



Data Staging: Moving large amounts of data around, and moving it close to compute resources

Digital Preservation Advanced Practitioner Course
Glasgow, July 19th 2013

c.cacciari@cineca.it





Outline

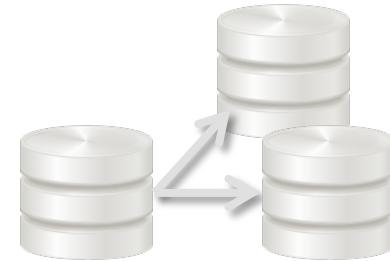
- Definition
- Starting point
- Moving data around
- Size/Performances
- The user
- Data selection
- Transfer options
- Numbers
- Registered data

Data movement: staging or replication?



Staging

- **Safe Replication** to enable communities easily create replicas of their scientific datasets in multiple data centres for improving data curation and accessibility
- **Data Staging** to facilitate communities to stage stored data onto external computational facilities, such as HPC resources



Moving large amounts of data around

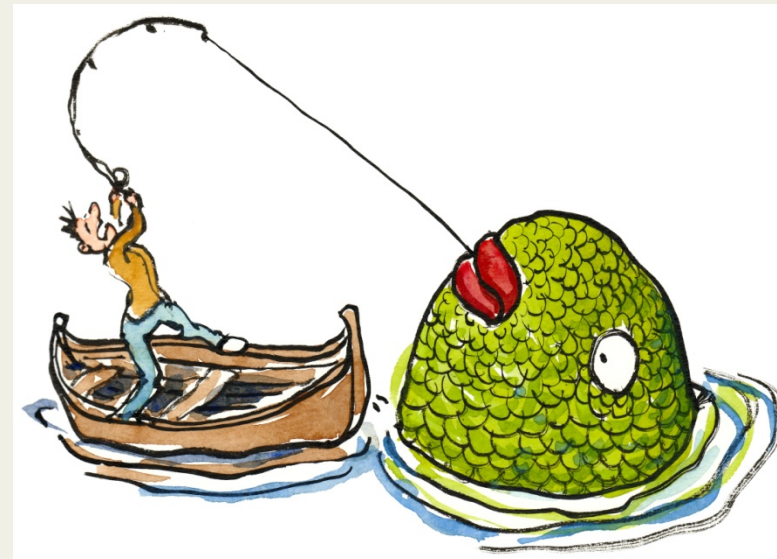
Data sets can be large both in terms of

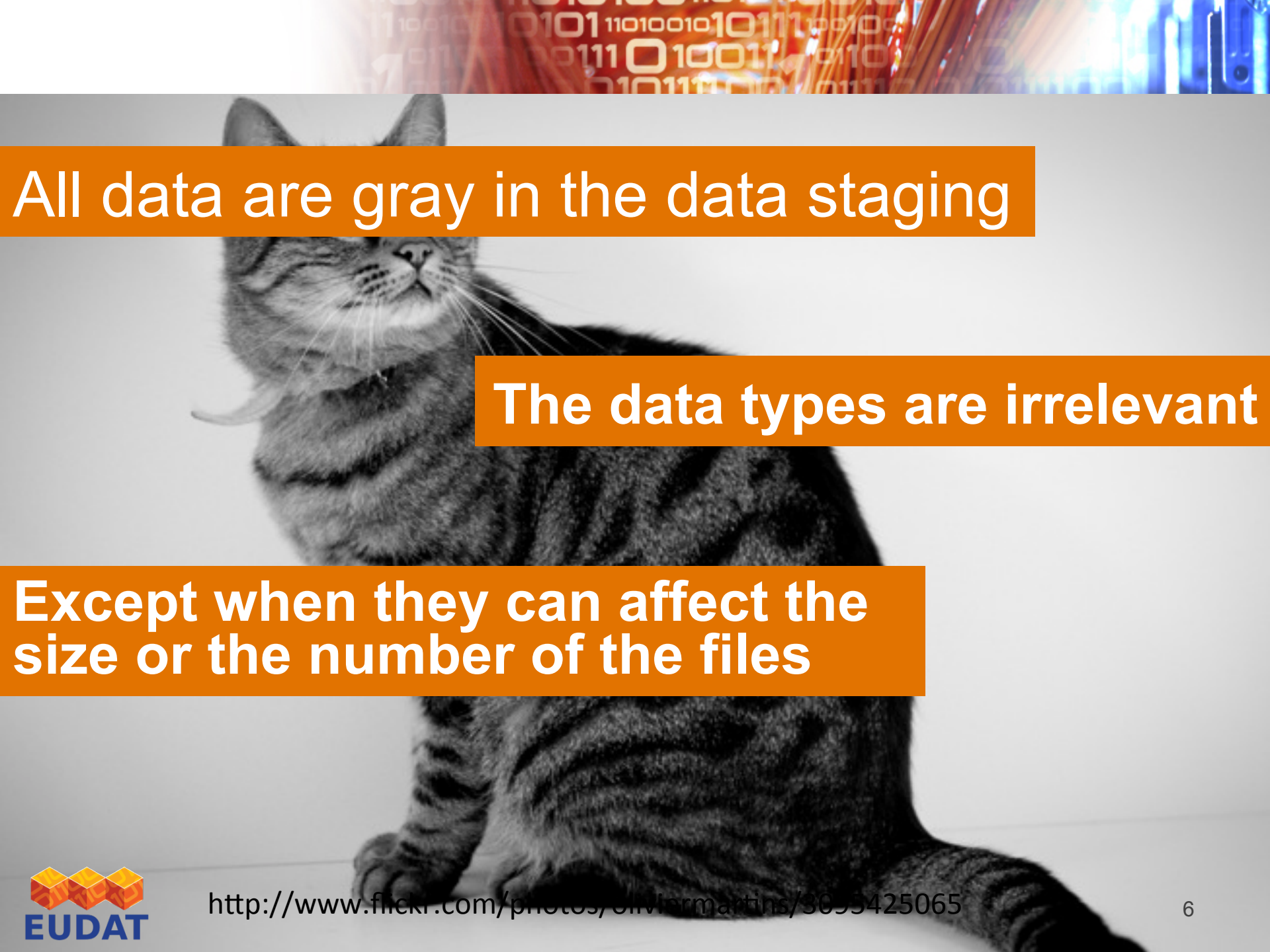
Numbers of objects



By HikingArtist.com

Single object size





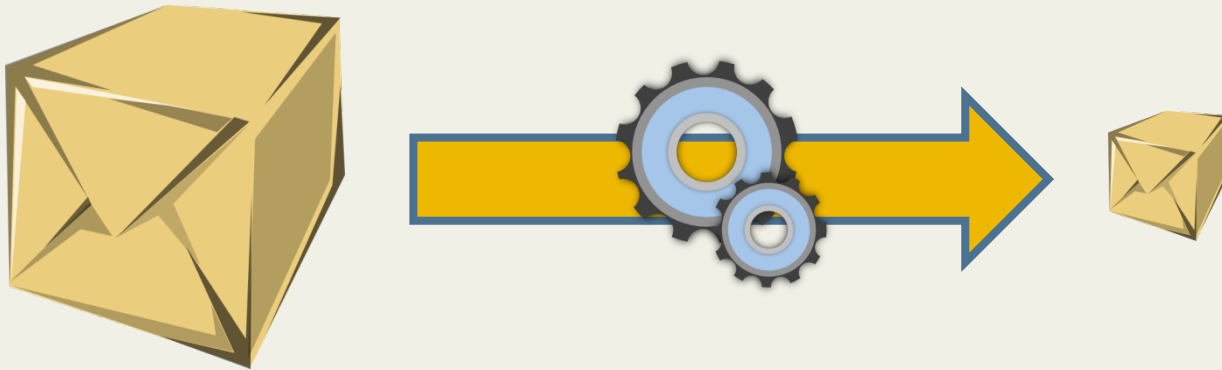
All data are gray in the data staging

The data types are irrelevant

Except when they can affect the size or the number of the files

When I should care about data types?

