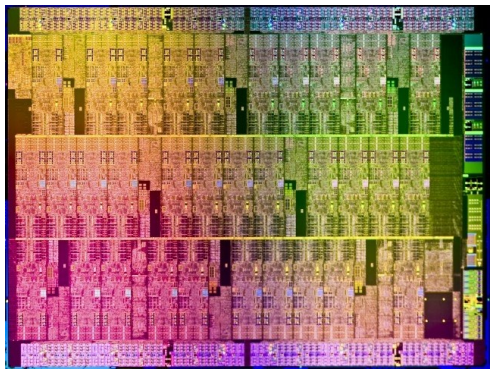
Integration of Intel Xeon Phi Servers into the HLRN-III Complex: Experiences, Performance and Lessons Learned

Florian Wende, Guido Laubender and Thomas Steinke

Zuse Institute Berlin (ZIB)



Cray User Group 2014 (CUG'14)

May 6, 2014

Outline

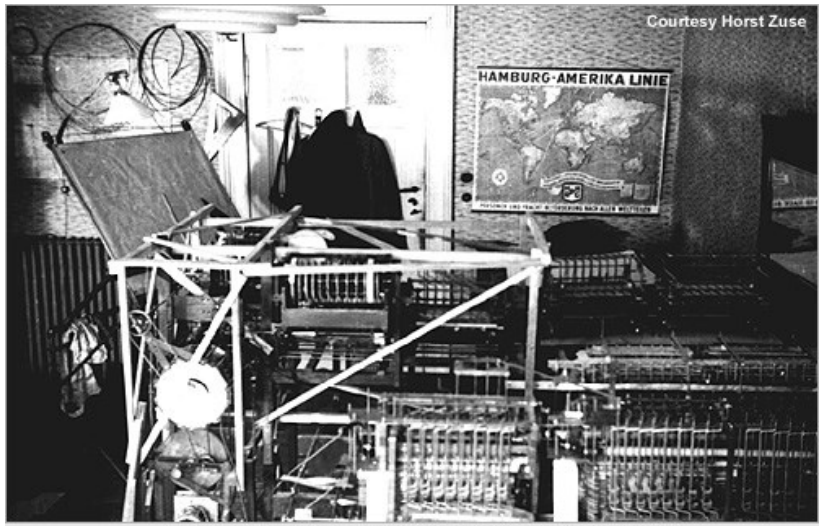
- Site Overview ZIB, IPCC & the HLRN III System
- Integration of a Xeon Phi cluster into HLRN complex @ ZIB
  - Workloads, research, challenges
- Performance: two example applications
- Lessons Learned

Site Overview ZIB, and HLRN

Norddeutscher Verbund für Hoch- und Höchstleistungsrechnen

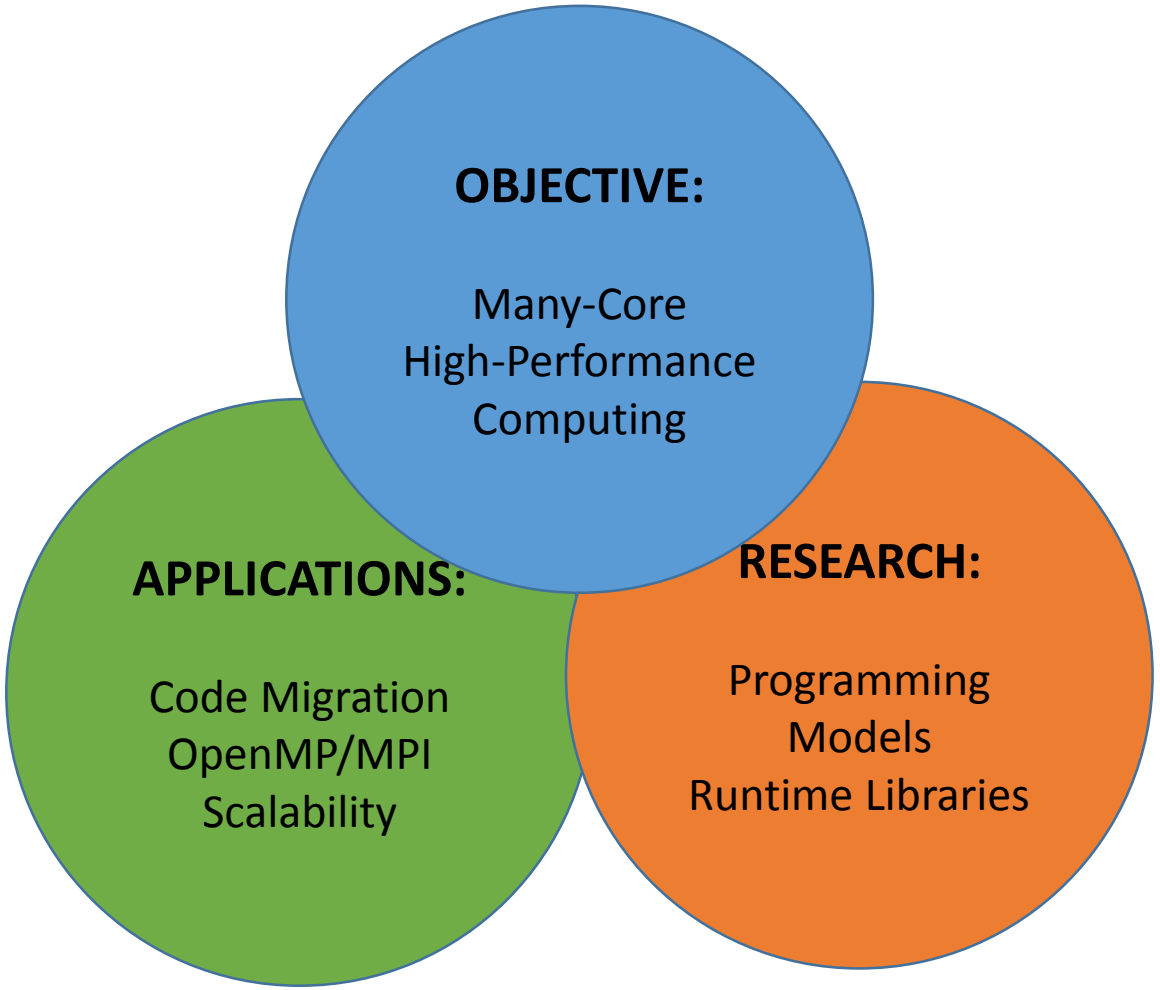
About the Zuse Institute Berlin

- non-university research institute
- founded in 1984
- Research domains:
  - Numerical Mathematics
  - Discrete Mathematics
  - Computer Science
- Supercomputing:
  - operates the HPC systems of the HLRN alliance
  - domain specific consultants
- Research: distributed systems, data management, many-core computing

