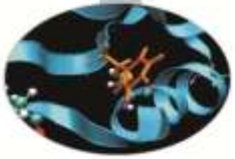
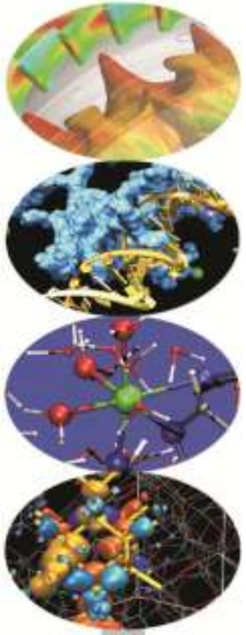


Management of large scientific data



Giovanni Morelli, Giuseppe Fiameni

{g.morelli, g.fiameni}@cineca.it

SuperComputing Applications and Innovation Department

Agenda

- **Bulk data transfer**
  - Tools and techniques
- **BigData techniques**
  - Hadoop/MapReduce
- **Data post-processing**
  - Remote visualization

Bulk data transfer is a software application feature that uses data ***compression, data blocking and buffering*** *to optimize transfer rates when moving large data files*

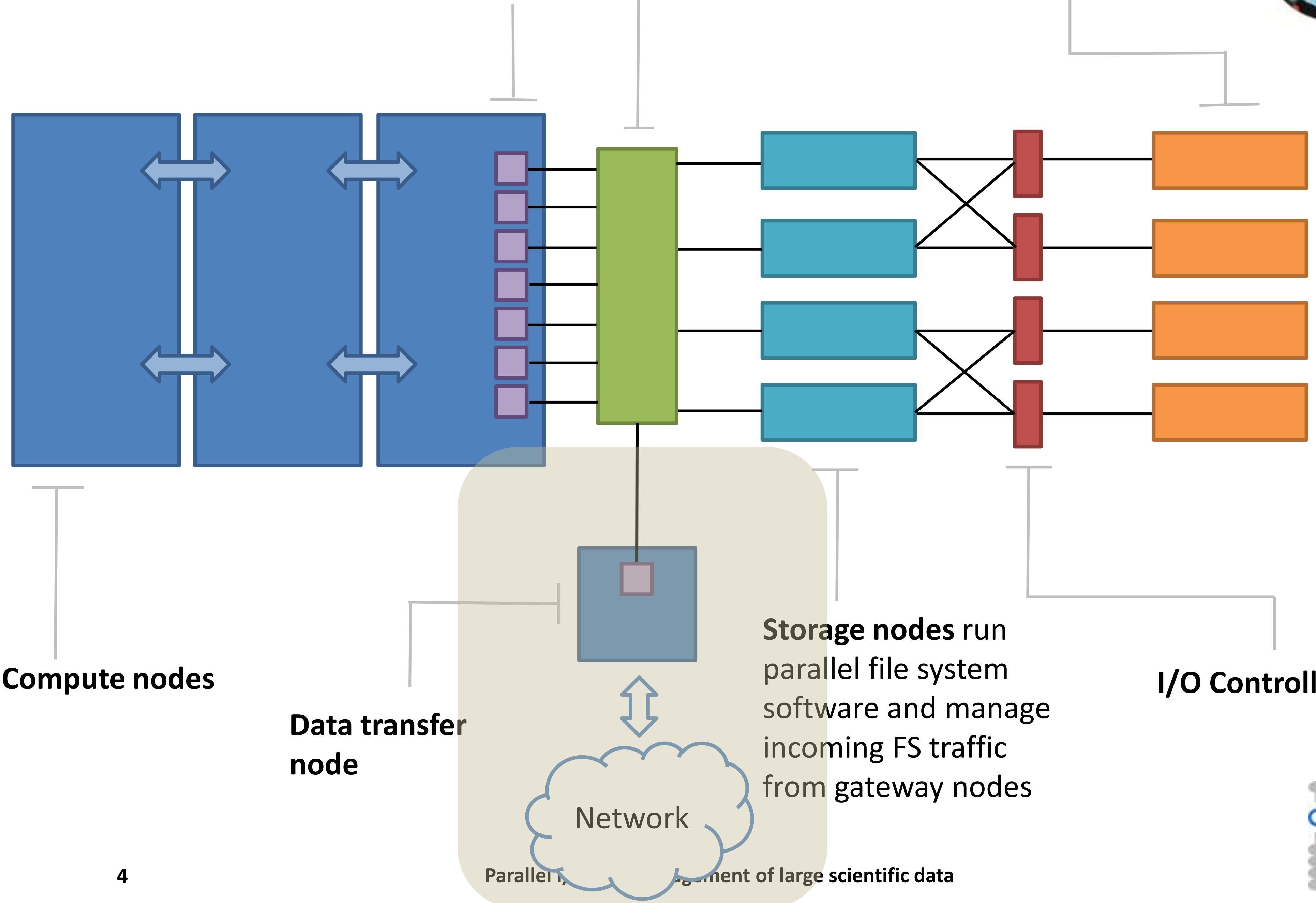
Agenda

- **Bulk data transfer**
  - Tools and techniques
- **BigData techniques**
  - Hadoop/MapReduce
- **Data post-processing**
  - Remote visualization

Gateway nodes run parallel file system client

Commodity network

Enterprise storage



Compute nodes

Data transfer node

Storage nodes run parallel file system software and manage incoming FS traffic from gateway nodes

I/O Controllers

Bulk data movement

- **The problem**
- Involved components
  - Network architecture
  - Dedicated hosts
  - Software tools

PHOTO: DAVIES & STARR

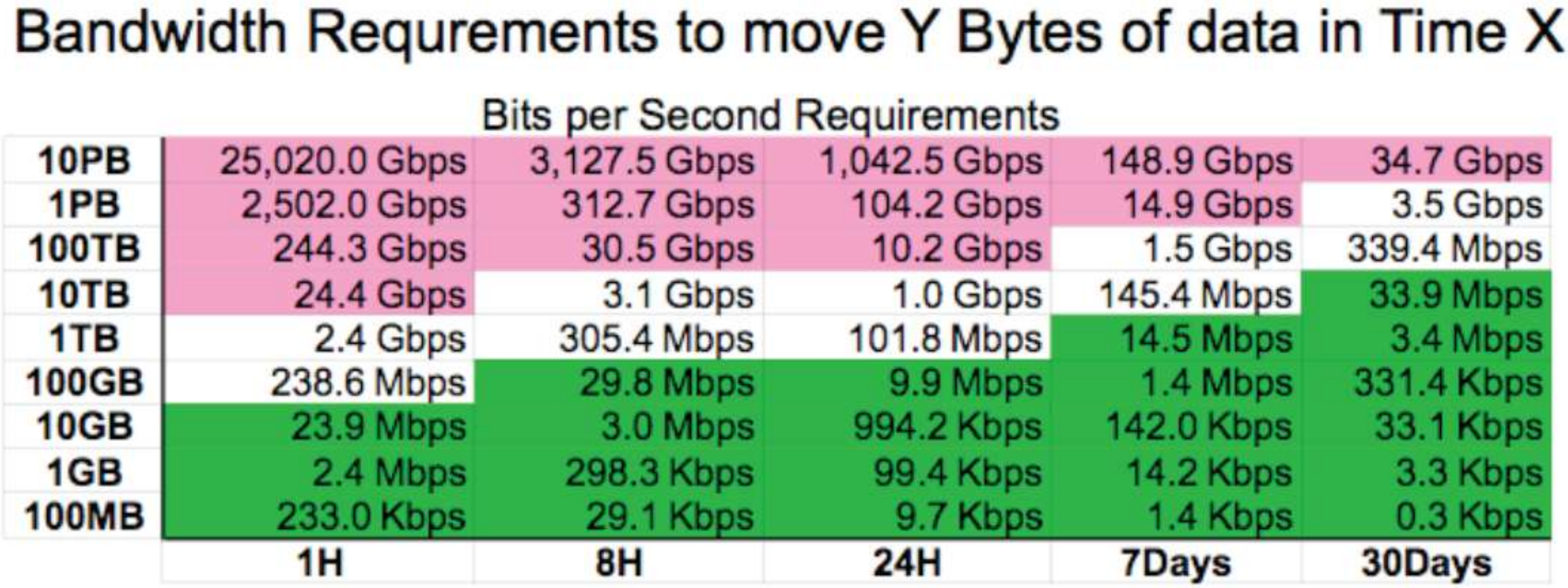
Bulk Data Movement

- Common task at all data scales
- Driven by collaboration, distributed resources
  - Computing centers
  - Facilities
  - Major instruments (e.g. LHC)
- Fundamental to the conduct of science (scientific productivity follows data locality)
- Data sets of 200GB to 5TB are now common
- Often a difficult task for various reasons
- Storage capacity grows faster with respect to Public Network bandwidth

Time to copy 1TB

- **10 Mb/s network:** 300 hrs (12.5 days)
- **100 Mb/s network:** 30 hrs
- **1 Gb/s network:** 3 hrs (are your disks fast enough?)
- **10 Gb/s network:** 20 minutes (need *really fast disks and* file system)
- **Compare these speeds to:**
  - USB 2.0 portable disk
    - 60 MB/sec (480 Mbps) peak
    - 20 MB/sec (160 Mbps) reported on line
    - 15-40 hours to load 1 Terabyte

Data Throughput – Transfer Times



Bandwidth Requrements to move Y Bytes of data in Time X

Bits per Second Requirements

	1H	8H	24H	7Days	30Days
10PB	25,020.0 Gbps	3,127.5 Gbps	1,042.5 Gbps	148.9 Gbps	34.7 Gbps
1PB	2,502.0 Gbps	312.7 Gbps	104.2 Gbps	14.9 Gbps	3.5 Gbps
100TB	244.3 Gbps	30.5 Gbps	10.2 Gbps	1.5 Gbps	339.4 Mbps
10TB	24.4 Gbps	3.1 Gbps	1.0 Gbps	145.4 Mbps	33.9 Mbps
1TB	2.4 Gbps	305.4 Mbps	101.8 Mbps	14.5 Mbps	3.4 Mbps
100GB	238.6 Mbps	29.8 Mbps	9.9 Mbps	1.4 Mbps	331.4 Kbps
10GB	23.9 Mbps	3.0 Mbps	994.2 Kbps	142.0 Kbps	33.1 Kbps
1GB	2.4 Mbps	298.3 Kbps	99.4 Kbps	14.2 Kbps	3.3 Kbps
100MB	233.0 Kbps	29.1 Kbps	9.7 Kbps	1.4 Kbps	0.3 Kbps

This table available at **http://fasterdata.es.net**

Bulk data movement

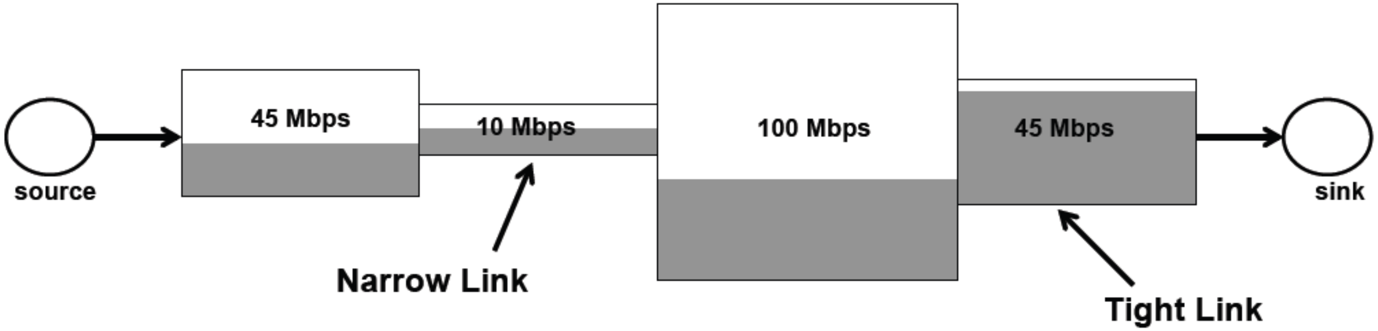
- The problem
- Involved components
  - **Network architecture**
  - Dedicated hosts
  - Software tools

PHOTO: DAVIES & STARR

Terminology

The term "Network Throughput" is vague and should be avoided

- **Capacity**: link speed
  - **Narrow Link**: link with the lowest capacity along a path
  - Capacity of the end-to-end path = capacity of the narrow link
- **Utilized bandwidth**: current traffic load
- **Available bandwidth**: capacity – utilized bandwidth
    - **Tight Link**: link with the least available bandwidth in a path
- **Achievable bandwidth**: includes protocol and host issues



Courtesy of the Berkley Lab

Network architecture

- Most LANs are not purpose-built for science traffic they carry many types of traffic
  - Desktop machines, laptops, wireless
  - VOIP
  - HVAC control systems
  - Financial systems, HR
  - *Some science data coming from someplace*
- Bulk data transfer traffic is typically very different than enterprise traffic

