SLURM at CEA

Matthieu Hautreux
(CEA/DAM/DIF)
matthieu.hautreux@cea.fr

Outline

- CEA Computing complex
- Focus on TERA-100
- Using SLURM on TERA-100

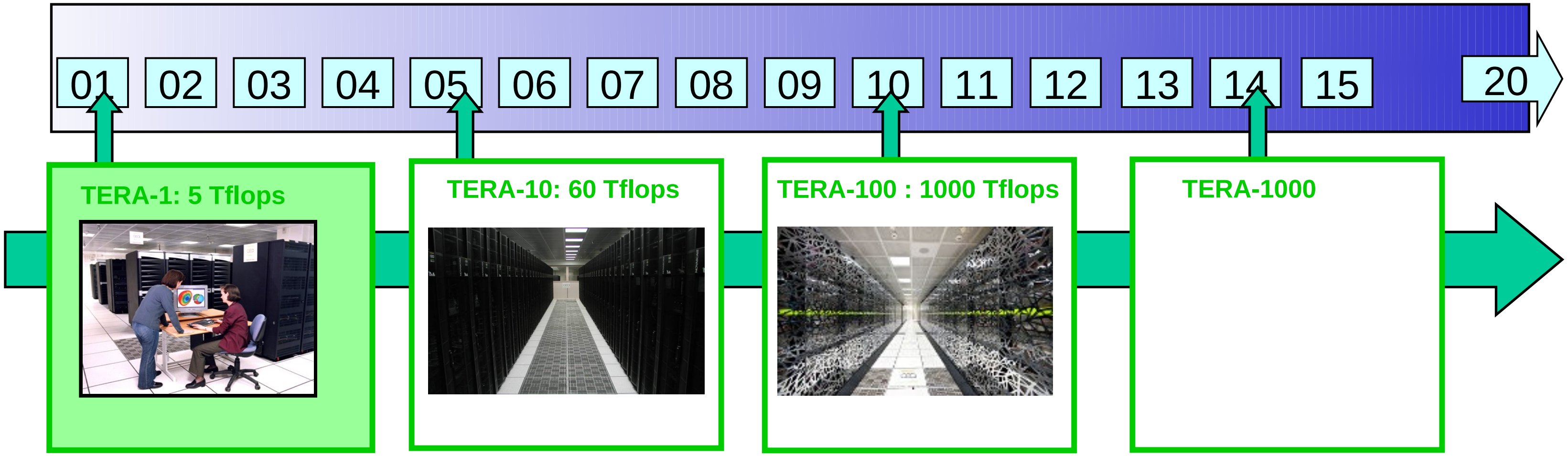
CEA Computing complex

Location

- CEA/DAM/DIF
  - Paris Area division of CEA defense pole
  - Bruyères-le-chatel (30km south of Paris)
  - Involved in 3 major HPC projects

