
The BaBar Experiment's Distributed Computing Model

Dominique Boutigny

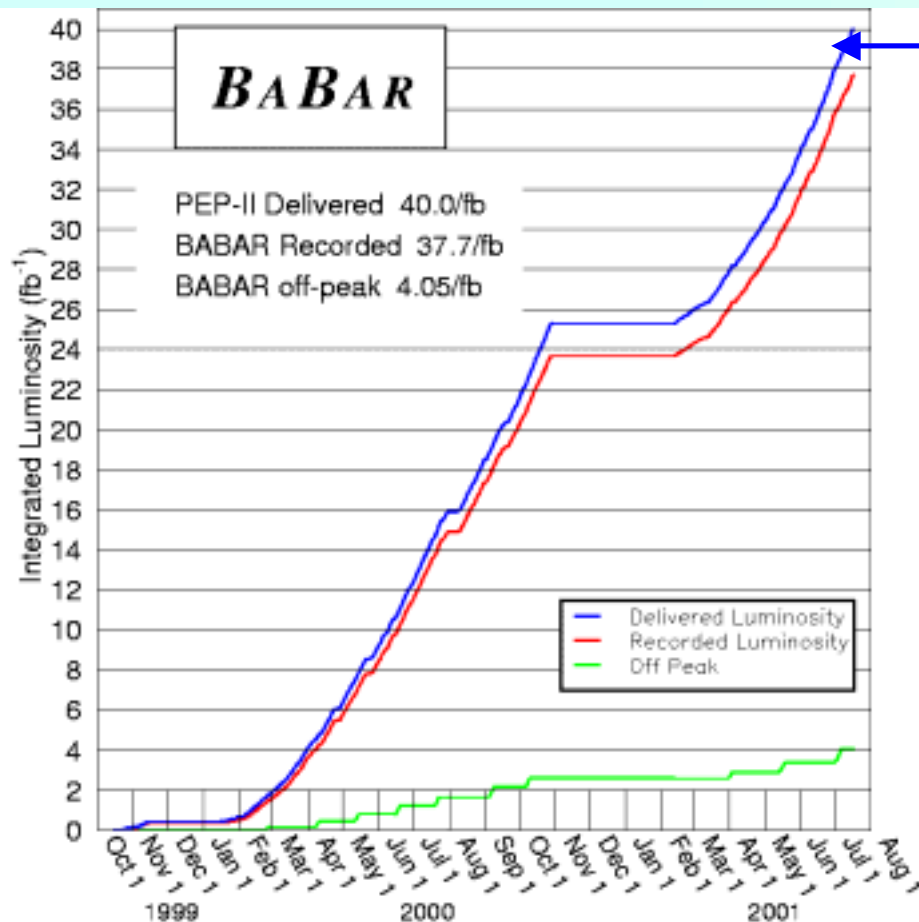
LAPP-CNRS/IN2P3

On behalf of the BaBar Collaboration's Computing Group

CHEP01 - Beijing

Physics Framework of the BaBar Experiment

- BaBar is recording Physics quality data since October 1999



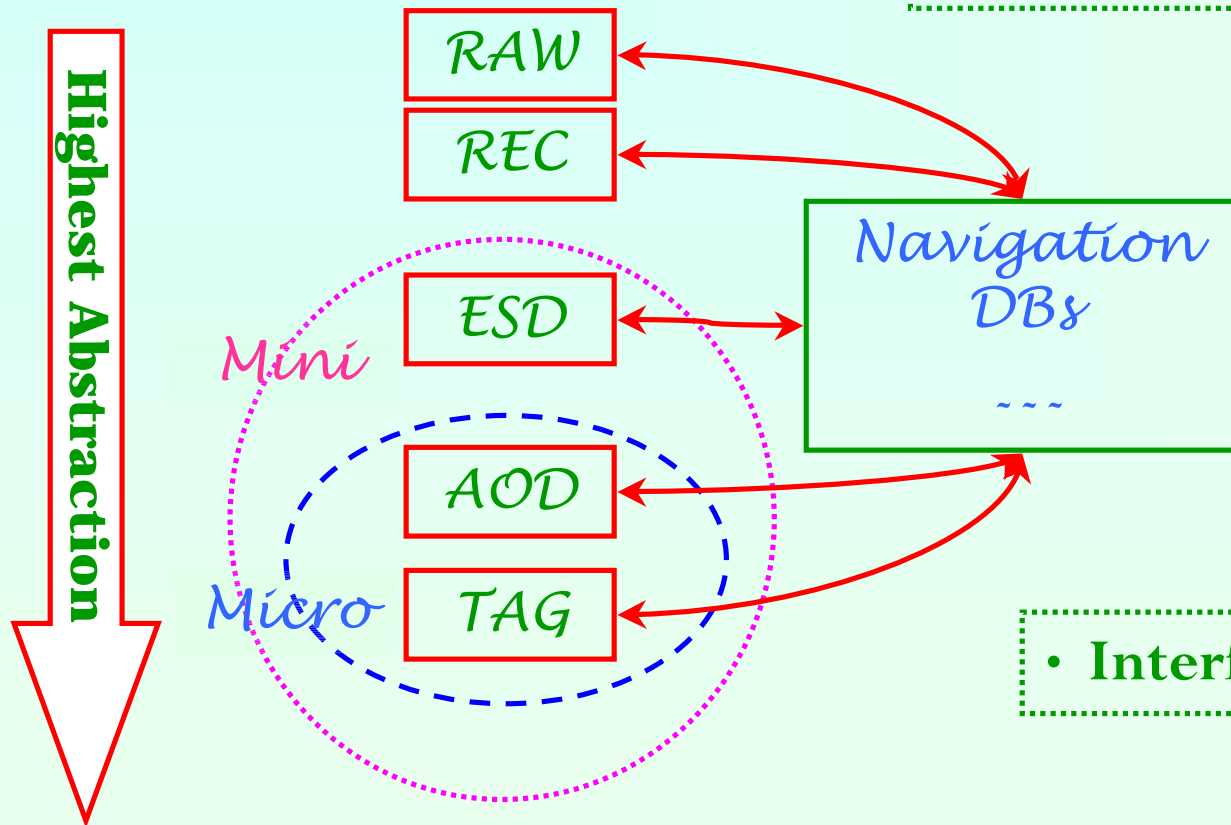
July 15

- Luminosities greater than the design are regularly achieved ($3.4 \cdot 10^{33} \text{cm}^{-2}\text{s}^{-1}$)
- Yearly integrated luminosity is expected to grow in the future
 - 25 fb⁻¹ in 2000
 - 225 fb⁻¹ expected in 2005 (total 620 fb⁻¹)

Chosen Technology for Data Storage and Access

- BaBar has chosen the Objectivity OO Database technology for its main event store

• Hierarchical structure



• Interfaced to HPSS

Event-Store Size

- RAW : ~ 50 KB /evt
- REC : ~ 150 KB /evt
- TAG : ~ 0.5 KB /evt
 - Suitable for fast physics filtering
- Micro : ~ 5.4 KB /evt (*including navigation databases*)
 - Suitable for standard physics analyses
- Mini : =Micro + ~4.7 KB /evt
 - Suitable for detailed physics analysis and standard detector studies

“evt” = Typical hadronic event

1999 + 2000 Event-store:

– Data

- ~27500 files
- ~128 TB

– Disk resident data: 7.3 TB

– Monte-Carlo simulation

- ~25300 files
- ~75 TB

– Disk resident data: 3.1 TB

Today:

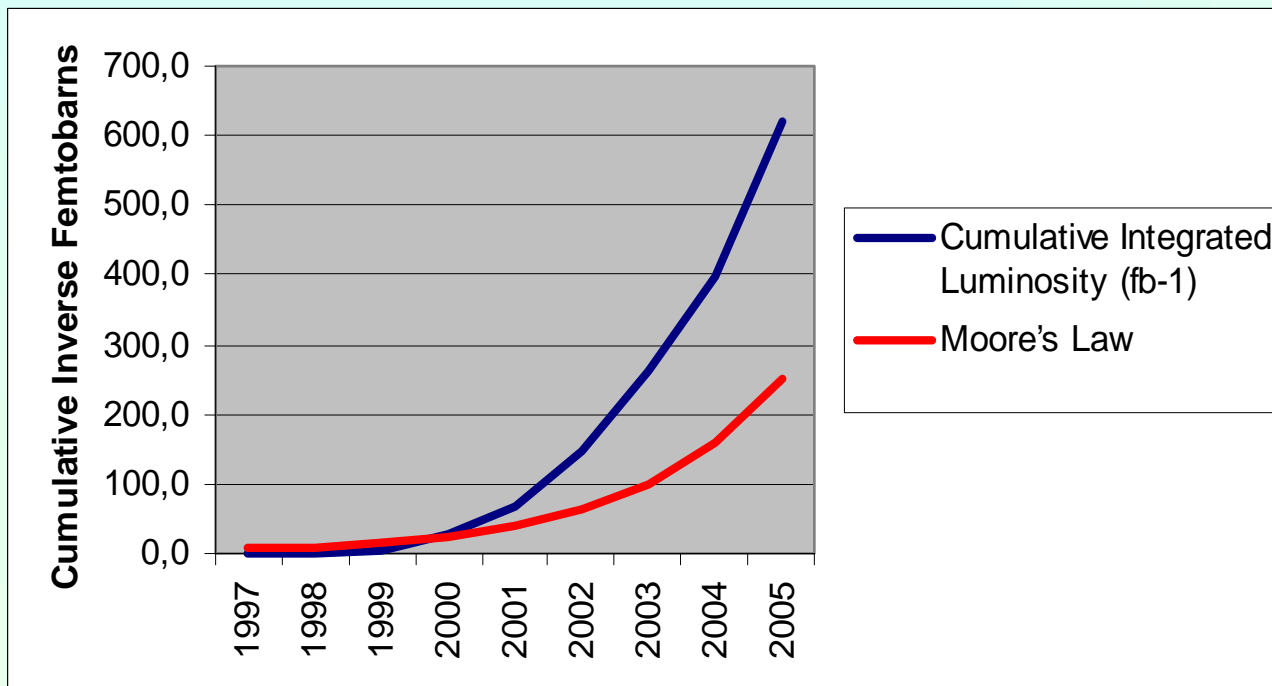
- 350 TB of data

Expected Evolution of Computing Power Needs

Observation:

- The BaBar integrated luminosity is expected to grow faster than the Moore's law (computing "power" increases X2 / 18 months)

→ The consequence is that the annual cost of the computing is expected to grow



→ Definition of a new computing model

Main Components of the BaBar Computing Model (1)

Data Format

- The Objectivity DB is the main BaBar event store
 - Easy creation of filtered events: “Skims”
 - Collections of pointers to real events
 - Little overhead
 - Only the navigation databases are needed
 - Powerful tool for site hosting a large fraction of the data
 - But exportation of skimmed collections is very inefficient.
 - Even worse due to Objectivity limitation of 64K DB / federation
 - Large databases (2 to 10 GB)
 - Exportation of 1 collection → Several hundreds of GB of data
 - Introduction of a compact format based on ROOT (KanGA)
 - Self contained events
 - In parallel :
 - Development of *multiple-federation*
 - Smaller databases (0.5 MB)
