Architect of an Open World™

Lustre SMP scalability and data I/O performance on bullx platforms

September 24, 2012

Grégoire PICHON

Parallel File Systems
Extreme Computing R&D

Agenda

- Bull Extreme Computing
- Lustre SMP Scalability
- Data I/O performance

Bull Extreme Computing

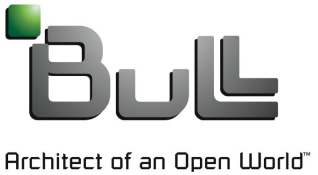
Actor of the HPC market

- 1st european manufacturer
- 3 supercomputers in the Top20
  - Curie (9th), Helios (12th), Tera-100 (17th)

Contributor to Lustre development

- EOFS founding member
- early adopter of lustre 2.0 and 2.1
- many bugs reported and several patches submitted
- working on Lustre static code analysis project

Lustre SMP Scalability

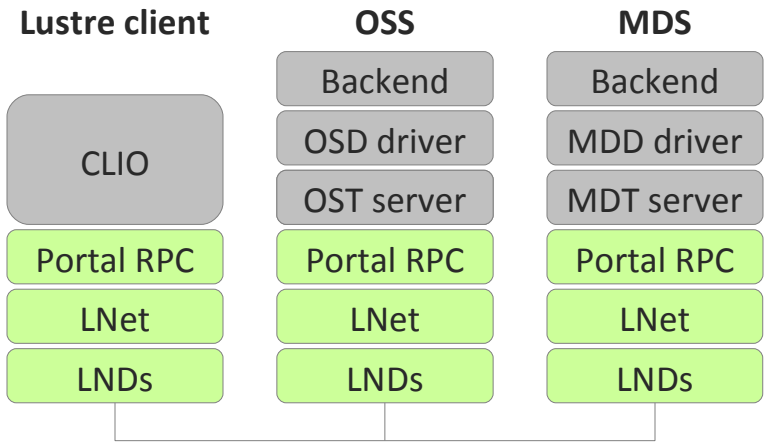


Lustre SMP scalability and affinity

Lustre SMP Node Affinity

- funded by OpenSFS
- landed to lustre 2.3, Liang Zhen (LU-56)
- goal
  - improve SMP scalability of Lustre, especially LNet
  - improve metadata performance for single MDS

Enhancement of Lustre networking Layers



Compute Partitions

What is it ?

- divide multi-CPU server into several processing partitions
- partition-local and per-partition data allocators
- Lustre threads can be bound to compute partitions

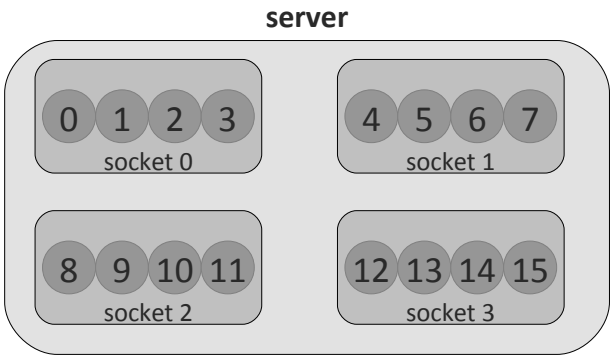
Compute Partitions

What is it ?

- divide multi-CPU server into several processing partitions
- partition-local and per-partition data allocators
- Lustre threads can be bound to compute partitions

Configuration parameters

- **cpu_npartitions** : # of CPU partitions
  - *options libcfs cpu_npartitions=2*
- **cpu_pattern** : CPU partitions pattern
  - *options libcfs cpu_pattern="0[0-7] 1[8-15]"*
  - *options libcfs cpu_pattern="N 0[0] 1[2]"*



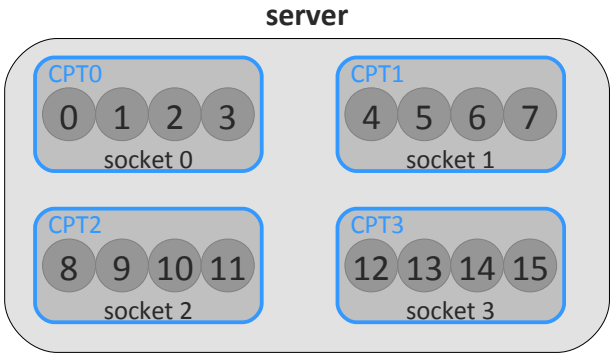
Compute Partitions

What is it ?

- divide multi-CPU server into several processing partitions
- partition-local and per-partition data allocators
- Lustre threads can be bound to compute partitions

Configuration parameters

- **cpu_npartitions** : # of CPU partitions
  - *options libcfs cpu_npartitions=2*
- **cpu_pattern** : CPU partitions pattern
  - *options libcfs cpu_pattern="0[0-7] 1[8-15]"*
  - *options libcfs cpu_pattern="N 0[0] 1[2]"*



Partitioned LNet and LND

- What is it ?
  - LND thread pool for each compute partition
  - LNet per-partition data
  - reduce lock contention

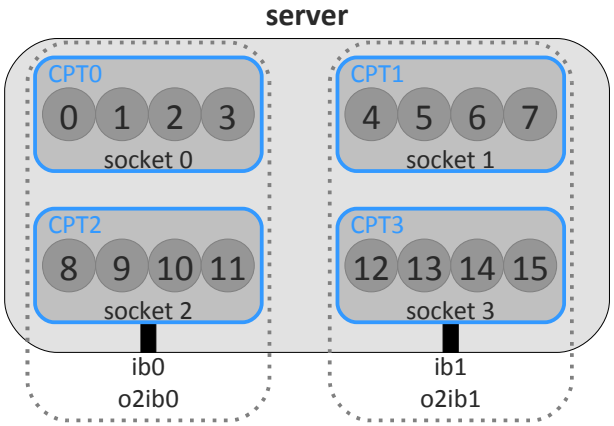
Partitioned LNet and LND

What is it ?

