High Performance Computing: Tools and Applications

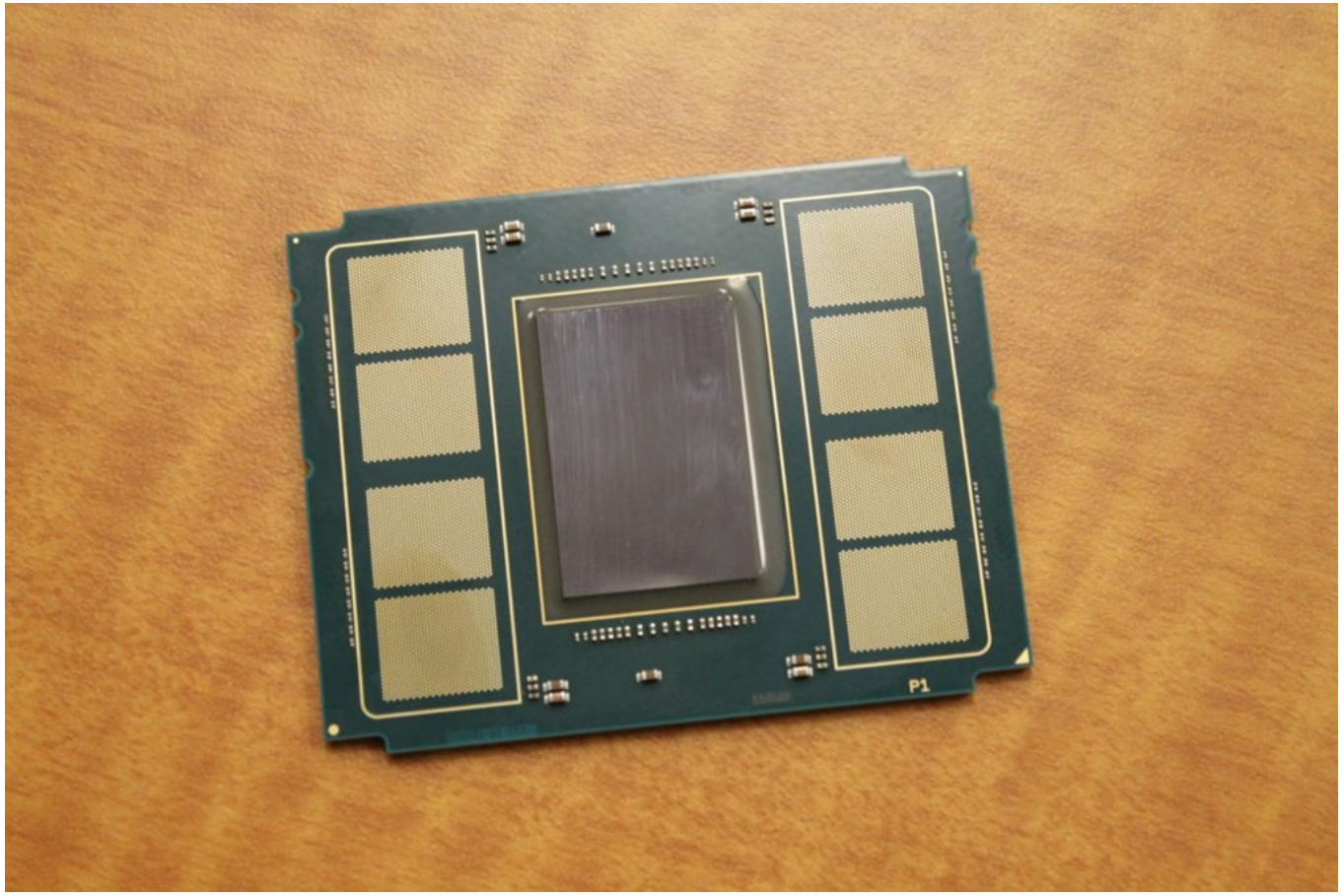
Edmond Chow
School of Computational Science and Engineering
Georgia Institute of Technology

Lecture 21

Intel MIC – Knights Landing (KNL)