 - ~20 self-contained physics streams (unit for data distribution in remote sites)
- KanGA withdrawal if the new system is
OK

Main Components of the BaBar Computing Model (2)

A multi-Tiers model “a la LHC”

- Tier A :
 - Main data repository
 - Main distribution center
 - Should host at least 30% of the total data sample
 - **No or small duplication among Tier A**
 - Any level of reconstruction available on disk or on mass-storage → **RAW data ...**
 - Any physics stream available
 - All type of analyses possible
 - Physics analyses
 - Detailed detector studies
- Tier B :
 - Serve a region
 - Secondary data distribution center
 - Host data in a format suitable for physics analyses (μ DST)
- Tier C :
 - Individual institutes and universities
 - Small sample of data \leftrightarrow physics interest of the site
 - **KanGA data now - Will move to Objectivity later**

Data Distribution (1)

Distribution of KanGA / Root data is described in Poster #4-023

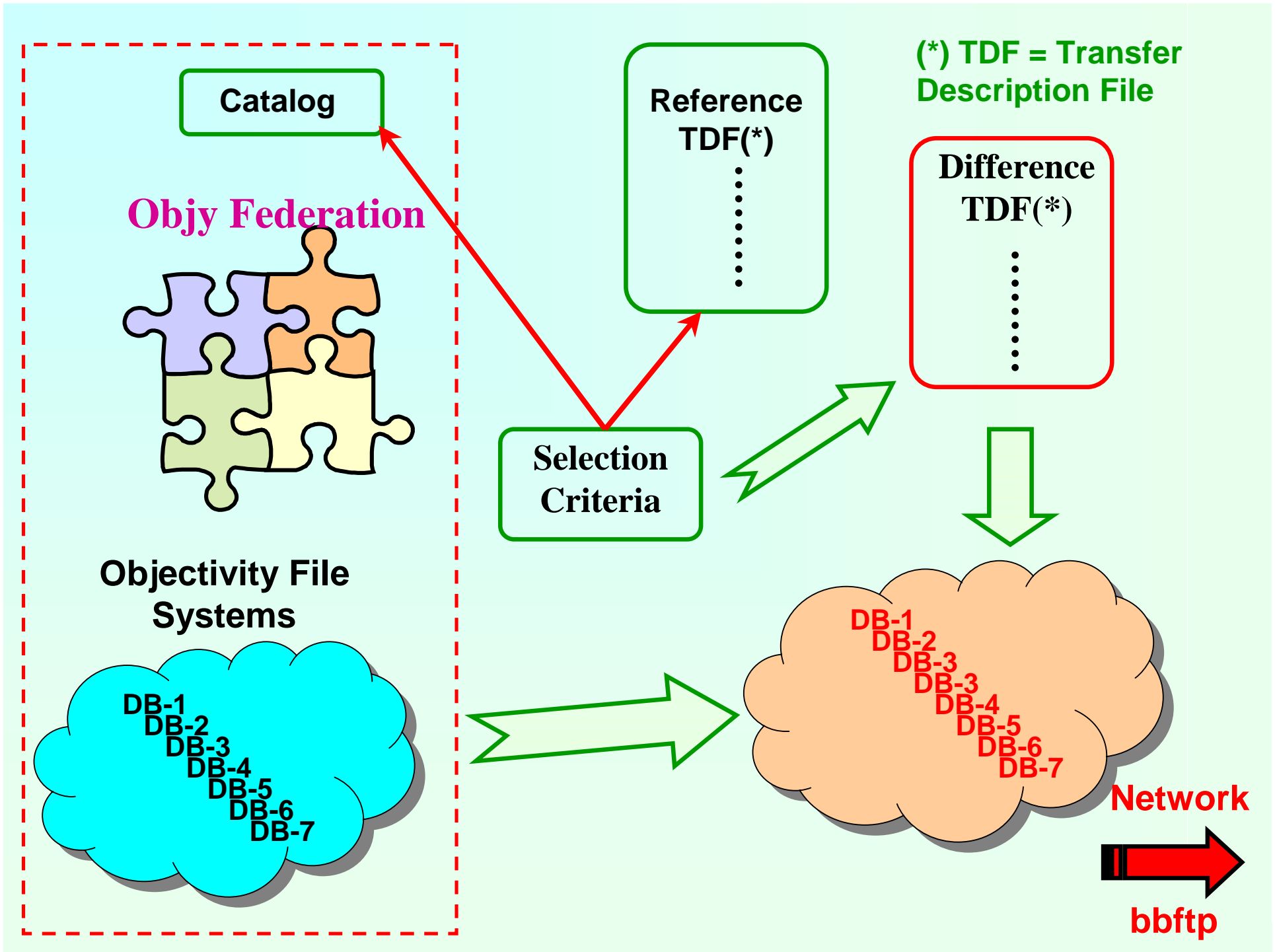
Expected amount of beam data to be transferred to Tier A sites (in TB)

