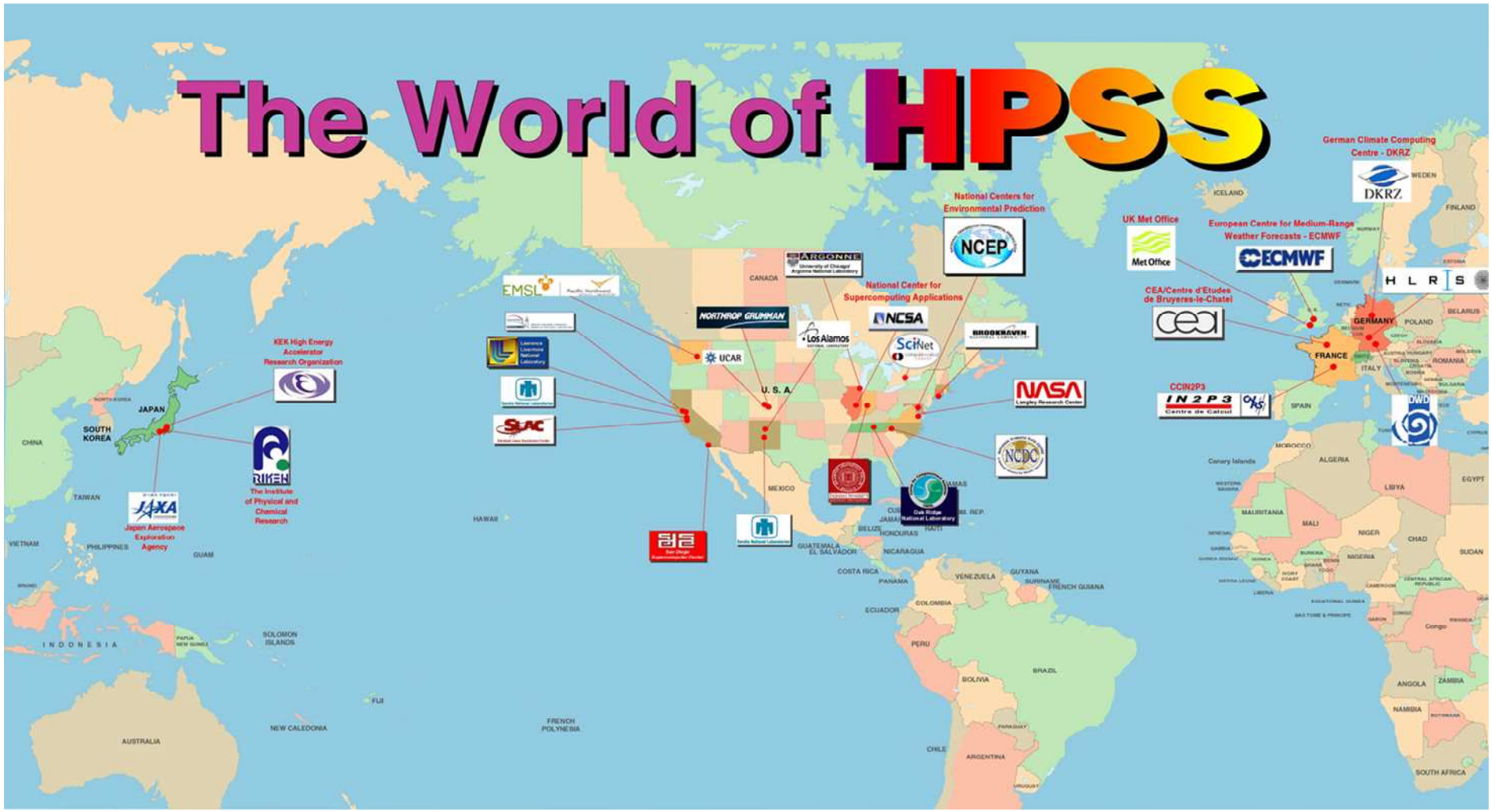
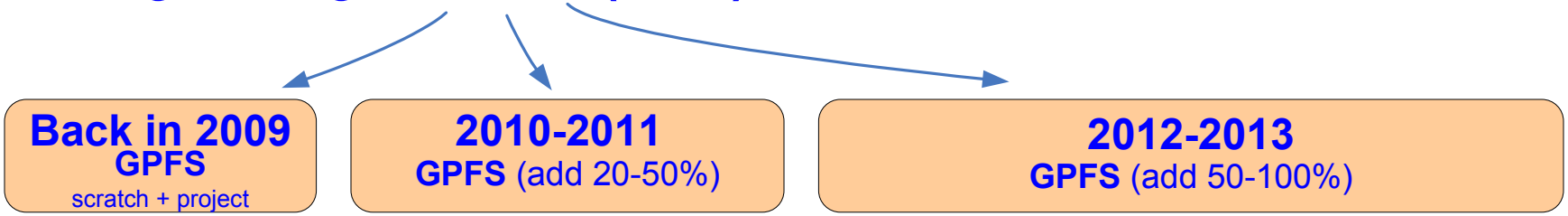

