Lustre at CEA: TERA-100

Stéphane Thiell, CEA – stephane.thiell@cea.fr

Contents

- **TERA-100**
  - Project Overview
  - Cluster Nodes
  - Cluster Architecture
  - Computing Center Integration
- **Lustre in TERA-100**
  - Storage
  - Servers
  - Network
  - Strategy
- **Shine**
  - TERA-100 and Shine
  - Project Status
  - Roadmap

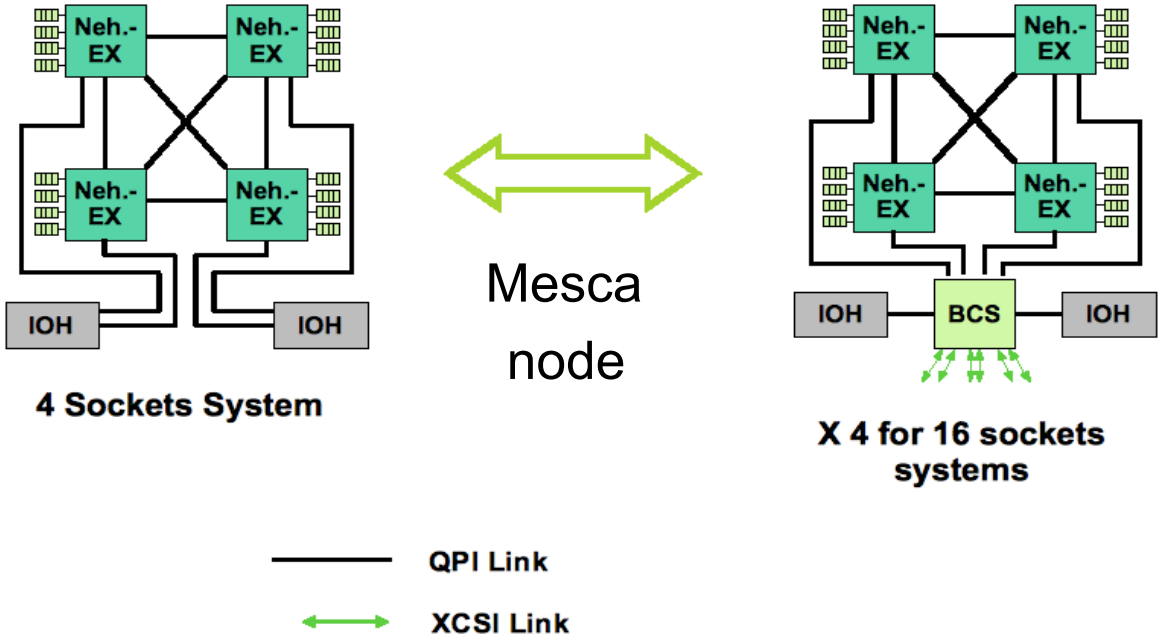
TERA-100 Project Overview

- **1 Pflop system**
  - Xeon based (100+ Kcores)
- **A larger cluster**
  - Thousands of nodes
  - Increase of storage needs
- **A new computing center architecture**
  - Data-centric architecture
- **Keep control on TCO**
- **System delivered in 2010**