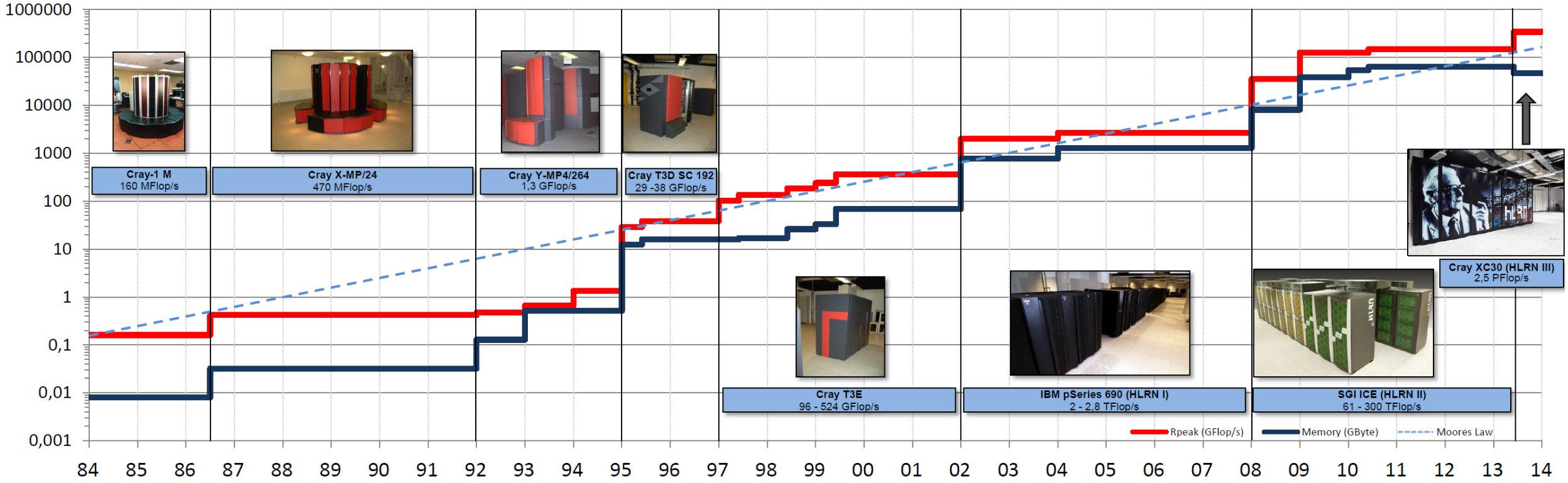


The original Z1 in Konrad's parent's living room circa 1938

Research Center for Many-Core High-Performance Computing @ ZIB



History of Supercomputing @ ZIB



HLRN – the North-German Supercomputing Alliance

- *Norddeutscher Verbund zur Förderung des Hoch- und Höchstleistungsrechnens – HLRN*
- joint project of seven North-German states (Berlin, Brandenburg, Bremen, Hamburg, Mecklenburg-Vorpommern, Niedersachsen and Schleswig-Holstein)
- established in 2001
- HLRN alliance jointly operates a distributed supercomputer system
  - hosted at Zuse Institute Berlin (ZIB) and at Leibniz University IT Service (LUIS), Leibniz University Hanover

