Using the CoolMUC-3 Cluster at LRZ

Dr. Volker Weinberg

PRACE Workshop: VI-HPS Tuning Workshop, LRZ, 23.4.- 27.4.2018

MEGWARE

First self-assembled Linux cluster (1999-2002)

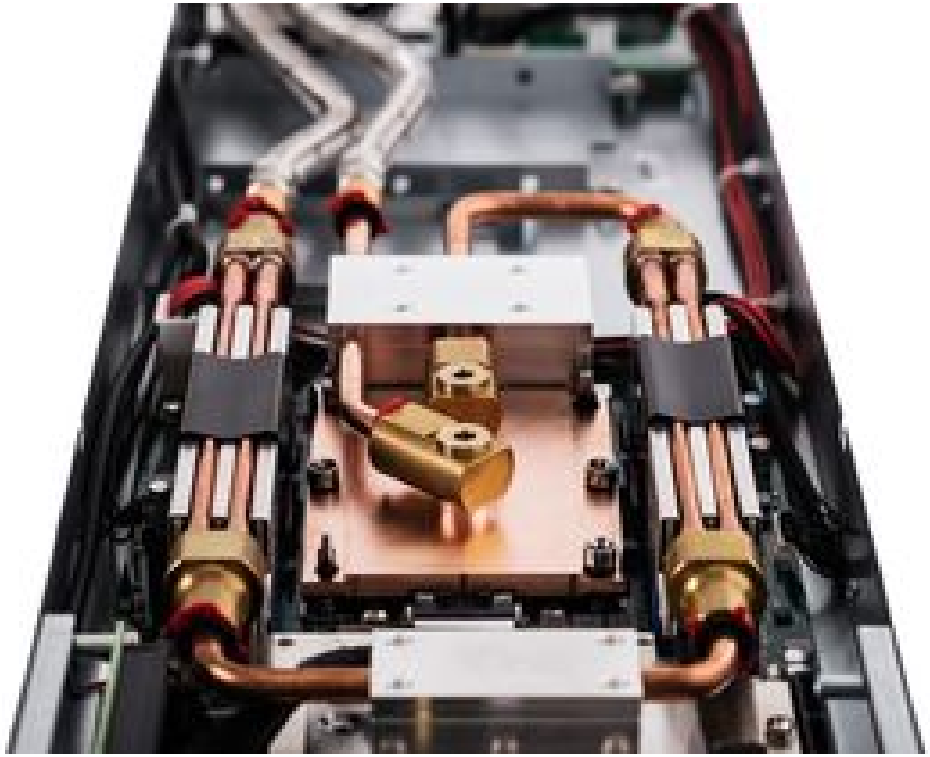
Cluster components (2012)

SGI UltraViolet with air guides in front to improve cooling efficiency (2012)

CoolMUC-2 (2015): The six racks to the left

CoolMUC-3 (2017): Really Cool

CoolMUC-3 (2017): Really Cool



Linux Cluster System Overview

Architecture				Total Numbers		Max Job Limits				How to get access	
System Name	CPU	Cores per Node	RAM per Node [GB]	Nodes	Cores	Nodes	Cores	Wall Time	Max. aggreg. RAM	Queue	Login Node (job submission)
Linux-Cluster CoolMUC-2	Intel Xeon E5-2697 v3 ("Haswell")	28	64	384	10752	60	1680	48h	3.8 TB	mpp2	lxlogin5.lrz.de, lxlogin6.lrz.de, lxlogin7.lrz.de
Linux-Cluster CoolMUC-2	Intel Xeon E5-2697 v3 ("Haswell")	28	64	1	28	1	28	96h	64 GB	serial	lxlogin5.lrz.de, lxlogin6.lrz.de, lxlogin7.lrz.de
Linux-Cluster Hugemem	Intel Xeon E5-2660 v2 ("Sandy Bridge")	20	240	7	220	1	20	168h	240 GB	hugemem	lxlogin5.lrz.de, lxlogin6.lrz.de, lxlogin7.lrz.de
Linux-Cluster Teramem	Intel Xeon E7-8890 v4	96	6144	1	96	1	96	48h	6.1 TB	inter	lxlogin5.lrz.de, lxlogin6.lrz.de, lxlogin7.lrz.de
Linux Cluster Many Core CoolMUC-3	Intel Xeon Phi (Knights Landing)	64	RAM:96 HBM:16	148	64	32	64	24h	RAM:2880GB HBM:512 GB	mpp3	lxlogin8.lrz.de

CoolMUC-3 System Overview

- Consists of **148** computational many-core Intel "**Knight's Landing (KNL)**" nodes (Xeon Phi 7210-F hosts).
- Connected to each other via an **Intel Omnipath high performance network** in fat tree topology.
- Theoretical **peak performance** is **400 TFlop/s** and the **LINPACK performance** of the complete system is **255 TFlop/s.**
- Standard Intel Xeon login node for development work and job submission.

