The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

1967-25

Advanced School in High Performance and GRID Computing

3 - 14 November 2008

High Bandwidth Data Transfer

BARRIOS HERNANDEZ Carlos Jaime
Laboratoire Informatique de Grenoble LIG
Montbonnot Sain Martin, 38330
Grenoble
FRANCE

High Bandwidth Data Transfer in Grids

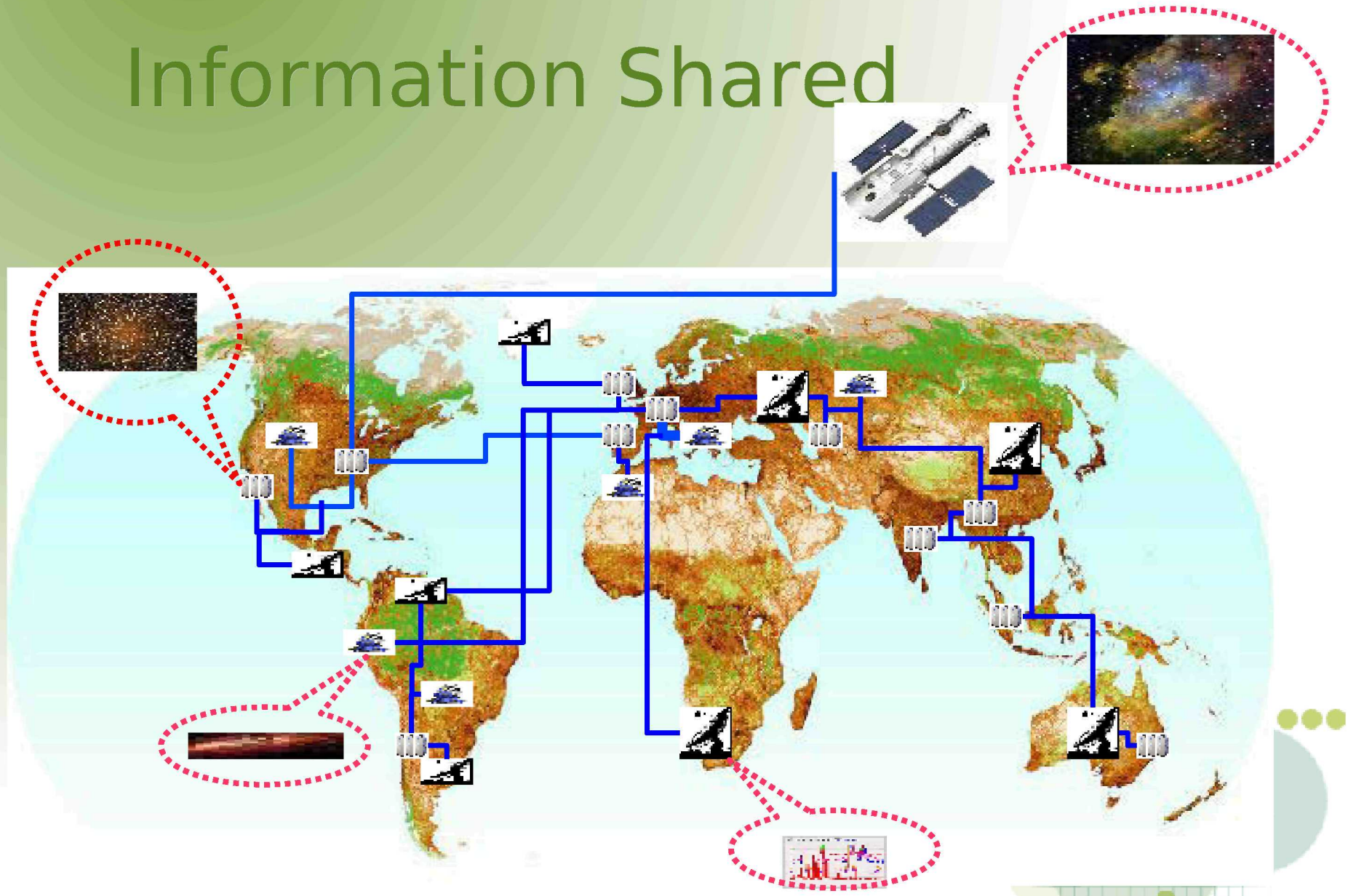
Carlos Jaime BARRIOS-HERNÁNDEZ

Laboratoire de Informatique de Grenoble
INRIA Equipe Mescal
Montbonnot-St Martin

Université de Nice à Sophia Antipolis
Laboratoire de Informatique, Signaux et Systèmes de Sophia Antipolis
Equipe Rainbow
Sophia-Antipolis

Advanced School in High Performance and GRID Computing
ICTP -Trieste, Italy, November 2008

Information Shared

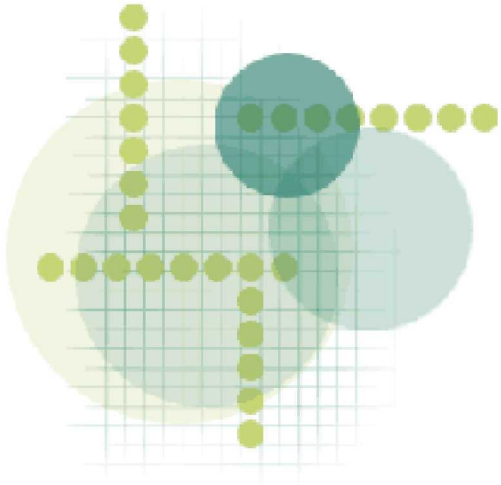


Information Transfer

- Data has different formats (and sizes)
- Data is transferred upon heterogeneous channels
  - Protocols
  - Physical links, devices, networks technologies
- Data are stored of different ways
  - File systems
- Data is treated (access – process) in concurrency
  - Synchronous and asynchronous ways
- Data Production and Consumption grows.
- Data transfer (storage and process) is not safe.

