Introduction

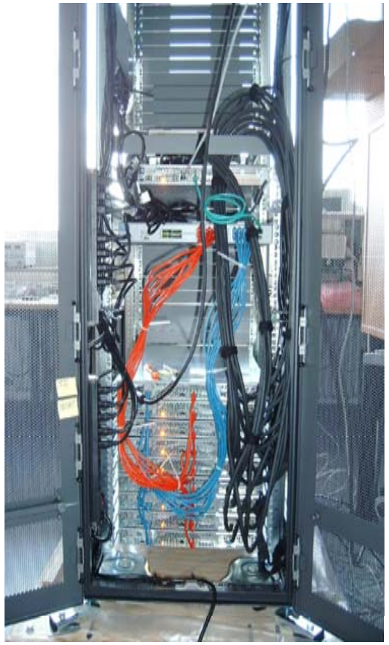
Innovative scalable HPC

Ulrich Bruening/Holger Fröning

June 2011

History

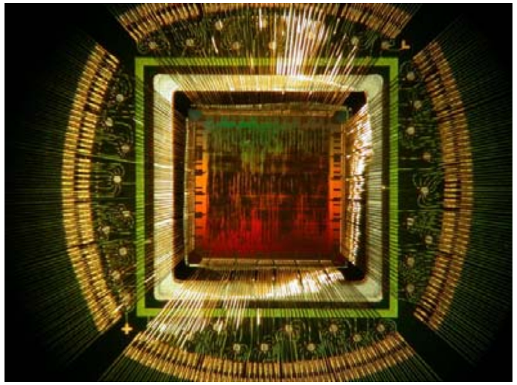
- Design of complex HW/SW systems, Computer Architecture Group,
  Prof. Dr. U. Brüning,
  University of Heidelberg
  - Computer architecture
  - Interconnection networks



ATOLL
cluster

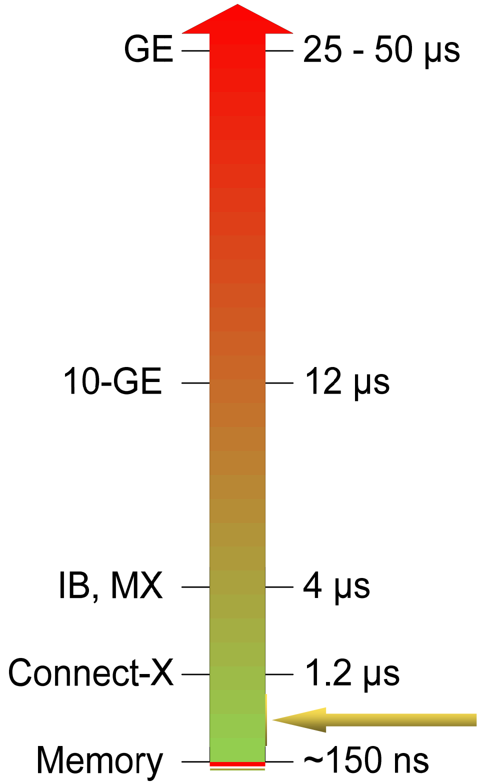
- EXTOLL project started in 2005
  - FPGA (Xilinx Virtex4 based) prototype since 2008

- Start-up company as a spin-off



ATOLL-Die
(bonded)

Introduction I



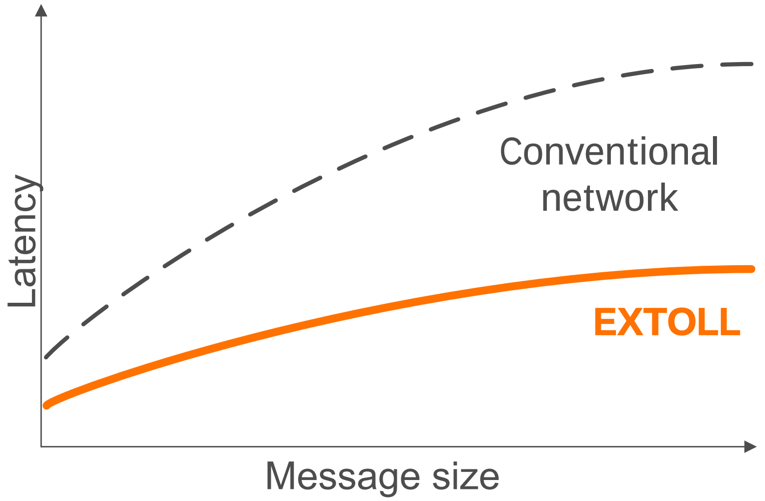
- Interconnection networks are the **key component** of parallel systems
- Prof. Patterson stated: “*Latency lags Bandwidth*”
- Need to significantly **lower communication latency** and improve communication in parallel systems
  - Finer grain parallelism
  - PGAS systems
  - Improve scaling
- **EXTOLL project at the CAG**