CoolMUC-3 System Overview

- CoolMUC-3 comprises of three **warm-water cooled racks**, using an inlet temperature of at least 40 C.
- Deployment of **liquid-cooled power supplies and Omni-Path switches**.
- **Thermal isolation of the racks** to suppress radiative losses.
- Liquid cooled racks operate **entirely without fans**.
- A complementary rack for aircooled components (e.g. management servers) uses less than 3% of the systems total power budget.
- With 4.96 GFlops/Watt (according to the strict Green500 level-3 measurement methodology) **CoolMUC-3 is one of the most efficient x86 systems worldwide**.
- Result of an established partnership between LRZ and Megware.

CoolMUC-3 System Overview

Hardware	
Number of nodes	148
Cores per node	64
Hyperthreads per core	4
Core nominal frequency	1.3 GHz
Memory (DDR4) per node	96 GB (Bandwidth 80.8 GB/s)
High Bandwidth Memory per node	16 GB (Bandwidth 460 GB/s)
Bandwidth to interconnect per node	25 GB/s (2 Links)
Number of Omnipath switches (100SWE48)	10 + 4 (each 48 Ports)
Bisection bandwidth of interconnect	1.6 TB/s
Latency of interconnect	2.3 µs
Peak performance of system	394 TFlop/s
Infrastructure	
Electric power of fully loaded system	62 kVA
Percentage of waste heat to warm water	97%
Inlet temperature range for water cooling	30 … 50 °C
Temperature difference between outlet and inlet	4 … 6 °C
Software (OS and development environment)	
Operating system	SLES12 SP2 Linux
MPI	Intel MPI 2017, alternativ OpenMPI
Compilers	Intel icc, icpc, ifort 2017
Performance libraries	MKL, TBB, IPP, DAAL
Tools for performance and correctness analysis	Intel Cluster Tools

