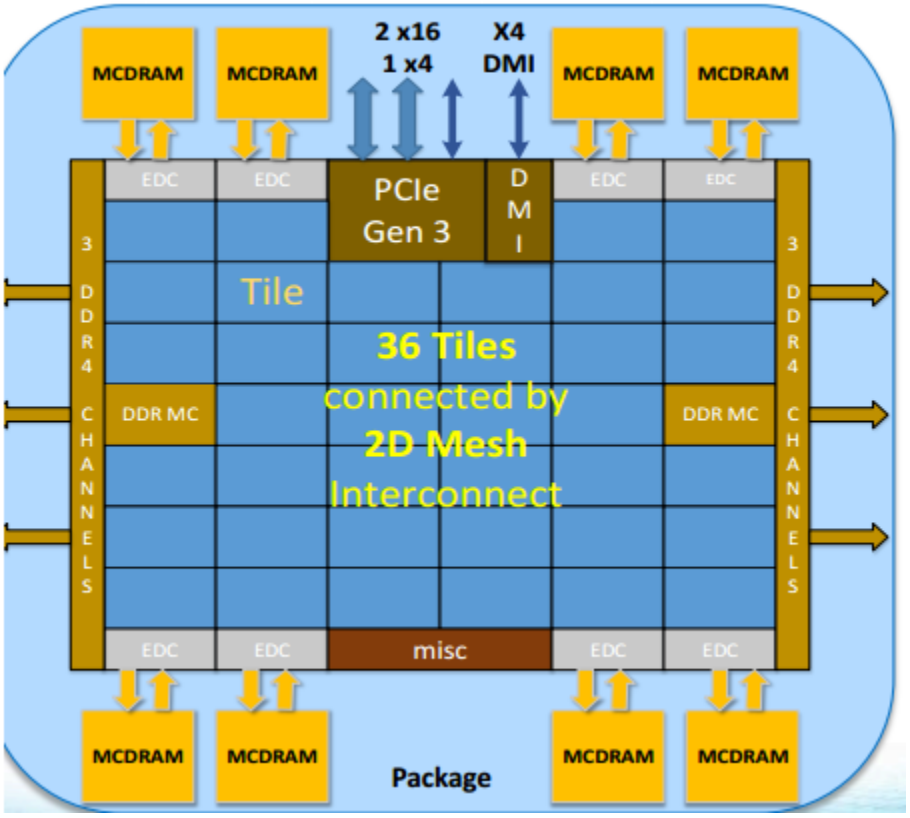
KNL Processor modes

The Xeon-Phi "Knights-Landing" 7250 processors in Marconi-knl have 68 cores, organized into 34 "tiles" (each tile comprising two cores and a shared 1MB L2 cache) placed in a 2D mesh, connected via an on-chip interconnect as shown in the following figure:



The KNL processor has 6 DDR channels, with controllers to the right and left of the mesh 8 MCDRAM channels, with controllers spread across 4 "corners" of the mesh.

