CLUSTER MODES

Adrian Jackson

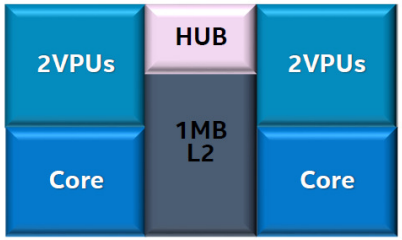
a.jackson@epcc.ed.ac.uk

@adrianjhpc

Some slides from Intel Presentations

Cache Coherency

KNL PROCESSOR TILE



CHA **C**aching/**H**ome **A**gent (or HUB)

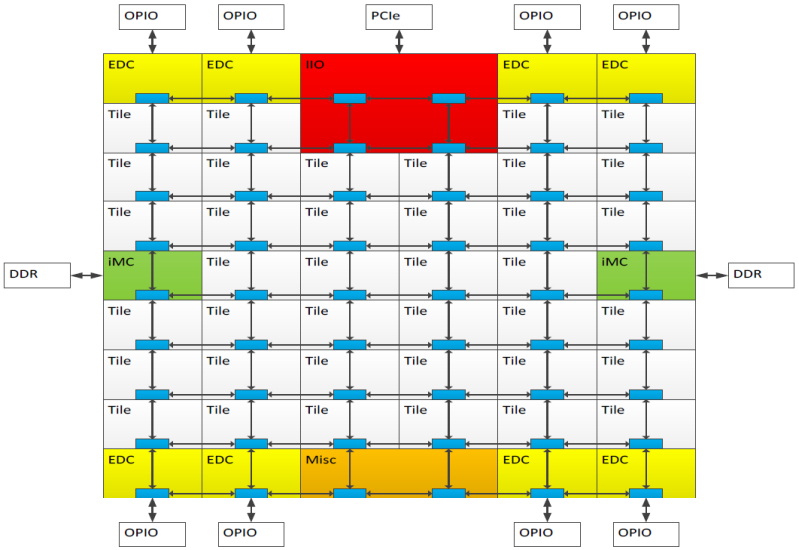
- 2D-Mesh connections for Tile
- Distributed Tag Directory to keep L2s coherent
- MESIF protocol

Cache coherency

- For memory loads/stores
  - Core (requestor) looks in local L2 cache
  - If not there it queries DTD for it:
    - Sends message to tile containing DTD (tag owner) entry for that memory address:
      - If it's not in any cache then data fetched from memory
        - DTD updates with requestor information
      - If it's in a tile's L2 cache then:
        - Tag owner sends message to tile where data is (resident)
        - Resident sends data to requestor

KNL

KNL MESH INTERCONNECT



Mesh of Rings

- Every row and column is a ring
- YX routing: Go in Y → Turn → Go in X
  - 1 cycle to go in Y, 2 cycles to go in X
- Messages arbitrate at injection and on turn

