HSM for Lustre : Data hierarchization for Parallel File Systems

Guy Chesnot – gchesnot@sgi.com

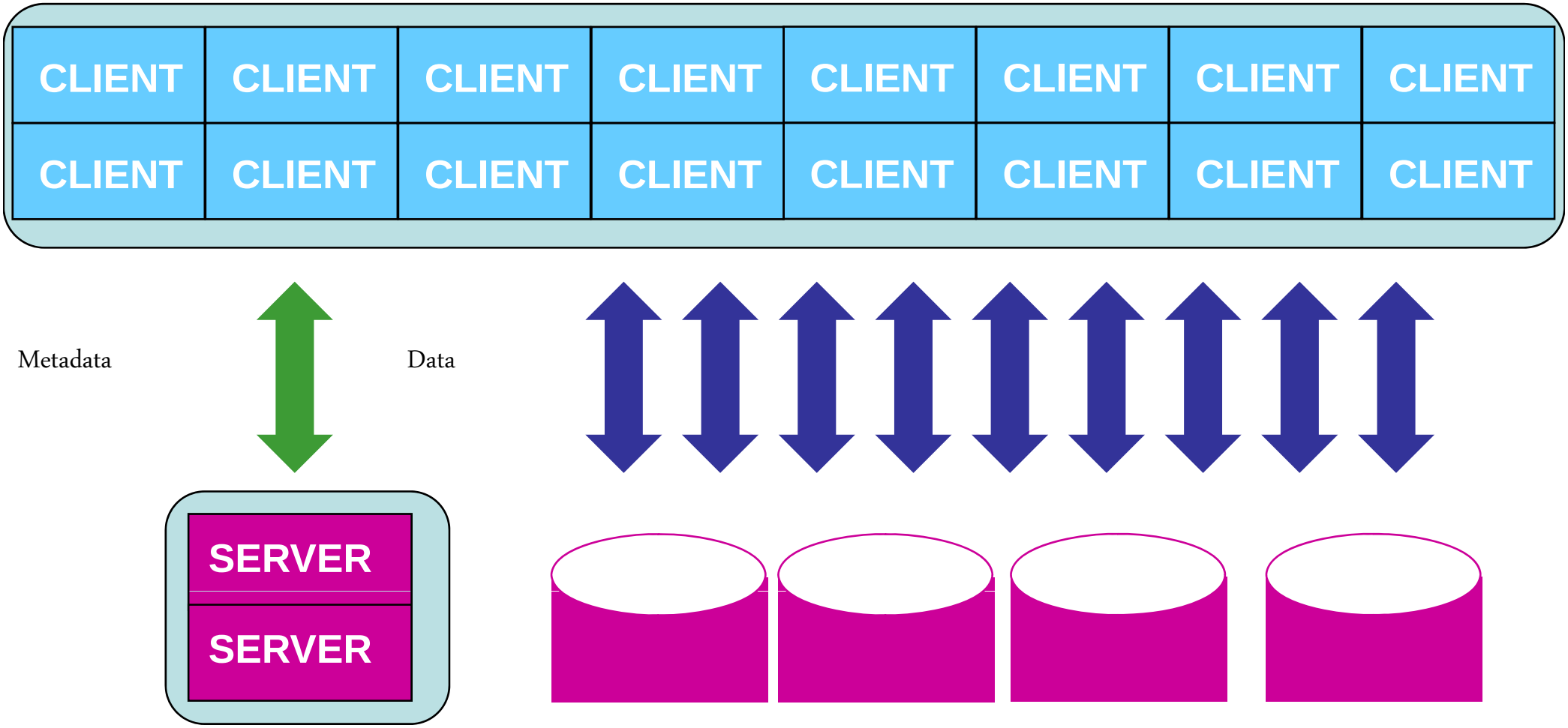
Agenda

- Introduction: Parallel File Systems advantages
- Large data capacity: Issues and goals
- Two answers: user managed, automated
- Common issues

Agenda

- **Introduction: Parallel File Systems advantages**
- Large data capacity: Issues and goals
- Two answers: user managed, automated
- Common issues

