


# Seminar Aspects of Distributed Systems

## Implementation of Energy Efficiency in HPC

**HLRN III Supercomputer  
Introduction + Guided Tour (26.05.2016)**

# Schedule

- 
- ~~07.04.16~~ — ~~Introduction, Impulse Presentation~~
  - ~~14.04.16~~ — ~~Assignment of Topics, Lab: Play with the System~~
  - ~~21.04.16~~ — ~~Lab: Work on Your Topic~~
  - ~~...~~ — ~~Lab: Work on Your Topic~~
  - **26.05.16**     **HLRN III Supercomputer**
  - ...     Lab: Work on Your Topic
  - 30.06.16     Presentation Session #1
  - 07.07.16     Presentation Session #2
  - ~~14.07.16~~ — ~~Presentation Session #3~~

# Der Norddeutsche Verbund für Hoch- und Höchstleistungsrechnen (HLRN)

- als Zusammenschluss der 6 nördlichen Bundesländer (Berlin, Bremen, Hamburg, Mecklenburg-Vorpommern, Niedersachsen und Schleswig-Holstein) gegründet
- Ende 2012 ist Brandenburg als 7. Mitglied beigetreten.



## Standorte der Hardware

- Berlin (Konrad-Zuse-Zentrum)
- Hannover (Leibniz Universität IT Services)



# Historie: HLRN-I und HLRN-II

## HLRN-I

- 40 Mio DM (2002-2008)
- 1024 CPUs (IBM Power)
- 5.2 TFLOP/s peak

## HLRN-II

- 30 Mio € (2008-2013)
- 25k CPU Kerne (Intel x86)
- 300 TFLOP/s peak

- (möglichst) symmetrischer Aufbau in Berlin und Hannover
- 10 Gbit/s Verbindung zwischen den Sites
- einheitliche Nutzer- und Projektverwaltung
- dezentrales Netzwerk von Fachberatern zur Unterstützung der Nutzer

# HLRN-III – zeitlicher Ablauf

November 2010	Antragsskizze an die DFG
Juli 2011	Genehmigung des Vollantrags
März 2012	EU weite Ausschreibung des HLRN-III im Verhandlungsverfahren
August 2012	Beginn der Verhandlungen mit IBM und Cray
Oktober 2012	Finale Angebote nach drei Verhandlungsrunden
Dezember 2012	Vertragsabschluss mit Cray
Oktober 2013	Betriebsbereitschaft HLRN-III Phase 1 in Berlin
Dezember 2013	Betriebsbereitschaft HLRN-III Phase 1 in Hannover
September 2014	Installation HLRN-III Phase 2
Dezember 2014	Betriebsbereitschaft HLRN-III Phase 2

# HLRN-III - Anforderungen

- Investitionssumme und Räumlichkeiten (30 Mio €)
- Limitierung der Betriebskosten (Wartung, Strom und Klima) (2 M € pro Jahr und Site für 5 Jahre)
- Erhalt der Expertise aus dem HLRN-II (Forderung mindestens ein Teil der MPP Komponente x86 CPU + Linux)
- extrem heterogener Anwendungsmix, d.h. die Leistungsbewertung kann nicht an einem singulären Benchmark festgelegt werden
- Anbieter mussten Leistungszusagen für 8 repräsentative Applikationen abgeben
- Garantierte Lauffähigkeit aller großen ISV und Open Source Pakete
- Vorgaben für weitere Komponenten: SMP Knoten, Pre/Post-Processing und Dateisysteme

# HLRN-III Hardware Phase 1

## MPP symmetrisch in Berlin + Hannover

- 4 Racks Cray XC30
  - 744 Knoten (2x 12core Intel Ivy Bridge, 64GB RAM)
  - Aries Interconnect
- 17856 Cores
- 46 TB Memory
- 329 TFlops peak
- 2 Pre/Post Knoten (4x 8core Intel Sandy Bridge, 768GB RAM)