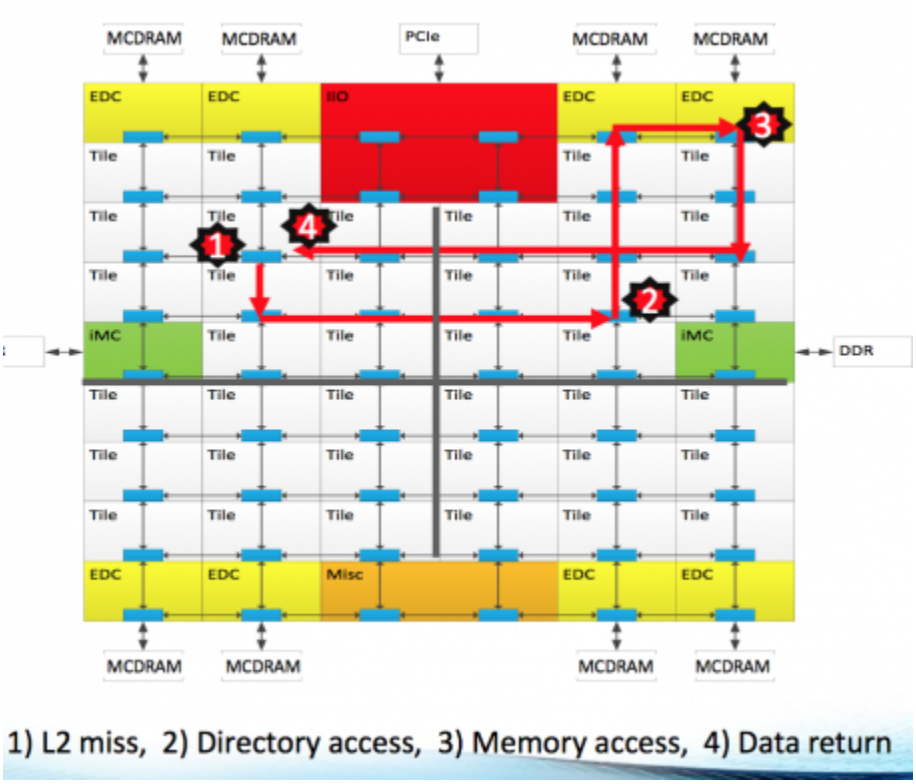
NUMA on KNL

NUMA stands for Non-Uniform Memory Access. It represents the situation where certain cores on a node can be considered "closer" to some part of the memory space than others or if some memory on the node has different latency or bandwidth to the cores. Both of these situations occur on the KNL nodes - there are two different types of memory available with different specs, and different channels of memory are closer to different cores in the 2d mesh.

NUMA Mode Options on KNL

The most useful NUMA modes on KNL are quadrant and sub-NUMA clustering (SNCx)

In quadrant mode the chip is divided into four virtual quadrants, but is exposed to the OS as a single NUMA domain. The diagram below illustrates the affinity of tag directory to the memory. In many cases, this mode is the easiest to use and will provide good performance.



In sub-NUMA mode each quadrant (or half) of the chip is exposed as a separate NUMA domain to the OS. In SNC4 (SNC2) mode the chip is analogous to a 4 (2) socket Xeon. This mode has potential for high performance, but software must be optimized for NUMA architectures to benefit. On our KNL node, with 68 cores, SNC4 can be complicated to fully utilize as there are non-homogeneous number of cores per quadrant (because the 34 tiles cannot be evenly distributed among 4 quadrants).

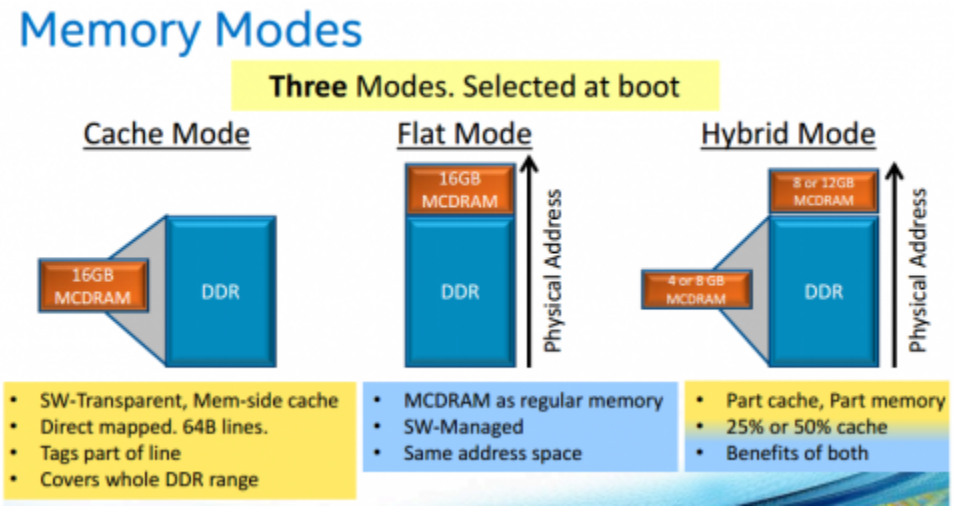
On Marconi, we have enabled only the quadrant mode configuration on production nodes. Other modes may be enabled in the future or on request.

MCDRAM Memory Options on KNL

There is no shared L3 cache on the KNL processor. However, the 16 GB of MCDRAM (spread over 8 channels) can be configured either as a direct-mapped cache or as addressable memory. When configured as a cache, recently accessed data is automatically stored in cache, similarly to an L3 cache on a Xeon processor. However, there are some notable differences:

- The cache (16GB) is significantly larger than a typical L3 cache on a Xeon processor (usually in the tens of MB).
- The cache is direct-mapped. Meaning it is non-associative - each cache-line worth of data in DRAM has one location it can be cached in MCDRAM. This can lead to possible conflicts for apps with greater than 16GB working sets.
- Data is not prefetched into the MCDRAM cache

The MCDRAM may be configured either as a cache, addressable memory or as a mix of the two. This is shown in the figure below:



When the MCDRAM is configured at least in part as addressable memory, it is presented to the operating system and the user as an additional NUMA domain (quadrant mode) as described above.

On Marconi, we have enabled only the cache mode configuration for the MCDRAM on production nodes. Other modes may be enabled in the future or on request.