- LND thread pool for each compute partition
- LNet per-partition data
- reduce lock contention

Configuration parameters

- **networks** : bind network interfaces to specified CPTs
  - *options lnet networks="o2ib0(ib0)[0,2],o2ib1(ib1)[1,3]"*
- **portal_rotor** : distribute received messages to CPTs
  - local or round-robin
  - *options lnet portal_rotor=3*

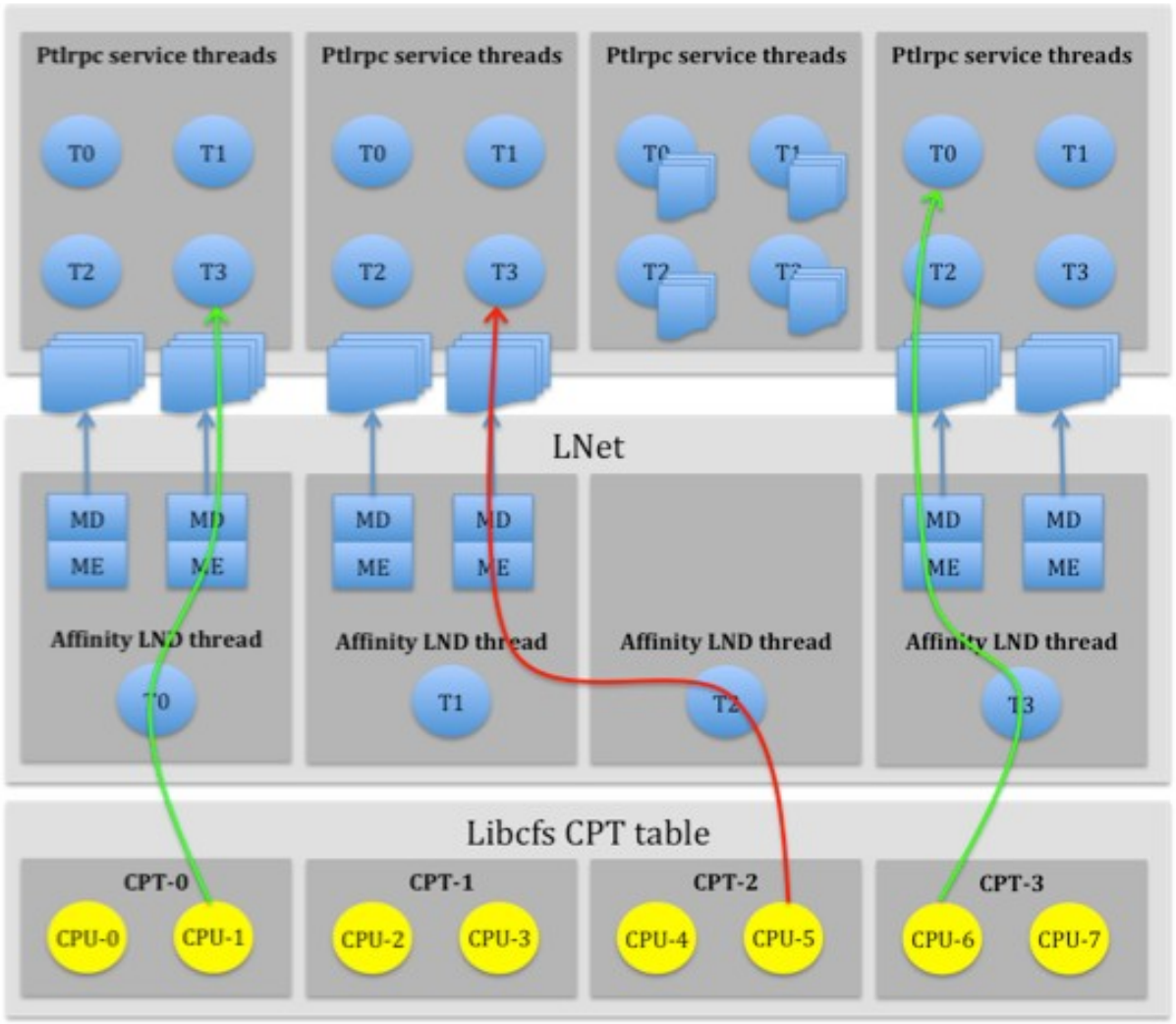


Functional Architecture

Portal Rotor

LNet network interface binding

Compute Partitions



source - *MDS SMP Node Affinity High Level Design*

MD: memory descriptor
ME: match entry
CPT: compute partition table

Partitioned ptlrpc service

What is it ?

- ptlrpc service thread pool for each partition
- request-queue and wait-queue for each partition
- to be completed with ptlrpc requests posting (lustre client)

Partitioned ptlrpc service

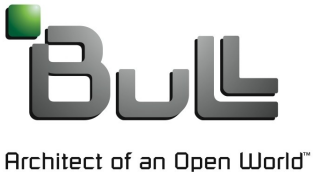
What is it ?

- ptlrpc service thread pool for each partition
- request-queue and wait-queue for each partition
- to be completed with ptlrpc requests posting (lustre client)

Configuration parameters

- CPU partitions threads should run on
  - MDS threads : **mds_num_cpts, mds_rdpg_num_cpts, mds_attr_num_cpts**
  - OSS threads : **oss_cpts, oss_io_cpts**
  - ldlm threads : **ldlm_cpts**

Data I/O Performance



Data I/O performance

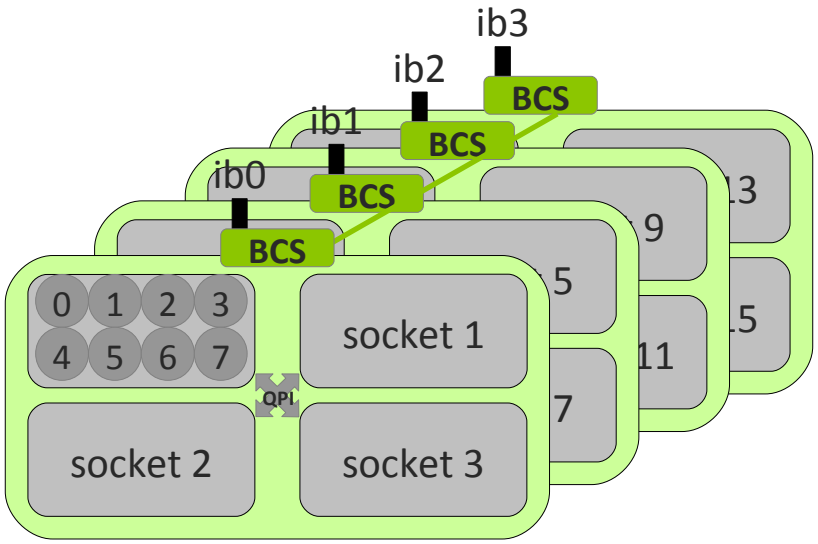
Goal

- evaluate the effect on data I/O performance
- measure the improvement on bullx Supernodes
  - large SMP node (up to 128 cores), highly NUMIOA
  - follow up of NUMIOA presentations at LUG 2010 and LUG 2011

