OpenSFS Project
Lustre SMP Node Affinity

Liang.zhen@intel.com

Aug, 28 2012

Agenda

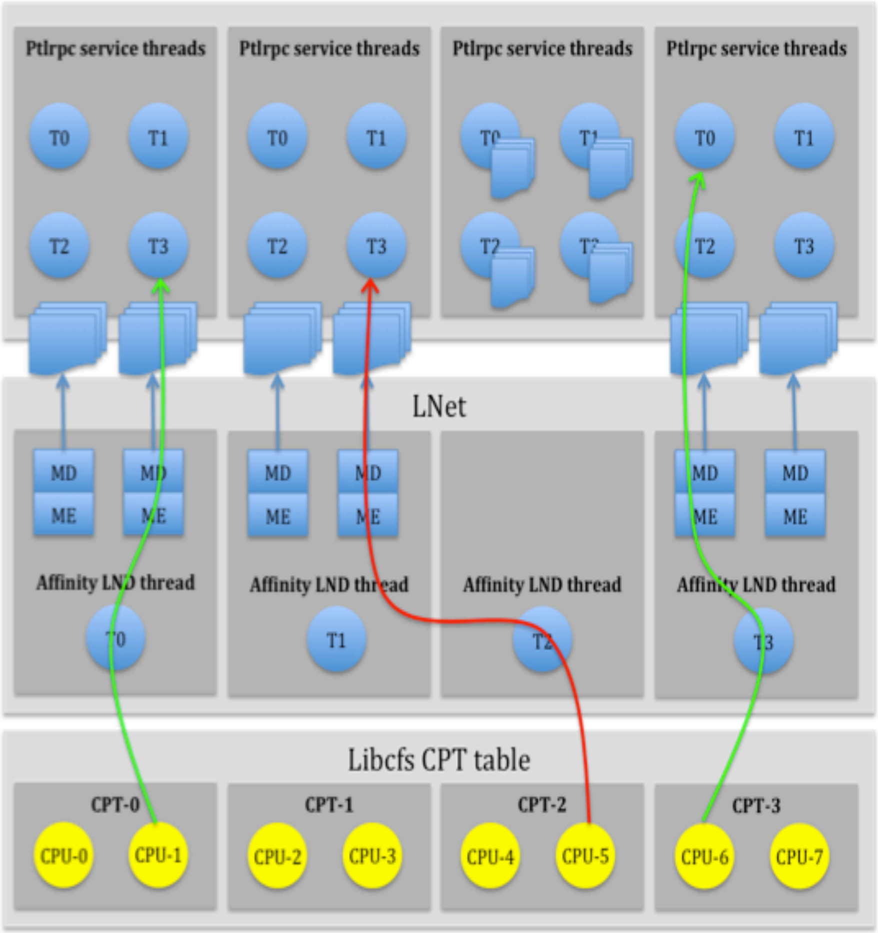
- Background
- Demonstration
- Tuning Lustre on SMP machine

Background

- Goal of this project
  - Improve SMP scalability of LNet
  - Improve metadata performance for single MDS
  - Funded by OpenSFS
- Code landed to 2.3
  - 16K+ LOC

Partitioned Lustre Server

- CPU Partition (CPT)
  - Similar to cpuset of linux
  - Can be easily used by kernel thread
- Partitioned LNet(LND)
  - LND thread-pool for each CPT
  - Core LNet has partition data
- Partitioned ptlrpc service
  - Ptlrpc service thread-pool for each CPT
  - Request-queue & wait-queue for each CPT