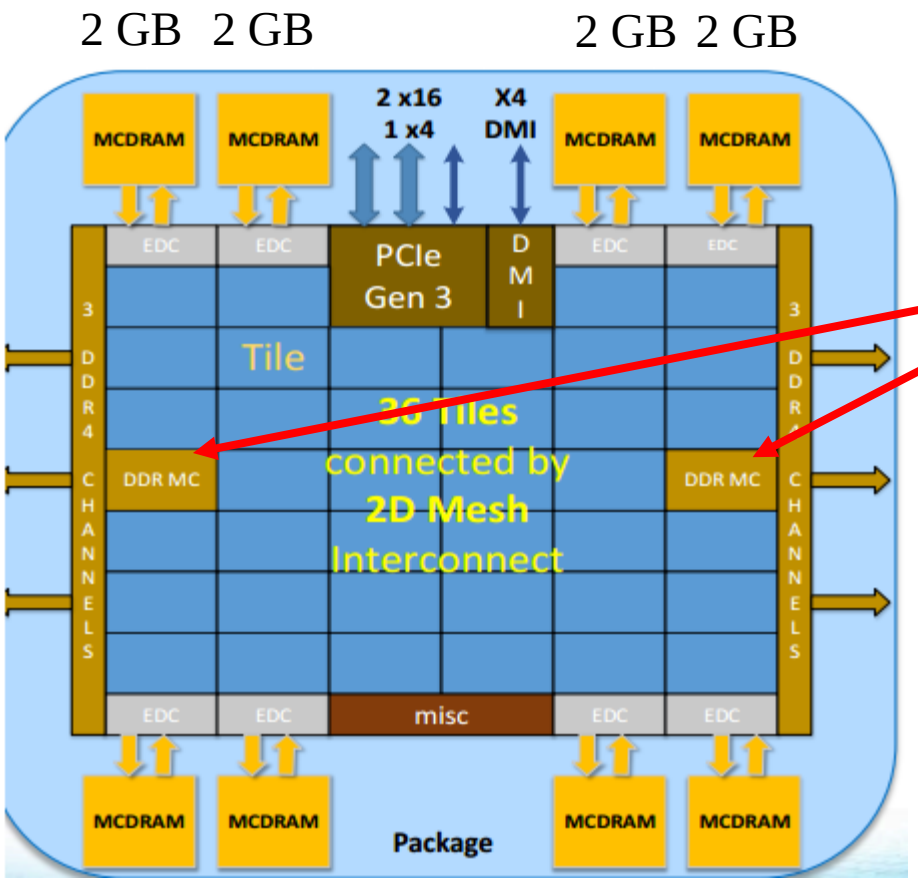
Knights Corner vs. KNL

- **Intel Many Integrated Cores (MIC**) is the code name for Intel's range of manycore CPUs

- **Intel Xeon Phi code-named Knights Corner (KNC)**
  - 1st generation of Xeon Phi 2012 – LRZ SuperMIC
  - Coprocessor
  - supporting 512 bit vectors
  - IMCI Instruction set

- **Intel Xeon Phi code-named Knights Landing (KNL)**
  - 2nd generation of Xeon Phi 2016 – LRZ CoolMUC-3
  - Processor
  - supporting 512 bit vectors
  - Intel AVX-512 Instruction set

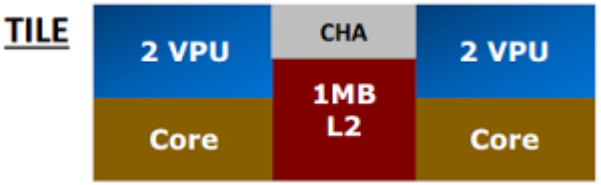
KNL Block Diagram



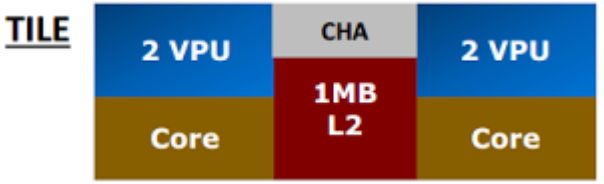
Cores are grouped in pairs (tiles)

- 36 possible tiles
- 2D mesh interconnect
- 2 DDR memory controllers
  - 6 channels DDR4
  - Up to 90 GB/s
- 16 GB MCDRAM
  - Up to 475 GB/s

CHA: Caching Home Agent



Tile



- Basic unit for replication
- Each tile consists of **2 cores**, **2 vector-processing units (VPU)** per core, a **1 MB L2 Cache** shared between the 2 cores
- **CHA** (caching/home agent)
  - Serves as the point where the **tile connects to the mesh**
  - Holds a portion of the **distributed tag directory structure**

KNL Memory: Overview

- **Memory hierarchy on KNL**:
  - DDR4 (96 GB)
  - MCDRAM (16 GB)
  - Tile L2 (1 MB)
  - Core L1 (32 KB)
- **Tile**: set of 2 cores sharing a 1MB L2 cache and connectivity on the mesh
- **Quadrant/Hemisphere**: virtual concept, not a hardware property. Way to divide the tiles at a logical level.
- **Tag Directory**: tracks cache line locations in all L2 caches. It provides the block of data or (if not available in L2) a memory address to the memory controller.

