150 Jahre
culture of
excellence

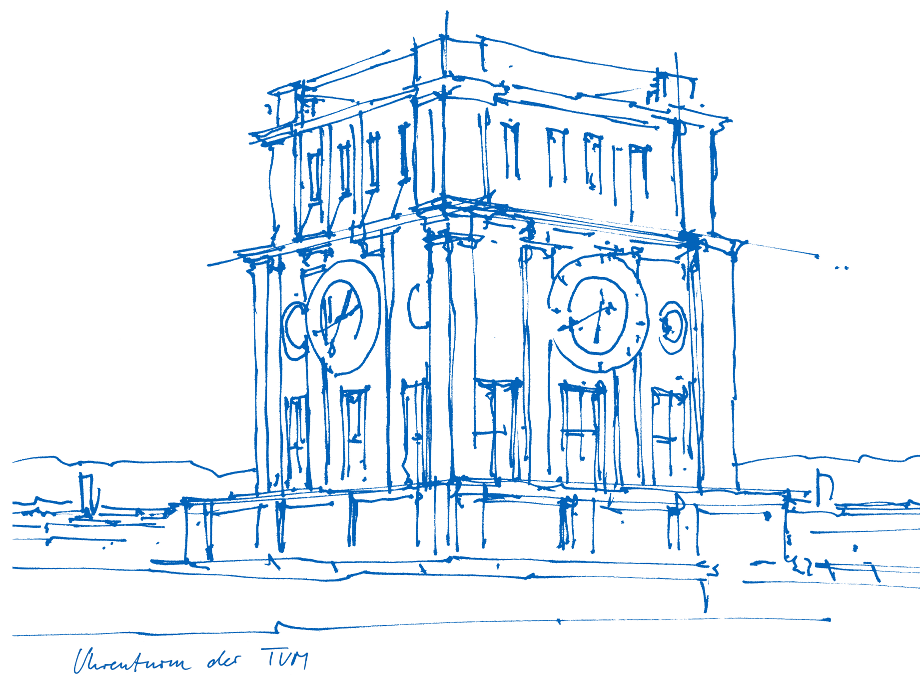
Optimization of a sparse grid-based data mining kernel for architectures using AVX-512

Paul-Cristian Sârbu, Hans-Joachim Bungartz

Technical University of Munich

Department of Informatics

Chair of Scientific Computing



1st HPML Workshop

September 2018, Lyon, France

Motivation

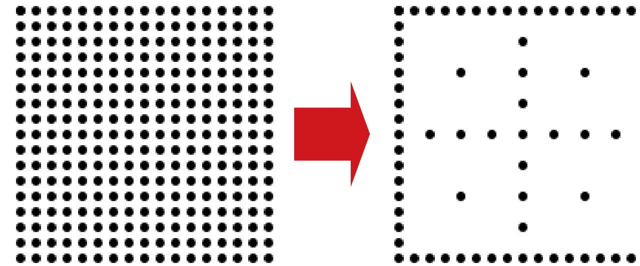
Intel® Parallel Computing Center (IPCC) at Leibniz Supercomputing Center (LRZ)[1]

GADGET	SeisSol
ls1-mardin	**SG++**

- Updating legacy code for AVX-512
- Studying influence of hardware specific features → Knights Landing
- Performance gains relative to Haswell implementation

[1] Intel® Parallel Computing Center at Leibniz Supercomputing Centre and Technische Universität München

Data mining with Sparse Grids



$$\hat{f}_N(\vec{x}) = \sum_{j=1}^{N} u_j \varphi_j(\vec{x})$$

$$(M\lambda I + BB^T)\vec{u} = B\vec{y},\ b_{j,i} = \varphi_j\left(\vec{x}_i\right)$$

→ solve with Conjugate Gradient

M = dataset size λ = regularization parameter

- Optimizing two matrix-vector operations

$$mult: \quad \vec{v} = B^T\vec{u} \qquad\qquad multT: \quad \vec{w} = B\vec{v}$$