## File Systeme in Berlin + Hannover

- 1.4 PB Work-Filesystem **lustre**  
~20GB/s Bandbreite
- 680TB /home /sw etc. (10GigE NFS/GPFS)

## SMP nur in Hannover

32 Knoten:



- 4x 8core Intel Sandy Bridge
- 256GB RAM
- 12 TB lokale Festplatten
- InfiniBand FDR (Lustre + MPI)

## HLRN-III Hardware Phase 2 1/2



### MPP in Berlin

- 6 Racks Cray XC40
  - 1128 Knoten (2x 12core Intel Haswell, 64GB RAM)
  - Aries Interconnect
- 27072 Cores
- 71TB Memory
- 1040 TFlops peak
- 2 Pre/Post Knoten (4x 10core Intel Ivy Bridge, 1.5TB RAM)

### MPP in Hannover

- 5 Racks Cray XC40
  - 936 Knoten (2x 12core Intel Haswell, 64GB RAM)
  - Aries Interconnect
- 22464 Cores
- 59TB Memory
- 863 TFlops peak
- 2 Pre/Post Knoten (4x 10core Intel Ivy Bridge, 1.5TB RAM)



## HLRN-III Hardware Phase 2 2/2



### SMP nur in Hannover





- 32 weitere Knoten:
- 4x 10core Intel Ivy Bridge
- 512GB RAM
- 24 TB lokale Festplatten
- InfiniBand FDR (Lustre + MPI)

### l.u.s.t.r.e. pro Site:

- 1.4 PB Work-Filesystem
- ~31GB/s Bandbreite

# HLRN-III Hardware Endausbau



- MPP Komponente
  - 3552 Knoten, 85248 Cores, 222TB RAM, ~2.5PFlops peak
- SMP Komponente 
  - 64 Knoten, 2304 Cores, 24TB RAM
- Hannover: 10 Nvidia K40 im SMP System 
- Berlin: 4 Knoten Intel Xeon Phi Mini-Cluster 
- Work Dateisystem 
  - ~7.2PB Kapazität, ~100GB/s I/O Bandbreite
- 8 Pre/Post Knoten, /home (680TB), Login Knoten, . . .

## HLRN-III vs. HLRN-II

- Anzahl Cores (MPP): 20k → 85k (**\*4.25**)
- größtes MPP Segment: 7680 cores → 44920 cores (**\*5.8**)
- RAM: 100TB → 222TB (**\*2.2**)
- Peak Performance: 240 TFlops → 2.5 Pflops (**~ \*10**)
- Work Filesystem: 1.6 PB → 7.2 PB (**\* 4.5**)
- Racks: 100 → 30 (**\* 0.3**)
  
- Leistungsaufnahme: **1.6 MW → 1.6 MW**
- Budget **30 Mio € → 30 Mio €**

## HLRN-III Betriebskonzept

- Zwei Standorte – ein System
- Einheitliche Nutzerverwaltung
- Ein Accounting für beide Systeme
- Black-Box
- Moab wird als Grid betrieben - Jobs aus H können auch in B laufen