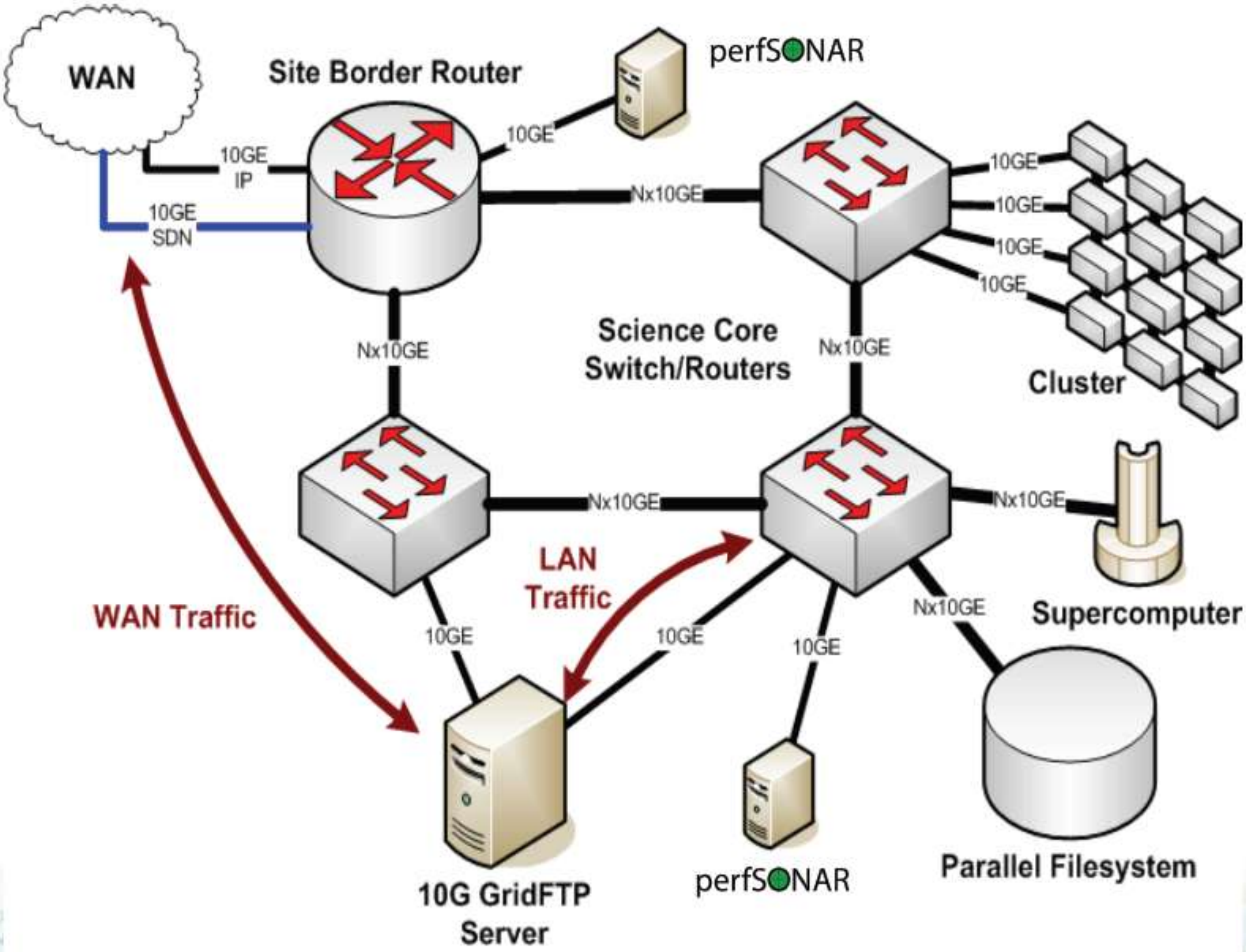
Bulk data movement

- The problem
- Involved components
  - Network architecture
  - **Dedicated hosts**
  - Software tools

Data transfer nodes

- **Reasons for dedicated hosts**
  - One thing to test and tune
  - One place for large WAN flows to go (it’s easier to give one host a special configuration than to do this for all workstations)
  - One set of firewall exceptions

Internal/external traffic

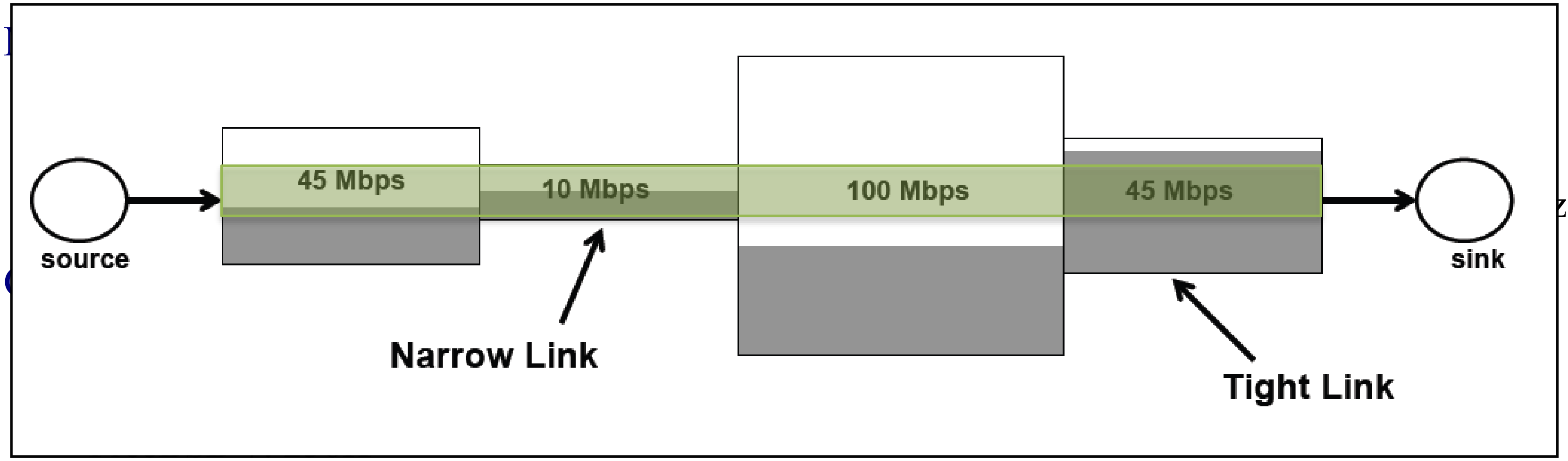


Host tuning - TCP

- TCP tuning commonly refers to the proper configuration of buffers that correspond to TCP windowing
- Historically TCP tuning parameters were host-global, with exceptions configured per-socket by applications
  - Applications had to understand the network in detail, and know how far away clients were
  - Some applications did this – most did not
- Solution: auto-tune TCP connections within preconfigured limits

Buffer autotuning

- To solve the buffer tuning problem, Linux OS added TCP Buffer autotuning
  - Sender-side TCP buffer autotuning introduced in Linux 2.4
  - Receiver-side autotuning added in Linux 2.6



  - Windows 7: 16M
  - Mac OSX 10.5: 8M
- Some defaults are still wrong!

Autotuning settings (Max 16MB)

- **Linux 2.6**

```
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216
# autotuning min, default, and max number of bytes to use
net.ipv4.tcp_rmem = 4096 87380 16777216
net.ipv4.tcp_wmem = 4096 65536 16777216
```

- **FreeBSD 7.0**

```
net.inet.tcp.sendbuf_auto=1
net.inet.tcp.recvbuf_auto=1
net.inet.tcp.sendbuf_max=16777216
net.inet.tcp.recvbuf_max=16777216
```

- **OSX 10.5 (“Self-Tuning TCP”)**

```
kern.ipc.maxsockbuf=16777216
```

Congestion control

- TCP senses network congestion by detecting packet loss
- Historically (TCP Reno) TCP used AIMD (Additive Increase, Mutiplicative Decrease) for window sizing in response to loss
- After loss, window opens back up very slowly
  - causes very poor performance
- Newer algorithms, available in Linux, offer higher performance than Reno
  - Cubic (now the default in several Linux distributions)
  - HTCP (Hamilton)

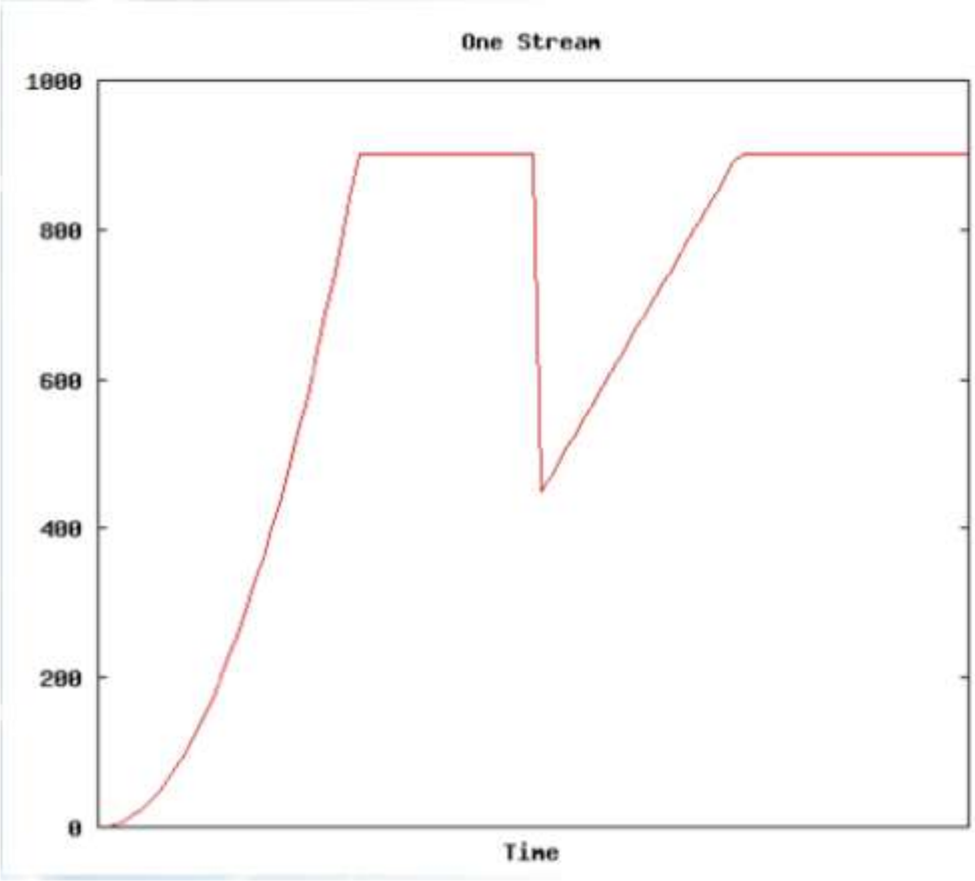
Bulk data movement

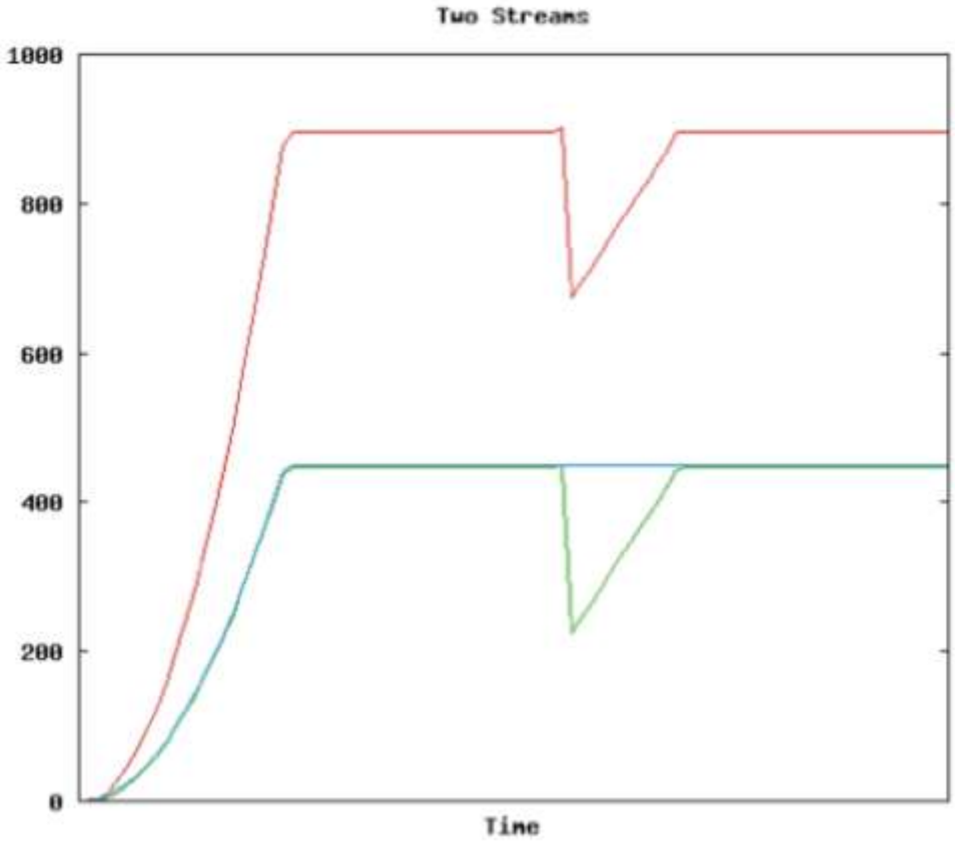
- The problem
- Involved components
  - Network architecture
  - Dedicated hosts
  - **Software tools**