When you can compress them



Data transfer efficiency is crucial in data staging

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 Requirements to move Y Bytes of data in Time X

Bits per Second Requirements					
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
	1H	8H	24H	7Days	30Days

This table available at <http://fasterdata.es.net>

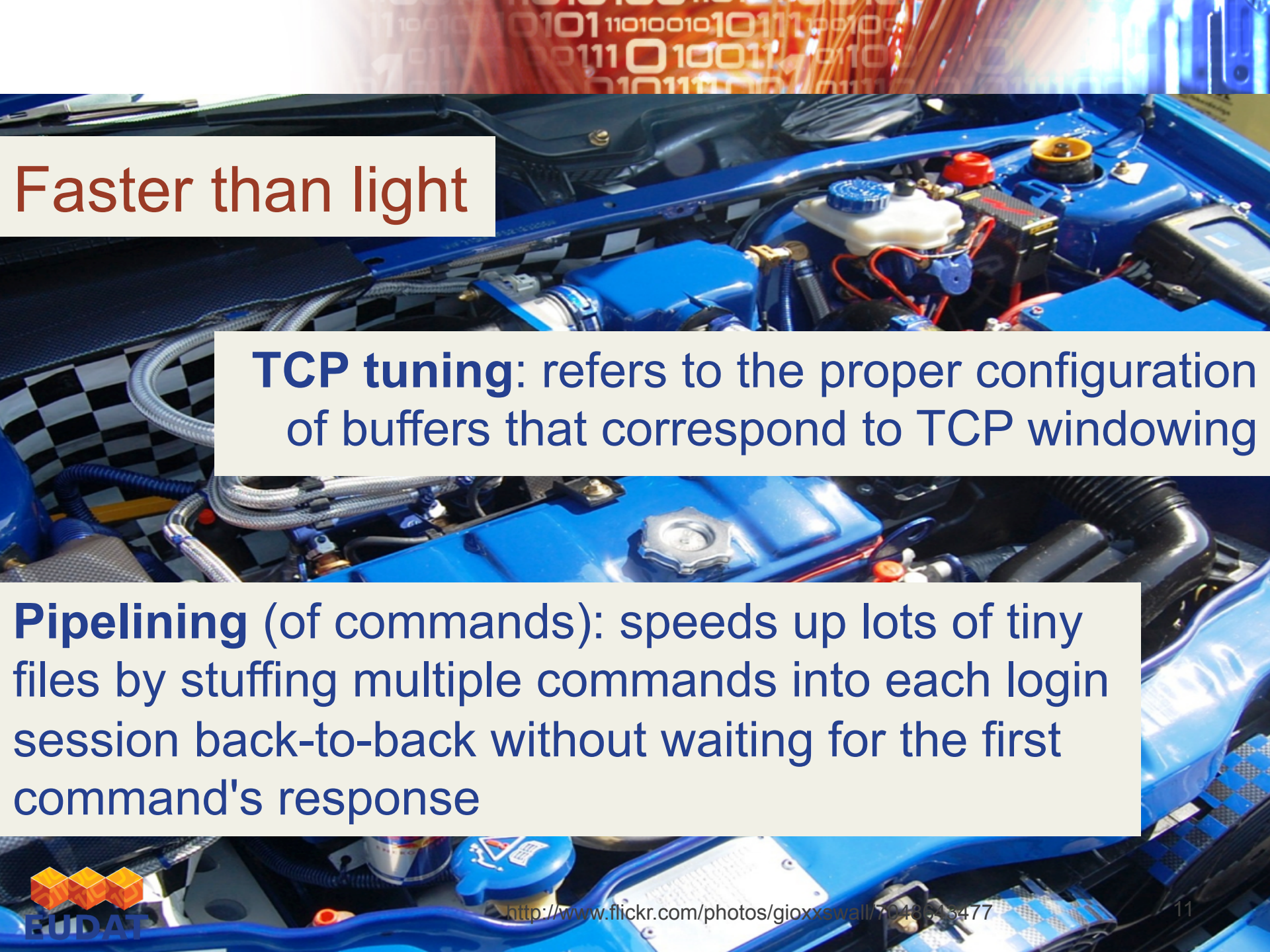


Data staging is Just in Time

You do not plan it

You can pre-stage... sometimes

**There are techniques to
improve the efficiency**



Faster than light

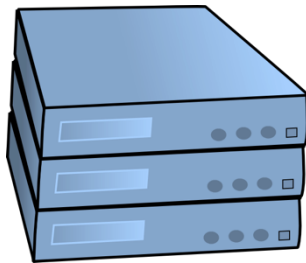
TCP tuning: refers to the proper configuration of buffers that correspond to TCP windowing

Pipelining (of commands): speeds up lots of tiny files by stuffing multiple commands into each login session back-to-back without waiting for the first command's response



EUDAT

Parallelizing: on wide-area links, using multiple TCP streams in parallel (even between the same source and destination) can improve aggregate bandwidth over using a single TCP stream



Striping: data may be striped or interleaved across multiple servers



Parallelism and TCP tuning are the keys

- It is much easier to achieve a given performance level with four parallel connections than with one
- A good TCP tuning can improve drastically performances



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

Use the right tool

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



But efficiency is not all

Easiness of use

High reliability

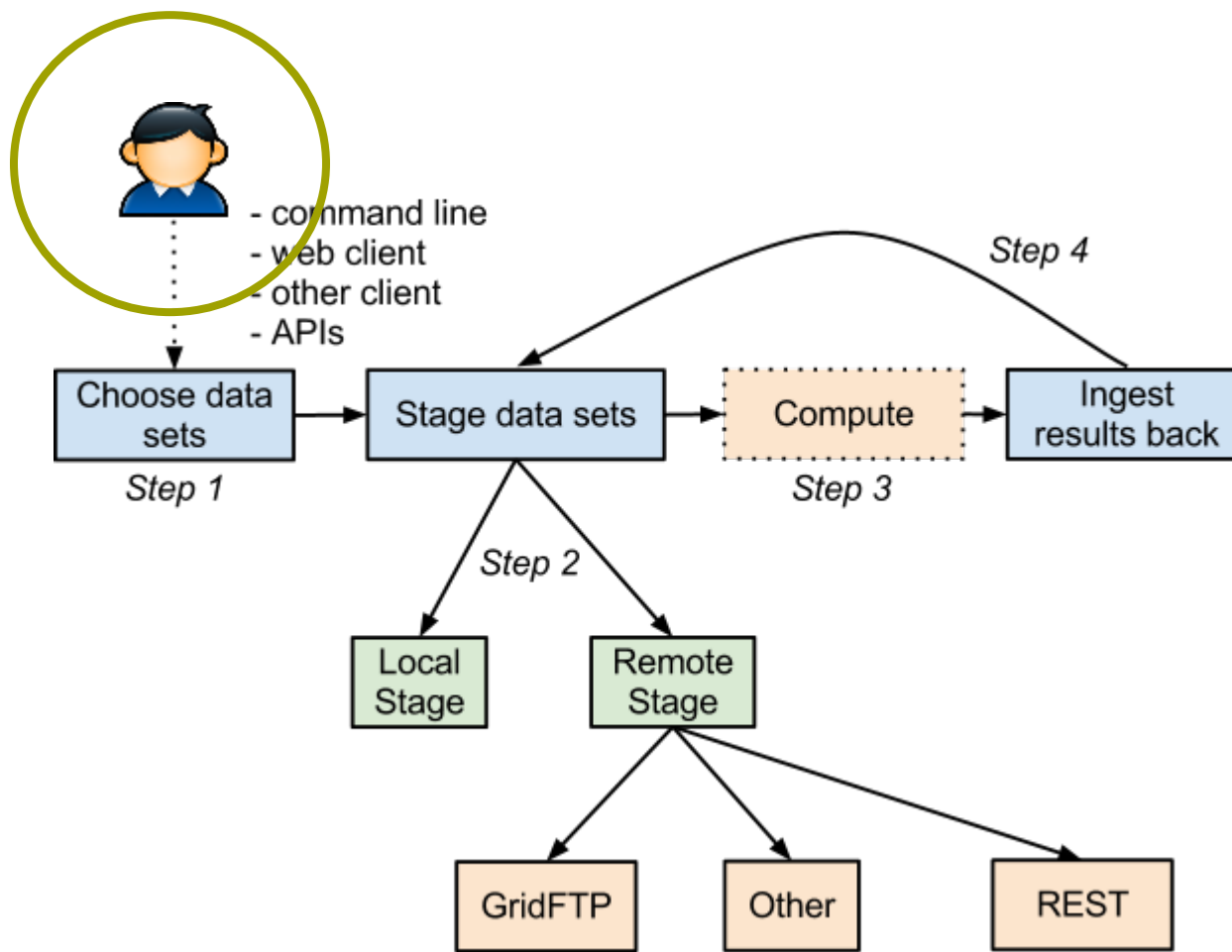
Third-party transfer

Possibility to control/limit the transfer throughput to avoid engulfing the network

Which priority?