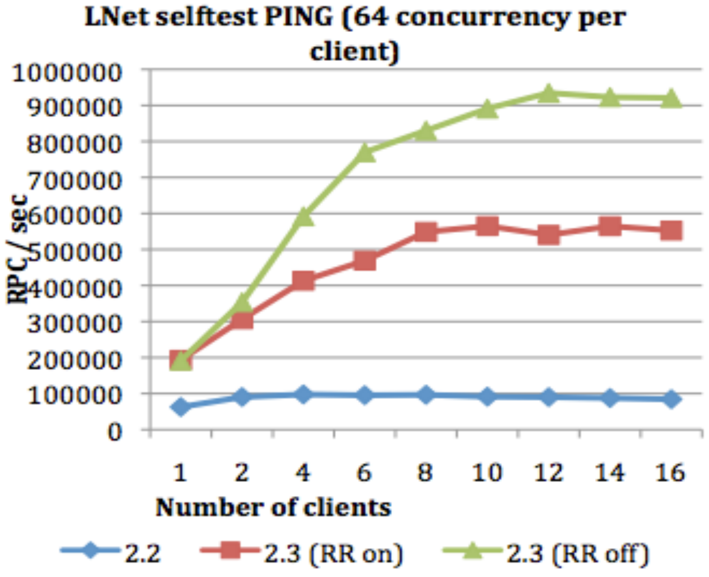


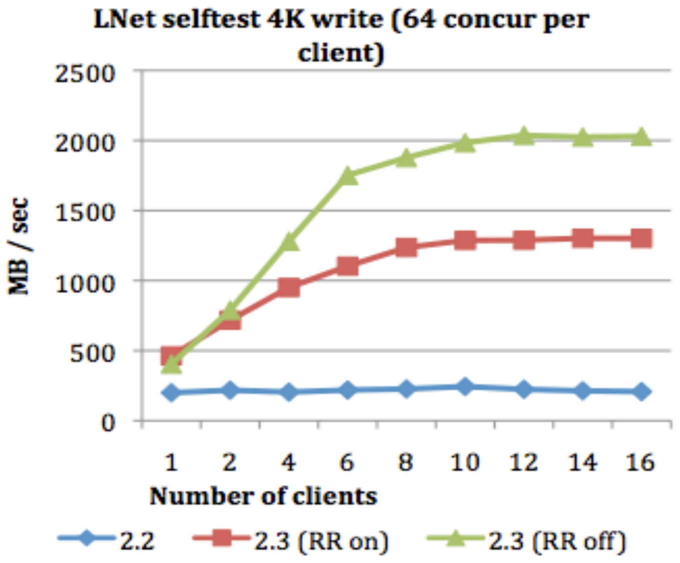
LNet performance tests

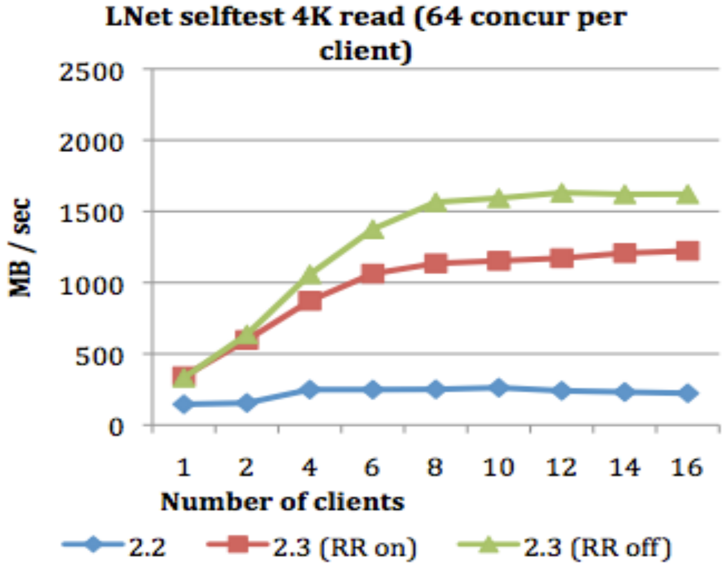
- Hardware
  - Server: 6-core CPU (2-HT), 2 sockets
  - Client: 4-core, 1 socket
  - QDR infiniband
- LNet selftest
  - Selftest ping
  - Selftest 4K read/write
  - Concurrency
- Portal Round-Robin (Portal RR)
  - NID affinity in LNet (LND)
  - Enable/disable NID affinity of incoming message for upper layer (ptlrpc service, or LNet selftest)

LNet performance

- 2.3 ping is 900% of 2.2 with Portal-RR OFF
- 2.3 ping is 600% of 2.2 with Portal-RR ON
- 2.3 4K-BRW is 600%-700% of 2.2 with Portal RR OFF
- 2.3 4K-BRW is 500% of 2.2 with Portal RR ON