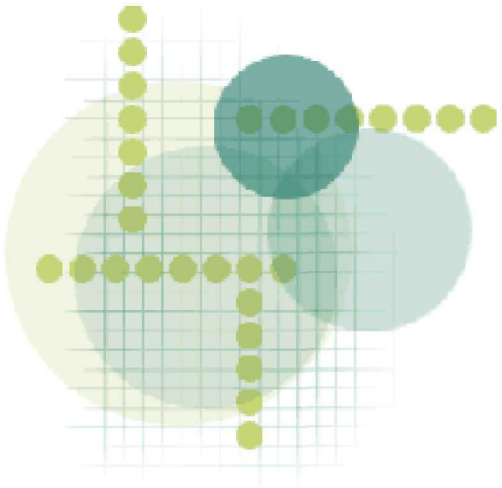
The Critical Point

- In Grid, frequently, the time to transfer data volumes is more important that the time to process or compute data.

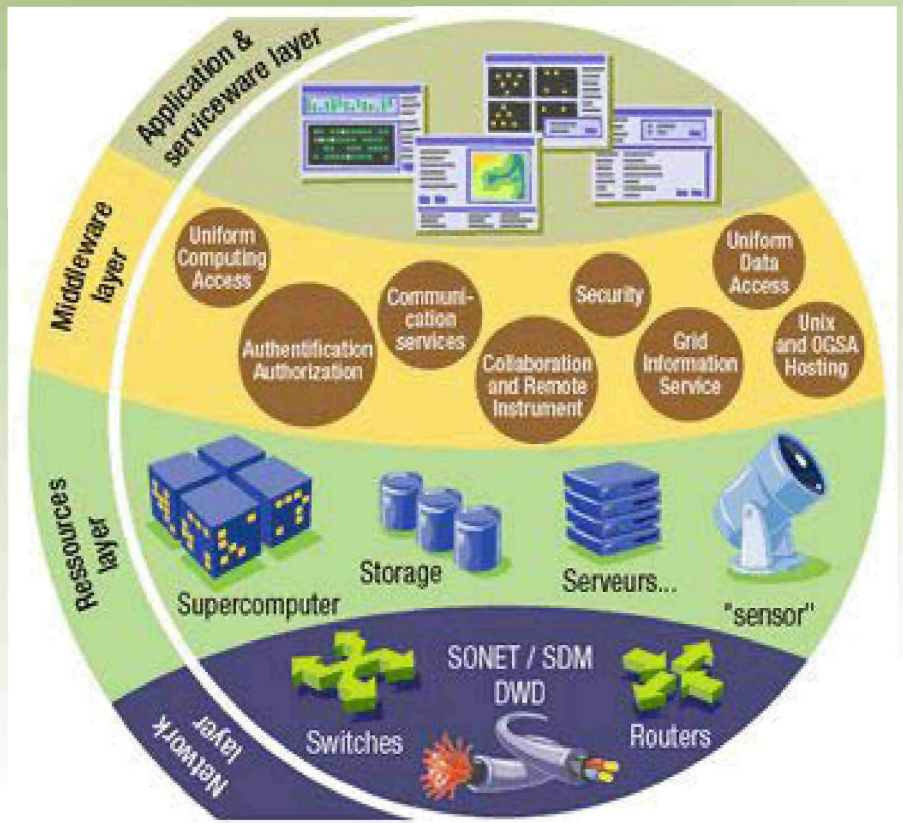


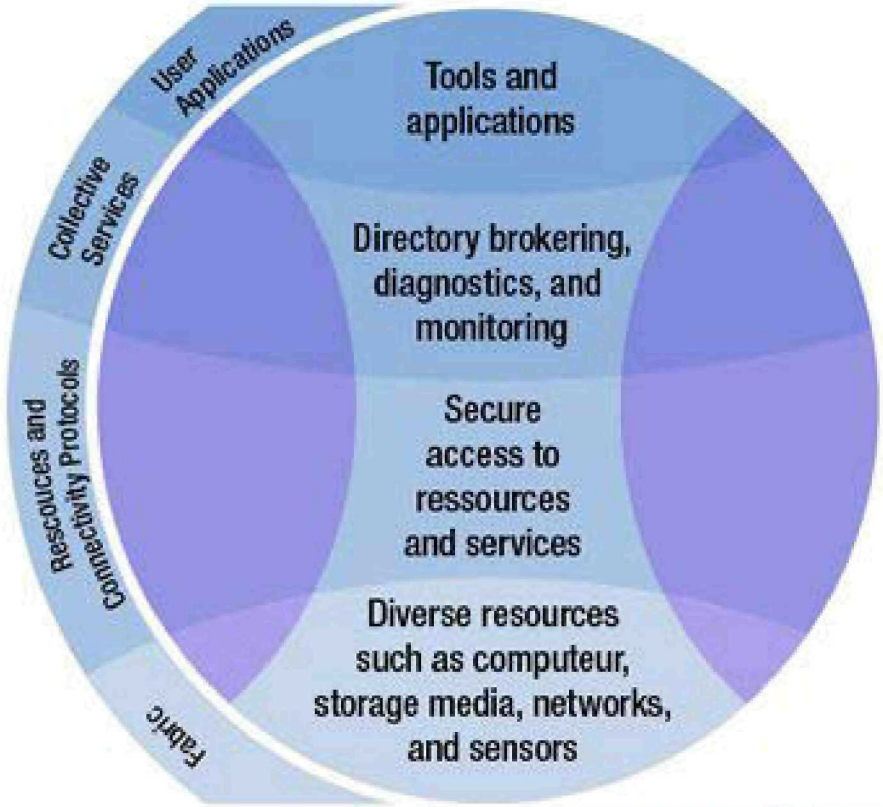
Outline

- High Bandwidth Data Transfer ... in Grid
  - Massive / Intensive Data Transfer
  - Concurrency
  - Synchronization
  - Storage
- Modeling
- Summary