Different scenarios, different needs

... and moving it close to compute resources

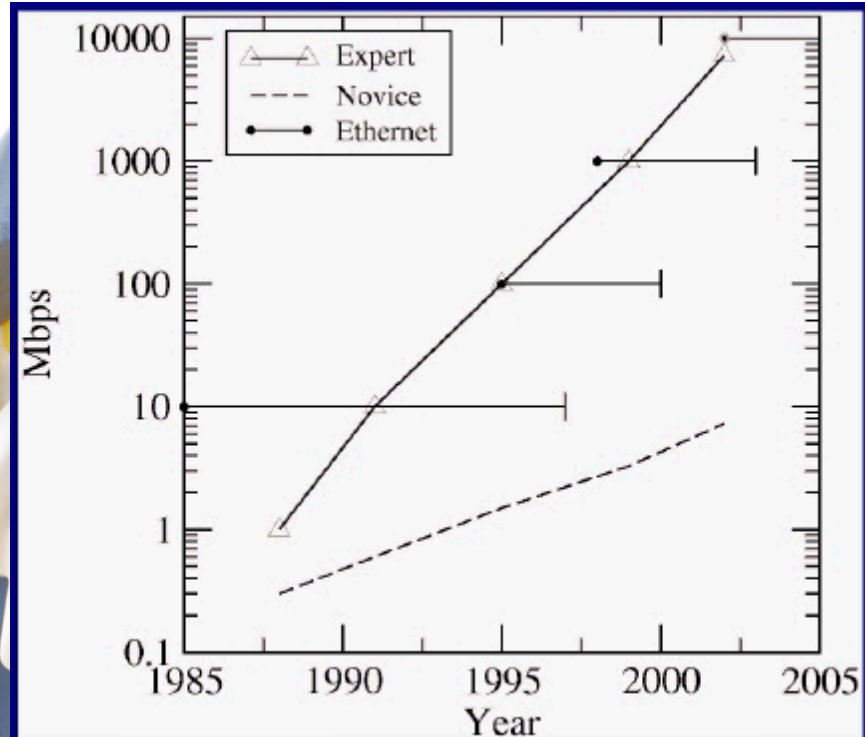




Who will move the data?

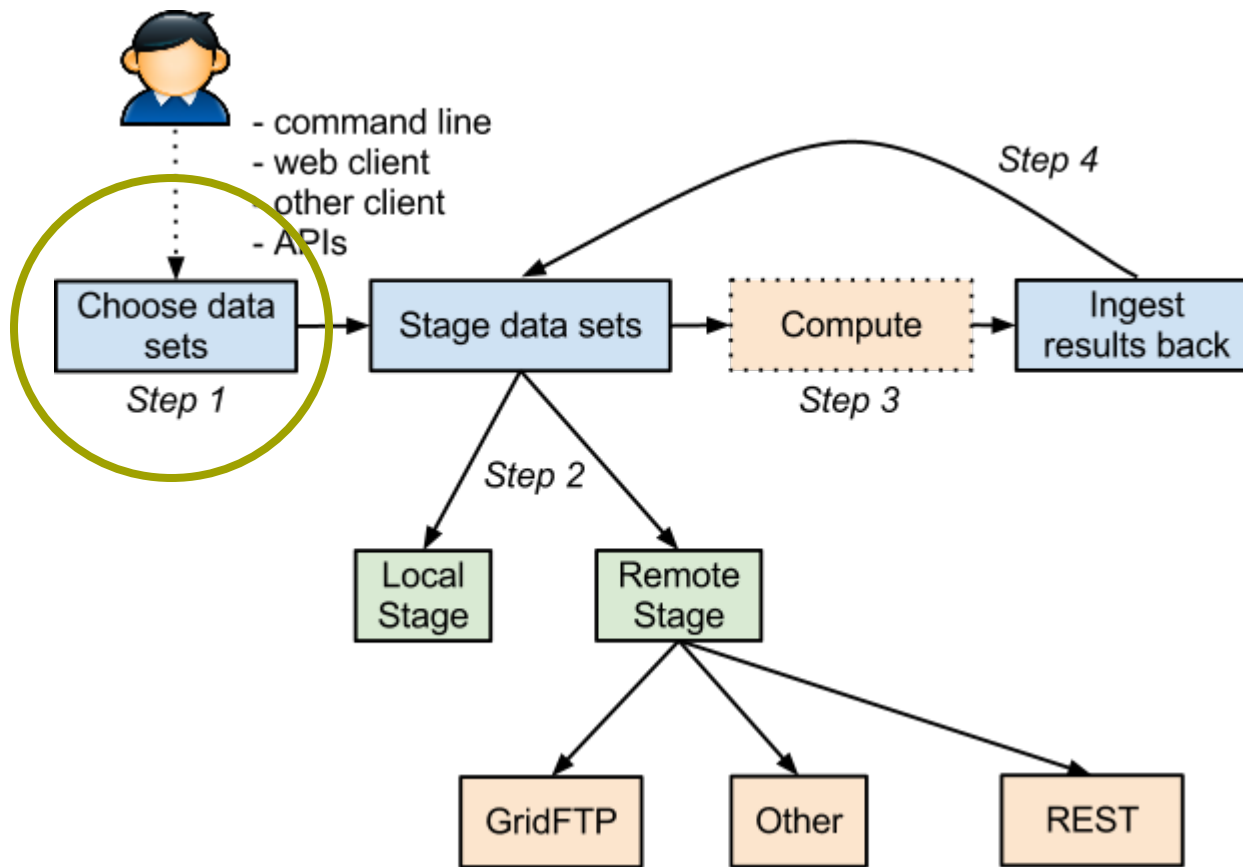
- Any user, part of a community or citizen scientist.
- In most part of the cases, a **domain expert**, such as a biologist, a linguist, a seismologist, not an expert of data transfer software.

Wizard gap



As Ethernet speeds have increased, there has been a widening gap between ability of the novice to fully use bandwidth capabilities, compared to that of network capabilities achieved by users who have had their network tuned by an expert ("Wizard").