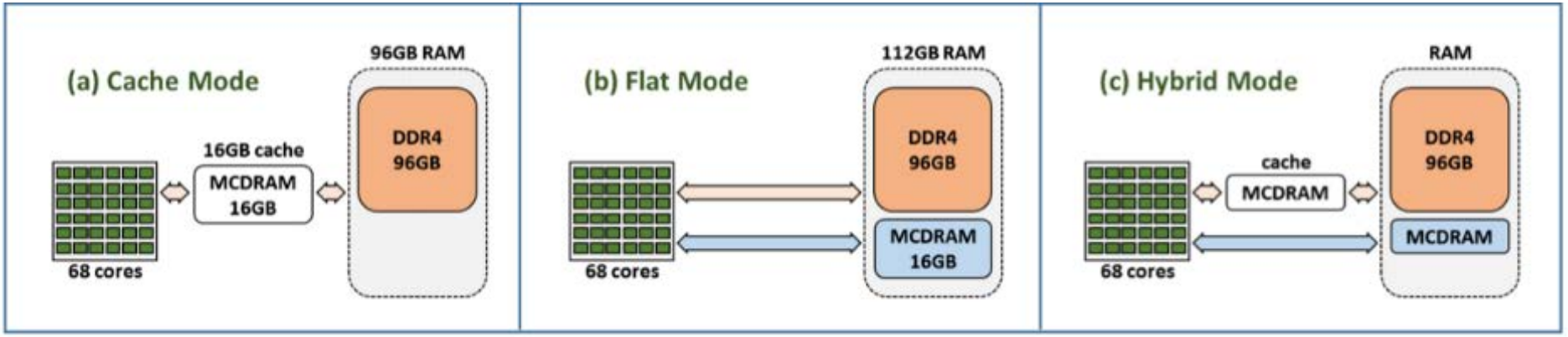
KNL Memory: MCDRAM

- **High-bandwidth** memory integrated **on-package**
- **8 MCDRAM** devices on KNL, each with 2 GB capacity -> **total 16 GB**
- Connected to EDC memory controller via **proprietary on-package I/O: OPIO**
- Each device has a **separate read and write bus** connecting it to its EDC (**E**mbedded **D**RAM **C**ontroller)
- **Aggregate Stream Triads Bandwidth** for the 8 MCDRAMS is **over 450 GB/s**
- Slighter higher latency than main memory (~10% slower)

KNL Memory: DDR4

- **High-capacity** memory **off-package**
- KNL has direct access to all of main memory
- **2 DDR4 memory controllers** on opposite sides of the chip, each controlling **3 DDR4 channels**
- Maximum total capacity is 384 GB
- **Aggregate Stream Triads Bandwidth** from all 6 DDR4 channels is around **90 GB/s**

KNL Memory Modes

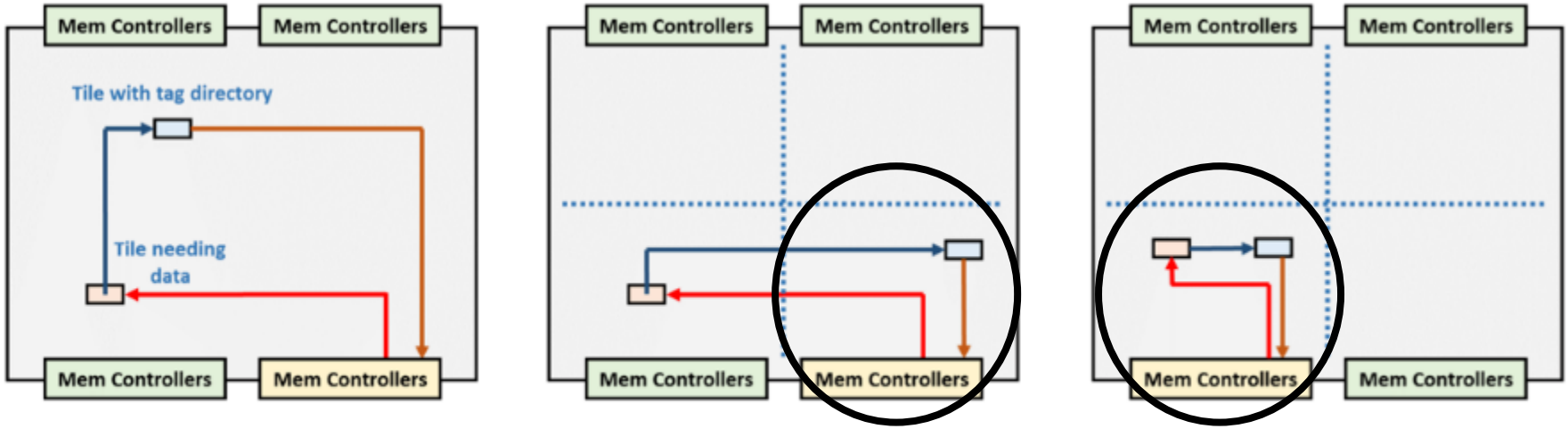


Cluster Modes

- **Cluster Modes modify the distance that L2 coherency traffic flows go through the mesh**
- 5 Cluster Modes supported:
  - All-to-all
  - Quadrant / Hemisphere
  - 2 **Sub-NUMA Cluster** modes: SNC-4 / SNC-2
- Regardless of the cluster mode selected, **all memory** (all MCDRAM and all DDR4) **is available to all cores**, and **all memory is fully cache-coherent.**
- What differs between the modes is whether the view of MCDRAM or DDR is **UMA** (Uniform Memory Access) or **NUMA** .