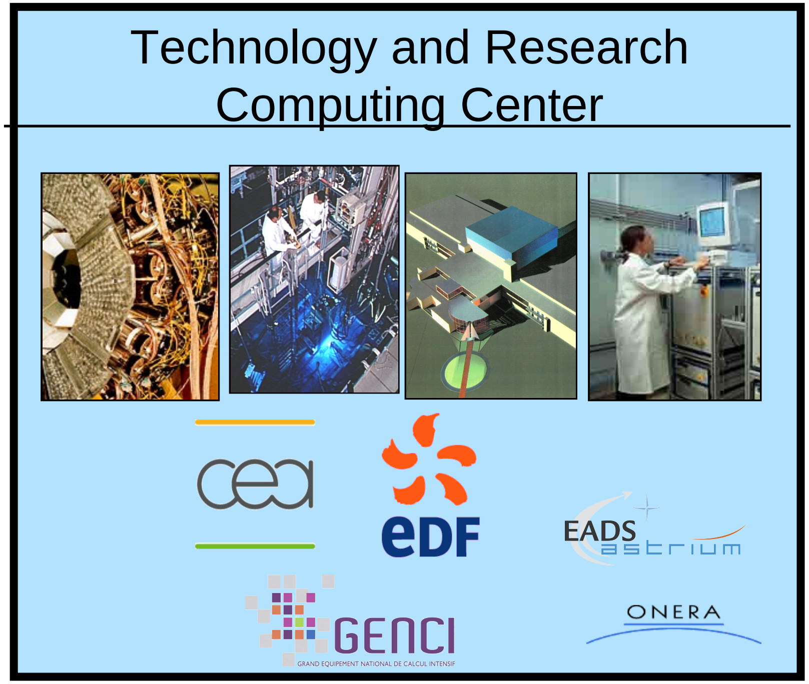
HPC Projects

- TERA
  - Defense computing center
  - Part of the Simulation project for French Nuclear Deterence
  - Project started in 1998



HPC Projects

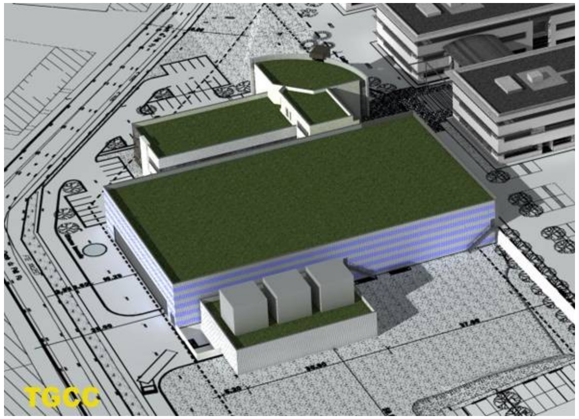
- CCRT
  - French Industrial and research partners shared computing center
  - Hosted at CEA/DAM/DIF since 2003


Technology and Research Computing Center

HPC Projects

energie atomique • energies alternatives

- PRACE (PaRtnership for Advanced Computing in Europe)
  - PRACE European project shared computing resources
  - New facility, TGCC, delivered October 4th 2010
  - Initial PRACE system to be deployed by end of 2010
  - Larger system to be deployed in 2011



PRACE

GENCI
GRAND EQUIPEMENT NATIONAL POUR LE CALCUL INTENSIF

energie atomique • energies alternatives

Focus on TERA-100

TERA-100 Objectives

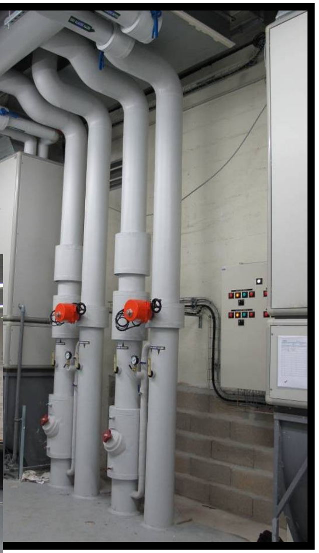
- Increase by ~20 TERA-10 computing power
  - Petaflopic cluster
- Keep Tera project macro-architecture
  - General purpose SMP cluster
    - ☛ One single cluster build with identical components
  - Supporting various programming model
    - ☛ MPI, OpenMP, Threads, CEA MPC
  - Supporting heterogeneous production workload
    - ☛ Daily CEA workload, Large computational challenges
  - Large sustained IO performances
  - Infrastructures constraints
    - ☛ Power consumption < ~ 5MW , Footprint < 750 m2

TERA-100 Objectives

- Planning
  - First prototype for CEA/DAM applications migration
    - ☛ Shipped mid-2009 (432 compute nodes, ~40 Tflops)
  - TERA-100 installation
    - ☛ Begins Q2-2010
  - TERA-100 CEA/DAM applications validation
    - ☛ End of 2010 / Begin of 2011