Data I/O performance

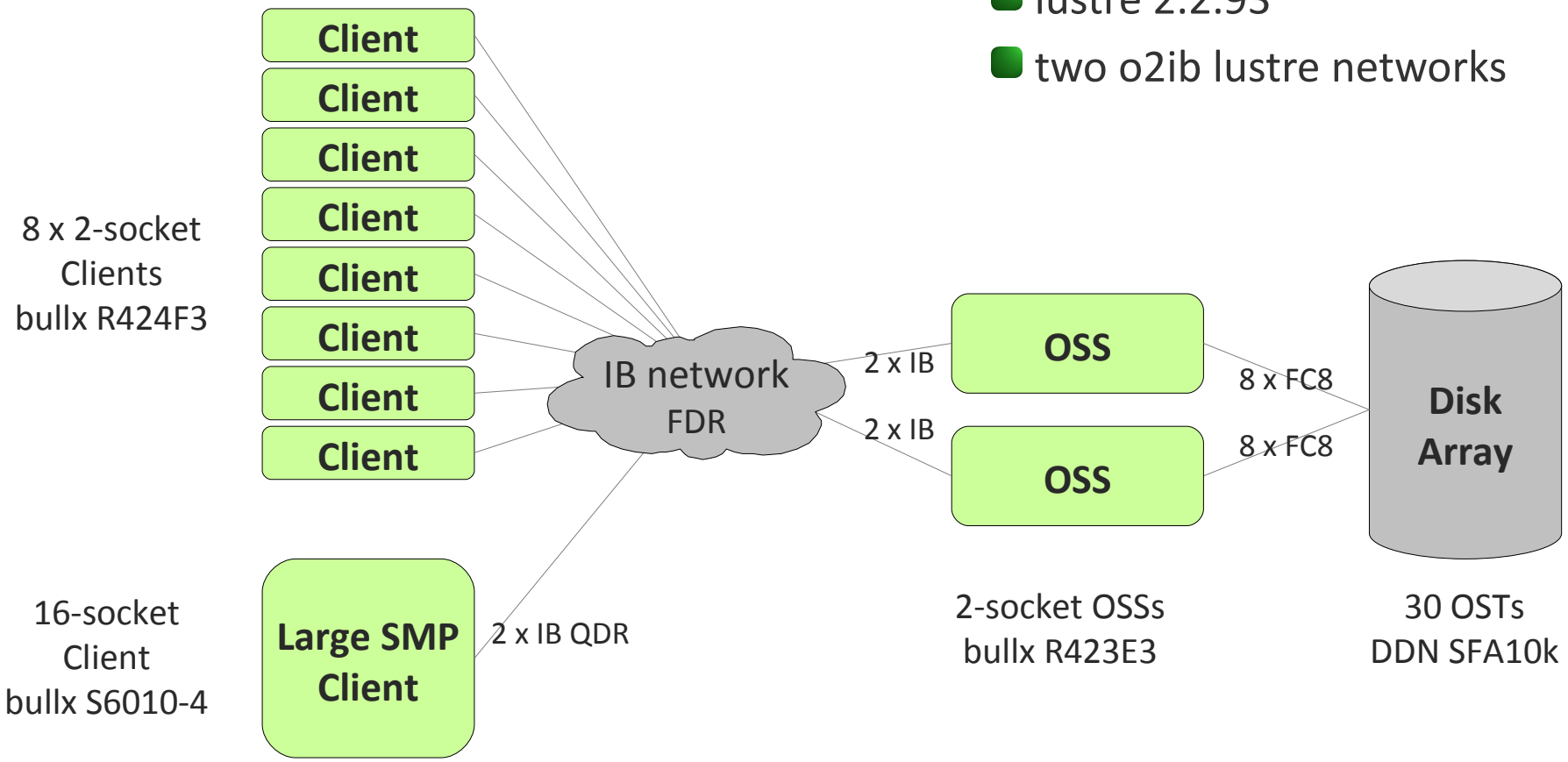
Goal

- evaluate the effect on data I/O performance
- measure the improvement on bullx Supernodes
  - large SMP node (up to 128 cores), highly NUMIOA
  - follow up of NUMIOA presentations at LUG 2010 and LUG 2011