Konrad @ ZIB

Gottfried @ LUIS

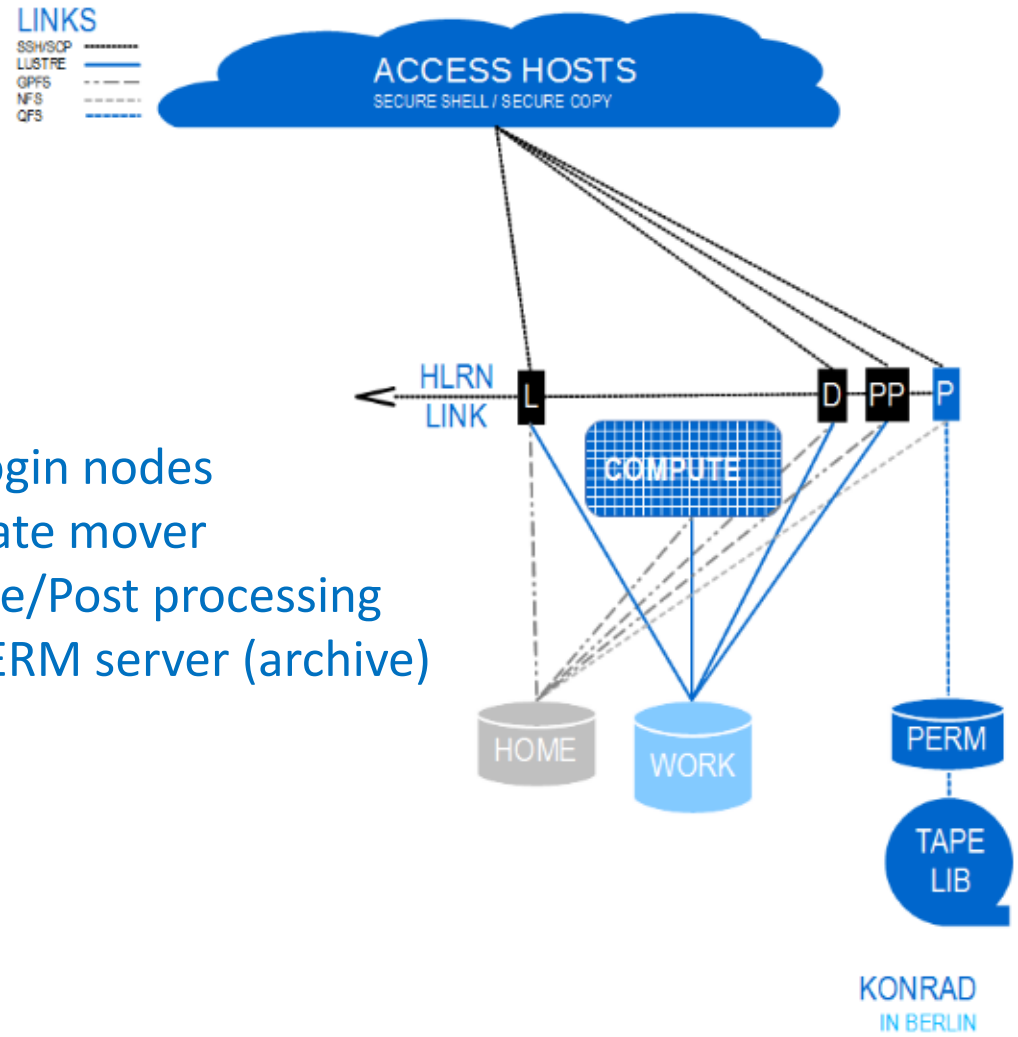
The HLRN-III System

Cray XC30 Systems in Q4/2014

HLRN-III Overall Architecture

Key Characteristic (Q4/2014)

- Non-symmetric installation
- @ZIB: 10 Cray XC30 cabinets
- @LUIS: 9 Cray XC30 cabinets
  + 64 four-way SMP nodes
- Global resource mgmnt & accounting (Moab)
- File systems
  - WORK: 2 x 3.6 PB, Lustre
  - HOME: 2 x 0.7 PB, NAS appliance



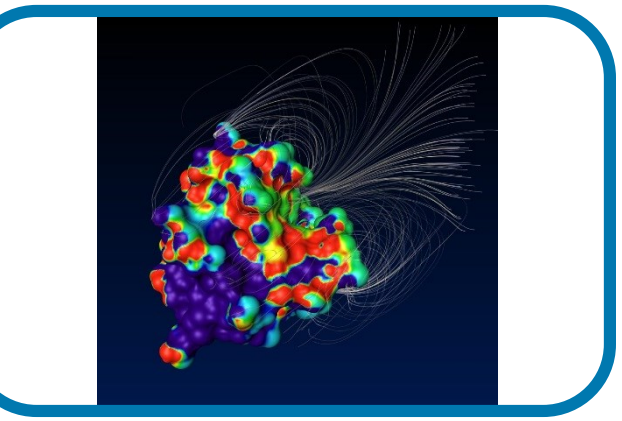
L Login nodes
D Date mover
PP Pre/Post processing
P PERM server (archive)

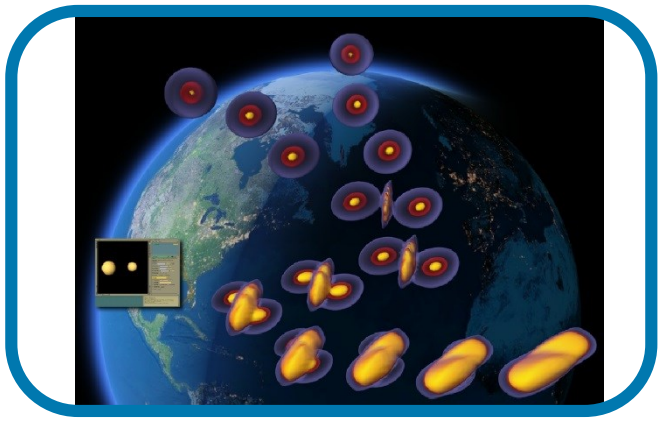
The HLRN-III Complex @ ZIB

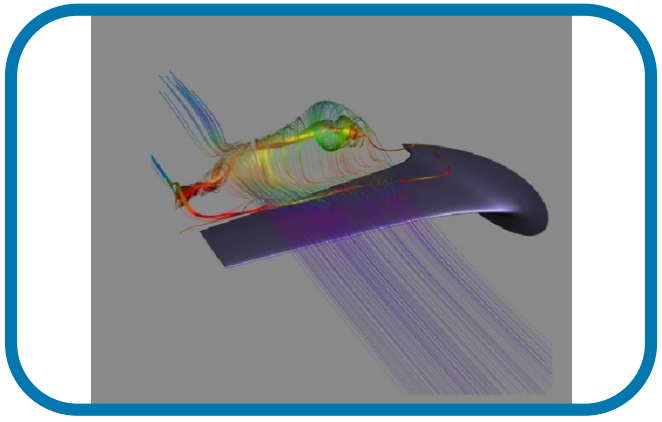
- Compute: Cray XC30 (Q4/2014)
- 744 XC30 nodes (1872 nodes)
  - 24 core Intel IVB, HSW
  - 64 GB / node
- 4 Xeon Phi nodes (7xxx series)
- Storage: Lustre + NAS
  - WORK (CLFS): 1.4 PB (3.6 PB)
  - HOME: 0.7 PB
  - DDN SFA12K