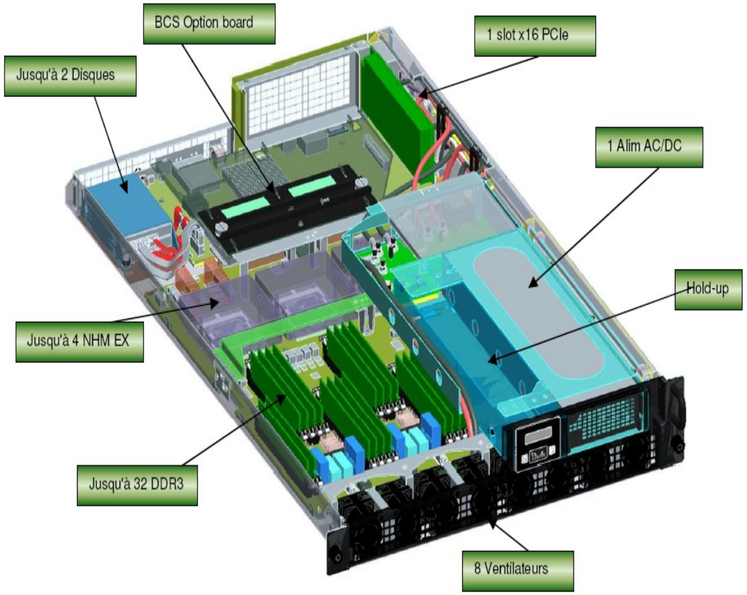
TERA-100 Hardware specificities

- Water cooled Racks
  - Up to 40 kW / rack



TERA-100 Hardware specificities

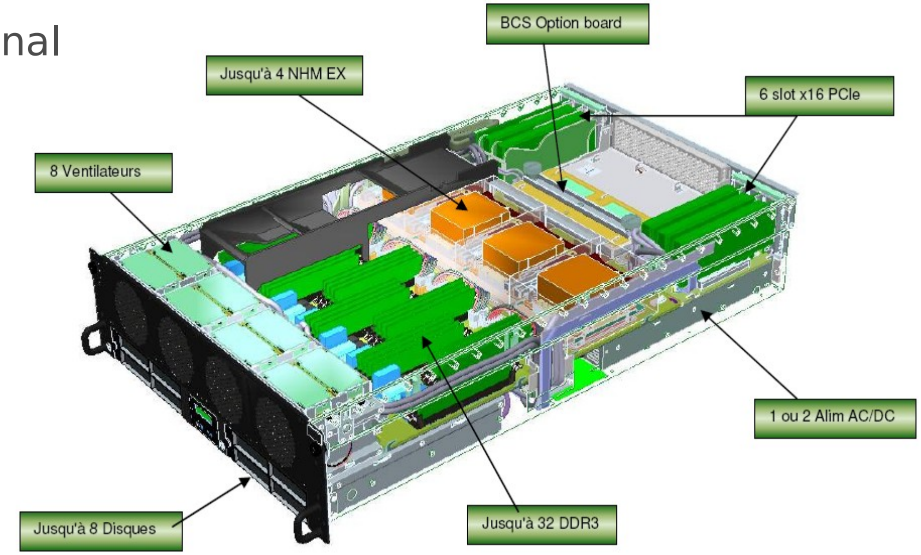
- Compute node
  - Bull Server MESCA* S6010 (1,5U)
  - 4 sockets 8 cores Nehalem EX 2.27 GHz : 290 Gflops
  - 2 or 4 GB/core = 64/128 GB
  - 1 port Infiniband ConnectX 4X QDR (40 Gb/s)
  - 1 port Gb ethernet
  - 1 or 2 SATA or SSD disks
  - 1 ultracapa
    - ☛ power dropout prevention



* Multiple Environment on SCalable Architecture

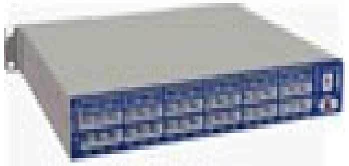
TERA-100 Hardware specificities

- Service Nodes (IO, Management, ...)
  - Bull Server MESCA S6030 (3U)
  - 4 sockets 8 cores Nehalem EX 2.27 GHz : 290 Gflops
  - 2 GB/core
  - 2 ports Infiniband ConnectX 4X QDR (40 Gb/s)
  - 2 ports Gb ethernet
  - 2+ SATA disks
  - 2 PCI-E 16X slots, 4 PCI-E 8X slots
    - ☛ For FC, 10 GE or additional IB connectivity