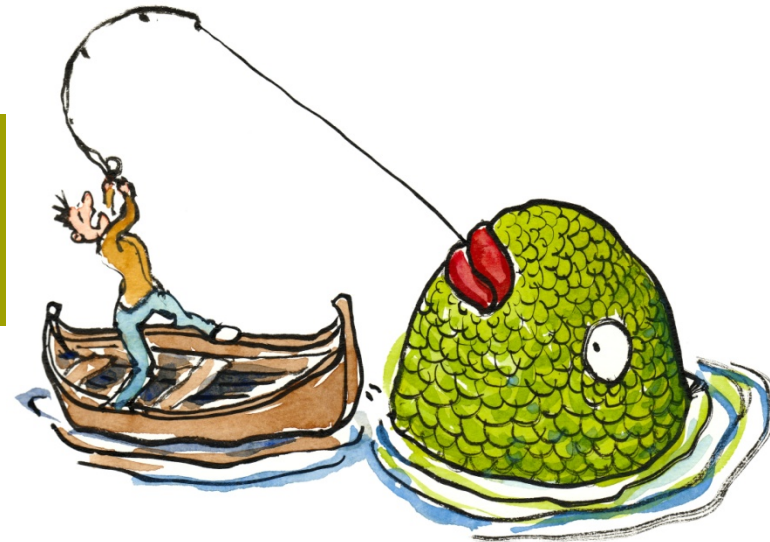
Data selection



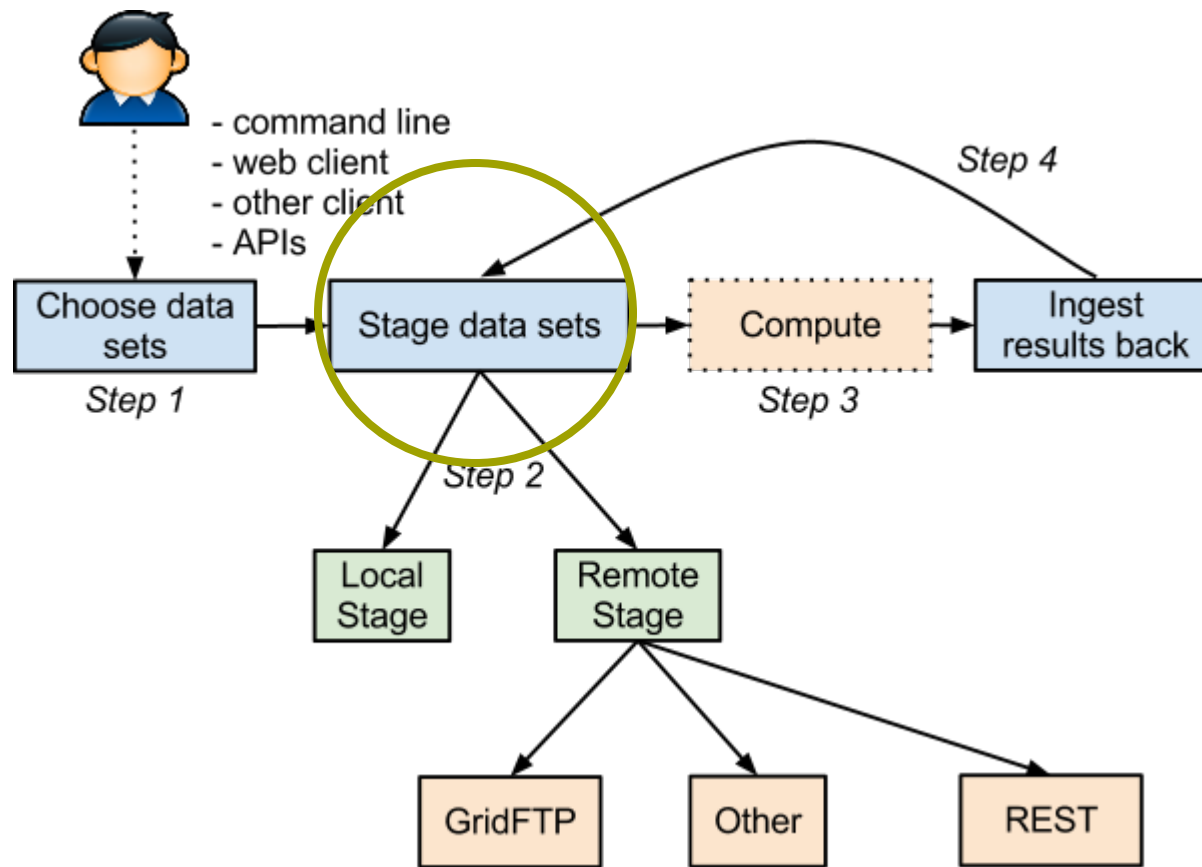
How the data sets are selected by the domain experts?

Through their **local tools**, where “local” means on the community side

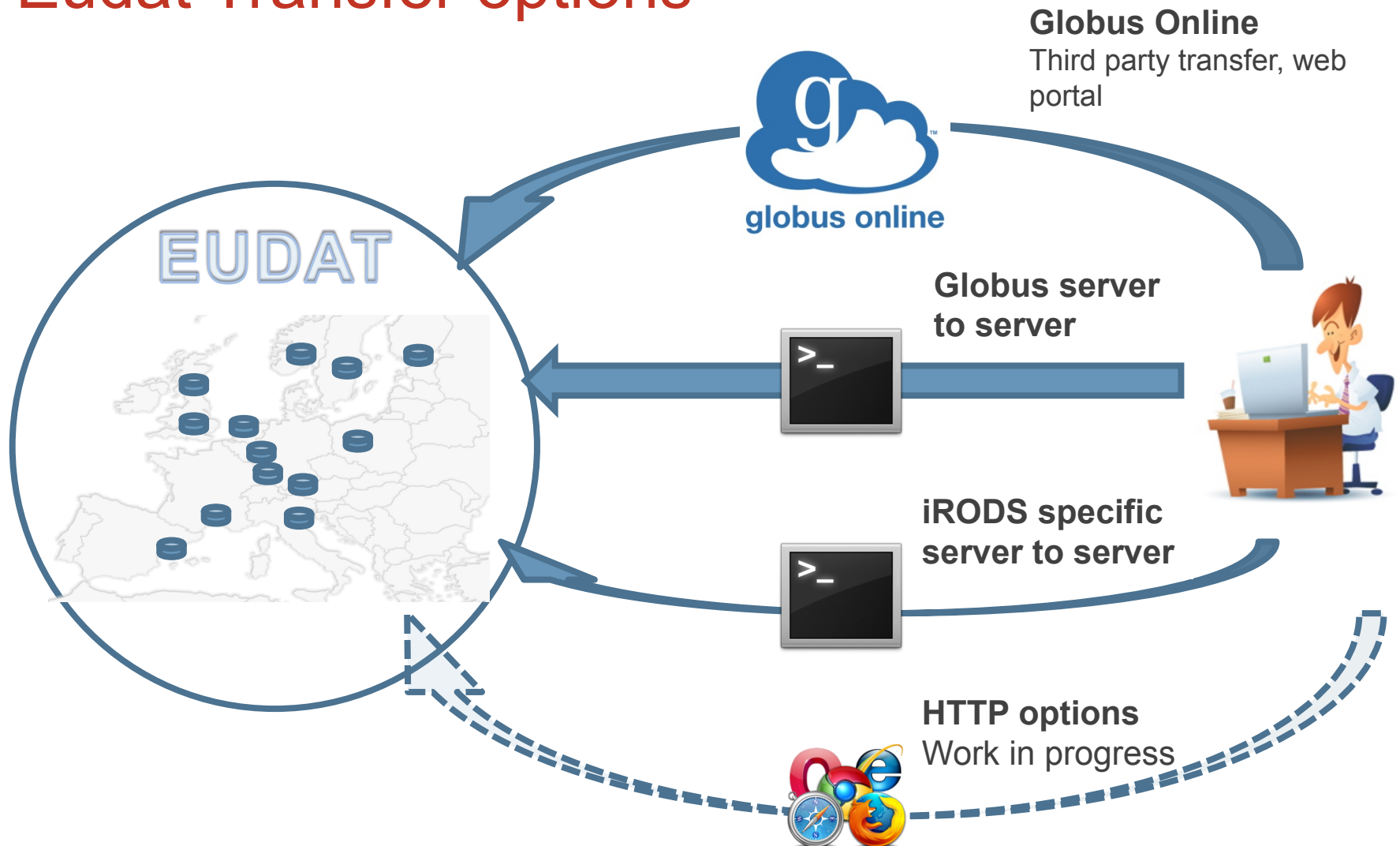
Through **remote services**, generic or specific per community