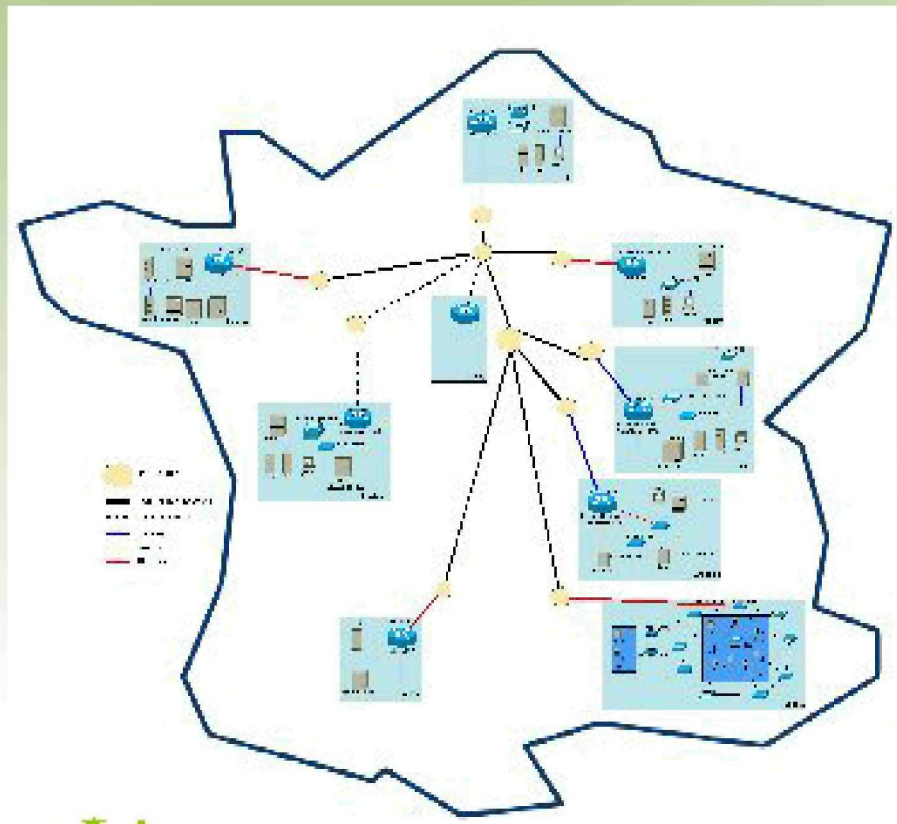
Grid Layers





From GridCafe.org

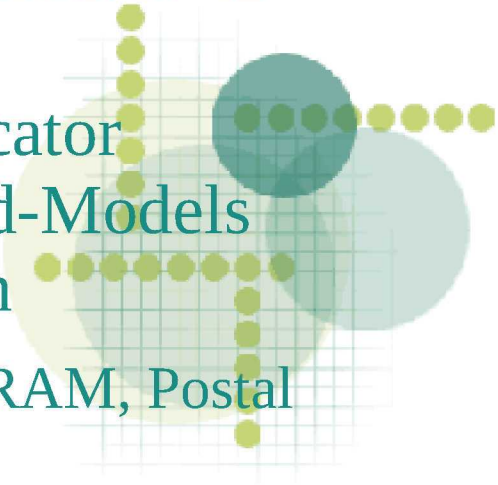
High Bandwidth Data Transfer... in Grids



- Data changes of environment
  - Local Platform Transfer (IntraCluster)
  - External Transfer (ExtraCluster)
- Data is massive / intensive
- Data are in concurrence

Data Treatment Perspective

- Intrastructure Point of View
  - Technological advantages to treat data volumes (devices, physical connexions)
  - Protocols
    - GridFTP, GXFER, TCP
  - File Systems
    - NFSp, PFS, dNFSp, Lustre, XtreemFS
- Application Point of View
  - Commicator Algortihms
    - Collective communicators implementations in MPI
  - Communicator Distributed-Models Adaptation
    - LogP, PRAM, Postal



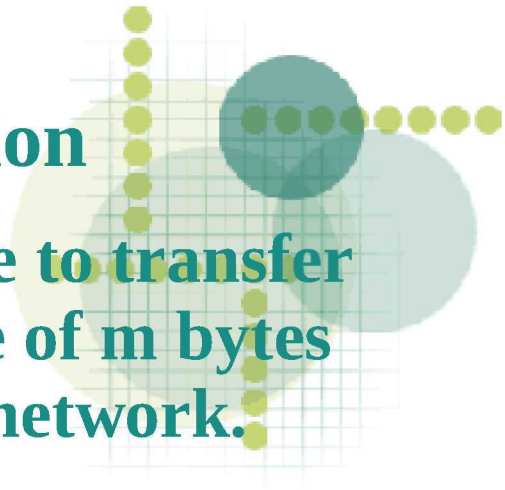
Data Treatment Perspective

- Intrastructure Point of View
  - Technological advantages to treat data volumes (devices, physical connexions)
  - Protocols
    - GridFTP, GXFER, TCP
  - File Systems
    - NFSp, PFS, dNFSp, Lustre, XtreemFS
- Application Point of View
  - Commicator Algortihms
    - Collective communicators implementations in MPI
  - Communicator Distributed-Models Adaptation
    - LogP, PRAM, Postal

IT EXISTS A MUTUAL INFLUENCE!!