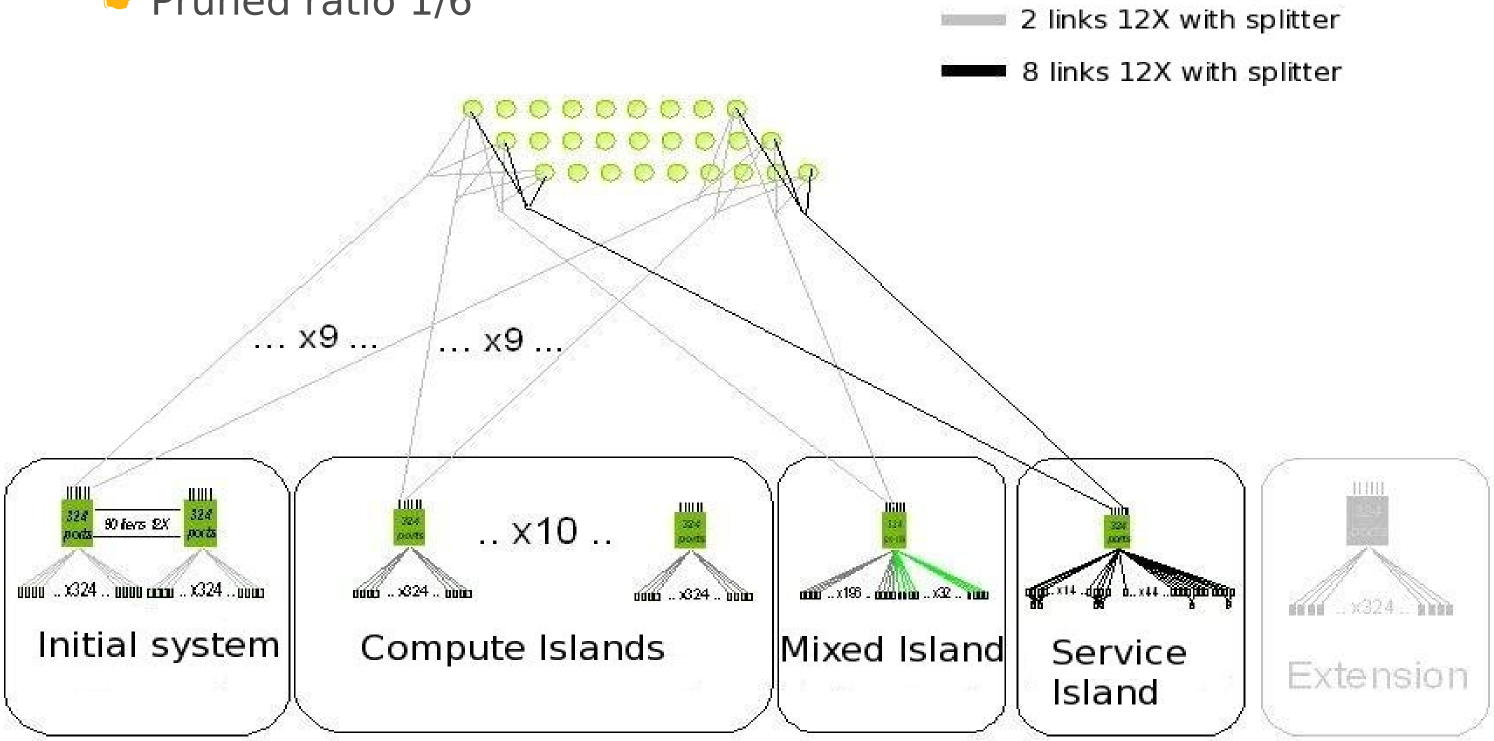
TERA-100 Hardware specificities

- Infiniband interconnect
  - Voltaire Grid Director 4700
    - ☛ 324 QDR (40 GB/s) ports (19U switch)
    - ☛ Ultra-low latency : 100-300 ns port-to-port
    - ☛ 51.8 Tbps non-blocking bandwidth
  - Voltaire Grid Director 4036
    - ☛ 36 QDR ports (1U switch)
    - ☛ 2.88 Tbps switching capacity
  - A bunch of fiber and copper cables