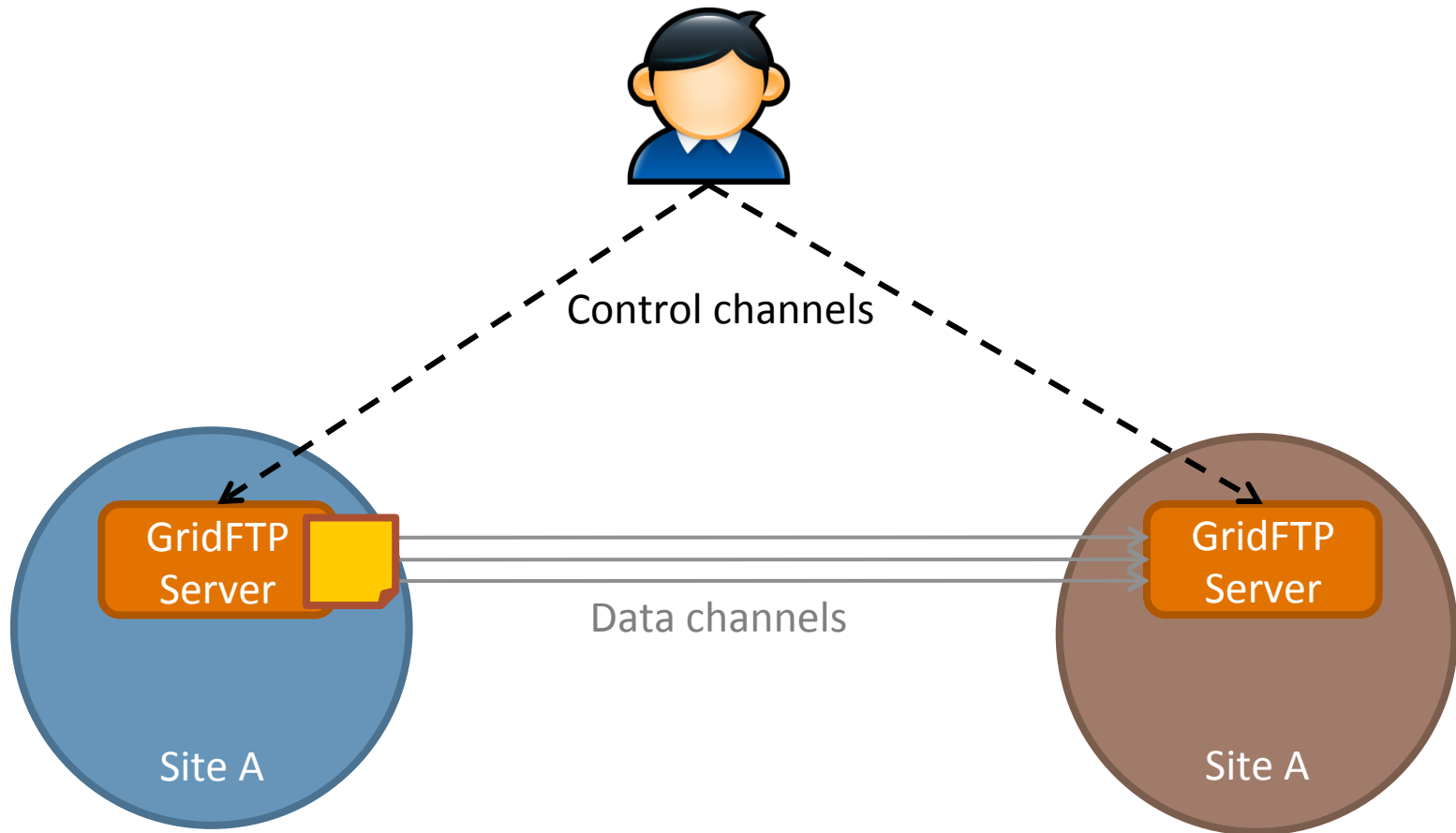
Move the data



Eudat Transfer options



GridFTP: third Party Transfer



Globus OnLine Service

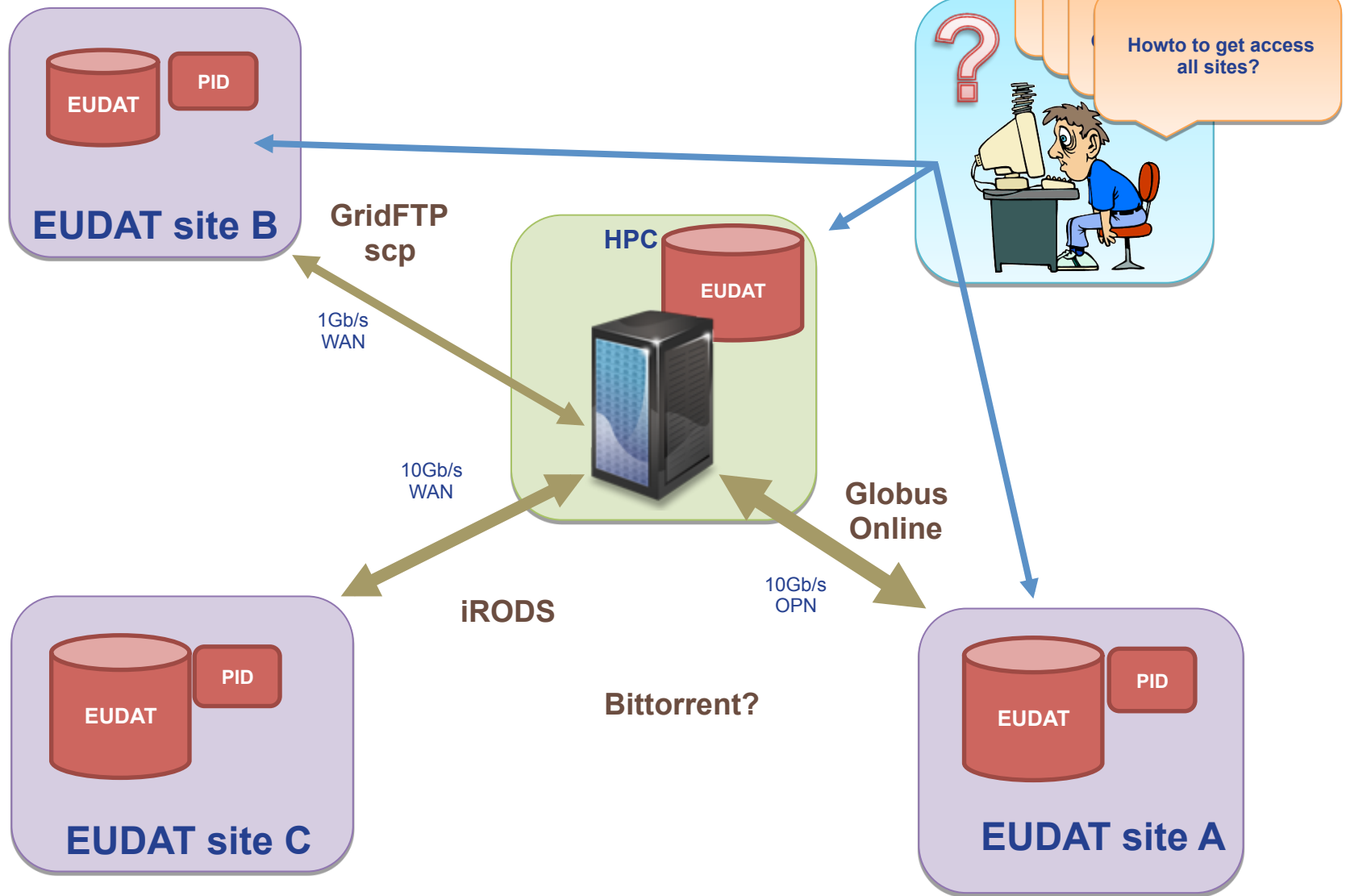
The screenshot displays the Globus OnLine Service web interface in a browser. The address bar shows the URL <https://www.globusonline.org/xfer/StartTransfer>. The page has a blue header with the Globus logo and navigation links. The main content area is titled "Transfer Files -" and includes a dropdown menu for "source overwrites files on destination".

Two endpoint selection panels are visible:

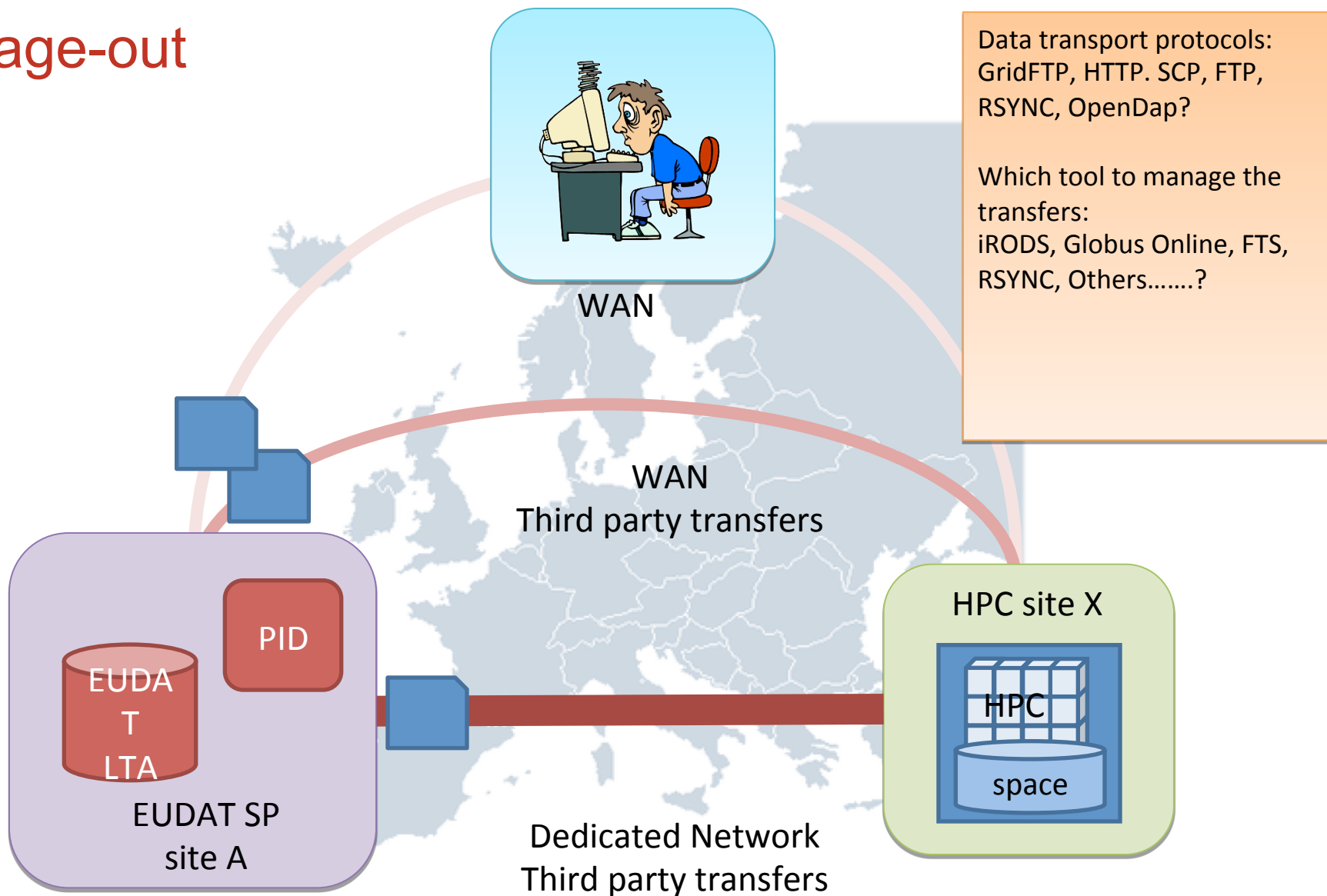
- Left Panel:** Endpoint is `mcarpene#pdl`, Path is `/~/`. The file list includes folders like "Documenti", "GSI-SSHTerm_IGE_for_PRACE_DGRID_LRZ-v1.3.2", "Immagini", "Modelli", "Musica", "Scaricati", "Scrivania", "Ubuntu One", "Video", "VirtualBox VMs", "globusconnect-1.4", "rpmbuild", "workspace", and files like "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), and "skype_2.2.0.35-0natty1_amd64.deb" (22.49MB).
- Right Panel:** Endpoint is `mcarpene#PLX`, Path is `/~/../`. The file list includes folders like "asdata", "cineca", "prod", "user_test", "useragip", "userbmwor", "usercorsi", "userdeisa", "userdompe", "userexternal", "userfercfd", "userforfait", "userfranc", "usergrant", "userhpe", "userhyper", "userinaf", "userincm", "userinternal", and "userjrc".

