



HPSS Status at CEA/DAM

(site of Bruyères le Châtel)

Philippe DENIEL
philippe.deniel@cea.fr

Who are we ?



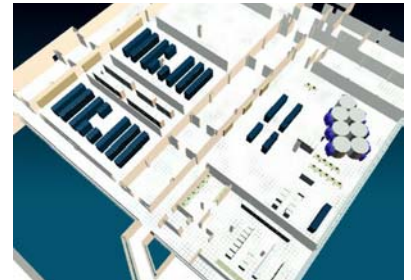
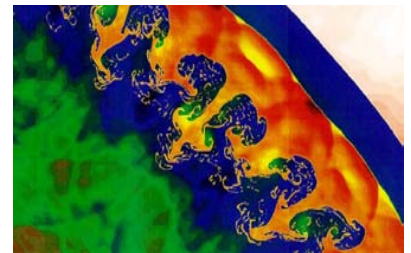
- CEA means "Commissariat à l'Energie Atomique"
- CEA is a research institute which handles much of the scientific research in the nuclear domain
- Within CEA, DAM (Direction des Applications Militaires) is a subdivision of CEA focused on military applications
- There are 4 CEA/DAM's sites in France,
 - DAM's Compute Center is located in Bruyères Le Châtel, south of Paris



The TERA project: overview



- Large compute power for High Performance computing
- Production codes will simulate physics within complex and critical systems
- Cluster of SMP with High-Speed network
- High Performance storage using the HPSS software with 2 tape storage classes
 - 1 PB for level 1
 - 5PB for level 2



The TERA architecture: the compute machine



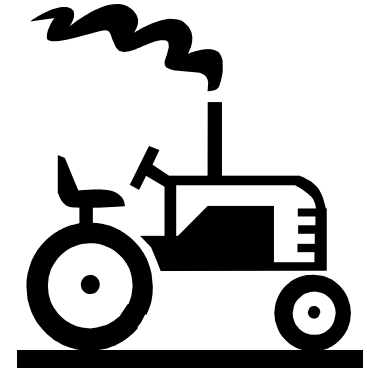
- Fully installed in summer 2002
- The compute machine:
 - SMP Cluster
 - 170 racks
 - 90 km of cables
 - 640 nodes with 4 processors each
 - Quadrix switches for interconnection
 - 50 TB disks (7.5 GB/s)
 - First run above 1 sustained Tflops : December 12, 2001
 - Linpack at 3.98 Tflops : April 12, 2002
- In full production since January 2003
- Mostly run parallel compute program