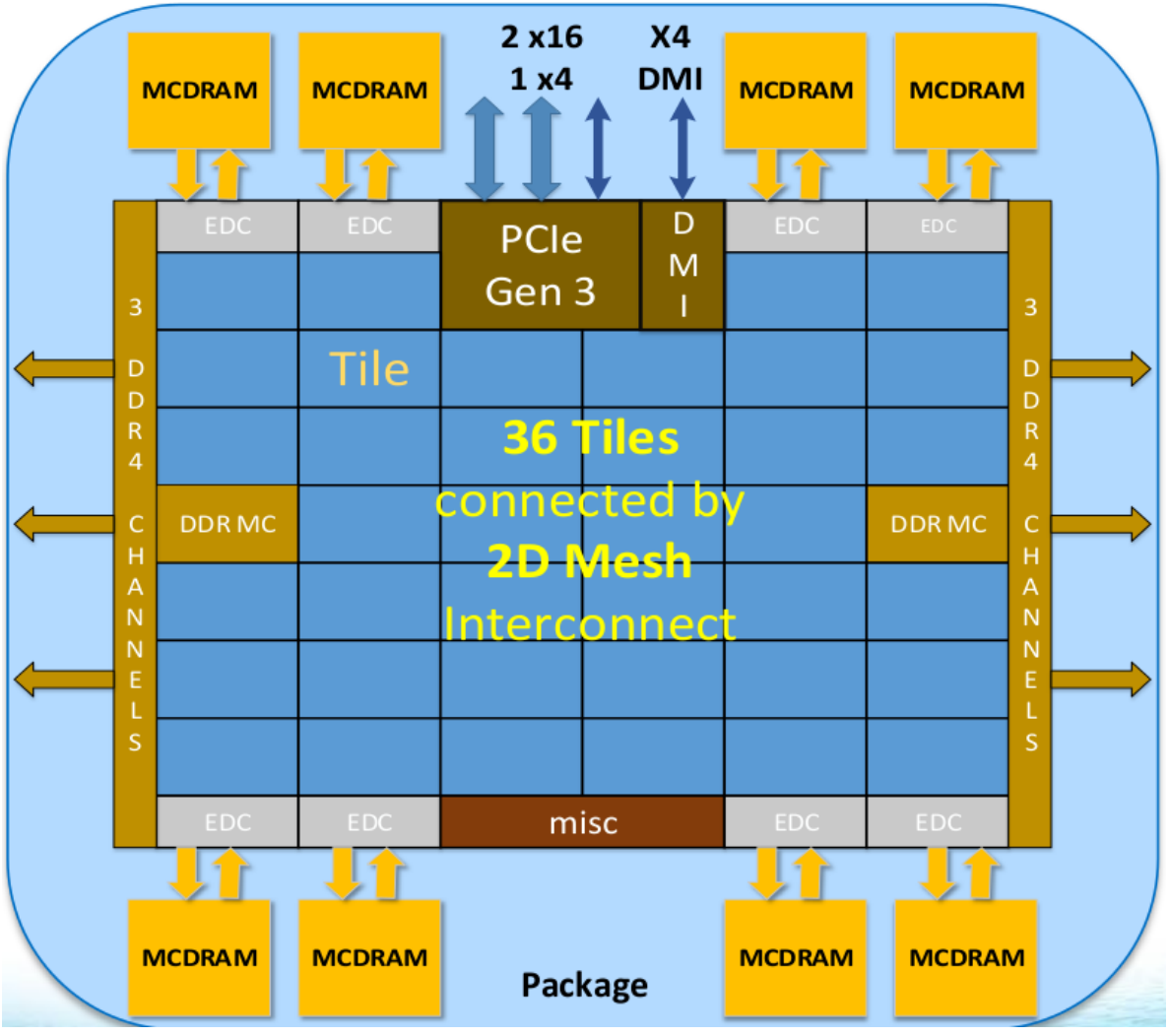
- KNL is a standalone processor
  - (KNC is a coprocessor connected via PCIe)
- Access to two types of memory: regular DRAM and MCDRAM
  - MCDRAM = new, on package, high bandwidth multi-channel DRAM
- AVX-512
  - 8 mask registers for vector predication
  - gather and scatter instructions
- KNL has approx 3$\times$ the performance of KNC,
  - e.g. peak 3 TFlops/s double precision
- Binary compatible

