



HPSS Status at CEA/DAM

(site of Bruyères le Châtel)

Philippe DENIEL
philippe.deniel@cea.fr



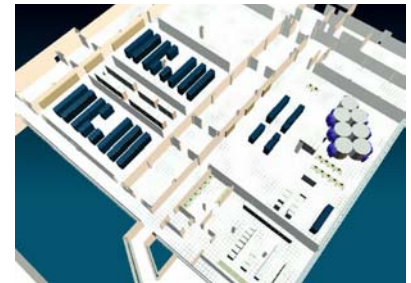
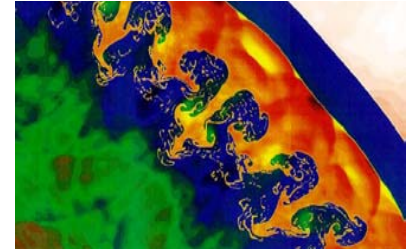
- CEA means "Commissariat à l'Energie Atomique"
- CEA is a research institute which handles much of the scientific research in the nuclear domain
- Within CEA, DAM (Division des Applications Militaires) is a subdivision of CEA focused on military applications
- There are 4 CEA/DAM's sites in France,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



CEA 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

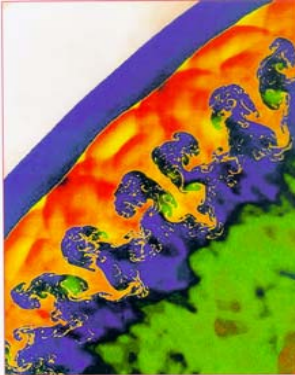


CEA The TERA project : milestones

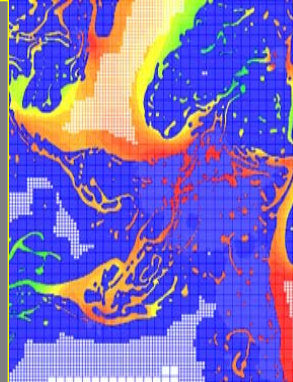


2D
computations

< 10^6 cells

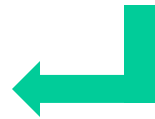
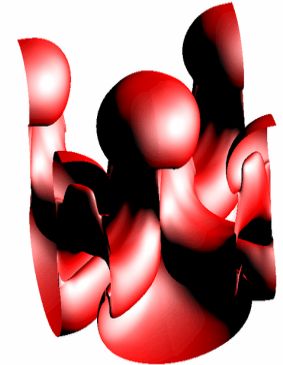


Ultra-fine
2D
computations
3D
validations
 10^7 cells



3D
computations

> 10^8 cells



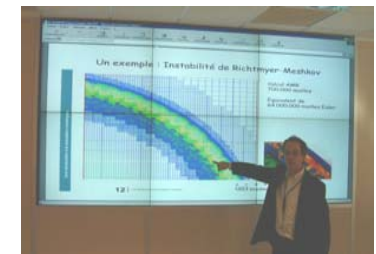
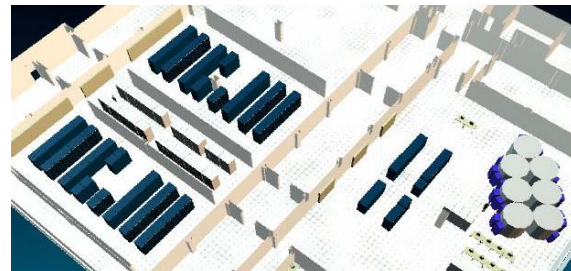
2005
10 Teraflops sustained
30/50 Teraflops peak



2009
100 Teraflops sustained
300/400 Teraflops peak



The largest
scientific computing
complex in Europe





- Installation started in October
- Full system delivered in beginning of December :
 - 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
- Largest compute machine in Europe