Mesh at fixed frequency of 1.7 GHz

Distributed Directory Coherence protocol

KNL supports Three Cluster Modes

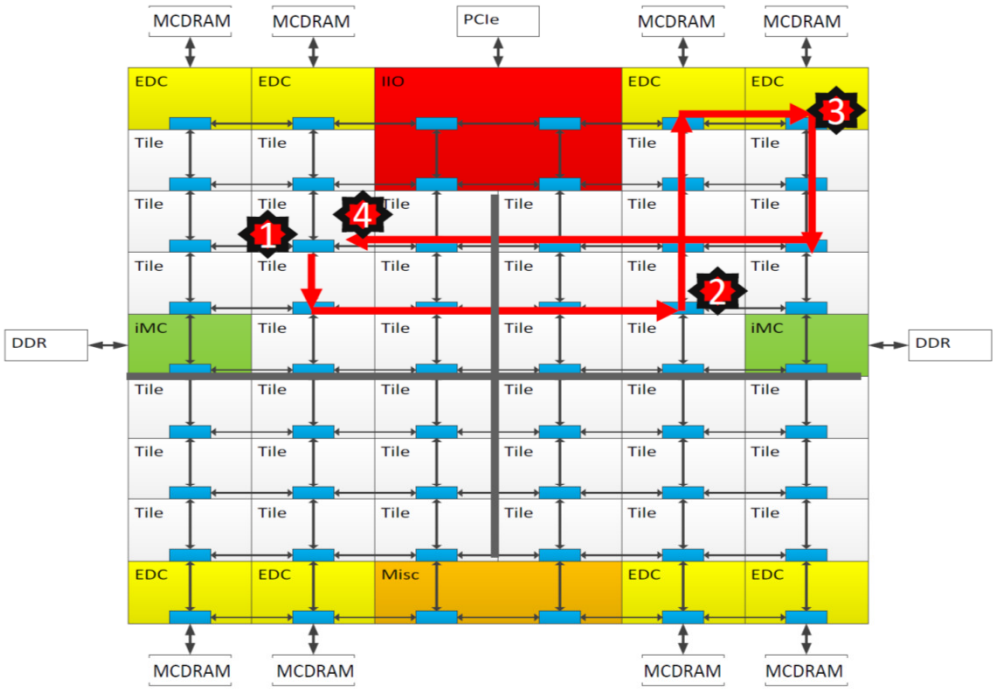
1) All-to-all
2) Quadrant
3) Sub-NUMA Clustering

Selection done at boot time.

intel

epcc | THE UNIVERSITY OF EDINBURGH

KNL
Cluster Mode: Quadrant



Chip divided into four virtual Quadrants

Address hashed to a Directory in the same quadrant as the Memory

Affinity between the Directory and Memory

Lower latency and higher BW than all-to-all. SW Transparent.

1) L2 miss, 2) Directory access, 3) Memory access, 4) Data return

Avinash Sodani CGO PPoPP HPCA Keynote 2016

epcc | THE UNIVERSITY OF EDINBURGH

Quadrant mode

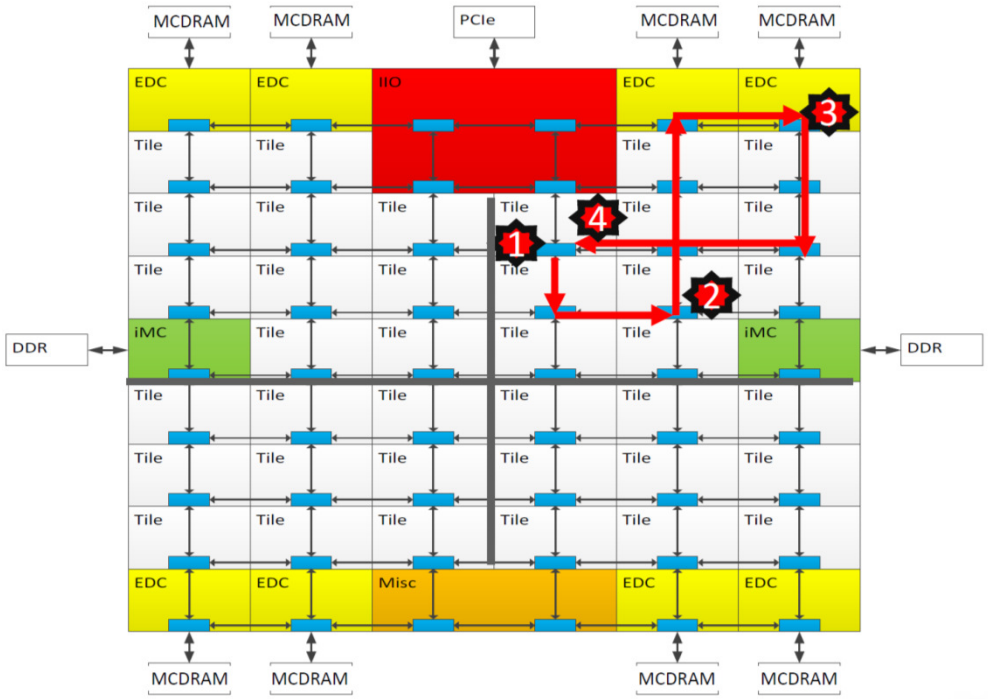
- One NUMA region for MCDRAM
- One NUMA region for main memory

KNL

If using only 1 MPI rank and OpenMP to fill up cores and also using SNC, have to enable all memory access, i.e.:

```
numactl -m 4,5,6,7
```

Cluster Mode: Sub-NUMA Clustering (SNC)



Each Quadrant (Cluster) exposed as a separate NUMA domain to OS.