At the bottom, there is a "Label This Transfer" section with a text input field and a "Get Globus Connect" link. A note states: "Turn your computer into an endpoint. The easiest and most convenient way to send and receive files on your machine."

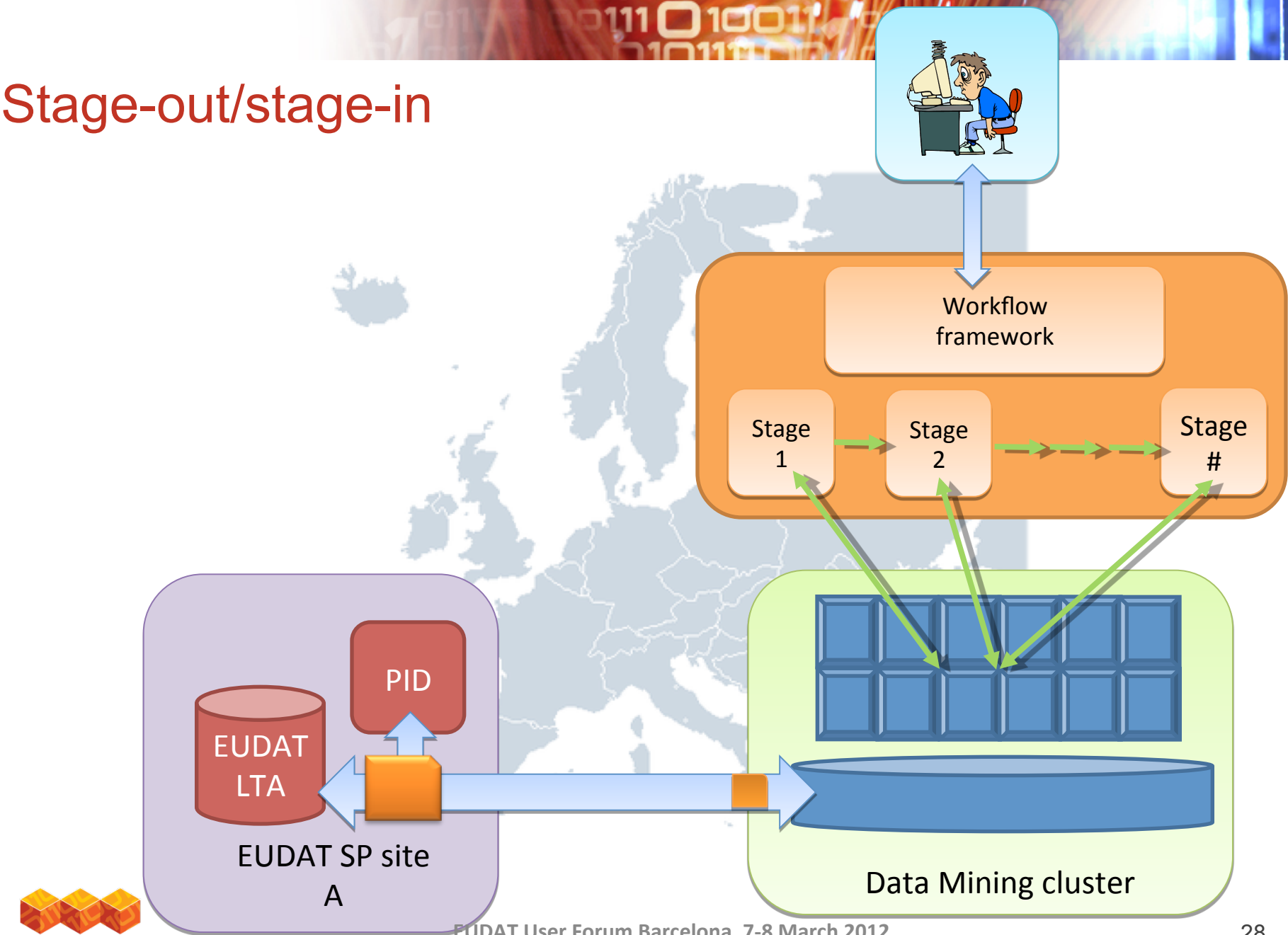
<http://www.globusonline.org>



Stage-out

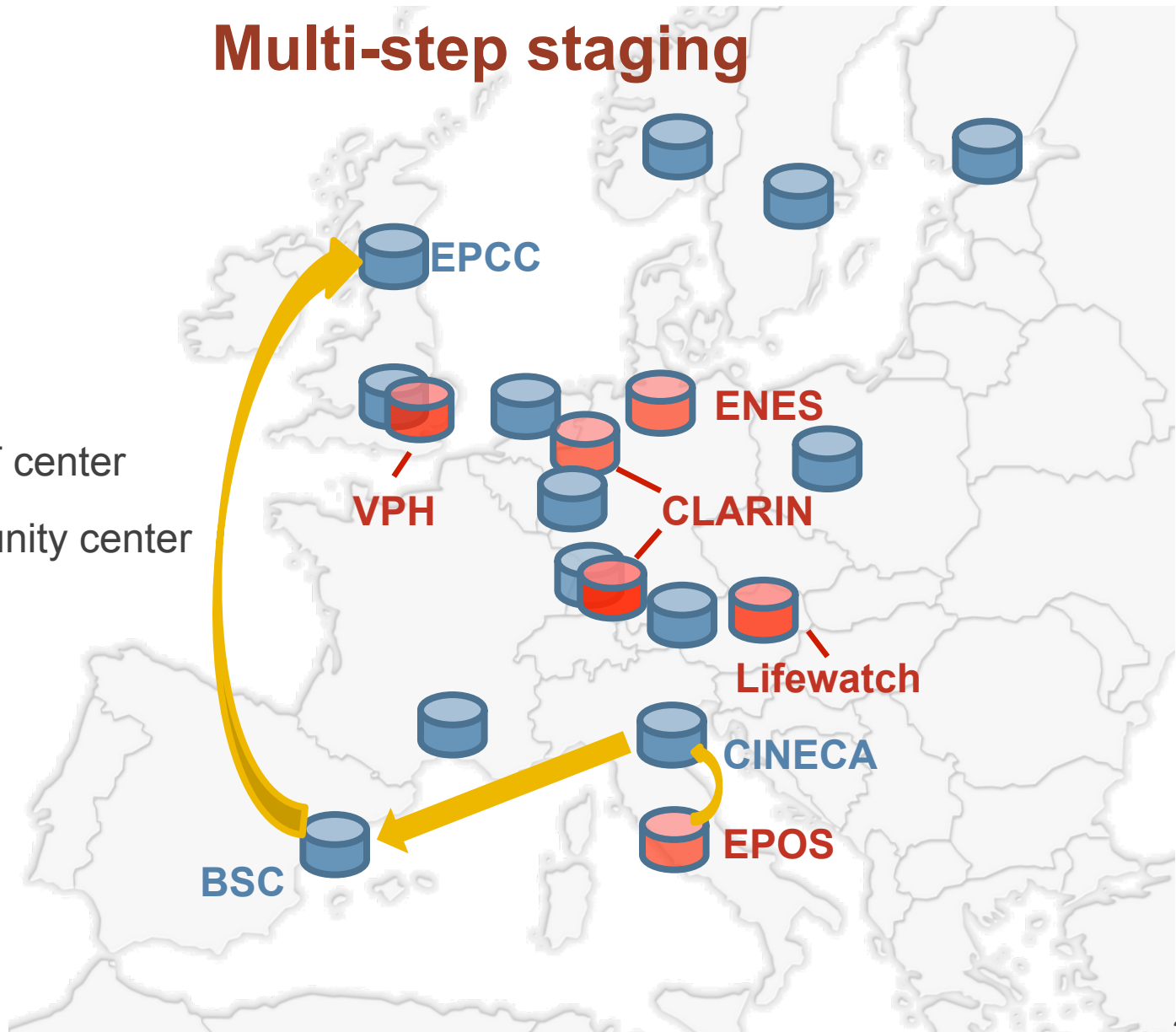


Stage-out/stage-in