The TERA project: HPSS System



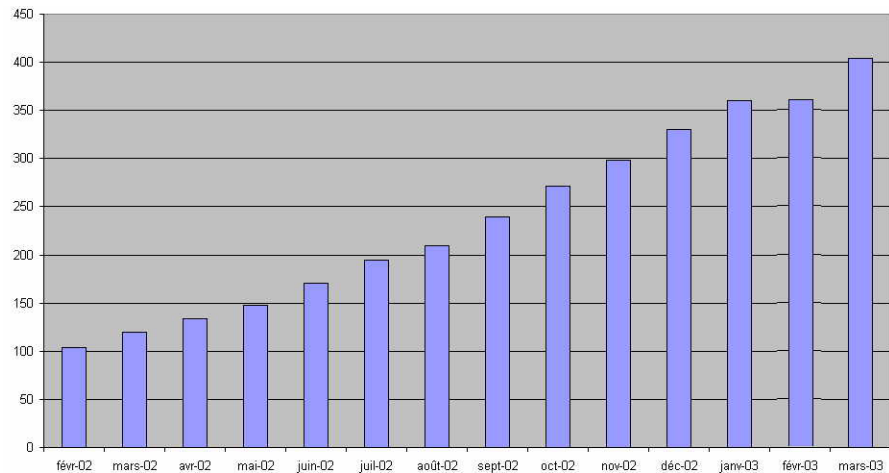
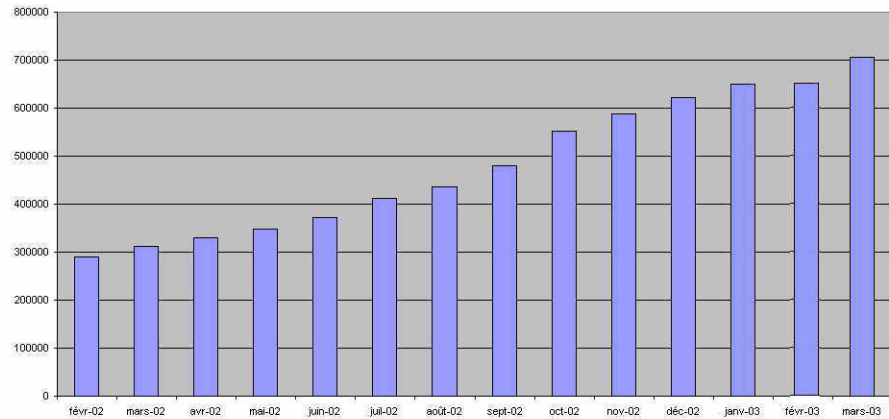
- HPSS Production system:
 - 1 WinterHawk Node: core server
 - 4 procs / 2 GB
 - 5 NightHawk I Nodes: Disk/Tapes Movers
 - 5 TB of disks (RAID 5+1)
 - 6 HiPPI adapters / node
 - 8 HBAs for SAN access to tapes drive / node
 - 8 procs / 4GB RAM
 - Interconnection via Colony Switch
 - 1 Sun Ultra 2: DCE CDS Server
 - 5 PowderHorn Silos with 20 x 9840B (to be upgraded to 7 silos and 48 x 9840C in 4H/2003))
 - 1 PB = 25000 x 40GB tapes (on STK 9840C technology) as first tape level (under deployment)
- System was upgraded to HPSS 4.3.0.2 in 1Q/2003
 - we are very satisfied with the stability of this version of HPSS
 - Very few local mods (located in the NFS daemon to prevent NFS direct access to data; NFS is used only for metadata management)



- A new compute center has to be deployed
 - The compute machine will be 10 times our production system in 4Q 2005
 - The HPSS server will have to cope with incoming/outgoing rates 10 times what they are now (both data and metadata)