The World of HPSS
National Centers for Environmental Prediction
NCEP
German Climate Computing Centre - DKRZ
DKRZ
UK Met Office
Met Office
European Centre for Medium-Range Weather Forecasts - ECMWF
ECMWF
National Center for Supercomputing Applications
NCSA
CEA/Centre d'Etudes de Bruyeres-le-Chatel
ceal
KEK High Energy Accelerator Research Organization
RIKEN
The Institute of Physical and Chemical Research
JAXA
Japan Aerospace Exploration Agency
CCIN2P3
IN2P3
Centre de Calcul
NASA
Langley Research Center
NCDC
Oak Ridge National Laboratory
EMSL
ARGONNE
University of Chicago Argonne National Laboratory
NORTHROP GRUMMAN
Los Alamos
BROOKHAVEN
SciNet
UCAR
SLAC
U. S. A.
CANADA
MEXICO
JAPAN
SOUTH KOREA
CHINA
TAIWAN
VIETNAM
PHILIPPINES
GUAM
HAWAII
INDONESIA
SOLOMON ISLANDS
NEW CALEDONIA
FIJI
AUSTRALIA
FRENCH POLYNESIA
PERU
BOLIVIA
BRAZIL
CHILE
ARGENTINA
COLOMBIA
ECUADOR
VENEZUELA
GUYANA
SURINAME
FRENCH GUIANA
COSTA RICA
PANAMA
NICARAGUA
HONDURAS
GUATEMALA
EL SALVADOR
ICELAND
FINLAND
WEDEN
NORWAY
ESTONIA
BELARUS
POLAND
GERMANY
FRANCE
SPAIN
ITALY
ROMANIA
MOROCCO
ALGERIA
LIBYA
EGYPT
MAURITANIA
MALI
NIGER
CHAD
SUDAN
NIGERIA
Congo
ANGOLA
ZAMBIA
NAMIBIA
SOUTH AFRICA
Canary Islands

Storage Capacity Expansion Plan (initial)

Storage Budget: **$ $ $ (5PB)**



Rationale:

* the longer we wait, the more we can buy with the same dollar amount (hopefully)
* add storage as demand increases

Usage management to date:

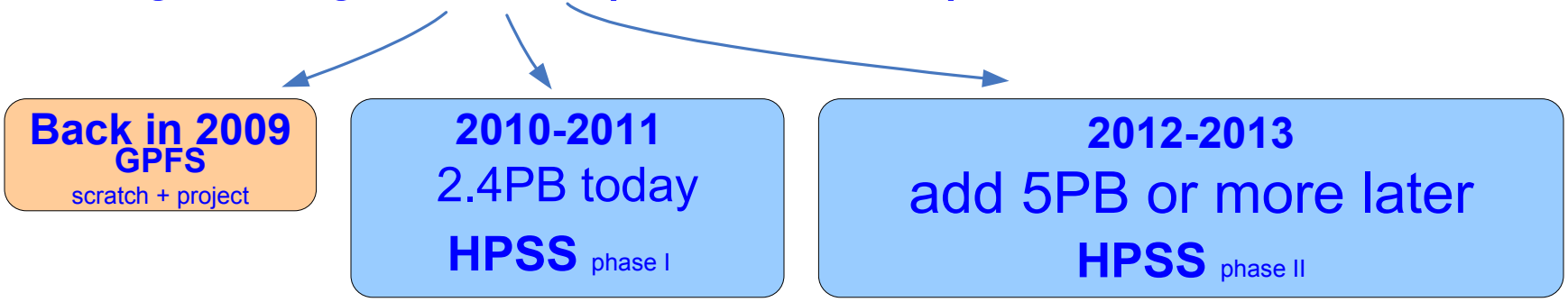
* allocations
* introduction of quotas
* HSM (limited data offload capability)
* regular purging (90 days old material)
* 97-98% → cleanup effort email

What have we learned in 2 years ?

* prices did not come down as we hoped for over this period
* GPFS still has problems and limitations at our scale
  (4000 nodes cluster)
* user data distribution patterns not GPFS or HPC friendly
* adding spinning disks to GPFS →
  more heat, higher electricity bill, more ”parked” data
* more users, more data, more files →
  more problems on active file systems
* 900+ users in Sep/2011 →
  we will need way more than 5PB and sooner

Storage Capacity Expansion Plan (revised)

Storage Budget: **$ $ $** **(5PB** or more)



Solution:

* near online storage with HPSS
  (tape-backed hierarchical storage system)

Usage management moving forward:

* allocations: GPFS + HPSS
* quotas & massive data offload to HPSS
* regular purging (possibly 60 days old material)
* less utilization of small files
* more use of tarballs in the regular workflow (new campaign)

High Performance Storage System

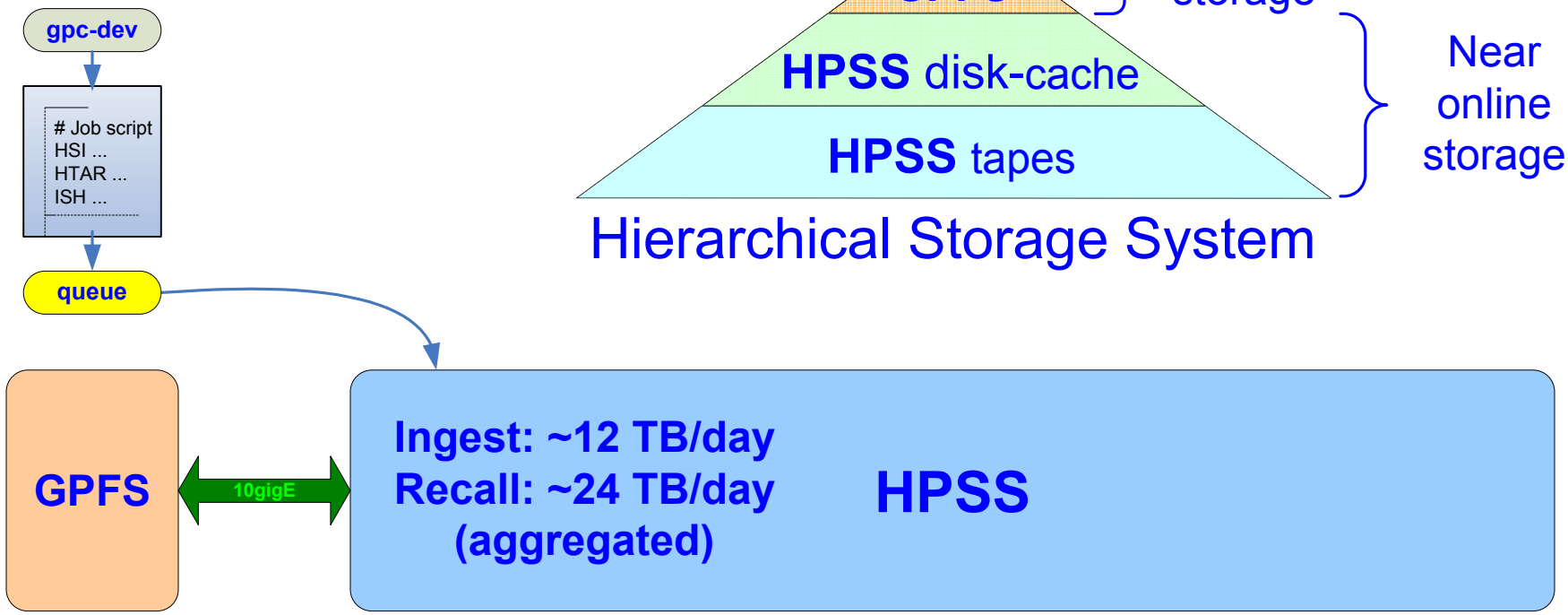
* 10+ years history
* used by 50+ facilities in the “Top 500” HPC list
* very reliable, data redundancy and data insurance built-in.
* highly scalable, reasonable performance at SciNet
* HSI/HTAR clients also used on several HPSS sites.