Parallel File System (PFS)



PFS Advantages

- Performance
  - Data transport
  - Bandwidth
  - Not for latency
  - Metadata: not yet
- Scalability
  - Bandwidth grows ~ linearly with capacity
- Costs

PFS Advantages for HPC

- Suits HPC requirements
  - High-speed data handling
  - More and more data

Agenda

- Introduction: Parallel File Systems advantages
- **Large data capacity: Issues and goals**
- Two answers: user managed, automated
- Common issues

The flood

- Tens of thousands clients
- So many files
- So many file sizes

Issues

- Capacity increase
- Data
  - Backup /
  - Short-mid-long term conservation /
  - Archiving
  - … whatever you name it
- Cost driven policies

First level issues

- Plain backup is a dead end
- Because of data volumes
- Because of transaction numbers
- Because of disk technology
  - Density, doubling every three years (average)
  - Hardly better access time (30% in 10 years)

- => HSM workflow

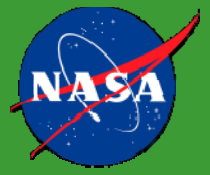
Agenda

- Introduction: Parallel File Systems advantages
- Large data capacity: Issues and goals
- **Two answers: user managed, automated**
- Common issues

« Simple » answer

- Low cost device: tape used as repository to duplicate PFS data
- **User managed data movement**
- Two (at least) levels hierarchy
  - First level: PFS disk storage
    - No management policy
  - Second and upper level: other disk space + tapes + remote system + etc.
    - Files metadata automatically ingested by HSM

User managed: NASA Ames



- Goals
  - Integration of Lustre & DMF (SGI HSM) as soon as possible
  - Performance:
    - 200 GB/s with Lustre
    - 10% (20 GB/s) to/from tapes
- Operational for a few months now
- Disk space management
  - No need of an HSM policy
  - NASA directs the data movement
- *"The biggest thing about any DLM system is reliability, reliability, reliability. You don't want to lose any data. That's really what drove us to implement DMF."*
  - Alan Powers, High End Computing Lead, NAS

Complex answer

- Low cost device: tape used as repository to duplicate PFS data
- **Automated movements between hierarchy levels**
- Two (at least) levels hierarchy: PFS disks + xxx
- HSM policy managing
  - PFS policy in charge of 1st level: PFS disks
  - HSM policy in charge of 2nd and other levels: disks, tapes, etc.

Automated: prospective customers

- Mostly all DMF customers wishing to protect all / part of their Lustre (or others POSIX) name space
  - Other Lustre NASA sites for instance
  - French Lustre (or others) & DMF customers … and other countries too

institut
national
de l'audiovisuel

Agenda

- Introduction: Parallel File Systems advantages
- Large data capacity: Issues and goals
- Two answers: user managed, automated
- **Common issues**