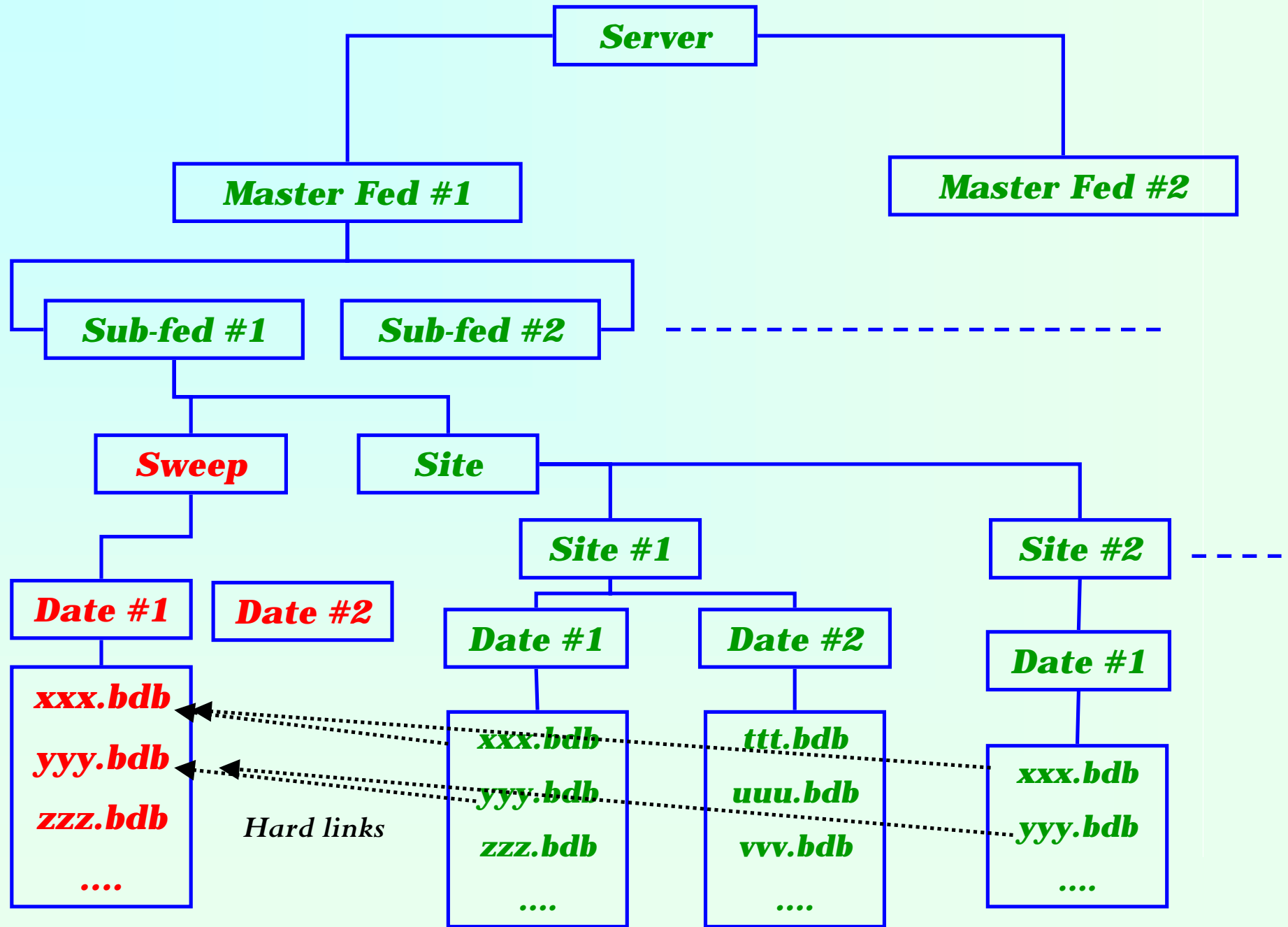
	2001	2002	2003	2004	2005
Tag	1,5	1,1	2	3	4,7
Micro	15	11	14	21	33
Mini	20	18	19	30	47
Rec	22	54	98	148	232
Raw	9	16	29	44	70
Sum	66 TB	101 TB	162 TB	246 TB	386 TB
Bandwidth	54 Mb/s	82 Mb/s	131 Mb/s	199 Mb/s	312 Mb/s