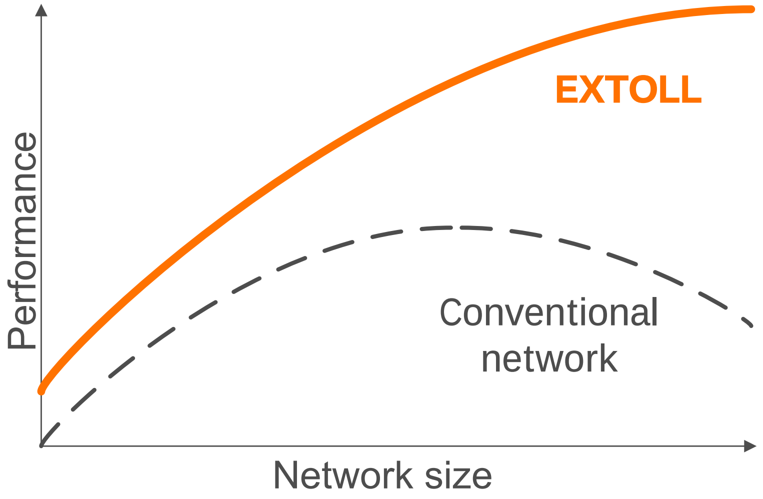
Vision: more performance, lower cost for HPC!

Introduction II

- Lowest latency
- Maximum message rate / s
- **Optimized for multi-core**
- Optimized CPU-Device interface

- Direct topology
- Efficiency
- **Innovative architecture**
- Optimized scalability