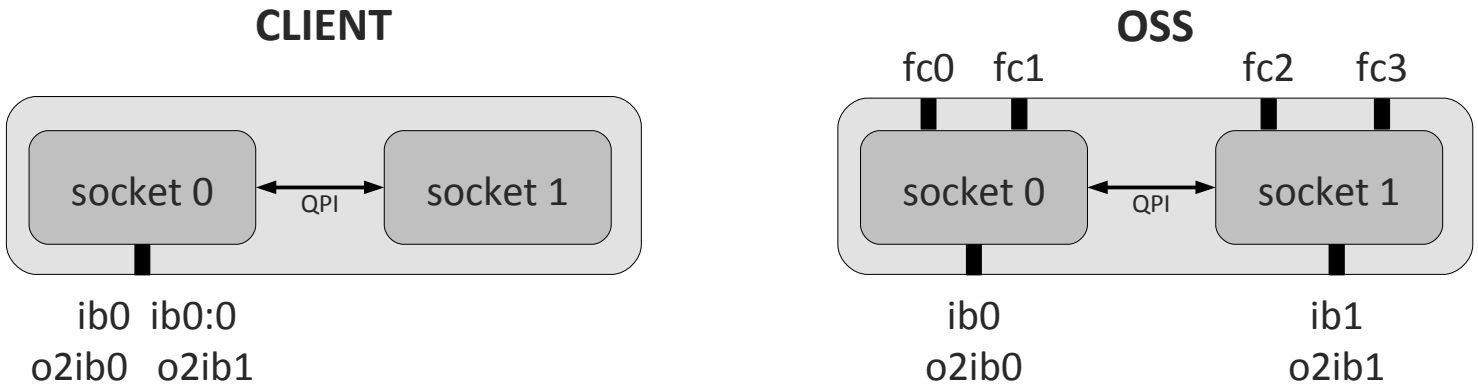
Test Cluster Architecture

- kernel 2.6.32
- OFED 1.5.4
- lustre 2.2.93
- two o2ib lustre networks



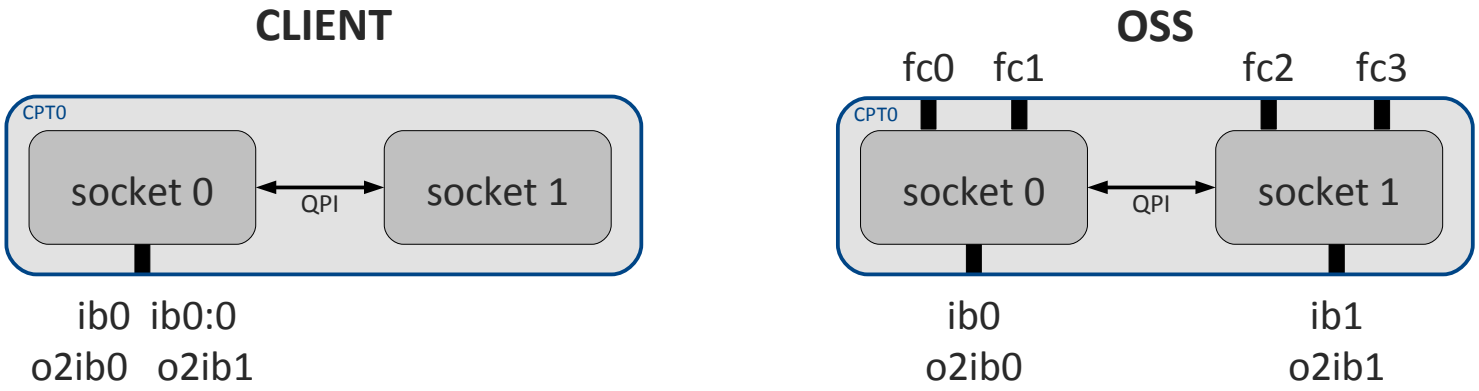
Lustre Configurations

- **nocpt** compute partitions disabled
- **auto** compute partitions automatic setup
- **cpt2** two compute partitions
- **cpt2lcl** two compute partitions, network interfaces bound to local CPT
- **cpt2rmt** two compute partitions, network interfaces bound to remote CPT
- **lustre21**



Lustre Configurations

- **nocpt** compute partitions disabled
- **auto** compute partitions automatic setup
- **cpt2** two compute partitions
- **cpt2lcl** two compute partitions, network interfaces bound to local CPT
- **cpt2rmt** two compute partitions, network interfaces bound to remote CPT
- **lustre21**

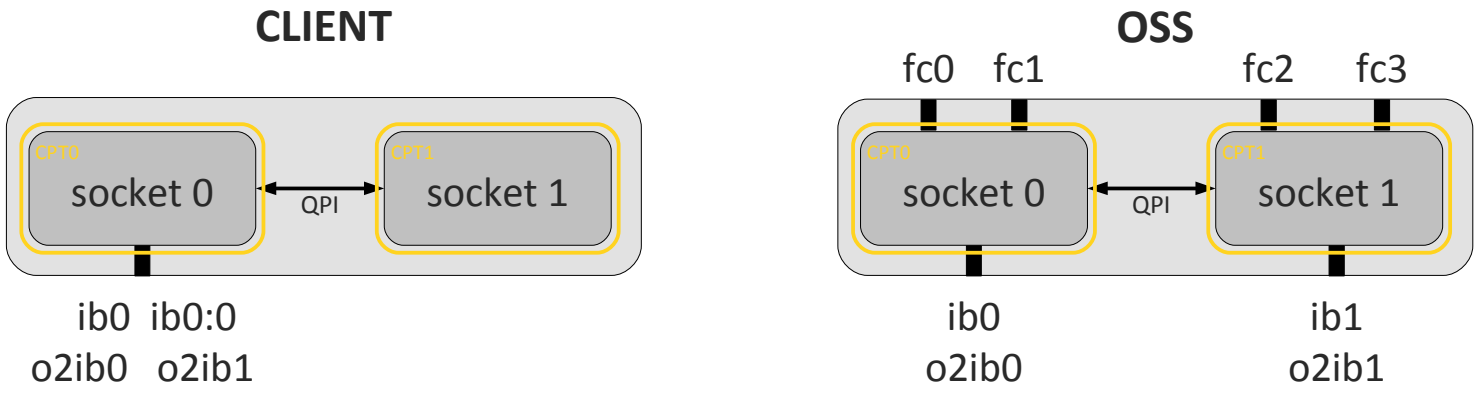


Lustre Configurations

- **nocpt** compute partitions disabled
- **auto** compute partitions automatic setup
- **cpt2** two compute partitions
- **cpt2lcl** two compute partitions, network interfaces bound to local CPT
- **cpt2rmt** two compute partitions, network interfaces bound to remote CPT
- **lustre21**

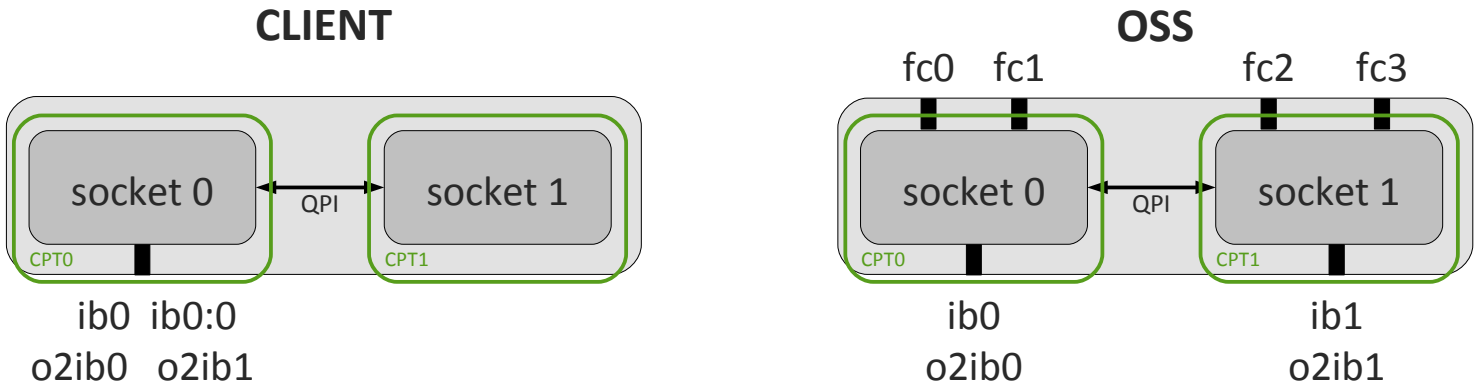
Lustre Configurations

- **nocpt** compute partitions disabled
- **auto** compute partitions automatic setup
- **cpt2** two compute partitions
- **cpt2lcl** two compute partitions, network interfaces bound to local CPT
- **cpt2rmt** two compute partitions, network interfaces bound to remote CPT
- **lustre21**