TERA-100 Nodes

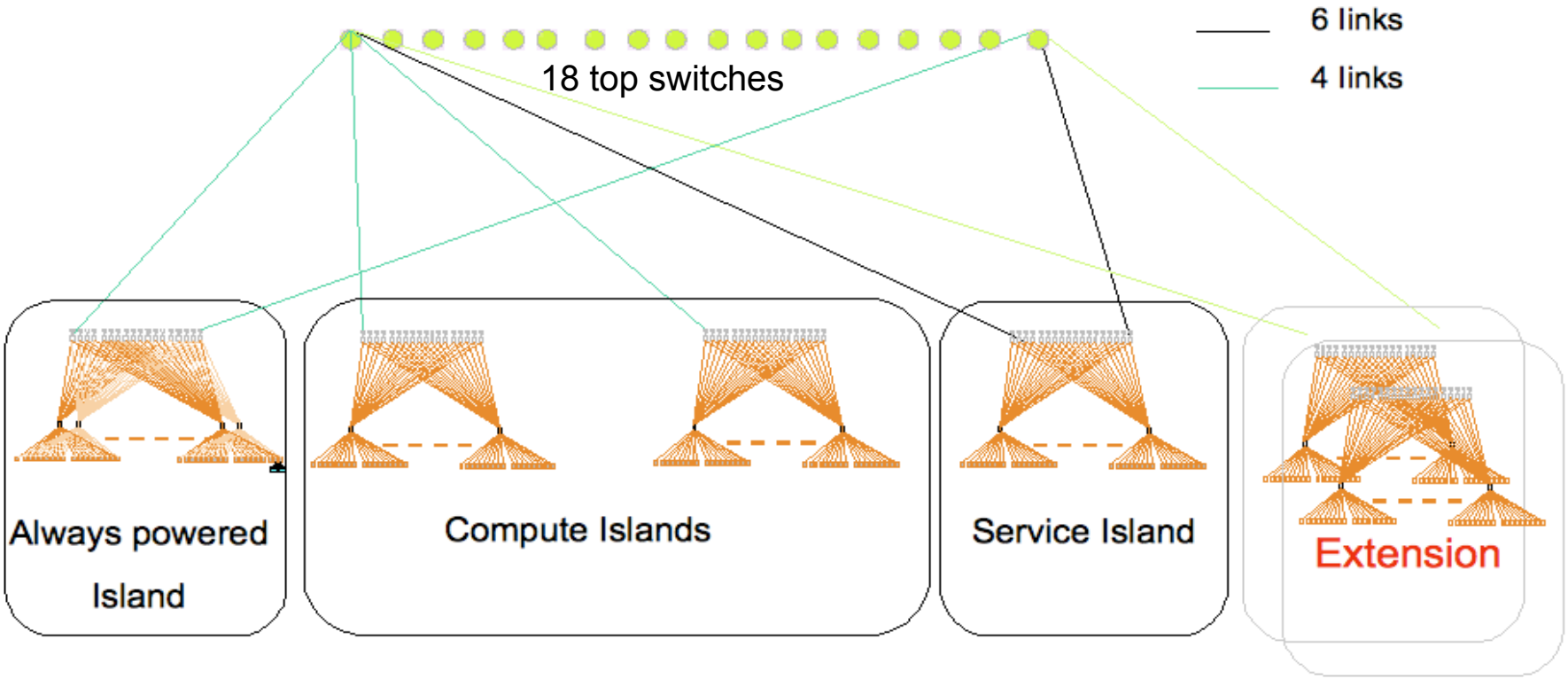
- **Most of the nodes are 4 processors nodes**
  - Optimized packaging (2 servers in 3U)

- **Some nodes are large SMP**
  - Nodes for multi-threaded applications and MPI computations
  - Development of a node of 4 to 16 sockets Nehalem-EX (Xeon)



TERA-100 Cluster Architecture

- **MPI and Lustre network**
  - Fat tree does not scale to petaflop system (complexity, cost)
  - Study island topology based on IB QDR
  - Adapt software stack to bring topology knowledge (MPI, ressource manager)



