Many core clients and KNL scalability

Lustre Developer Summit 2016

Grégoire Pichon

22-09-2016

Bull
atos technologies

Test System

- Hardware
  - Bull Sequana X1210 blade (prototype)
    - 1 socket Intel Xeon Phi 000A @1.20GHz
      66 cores / 264 cpus
    - 192GB DDR4 memory
    - 16GB MCDRAM memory
    - 1 EDR Infiniband interface
    - socket mode: SNC-4 Sub-NUMA Clustering
- Software
  - Kernel 3.10.0-327.28.3.el7.x86_64
  - MOFED 3.3-OFED.3.3.1.0.0.1
  - Lustre 2.8.57

```
# numactl -H
available: 8 nodes (0-7)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 66 67 68 69 70 71
72 73 74 75 76 77 78 79 80 81 82 83 132 133 134 135 136 137 138 139 140
141 142 143 144 145 146 147 148 149 198 199 200 201 202 203 204 205
206 207 208 209 210 211 212 213 214 215
node 0 size: 48353 MB
node 0 free: 45433 MB
node 1 cpus: 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 84 85 86 87
88 89 90 91 92 93 94 95 96 97 98 99 150 151 152 153 154 155 156 157 158
159 160 161 162 163 164 165 216 217 218 219 220 221 222 223 224 225
226 227 228 229 230 231
node 1 size: 49152 MB
node 1 free: 47586 MB
node 2 cpus: 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 100 101 102
103 104 105 106 107 108 109 110 111 112 113 114 115 166 167 168 169
170 171 172 173 174 175 176 177 178 179 180 181 232 233 234 235 236
237 238 239 240 241 242 243 244 245 246 247
node 2 size: 49152 MB
node 2 free: 47758 MB
node 3 cpus: 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 116 117 118
119 120 121 122 123 124 125 126 127 128 129 130 131 182 183 184 185
186 187 188 189 190 191 192 193 194 195 196 197 248 249 250 251 252
253 254 255 256 257 258 259 260 261 262 263
node 3 size: 49152 MB
node 3 free: 47840 MB
node 4 cpus:
node 4 size: 4096 MB
node 4 free: 3955 MB
node 5 cpus:
node 5 size: 4096 MB
node 5 free: 3955 MB
node 6 cpus:
node 6 size: 4096 MB
node 6 free: 3955 MB
node 7 cpus:
node 7 size: 4096 MB
node 7 free: 3953 MB
…
```

Lustre Compute Partitions and Worker Threads

- Default configuration
  - 12 compute partitions, 22 cpus each (4 numa nodes: 72/64/64/64 cpus)
  - ~380 worker threads (ptlrpcd, ptlrpc_hr, kiblnd_sd, ldlm_cb, …)

  ⇒ does not fit the hardware architecture (thread quartets, pairs of core: tile)

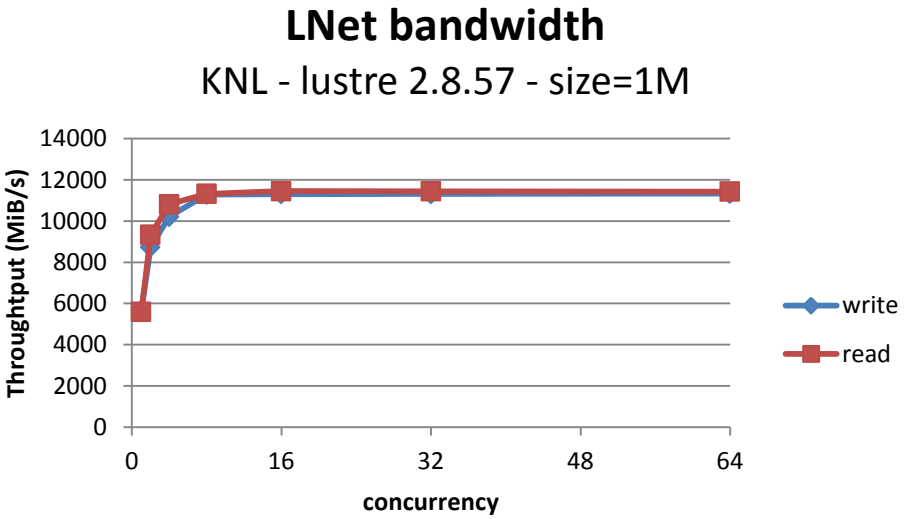
  ⇒ too much working threads

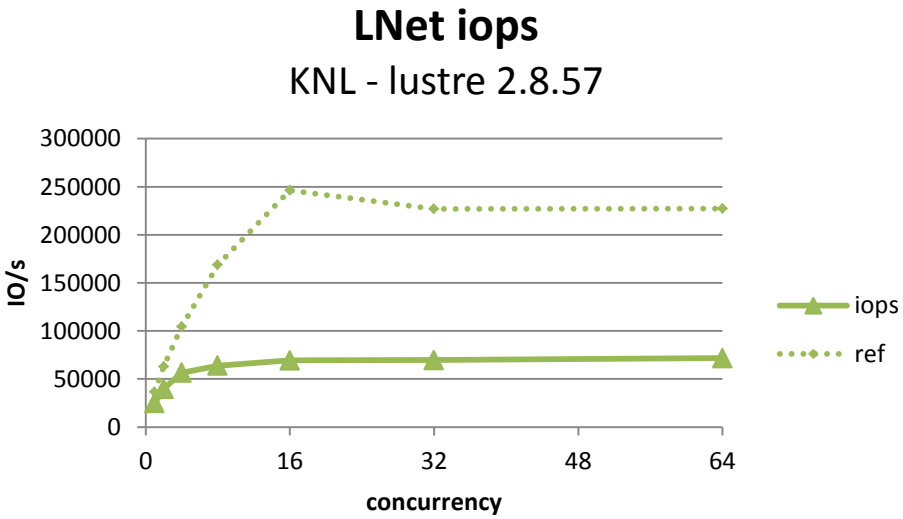
- Used configuration

  options libcfs cpu_pattern="0[0-15] 1[18-33] 2[34-49] 3[50-65]"
  - 4 compute partitions, 16 cpus each, 1 cpu per core
  - ~107 worker threads

