Hawk Interconnect Network

Björn Dick (HLRS), Thomas Bönisch (HLRS), Bernd Krischok (HLRS)
• **InfiniBand HDR**
  • 200 Gbit/s bidirectional bandwidth per link, also individual nodes are connected to the network with 200 Gbit/s links!
  • MPI Latency \(\sim 1.3 \mu s\) (nearest neighbor)

• **Per switch chip:**
  • 40 Ports:
    • 16 nodes
    • 23 ports used to connect switches as a hypercube
    • one switch in a rack uses remaining port to attach filesystem
  \(\rightarrow\) fully non-blocking communication among 16 attached nodes
Interconnect topology

- 16 nodes connected to a common switch (represented by bullets)
- Switches arranged as a (partial enhanced) 9D hypercube
- i.e. by iteratively
  1. Doubling existing structures
  2. Connecting corresponding nodes
- More links (→ enhanced B/W) on lower dimensions (thicker lines)

1D line 4 links
2D square 4 links
3D cube 3 links
4D hypercube 2 links
... 2 links
9D (partial) hypercube

Established by an entire rack
Remarks

• On 3D computational domains, remaining 6 dimensions can be used to maintain proximity.

• We plan to deploy topology aware scheduling and MPI placement.
How to imagine higher dimensions?

- E.g. represent a 3D (hyper)cube by a single bullet.
- And also a 2$^\text{nd}$ 3D (hyper)cube.
- Connect the bullets in order to represent all the links between corresponding nodes of the 3D (hyper)cube.
- Now those “hyper”-nodes can be combined as seen before.
A bullet may represent a 5D hypercube.  
Then dimensions 6 to 8 can be visualized as a cube.  
Dimension 9 connects 8192 compute nodes.  
However, Hawk incorporates 5632 nodes only.  
So the 9D hypercube is truncated.