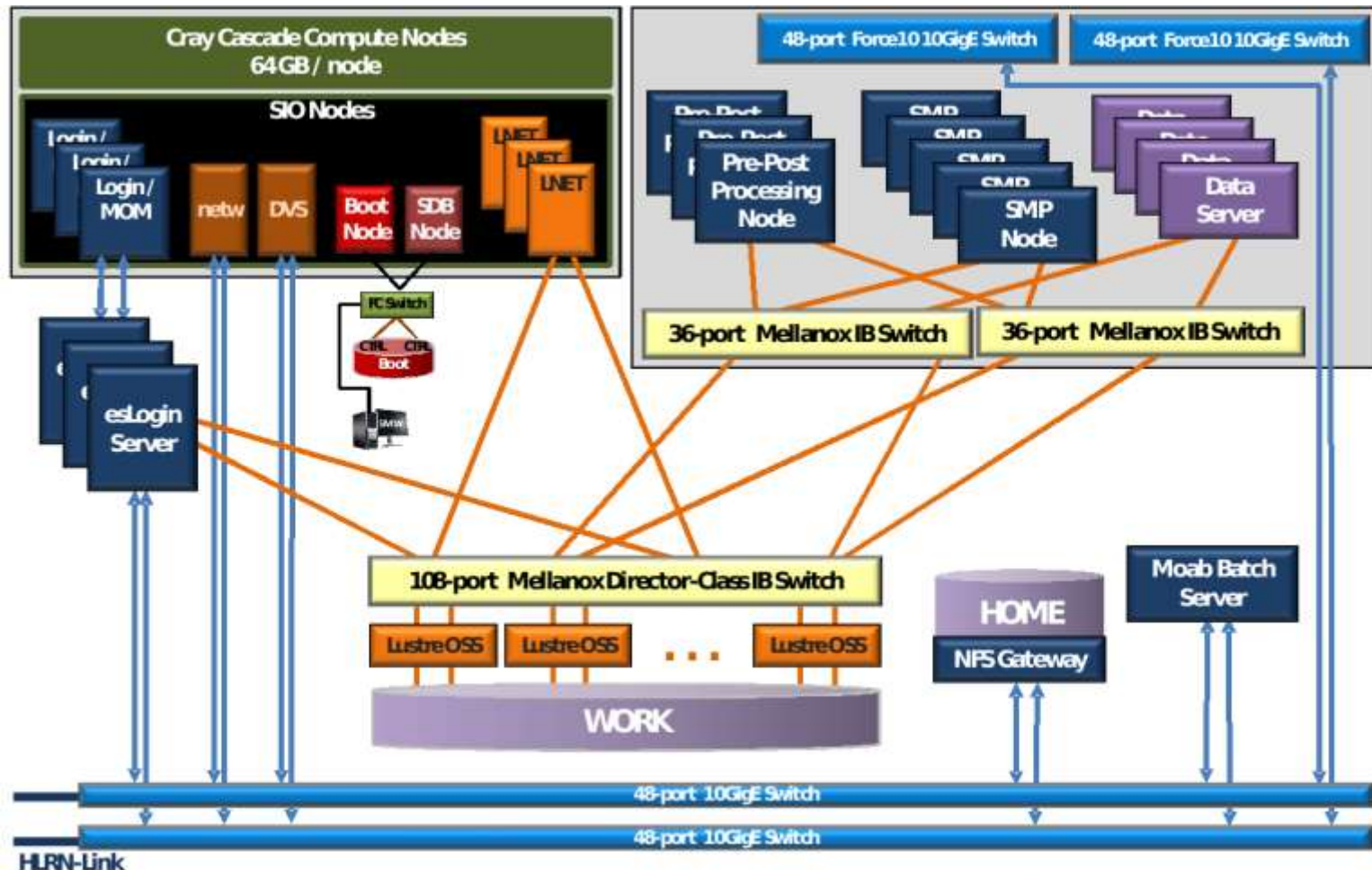
### Vorteile:

- mehr Personal
- Raid1 – Effekt bei Ausfällen und Wartung
- doppelte Datenhaltung möglich
- Betriebskosten werden geteilt