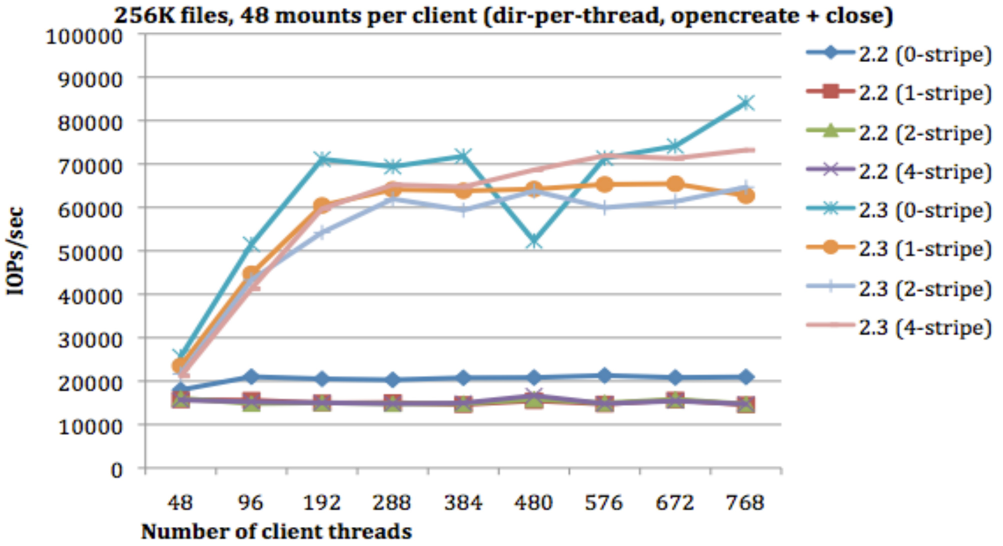


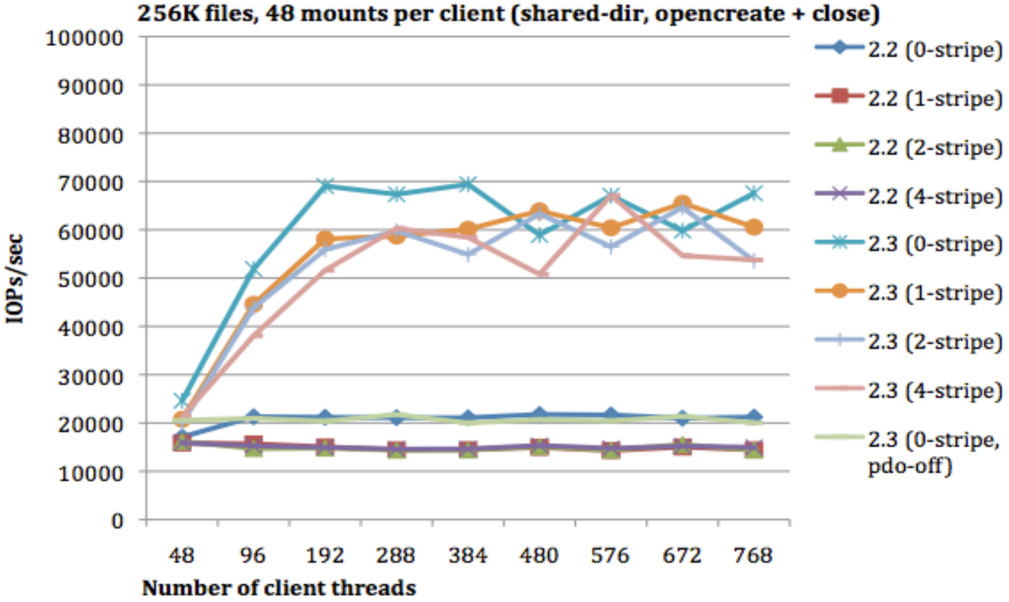
mdtest

- Hardware
  - MDS
    - 6-core CPU (2-HT), 2 sockets
    - 8G SSD as MDT journal
  - OSS
    - 3 OSSs, 6 OSTs per OSS
  - Client: 4-core, 1 socket
  - QDR infiniband
- Mdtest patches
  - multi-mount
    - Simulate high work load with small number of clients
    - Disable mdc_rpc_lock can't help shared directory tests
  - 0-stripecount file
    - w/o OST object creation