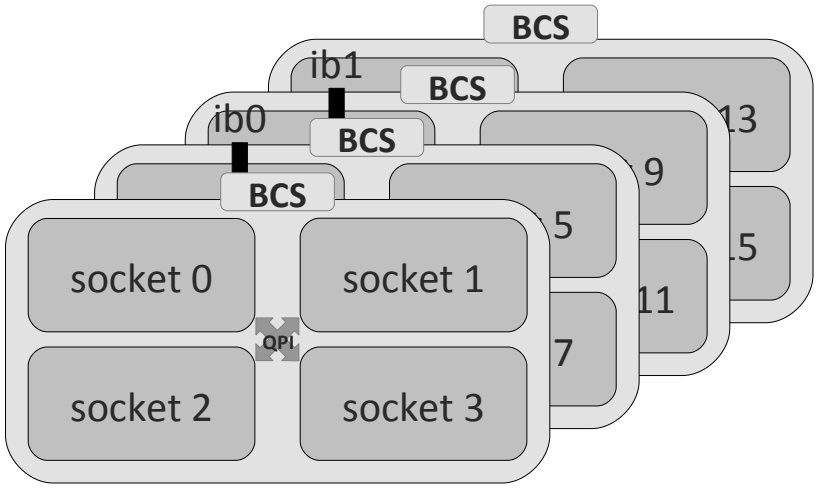
Lustre Configurations

- **nocpt** compute partitions disabled
- **auto** compute partitions automatic setup
- **cpt2** two compute partitions
- **cpt2lcl** two compute partitions, network interfaces bound to local CPT
- **cpt2rmt** two compute partitions, network interfaces bound to remote CPT
- **lustre21**



Lustre Configurations

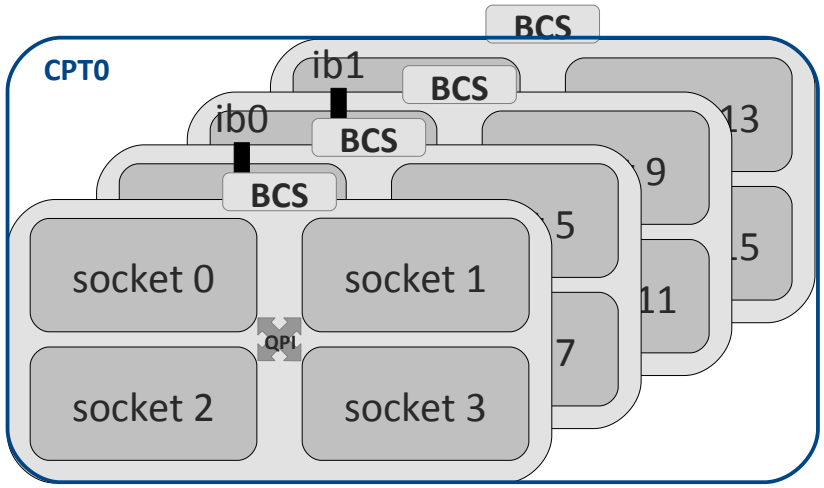
- **nocpt** compute partitions disabled
- **auto** compute partitions automatic setup
- **cpt2** two compute partitions
- **cpt2lcl** two compute partitions, network interfaces bound to local CPT
- **cpt2rmt** two compute partitions, network interfaces bound to remote CPT
- **lustre21**



Large SMP CLIENT

Lustre Configurations

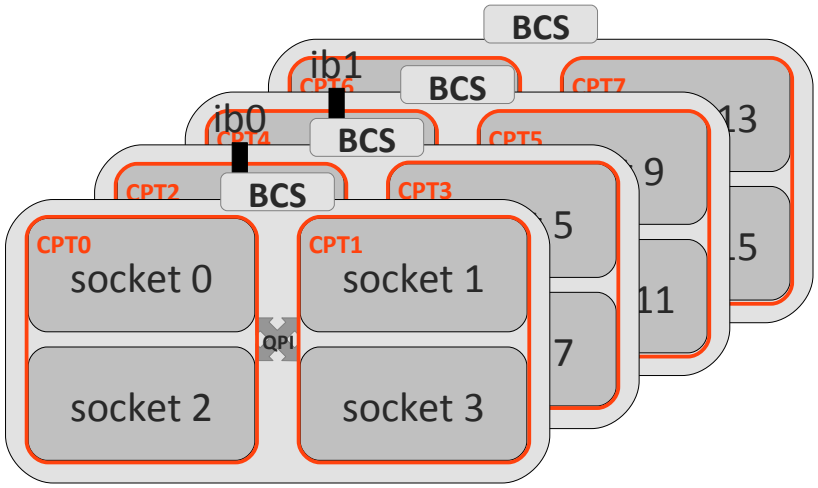
- **nocpt** compute partitions disabled
- **auto** compute partitions automatic setup
- **cpt2** two compute partitions
- **cpt2lcl** two compute partitions, network interfaces bound to local CPT
- **cpt2rmt** two compute partitions, network interfaces bound to remote CPT
- **lustre21**



Large SMP CLIENT

Lustre Configurations

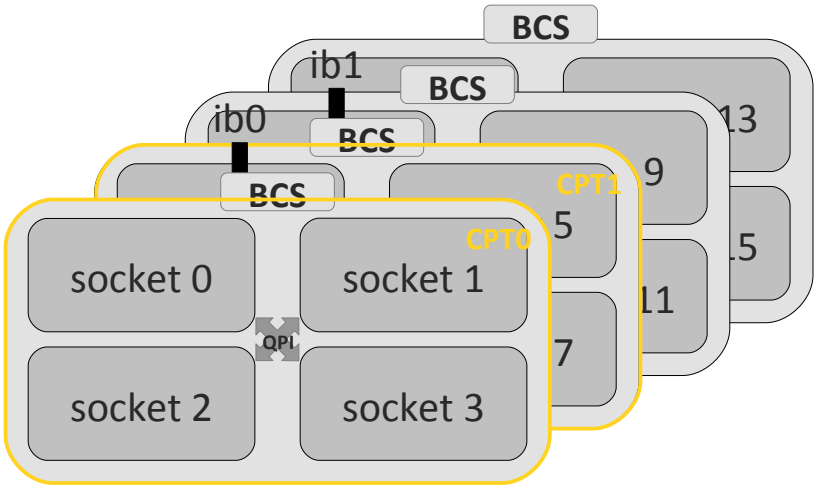
- **nocpt** compute partitions disabled
- **auto** compute partitions automatic setup
- **cpt2** two compute partitions
- **cpt2lcl** two compute partitions, network interfaces bound to local CPT
- **cpt2rmt** two compute partitions, network interfaces bound to remote CPT
- **lustre21**