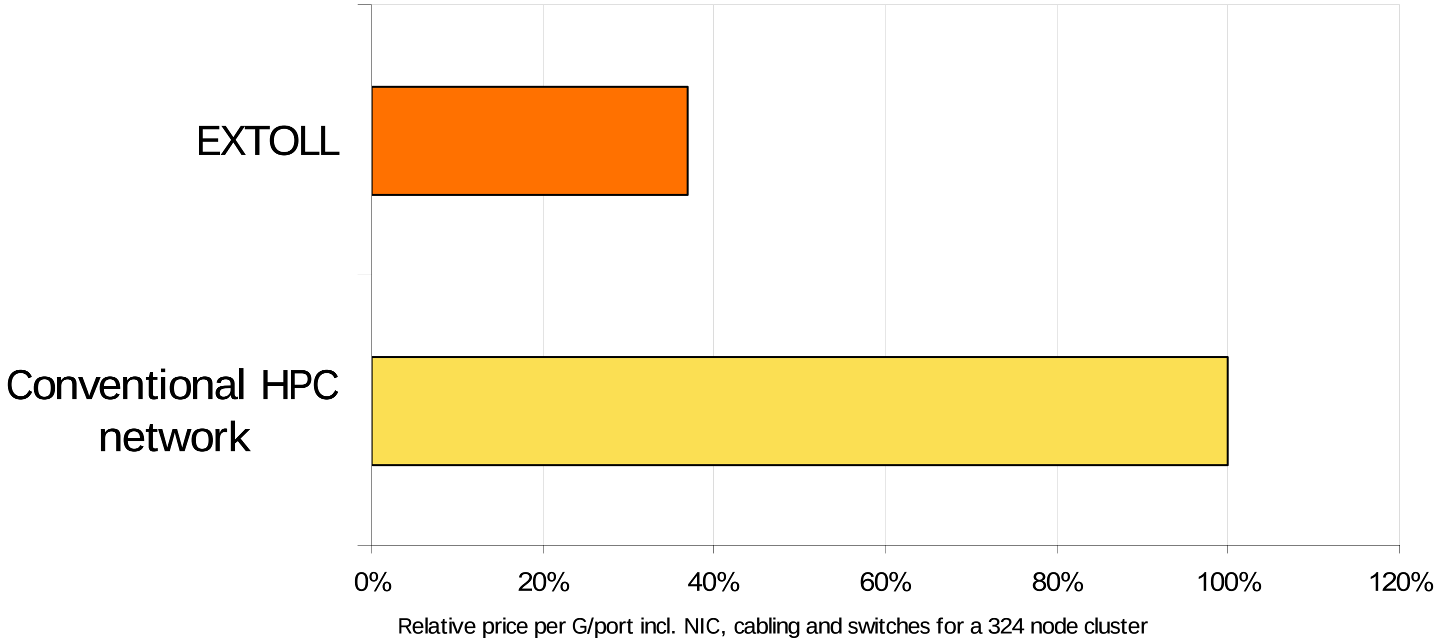




More Performance for HPC customer…

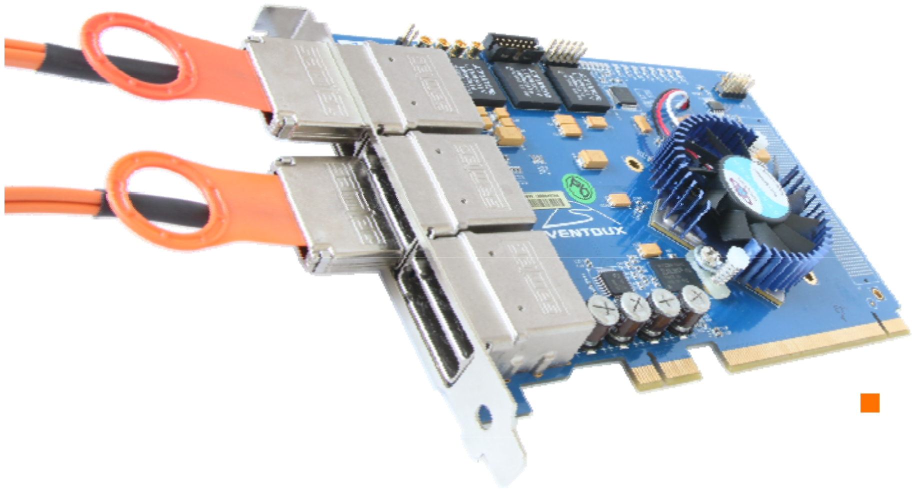
Introduction III

- **No** external switches
- Complete **own** IP
- **Lower** cost
- **Lower** energy consumption