HSI is a client with an ftp-like interface
- can be used to archive and retrieve large files
- also useful for inspecting the contents of HPSS.

HTAR creates tar formatted archives directly into HPSS.
- It creates a separate index file (.idx) that can be accessed quickly.

ISH is a utility to perform an inventory of contents in your tarballs and HPSS contents (developed at SciNet)

How it works ...



* access and transfer done through the GPC queue system
* end-user interaction via HSI/HTAR/ISH calls on job scripts

Scripted File Transfers

```
#!/bin/bash
#PBS -l walltime=72:00:00
#PBS -q archive
#PBS -N htar_create_tarball_in_hpss
#PBS -j oe
#PBS -m e

echo "Creating a htar of finished-job1/ directory tree into HPSS"

trap "echo 'Job script not completed';exit 129" TERM INT
# Note that your initial directory in HPSS will be /archive/<group>/<user>/

cd /scratch/$(whoami)/workarea/
htar -cpf /archive/$(id -gn)/$(whoami)/finished-job1.tar finished-job1/
status=$?

trap - TERM INT

if [ ! $status == 0 ]; then
   echo 'HTAR returned non-zero code.'
   /scinet/gpc/bin/exit2msg $status
   exit $status
else
   echo 'TRANSFER SUCCESSFUL'
fi
```

headers

htar
hsi
ish

trap

status

Note: Make sure to check the job's exit code for errors after all data transfers and any tarball creation process

Pending jobs can be monitored with showq

```
showq -w class=archive
```

ISH used from the command line:

```
rzon@scinet02:~$ ls
data.tgz
rzon@scinet02:~$ /scinet/gpc/bin/ish
ish 0.98
Ramses van Zon - SciNet/Toronto/Canada/July 8, 2011
[ish]hpss.igz> index data.tgz
[ish]data.tgz.igz> ls -l
drwxr-xr-x rzon/scinet            0 2011-02-10 13:57:01 data/
-rw-r--r-- rzon/scinet        16714 2010-10-05 12:41:45 input.ini
-rwxr-xr-x rzon/scinet          293 2011-06-30 12:42:57 submit.pbs
[ish]data.tgz.igz> cd data
[ish]data.tgz.igz> ls
run1/   run2/
[ish]data.tgz.igz> find important*.dat
run1/important01.dat  run1/important02.dat  run1/important03.dat
run1/important04.dat  run1/important05.dat  run1/important06.dat
run2/important01.dat  run2/important02.dat  run2/important03.dat
[ish]exit
rzon@scinet02:~$
```

