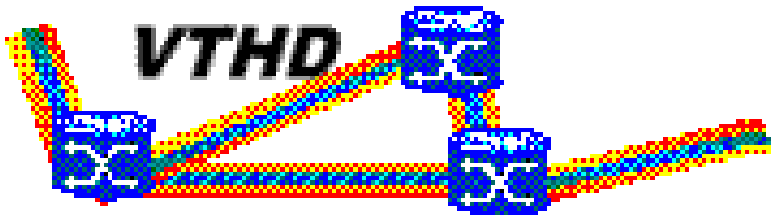
Middleware for Grid computing: experiences using the VTHD Network (2.5 Gbit/s)

Frédéric DESPREZ
ENS-Lyon/INRIA
&
Thierry PRIOL
IRISA/INRIA

Contents:

- The RNRT VTHD project
- How to exploit a high bandwidth network transparently ?
  - High-performance distributed objects platform (Paco/PADICO)
  - Global address space for the Grid (MOME)
- Computational servers on a High Speed WAN (DIET)



VTHD project,

"*Vraiment Très Haut Débit*"

Coordinator: Christian Guillemot **(**FTR&D)

christian.guillemot@francetelecom.com

http://www.vthd.org

Project organisation

SP1: Platform deployment (FTR&D, INRIA, ENST,ENST-Br, INT, EURECOM)

- IP engineering and WDM, Back-office IP, « best-effort » IPv4 service

SP2: Flow and congestion control (INRIA, FTR&D)

- TCP, Web traffic genetation and monitoring (WAGON/ INRIA software).

SP3: Multi-service high-bandwidth platform (ENST, ENST-Br, INT, FTR&D)

- QoS service class, RPV service, Tera-routers.

SP4: Training and medical applications (ENST, INRIA, ENST-Br, INT)

- Contents server, remote training (ENST-Br,INT,ENST).
- Medical « white board » , Data transmission associated with the use of medical robots (INRIA, HEGP).

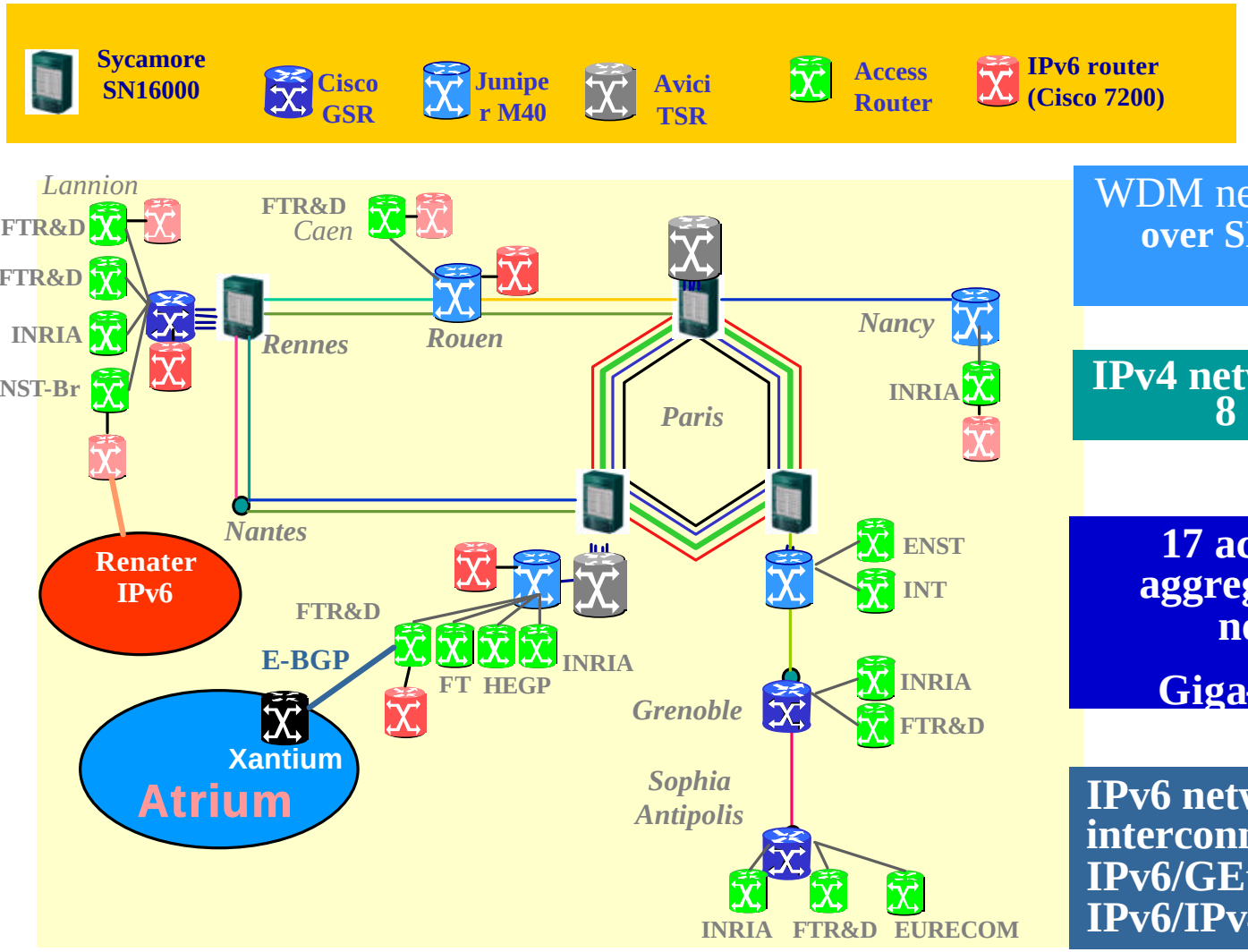
SP5: Distributed computing and visualisation resources (INRIA)