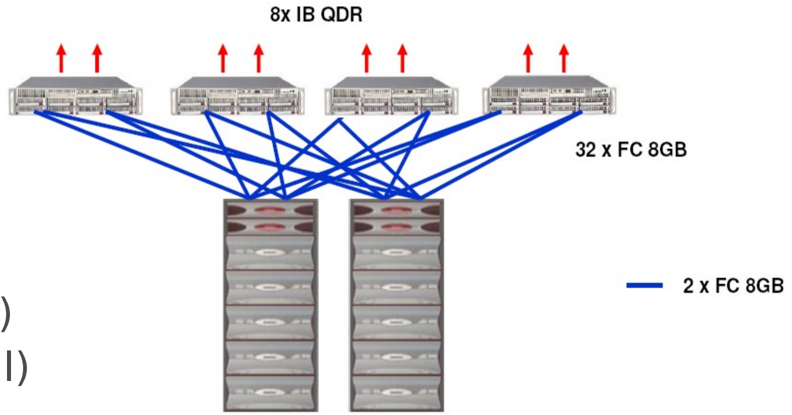
TERA-100 Hardware specificities

- Interconnect topology
  - Islands of nodes connected in fat tree
    - Up to 324 nodes per island using Voltaire Grid Director 4700
  - Cluster of islands building a pruned tree
    - 27 Voltaire Grid Director 4036
    - Pruned ratio 1/6



TERA-100 Hardware specificities

- Private Storage
  - 68 IO nodes (S6030)
    - Using Lustre 2.0 FS
    - 1 MDS IO cell (4 nodes per cell)
    - 16 OSS IO cell (4 nodes per cell)
    - Managed using Shine (Bull/CEA open source project)