LNet Performance

Lnet_selftest

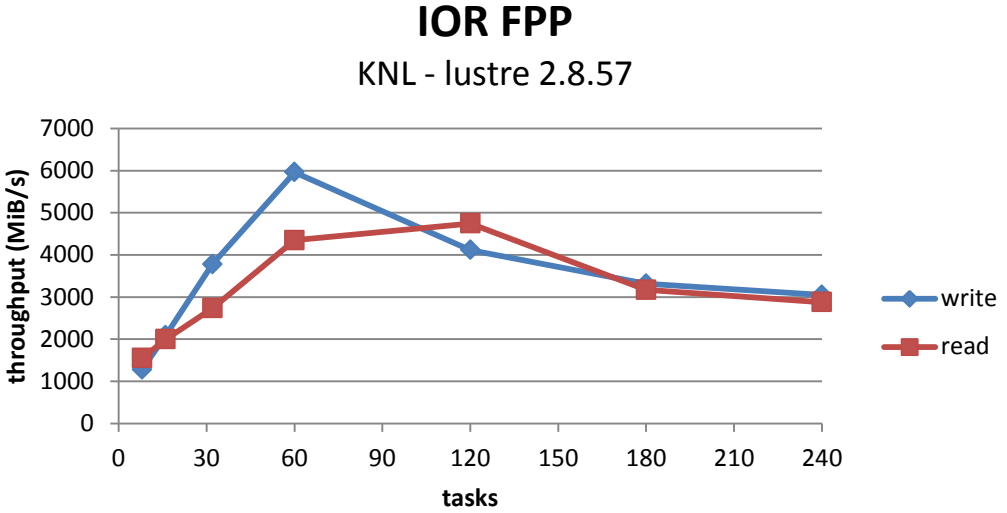


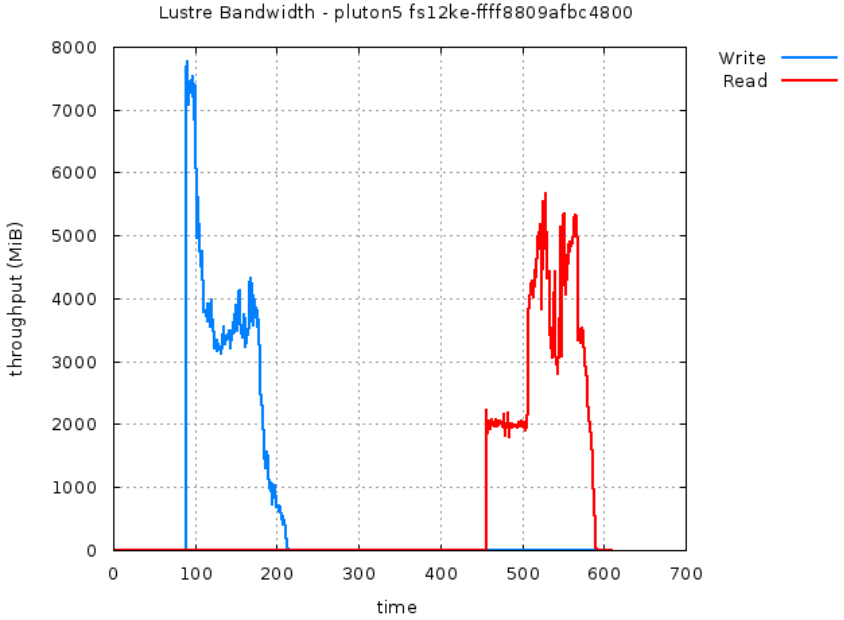


- Lustre Network performance on KNL platform
  - large data transfer bandwidth is OK
  - but small operations rate is 3 times slower

Large Sequential IOs – File Per Process

IOR





lustre bandwidth
IOR FPP 240 tasks

Large Sequential IOs – File Per Process

Profiling - 240 tasks

perf.report.ior.write.240tasks.lu2.8.57.txt

24.53%	IOR	[kernel.vmlinux]	[k] _raw_spin_lock_irqsave
1.56%	IOR	[kernel.vmlinux]	[k] copy_user_enhanced_fast_string
0.99%	IOR	[kernel.vmlinux]	[k] _raw_spin_lock
0.92%	IOR	[osc]	[k] osc_queue_async_io

perf.report.ior.read.240tasks.lu2.8.57.txt

39.31%	IOR	[osc]	[k] osc_lru_alloc
6.60%	IOR	[kernel.vmlinux]	[k] _raw_spin_lock_irqsave

- major contention on
  - page_zone(page)->lru_lock
- minor contention on
  - cl_object->coh_attr_guard
  - client_obd ->cl_loi_list_lock
  - osc_object->oo_lock
  - obd_import->imp_lock
- osc_lru_alloc() & osc_lru_reclaim() logic appears to be expensive