Legacy implementation

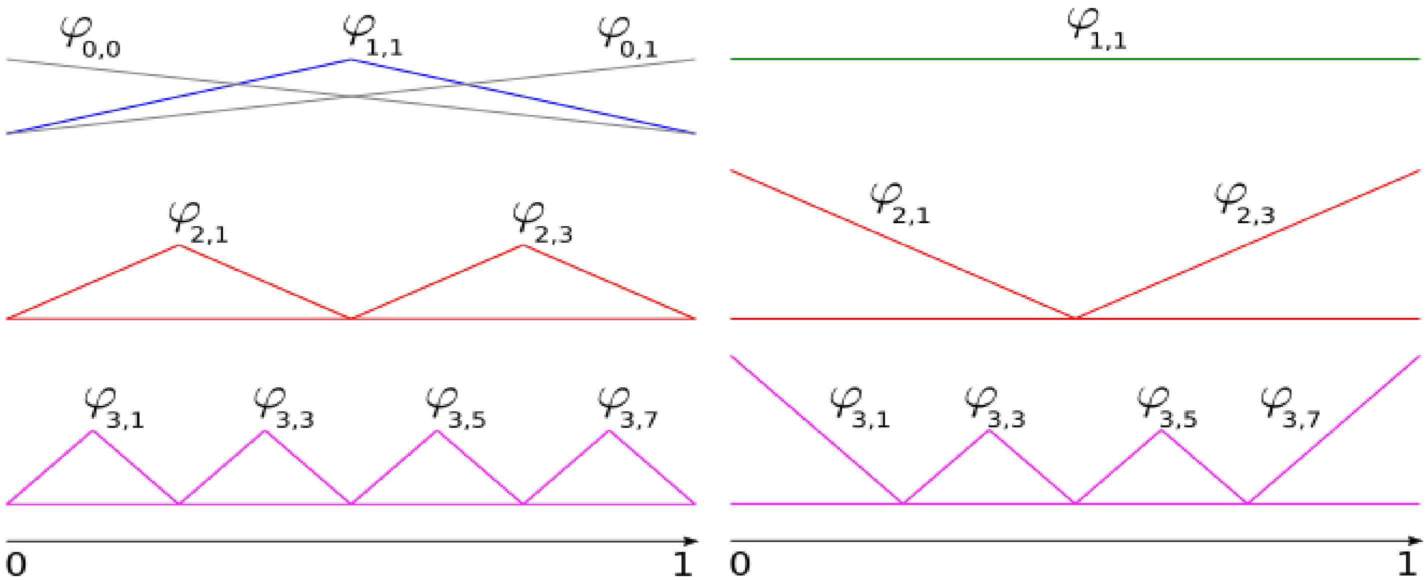
3 loops, iterative scheme, manual vectorization, optimized for AVX2

```
For chunk of data {
    Initialize
    For chunk of grid {
        Broadcast
        For dimension {
            Inner kernel
        }
        Vertical add
    }
}
```

$mult$

```
For chunk of grid {
    Initialize
    For chunk of data {
        Broadcast
        For dimension {
            Inner kernel
        }
        Horizontal add
    }
}
```

$multT$



Basis	Masking needed	Peak performance	Grid points	Time to solution
linear	No	↗	↗	↗
modlinear	Yes	↘	↘	↘

Hardware specifications – Knights Landing

- CooLMUC-3 cluster at LRZ[1]

  148 nodes of 64-cored Xeon Phi 7210-F @ 1.30 Ghz, Turbo Mode ON

- CooLMUC-2 cluster at LRZ[1]

  28-core Xeon E5-2697 v3 (Haswell) nodes @ 2.60 Ghz, Turbo Mode ON

  ↑ ***baseline architecture***

- Why KNL?
  - AVX-512 instruction set
  - Level 2 cache (L2) format → bidirectional 2D mesh
  - 96GB DDR4 + 16GB MCDRAM
  - Clustering and memory modes
  - ...