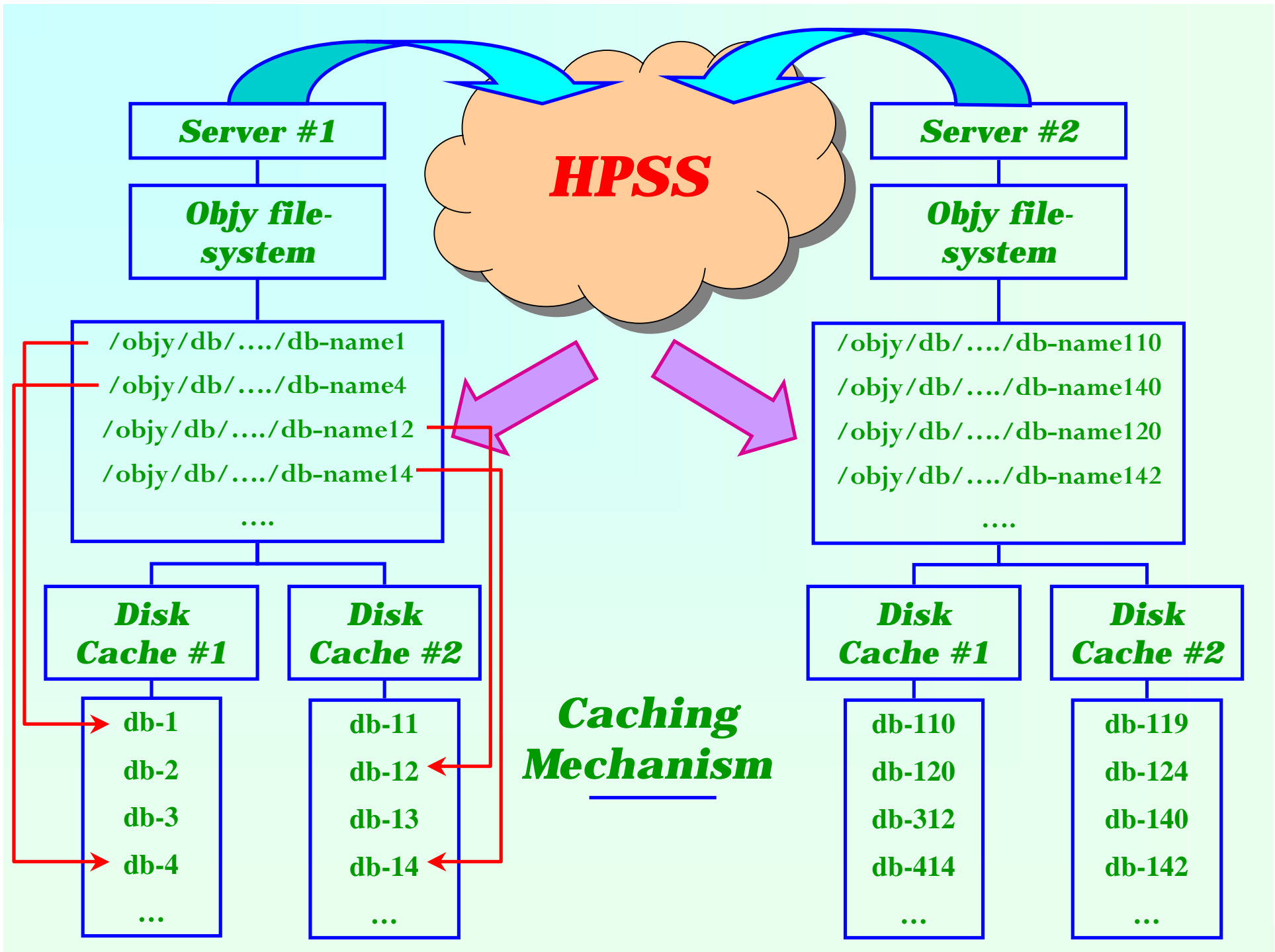
- Transfer of MC data will multiply these numbers by a factor 2 to 3
- OC12 links will soon be necessary