High Bandwidth Data Transfer

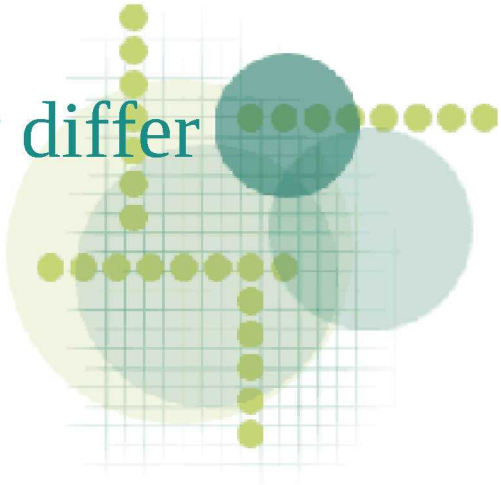
- **Bandwidth**
  - Capacity for a given system to transfer data over a connection (wikipedia).
- Normally is **confused** with **throughput**
  - Average rate of successful message delivery over a communication channel (wikipedia)
- **Latency**
  - Time to send a message of m bytes between emisor to receptor.
- Normally is confused with **Time Transmission**
  - **Total time to transfer a message of m bytes among a network.**



Interpeting Latency and Bandwidth

From W.Groop Modeling Performance Course

- Simplest model $s + r\,n$
- s includes both hardware (gate delays) and software (context switch, setup)
- r includes both hardware (raw bandwidth of interconnection and memory system) and software (packetization, copies between user and system)
- head-to-head and pingpong values may differ



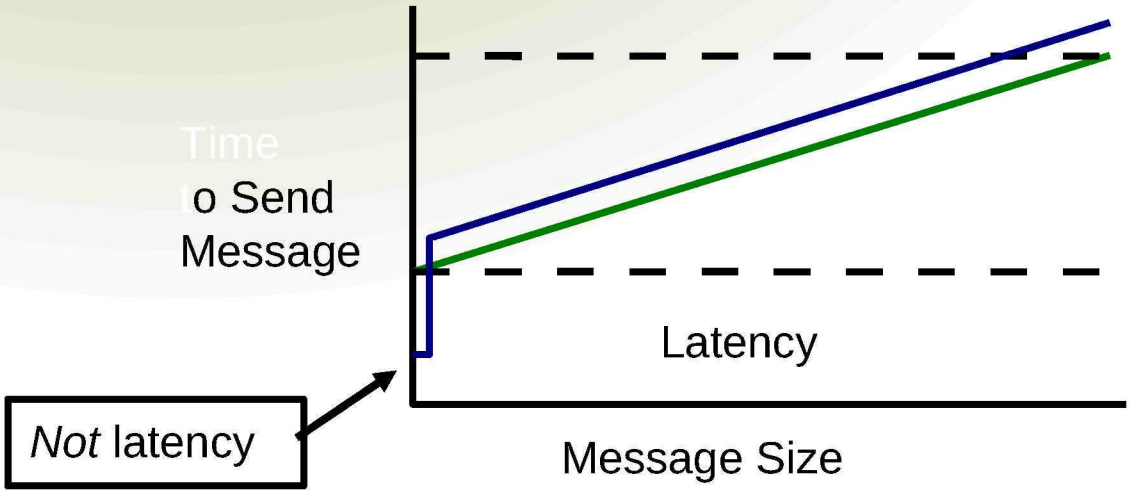
Interpreting Latency and Bandwidth

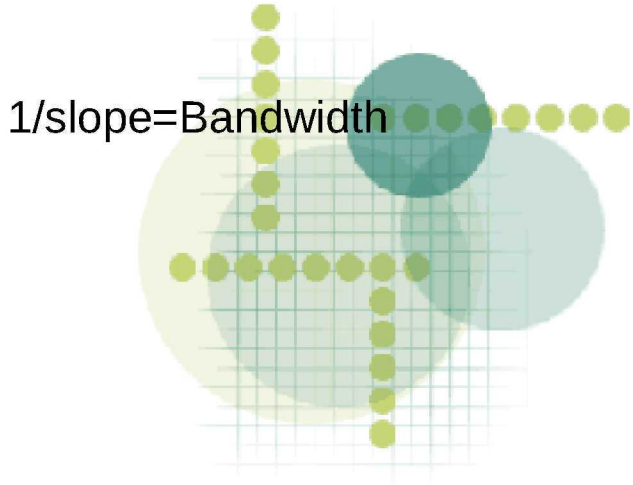
From W.Groop Modeling Performance Course