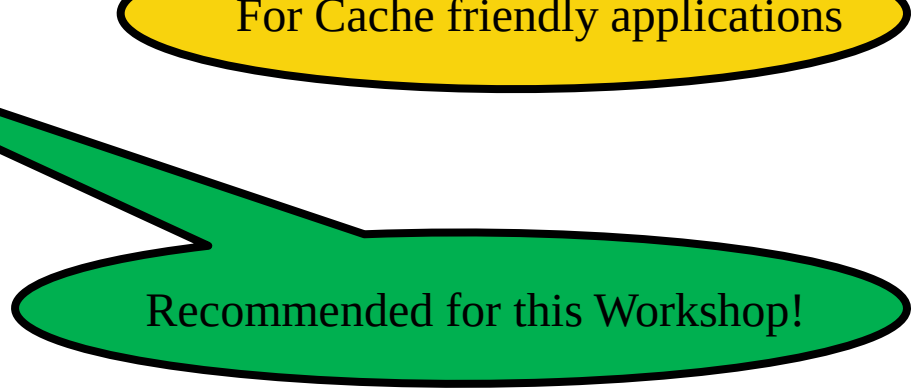
Cluster Modes: Overview

Cluster modes modify the distance that coherency traffic flows through mesh!



Cluster-Modes & Memory Modes Combinations

- 5 Flat Memory Mode Variants:
  - Flat-A2A
  - Flat-Quadrant
  - Flat-Hemisphere
  - Flat-SNC4
  - Flat-SNC2
- 5 Cache Memory Mode Variants
  - Cache-A2A
  - **Cache-Quadrant**
  - Cache-Hemisphere
  - Cache-SNC4
  - Cache-SNC2
- 5 x 3 = 15 Hybrid Variants



Using numactl

- use only DDR (default)
  `numactl --membind=0 ./a.out`
- use only MCDRAM in flat-quadrant mode
  `numactl --membind=1 ./a.out`
- use MCDRAM if possible in flat-quadrant mode; else DDR
  `numactl --preferred=1 ./a.out`

- show numactl settings
  `numactl --hardware`
- list available numactl options
  `numactl --help`

Compilation for KNL

- For reasonable optimization including SIMD vectorization for the KNL compute nodes, use options
  `-O3 -xmic-avx512`
- To optimise both for Broadwell (AVX2) and KNL (AVX512) use options
  `-xcore-avx2 -axmic-avx512`

Linux Cluster System Access

ssh -Y lxlogin5.lrz.de -l xxyyyzz	Haswell (CoolMUC-2) login node
ssh -Y lxlogin6.lrz.de -l xxyyyzz	Haswell (CoolMUC-2) login node
ssh -Y lxlogin7.lrz.de -l xxyyyzz	Haswell (CoolMUC-2) login node
gsissh -Y lxgt2.lrz.de	login node for Gsi-SSH
ssh -Y lxlogin8.lrz.de -l xxyyyzz	KNL Cluster (CoolMUC-3 login node)

Linux Cluster System Documentation

- Linux-Cluster:
  - https://www.lrz.de/services/compute/linux-cluster/
- CoolMUC-3:
  - https://www.lrz.de/services/compute/linux-cluster/coolmuc3/initial_operation/
  - https://www.lrz.de/services/compute/linux-cluster/coolmuc3/overview/

Linux Cluster System Documentation

- Using CoolMUC-3 for Training
- https://www.lrz.de/services/compute/courses/Using-CoolMUC-3/

https://goo.gl/pKJPwd

CoolMUC-3 SLURM

- Submit a job:
  `sbatch --reservation=TuningWorkshop job.sh`
- List own jobs:
  `squeue –u hpckurs??`
- Cancel jobs:
  `scancel jobid`
- Interactive Access:
- `salloc --nodes=1 --time=02:00:00`
- `--constraint=cache,quad`
- `--reservation=TuningWorkshop --partition=mpp3_batch`
- `srun --reservation=TuningWorkshop --pty bash`

CoolMUC-3 SLURM Simplest Batch File

```
-> /lrz/sys/courses/KNL/batch-cache-quad.sh

#!/bin/bash
#SBATCH -o /home/hpc/a2c06/hpckurs01/test.%j.%N.out
#SBATCH -D /home/hpc/a2c06/hpckurs01/
#SBATCH -J jobname
#SBATCH --clusters=mpp3
#SBATCH --get-user-env
#SBATCH --time=02:00:00
#SBATCH --constraint=cache,quad


commands
```

And now …

Enjoy the course!