Lustre as a buffer and exchange area at TOTAL

Guy Chesnot - September 24th 2012

Agenda

- In the beginning
- Usual requirements
- Unique requirements
- Possible solutions
- Obligations and drawbacks
- Some difficulties on the way
- In the end: benefits

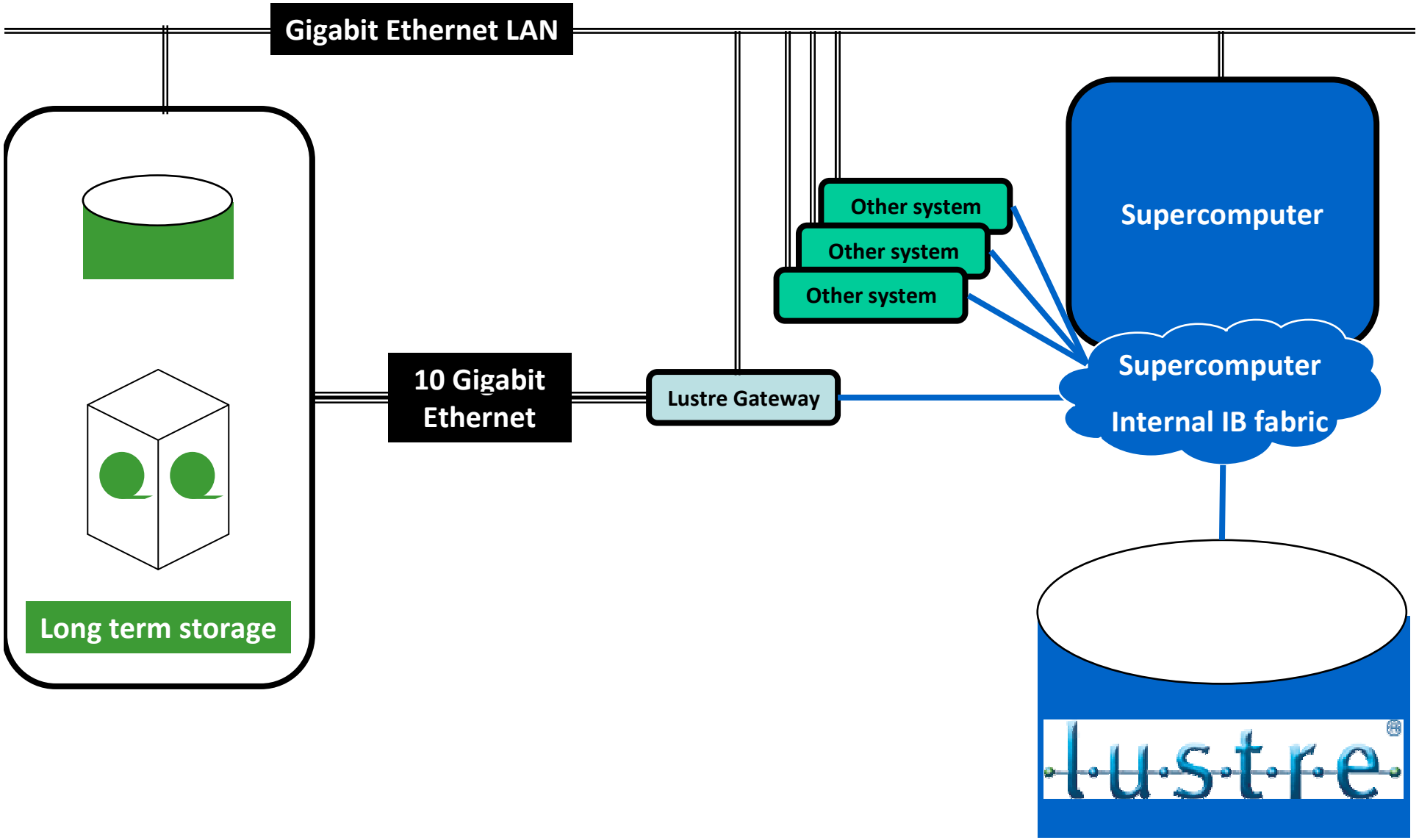
Agenda

- **In the beginning**
- Usual requirements
- Unique requirements
- Possible solutions
- Obligations and drawbacks
- Some difficulties on the way
- In the end: benefits