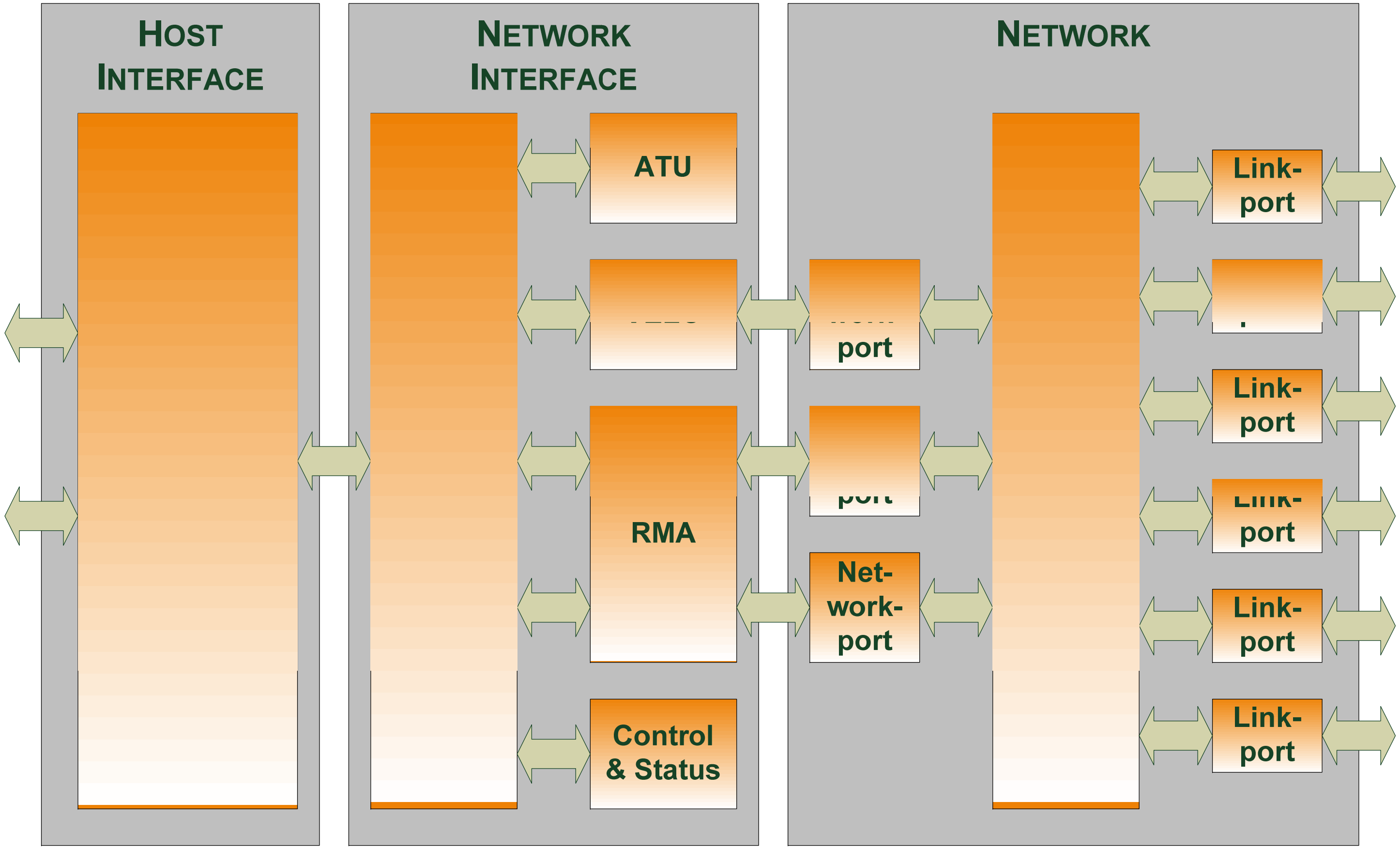
...lower cost

Architecture - Ideas

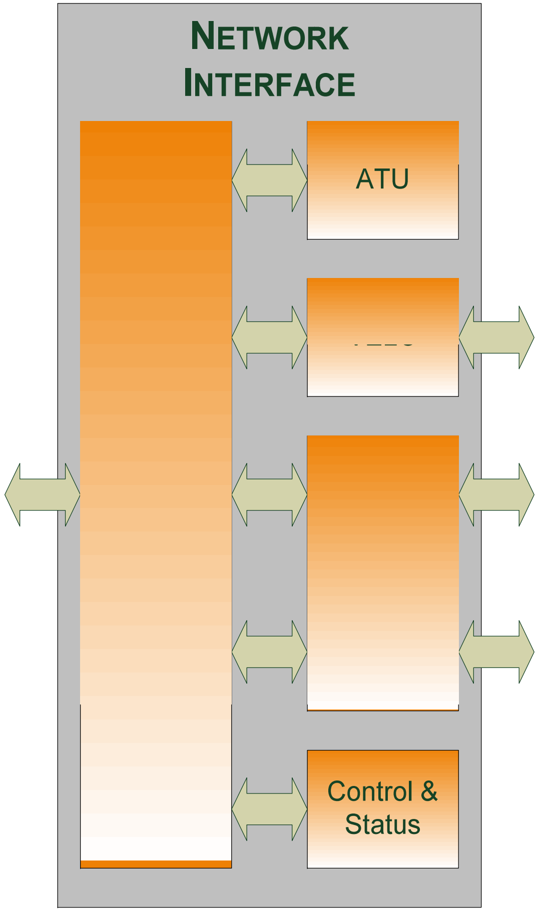


- Lean network interface
  - Ultra low latency message exchange
  - Extremely high **hardware** message rate
  - Small memory footprint
- **Switchless-Design** - 3D Torus Direct Network
  - Reliable network
  - High Scalability
  - Extremely efficient, pipelined hardware architecture