File creation performance

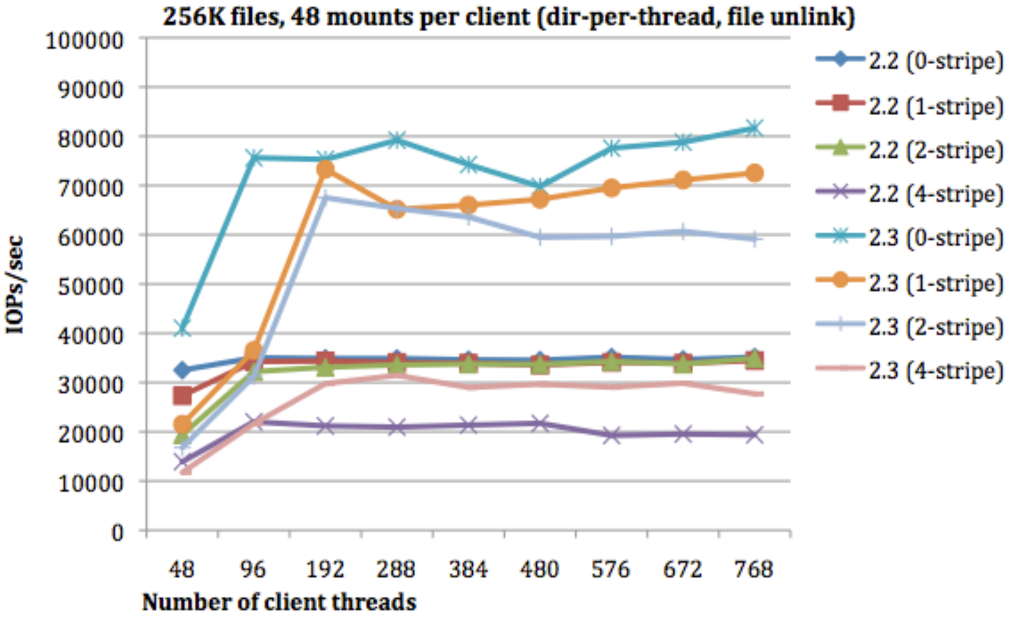
- Iterate over 1,2, 6, 4, 8,10, 12,14, 16 clients
- Each client has 48 threads
- Each thread is running under a private mount
- 2.3 opencreate performance is 350%-400% of 2.2
- OST object pre-creation works pretty good
- Turning off PDO, shared directory opencreate performance of 2.3 is similar to 2.2