Data Distribution (2)

- The Data Distribution System is using the BaBar/Objectivity C++ API to interact with the event store.
- + collection of shell, tcl, perl scripts for file manipulation and book-keeping
- 1. Dump the Objy catalog
- 2. Compare each entry with a reference file
- 3. Define which databases should be exported → **Criteria**
- 4. Extract the database files
- 5. Update the reference file
- 6. Transfer the database files
- Database selection criteria are based on
 - DB type
 - Physics stream
 - Data / Monte-Carlo
 - Federation #
- Each site interested in receiving data (*only 1 at the moment*) can have different selection criteria
- Part of the export procedure is automatic but could also be controlled by simple E-Mail to the data distribution server







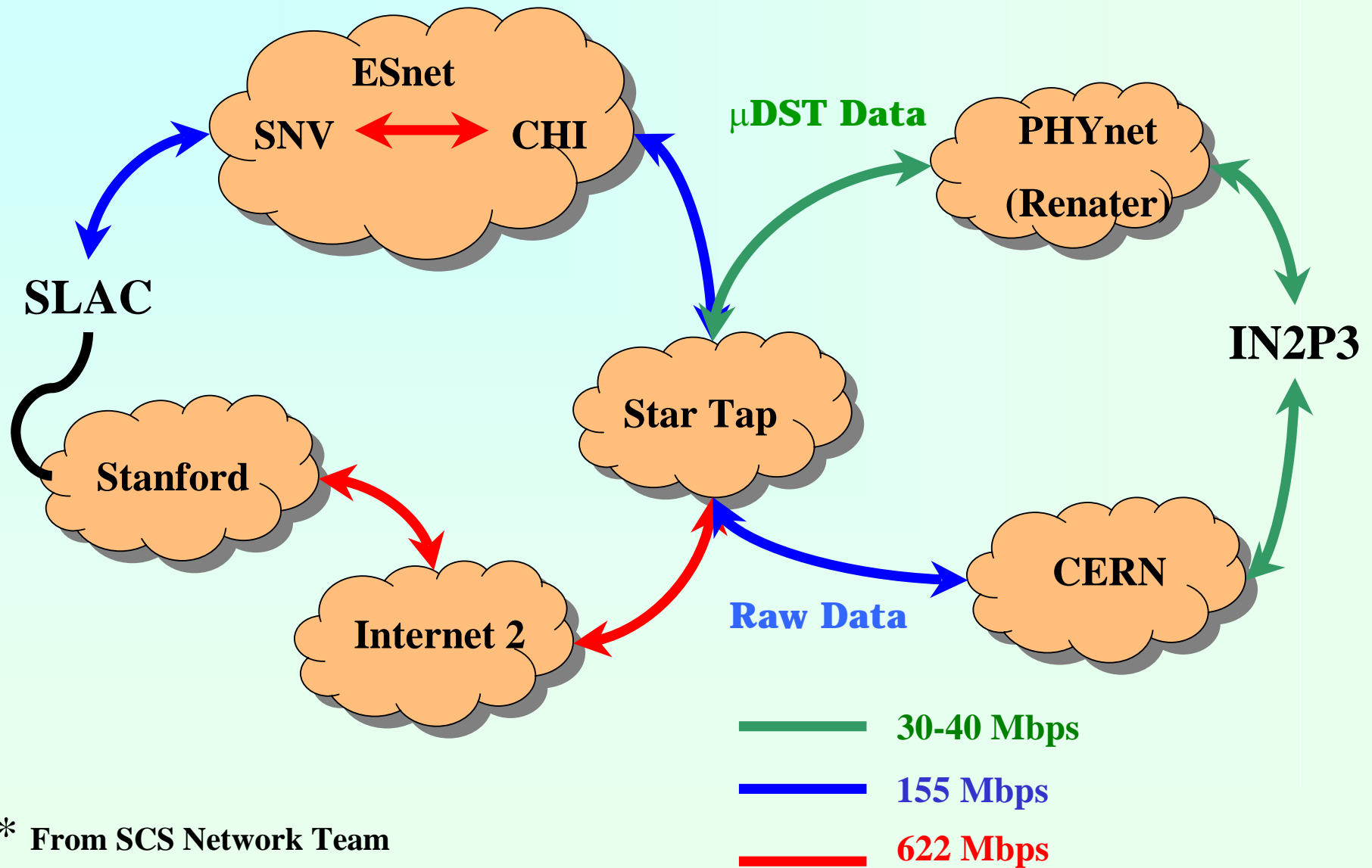
Importation Procedure at IN2P3

- In2p3 Objectivity current configuration consists in:
 - 6X 4-CPU Sun Netra servers
 - 2X 1 TB of disk / server
 - lock, journal servers
- Each server is using the caching mechanism (see previous)
- The import procedure is using a Java application