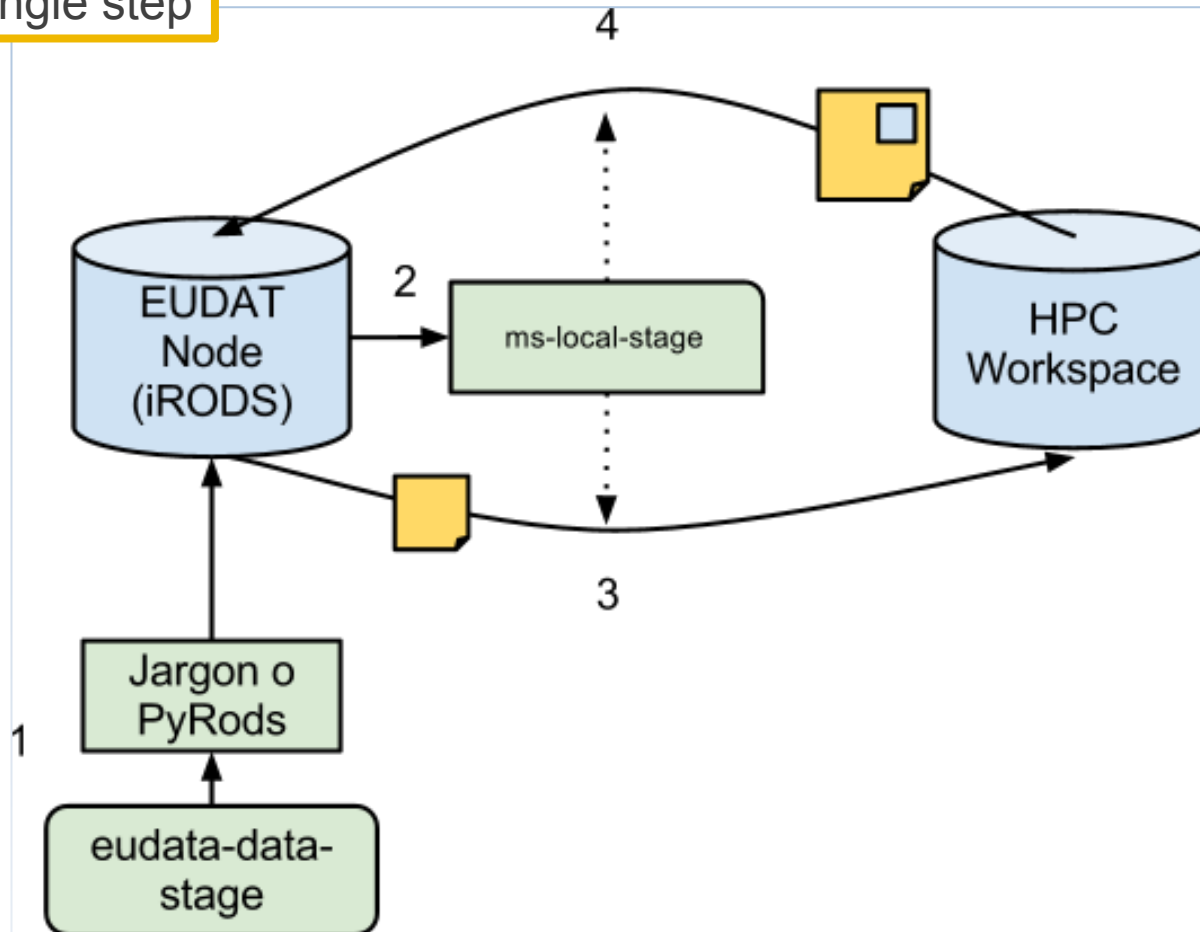
Multi-step staging

 EUDAT center
 Community center



Local staging targets

1. Single step





Sample Data Transfer Results

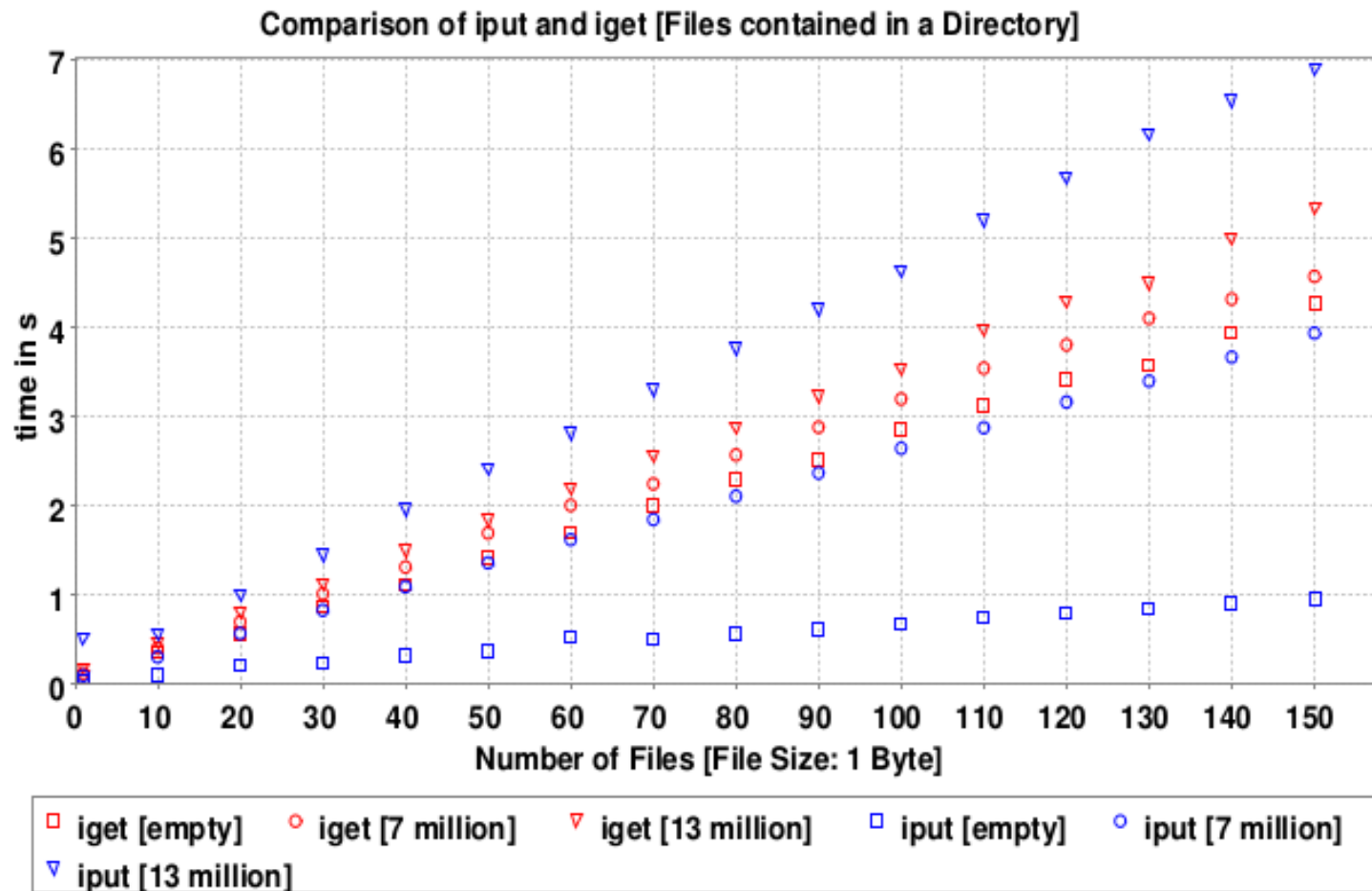
- Sample Results:
 - RTT = 53 ms, network capacity = 10Gb/s.
- Tool Throughput
 - **scp**: 140 Mb/s
 - **HPN patched scp**: 1.2 Gb/s
 - **FTP**: 1.4 Gb/s
 - **GridFTP**, 4 streams 5.4 Gb/s
 - **GridFTP**, 8 streams 6.6 Gb/s