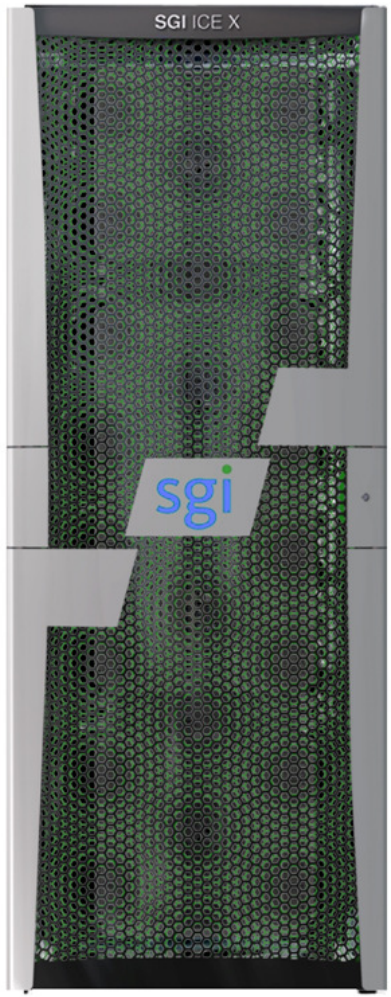
A standard Lustre configuration

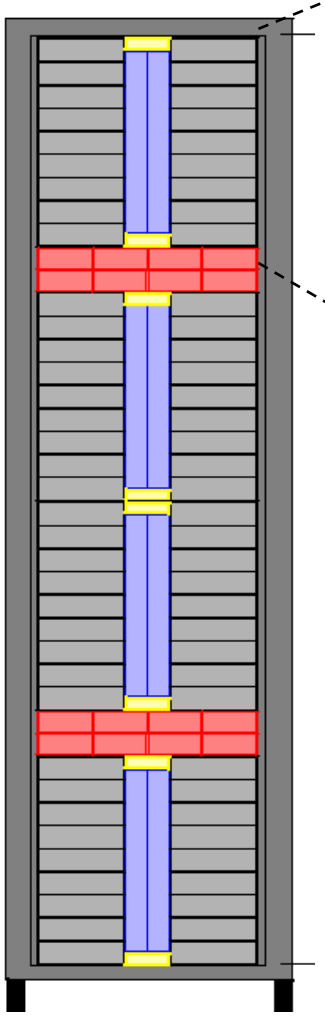
- A Lustre infrastructure based on the supercomputer own interconnect fabric
  - InfiniBand
- => Only one data class
- Time passing by
  - New systems attached to Lustre infrastructure
  - Therefore on supercomputer internal fabric

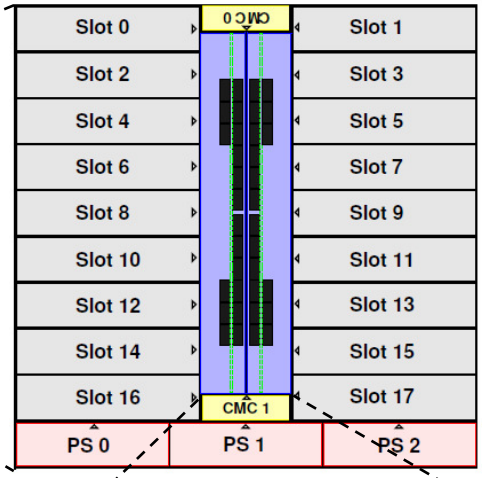
A standard Lustre configuration (cont.)