Architecture - Overview

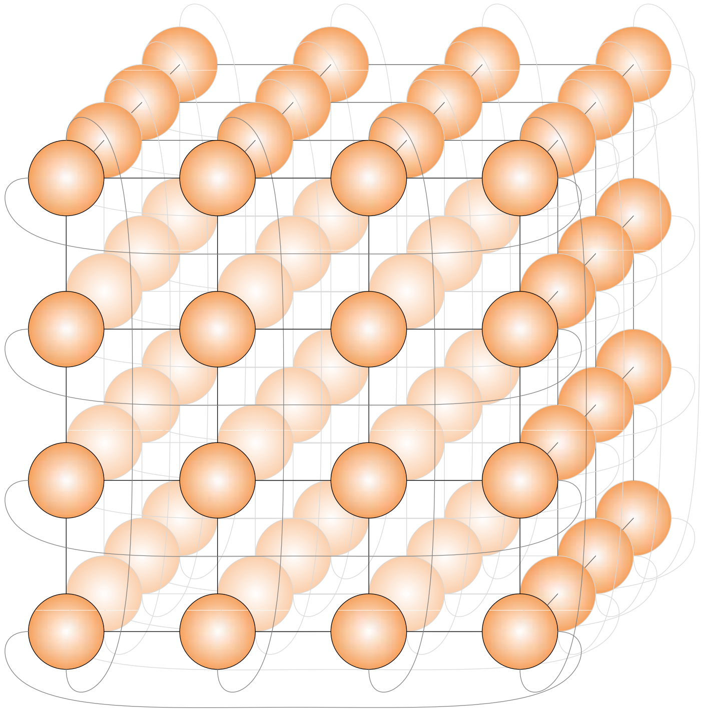


NIC Features



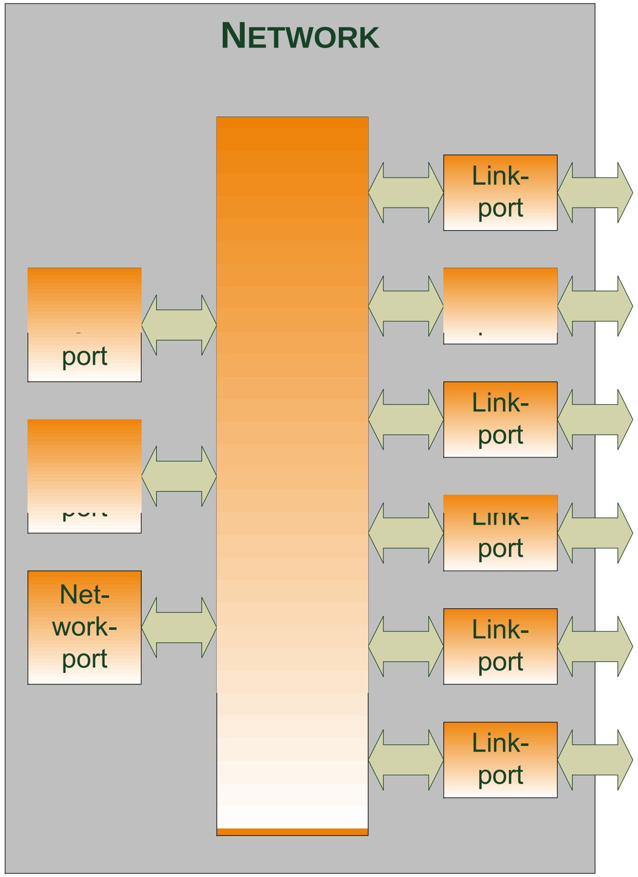
- VELO: Very fast two sided messaging
- RMA: Optimized access to remote memory
  - Local and remote completion notifications for all RMA operations
- Fully virtualized
  - user space and/or different virtual machines
  - secure
- Hardware address translation unit (ATU) including TLB

Network Features I



- 64k nodes
- hundreds of endpoints per node
- Efficient network protocol
  - low overhead even for very small packets
- Support for arbitrary direct topologies
- Choice for implementation: 6 links
- Natural topology 3-d torus

Network Features II

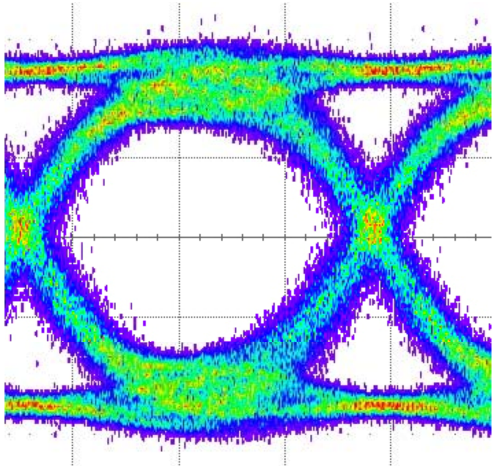


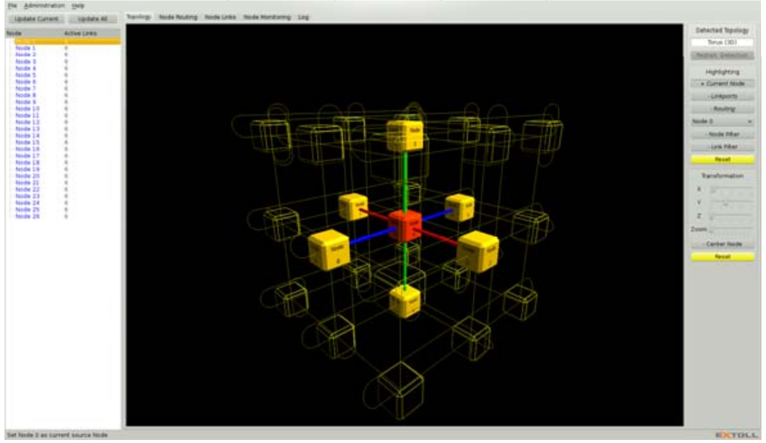
- Adaptive and deterministic routing
- Packets routed deterministically are delivered in-order

- Three virtual channels (VCs)
- Four independent traffic classes (TCs)

- Support for remote configuration and monitoring of EXTOLL nodes without host interaction

Reliability & Security Features





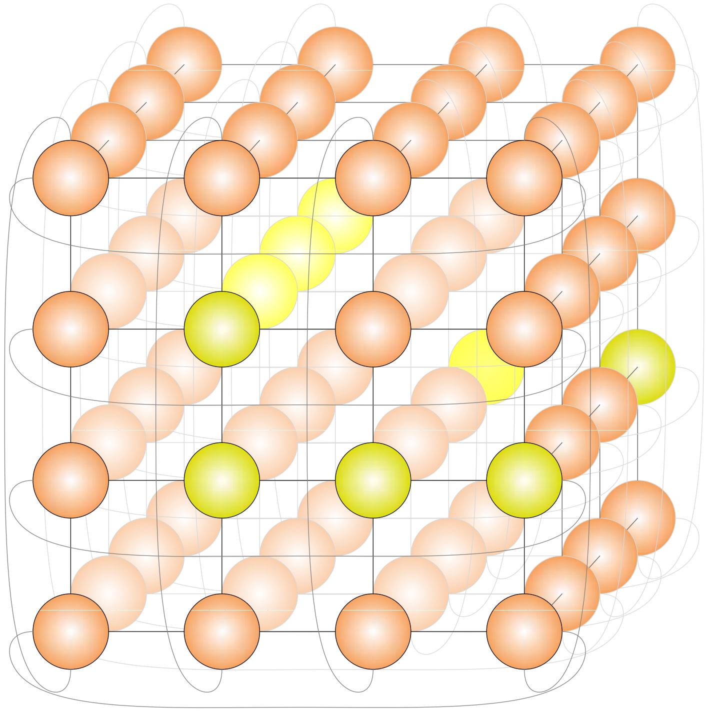
Node	Active Links
[illegible]	[illegible]
Node 1	6
Node 2	6
Node 3	6
Node 4	6
Node 5	6
Node 6	6
Node 7	6
Node 8	6
Node 9	6
Node 10	6
Node 11	6
Node 12	6
Node 13	6
Node 14	6
Node 15	6
Node 16	6
Node 17	6
Node 18	6
Node 19	6
Node 20	6
Node 21	6
Node 22	6
Node 23	6
Node 24	6
Node 25	6
Node 26	6

