343] main video output warning: late picture skipped (35204)
[00000343] main video output warning: late picture skipped (-14835)
[00000368] main private debug: stream periodicity changed from P[5] t
[00000343] main video output warning: late picture skipped (68561)
[00000343] main video output warning: late picture skipped (35204)
[00000343] main video output warning: late picture skipped (-14835)
[00000368] main private debug: stream periodicity changed from P[5] to P[3]
[00000343] main video output warning: late picture skipped (17784)
[00000343] main video output warning: late picture skipped (366180)
[00000343] main video output warning: late picture skipped (316228)
[00000343] main video output warning: late picture skipped (282864)
[00000343] main video output warning: late picture skipped (232825)
[00000343] main video output warning: late picture skipped (199468)
[00000343] main video output warning: late picture skipped (177272)
[00000343] main video output warning: late picture skipped (60603)
  
```

The video player window shows a scene of a small boat on a dark sea at night, with a large fire or explosion visible in the foreground, illuminating the water and the boat.

- Which Layer must respond to failures ?
- Which layer more efficiently overcomes which failures ?
- Inter-layer coordination between application layer, middleware, and grid network services resilient schemes
- Can recovery times typical of Clusters be achieved in Global Grid Computing ?
- Which are the expenses to make a Global Grid Cluster perform as well as Local Area Network Cluster ?



Thanks to ... the people (in alphabetical order)

- Filippo Cugini, CNIT
- Luca Foschini, SSSUP
- Domenico Laforenza, CNR
- Francesco Paolucci, CNIT
- Marco Vanneschi, UNIPI