### Nachteile:

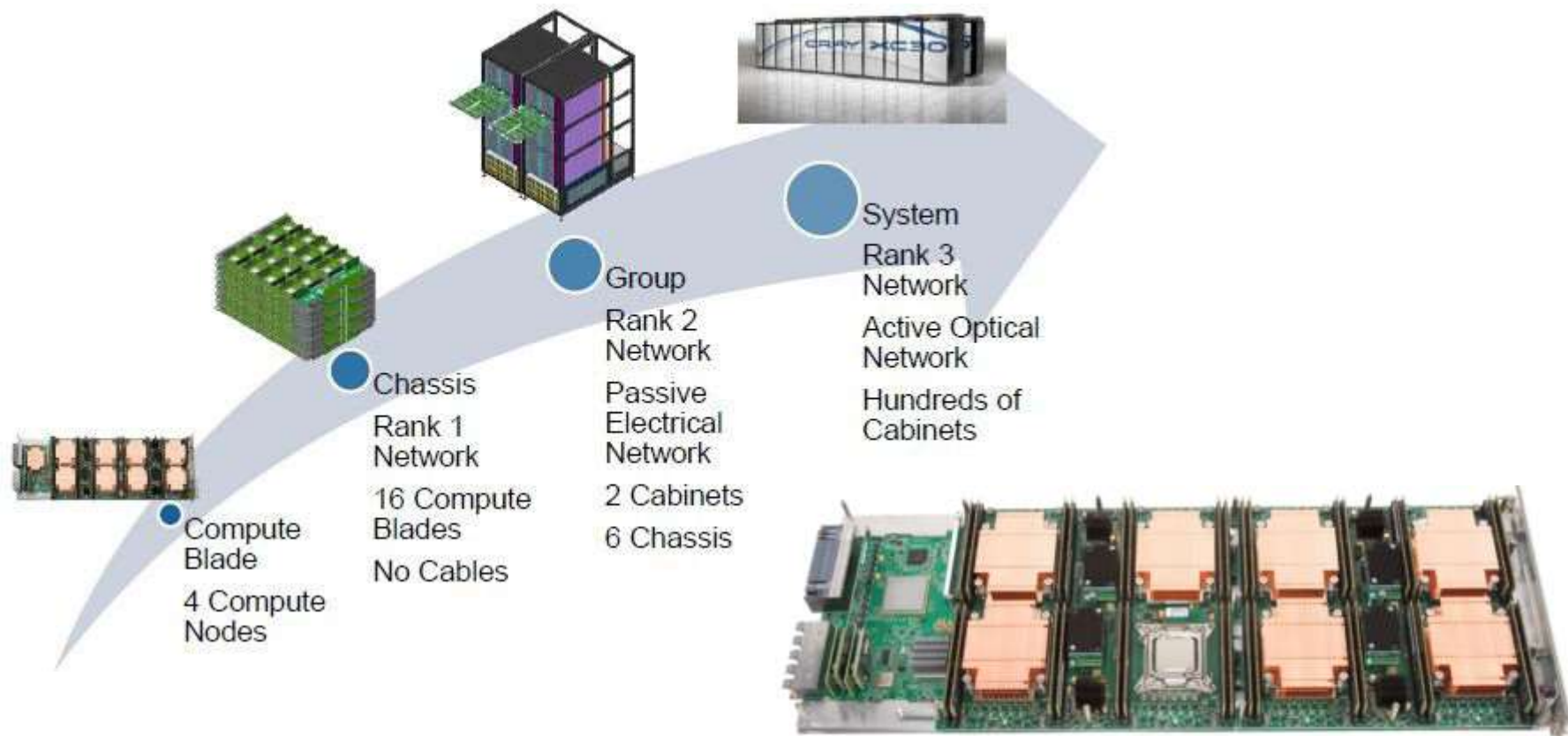
- Partitionierung – maximale Problemgröße halbiert
- Sites nach einiger Zeit leicht abweichend konfiguriert
- doppelte Datenhaltung nötig
- mehr Infrastruktur