Set Node 0 as current source Node

- Reliable message transport
- Link level retransmission protocol for reliable data transmission
- All network protocol elements are secured by strong CRCs
- All internal memory protected by ECC for high reliability

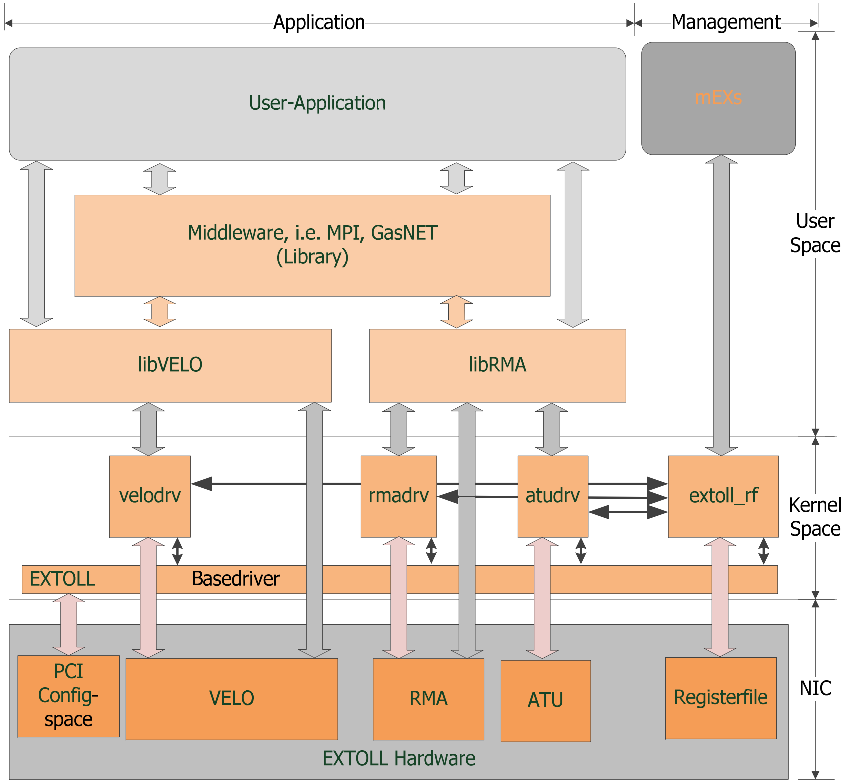
- Isolation of communication groups in the network
- Process isolation on the node

Additional Features



- Scalable hardware barriers implemented completely in hardware
- Global interrupts with low skew
- Hardware support for multicast operations
- Non-coherent shared memory features (for PGAS)