Large SMP CLIENT

Lustre Configurations

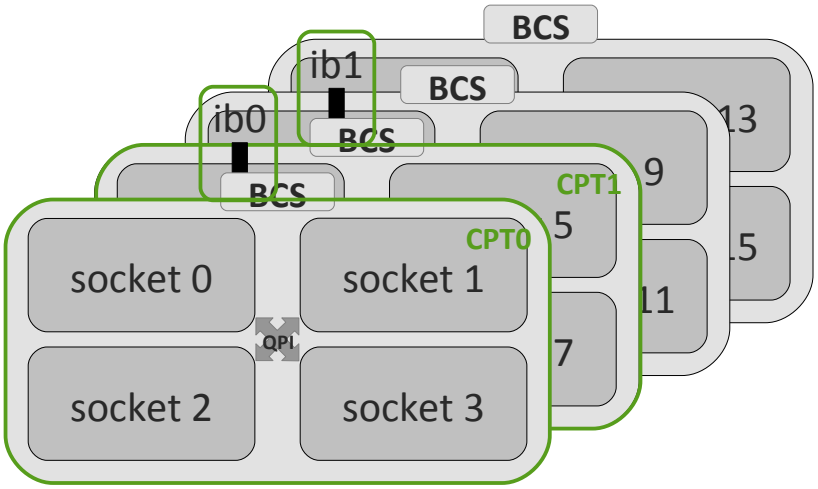
- **nocpt** compute partitions disabled
- **auto** compute partitions automatic setup
- **cpt2** two compute partitions
- **cpt2lcl** two compute partitions, network interfaces bound to local CPT
- **cpt2rmt** two compute partitions, network interfaces bound to remote CPT
- **lustre21**



Large SMP CLIENT

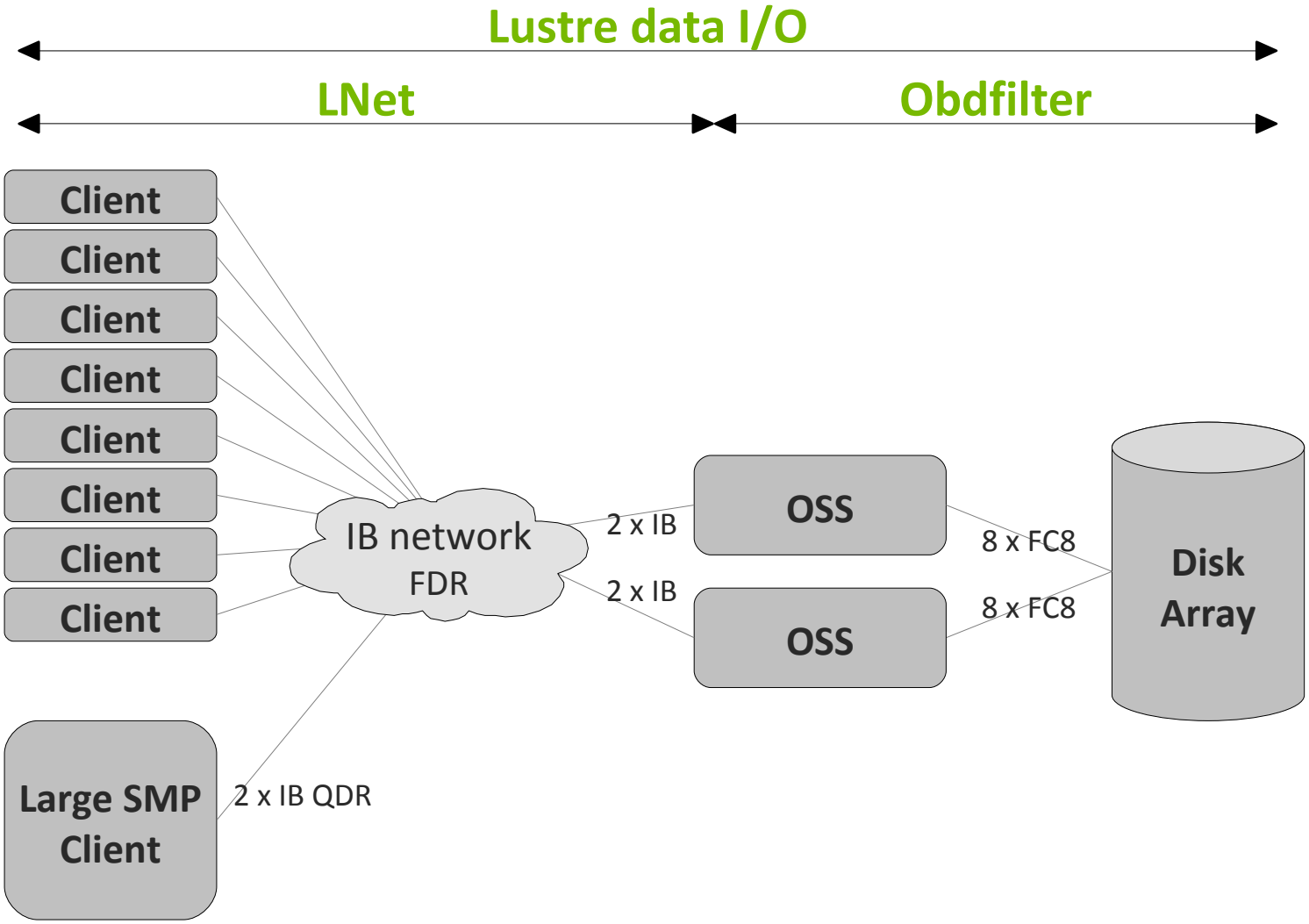
Lustre Configurations

- **nocpt** compute partitions disabled
- **auto** compute partitions automatic setup
- **cpt2** two compute partitions
- **cpt2lcl** two compute partitions, network interfaces bound to local CPT
- **cpt2rmt** two compute partitions, network interfaces bound to remote CPT
- **lustre21**