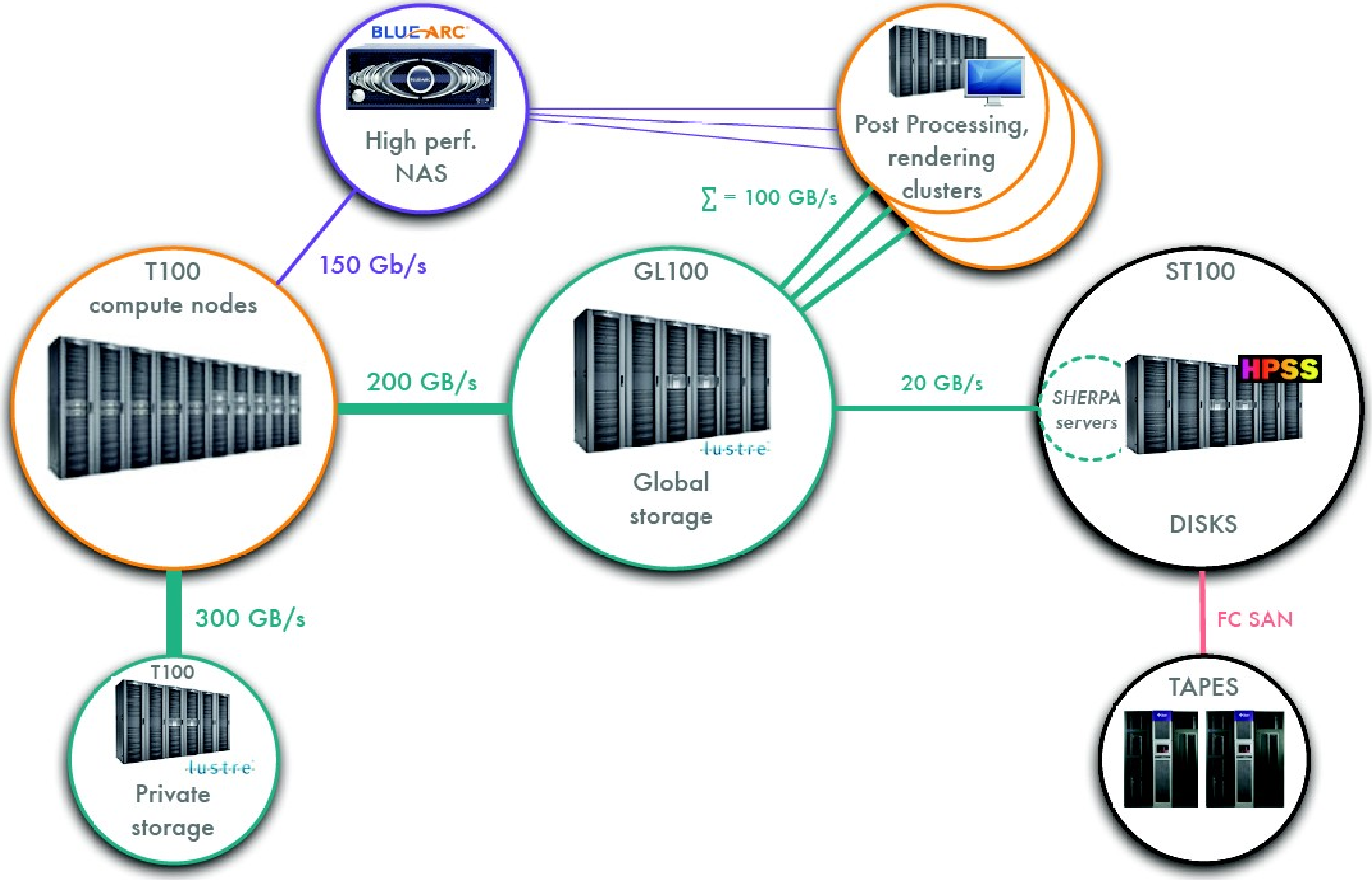


  - Data Direct Network™ SFA 10K backend
    - Metadata : 1 SFA 10K for a total of 11 TB
    - Data : 32 SFA 10K for a total of 9 PB

TERA-100 Software specificities

- BULL XBAS Linux distribution (based on RHEL6)
  - Kernel improvements (clock sync, noise reduction)
- BULL Petaflopic Cluster Management Tool
  - Deployment, Power management, Monitoring, ...
- OFED 1.5 Infiniband stack
  - With BULL contributions (OpenSM, diagnostic tools,..)
- BULL MPI stack (OpenMPI based)
  - Optimized for Petaflopic and production cluster
- Lustre 2.0 Parallel FS
  - Managed using Shine (BULL/CEA open source project) http://sourceforge.net/projects/lustre-shine/

TERA-100 Ecosystem



TERA-100 Some figures

- Peak performance : 1,25 Pflops
- Global Memory : 291 TB
- Private storage capacity : 8,64 PB
- Aggregated IO bandwidth : 300 GB/s
- Storage network bandwidth : 200 GB/s
- Backbone network bandwidth : 150 Gb/s

Using SLURM on TERA-100

TERA-10 feedback

- TERA-10 batch environment
  - In-house LSF/RMS (Platform/Quadrics) hybrid approach
    - LSF for batch submission
    - RMS for efficient parallel execution
    - Allocation at core level (10K cores) using RMS
    - 2 schedulers, hard to be deterministic
  - In-house Metascheduler
    - Automatic fairshare scheduler with long term provisioning
    - End User workflow oriented GUI
- TERA-10 Post-processing environment
  - Dedicated clusters
  - Access data produced by TERA-10
  - First usage of slurm at CEA
    - Starting in 2005
    - Interactive usage only
    - Allocation at node level