Data transfer tools

- **Parallelism is key**
  - It is much easier to achieve a given performance level with four parallel connections than one connection
  - Several tools offer parallel transfers
- **Latency interaction is critical**
  - Wide area data transfers have much higher latency than LAN transfers
  - Many tools and protocols assume a LAN
  - Examples: SCP/SFTP, HPSS mover protocol

Parallel Streams Help With TCP Congestion Control Recovery Time





Sample data transfer rate

Using the right tool is very important

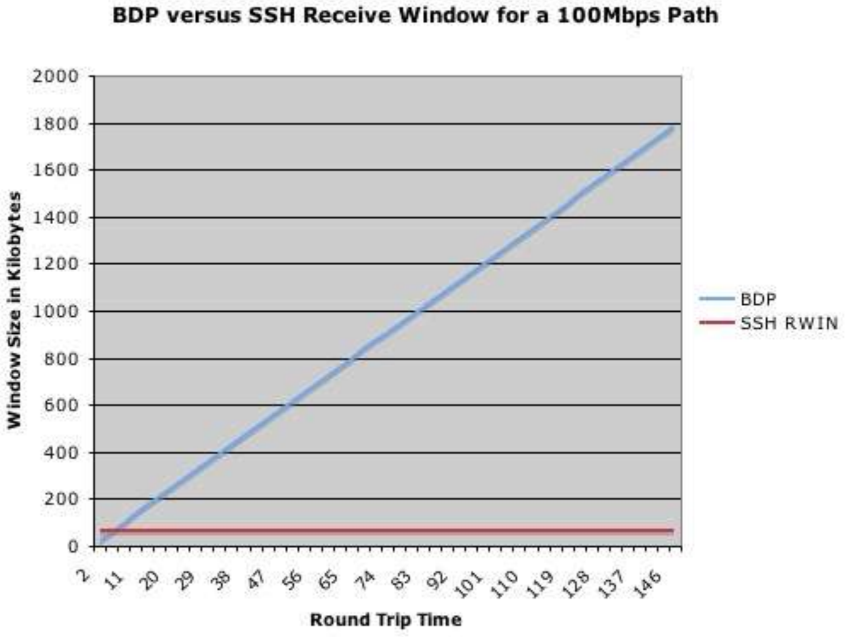
- **SCP/SFTP: 10 Mb/s**
  - standard Unix file copy tools
  - fixed 1 MB TCP window in OpenSSH
    - only 64 KB in OpenSSH versions < 4.7
- **FTP: 400-500 Mb/s**
  - assumes TCP buffer autotuning
  - Parallel stream FTP: 800-900 Mbps

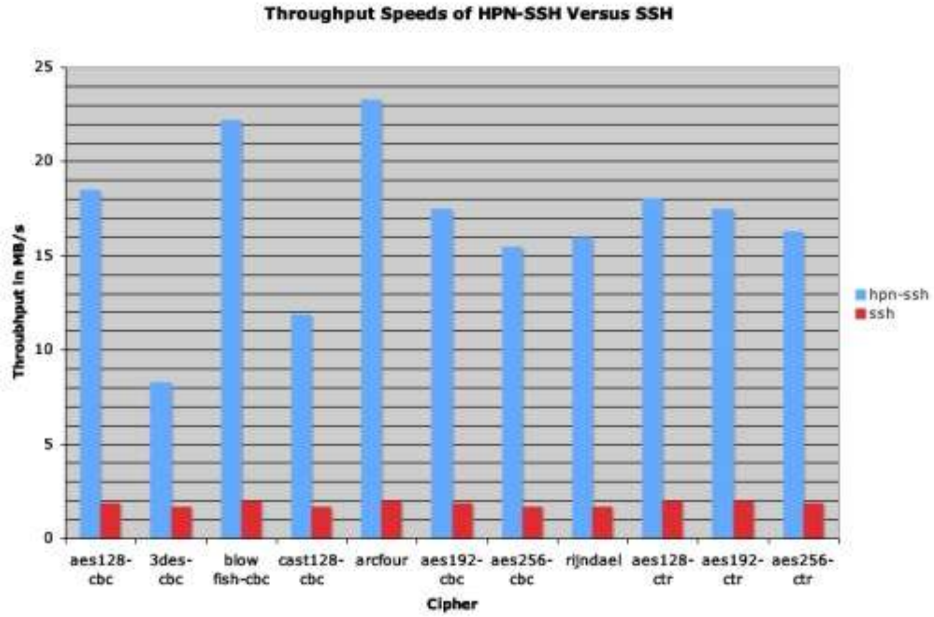
Why Not Use SCP or SFTP?

- **Pros:**
  - Most scientific systems are accessed via OpenSSH
  - SCP/SFTP are therefore installed by default
  - Modern CPUs encrypt and decrypt well enough for small to medium scale transfers
  - Credentials for system access and credentials for data transfer are the same
- **Cons:**
  - The protocol used by SCP/SFTP has a fundamental flaw that limits WAN performance
  - CPU speed doesn't matter – latency matters
  - Fixed-size buffers reduce performance as latency increases
  - It doesn't matter how easy it is to use SCP and SFTP – they simply do not perform
- **Verdict: Do Not Use Without Performance Patches**

Why Not Use SCP or SFTP?

- PSC has a patch set that fixes problems with SSH
  - http://www.psc.edu/networking/projects/hpnssh/
- Significant performance Increase
- Advantage – this helps rsync too





What’s about SFTP?

- Uses same code as SCP, so don't use SFTP for WAN transfers unless you have installed the HPN patch from PSC
- But even with the patch, SFTP has yet another flow control mechanism
  - By default, SFTP limits the total number of outstanding messages to 16 (32KB) messages
  - Since each datagram is a distinct message you end up with a 512KB outstanding data limit
  - You can increase both the number of outstanding messages ('-R') and the size of the message ('-B') from the command line though
- Sample command:
  - sftp -R 512 -B 262144 user@host:/path/to/file outfile

GridFTP

- **GridFTP from ANL has everything needed to fill the network pipe**
  - Buffer Tuning
  - Parallel Streams
- **Supports multiple authentication options**
  - Anonymous
  - X.509 (Personal certificates)
- **Ability to define a range of data ports**
  - helpful to get through firewalls
- **Sample Use:**
  - globus-url-copy -p 4 sshftp://data.lbl.gov/home/mydata/myfile file://home/mydir/myfile!
- Available from: http://www.globus.org/toolkit/downloads/

GridFTP new features

- ssh authentication option
  - Not all users need or want to deal with X.509 certificates
  - Solution: Use SSH for Control Channel
    - Data channel remains as is, so performance is the same
- Optimizations for small files
  - Concurrency option (-cc)
    - establishes multiple control channel connections and transfer multiple files simultaneously
    - Pipelining option:
      - Client sends next request before the current completes
  - Cached Data channel connections
    - Reuse established data channels (Mode E)
    - No additional TCP or GSI connect overhead
- Support for UDT protocol

GridFTP bottleneck detector

- new command line option for globus-url-copy, "-nlb”
  - nlb = NetLogger bottleneck
  - Uses NetLogger libraries for analysis of network and disk I/O
    - http://acs.lbl.gov/NetLogger
- Possible "Bottleneck:" results are:
  - network: somewhere in the network
  - disk read: sender's disk
  - disk write: receiver's disk
  - unknown: disk/network are about the same and/or highly variable

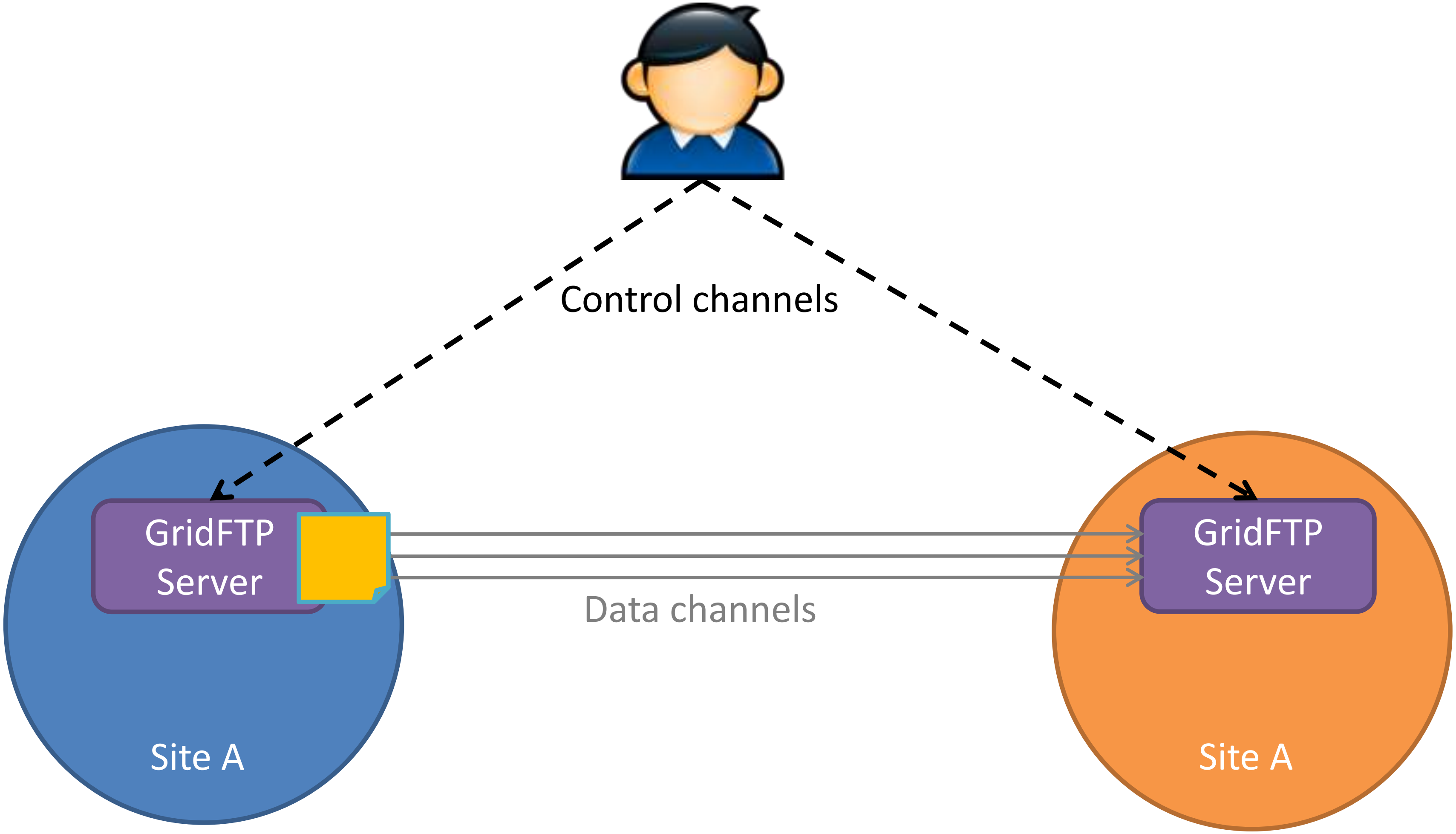
GridFTP bottleneck detector (cont.)

- **Sample Output:**
  - Total instantaneous throughput:
    - disk read = 1235.7 Mbits/s
    - disk write = 2773.0 Mbits/s
    - net read = 836.3 Mbits/s
    - net write = 1011.7 Mbits/s
  - **Bottleneck: network**
- Ignore the "net write" value (strongly influenced by system and TCP buffer artifacts)
- ***instantaneous throughput is*** **the average # of bytes divided by the time spent blocking on the system call**
- ***instantaneous throughputs*** **are higher than the overall throughput of the transfer:**
  - does not include the time waiting for data to be available
  - primarily useful for comparison and not as absolute numbers