- Bandwidth is the inverse of the slope of the line

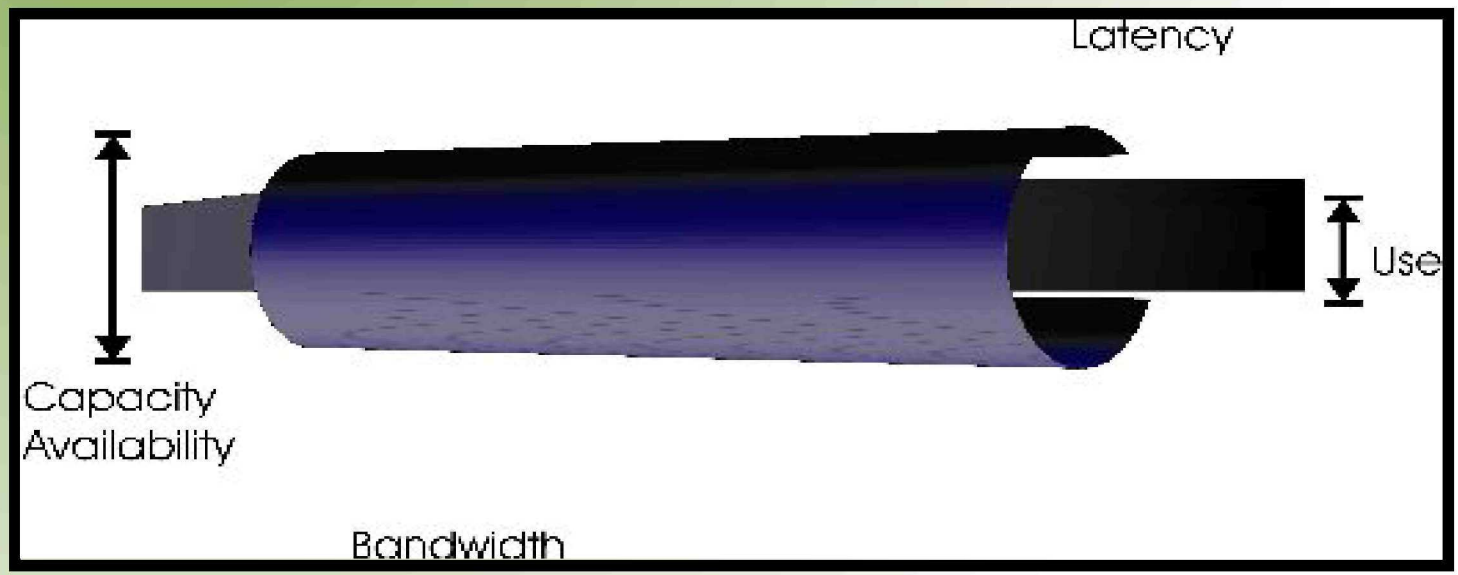
  time = latency + (1/rate) size_of_message

- Latency is sometimes described as “time to send a message of zero bytes”. This is true *only* for the simple model. The number quoted is sometimes misleading.