TERA-100 R&D phase

- Evaluation of promising solutions
  - Launched after Tera-10 installation
  - Both Hardware and software aspects
- New batch environment research
  - Simplify scheduling logic with large number of cores
  - Move to open source software to understand/adapt when necessary
  - Comply with CEA production requirements

TERA-100 R&D phase

- SLURM elected the best candidate
  - Good performances and scalability
  - Already known by CEA sysadmins
  - High modularity (plugins, SPANK framework)
  - Good community support
  - Some gaps but nothing unmanageable

- SLURM study beginning (2008)
  - Starting with slurm-1.2
  - Identify ways of improvements
  - Discuss evolutions and roadmaps with LLNL
    - ☛ Core level allocation and binding
  - Start developments and patches sharing
    - ☛ HA enhancement, Cpusets, Kerberos support, ...

TERA-100 R&D phase

- SLURM 2.x study
  - CEA patches proposals on specific aspects
    - ☛ Gently integrated or modified by Moe and Danny
  - Joint study with BULL on other aspects
    - ☛ Part of the TERA-100 contract
    - ☛ To comply with CEA expressed requirements
    - ☛ To comply with BULL Cluster Management solution
    - ☛ Main objective : reduce official release drift
  - BULL 2.2.0 flavor as the target for TERA-100
    - ☛ Complete core/memory level allocation for jobs and job steps
    - ☛ Tree topology support with fragmentation reduction
    - ☛ Tree topology awareness for MPI layer performance
    - ☛ Linux cgroups for process tracking, confinement and tasks affinity
    - ☛ BULL additional logic for tight integration in their petaflopic solution

TERA-100 SLURM status

- Current configuration
  - BULL packaging of slurm-2.2.0.pre9
  - CEA/LLNL additional patches (from pre10)
  - Consumable resources selection algorithm
    - ☛ Topology/tree plugin
    - ☛ Best-fit selection of switches
    - ☛ Best-fit selection of nodes
    - ☛ Block distribution of cores across nodes (fragmentation optimization)
    - ☛ Tasks topology address tagging for MPI optimization
  - Core/Memory level allocation
    - ☛ Using a block distribution by default
    - ☛ With best-fit selection across sockets
    - ☛ HyperThreading disabled (by choice, interest still in evaluation)
  - Scheduler
    - ☛ With backfilling
    - ☛ Multiple partitions sharing the same resources with different limits

TERA-100 SLURM status

- Current configuration
  - Process tracking using cgroup
    - Freezer subsystem for atomic suspend/resume
  - Resources confinement using cgroup
    - Only cores for now
    - Memory confinement with cgroup not mature when tested
  - Tasks binding using cgroup
    - Cpusets subsystem
  - Slurmdbd
    - In HA with a MySQL DB backend
    - For limits and account enforcement
    - For accounting and in-house Metascheduler feeding
  - Sview
    - For day-to-day production usage (drain, resume, cancel, ...)

TERA-100 SLURM status

- Current extensions
  - Setsched SPANK plugin
    - Allow on demand alternative Kernel scheduler selection
    - Used to automatically leverage BULL noise reduction primitives
    - CEA contribution to slurm-spank-plugins project
      http://code.google.com/p/slurm-spank-plugins/
  - X11 remote display SPANK plugin
    - Allow X11 display access in SLURM jobs (both batch and interactive)
    - Based on OpenSSH X11 tunneling (requires Single Sign On)
    - CEA in-house development

TERA-100 SLURM status

- Current extensions
  - AUKS SPANK plugin
    - ☛ Provide Kerberos credential support (forwarding and renewal)
    - ☛ Based on and part of AUKS (CEA open source project) http://sourceforge.net/projects/auks/
  - Bridge
    - ☛ CEA in-house development
    - ☛ Abstraction layer on top of batch system / resource managers
    - ☛ Reduce user vision of underlying systems
    - ☛ Ease systems migration and heterogeneous clusters usage

Thank you for your attention