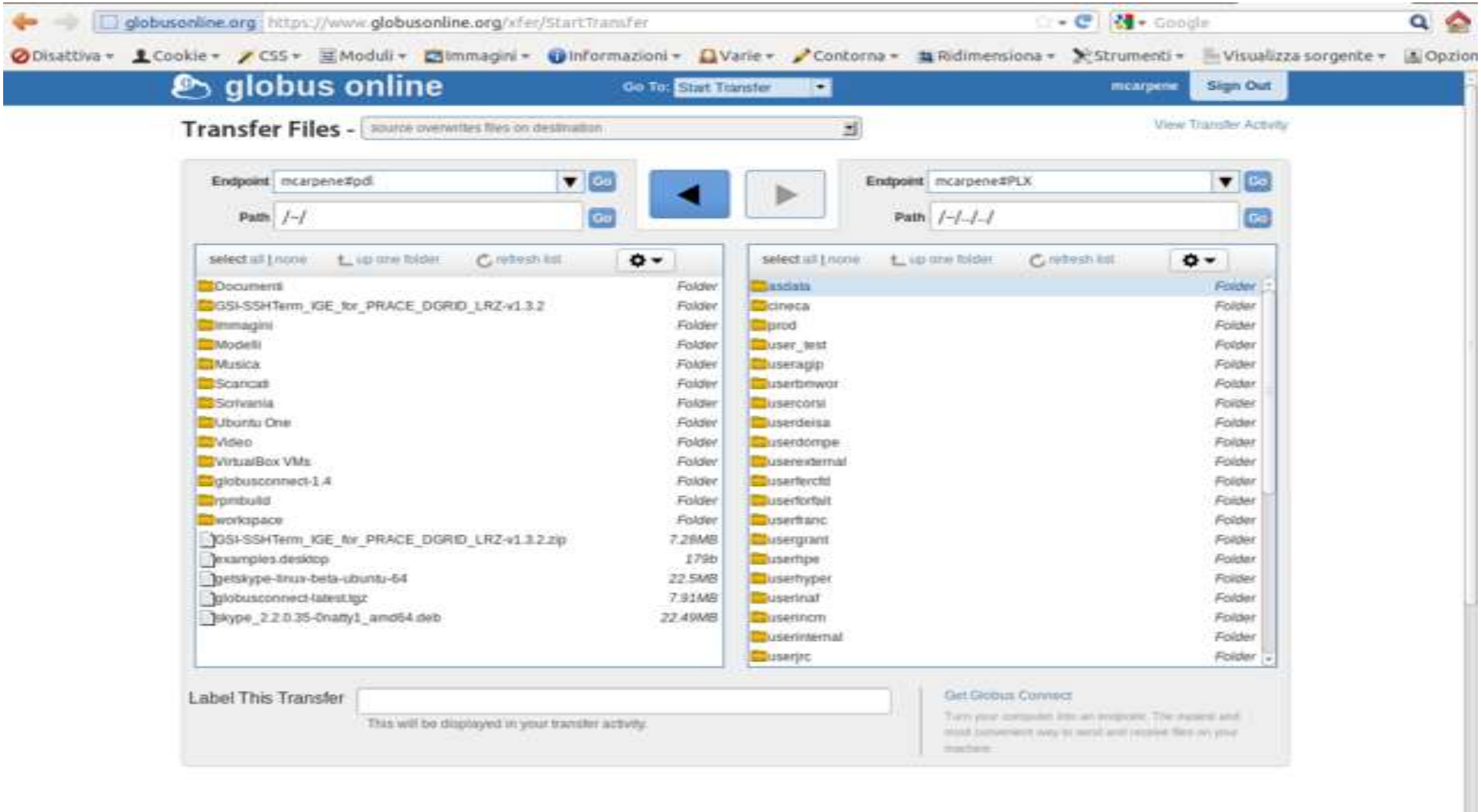
Sample Data Transfer Results

- Using the right tool is very important
- Sample Results:
  - RTT = 53 ms, network capacity = 10Gb/s.
- **Tool Throughput**
  - **scp:** 140 Mb/s
  - **HPN patched scp:** 1.2 Gb/s
  - **FTP:** 1.4 Gb/s
  - **GridFTP**, 4 streams 5.4 Gb/s
  - **GridFTP**, 8 streams 6.6 Gb/s

GridFTP: third Party Transfer



Globus OnLine Service



Transfer Files - source overwrites files on destination

Endpoint: mcarpene#pdl
Path: /~/

Documenti	Folder
GSI-SSHTerm_IGE_for_PRACE_DGRID_LRZ-v1.3.2	Folder
Immagini	Folder
Modelli	Folder
Musica	Folder
Scaricati	Folder
Scrivania	Folder
Ubuntu One	Folder
Video	Folder
VirtualBox VMs	Folder
globusconnect-1.4	Folder
rpmbuild	Folder
workspace	Folder
GSI-SSHTerm_IGE_for_PRACE_DGRID_LRZ-v1.3.2.zip	7.28MB
examples.desktop	179b
getskype-linux-beta-ubuntu-64	22.5MB
globusconnect-latest.tgz	7.91MB
skype_2.2.0.35-0natty1_amd64.deb	22.49MB

Endpoint: mcarpene#PLX
Path: /~/../../

asdata	Folder
cineca	Folder
prod	Folder
user_test	Folder
useragip	Folder
userbmwor	Folder
usercorsi	Folder
userdeisa	Folder
userdompe	Folder
userexternal	Folder
userfercfd	Folder
userforfait	Folder
userfranc	Folder
usergrant	Folder
userhpe	Folder
userhyper	Folder
userinaf	Folder
userincm	Folder
userinternal	Folder
userjrc	Folder

Label This Transfer

This will be displayed in your transfer activity.

Turn your computer into an endpoint. The easiest and most convenient way to send and receive files on your machine.

http://www.globusonline.org

Other tools

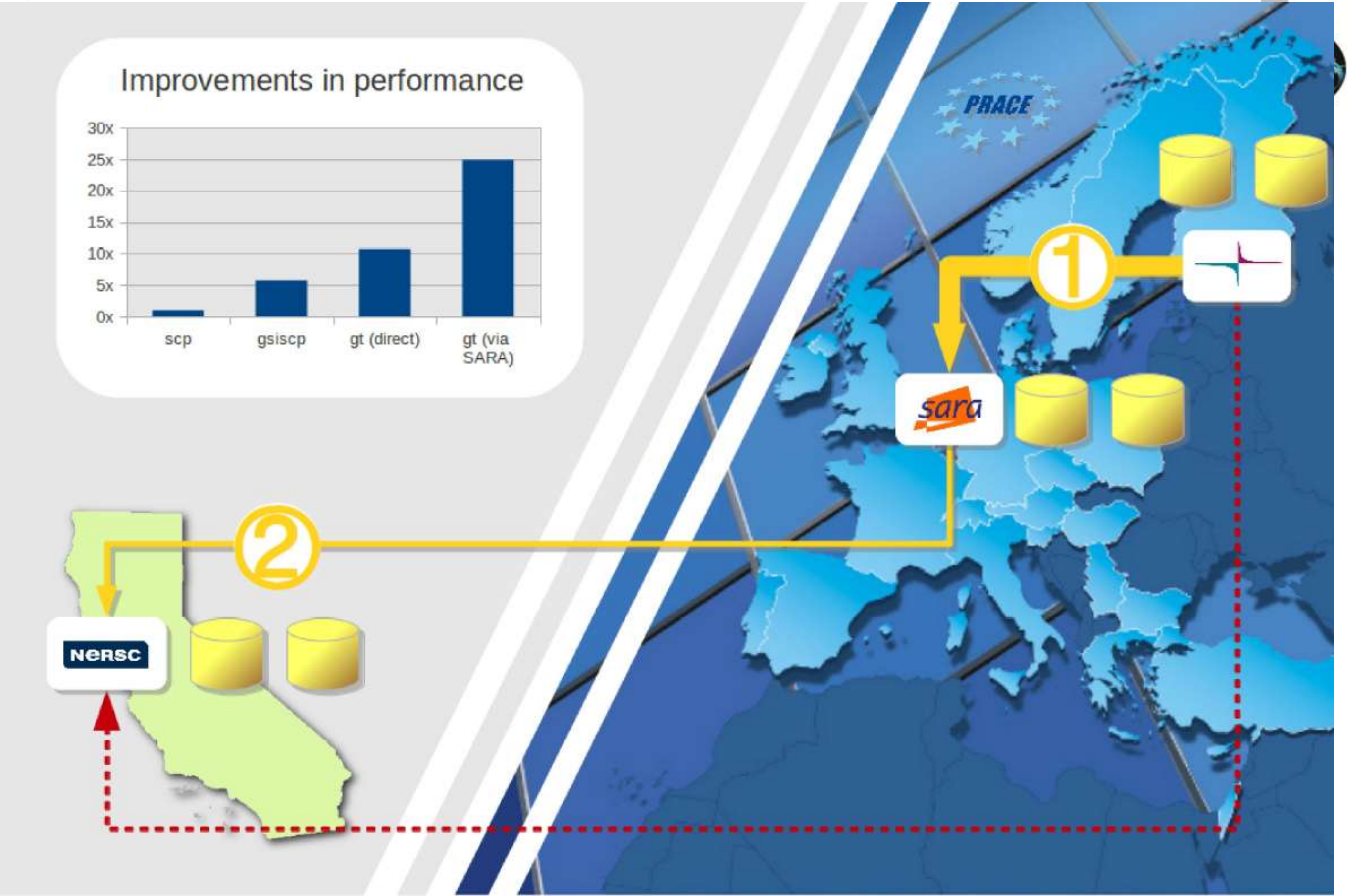
- **bbcp**: http://www.slac.stanford.edu/~abh/bbcp/
  - supports parallel transfers and socket tuning

  -
    ```
    bbcp -P 4 -v -w 2M myfile remotehost:filename
    ```
- **lftp**: http://lftp.yar.ru/
  - parallel file transfer, socket tuning, HTTP transfers, and more.

  -
    ```
    lftp -e 'set net:socket-buffer 4000000; pget -n 4
    [http|ftp]://site/path/file; quit'
    ```
- **axel**: http://axel.alioth.debian.org/
  - simple parallel accelerator for HTTP and FTP.

  -
    ```
    axel -n 4 [http|ftp]://site/file
    ```
- **rsync:** http://rsync.samba.org/

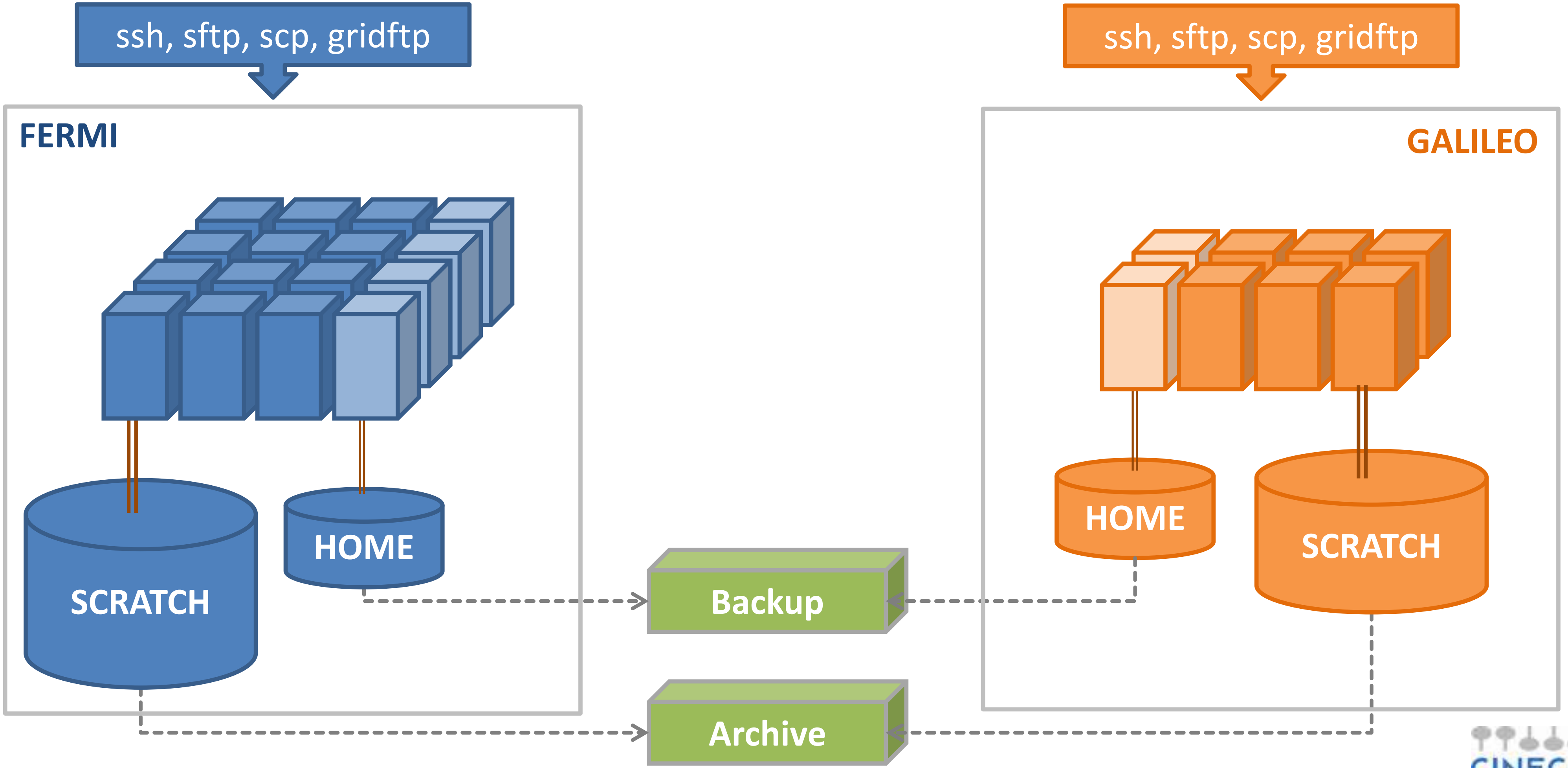
  -
    ```
    rsync --timeout=600 -avHS -r --numeric-ids --
    bwlimit=80000 --block-size=1048576  --progress
    $CINECA_SCRATCH/path/file $CINECA_DATA/path/
    ```



Improvements in performance

CINECA data resources

All CINECA machines (GALILEO and FERMI) have similar file system organization



Network resources

- The clusters are reachable from the public network through GARR (Italian NREN) facility (1Gb/s)
- The PRACE infrastructure has a dedicated private network which provides 10Gb/s guaranteed bandwidth (available on FERMI)