[1] LRZ Linux-Cluster

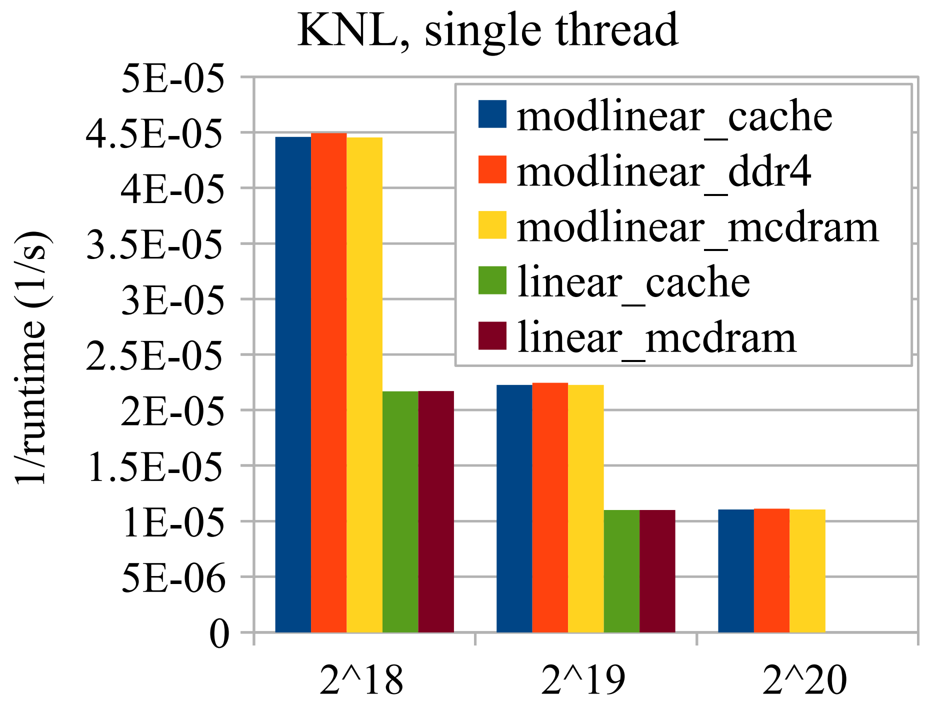
Optimization results

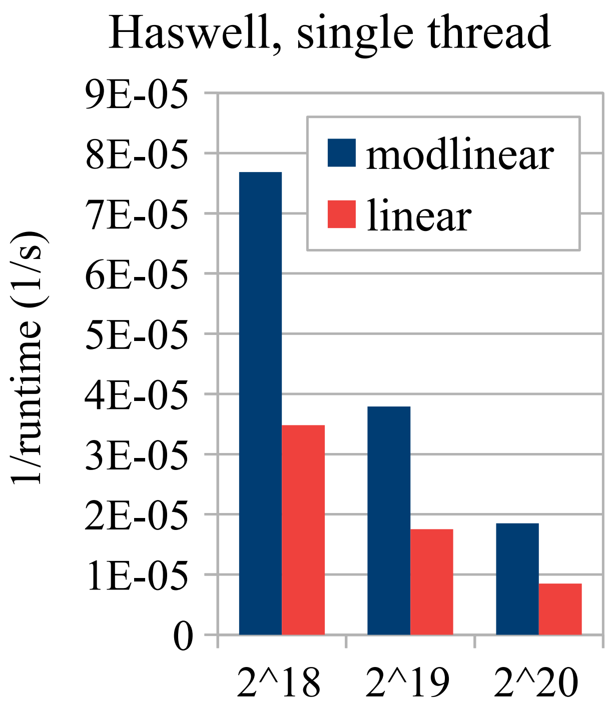
- Intrinsics adaptation
  - embedded broadcasts, but no AVX-512DQ → Skylake
- OpenMP scheduling study
  - dynamic better for few cores, static still better for whole node
- Chunk size study → 2x SIMD registers available