KNL



KNL

- 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



KNL cores

- 2 vector units per core
- 32 KB L1D cache
- 1 MB L2 cache shared between 2 cores
- 4 hyperthreads
- out-of-order capability

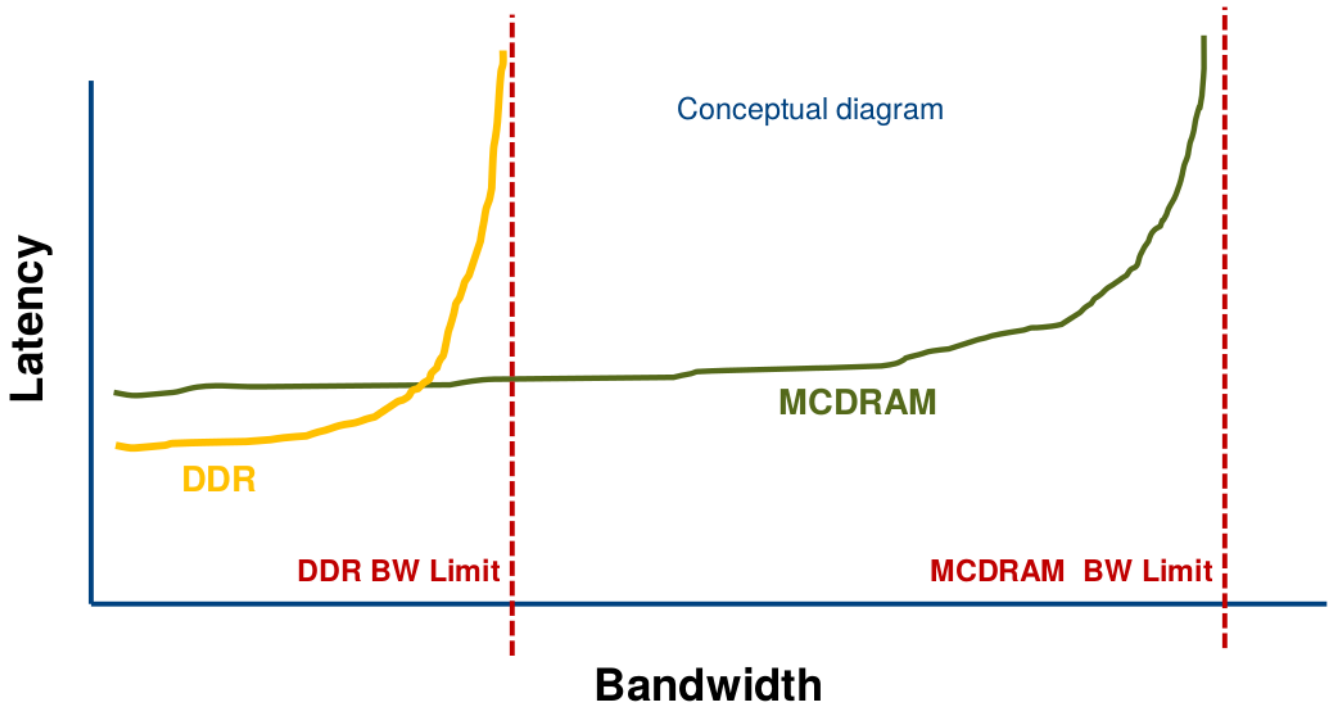
2D mesh interconnect

- 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

DDR vs. MCDRAM



MCDRAM latency greater than DDR latency at low loads but much less at high loads

Figure: Intel

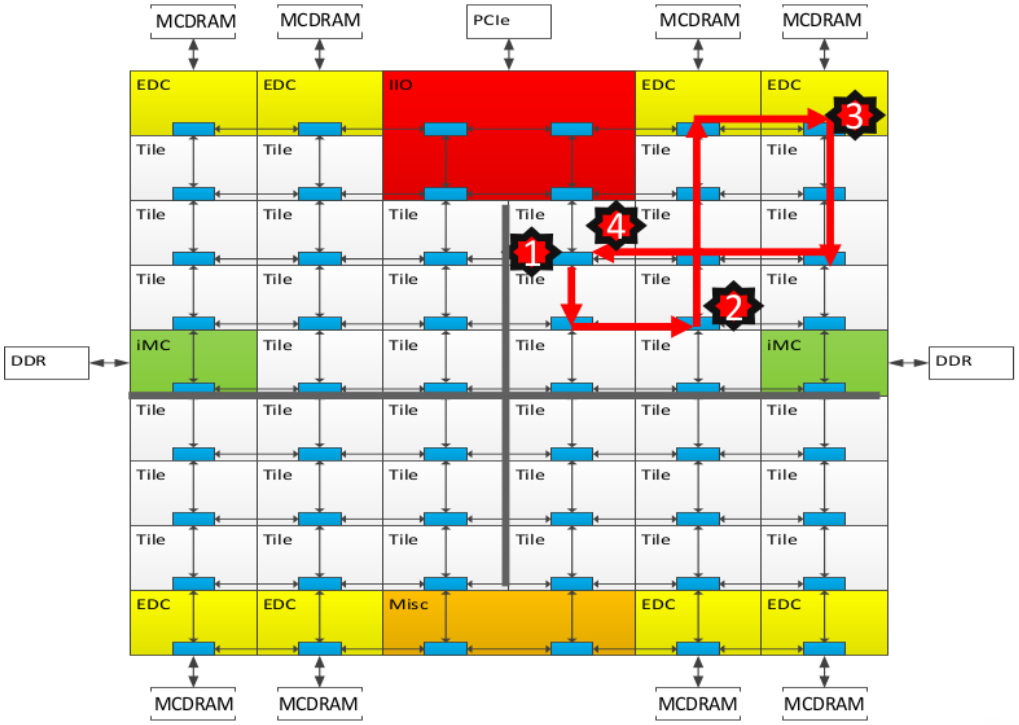
MCDRAM

- 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.

Memory access and cluster mode



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

QUADRANT with CACHE

```
%% 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)

```
available: 8 nodes (0-7)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 64 65 ... 207
node 0 size: 16280 MB
node 0 free: 15413 MB
node 1 cpus: 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 80 81 ... 223
node 1 size: 16384 MB
node 1 free: 15818 MB
node 2 cpus: 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 96 97 ... 239
node 2 size: 16384 MB
node 2 free: 15617 MB