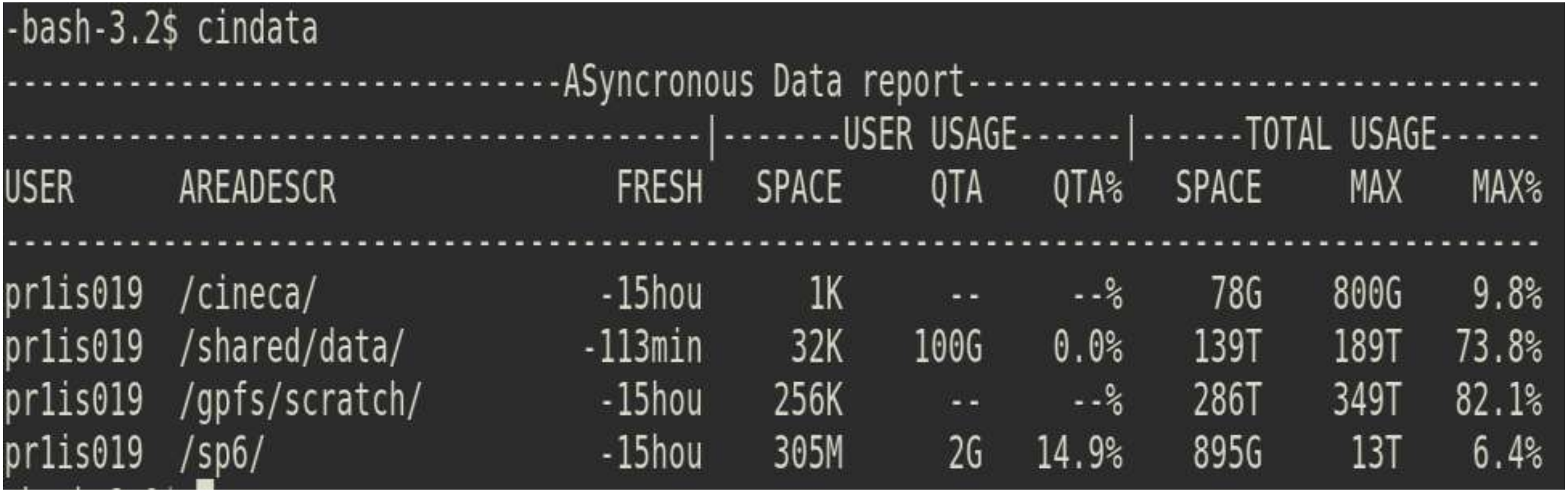
Already available at CINECA

- A public installation for GALILEO (**without -stripe** option) is available for CINECA users. It is reachable at:
  - **gsiftp://gftp.galileo.cineca.it:2811**

- A public installation for FERMI (**with -stripe** option) is available for CINECA users. It is reachable at:
  - **gsiftp://gftp-fermi.cineca.it:2811 (public network)**
  - **gsiftp://gftp-prace.cineca.it:2811 (PRACE network)**

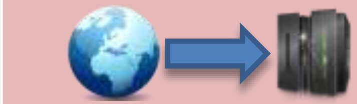
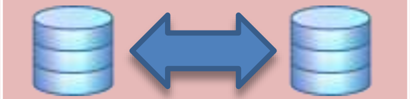
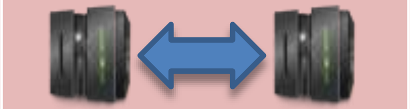
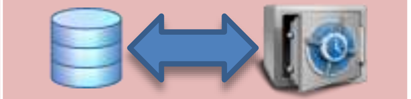
CINECA “cindata” command

- What’s about storage’s status?



```
-bash-3.2$ cindata
-------------------------------ASyncronous Data report-------------------------------
---------------------------------------|-------USER USAGE------|------TOTAL USAGE------
USER       AREADESCR                  FRESH   SPACE     QTA     QTA%   SPACE     MAX    MAX%
---------------------------------------------------------------------------------------
pr1is019   /cineca/                   -15hou     1K      --      --%     78G    800G    9.8%
pr1is019   /shared/data/             -113min    32K    100G     0.0%    139T    189T   73.8%
pr1is019   /gpfs/scratch/             -15hou   256K      --      --%    286T    349T   82.1%
pr1is019   /sp6/                      -15hou   305M      2G    14.9%    895G     13T    6.4%
```

Tools: comparative table

				
cp/mv		✓		
scp/sftp	✓		✓	
rsync		✓✓	✓✓	
GridFTP	✓✓		✓✓	
LTFS				✓✓

Extreme solution…

PHOTO: DAVIES & STARR

Bulk Data Transfer

Summary

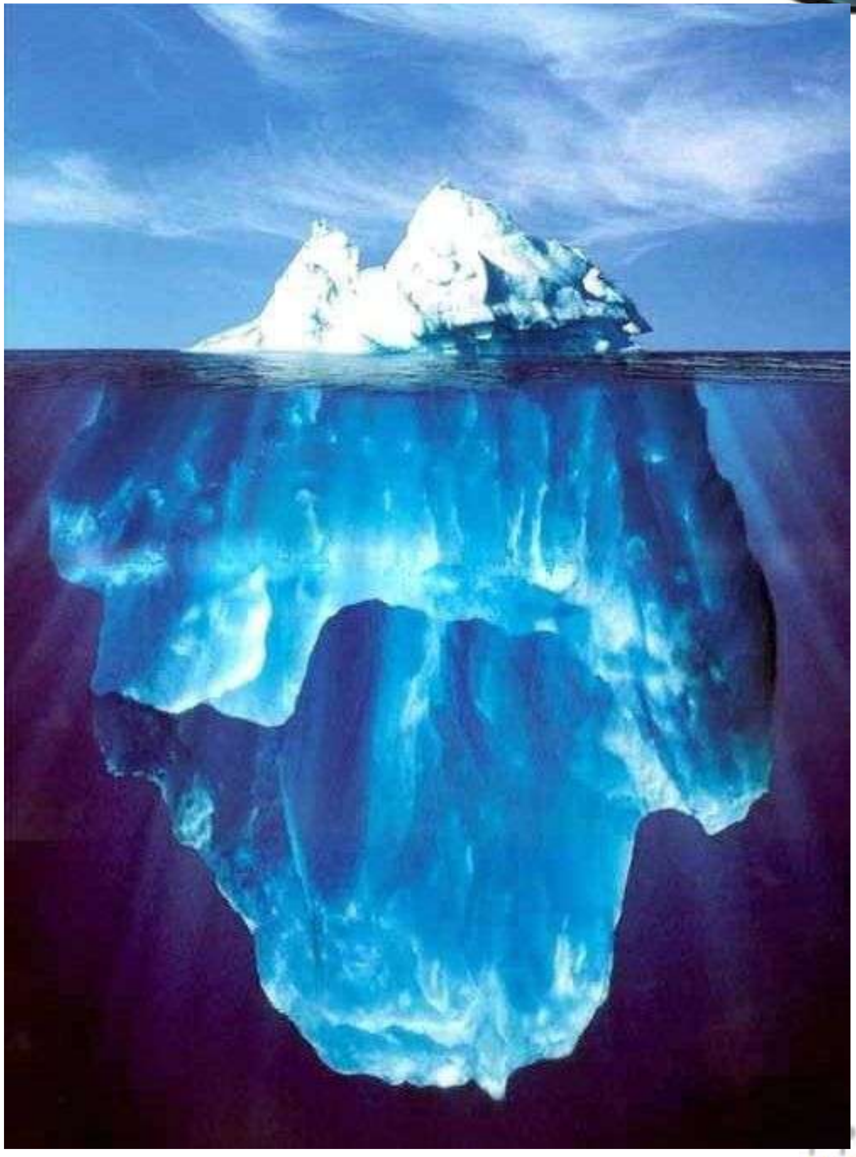
- **TCP tuning is critical, but is now easy**
  - Four lines in /etc/sysctl.conf to give autotuning
  - Make sure you’re not stuck with TCP Reno
- **Build one host for WAN data transfers, make sure it’s right**
  - Make sure TCP parameters are configured
- **Plug your hosts into the right place in the network**
- **Use the right tools**
  - Parallelism is a key
  - GridFTP, BBCP, HPN-SSH

Agenda

- Bulk data transfer
  - Tools and techniques
- **BigData techniques**
  - Hadoop/MapReduce
- Data post-processing
  - Remote visualization

Pyramid or Iceberg

- PRACE addresses the top of the pyramid
- What happens with other communities having modest HPC requirements?
- New technologies might facilitate **big** data analysis
- New scientists deal with other programming languages (Python, Java, etc.)
- New opportunities ahead us
- How to make the submerged part ramp up?

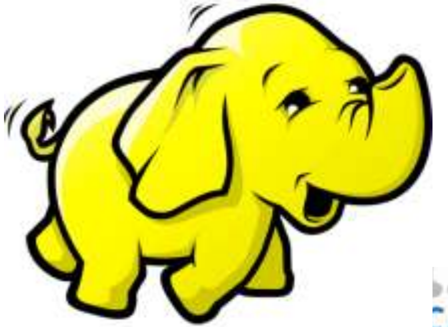


Big Data

- Extremely large datasets that are hard to deal with using Relational Databases (structured data)
  - Storage/Cost
  - Search/Performance
  - Analytics and Visualization
- Need for parallel processing on hundreds of machines

Hadoop design principles

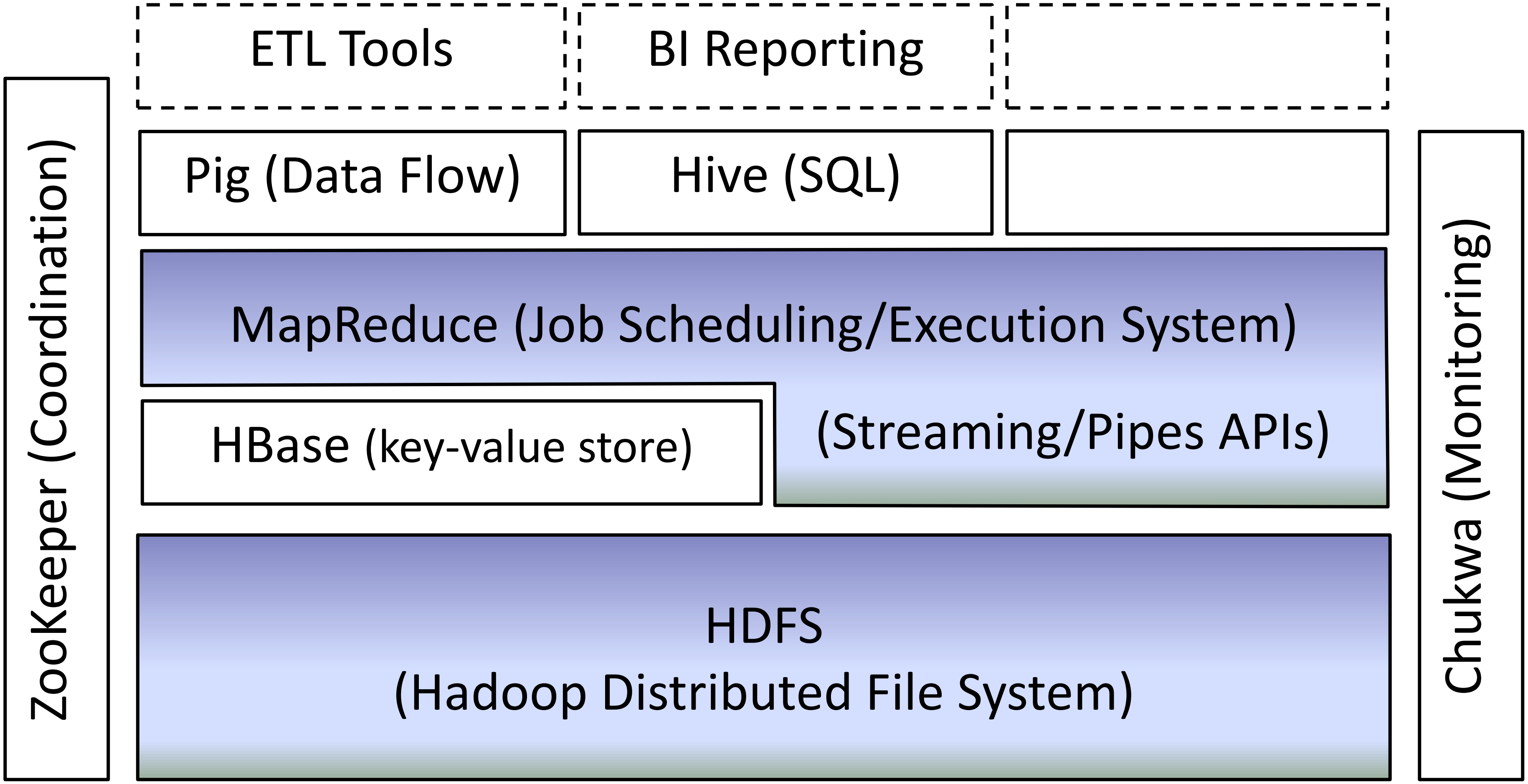
- System shall manage and heal itself
  - Automatically and transparently recover from failures
  - Speculatively execute redundant tasks if certain nodes are detected to be slow
- Performance shall scale linearly
  - Proportional change in capacity with resource change
- Compute should move to data
  - Lower latency, lower bandwidth
- Simple core, modular and extensible