	Grid level	Grid index	Result vectors	Data vectors	Buffers
AVX2	1	1	6	6	2
AVX-512	1	1	~~13~~ 12	~~13~~ 12	~~2~~ 4

Optimization results

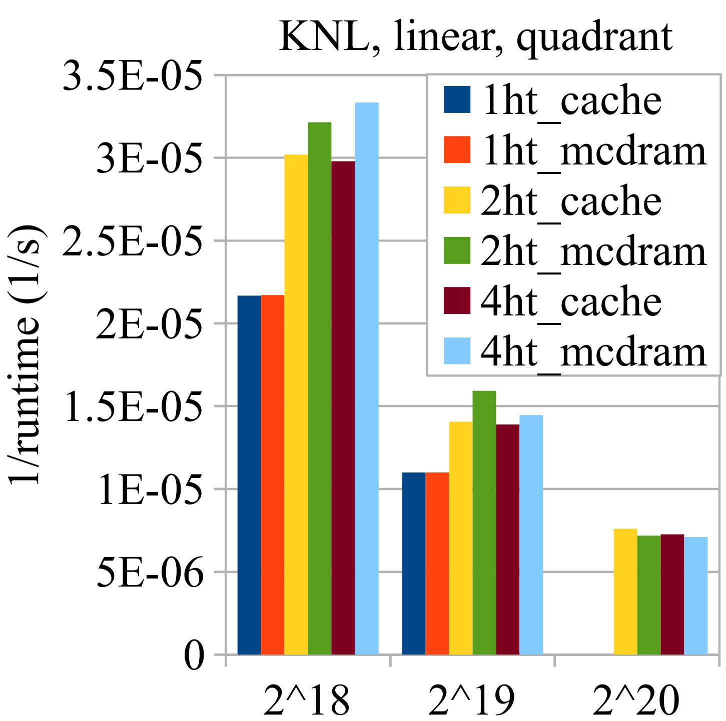
- 5D binary classification problem, chessboard dataset, up to 2^{28} data points (≈ 20GB)
- Thread-level runs