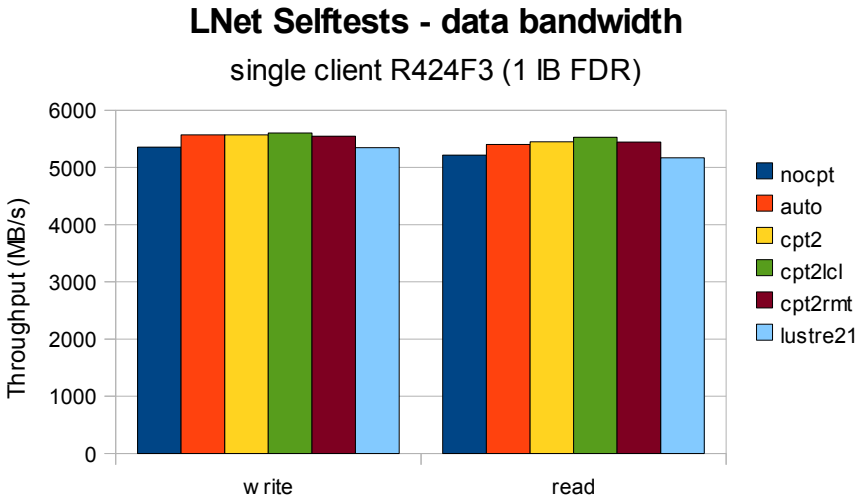
Large SMP CLIENT

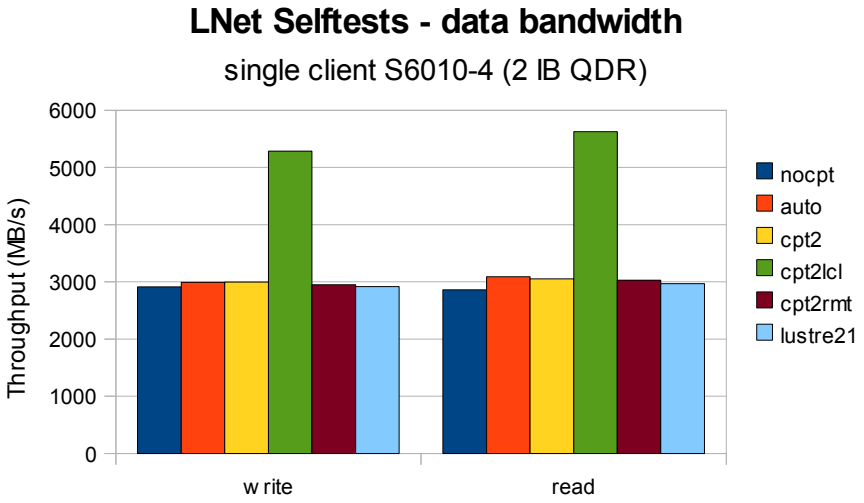
Data I/O performance measurements



LNet data bandwidth

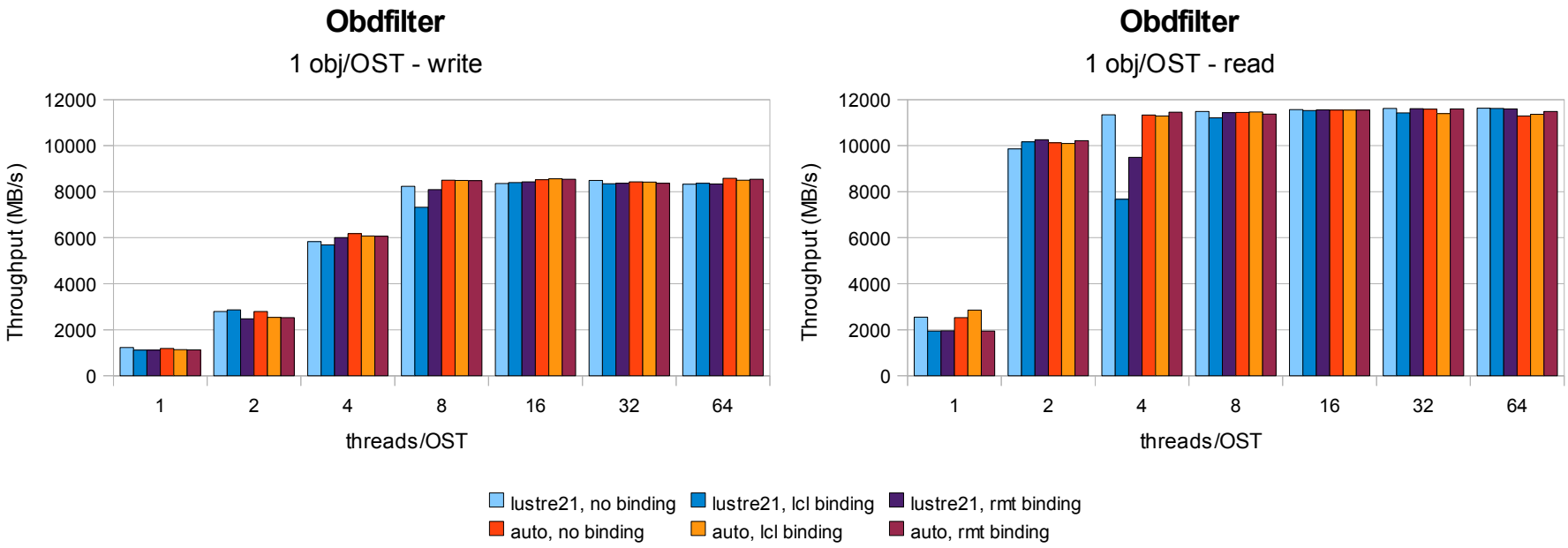
- single client
  - 2-socket client : 5% improvement vs. lustre21 or no CPT
  - large SMP client : 80% improvement