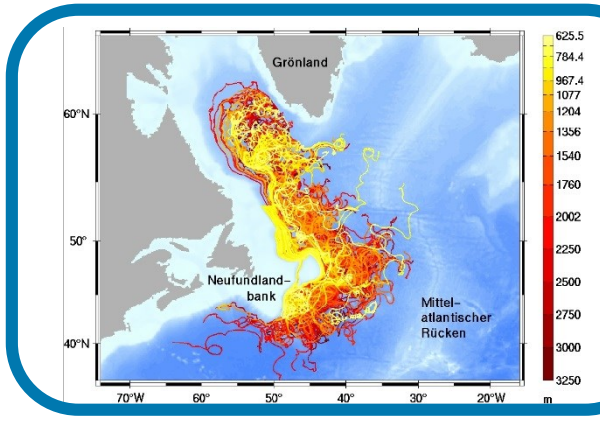
Current Cray XC30 installation @ ZIB

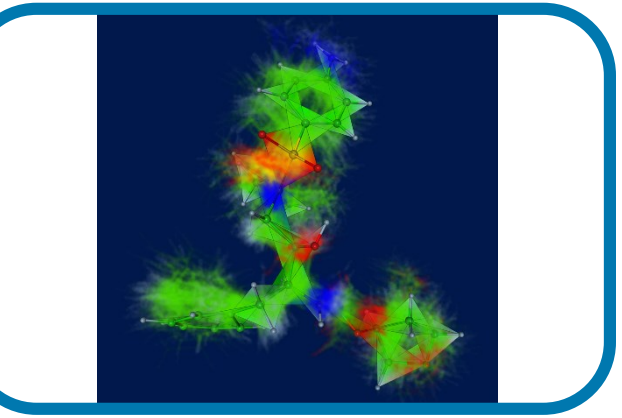
Workloads on HLRN System

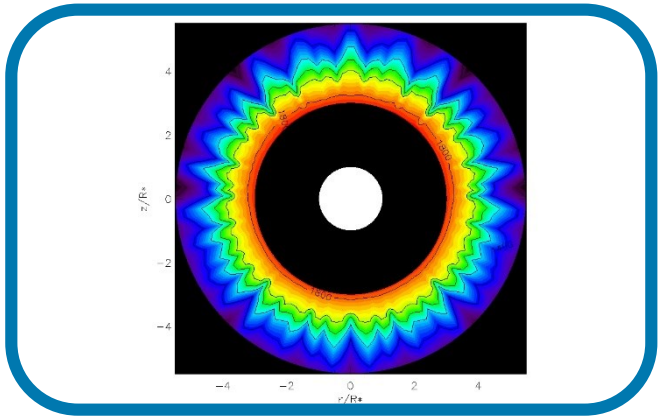


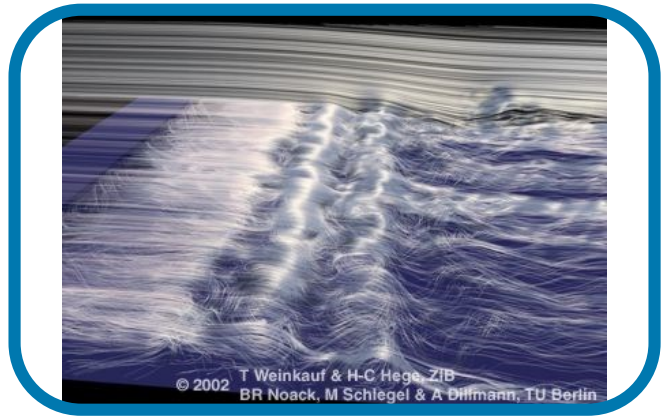


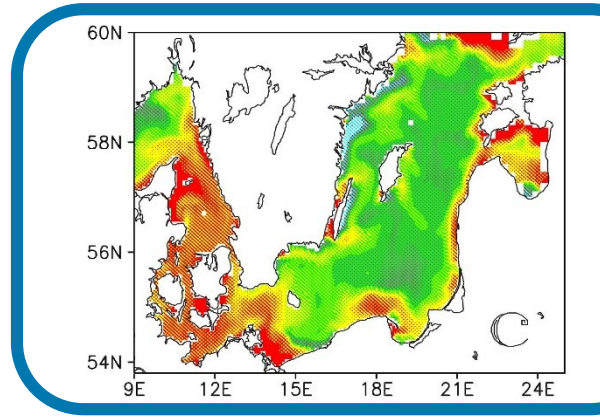












- **Diverse job mix, various workloads**
- **Codes: self-developed codes + community codes + ISV codes**

Preparing for Many-Core in HPC

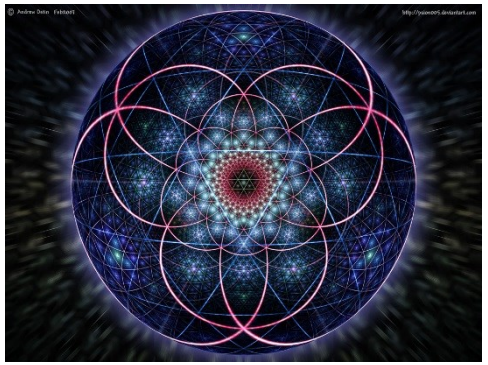
Integration of a Xeon Phi Development Cluster into HLRN-III Complex

Our Approach with Given Constraints

- Goal: Evaluation, migration, optimization of selected workloads
- Status: Research experiences with accelerator devices since ~2005
  - FPGA (Cray XD1,…), ClearSpeed, CellBE, now GPGPU + MIC
- Challenges: productivity, easy-of-use, “programmability”
  - limited personal resources for optimizing production workloads
  - additional funding extremely important

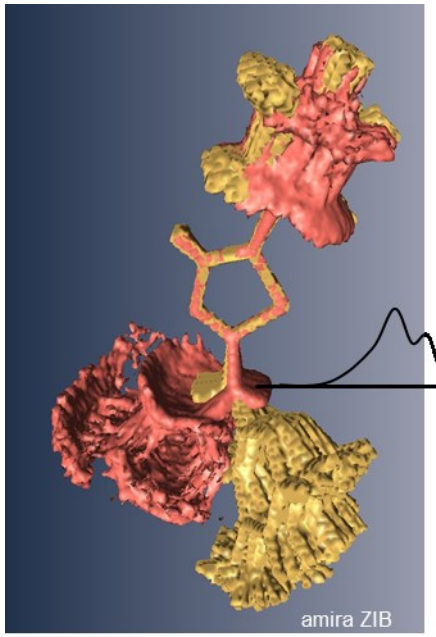
- Collaboration with Intel (IPCC)
  - Push many-core capabilities with MIC
  - Optimization of workloads and many-core research

Workloads Considered



BQCD

Raasch, Uni Hanover



GLAT

Work in Progress…

Workload	Key Results (Status)	Issues/Challenges	Solutions	Tools/Approaches
BQCD	OpenMP with LEO	SIMD with MPI data layout	AoSoA	• VTune • Data layout redesign
GLAT	• CPU+Acc code • OpenMP + MPI • Concurrent kernel execution	• Concurrent kernel exec • Vectorization	• LEO and MPI • HAM Offload • Intrinsics	• SIMD on CPU based on MIC code • Offload (LEO, OpenMP4, HAM)
HEOM	MIC-friendly data layout	Auto-vectorization in OpenCL	Flexible data models	• Data layout (SIMD) for OpenCL
VASP	• Extensive profiling • Major call-trees for HFXC	• Introducing OpenMP parallelism • Data layout	• Thread-safe functions	• VTune, Cray PAT • in progress
PALM	Test bench	working OpenMP test set		