Architecture -Software

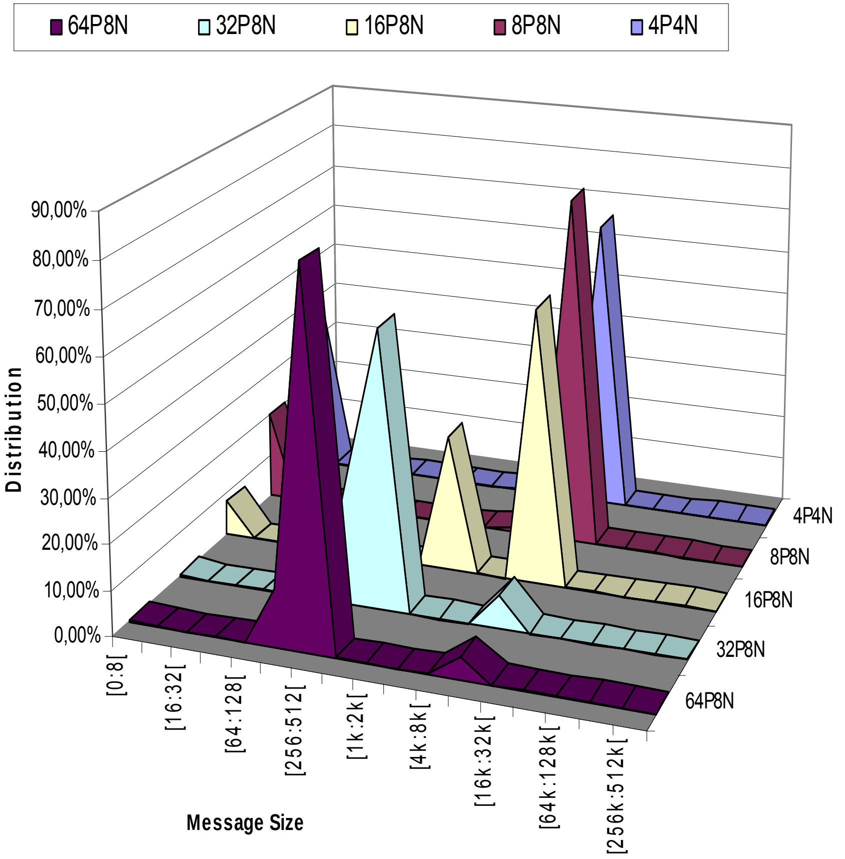


- Optimized for HPC-Users
- OS bypass
- Linux kernel drivers manage resources
- Low-level API libraries
- MPI support
  - OpenMPI
- PGAS support
  - GASNet (prototype)
- Management software

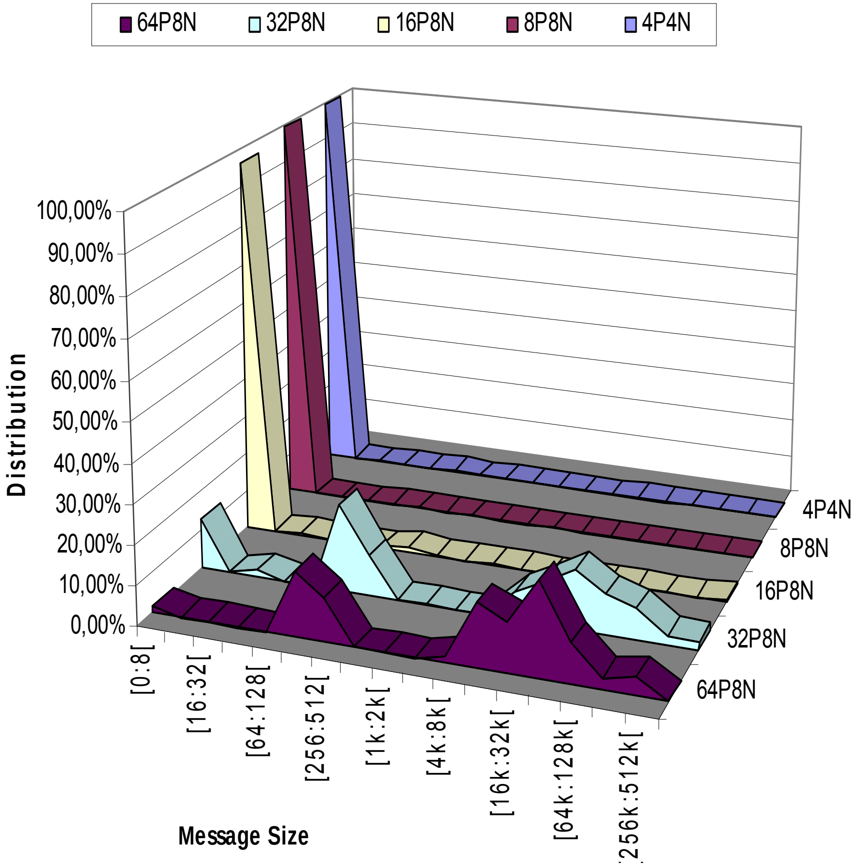
Why do I actually need support for small messages?

Use of Small Messages in HPC

HPCC RandomAccess - Message Size Histogram



WRF - Message Size Histogram



Shift towards **smaller messages** with increasing process count

About **30%** of all messages have a size below 512B

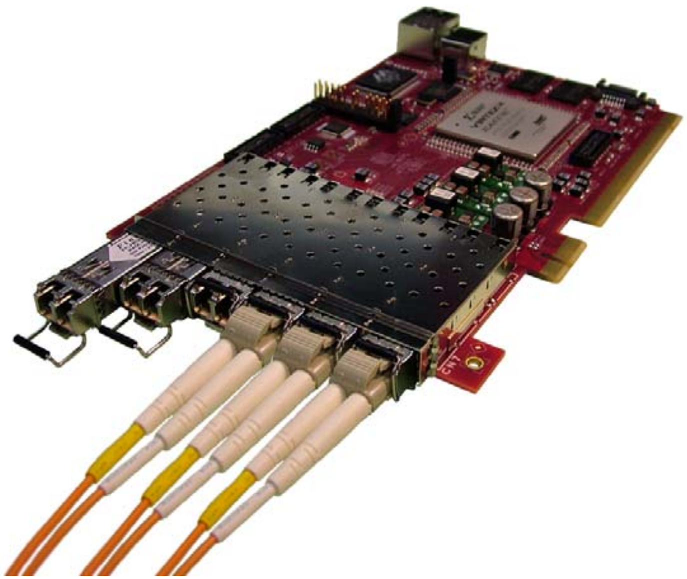
EXTOLL

A Small Test Drive of our Prototype...