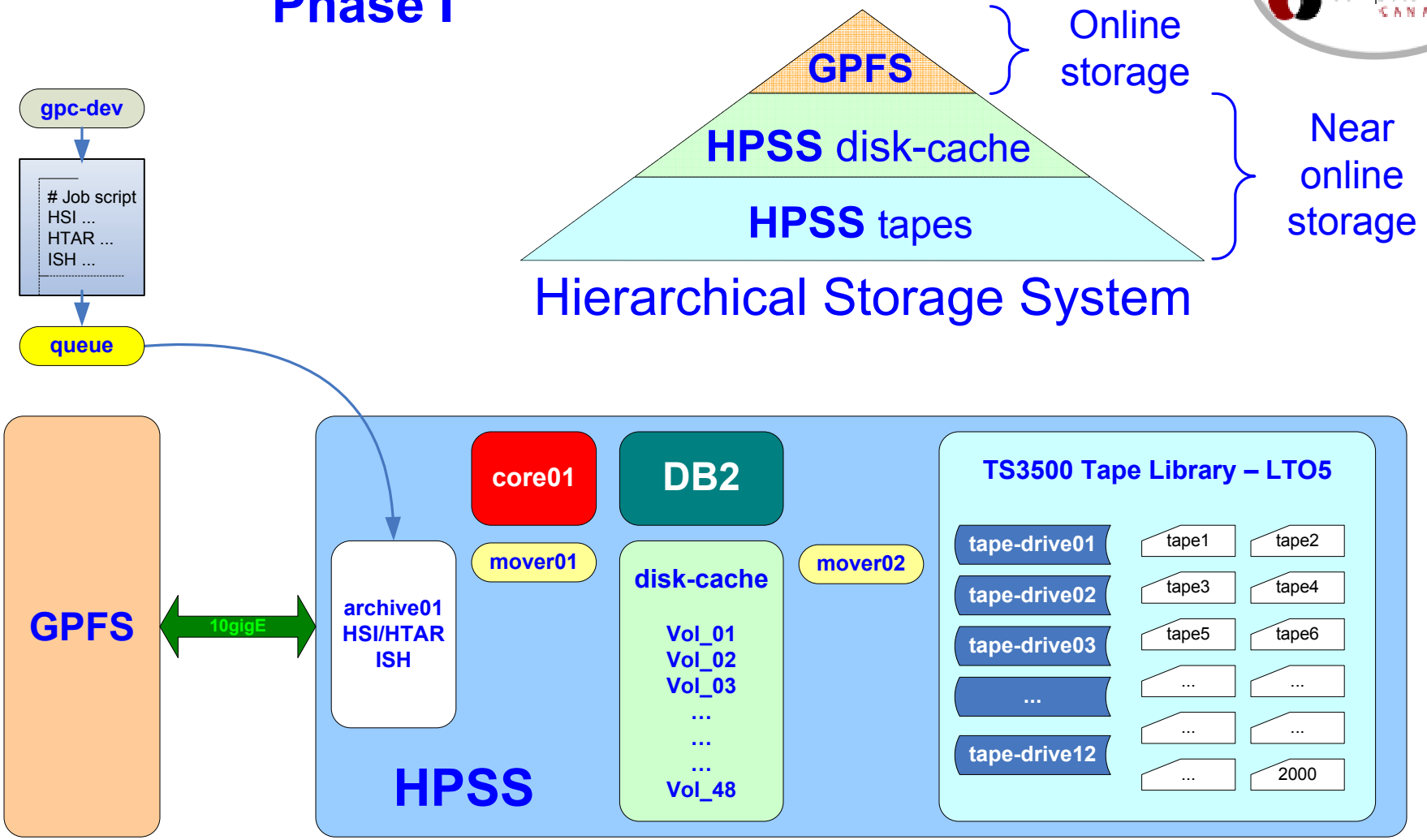
ISH used from a job script:

```
#!/bin/bash
# This script is named: data-list.sh
#PBS -q archive
#PBS -N hpss_index
#PBS -j oe
#PBS -m e
/scinet/gpc/bin/ish hindex
```