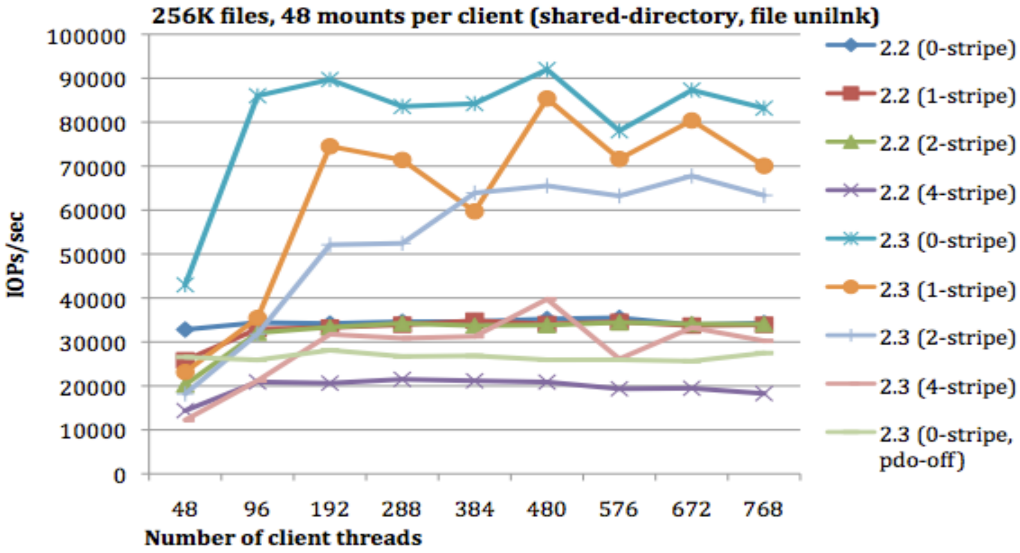




File unlink performance

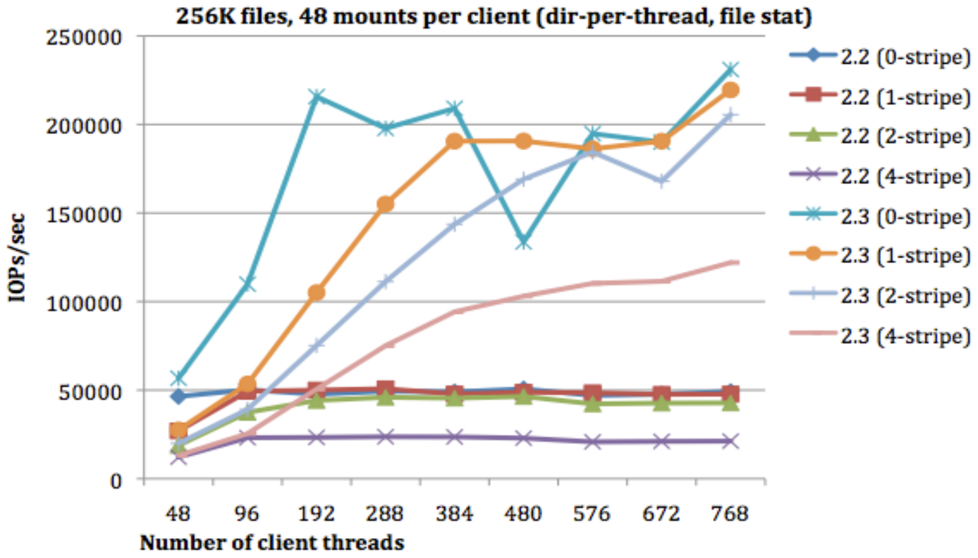
- Iterate over 1,2, 6, 4, 8,10, 12,14, 16 clients
- Each client has 48 threads
- Each thread is running under a private mount
- 2.3 unlink performance is 150%-300% of 2.2
- Client needs to send RPC to destroy each OST object
- Turning off PDO, shared directory opencreate performance of 2.3 is even worse than 2.2

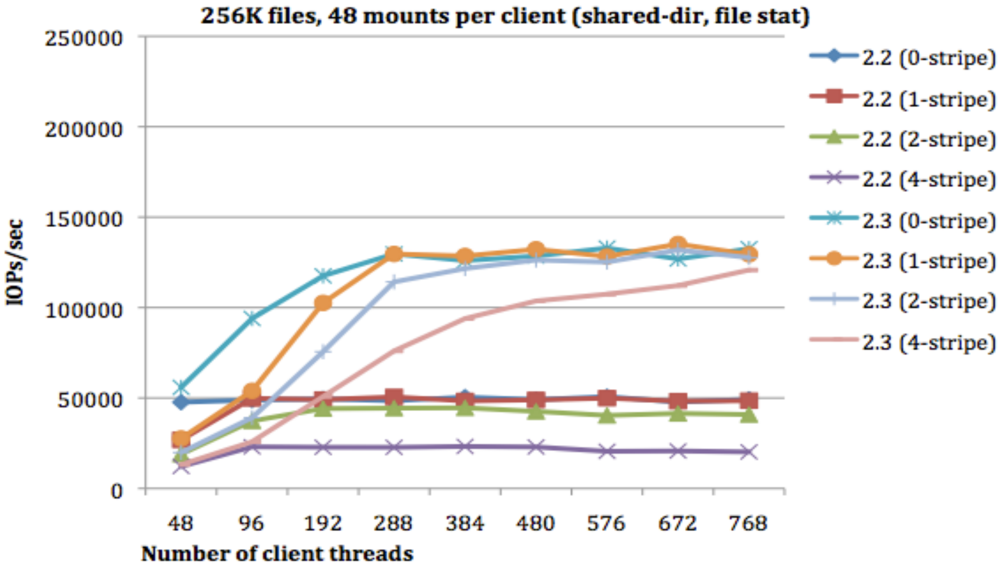




File stat performance

- Iterate over 1,2, 6, 4, 8,10, 12,14, 16 clients
- Each client has 48 threads
- Each thread is running under a private mount
- 2.3 stat performance is 200%-400% of 2.2
- Client needs to send RPC to stat each OST object