Ongoing Research Work

- Programming Models: Heterogeneous Active Messages (HAM) *(M. Noack)*
- Throughput Optimization: Concurrent Kernel Execution framework *(F. Wende)*

→ *prepared for new application (de)composition schemes*

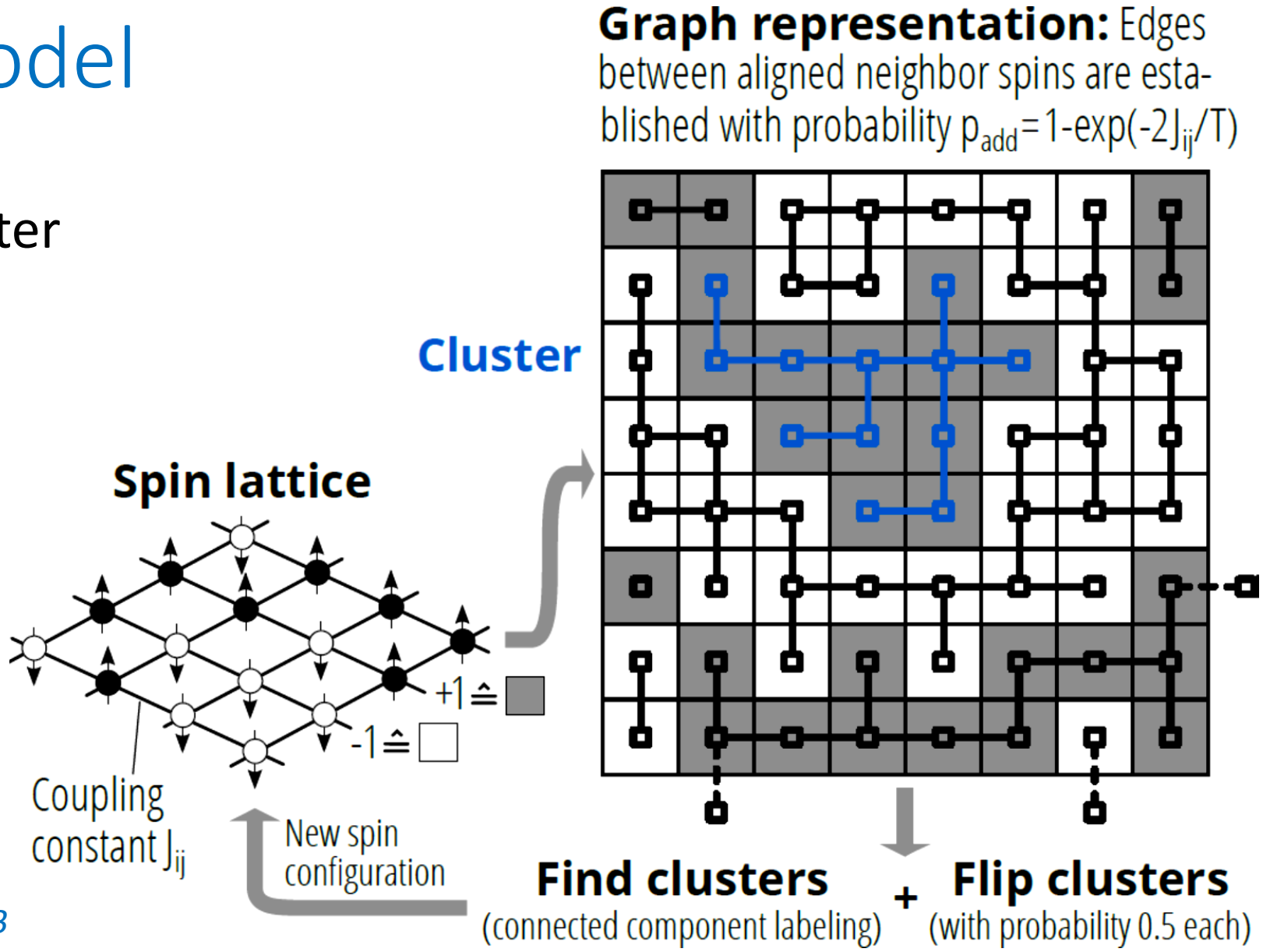
- designs rely on C++ template mechanism
- work on Intel Xeon Phi and Nvidia GPUs
- interface to Fortran / C
- performance studies with real-world app

see SAAHPC12 paper
and SC14 & EuroPar14 (submitted)