iRODS

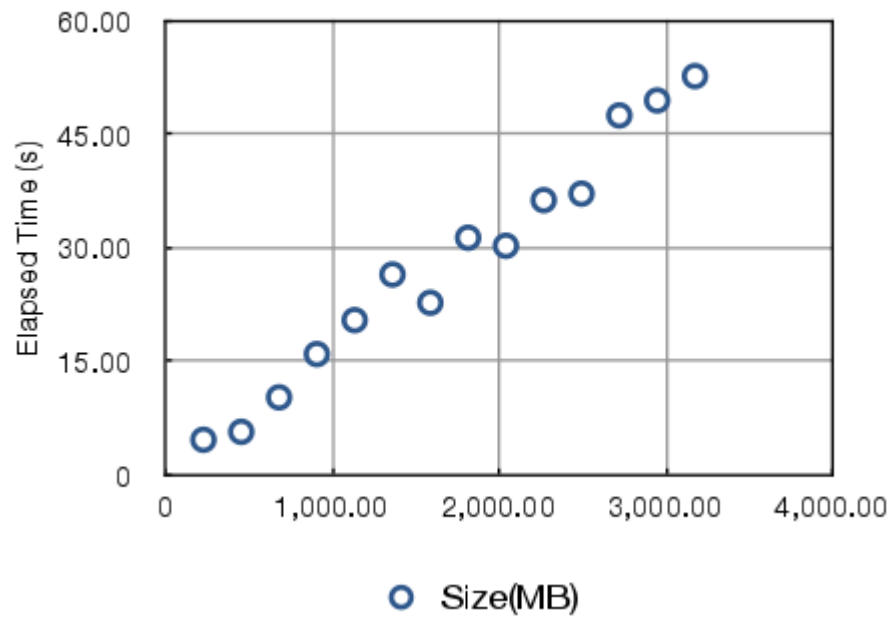
- iRODS is a data management system, which integrates a **rule engine**
- **The rules can be triggered automatically** based on specific events (for example a data set is moved to a particular directory)
- **Invoked from remote** via iRODS command line client or integrated into applications based on iRODS java (Jargon) and python (PyRods) API

Metadata!

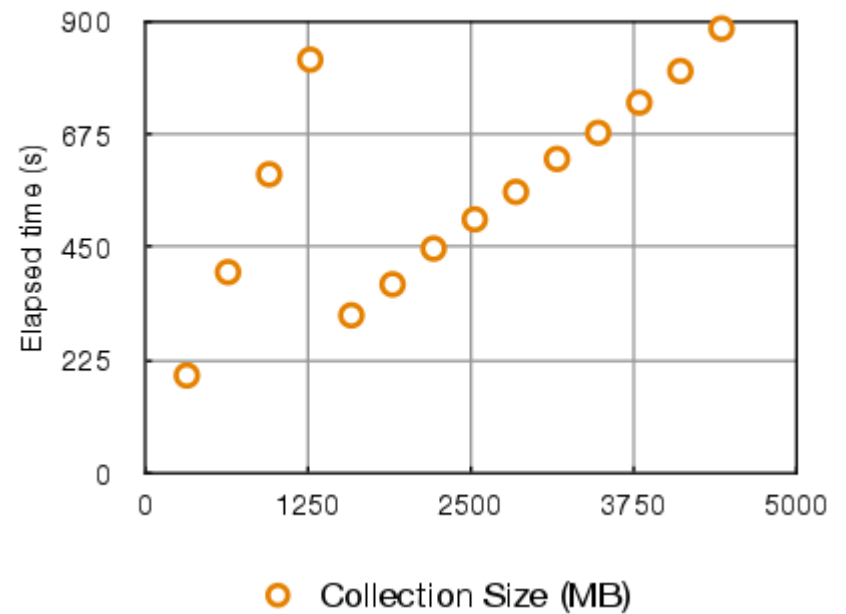




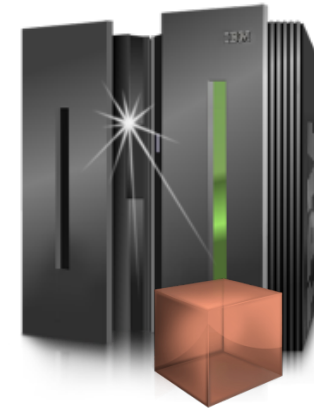
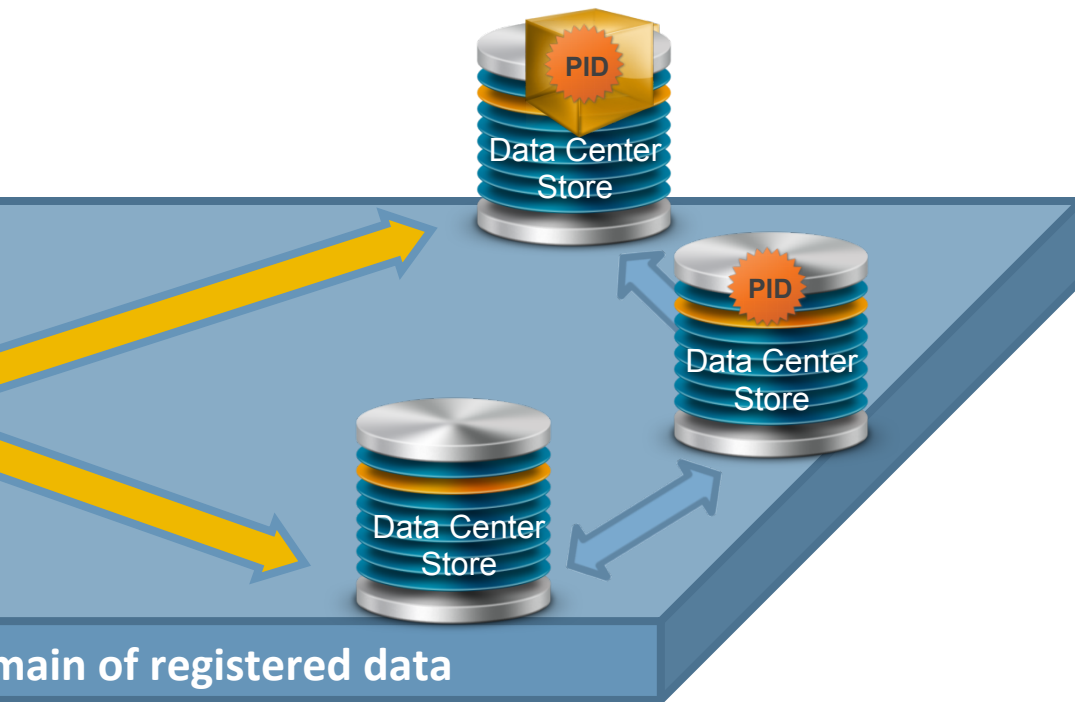
Transfer Time vs File Size (Tar-gzipped Collection)



Transfer Time vs Collection Size



Data Staging and registered data





Conclusions

- The way to move data has to be enough **flexible** to accomodate different transfer protocols, different access mechanisms.
- Flexibility means also that the transfer tools can be used as they are, with default parameters, for average performances, but also **fine tuned** by experts **for faster transfers**.
- No solution fits all, so different services are provided

Contact Us



<http://www.eudat.eu>



eudat-info@postit.csc.fi

Project Coordinator: Kimmo Koski
kimmo.koski@csc.fi

Scientific Coordinator: Peter Wittenburg
peter.wittenburg@mpi.nl

Project Manager: Damien Lecarpentier
damien.lecarpentier@csc.fi

Dissemination Manager: Nagham Salman
nagham.salman@bsc.es

Industry Task Force: David Manset
dmanset@maatg.fr





Thank you!