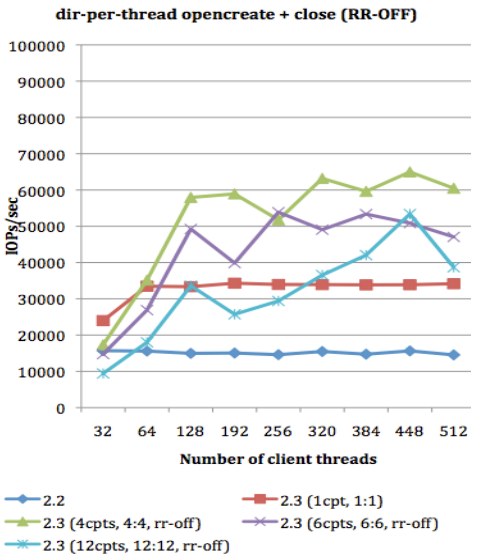


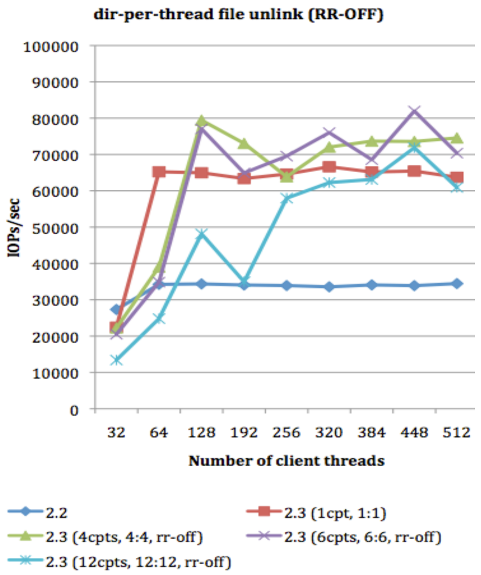


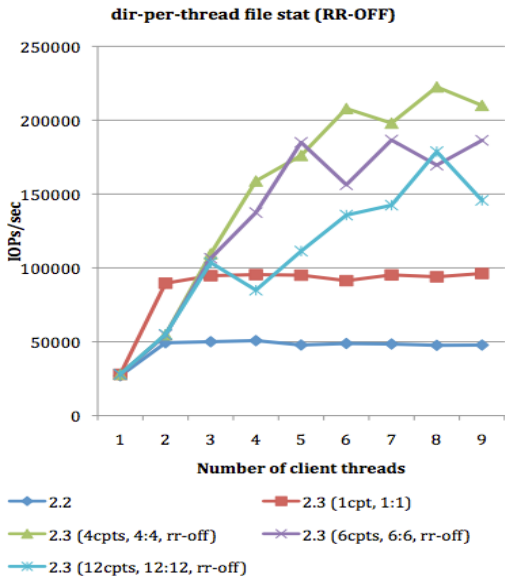
Performance of different CPT configurations

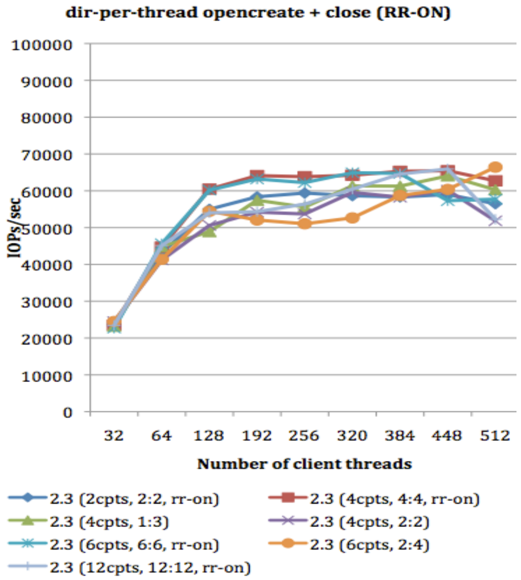
- MDS has 12 cores (24 HTs)
- 1 CPT
- 2 CPTs
  - Portal-RR ON
- 4 CPTs (default)
  - Portal-RR ON & OFF
  - 2 CPTs for LNet, 2 CPTs for ptlrpc service
  - 1 CPT for LNet, 3 CPTs for ptlrpc service
- 6 CPTs
  - Portal-RR ON & OFF
  - 2 CPTs for LNet, 4 CPTs for ptlrpc service
- 12 CPTs
  - Portal-RR ON & OFF