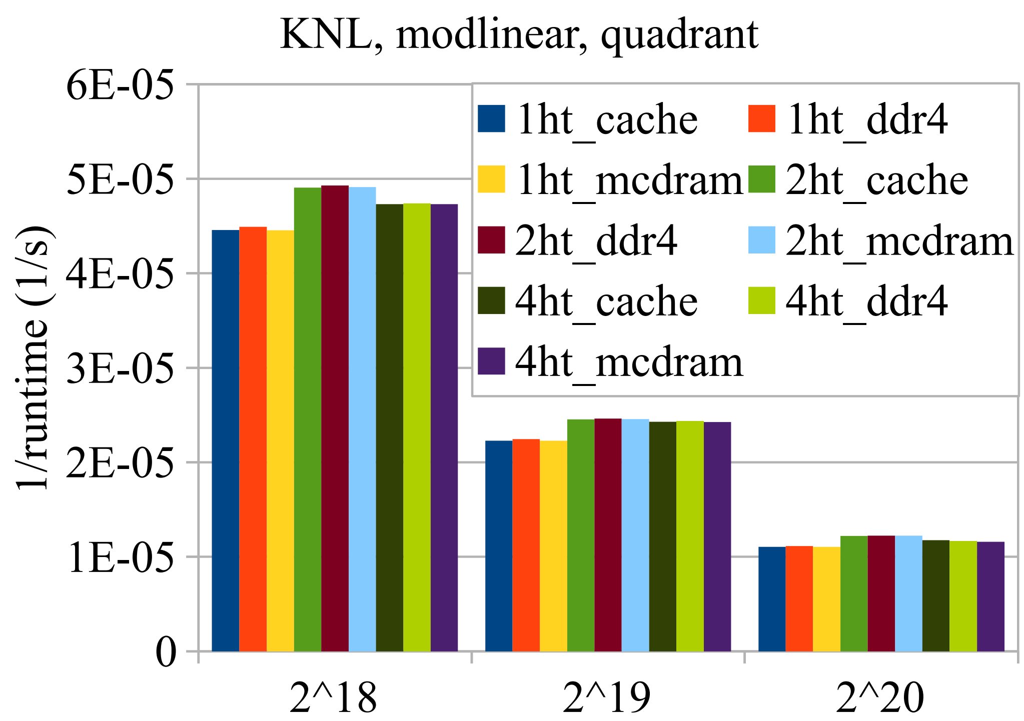




Optimization results

- Core-level runs

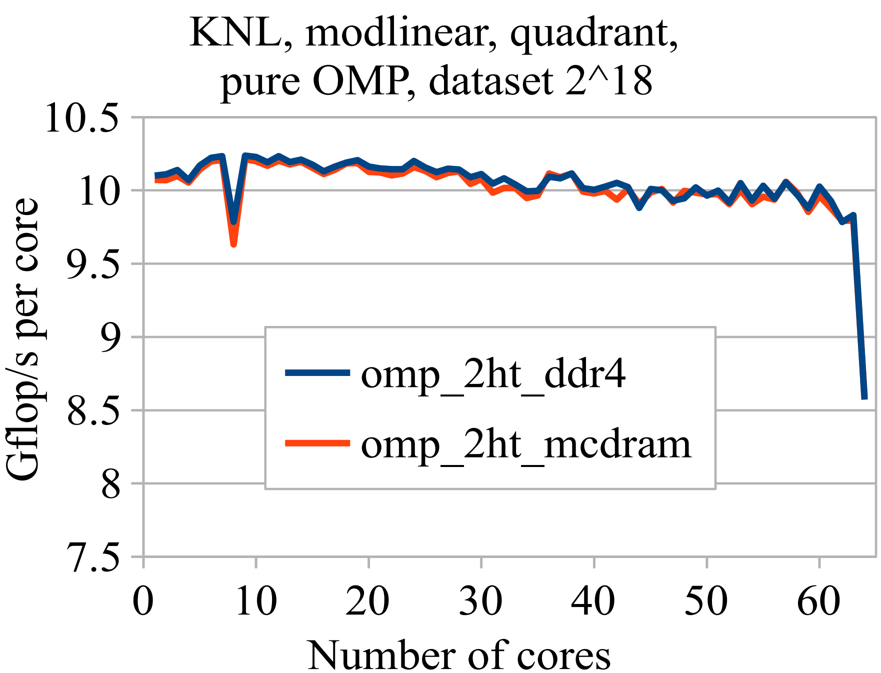


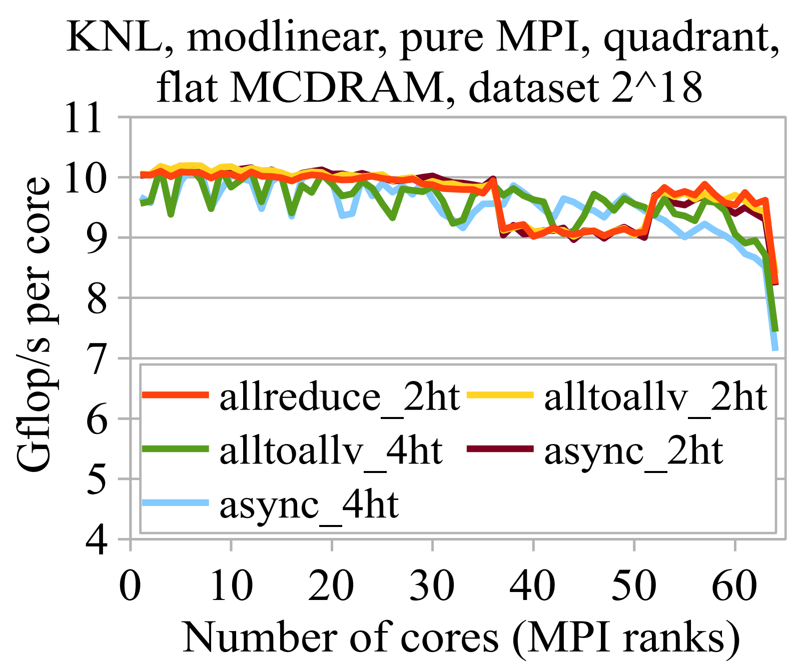


Optimization results

- Node-level runs
  - best configuration: 63 cores x 2 hyperthreads, pure OpenMP, quadrant mode