Two Example Apps on Xeon Phi

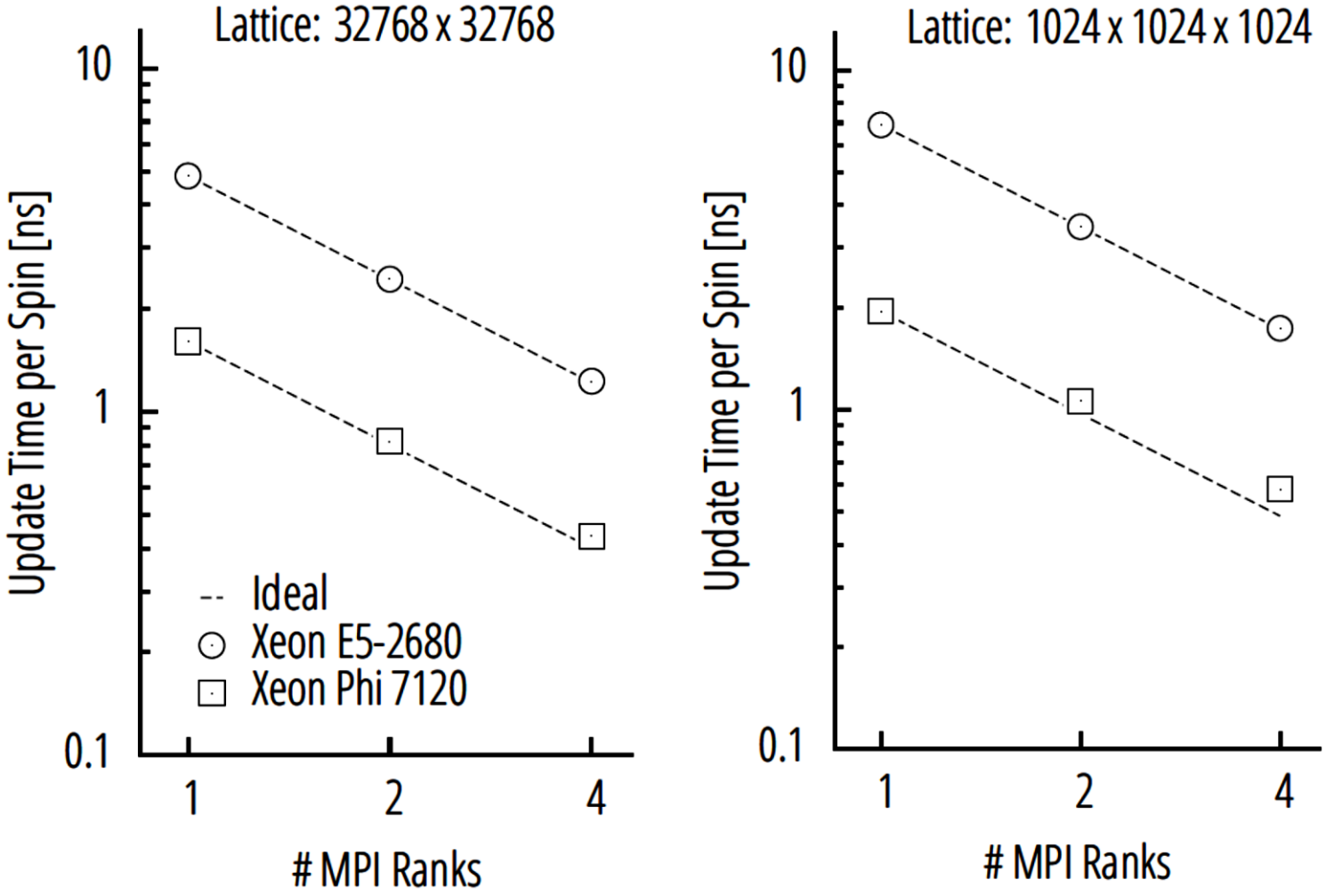
2D/3D Ising Model

- Swendsen-Wang cluster algorithm



Work of F. Wende, ZIB

Performance: Device vs. Device (Socket)



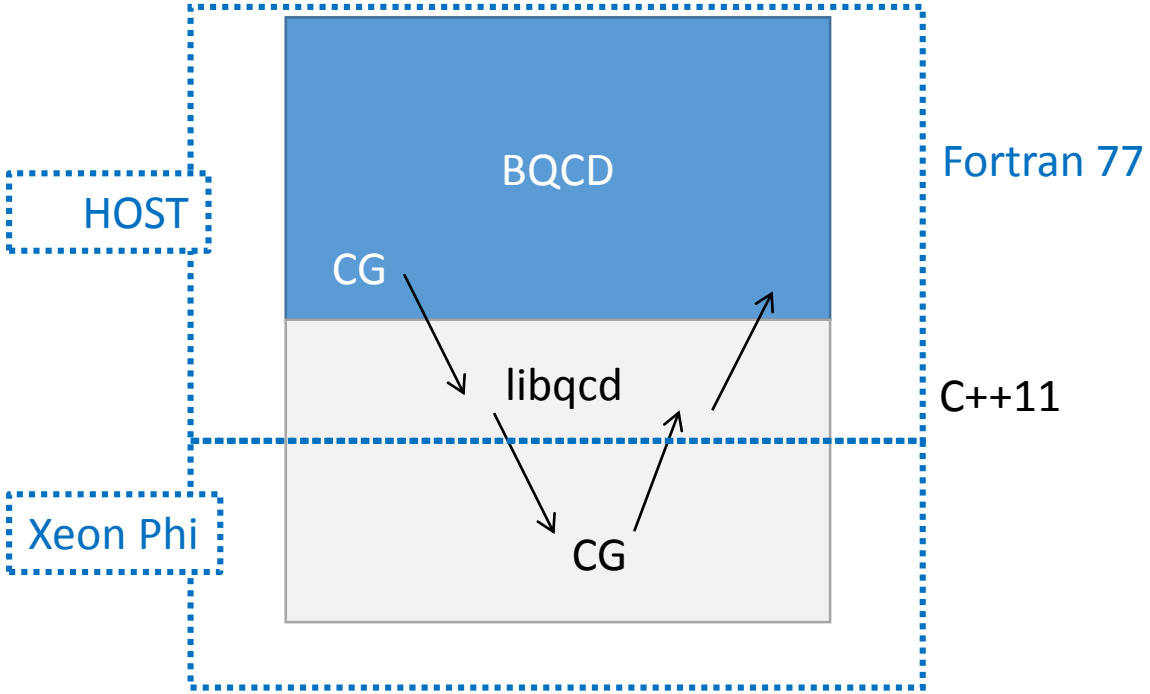
- one MPI rank per device/host
- OpenMP
- native exec on Phi
- Phi: SIMD intrinsics
  Host: SIMD by comp.
- Phi: 240 threads
  Host: 16 threads
- **~ 3 x speedup**