- Grid computing, distributed numerical simulation, remote visualisation.

SP6: Distributed cache systems (EURECOM, ENST, INRIA)

- Data localisation, benchmarking WAGON.

VTHD network



WDM network with « **packet over SDH** » **channels** 2,5 Gbit/s.

IPv4 network supported by 8 giga-routers

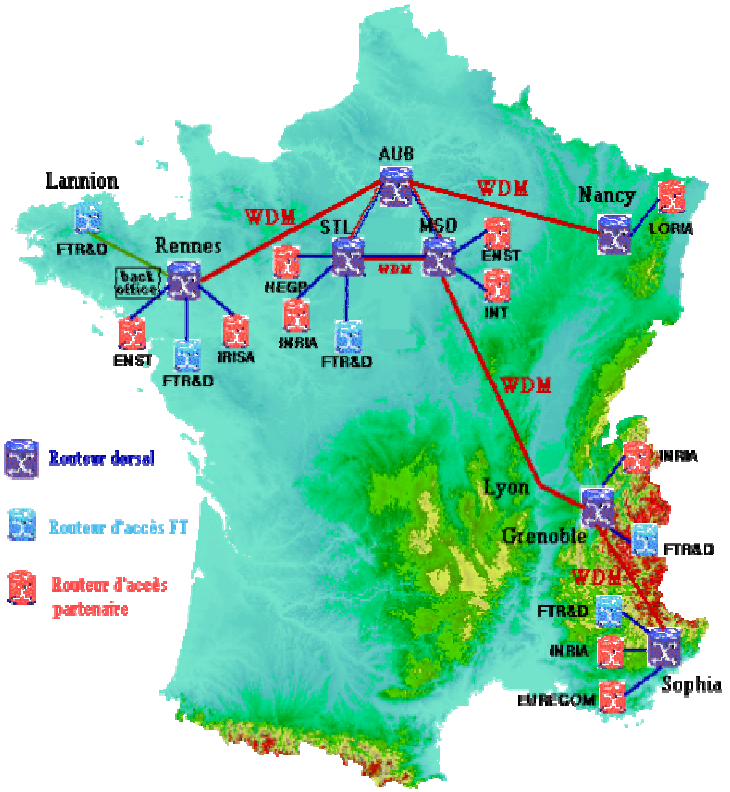
17 access routers to aggregate local traffic network over

Giga-Ethernet fiber

IPv6 network interconnected by IPv6/GEth/MPLS or IPv6/IPv4 tunnels

VTHD computational grid

Inria Rennes
34 nodes cluster
SCI / Myrinet / Gigabit Ethernet



Inria Rhône-Alpes
HP Cluster - 225 nodes

Inria Sophia-Antipolis
16 nodes cluster

Context of the experiments

❑ At the beginning of the project

- **« 2.5 gigabits is probably too much for PCs that are connected to the VTHD network with 100 Mbits/s Ethernet interfaces … »**

❑ Goal of the research works

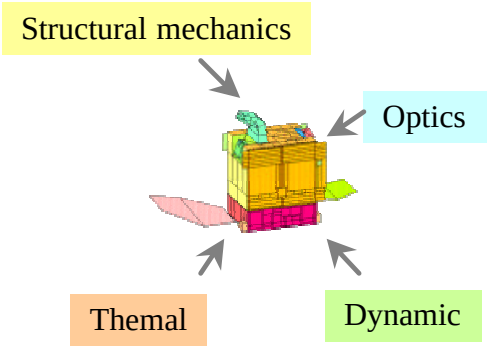
- **Grid = cluster of cluster**
- **Design middleware that allow transparent exploitation of high-bandwidth network**
  - **For scientific applications (parallel codes)**