# Schema einer Site

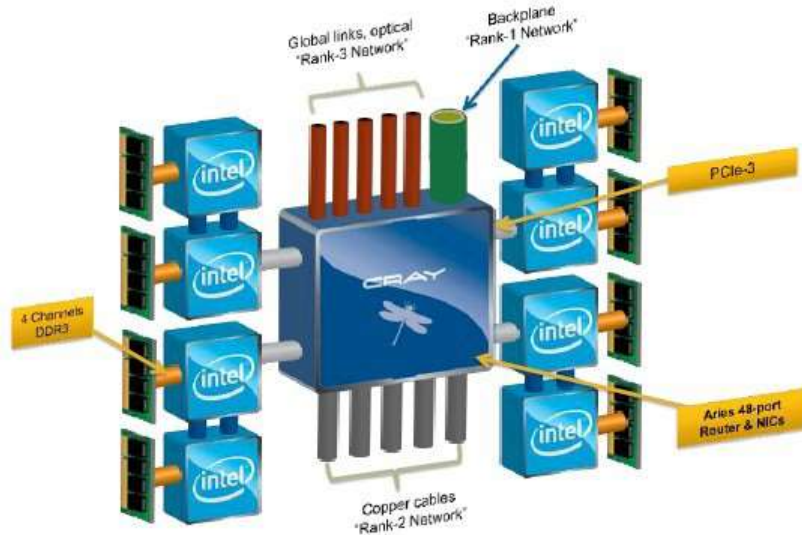


# HLRN-III Architektur

## ■ Aufbau der Cray XC30 (Cascade)



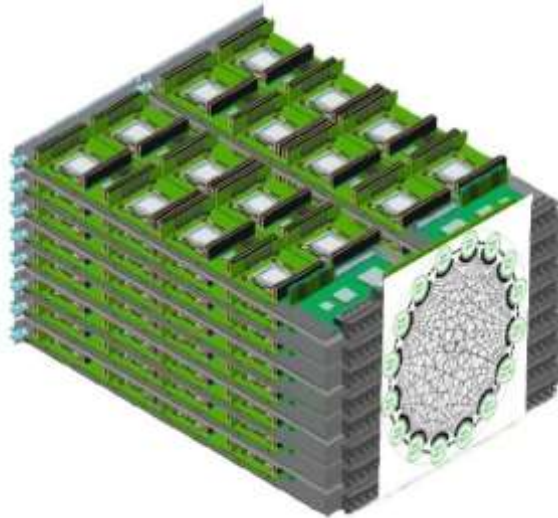
# Aries Netzwerk 1/2



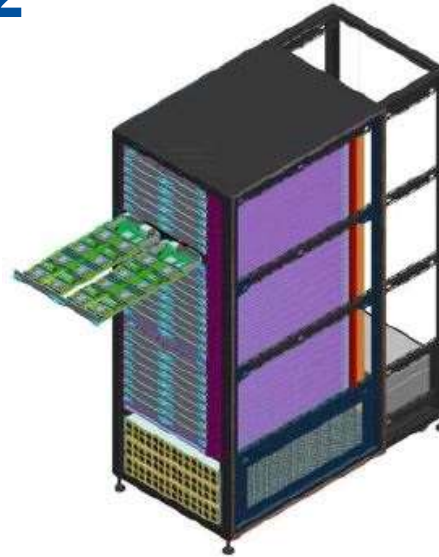
- Ein Aries Router pro Blade
  - direkt mit den PCIe 3.0 x16 Lanes eines Knotens verbunden.
- 1.3  $\mu$ s Latenz
- ~16GB/s Durchsatz (bidirektional) Knoten  $\leftrightarrow$  Router



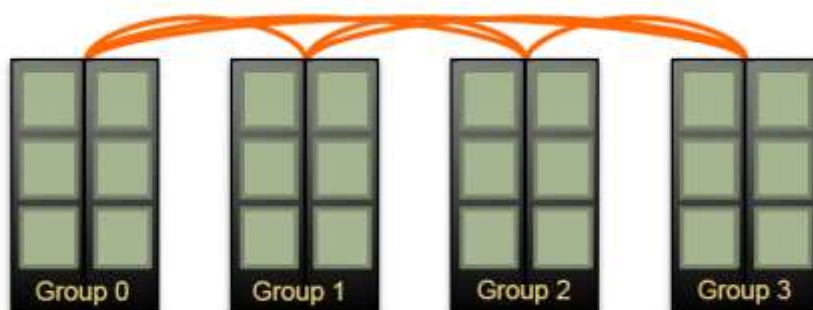
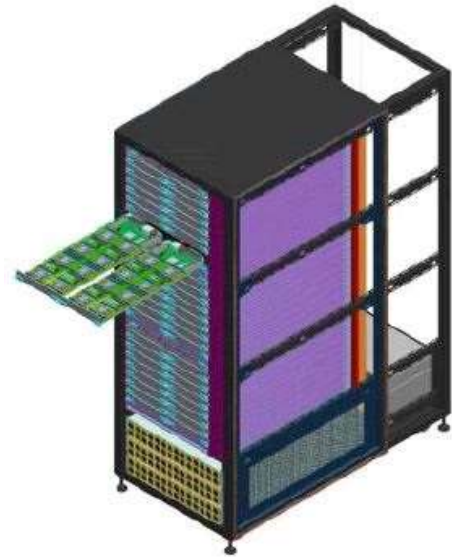
## Aries Netzwerk 2/2