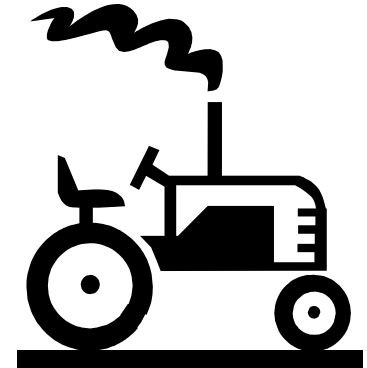




The TERA project: HPSS Hardware



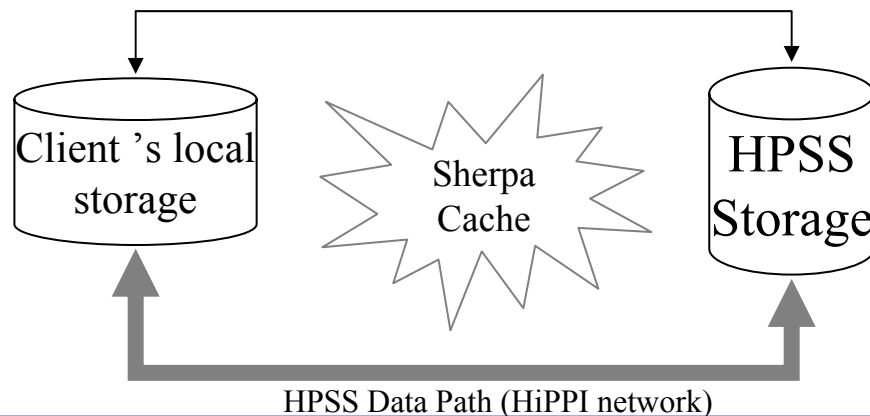
- HPSS Production system:
 - 1 WinterHawk Node: core server
 - 4 procs / 2 GB
 - 5 NightHawk I Nodes: Disk/Tapes Movers
 - 5 TB of disks (RAID 5+1)
 - 3 HiPPI adapters / node
 - LP9000 HBA for SAN access to tapes drive / node
 - 8 procs / 4GB RAM
 - Interconnection via Colony Switch
 - 1 Sun Ultra 2: DCE CDS Server
 - 5 PowderHorn Silos with 20 x 9840B (to be upgraded to 7 silos and 48 x 9840C)
 - 1 PB = 25000 x 40GB tapes (on STK 9840C technology) as first tape level (under deployment)
 - There will be soon an RFP for level 2





- Files exist and are kept within the HPSS namespace, their metadata are manipulated via NFS : taking benefits of HPSS very large namespace
- Files are accessed locally through a caching device: taking benefits of the Cluster's file system feature (High Performance, quick MPIIO support)
- Caching device uses optimized transfer tool for quick access.
 - NFS used as metadata/control path for the caching device
 - local transfer tool for data path
 - could be used with any HSM with NFS support (was previously used as frontend to DMF)
- Read and write are prohibited via NFS by a local modification in the NFS Server (NFS_READ and NFS_WRITE always returns EDQUOT).
- Upgrade to HPSS 4302 will provide NFS V3 support for files larger than 4GB management.

HPSS Control path and NFS metadata consultation (compute center backbone)





- Test system is running 4.3.0.2
- Production system is running 4.1.1.1
- An upgrade of the production system to 4.3.0.2 is scheduled for mid-term range
- Test will be made with very heterogenous HPSS systems (on Sun, SGI, Compaq..)

Daily production

- Intensive compute results production.
- Intensive NFS traffic (inode only)
- Unitary performances around 39 GB/s between HPSS and Cluster's local file system, through HiPPI network
- These numbers are with the TERA intermediate system (10 times smaller than the final one)



- Extend the support of NFS within HPSS (NFS V4 ?)
- Generalize the capability of HPSS to communicate with files systems which can use DMAPI
- Make use of native fiber channel protocol available as data path in the mover protocol (for data migration and access to file from non-HPSS agents which can use the mover protocol)
- Crash Can ? (kind of implementation of « soft delete »)





- **hpss_rcp: rcp oriented transfer tool**
 - uses ONC RPC and POSIX threads: Client can be compiled on any Unix system, hpss_rcpd daemon runs on Core Server and uses CLAPI
 - hpss_rcp_slave daemons run on clusters and uses mover protocol to access HPSS
 - client takes consideration of HPSS volumes striping
- **Token Server**
 - used to limitate simultaneous accesses to HPSS systems (to avoid system overload): kind od « network-wide semaphore »
 - give full control on the incoming storage request flow
- **Casimir: Generic Error Monitoring Program**





- Objective: to automatize basic emergency system actions and to have a central system monitoring tool
- Casimir is fully written in PERL. It is so fully portable
- For each resource to be watched, a specialized tool (generally a script) generates a log that indicates the resource's state
- Each log is in a standardized format
 - `04/05/2001 10:53:05 : hpss_rcpd-45536 :
hpss_ReadList status is HPSS_EIO`
- The casimir program gathers all the log and search for "pathological" regular expressions. The RegExp found can be combine in both boolean and chronological way to define "events".
- An event is symptomatic of a critical situation, an action is associated with it and executed (genrally a launch of an external script)
- Casimir is used to watch the HPSS system (via modified whpss) , but also the clusters and the distributed systems (server and workstations) at CEA
- Casimir is used in addition to remote syslog protocol to centralize and act on local and remote errors.