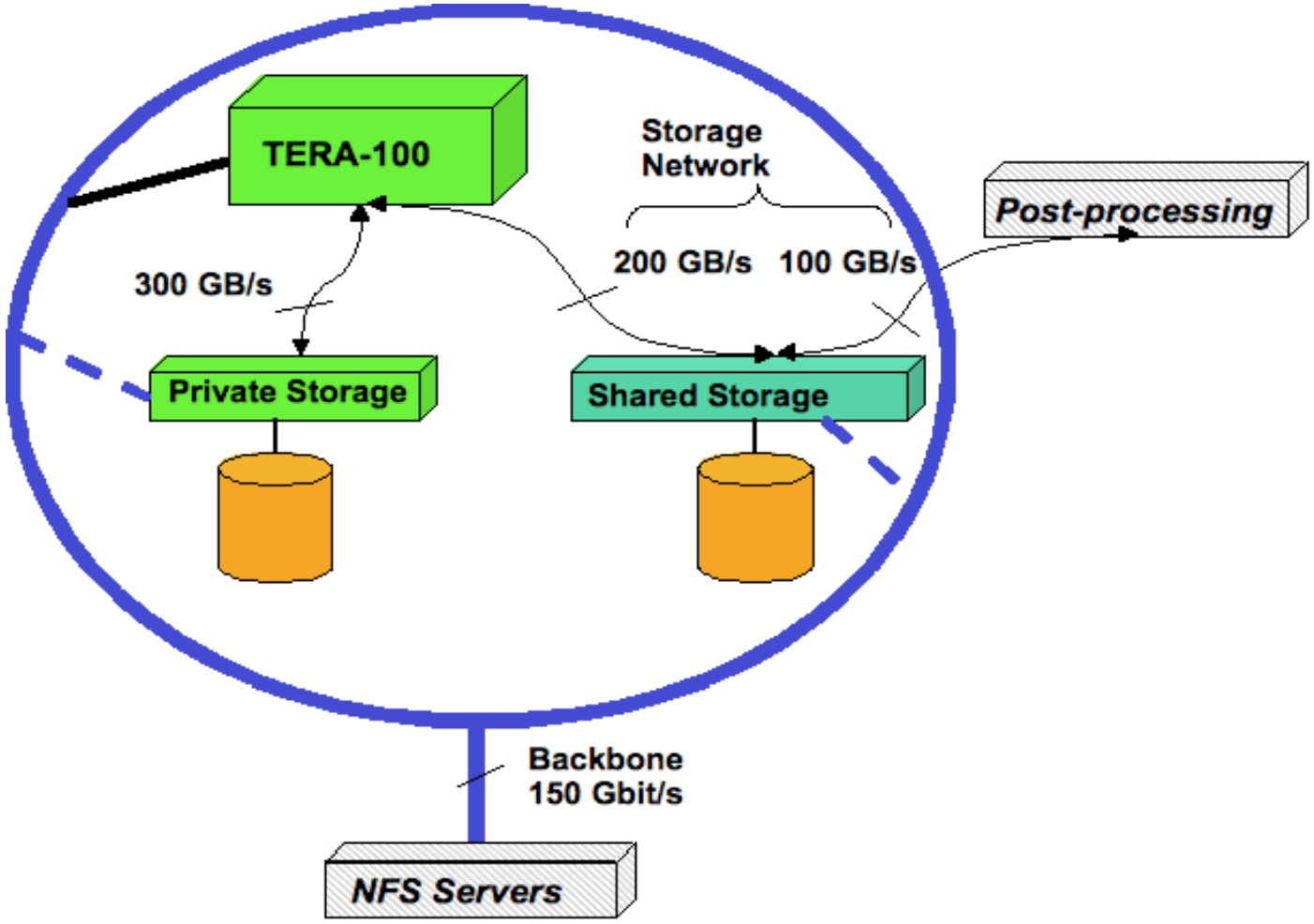
TERA-100 Computing Center Integration

- **Goal**
  - All (= thousands of) client nodes use computing center shared resources (Lustre + NFS filers)
- **Studies**
  - Mechanisms to control flow from compute cluster
  - Lustre routers
  - Services proxy (NFS, …)

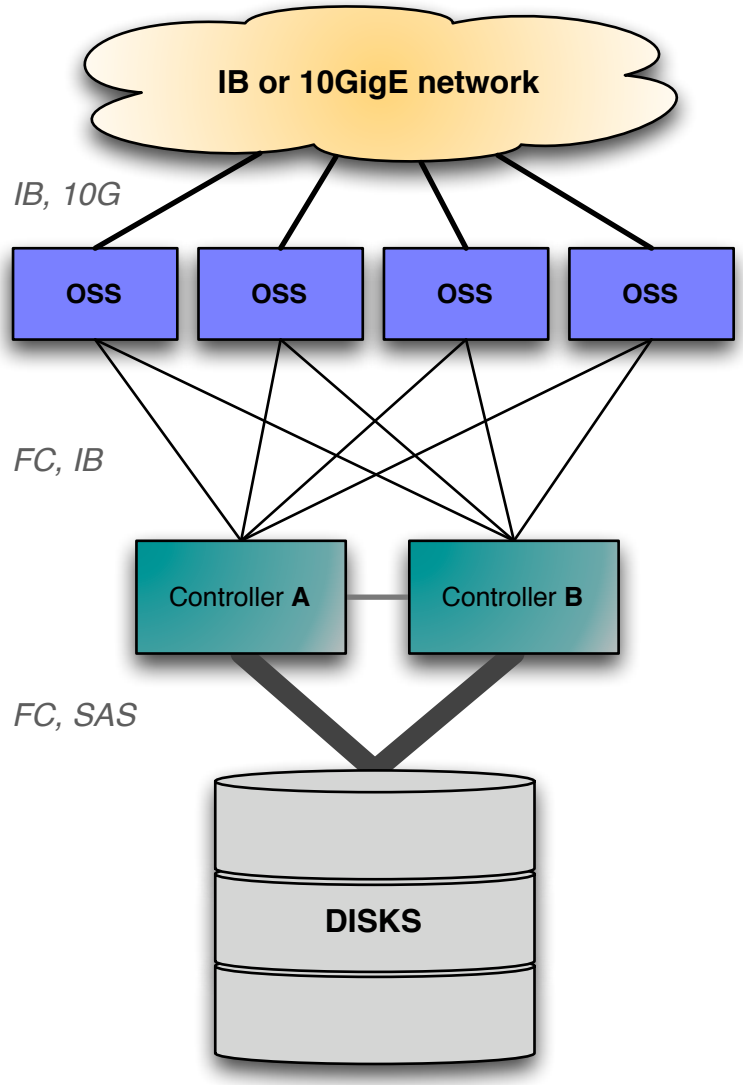
Lustre in TERA-100

TERA-100 Lustre Storage

- **Goal: 300 GB/s internally and 200 GB/s externally**



TERA-100 Lustre Servers



- **Nodes**
  - Mesca in compute cluster
  - Thin nodes (EP based) for storage cluster ?
- **4 nodes HA architecture**

TERA-100 Storage Network

- **2 choices for 200 GB/s**
  - InfiniBand
    - QDR or DDR ?
    - Large switches or 36 ports switches ?
  - 10 GigE
    - With iWARP NICS (Chelsio) ?
    - First LNET tests with iWARP are promising
    - L2 switches (fully connected): Cisco 5000 ?