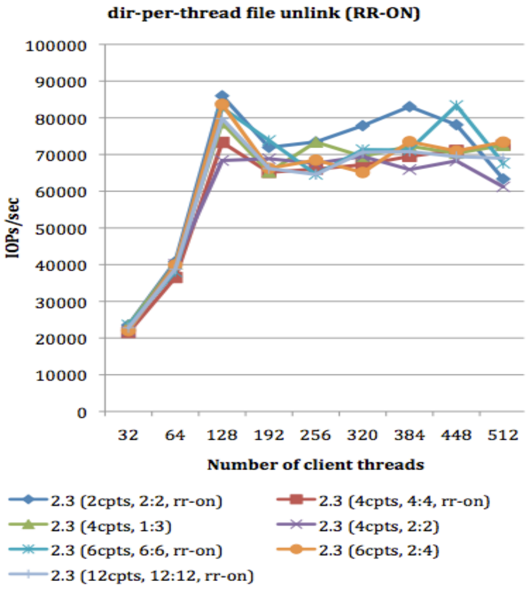
Performance of different CPT configurations

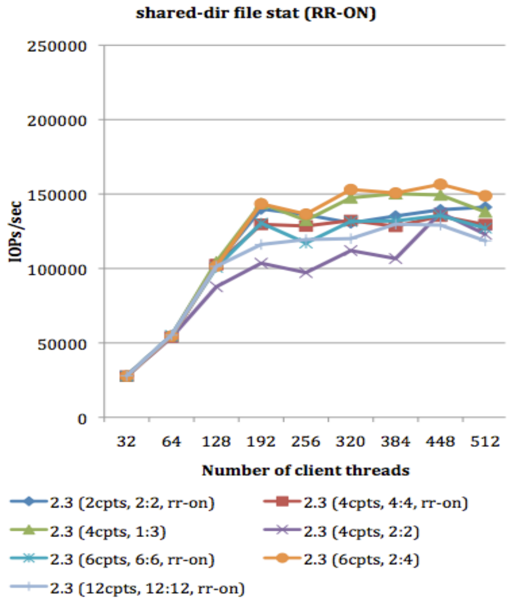

dir-per-thread opencreate + close (RR-OFF)


dir-per-thread file unlink (RR-OFF)


dir-per-thread file stat (RR-OFF)


dir-per-thread opencreate + close (RR-ON)


dir-per-thread file unlink (RR-ON)


shared-dir file stat (RR-ON)

Lustre SMP configurations (libcfs)

- Many chip types
  - Server-1: Dual-core CPU, 8 sockets
  - Server-2: 50 cores, 1 socket
  - Server-3: 4 sockets, 2 NUMA nodes
  - Server-4: 2 sockets, 4 NUMA nodes
- Default
  - Preferred value “N”
    - 2 * (N - 1)^2 < NCPUS <= 2 * N^2
  - Adjust “N” based on number of sockets or NUMA nodes
- Configure CPU partitions for libcfs
  - Libcfs cpu_npartitions=NUMBER
    - Prefer to put siblings in same CPT
  - Libcfs cpu_pattern=STRING_PATTERN
    - Example: libcfs cpu_pattern=“0[0-6/2] 1[1-7/2]”
    - Example: libcfs cpu_parttern=“N 0[0,2] 1[1,3]”

Lustre SMP configurations (LNet)

- NID affinity
  - Hash NID by default
  - Bind NI on CPTs
    - O2ib0(ib0)[0, 1], tcp(eth0)[2, 3]
- Credits
  - NI credits
  - Router buffer credits
- Portal Round-Robin
  - /proc/sys/lnet/portal_rotor
- LND threads number
  - Decrease default threads number
  - Add extra threads for multiple interfaces

Lustre SMP configurations (Lustre server & client)

- Bind service on CPTs
  - Both for MDS and OSS
- Use-cases
  - 32 cores machine, 4 sockets
  - Default
    - 4 partitions, LNet and ptlrpc services can run on all partitions
  - Config-1, one IB interface MDS
    - Lnet networks="o2ib0(ib0)[0]"
    - Mdt mdt_num_cpts="[1,2,3]"
  - Config-2, user only want to run Lustre client on one socket.
    - Libcfs cpu_pattern="0[0-31/4]"
    - Need some changes to set affinity for client threads

Thank You