System benchmark results on KNL – STREAM and HPL.

By Garima Kochhar. December 2016. HPC Innovation Lab

The Intel Xeon Phi bootable processor (architecture codenamed "Knights Landing" – KNL) is ready for prime time. The HPC Innovation Lab has had access to a few engineering test units, and this blog presents the results of our initial benchmarking study. [We also published our results with Cryo-EM workloads on these systems, and that study is available here.]

The KNL processor is from the Intel Xeon Phi product line but is a ***bootable processor***, i.e., the system does not need another processor in it to power on, just the KNL. Unlike the Xeon Phi coprocessors or the NVIDIA K80 and P100 GPU cards that are housed in a system that has a Xeon processor as well, the KNL is the only processor in the server. This necessitates a new server board design and the PowerEdge C6320p is the Dell EMC platform that supports the KNL line of processors. A C6320p server includes support for one KNL processor and six DDR4 memory DIMMs. The network choices include Mellanox InfiniBand EDR, Intel Omni-Path, or choices of add-in 10GbE Ethernet adapters. The platform has the other standard components you'd expect from the PowerEdge line including a 1GbE LOM, iDRAC and systems management capabilities. Further information on C6320p is available here.

The KNL processor models include 16GB of on-package memory called ***MCDRAM***. The MCDRAM can be used in three modes – memory mode, cache mode or hybrid mode. The 16GB of MCDRAM is visible to the OS as addressable memory and must be addressed explicitly by the application when used in memory mode. In cache mode, the MCDRAM is used as the last level cache of the processor. And in hybrid mode, a portion of the MCDRAM is available as memory and the other portion is used as cache. The default setting is cache mode as this is expected to benefit most applications. This setting is configurable in the server BIOS.

The architecture of the KNL processor allows the processor cores + cache and home agent directory + memory to be organized into ***different clustering modes***. These modes are called all2all, quadrant and hemisphere, Sub-NUMA Clustering-2 and Sub-NUMA Clustering 4. They are described in this Intel article. The default setting in the Dell EMC BIOS is quadrant mode and can be changed in the Dell EMC BIOS. All tests below are with the quadrant mode.

The configuration of the systems used in this study is described in Table 1.

Table 1 - Test configuration

Server	12 * Dell EMC PowerEdge C6320p
Processor	Intel Xeon Phi 7230. 64 cores @ 1.3 GHz, AVX base 1.1 GHz.
Memory	96 GB at 2400 MT/s [16 GB * 6 DIMMS]
Interconnect	Intel Omni-Path and Mellanox EDR
Software	
Operating System	Red Hat Enterprise Linux 7.2
Compilers	Intel 2017, 17.0.0.098 Build 20160721

MPI	Intel MPI 5.1.3
XPPSL	Intel XPPSL 1.4.1

STREAM

The first check was to measure the memory bandwidth on the KNL system. To measure memory bandwidth to the MCDRAM, the system must be in “memory” mode. A snippet of the OS view when the system is in quadrant + memory mode is in Figure 1.

Note that the system presents two NUMA nodes. One NUMA node contains all the cores (64 cores * 4 logical siblings per physical core) and the 96 GB of DDR4 memory. The second NUMA node, node1, contains the 16GB of MCDRAM.

```
# numactl -H
available: 2 nodes (0-1)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76
77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101
102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139
140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158
159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177
178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196
197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215
216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234
235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253
254 255
node 0 size: 98199 MB
node 0 free: 94208 MB
node 1 cpus:
node 1 size: 16384 MB
node 1 free: 15910 MB
node distances:
node   0   1
  0:  10  31
  1:  31  10
```

Figure 1 – NUMA layout in quadrant+memory mode

On this system, the `dmidecode` command shows six DDR4 memory DIMMs, and eight 2GB MCDRAM memory chips that make up the 16GB MCDRAM.

STREAM Triad results to the MCDRAM on the Intel Xeon Phi 7230 measured between **474-487 GB/s** across 16 servers. The memory bandwidth to the DDR4 memory is between **83-85 GB/s**. This is expected performance for this processor model. This link includes information on running stream on KNL.

When the system has MCDRAM in cache mode, the STREAM binary used for DDR4 performance above reports memory bandwidth of 330-345 GB/s.

XPPSL includes a `micprun` utility that makes it easy to run this micro-benchmark on the MCDRAM. "`micprun -k stream -p <num cores>`" is the command to run a quick stream test and this will pick the MCDRAM (NUMA node1) automatically if available.

HPL

The KNL processor architecture supports AVX512 instructions. With two vector units per core, this allows the processor to execute 32 DP floating point operations per cycle. For the same core count and processor speed, this **doubles** the floating point capabilities of KNL when compared to Xeon v4 or v3 processors (Broadwell or Haswell) that can do only 16 FLOPS/cycle.