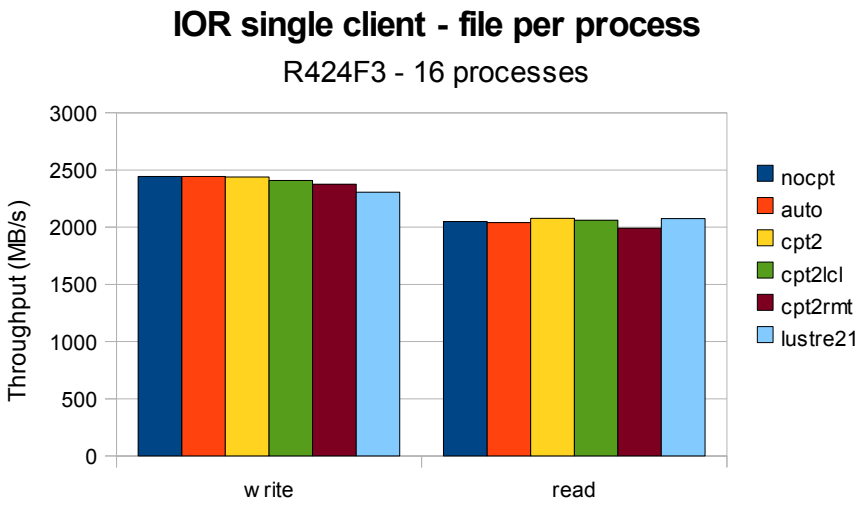
Obdfilter bandwidth

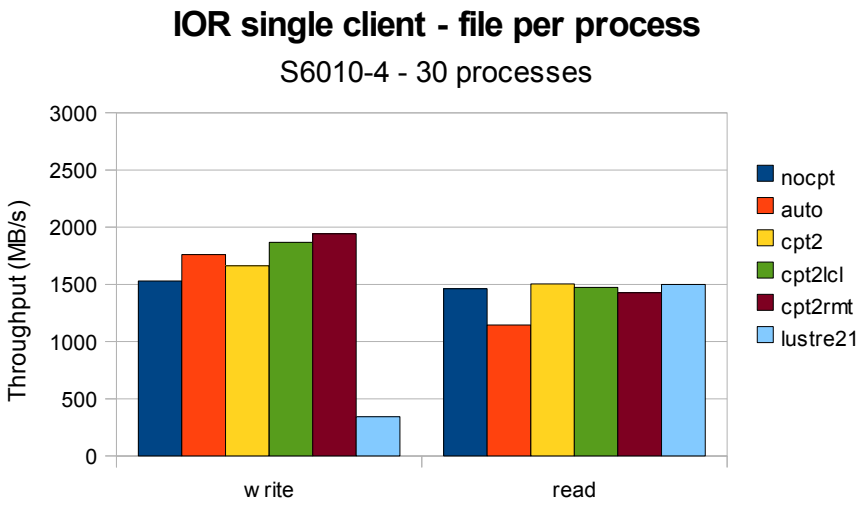
- test process placement according to NUMA node of each OST
  - enhanced version of obdfilter-survey
  - no NUMIOA effect observed



File System data I/O Bandwidth

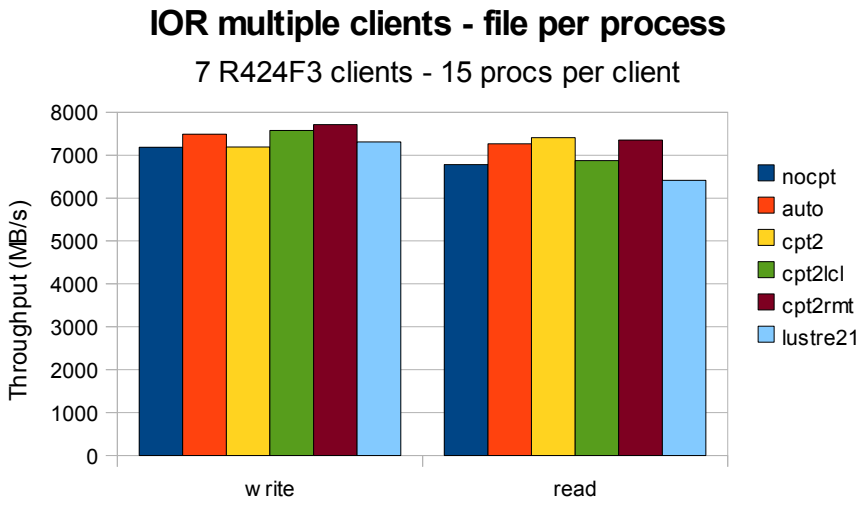
- single client
  - far below LNet bandwidth:
    performance is limited by other Lustre client components
  - LU-744 Single client's performance degradation on 2.1





File System data I/O Bandwidth

- multiple clients
  - OSSs have a low NUMIOA effect (obdfilter results)
  - at least no regression



Conclusion

SMP scalability and affinity feature

- efficient APIs to run Lustre on SMP and NUMIOA machines
- configuration is manual
  - "dynamic LNet configuration" project

Data I/O performance results

- at LNet level: visible improvement
- at filesystem level:
  - at least no regression
  - masked by client limitations

Architect of an Open World™

Hardware / Software configuration

OSS

- 2 bullx R423E3 - Intel SandyBridge E5-2660, 2.20GHz, 16 CPU cores, 32GB Memory, 2 x IB FDR adapters, 4 x bi-port FC8 adapters
- DDN SFA10k - 16 x FC8, 300 x 940GB SATA disks, 30 x RAID6 pools
- 30 OSTs - 30 data + 30 journal LUNs

Clients

- 8 bullx R424F3 - Intel SandyBridge E5-2660, 2.20GHz, 16 CPU cores, 32GB Memory, IB FDR adapters
- 1 bullx S6010-4 - Intel NehalemEX X7550, 2.00GHz, 128 CPU cores, 256GB Memory, 2 x IB QDR adapters

Network

- InfiniBand FDR
- two o2ib lustre networks
- OSTs restricted to one of the lustre network

Software Stack

- bullxlinux 6.2 (based on rhel 6.2)
- OFED 1.5.4
- lustre 2.2.93

Test parameters

- LNet Selftests
  - brw read/write, o2ib0+o2ib1, size=1M, concurrency=64, check=none
- Obdfilter-survey
  - case=disk, size=10GB, recordsize=1M
  - nobjlo=1, nobjhi=64, thrlo=1, thrhi=256
  - verify=0
  - numactl=none/local/remote

    numactl --physcpubind=xxx --localalloc lctl test_brw ...
- IOR
  - stripe count=1, stripe size=1M
  - api=POSIX, filePerProc=1, fsync=1, transferSize=1MB
  - single R424F3 client: blockSize=4GB, numTasks=16
  - single S6010-4 client: blockSize=2GB, numTasks=30
  - multiple R424F3 clients: blockSize=4GB, numTasks=15*#clients

Lustre SMP Node Affinity - MDS results

from Liang Zhen's presentation - Aug, 28 2012

LNet Selftests

- lustre 2.3 ping is 900% (600%) of lustre 2.2 with Portal-RR OFF (ON)
- lustre 2.3 4K-BRW is 500%-700% of lustre2.2

mdtest

- lustre 2.3 open-create performance is 350%-400% of lustre 2.2
- lustre 2.3 unlink performance is 150%-300% of lustre 2.2
- lustre 2.3 stat performance is 200%-400% of lustre 2.2