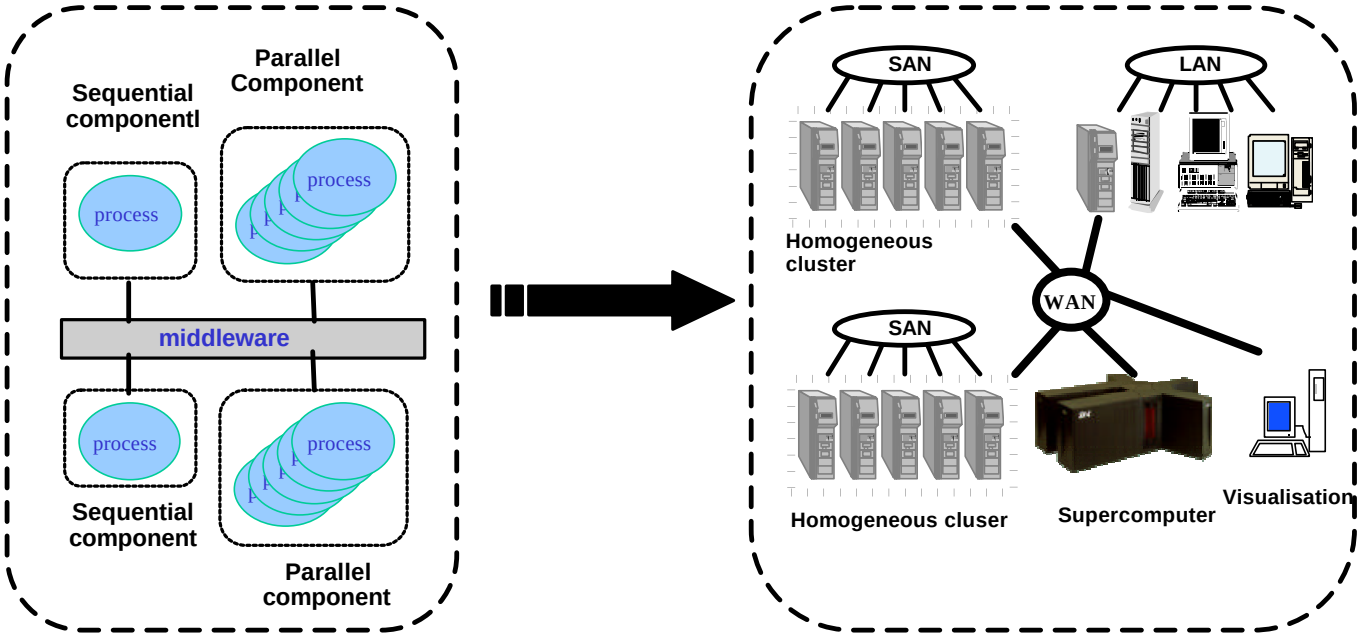
❑ Principle

- **Exploit data distribution within a parallel code to exploit high-bandwidth networks**