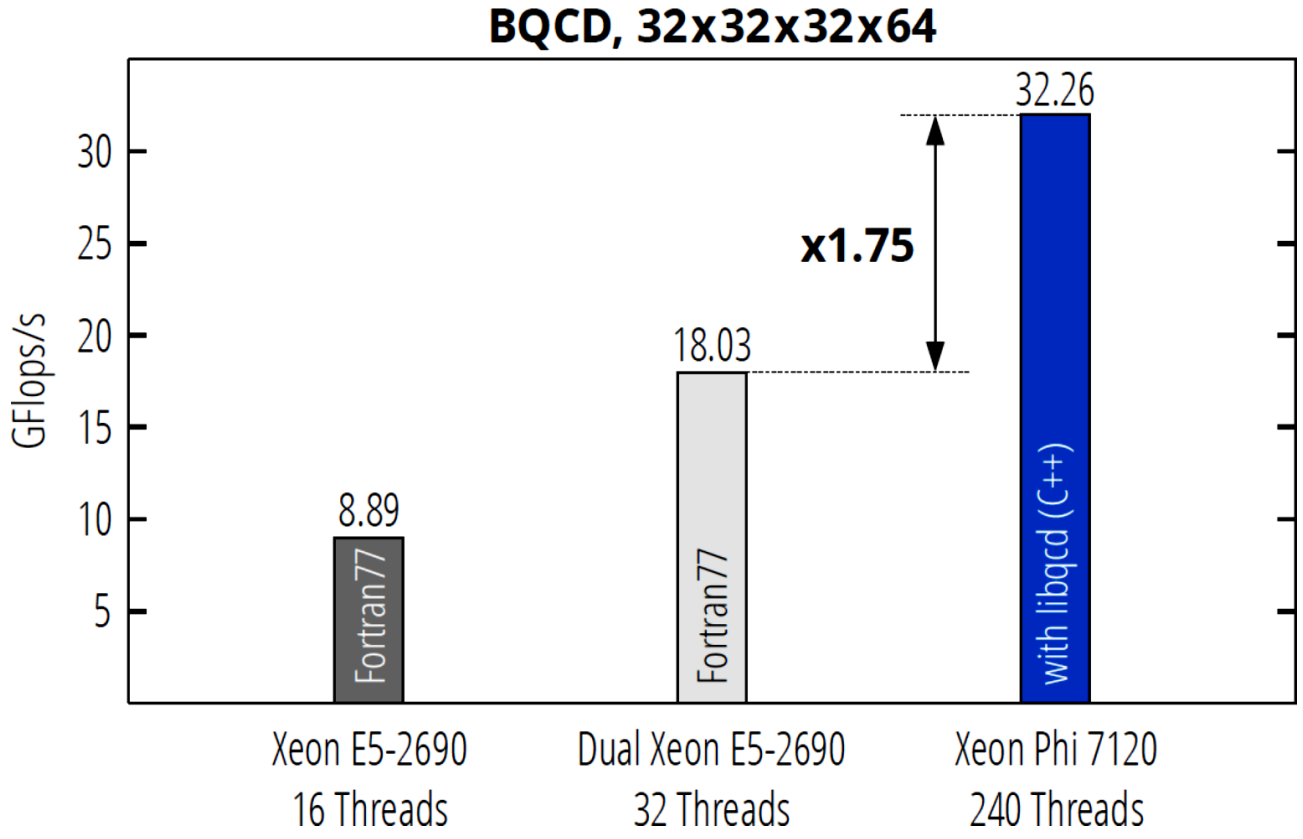
F. Wende, Th. Steinke, SC13, pp 83:1/12

BQCD - *Berlin Quantum Chromodynamics*

Solve Ax=b with CG

Offload Architecture for Xeon Phi (Intel LEO)





Original code developed by H. Stüben, Y. Nakamura

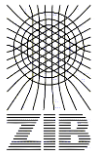
Vectorization: **AoS → AoSoA**

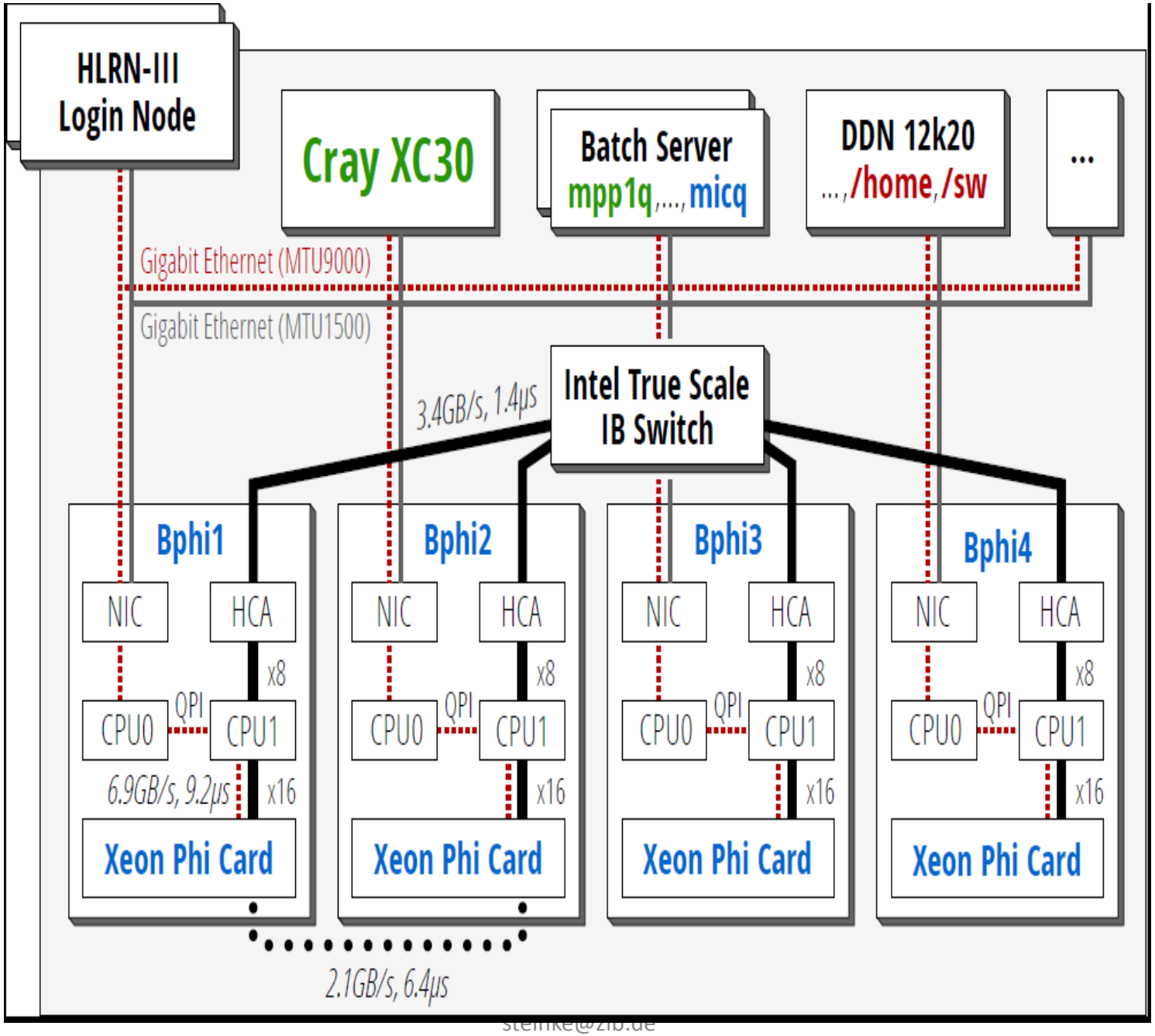
libqcd *by Th Schütt (ZIB)*

Lessons Learned

If Non-Sysadmins Have to Build and Configure a Xeon Phi Cluster…

(consequences of “bad timing”: concurrent HLRN-III and Phi cluster installation)