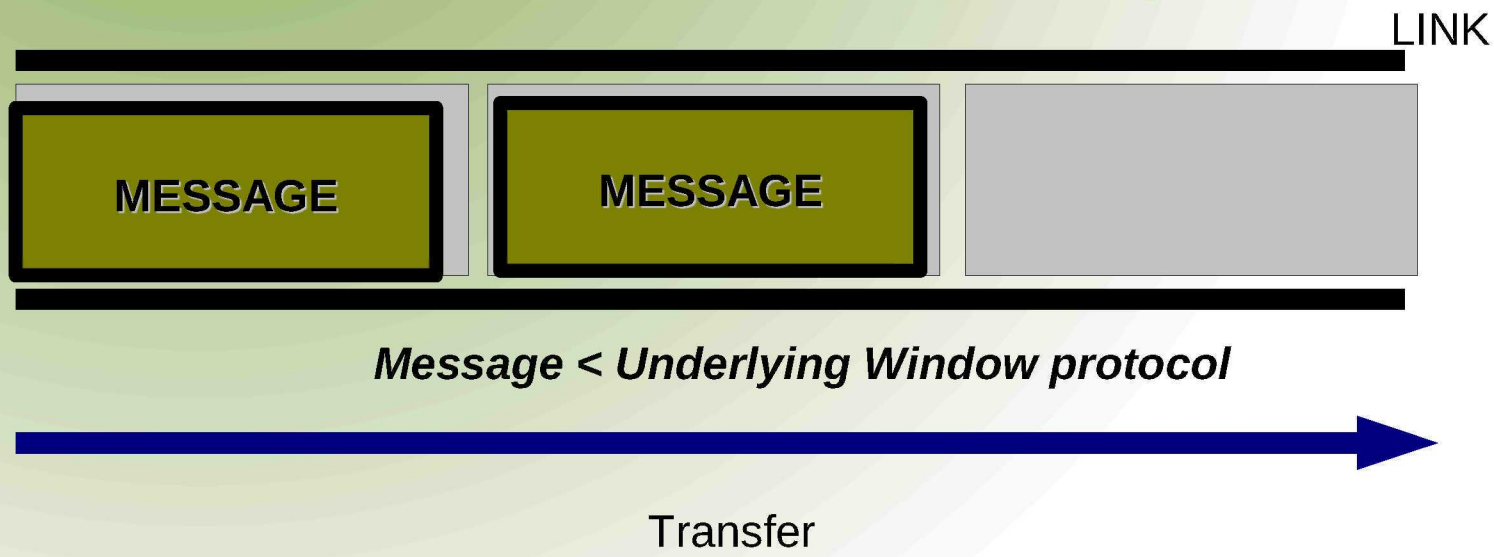


Intepreting Latency and Bandwidth



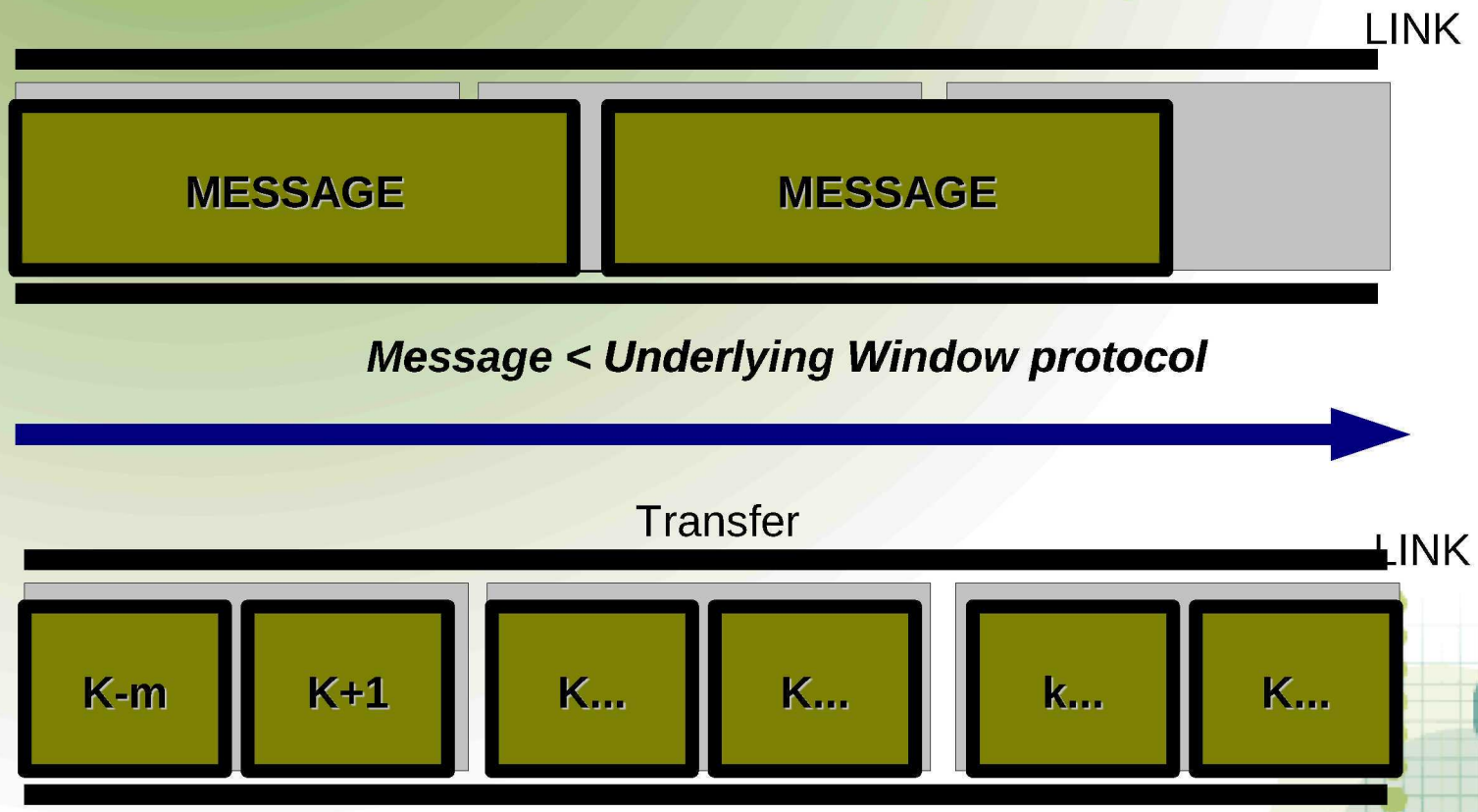
- **Bandwidth :**
  - Capacity ; Availability
- **Latency:**
  - Use
- **Time Transmission:**
  - Time to Send + Time to Receive + Latency + (Network or Applications Delays) (Minimalistic conception)

Massive and Intensive Data Transfer (1/3)



- In a transfer, normally, the message is not very great that the underlying transmission capacity of the communication protocol (i.e. tcp window)

Massive and Intensive Data Transfer (2/3)



- A very great size message (or long) is cut in “packets” to be transferred among the link.

Transfer

Massive and Intensive Data Transfer (3/3)

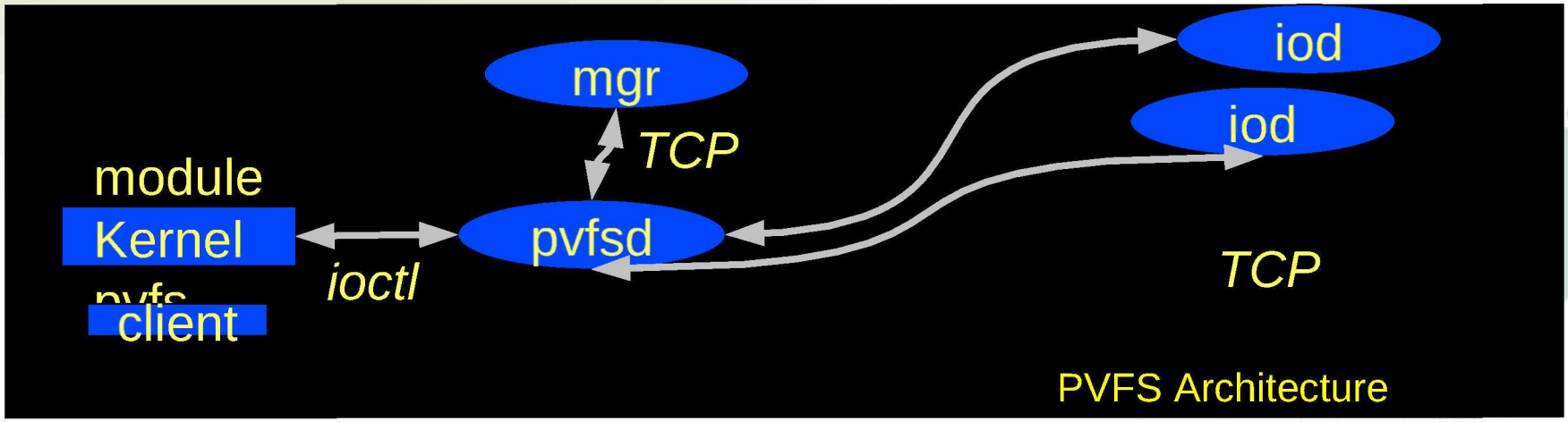
- Massive Data Transfer occurs when a great message (m) is transferred.
- As m is cut in packets, there are a consecutive transfer of k-packets of m, then a massive data transfer can be described in terms of massive data transfer.
- There are additional costs related with massive/intensive data transfer:
  - Gap, overheads, contention, delays due to saturation.
- However, Grid Infrastructure guarantees high bandwidth transmission (in general)