What is Hadoop

- A scalable fault-tolerant system for data storage and processing
  - Commodity hardware
  - HDFS: Fault-tolerant high-bandwidth clustered storage
  - MapReduce: Distributed data processing
  - Works with structured and unstructured data
  - Open source, Apache license
  - Master (named-node) – Slave architecture

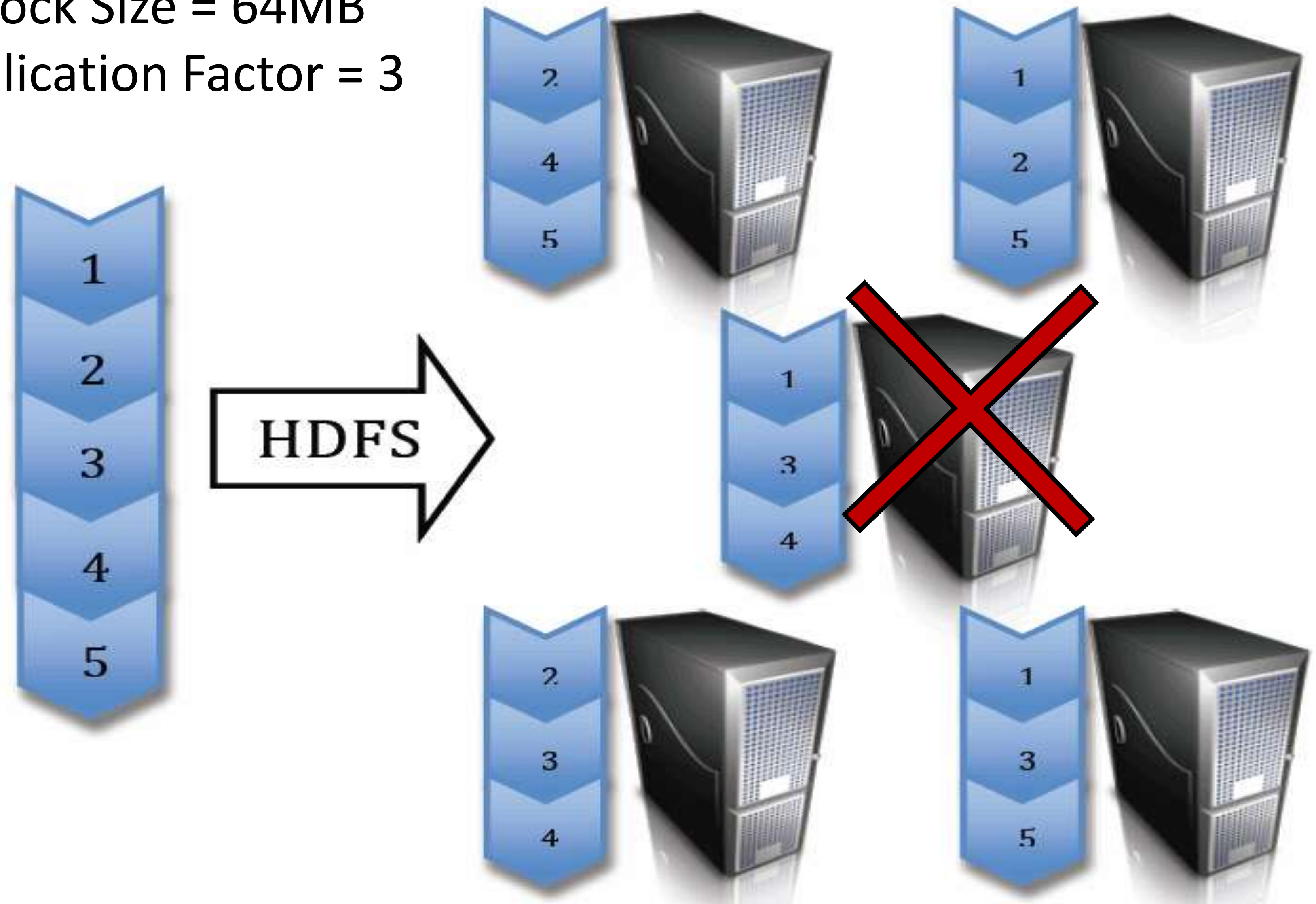
Hadoop Projects



HDFS: Hadoop Distributed FS

Block Size = 64MB

Replication Factor = 3

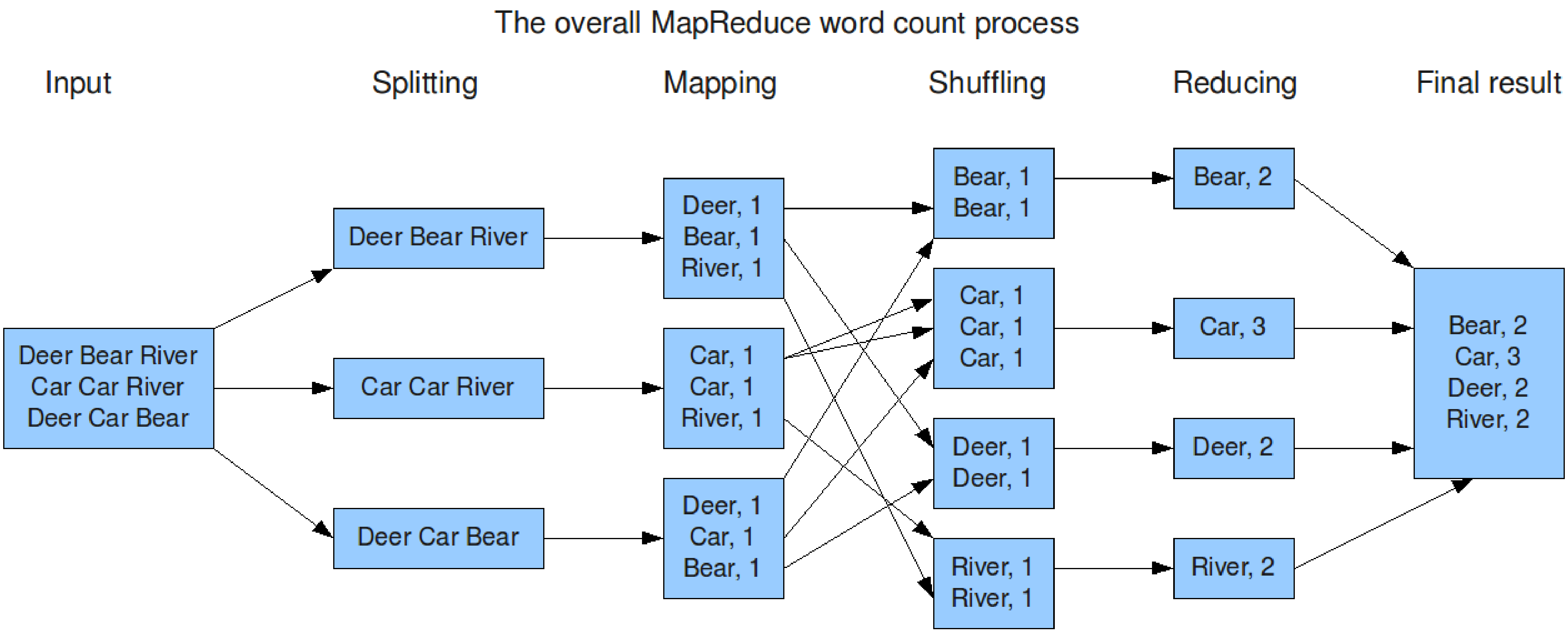


MapReduce

- Patented Google framework
- Distributed processing of large datasets

```
map (in_key, in_value) ->
  list(out_key, intermediate_value)
reduce (out_key,
  list(intermediate_value)) ->
  list(out_value)
```

MapReduce


The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reducing
Final result
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

Where and When using Hadoop

Where

- Batch data processing, not real-time / user facing
- Highly parallel data intensive distributed applications
- Very large production deployments (GRID)

When

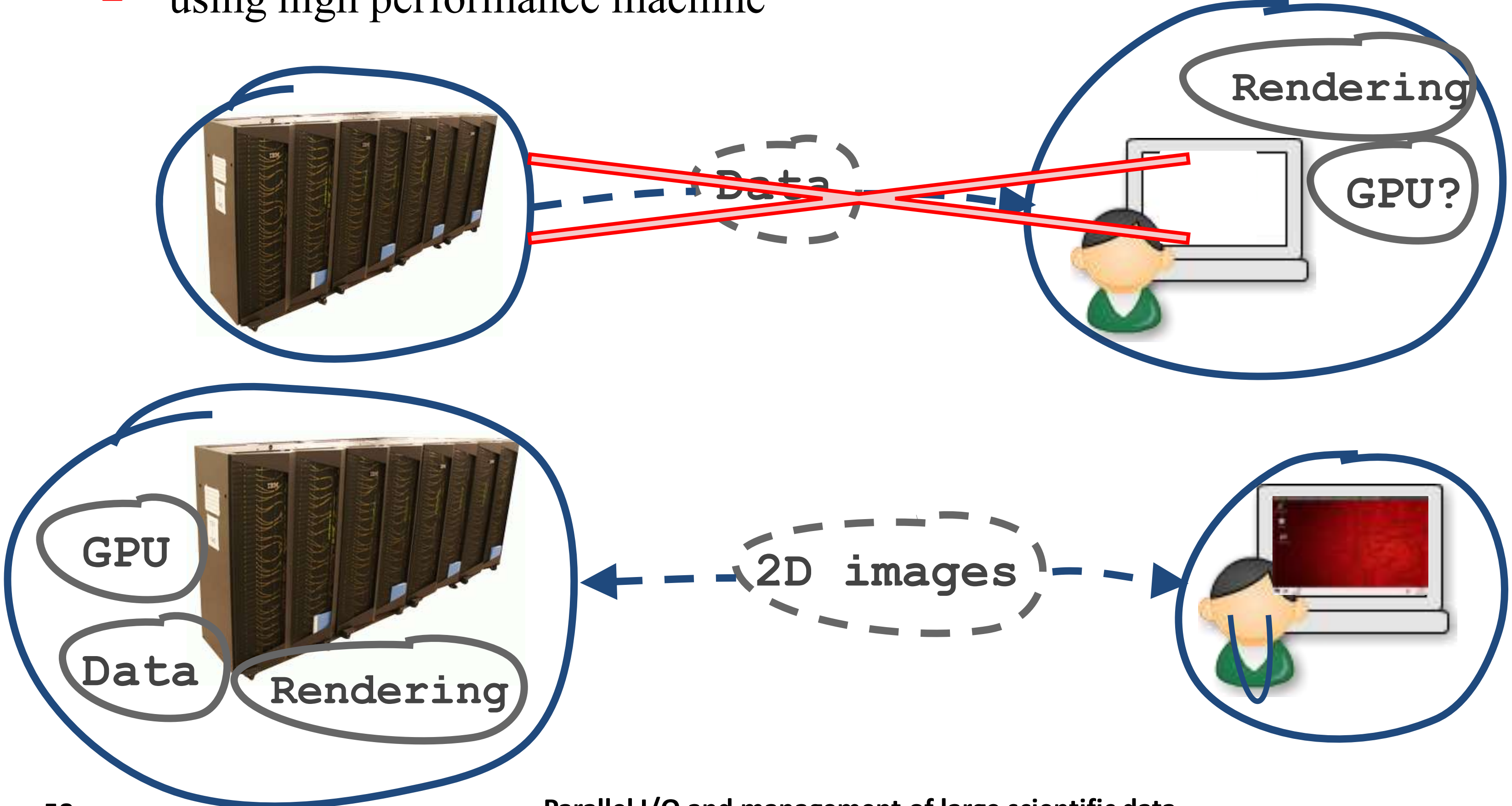
- Process lots of unstructured data
- When your processing can easily be made parallel
- Running batch jobs is acceptable
- When you have access to lots of cheap hardware

Agenda

- Bulk data transfer
  - Tools and techniques
- BigData techniques
  - Hadoop/MapReduce
- **Data post-processing**
  - Remote visualization

Remote Visualization

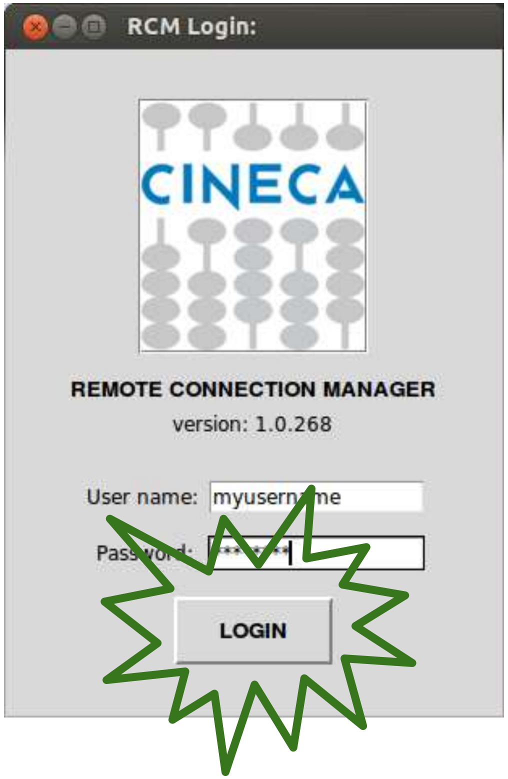
- Perform scientific visualization on large amounts of data produced on HPC systems
  - without moving data
  - using high performance machine