16 Router werden über die Chassis-Backplane verbunden



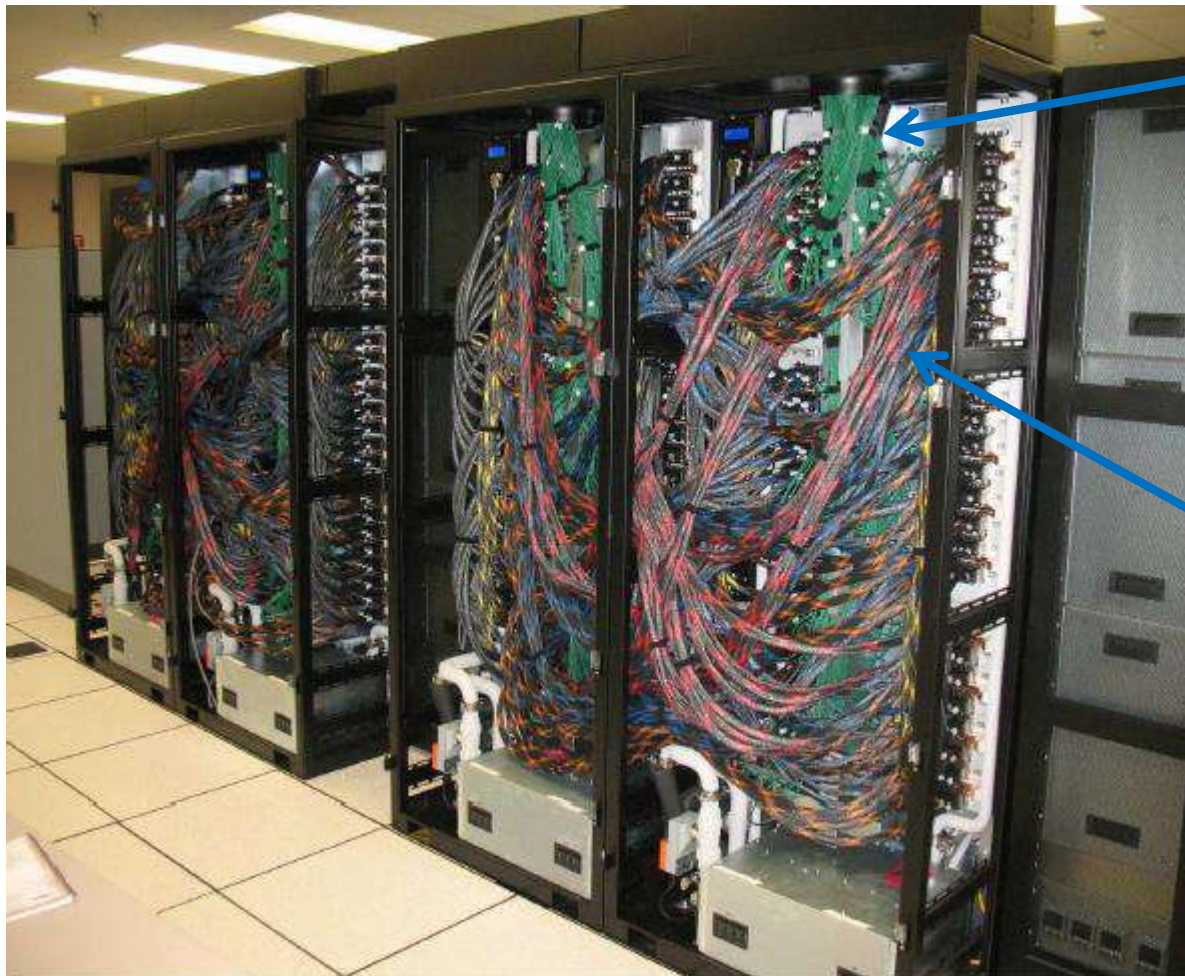
6 Chassis werden elektrisch verkabelt  
Rank-2 Copper Network