For more details and examples please consult the wiki page:
https://support.scinet.utoronto.ca/wiki/index.php/HPSS

HPSS – main components

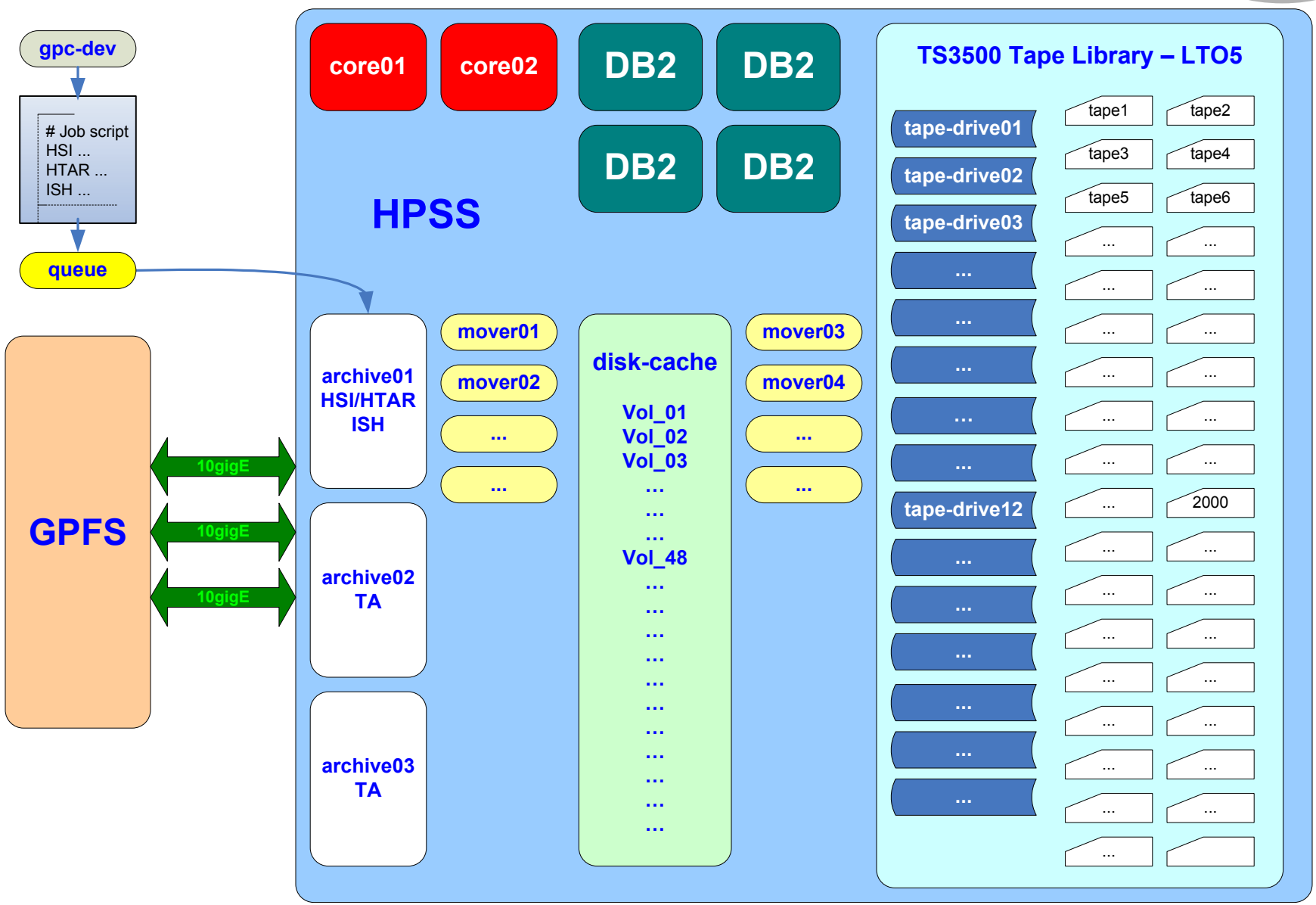
Phase I



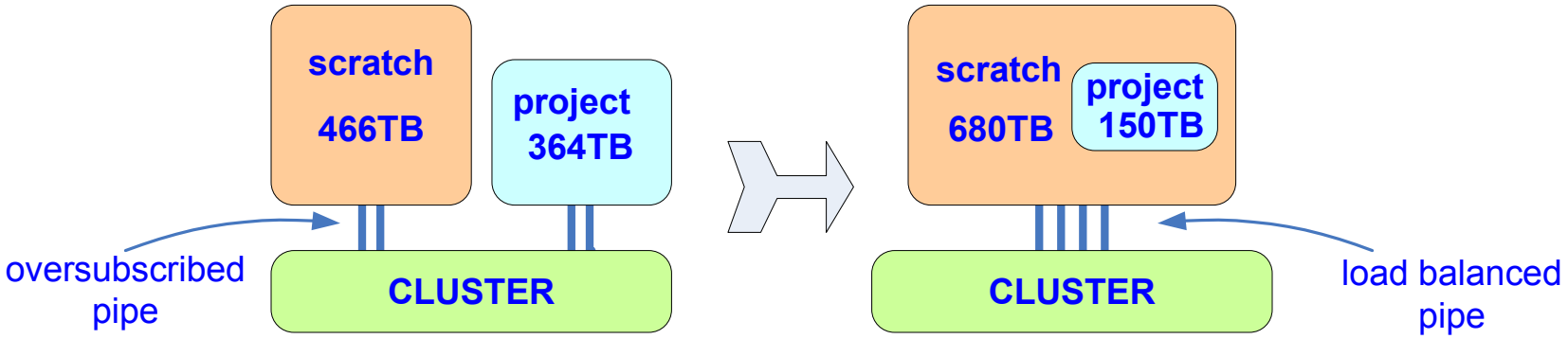
HPSS (broad use of the term) =
nodes + disks + network + FC + HPSS + DB2 + HSI + HTAR + ISH + Library + tapes + services

HPSS – scaling potential

Possible Phase II (TBD)



Merger of scratch and project (in 1 month)



Objectives:

* Migrate user's data from project to HPSS
* Increase size and performance of scratch

Breakdown:

830TB total available on the new filesystem

(4 controllers, load balanced)

scratch 680TB , reorganized on a per group hierarchy

project 150TB , same mounting point as before

Transition Plan

* "freeze" /project (i.e. make it read-only) for a period of roughly 1-2 weeks
* temporarily back-up the /project data to two completely separate tape systems
* /project disks will then be reconfigured during scheduled system downtime
* groups with allocations of 5 TB or less will see no difference
* groups with > 5TB allocations will find that they have an empty /project with 5TB of available space and all their former /project data will live in HPSS
* accounts will be relocated inside /home, /scratch, /project
* users should start to adapt scripts to use env. variables:
  - $HOME
  - $GROUP
  - $SCRATCH
  - $PROJECT

Policies

scratch (same as before)

- 90 days purgeable (for now)
- 20TB/user but max 80TB/group (quota enforced)
- 1M/user and 10M/group file limit

project allocated by RACs

- non-purgeable
- 5TB max/group, 1M file limit (quota enforced)
- excess must be migrated to HPSS

hpss allocated by RACs

- quota enforced
- non-deleted by SciNet
- beyond RAC: users can buy tape for one-time cost of ~$120/TB/copy

RAC Applications

*PIs can request "dedicated" (never purged) storage
for their groups for 2012
* up to 5 TB (per group) will be allocated on disk (/project),
remainder on HPSS
* project space which is not reallocated in the next RAC call is
migrated to HPSS and then deleted from disk

RAC Q&A will be scheduled for early October