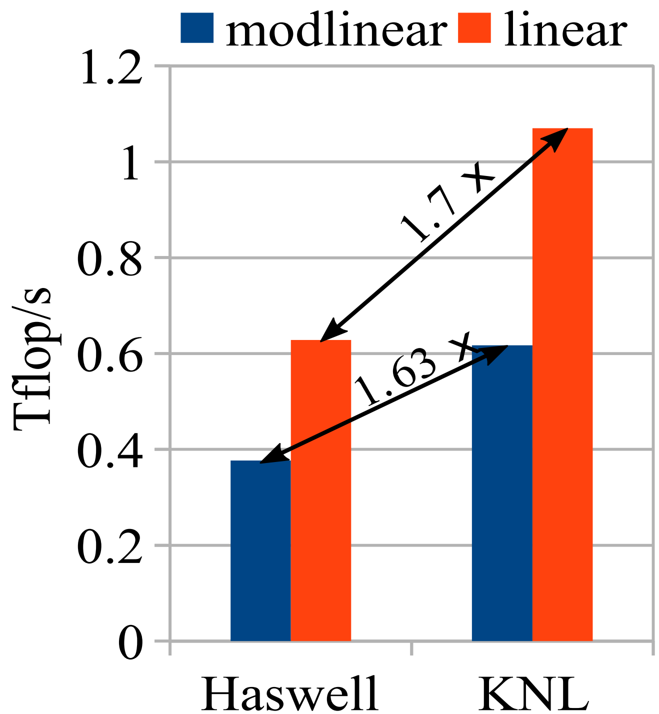
    > *97% parallel efficiency*





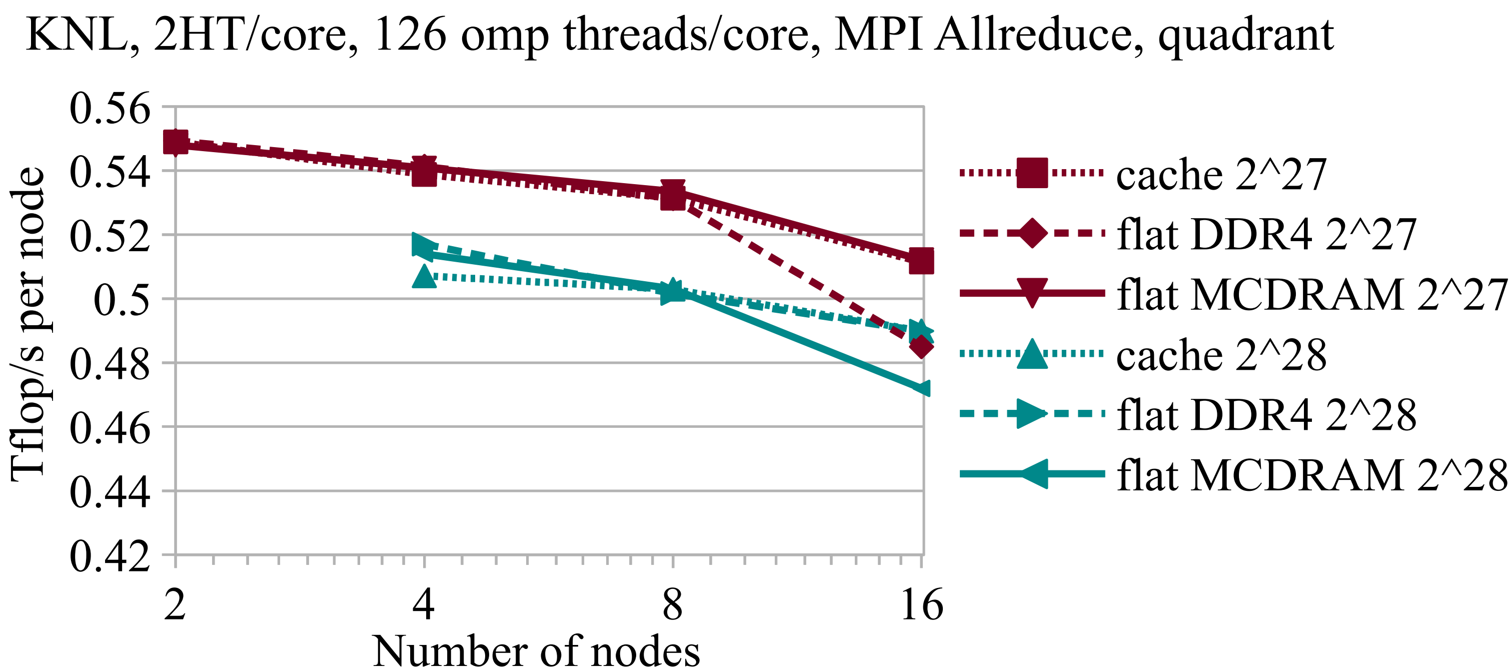
Optimization results

- Node-level runs
  - 1.63 – 1.7x speedup versus Haswell



Optimization results

- Cluster-level runs at MCDRAM size limit



<*16GB* *flat* MCDRAM mode

>*16GB* *cache* or *flat* DDR4 mode

Conclusions

- Successful node-level optimization on KNL
- Considerable speedup versus Haswell
- Good behavior at the HBM size limit

Outlook

- Cluster-level runs (close to DDR4 size limit) on larger machines
- Code ready for Skylake runs

Conclusions

- Successful node-level optimization on KNL
- Considerable speedup versus Haswell
- Good behavior at the HBM size limit

Thank you for your attention!

Contact: sarbu@in.tum.de

References

- H.-J. Bungartz and M. Griebel - „Sparse grids“, Acta Numerica, vol. 13, pp. 147–269, 2004.
- D. Pflüger - „Spatially Adaptive Sparse Grids for Higher-Dimensional Problems“, Verlag Dr. Hut, München, 2010. ISBN 9-783-868-53555-6
- D. Pflüger, B. Peherstorfer, and H.-J. Bungartz - „Spatially adaptive sparse grids for high-dimensional data-driven problems“, Journal of Complexity, Volume 26, Issue 5, 2010
- Alexander Heinecke - „Boosting Scientific Computing Applications through Leveraging Data Parallel Architectures“, Verlag Dr. Hut, pp. 1-231, 2014. ISBN 978-3-8439-1408-6

Hardware specifications – Knights Landing - extra

Clustering modes

- *quadrant*
  - KNL as symmetric multi-processor
  - software transparent
  - low L2 miss latencies
- *sub-NUMA clustering*
  - heavily reliant on non-uniform memory access model
  - boost for memory bound codes
- *all-to-all*
  - default debug mode

Memory modes

- *cache*
  - MCDRAM as level 3 cache
  - software transparent
  - expensive L3 misses
- *flat*
  - straight-forward address mapping
  - memory pinning required
  - reduced data access time
- *hybrid*
  - mixed strategy; fine control needed