Concurrence

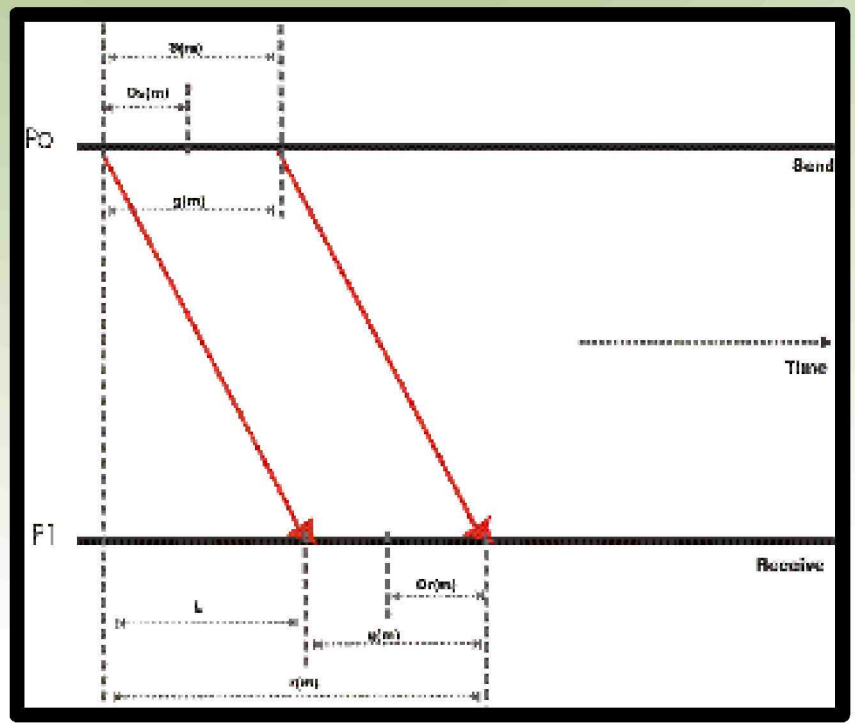
- Programs may be generate collective communications
- In Grid Platforms, never we alone.
- Transfer may be asynchronous or synchronous.
  - For example, when you use mpi barriers.
- « Users » can be to need the sames data sets or data volumes.

Storage

- Data access among high bandwidth networks
- It exist solutions with distributed storage systems
  - Expensives (in technical terms)
    - New Protocols, new File Systems
    - PVFS, NFSp, dNFSp



Modeling
(Example)

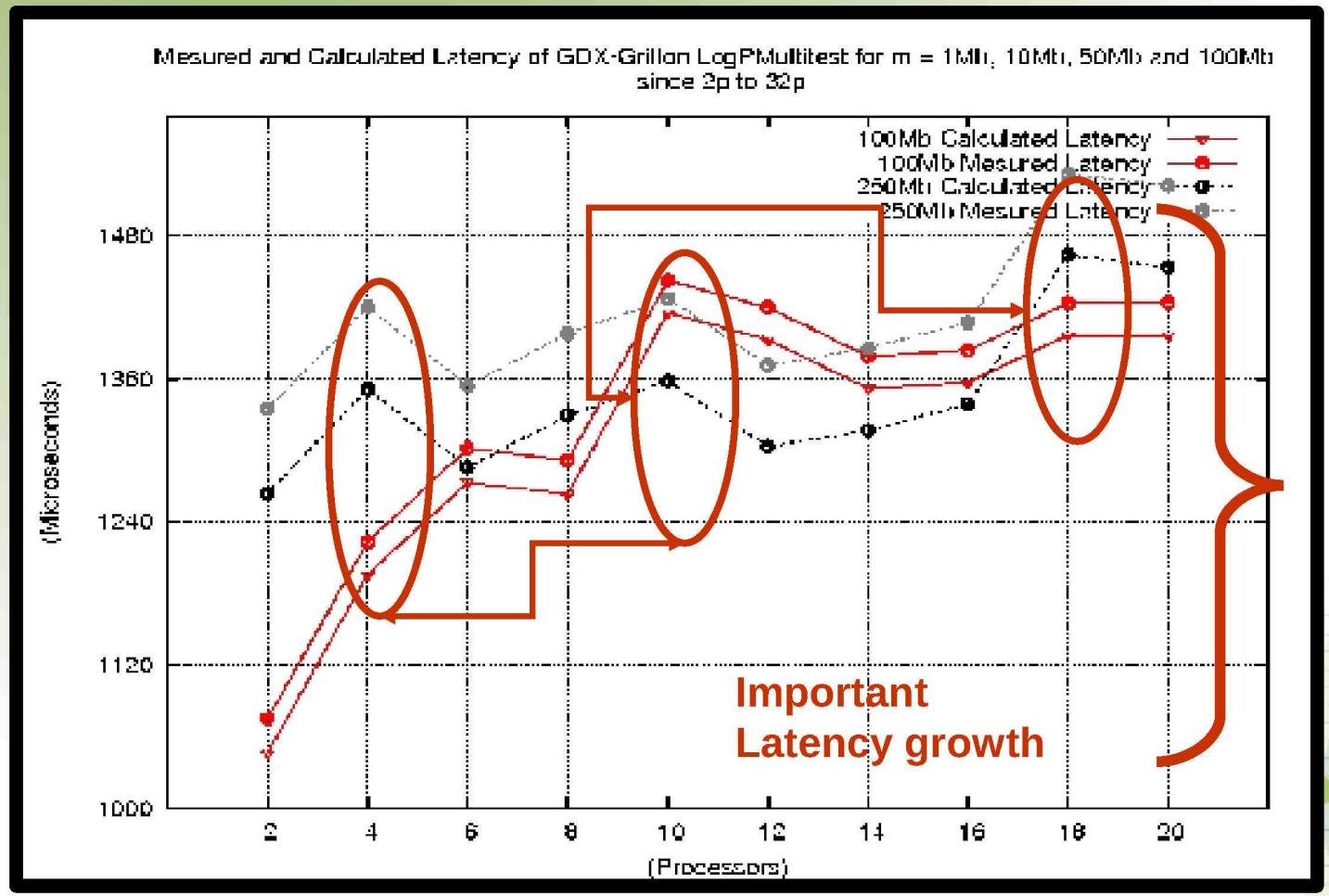


LogP/LogGP	LogGp Parameters
L	= L + g(1)-Os(1)-Or(1)
o	= (Os(1) +Or(1)/2
g	= g(1)
G	= g(m)/m

