High Performance Computing: Tools and Applications

Edmond Chow
School of Computational Science and Engineering
Georgia Institute of Technology

Lecture 21
KNL is a standalone processor
  ▶ (KNC is a coprocessor connected via PCIe)

Access to two types of memory: regular DRAM and MCDRAM
  ▶ MCDRAM = new, on package, high bandwidth multi-channel DRAM

AVX-512
  ▶ 8 mask registers for vector predication
  ▶ gather and scatter instructions

KNL has approx 3× the performance of KNC,
  ▶ e.g. peak 3 TFlops/s double precision

Binary compatible
cores are arranged in a tile architecture with a 2D mesh network (KNC has a 1D ring network), with 2 cores per tile

max 72 cores (36 tiles), but manufactured with 38 tiles
2 vector units per core
32 KB L1D cache
1 MB L2 cache shared between 2 cores
4 hyperthreads
out-of-order capability
2D mesh interconnect

- 700 GB/s aggregate bandwidth
- x-hops take 2 clocks; y-hops take 1 clock
Two types of memory

- **MCDRAM**
  - on package memory
  - high bandwidth
  - total 16 GB arranged in 8 devices
  - stream benchmark for MCDRAM: 450 GB/s

- **DDR**
  - off package
  - large capacity
  - two DDR controllers, 3 channels each
  - stream benchmark: 90 GB/s
MCDRAM latency greater than DDR latency at low loads but much less at high loads

Figure: Intel
Three modes (configured at boot)
- cache mode, i.e., direct mapped cache for DDR
- flat mode (same address space as DDR)
- hybrid mode
Flat/hybrid mode: How to use different types of memory

- memkind library
  - hbw_malloc – allocate high bandwidth memory
  - etc.
1. Memory access miss in local L2. Send message to tag directory on another tile.
2. Tag directory specifies cache line or miss. Send message to L2 on third tile, or to L3 memory controller (MCDRAM)
3. Access requested data in memory
4. Data is returned to original tile
Cluster modes (set at boot time via BIOS)

Cores, tag directories, and memory can be *clustered* to improve performance (shorten message paths on L2 miss)

- quadrant mode (default)
  - tag directory and data reside in the same quadrant as memory controller
  - UMA to MCDRAM (essentially)
  - UMA to DDR (essentially)
  - variation: hemisphere

- sub-NUMA clustering (SNC)
  - SNC-4: cores and memory are divided into 4 quadrants (could be appropriate if 4 MPI processes run on the KNL processor)
  - variation: SNC-2

- all-to-all
  - addresses hashed uniformly across memory
  - allows for irregular DIMM configurations
Checking the cluster mode

user@knl% sudo hwloc-dump-hwdata

Dumping KNL SMBIOS Memory-Side Cache information:

... Getting MCDRAM KNL info. Count=8 struct size=12
MCDRAM controller 0
Size = 2048 MB
MCDRAM controller 1
Size = 2048 MB

... Total MCDRAM 16384 MB
Cluster mode: SNC-4
Memory Mode: Flat
Flat Mode: No MCDRAM cache available, nothing to dump.
NUMA domains

- MCDRAM and DDR reside in separate locality domains
%% numactl -H
available: 1 nodes (0)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 ... 255
node 0 size: 65432 MB
node 0 free: 62887 MB
node distances:
node 0
 0: 10
### SNC-4 with FLAT (0-3: DDR nodes, 4-7 MCDRAM nodes)

<table>
<thead>
<tr>
<th>Node</th>
<th>CPUs</th>
<th>Size</th>
<th>Free</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 64 65 ... 207</td>
<td>16280 MB</td>
<td>15413 MB</td>
</tr>
<tr>
<td>1</td>
<td>16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 80 81 ... 223</td>
<td>16384 MB</td>
<td>15818 MB</td>
</tr>
<tr>
<td>2</td>
<td>32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 96 97 ... 239</td>
<td>16384 MB</td>
<td>15617 MB</td>
</tr>
<tr>
<td>3</td>
<td>48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 112 113 ... 255</td>
<td>16384 MB</td>
<td>15886 MB</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>4096 MB</td>
<td>3983 MB</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>4096 MB</td>
<td>3982 MB</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>4096 MB</td>
<td>3982 MB</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>4096 MB</td>
<td>3979 MB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Node</th>
<th>CPU Mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10 21 21 21 31 41 41 41</td>
</tr>
<tr>
<td>1</td>
<td>21 10 21 21 41 31 41 41</td>
</tr>
<tr>
<td>2</td>
<td>21 21 10 21 41 41 31 41</td>
</tr>
<tr>
<td>3</td>
<td>21 21 21 10 41 41 41 31</td>
</tr>
<tr>
<td>4</td>
<td>31 41 41 41 10 41 41 41</td>
</tr>
<tr>
<td>5</td>
<td>41 31 41 41 10 41 41 41</td>
</tr>
<tr>
<td>6</td>
<td>41 41 31 41 41 41 10 41</td>
</tr>
<tr>
<td>7</td>
<td>41 41 41 31 41 41 41 10</td>
</tr>
</tbody>
</table>
Where to allocate high bandwidth memory?

<table>
<thead>
<tr>
<th>node</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0:</td>
<td>10</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>31</td>
<td>41</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>1:</td>
<td>21</td>
<td>10</td>
<td>21</td>
<td>21</td>
<td>41</td>
<td>31</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>2:</td>
<td>21</td>
<td>21</td>
<td>10</td>
<td>21</td>
<td>41</td>
<td>41</td>
<td>31</td>
<td>41</td>
</tr>
<tr>
<td>3:</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>10</td>
<td>41</td>
<td>41</td>
<td>41</td>
<td>31</td>
</tr>
<tr>
<td>4:</td>
<td>31</td>
<td>41</td>
<td>41</td>
<td>41</td>
<td>10</td>
<td>41</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>5:</td>
<td>41</td>
<td>31</td>
<td>41</td>
<td>41</td>
<td>41</td>
<td>10</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>6:</td>
<td>41</td>
<td>41</td>
<td>31</td>
<td>41</td>
<td>41</td>
<td>41</td>
<td>10</td>
<td>41</td>
</tr>
<tr>
<td>7:</td>
<td>41</td>
<td>41</td>
<td>41</td>
<td>31</td>
<td>41</td>
<td>41</td>
<td>41</td>
<td>10</td>
</tr>
</tbody>
</table>

- node 0 is closest to MCDRAM node 4, etc.
Running an entire application in MCDRAM

If application will fit in 16 GB memory:

```
$ numactl -m 4,5,6,7 ./myApp
```
KNL as a coprocessor

- Flat mode only for MCDRAM
- No access to DDR (MCDRAM only)