RCM - Login

USER

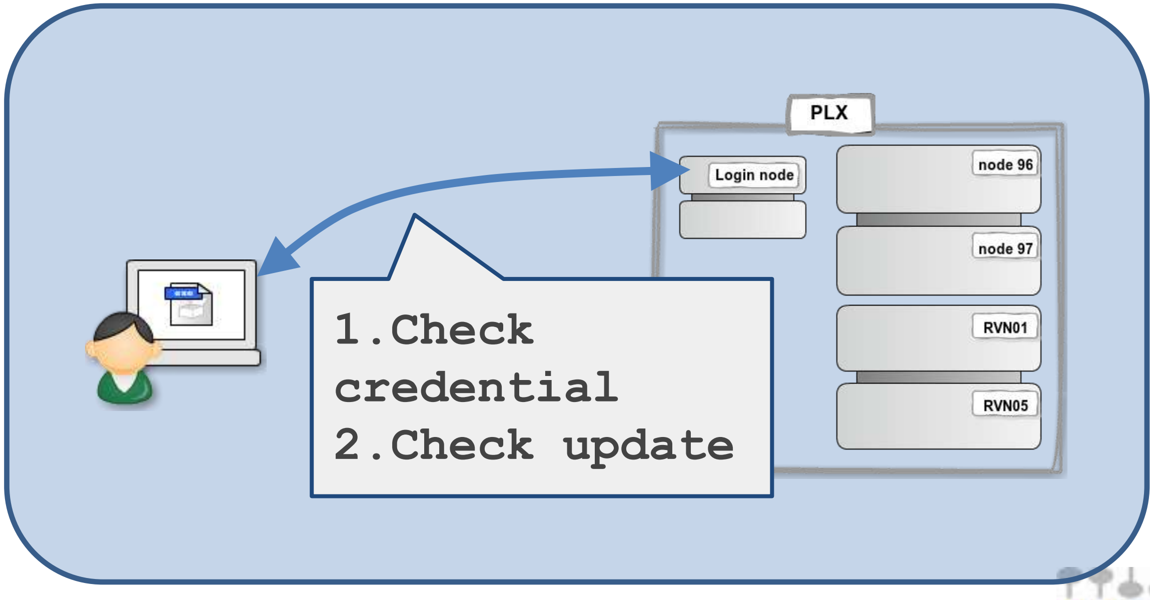
SYSTEM



RCM Login:

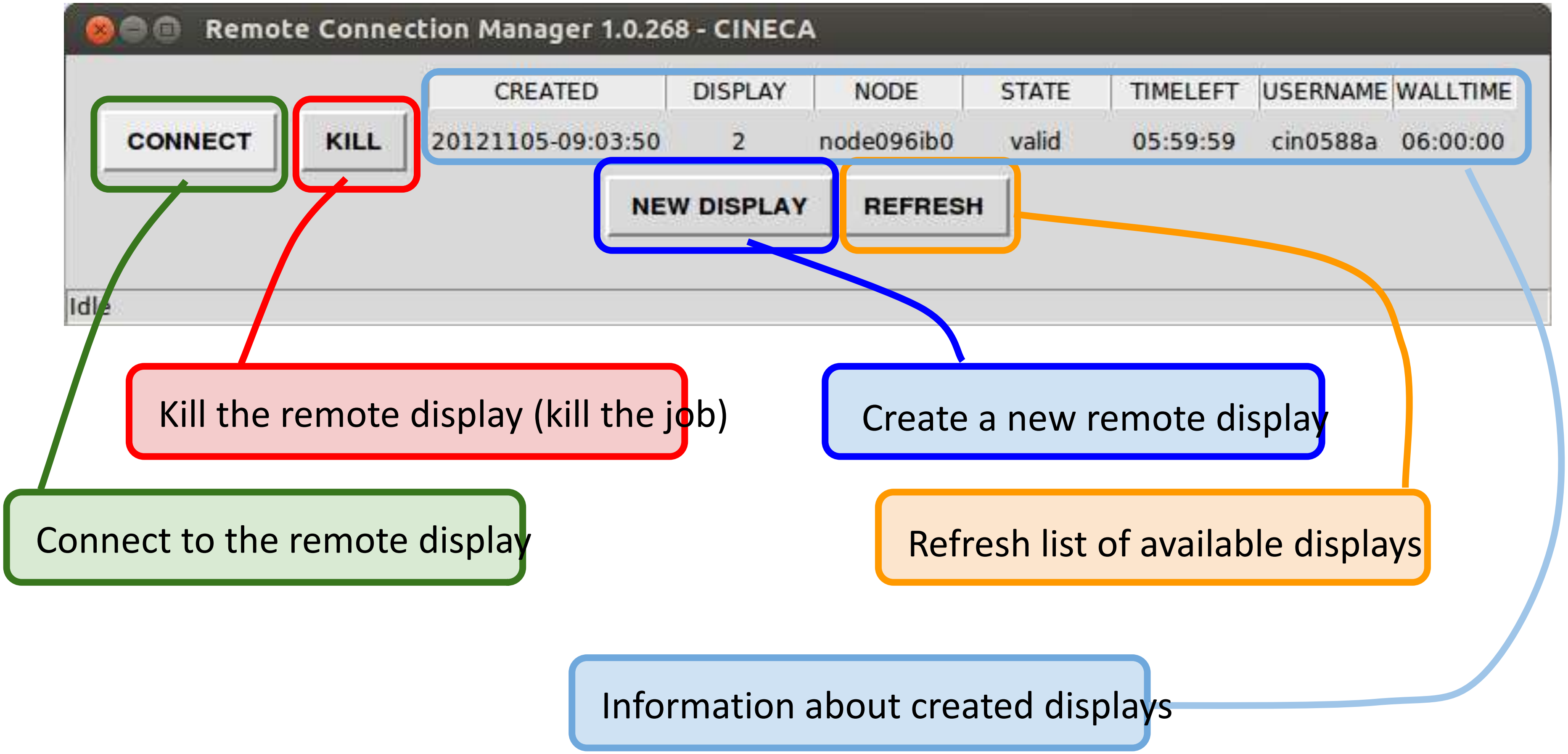
REMOTE CONNECTION MANAGER
version: 1.0.268

User name: myusername
Password: *******



1. Check credential
2. Check update

RCM - Display info



CREATED	DISPLAY	NODE	STATE	TIMELEFT	USERNAME	WALLTIME
20121105-09:03:50	2	node096ib0	valid	05:59:59	cin0588a	06:00:00

Kill the remote display (kill the job)

Create a new remote display

Connect to the remote display

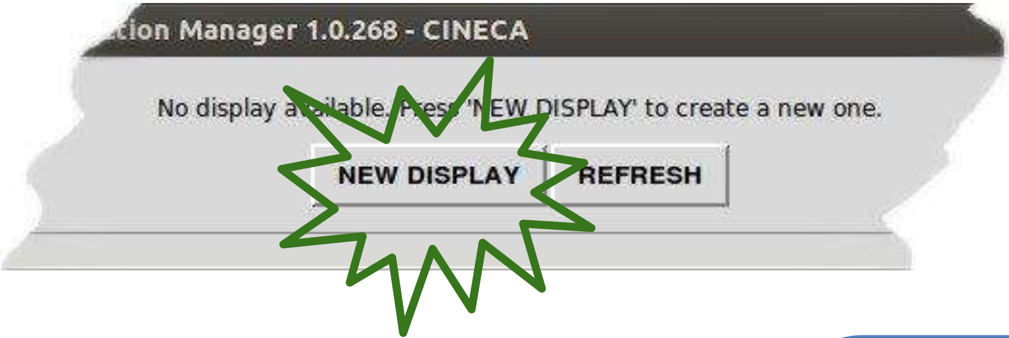
Refresh list of available displays

Information about created displays

RCM - New display (1)

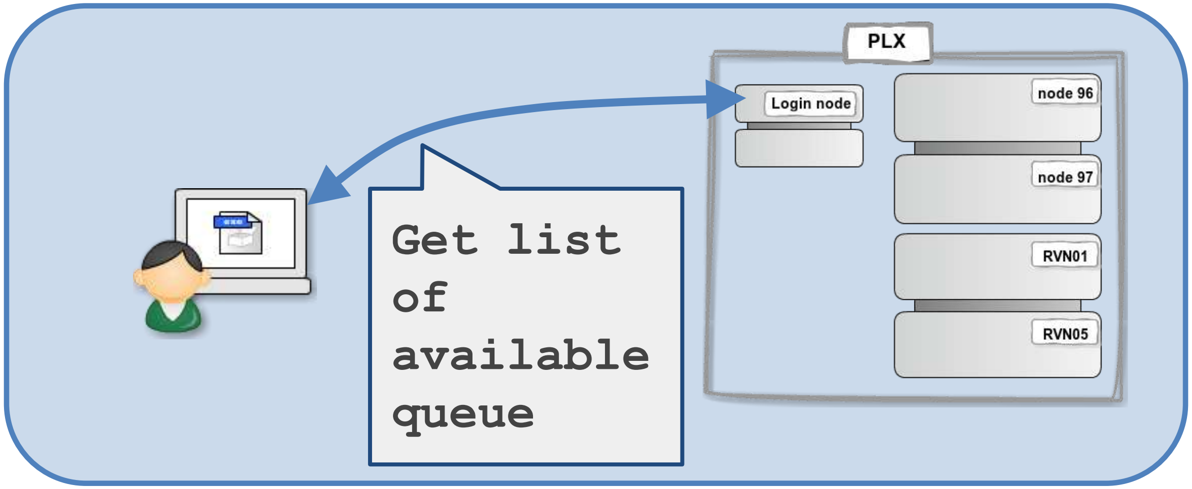
USER

SYSTEM

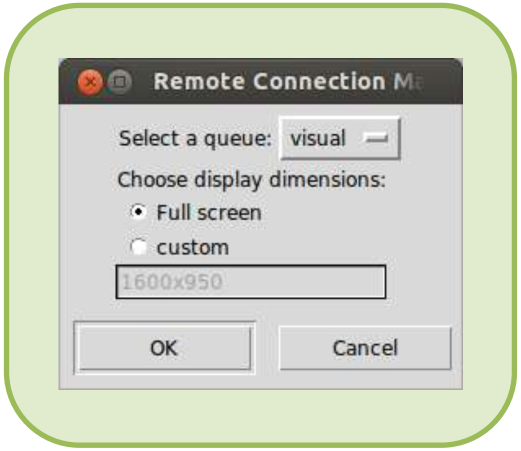


tion Manager 1.0.268 - CINECA

No display available. Press 'NEW DISPLAY' to create a new one.



Get list of available queue



Remote Connection M

Select a queue: visual

Choose display dimensions:

● Full screen

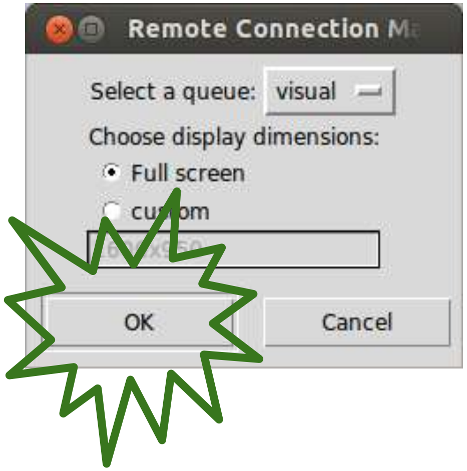
○ custom

1600x950

RCM - New display (2)

USER

SYSTEM



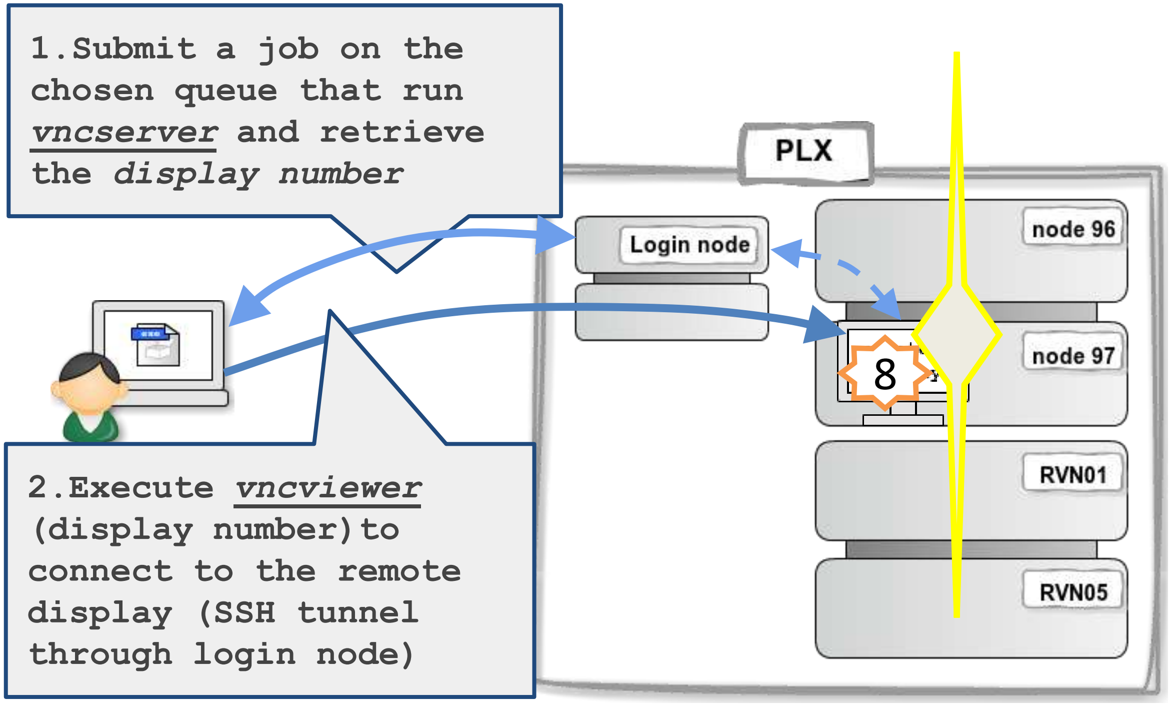
Remote Connection M

Select a queue: visual

Choose display dimensions:

● Full screen

○ custom



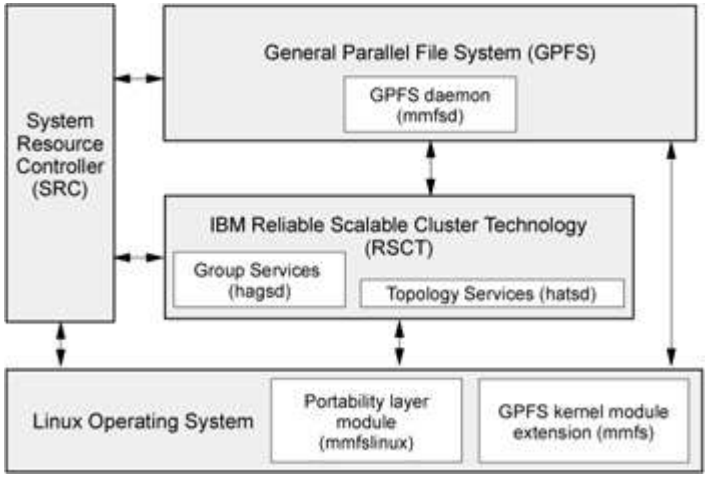
1. Submit a job on the chosen queue that run *vncserver* and retrieve the *display number*

2. Execute *vncviewer* (display number) to connect to the remote display (SSH tunnel through login node)

References

- CINECA services and documentation
  - http://www.hpc.cineca.it/services
- Get in touch
  - hpc-service-int@cineca.it

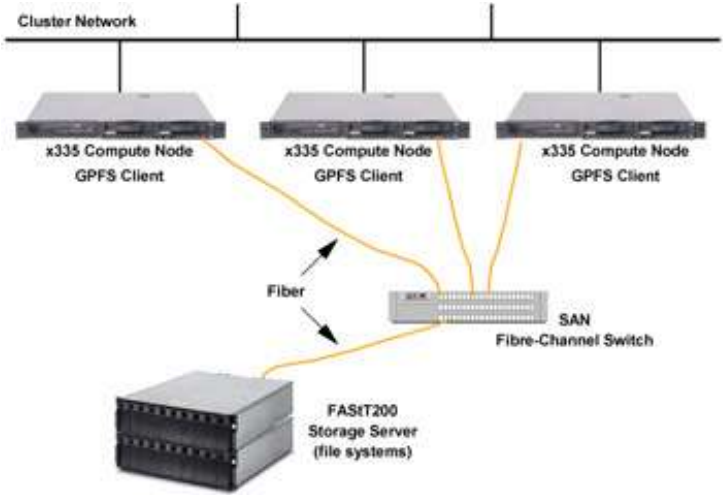
GPFS


System Resource Controller (SRC)
General Parallel File System (GPFS)
GPFS daemon (mmfsd)
IBM Reliable Scalable Cluster Technology (RSCT)
Group Services (hagsd)
Topology Services (hatsd)
Linux Operating System
Portability layer module (mmfslinux)
GPFS kernel module extension (mmfs)

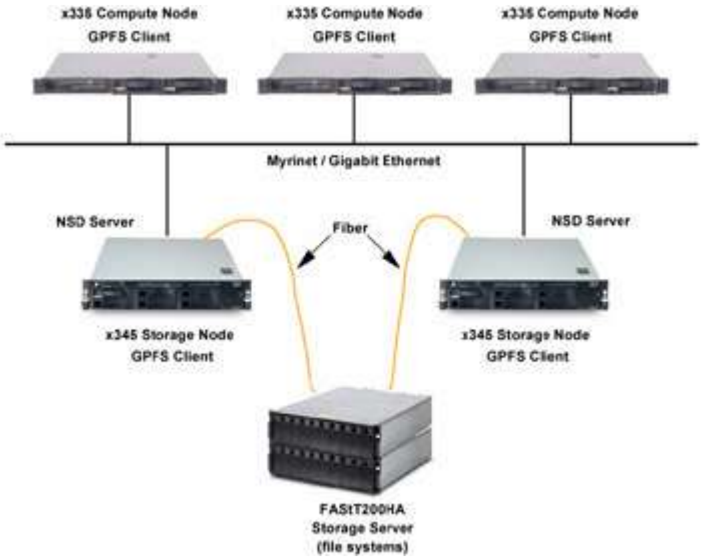
GPFS is based on the following components:

- A kernel extension
- GPFS daemon
- RSCT daemons
- Portability layer module

GPFS



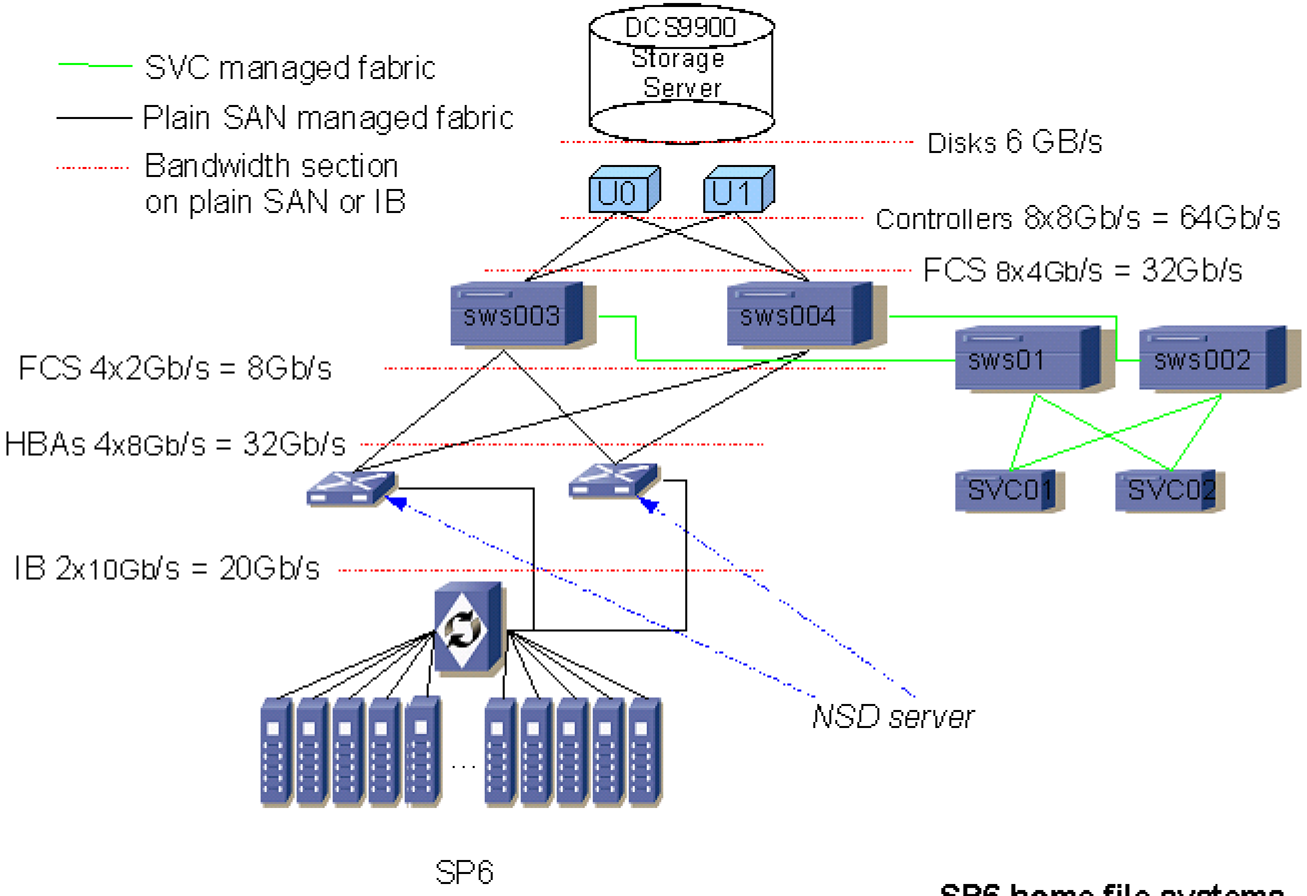
Direct attached disks



Primary and secondary servers

GPFS

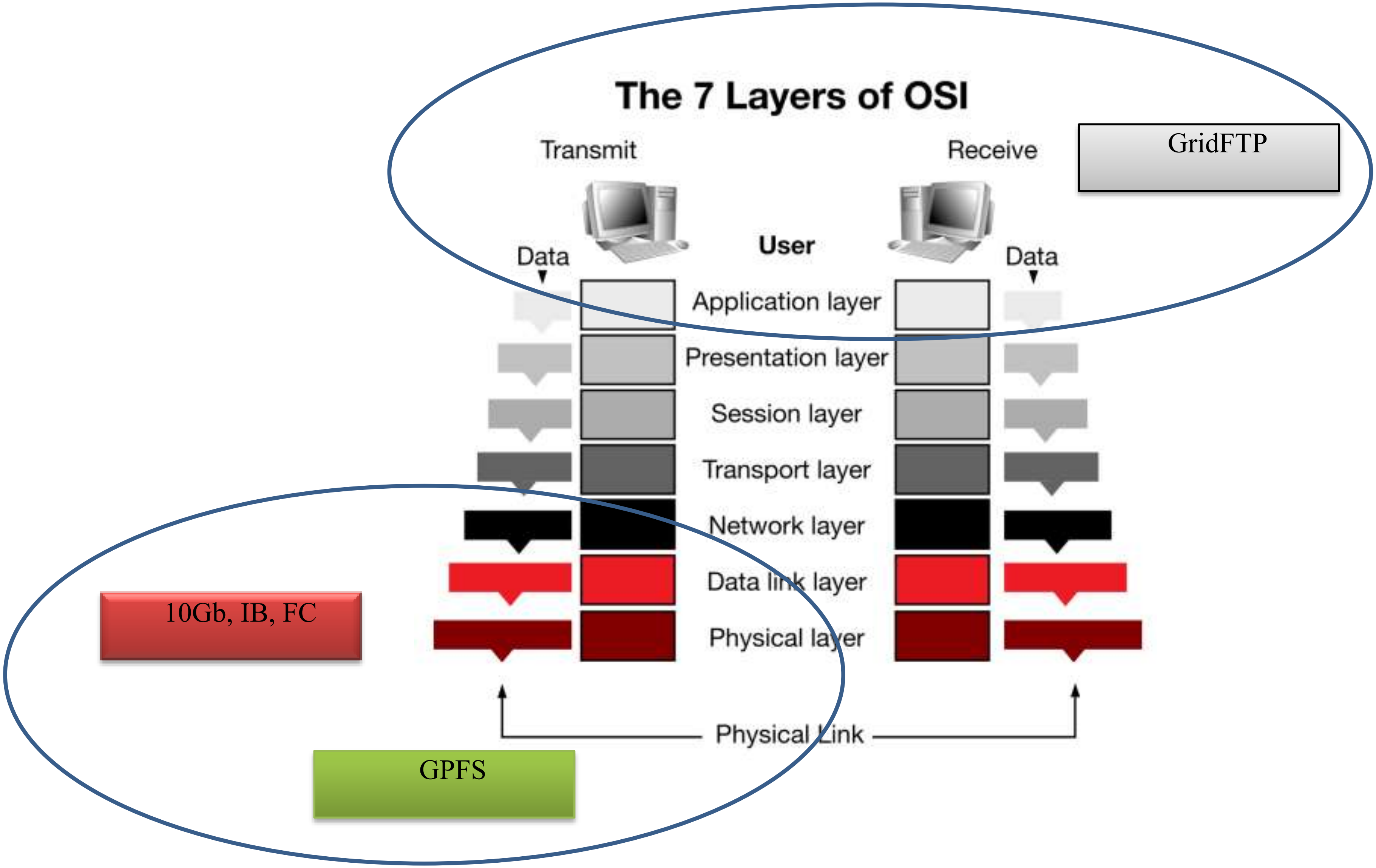
...an hold example



SP6 home file systems

Bulk data movement

Different optimization at different scale



GridFTP

10Gb, IB, FC

GPFS

Credits

- NICS Scientific Computing Group
  - http://www.nics.tennessee.edu/
- Energy Sciences Network
  - http://fasterdata.es.net
- Lawrence Berkeley National Laboratory
  - http://www.lbl.gov/
- Argonne National Laboratory
  - www.anl.gov