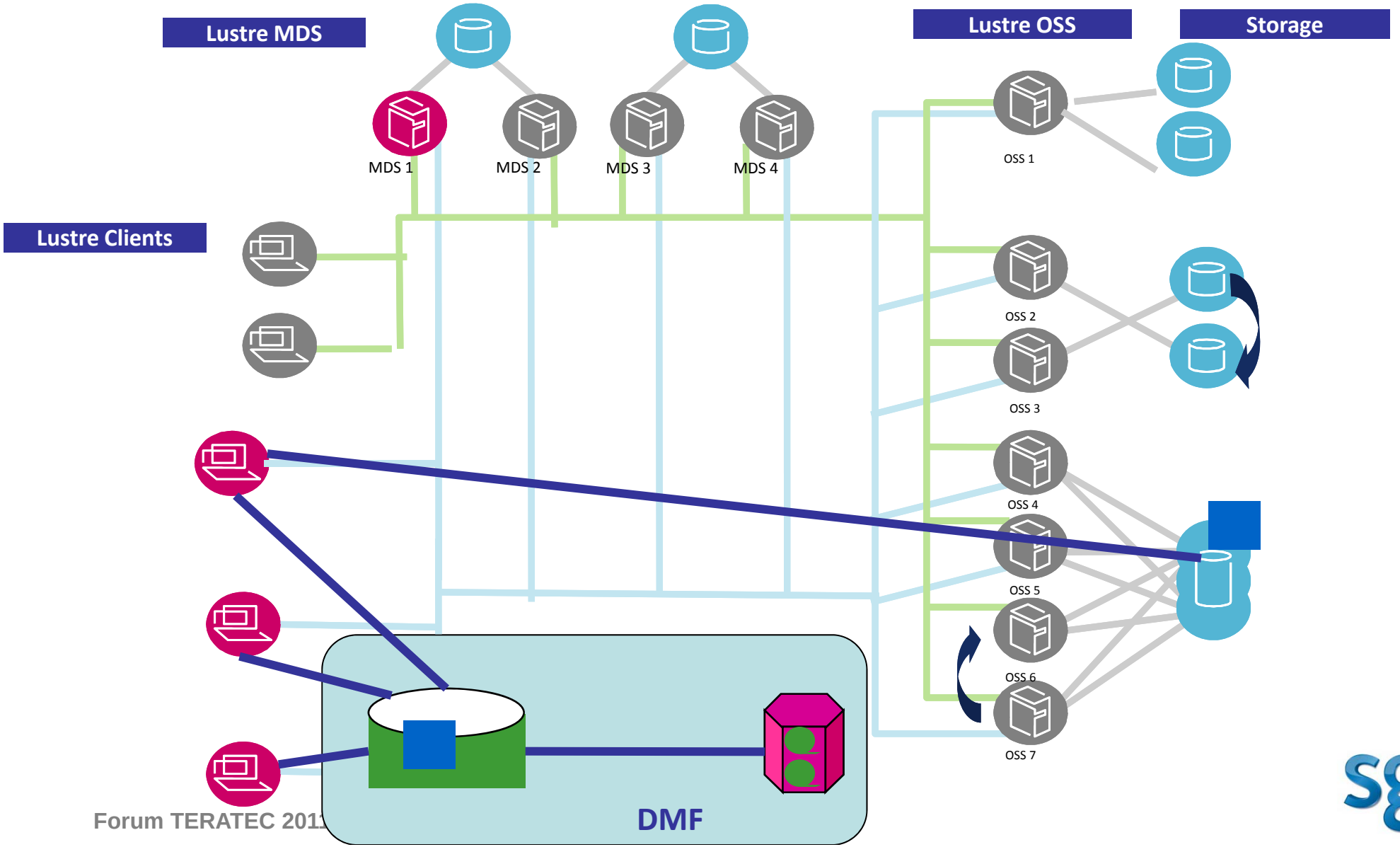
Main issue

- **Performance**
- To / from tapes
- Bandwidth
- Tape latency cannot be bypassed

Three levels architecture

- PFS disks – HSM disks – HSM tapes
- Files copied from primary filesystem disks to HSM disks
  - Migrated & freed immediately
  - Later recalled, copied to primary, freed again
- HSM policy implemented by PFS