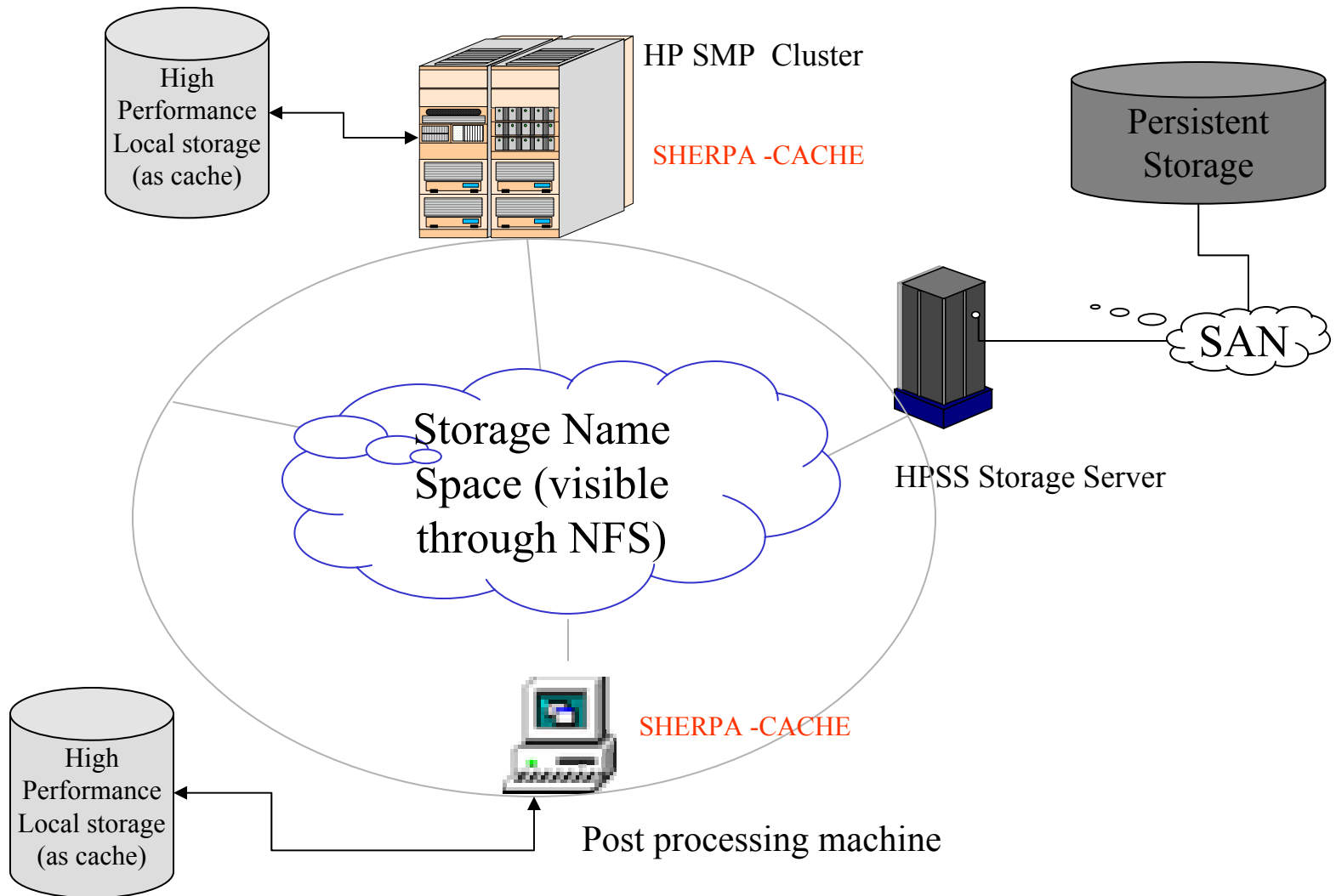


Numbers of files / volume stored within HPSS

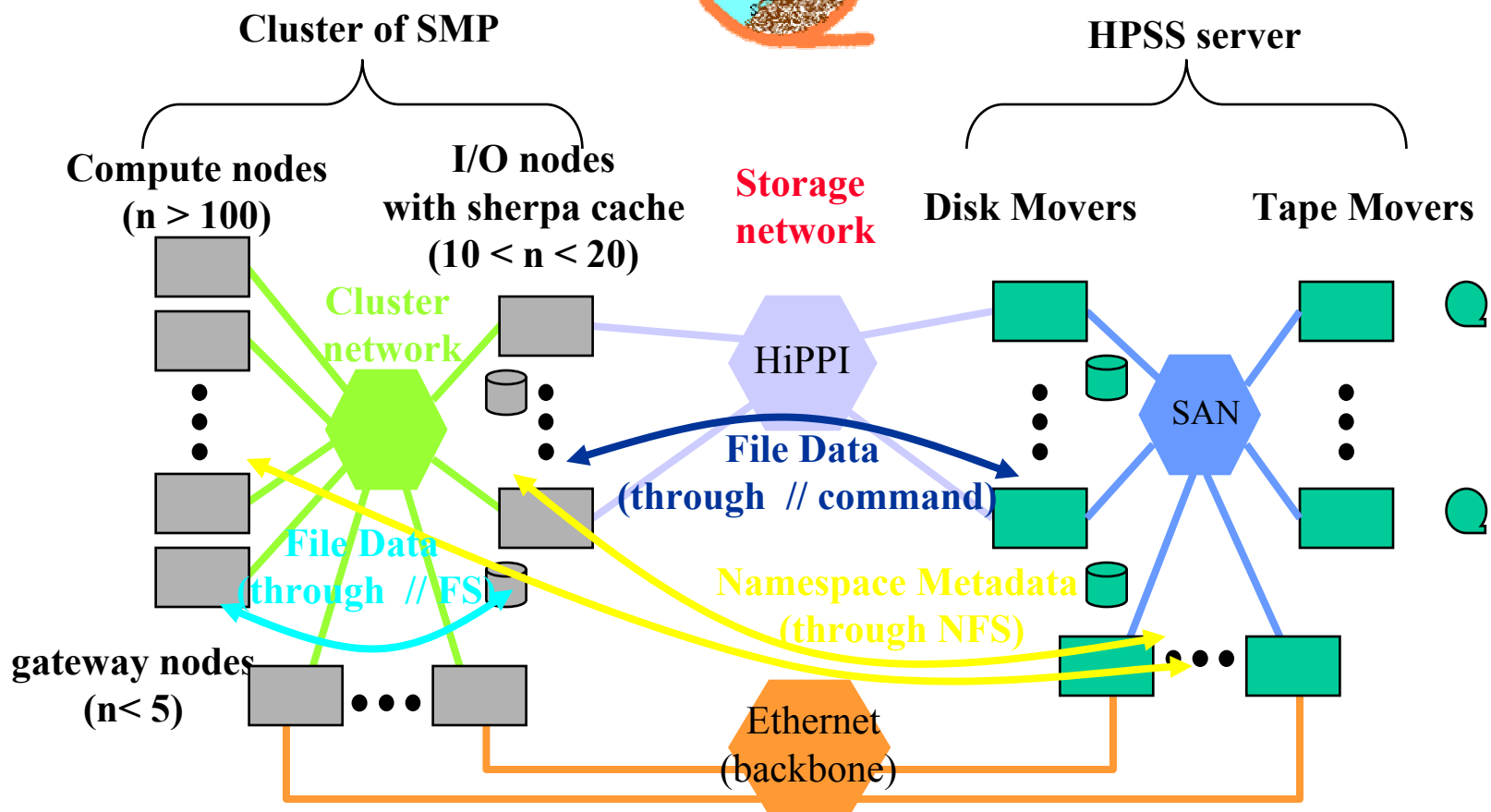


File Size is a quasi constant with CEA's production code

The TERA project: A Storage centric architecture



SHERPA Mechanism



The need for NFS

- The SHERPA-Cache uses NFS as protocol for metadata manipulation and consultation
- NFS is widespread, it provides us with a portable and generic interface to access HPSS Namespace, that is fully integrated in the OS. This makes our caching device easily portable and deployable to lots of OS
- Users just perform basic Unix commands to view the state of their files (cd, ls, ...) without having to learn how to use a new tool
- NFS is supported by every vendor on both client and server side.
- CEA did the development of NFS V3 in the HPSS/NFS daemon and we are very familiar with this protocol



- NFS V3 is very close to NFS V2. Mostly the 64 bits semantic and an enhancement management of client caching capability were added in V3
- NFS V4 is very different from this and is really a brand new protocol
- It provides enhanced security and scalability by making the communications between client and server much more compact. Client caching capability becomes naturally one of the keynote for NFS V4
- NFS V2 and V3 were very close to unix file systems semantic and behavior
- NFS V4 is designed to fit every file system, some features will provide lots of improvement when using with an HSM
- NFS V3 will not fit the increase of production we expect in the next three years
- NFS V4 is a natural evolution of HPSS/NFS to our opinion
- I would be very pleased to discuss with you about this topic. Do not hesitate to start the discussion

- Extend the support of NFS to NFS V4
- Generalize the capability of HPSS to communicate with files systems which can use DMAPI
- Crash Can ? (kind of implementation of « soft delete »)
- Enhancing the capability to control every incoming and outgoing stream from/to the HPSS system, in order to protect the system from being overflowed when a client application goes mad





- All of HPSS running on Linux Boxes
- Support of disks with PBs of capacity
- Support of 10Gbits Ethernet or InfiniBand (native SDP support ?)
- Disk and Tape Movers should be able to mutualize storage
Devices and to IO load balancing



Questions ?