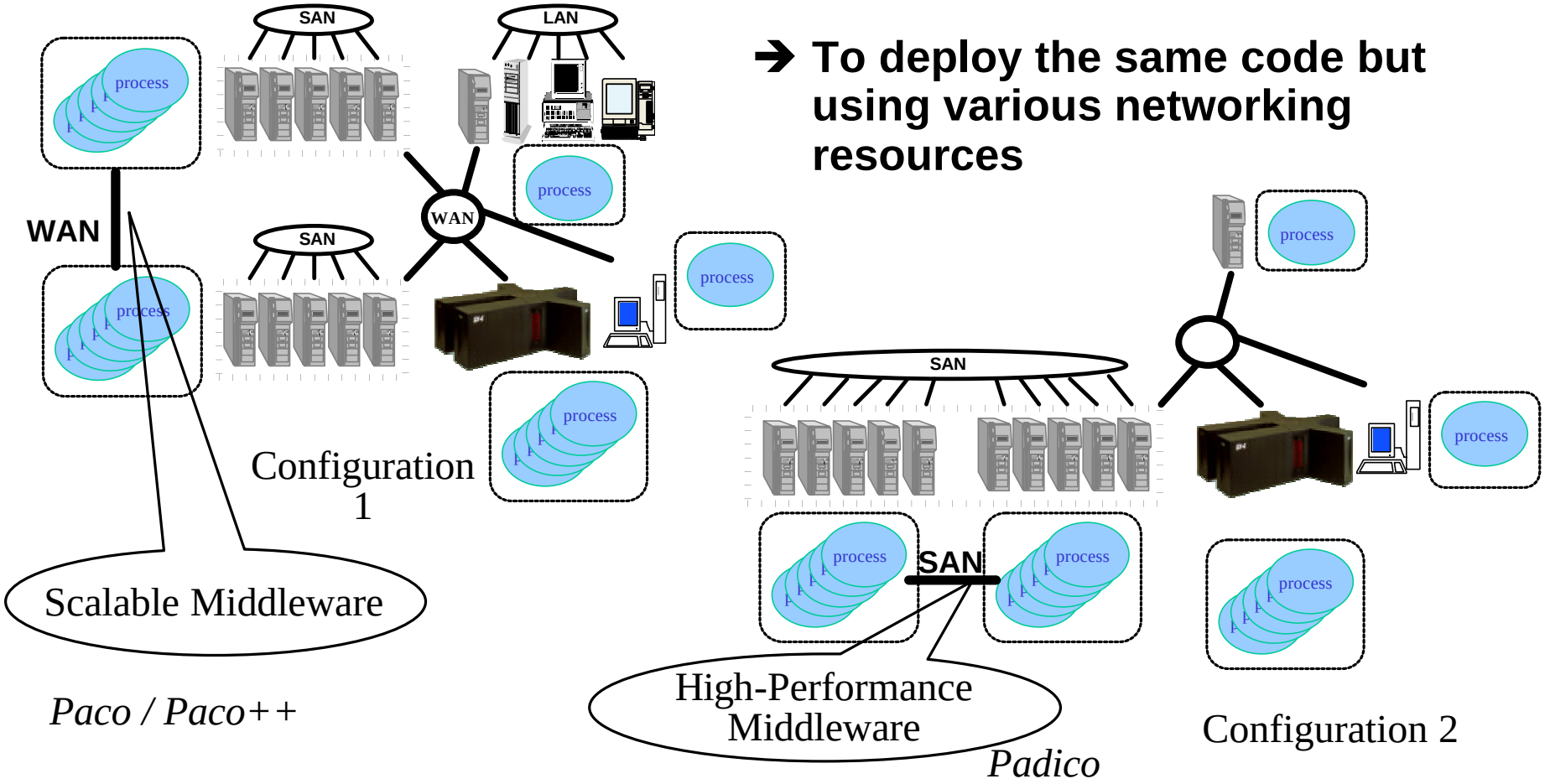
Distributed objects oriented platforms

- **Distributed objects / Components**
  - **To enforce users to structure their applications**
  - **Parallel codes encapsulation**
- **Coupling**
  - **Through a scalable middleware**





Mapping of components onto resources

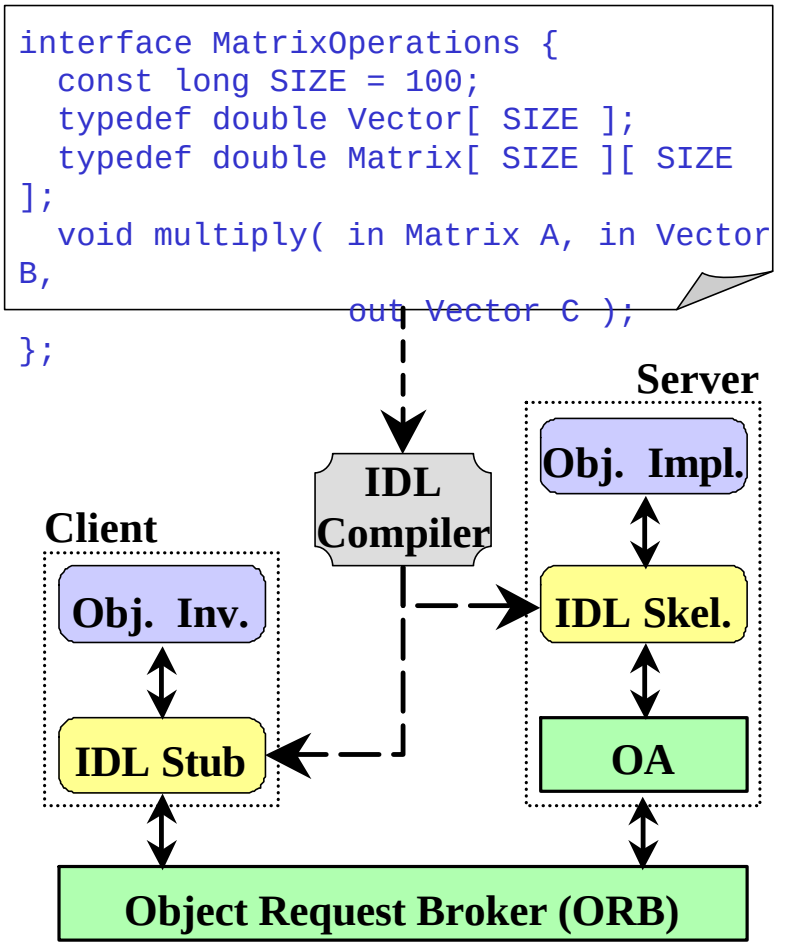


➔ **To deploy the same code but using various networking resources**

➔ **The middleware should adapt itself to the available networking resources**

CORBA : a middleware for scientific computing

- CORBA: Common Object Request Broker Architecture
- Open standard for distributed object computing by the OMG
  - Software bus, object oriented
  - Remote invocation mechanism
  - Hardware, operating system and programming language independence
  - Vendor independence (interoperability)
- Problems to face
  - Performance issues (not true)
  - Poor integration of high performance computing environments