[*] LogGP (Alexandrov et al, pLogP (Kielman et al)

Example of use of the model to Performance Evaluation

- **GDX-Grillon Calculated and Measured Latency**

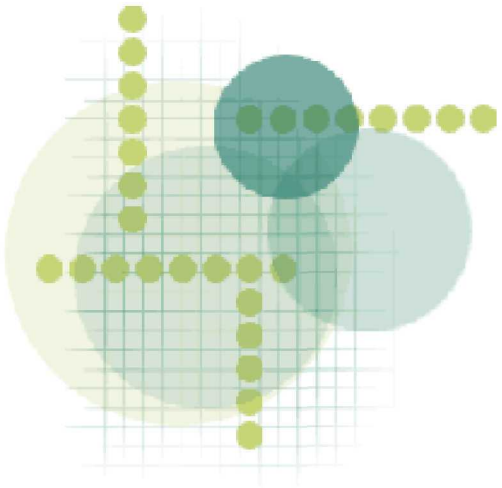


Unstable-Regular behavior

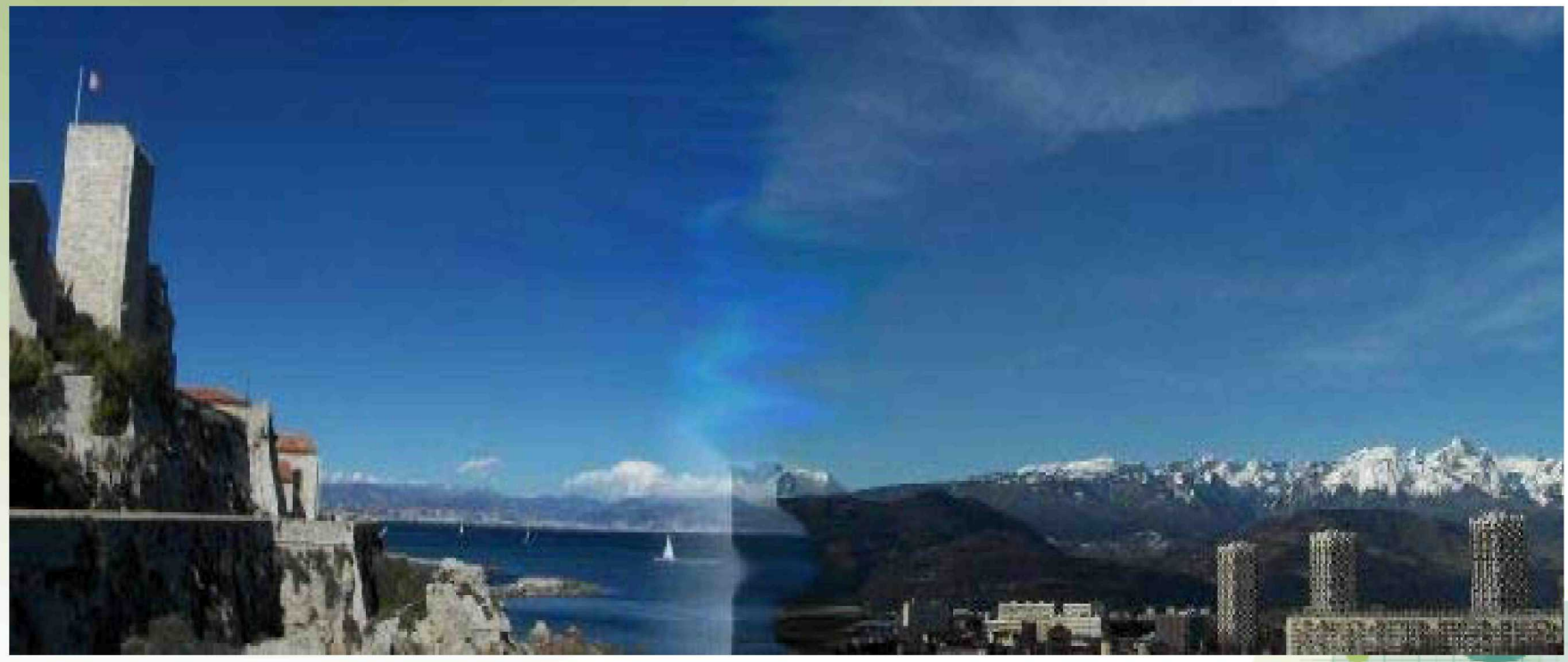
From Barrios and Denneulin EXPEGRID HPDC 2006

Summary

- In Grid , Data transfer is not only an infrastructure or technology problem.
- In Grid, Data Storage is a technology problem.
- Knowledge grows, information grows, data transfer grows, data storage grows... and the efficiency in data/storage?



The problem is not to know too much... it's not to know what make with the information.



Thanks, merci, gracias, grazie

Next Year

- Latin American Conference on High Performance Computing (CLCAR 2009)
  - **Choroni, Venezuela, 13-19 September 2009**
  - Call for papers open in december 2008 to May 2009
  - For more information:
    - cjbarrioshernandez@ieee.org
    - nunez@ula.ve