 - Detect available export at SLAC
 - Create description file for the transfer tool (bbftp)
 - Transfer DB files to a dedicated import server
 - Create federation if needed
 - (fast)Attach DB to federation
 - Move DB files to HPSS
 - If DB is an update / existing DB
 - ➔ Delete the old DB from disk
 - The new version will be recalled automatically by HPSS when needed

The **bbftp** Network Transfer Tool

- Network transfers are done with **bbftp** (original author: Gilles Farrache from CCIN2P3) <http://doc.in2p3.fr/bbftp/index.html>
 - High performance
 - Multi-Stream
 - Adjustable buffer and TCP window size
 - Automatic error recovery
 - Customizable security module (ssh, kerberos, Unix pwd...)

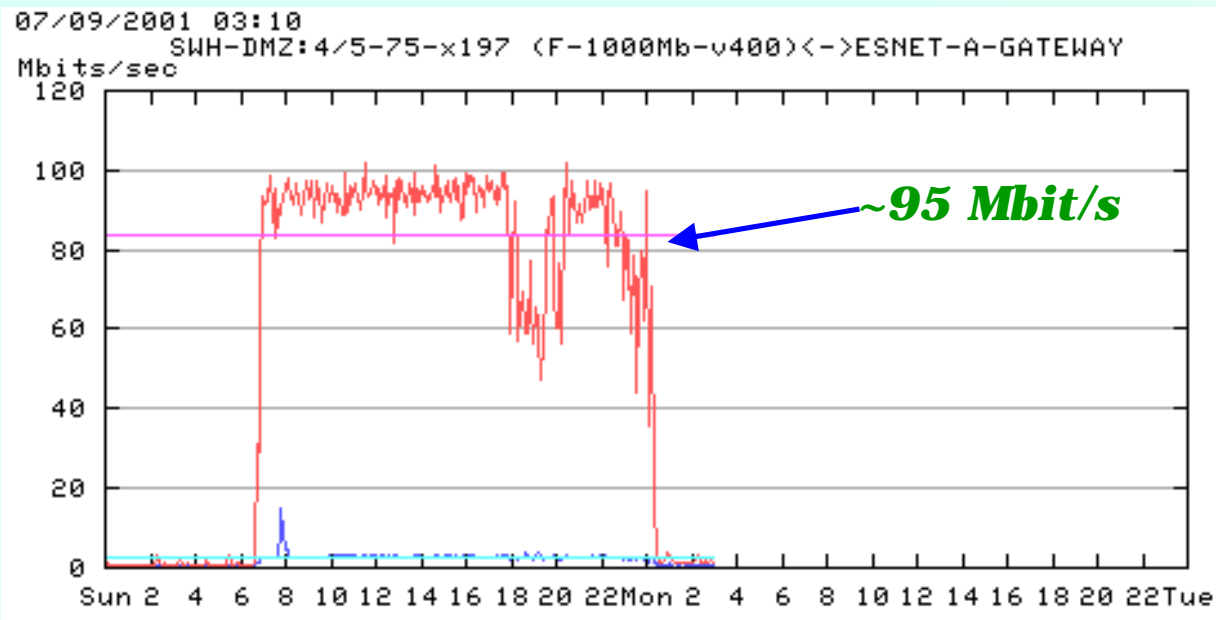
Network Connection between SLAC and IN2P3 (*)



* From SCS Network Team

Network Performance

- While transferring *on both links* we get the following transfer speed on the SLAC-ESNET link
- On the ESNET-CERN link we are not able to get more than 70 Mb/s over 155 Mb/s. The reason is unknown : Bottleneck, protocol ... ?