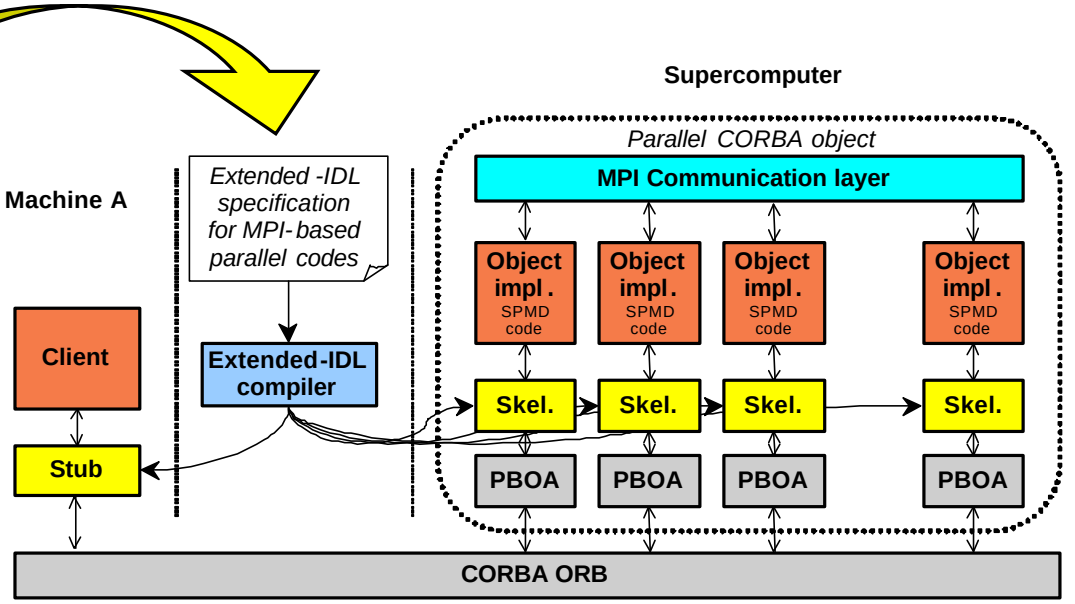


Parallel CORBA object concept

❑ Goal

- ◆ Encapsulation of parallel codes into CORBA objects
- ◆ Scalable "connection" between parallel CORBA objects

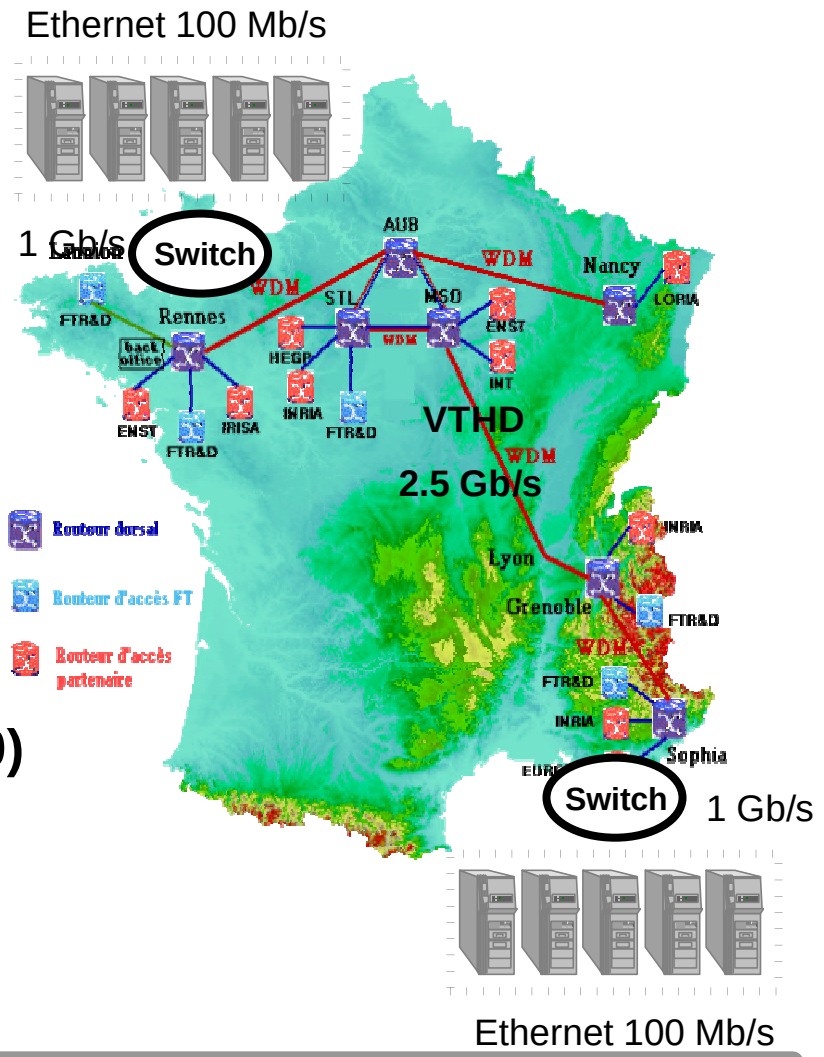
```
interface[*:2*n] MatrixOperations {
  const long SIZE = 100;
  typedef double Vector[SIZE];
  typedef double Matrix[SIZE][SIZE];
  void multiply(in dist[BLOCK][*] Matrix A,
                in Vector B,
                out dist[BLOCK] Vector C );
  void skal(in dist[BLOCK] Vector C,
            out csum double skal);
};
```