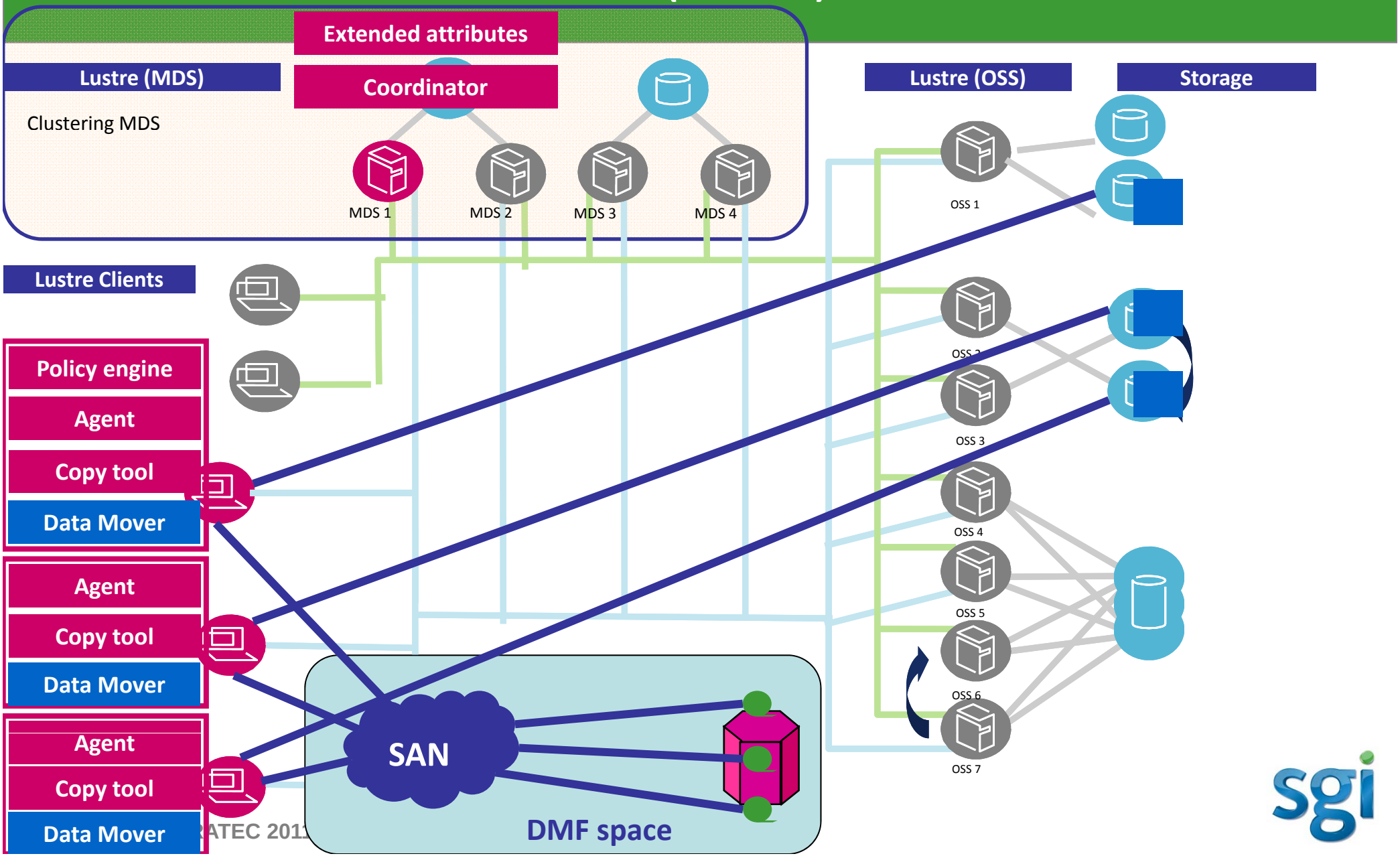
Three levels architecture (cont.)



Two levels architecture

- Direct-to-tape
  - Data moved directly from primary filesystem to
    - Tape or Disk or remote system
  - HSM filesystem used only as a namespace
    - Low capacity & bandwidth requirements
  - Primary filesystem (PFS disks) can be any POSIX filesystem
- Direct-from-tape
  - copy to non-HSM native filesystem
- Available with SGI Data Migration Facility, DMF

Two levels architecture (cont.)



Performance

- Parallelism inside HSM: disks, tapes
- Parallelized HSM
  - Numerous data movers

Performance: tape bandwidth

- Tape drive scheduling

- Library
- Robot
- Tape

} load balancing

Performance: tape bandwidth (cont.)

- **Rules for tape drives scheduling**
- Select the least used tape drive, with some constraints
  - Use same robot as the tape cartridge
  - Use same bay as the tape cartridge to avoid unnecessary cartridge movement
- Per data mover
  - Select port with greatest remaining bandwidth
- Globally
  - Select data mover with the most remaining bandwidth