HPL performance on KNL is slightly better (up to 5%) with the MCDRAM in memory mode when compared to cache mode and when using the HPL binary packaged with Intel MKL. Therefore the tests below are with the system in quadrant+memory mode.

On our test systems, we measured between ***1.7 – 1.9 TFLOP/s HPL performance*** per server across 16 test servers. The `micprun` utility mentioned above is an easy way to run single server HPL tests. "`micprun -k hplinpack -p <problem size>`" is the command to run a quick HPL test. However for cluster-level tests, the Intel MKL HPL binary is best.

HPL cluster level performance over the Intel Omni-Path interconnect is plotted in Figure 2. These tests were run using the HPL binary that is packaged with Intel MKL. The results with InfiniBand EDR are similar.

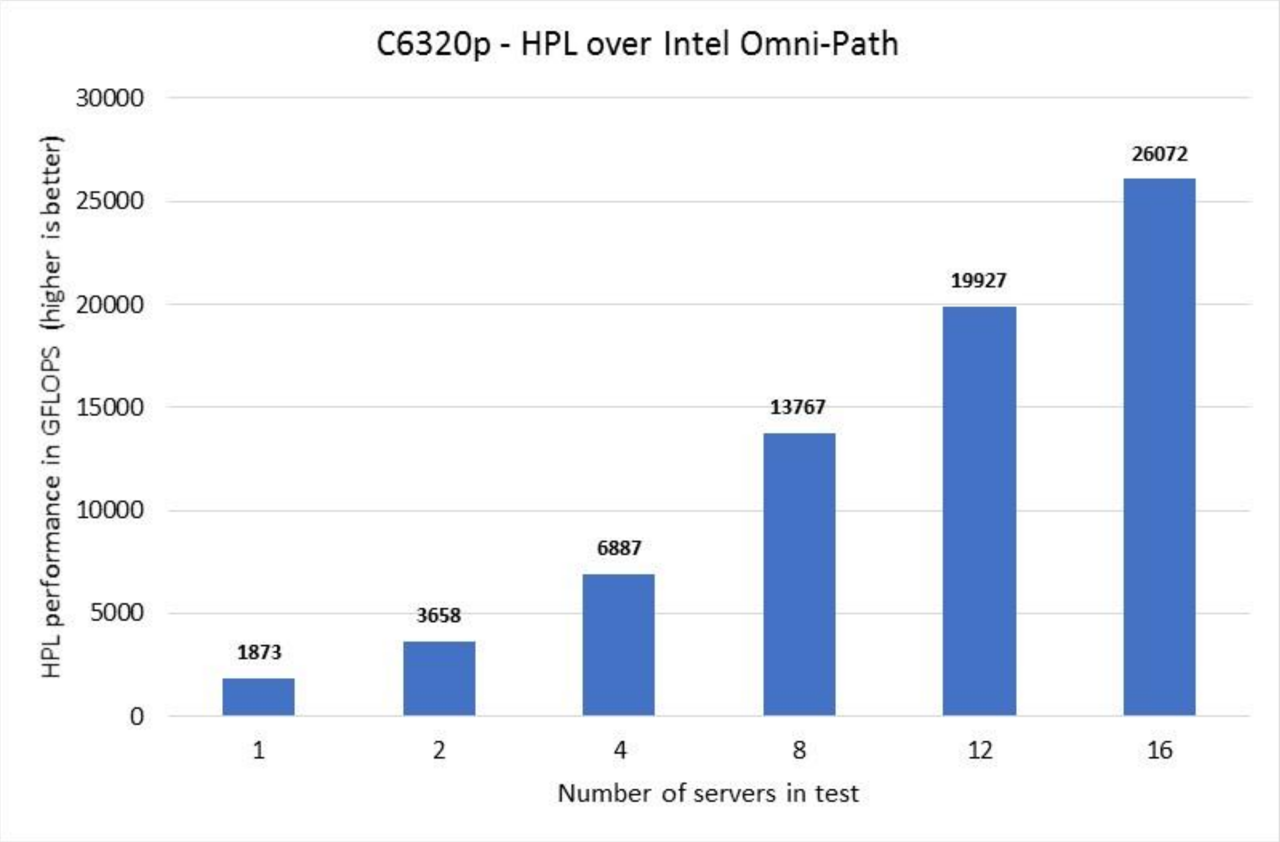

Figure 2 - HPL performance over Intel Omni-Path

Conclusion and Next Steps

The KNL-based system is a good platform for highly parallel vector applications. The on-package MCDRAM helps balance the enhanced processing capability with additional memory bandwidth to the applications. KNL introduces the AVX512 instruction set which further improves the performance of vector operations. The PowerEdge C6320p provides a complete HPC server with multiple network choices, disk configurations and systems management.

This blog presents initial system benchmark results. Look for upcoming studies with HPCG and applications like NAMD and Quantum Espresso. We have already published our results with Cryo-EM workloads on KNL and that study is available here.