node 3 cpus: 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 112 113 ... 255
node 3 size: 16384 MB
node 3 free: 15886 MB
node 4 cpus:
node 4 size: 4096 MB
node 4 free: 3983 MB
node 5 cpus:
node 5 size: 4096 MB
node 5 free: 3982 MB
node 6 cpus:
node 6 size: 4096 MB
node 6 free: 3982 MB
node 7 cpus:
node 7 size: 4096 MB
node 7 free: 3979 MB
node   0   1   2   3   4   5   6   7
  0:  10  21  21  21  31  41  41  41
  1:  21  10  21  21  41  31  41  41
  2:  21  21  10  21  41  41  31  41
  3:  21  21  21  10  41  41  41  31
  4:  31  41  41  41  10  41  41  41
  5:  41  31  41  41  41  10  41  41
  6:  41  41  31  41  41  41  10  41
  7:  41  41  41  31  41  41  41  10
```

Where to allocate high bandwidth memory?

```
node   0   1   2   3   4   5   6   7
  0:  10  21  21  21  31  41  41  41
  1:  21  10  21  21  41  31  41  41
  2:  21  21  10  21  41  41  31  41
  3:  21  21  21  10  41  41  41  31
  4:  31  41  41  41  10  41  41  41
  5:  41  31  41  41  41  10  41  41
  6:  41  41  31  41  41  41  10  41
  7:  41  41  41  31  41  41  41  10
```

- 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)