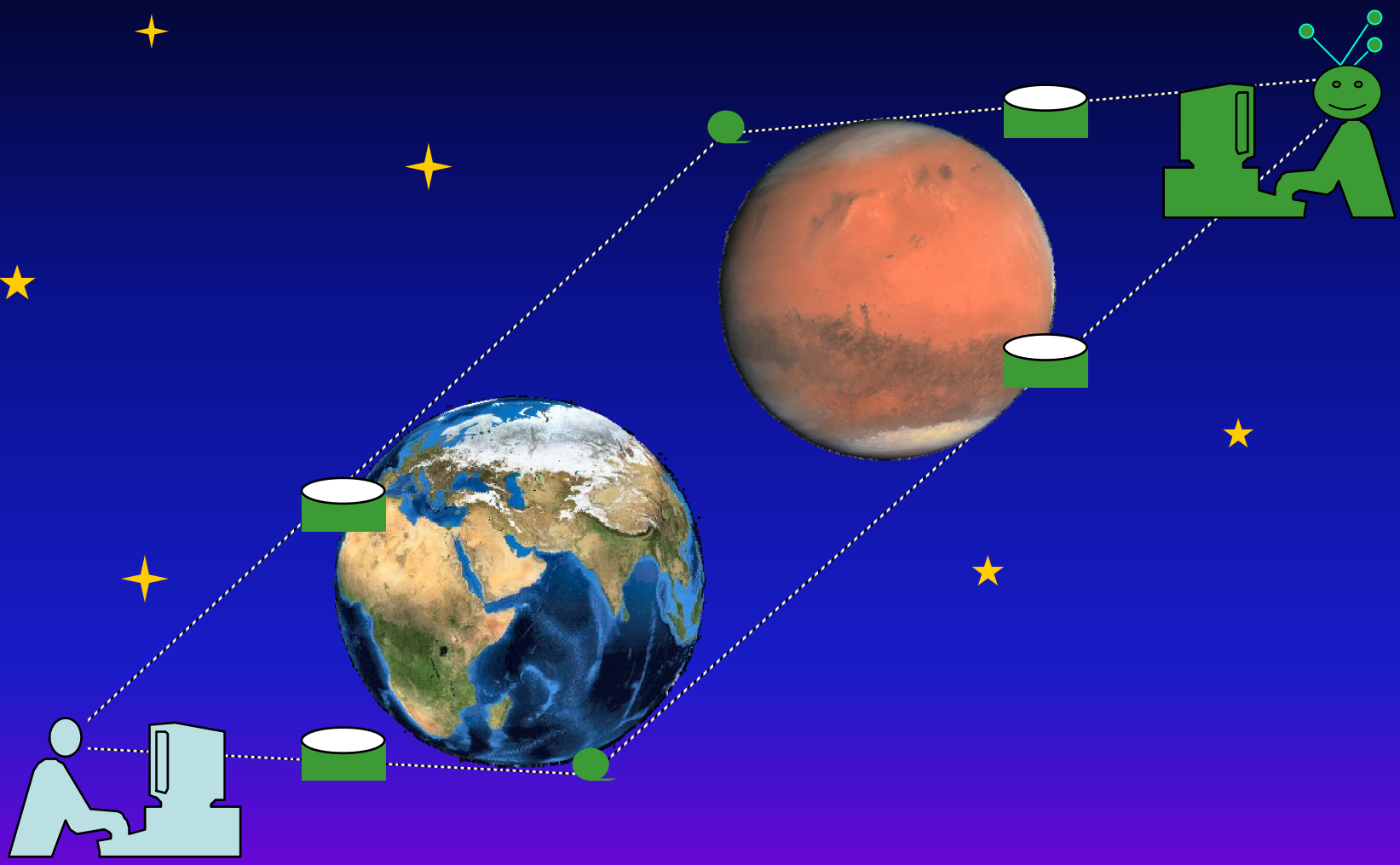
Performance: PFS MDS

- PFS MDS in charge of HSM policy
- Too much load
- Future?
  - Split MDS's
  - Dedicated MDS for HSM managed files
  - => multi level metadata

Concluding remark: Savings!

- Low acquisition cost
  - Tape cassettes (& tape drives if not enough bandwidth)
  - A few data movers: plain small x86 servers
  - HSM license
- Few admin
  - No backup pain
  - Less users' complains, because they lost data.

Prospective future




sgi
accelerating results