TERA-100 Lustre Strategy

- **Lustre version**
  - 1.8 minimum
  - **2.0 targeted**
- **LNET**
  - **o2ib** for QDR
  - **o2ib** in iWARP mode for 10 GigE
- **Lustre HSM**
  - Will be used when integrated in Lustre release

Shine
(Lustre Administration Tool)

Shine and TERA-100

- **Lustre administration Python library and tool**
  - User-friendly configuration files
  - Cluster-wide, for small or large installations
  - Evolutive
- **Open source project in collaboration with Bull**
  - http://lustre-shine.sourceforge.net
- **Needs for TERA-100**
  - 10K nodes scalability (with ClusterShell v2)
  - Common tool to manage Lustre components in the Computing Center
  - Support all Lustre features and quickly conform with Lustre configuration and tuning changes

Shine Project Status

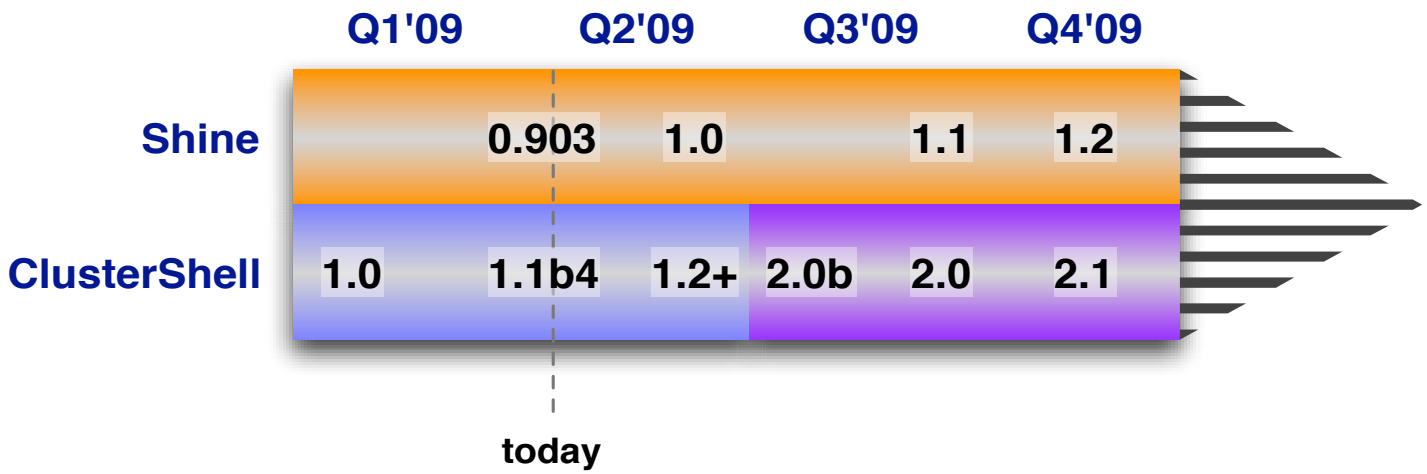
- **LUG 2009 : version 0.903**
  - Commands: `install, remove, format, status, start, stop, mount, umount, tune`
  - Partial HA support (configuration files and format)
  - Code design change to be used as a Python Library by scripts (eg. HA scripts)

- **ClusterShell 1.1 beta4**
  - Event-based Python library to execute local or remote shell commands
  - CEA open source project, requirement for Shine
  - Requires only ssh (since 1.1)
  - Scales up to 1K nodes

- **ClusterShell v2.0**
  - Will scale up to 10K+ nodes
  - Studying the best approach (one proof of concept is already working)
  - v2 move will have no impact on Shine's code
  - Needed for TERA-100

Shine Project Roadmap

- **Version 0.903 (1.0 beta) available today**
  - Along with ClusterShell 1.1b4 (required for Shine)
- **1.0 GA in June 2009**
  - Full HA support, `fsck` and `update` commands
- **1.1 by the end of Q3 2009**
  - OST Pools and Routers support
- **1.2 by the end of the year**
  - Multi-NIDs support




ccrt
NOVASCALE
Questions ?