❑ Submitted to the OMG as a response to a RFI

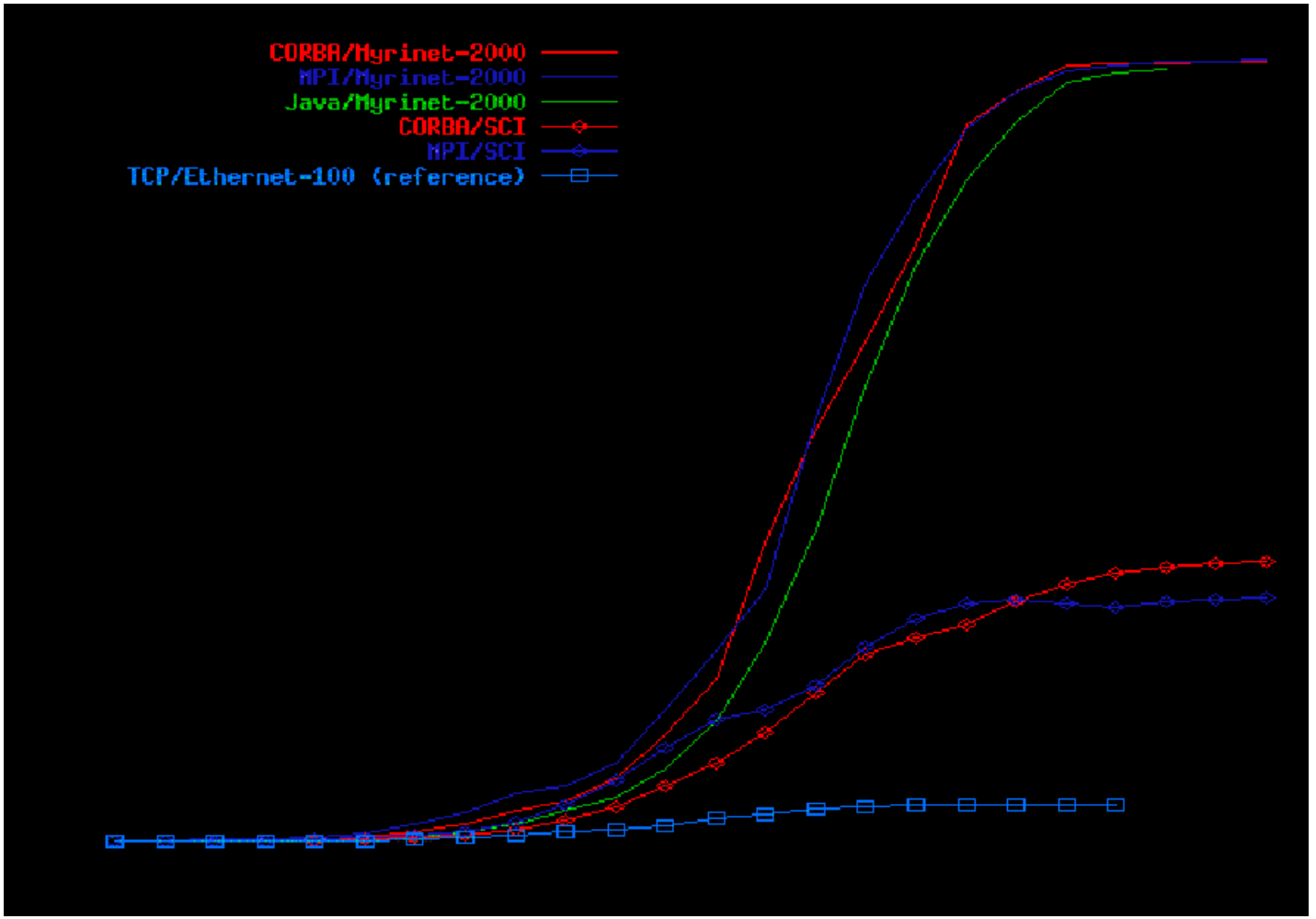
VTHD Experiment

- **Protocols**
  - Two MPI codes
  - Encapsulation using PaCO++
- **VTHD Performance**
  - 11 nodes to 11 nodes
  - 826 Mb/s (103 MB/s)
  - P2P: 75 Mb/s (9.4 MB/s)
  - Near the switch limit
- **Cluster Performance**
  - 8 nodes to 8 nodes (Myrinet 2000)
  - 12 Gbit/s (1.5 GB/s)
  - P2P: 1.5 Gbit/s (187 MB/s)