Export Status

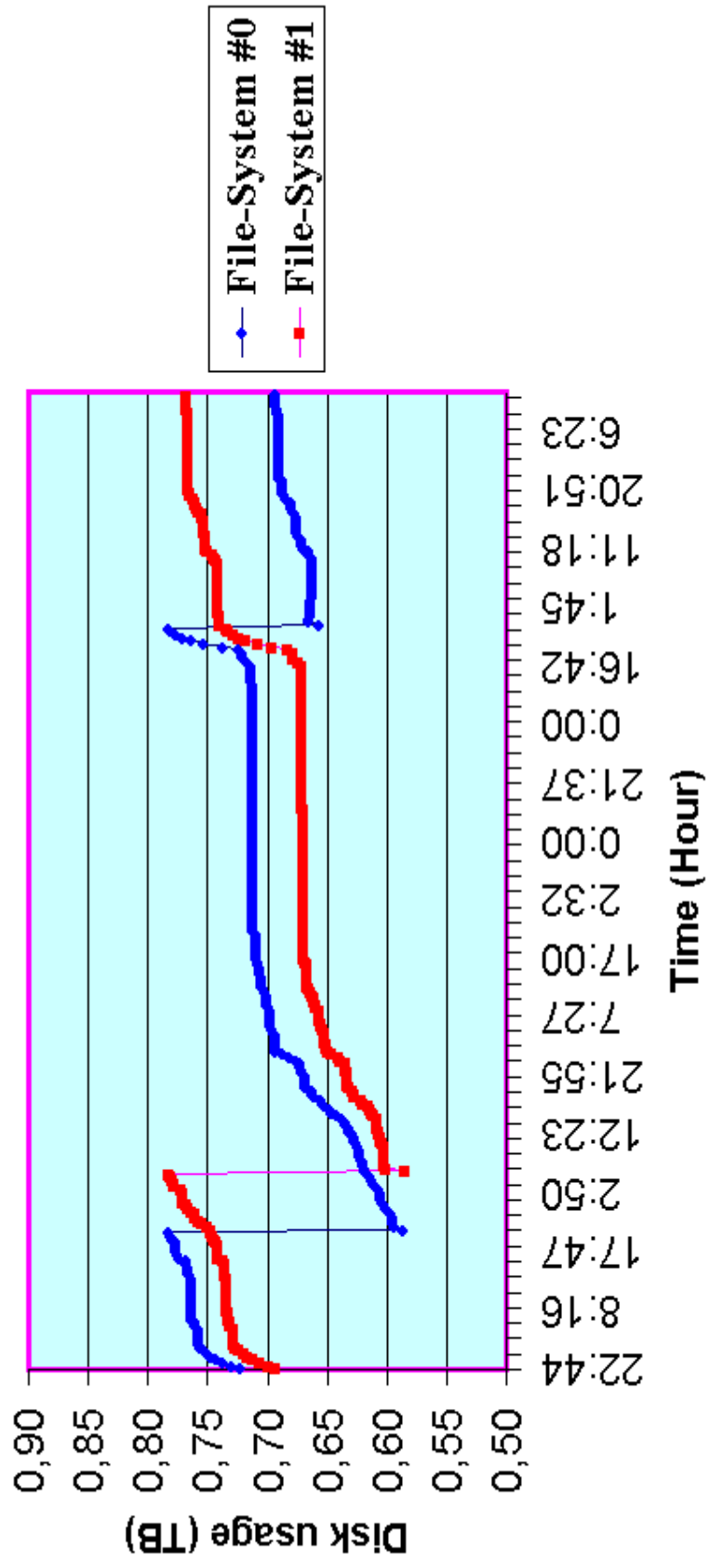
- Size of weekly μ -DST export: 200–900 GB
- Extraction speed
 - Average 5 MB/s
 - Maximum 12 MB/s
 - ➔ Need to parallelize !
- In 2001: 7.4 TB of Micro DST data extracted and transferred to IN2P3
 - Overhead due to re-exportation of updated DB: ~25%
 - Delay between SLAC and IN2P3 < 48H
- The RAW and REC data are extracted and transferred ~continuously from the SLAC HPSS
 - 4.6 TB already exported

Total: 12 TB SLAC ➔ IN2P3 in 2001

One Word on HPSS at IN2P3

- IN2P3 has activated the automatic purging / staging system for all Objectivity data
- The BaBar/Objy Advanced Multi-threaded Server (AMS) is able to retrieve automatically the staged data
- Performances are OK, but the number of simultaneous jobs is still low → **Scaling ?**
- IN2P3 is also providing user space in HPSS for n-tuple storage :
 - Accessible through RFIO
 - Performances are very good up to now
 - Very flexible
 - Allows “infinite” storage for analysis

**File-system occupancy
With HPSS automatic staging**



Conclusions

- BaBar is evolving toward a distributed multi-Tier computing model
- SLAC is the primary Tier-A
- IN2P3 is ramping up and is already providing significant data analysis resources
- Other Tier-A will come soon
- Next step: Stop the duplication between sites
- Data distribution tools have been setup to transfer data in a timely manner
 - Present tools will be difficult to scale → **GRID tools**
- Network is critical and 622 Mb/s links will soon be necessary
 - Will require development to use the full bandwidth
 - Data transfer tools
 - Objectivity data extraction software

Distributed data analysis will require to develop tools for remote job submission using the GRID technology → Starting now