A bit of explanation
SGI Integrated Cluster Architecture







- Built-in IB fabric
- 36 port FDR IB switch ASIC
- Used for building fat tree, hypercube, enhanced hypercube and all-to-all topologies

sgi

Agenda

- In the beginning
- **Usual requirements**
- Unique requirements
- Possible solutions
- Obligations and drawbacks
- Some difficulties on the way
- In the end: benefits

Data classes

- Different lifecycles and usage patterns
  - From some days to some years
  - Short term (some days)
    - Scratch data
    - Very high performance required
    - As close as possible to the supercomputer
  - Mid term (some months): potential migration / unmigration
    - Whatever the supercomputer load and status
    - Used mostly by SMP platforms
  - Long term (some years)
- Data zones
  - Each user should be allowed to access only some datasets

Other needs

- Current status: large Lustre infrastructure
  - Stuck to a supercomputer that might
    - Stop (for maintenance purpose)
    - Change (new system, new provider, …)
- Need of a new architecture for
  - Fault tolerance
  - Ease of operation
  - Ease of extensions: new systems, new data features
- In summary, get rid of supercomputer's grip

Agenda

- In the beginning
- Usual requirements
- **Unique requirements**
- Possible solutions
- Obligations and drawbacks
- Some difficulties on the way
- In the end: benefits

Larger and faster data streams

- Larger input datasets
- Larger streams between scratch data space and long term data space
  - Current pipes are too narrow
    - Datasets cannot move and escape from Lustre infrastructure (tens of Terabytes)
  - Both sides streams
    - Large compute projects may restart later on
- Flush and fill quickly the Lustre space
  - Occupancy rate: 70% to 80%

Agenda

- In the beginning
- Usual requirements
- Unique requirements
- **Possible solutions**
- Obligations and drawbacks
- Some difficulties on the way
- In the end: benefits

Possible solutions

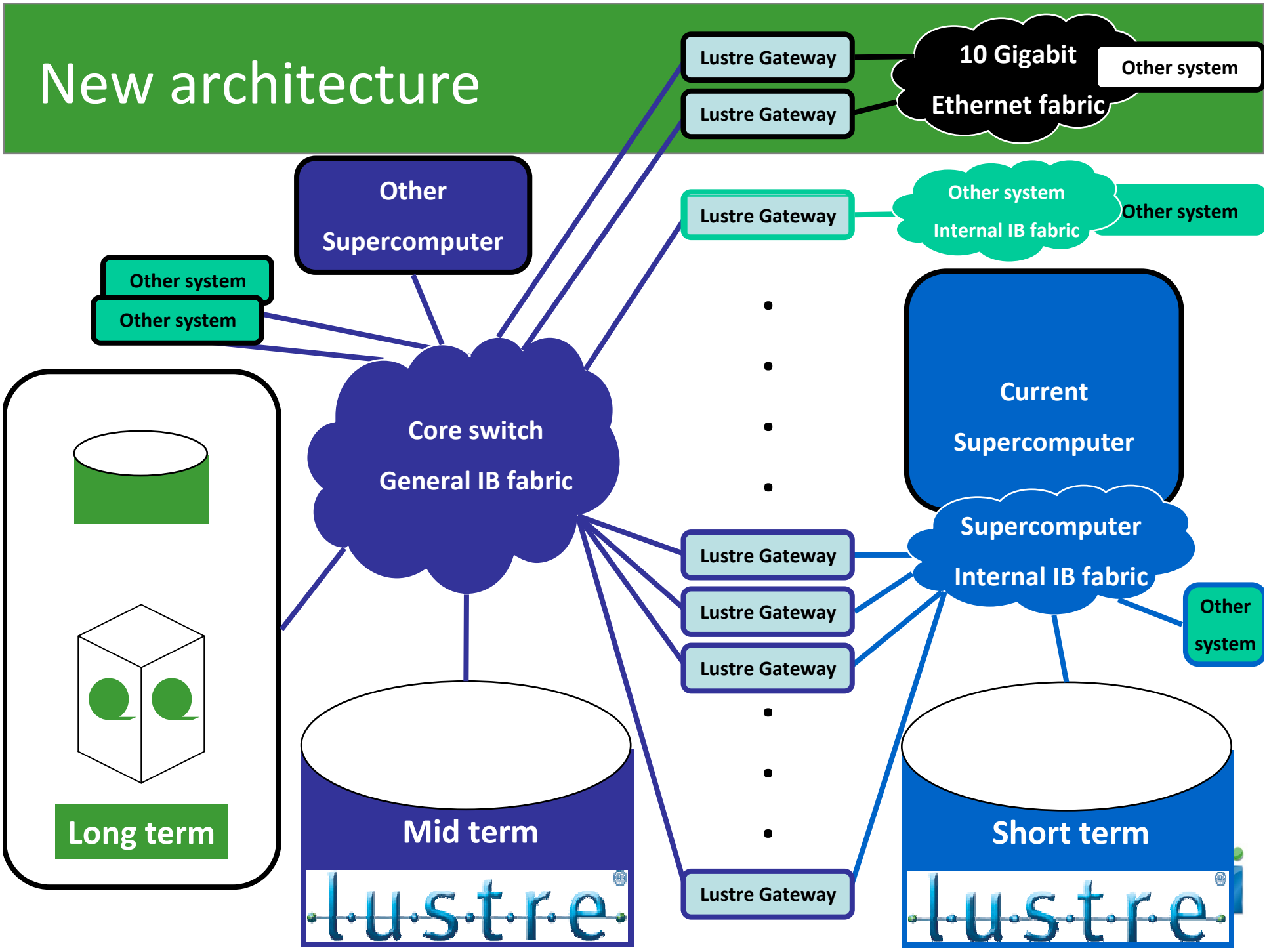
- Lustre with several attachments
  - Pros: easier implementation
  - Cons: restricted access from beyond security area