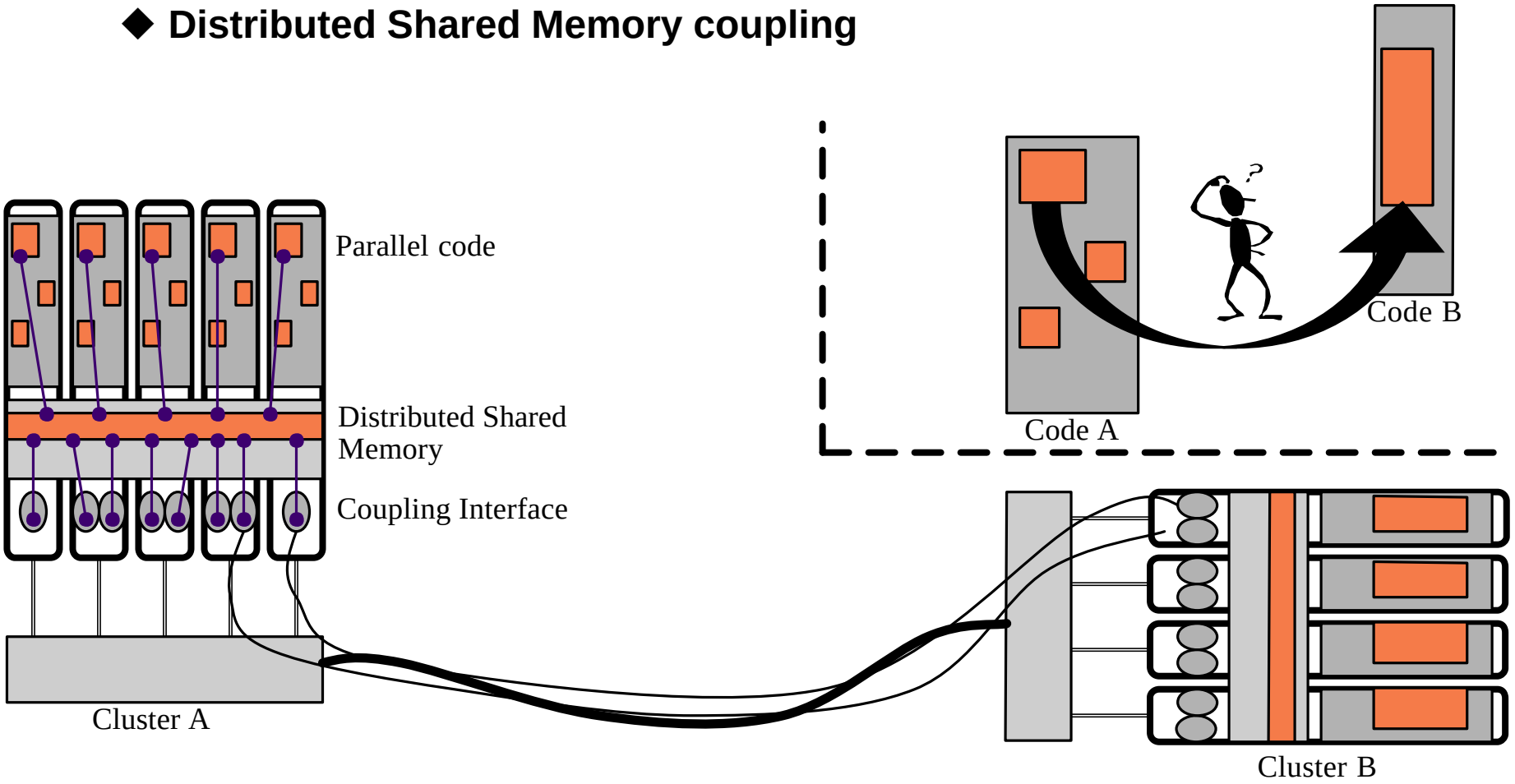
Cluster performance

- **PADICO: an Open Integration platform for communicating middleware and runtimes**