HLRN-III
Login Node
Cray XC30
Batch Server
mpp1q,...,micq
DDN 12k20
.../home,/sw
...
Gigabit Ethernet (MTU9000)
Gigabit Ethernet (MTU1500)
3.4GB/s, 1.4µs
Intel True Scale
IB Switch
Bphi1
Bphi2
Bphi3
Bphi4
NIC
HCA
x8
QPI
CPU0
CPU1
6.9GB/s, 9.2µs
x16
Xeon Phi Card
2.1GB/s, 6.4µs

„Challenges“ (1)

Batchsystem:

- Torque client supports MIC (re-compile)
- smooth integration with HLRN-III config
  - introduce new Moab class & feature “mic”

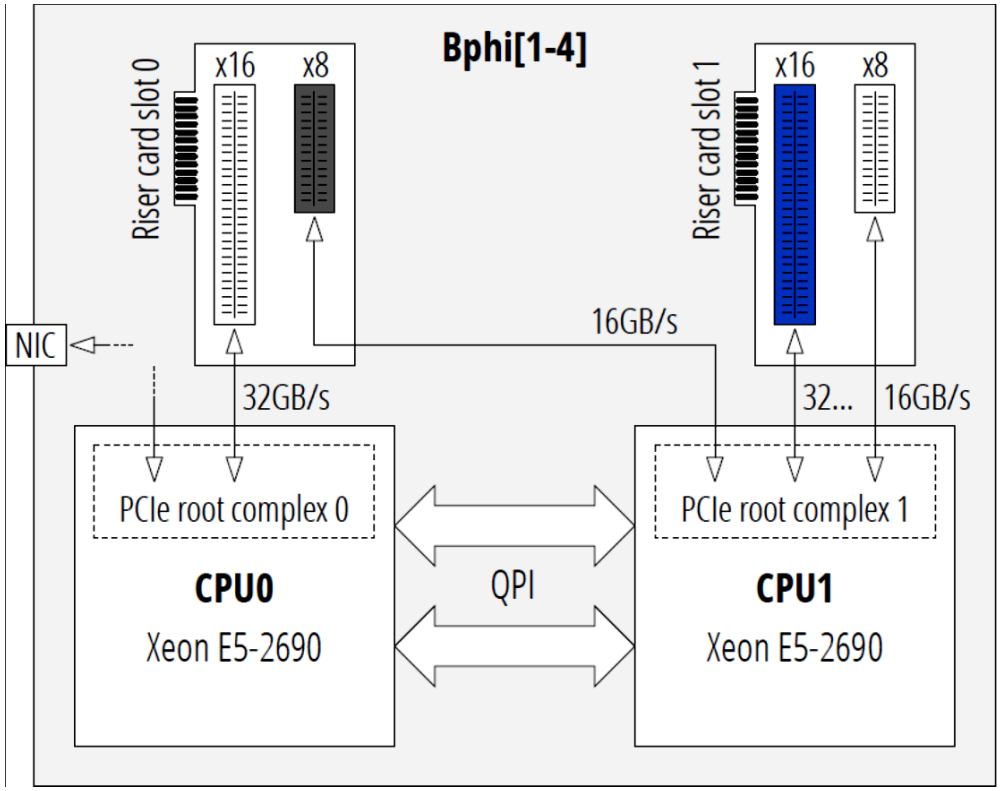
- Torque prologue/epilogue scripts for handling Phi card access:
  - prologue: enable temporary user access on Phi card
  - epilogue: remove user from Phi OS, re-boot Phi OS

„Challenges“ (2)

- Authentication:
  - LDAP integration host-side smoothly
  - card-side not supported (MPSS 3.1)

Cluster Assembling…

- Initial HW configuration showed serious MPI performance issues
  - Beginner's mistake: the PCIe root complex story



Intel True Scale Fabric InfiniBand host adapter QLE7340 (HCA)

Intel Xeon Phi 7120

theoretical bandwidths for bi-directional communication (full duplex)

… Solved: Intel MPI Benchmark Results

	Fabric	# Ranks	Rate [GB/s]	Latency [us]
(A) Host to Host	TMI	2	1.8	**1.4**
		16	**3.0**	8.0
(B) Host to Phi	SCIF	2	5.7	**9.2**
		16	**6.9**	62.0
(C) Phi to Phi	TMI	2	0.4	**6.4**
		16	**2.1**	9.3

IMB v. 3.2.4
MPSS 3.1

Almost Last Words…

- Security:
  - MPSS supports old CentOS kernel → access to Phi host from HLRN login nodes where HLRN access policies are in effect
  - /sw mounted read-only
  - access granted from offload programs (COI daemon)?

- Transition into the HPC SysAdmin group done.

IPCC @ ZIB is a Significant Instrument…

- Many-cores in future data processing architectures
  - prepare HLRN community for future architectures
- Xeon Phi = flexible architecture
  - optimization & clear designs → **beneficial for standard CPU too!**
  - for R&D in computer science (MPI, SCIF, …)
- pushes re-thinking: algorithms, architectures, HW/SW partitioning,…
- support for ZIB/HLRN community by Intel

Thank You!

ACKNOWLEDGEMENT

- Thorsten Schütt
- Intel:
  - Michael Hebenstreit, Thorsten Schmidt
  - Michael Klemm, Heinrich Bockhorst, Georg Zitzelsberger