Looks analogous to 4-Socket Xeon

Affinity between Tile, Directory and Memory

Local communication. Lowest latency of all modes.

SW needs to NUMA optimize to get benefit.

1) L2 miss, 2) Directory access, 3) Memory access, 4) Data return

Avinash Sodani CGO PPoPP HPCA Keynote 2016

epcc | THE UNIVERSITY OF EDINBURGH

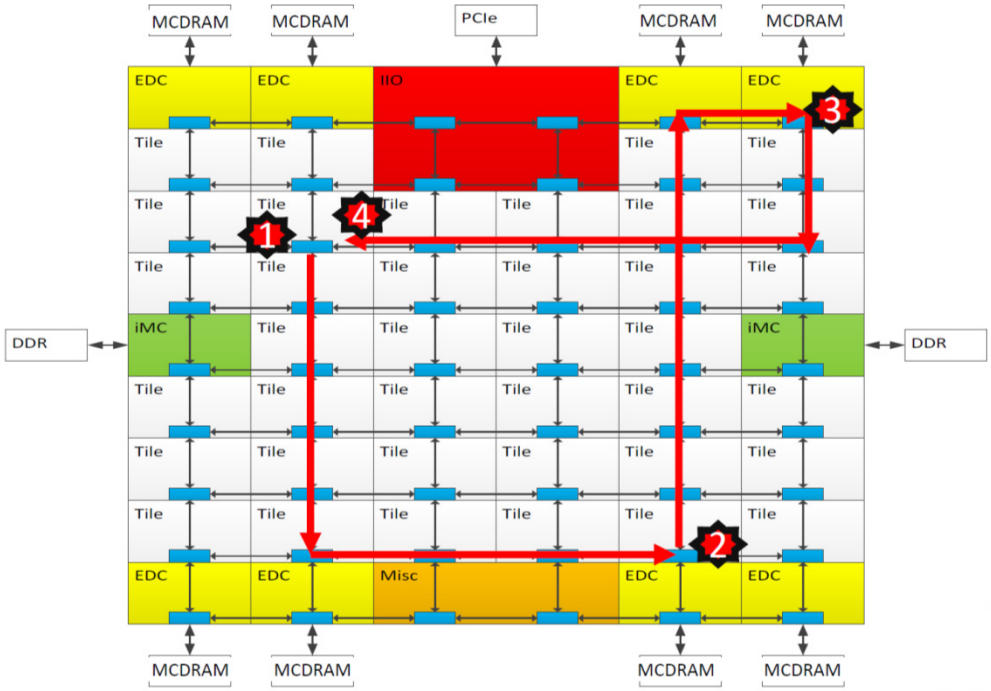
SNC-4

- Four NUMA regions for MCDRAM
- Four NUMA regions for main memory

KNL

Don’t use, fallback/for broken hardware mode

Cluster Mode: All-to-All



Address uniformly hashed across all distributed directories

No affinity between Tile, Directory and Memory

Most general mode. Lower performance than other modes.

Typical Read L2 miss

1. L2 miss encountered
2. Send request to the distributed directory
3. Miss in the directory. Forward to memory
4. Memory sends the data to the requestor

Cluster modes

- Cluster modes are really just part of the memory modes
- Two ones that may be of interest
  - Quadrant and SNC-4
  - Quadrant will always give reasonable performance
  - SNC-4 should give a bit better performance if code is properly NUMA aware
    - Will give worse performance if your code goes beyond the NUMA regions
    - May require careful pinning if running less processes than numa regions
- Ignore alltoall, hemisphere, SNC-2
- Changing either cluster mode or memory mode requires rebuild of tag directories
  - Requires reboot
  - Takes ~15-20 minutes