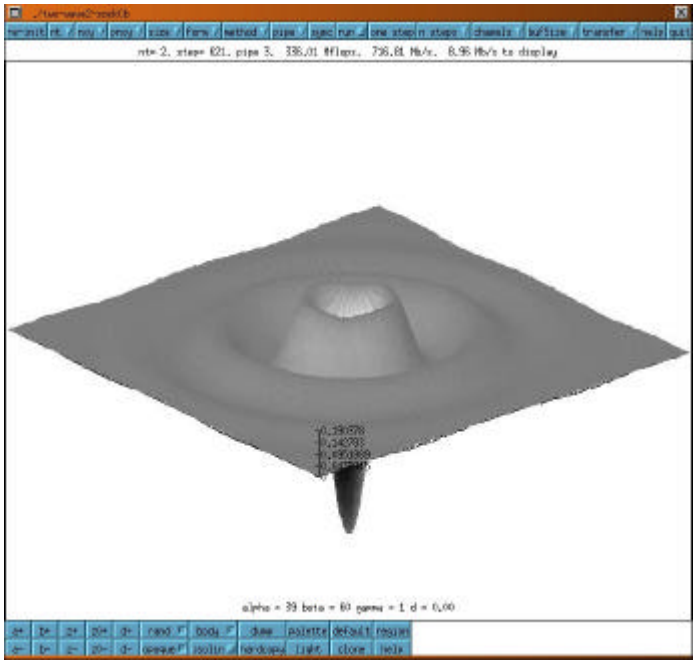
Global address space for the Grid (MOME)

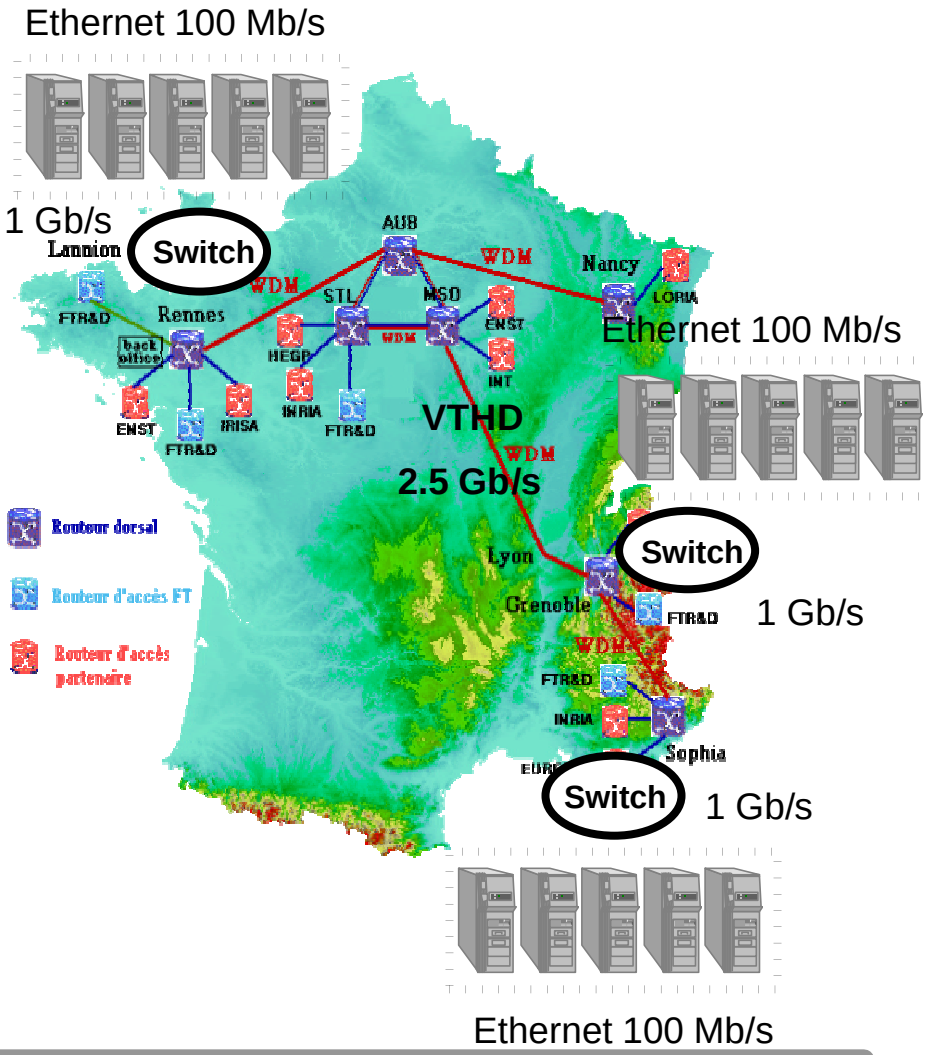
- **Goal**
  - **To transfer data between DSM-based parallel codes**
  - **Distributed Shared Memory coupling**



VTHD Experiment

- **Protocol**
  - **Three codes coupled with MOME (Simulation + visualisation)**
- **VTHD Performance**
  - **11 nodes to 11 nodes**
  - **850 Mb/s (106 MB/s)**





The JACO3 Environment

- A set of CORBA services to support coupled simulation applications