optische Verkabelung zwischen den Gruppen  
Rank-3 Network



# Copper & Optical Cabling



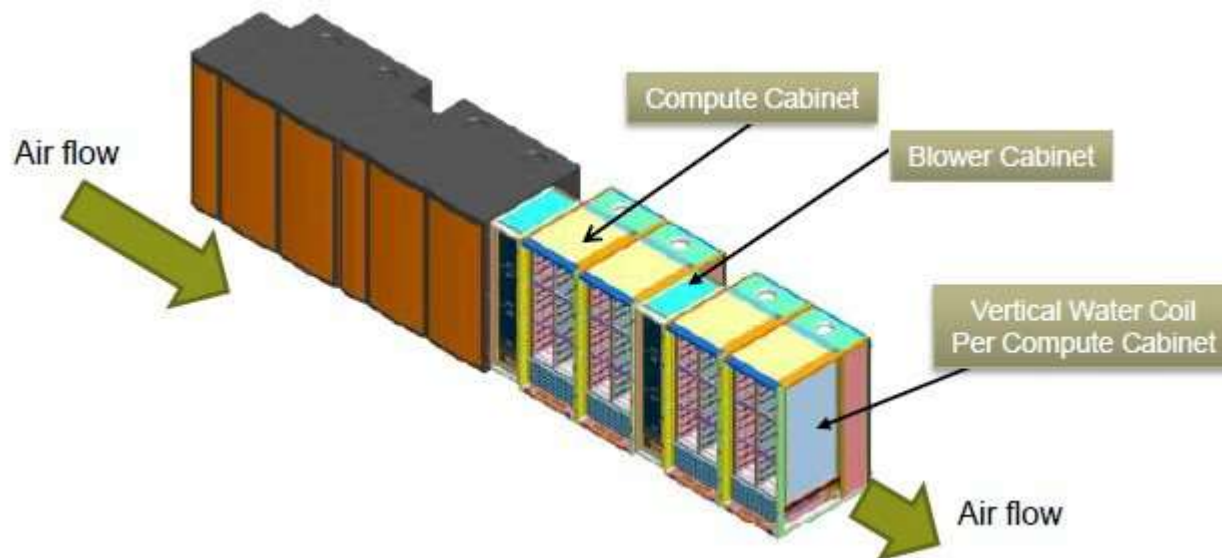
Rank-3  
Active Optics



Rank-2  
Passive CU

# Cray XC30 Kühlung

- erfolgt quer



# Notstromversorgung / USV



## HLRN-III Software (System)

- Betriebssysteme:
  - Cray Linux Environment (CLE) auf der XC30, XC40
  - SLES auf den SMP und Pre/Post Knoten
- Compiler: GNU, Intel und Cray
- MPI: Intel MPI und Cray mpt
- Intel MKL + Cray Scientific Libraries
- Allinea DDT Debugger
- Intel und Cray Performance und Analysis Tools
- Batch: Moab/Torque HPC Suite – Enterprise Edition

Ein Cray Mitarbeiter zur Anwenderunterstützung und Portierung von Open Source Paketen und Nutzerprogrammen

## HLRN-III Software (User)

- Bereitstellung von optimierten Anwenderprogrammen durch Module
- Ingenieurwissenschaft – abaqus, fluent, cfx . . .
- Chemie – cp2k, gromacs, namd, vasp . . .
- Bibliotheken – boost, netcdf, hdf5, fftw . . .
- Unterstützung beim Bau von Software durch Cray, Betreiberzentren und Fachberater