Possible solutions (cont.)

- **Independent Lustre infrastructure**
  - IB switch
  - Lustre gateways
- Pros
  - All requirements satisfied
  - IB topology is customized according to usage
    - Double hypercube for scratch
    - Non blocking for Mid term (could not be achieved on supercomputer fabric for every usage)
  - Data is at the center of stage
    - Several islands revolving around
- Cons: performance requirement harder to achieve

sgi

New architecture



Agenda

- In the beginning
- Usual requirements
- Unique requirements
- Possible solutions
- **Obligations and drawbacks**
- Some difficulties on the way
- In the end: benefits

Obligations

- Mixing Lustre clients is difficult
  - Theory and real life
  - Which Lustre release?
    - 1.8 was preferred to 2.1 (beginning of 2012)
- Mixing several IB: QDR, DDR, FDR to come
- « Big » Data
  - Large datasets
  - High transfer rates: 20+ GB/s
- By the way let’s move to parallel DMF (SGI HSM) ?!

Obligations: performance

- Several computer rooms … on several floors
  - 100 m IB links
    - Supercomputer -> long term storage -> SMP usage
    - => Higher latency. Is it acceptable or not?
    - => Mid term only, not for scratch
- Performance
  - Mid term -> supercomputer
  - Mid term <-> Long term storage
  - => numerous gateways
  - => large IB core switch

Drawbacks

- More Lustre gateways on data paths
  - Means higher latencies
  - Balanced by more potential features

Agenda

- In the beginning
- Usual requirements
- Unique requirements
- Possible solutions
- Obligations and drawbacks
- **Some difficulties on the way**
- In the end: benefits

Implementation not so easy

- No data loss on the way
  - Data replication at every step
- Temporary compute unavailability in some areas when routing rules are applied
- Process per 1 PB slices (file systems)
  - Freeze
  - Move
  - Restart

Agenda

- In the beginning
- Usual requirements
- Unique requirements
- Possible solutions
- Obligations and drawbacks
- Some difficulties on the way
- **In the end: benefits**

Benefits

- Lustre operation is free from Supercomputer
  - Halt / maintain supercomputer whenever needed
  - Think ahead of a new supercomputer or other evolutions
    - Many free IB ports on fabric
  - Control users
    - Partition
    - Foreign users
- Migration performance
  - From 10 TB to 120+ TB / day
  - Up to 175 TB -> tape library / day
    - Bottleneck is now the tape library

Benefits (cont.)

- A direct path to Lustre / HSM integration
  - Since Long term island data is already in place
  - And explicit transfers already work

- SGI does not hear of it (should work)!


sgi
accelerating results