Prototype Performance Results

EXTOLL FPGA Prototyp

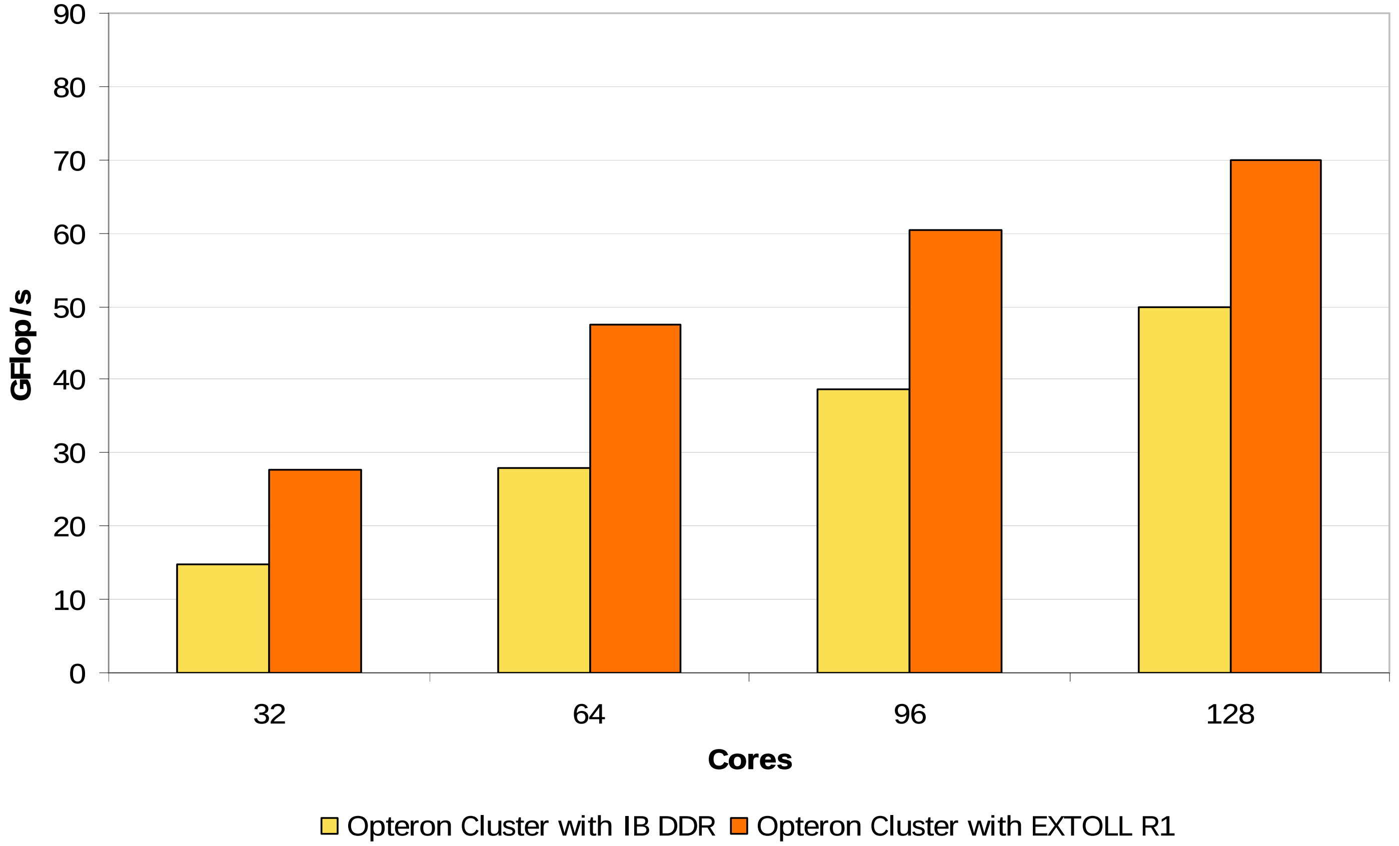
- Xilinx Virtex 4 based
- HT400 16-bit
- 150 MHz core frequency
- 6 optical links
- 6 Gbit/s link bandwidth
- ~ 1 µs end-to-end latency
- FPGA 90% filled



Test machines

- 2 node Quad Socket Quad-Core Opteron 8354 „Barcelona“ 2.2 GHz
  - 16GB, Supermicro 1041MT2B/H8QME-2+, Open MPI 1.3.3
  - Mellanox ConnectX IB SDR and SDR IB Switch, Open MPI 1.3.2
- 16 node dual-socket Opteron 8380 „Shanghai“ 2.5 GHz

EXTOLL Virtex4: WRF V3



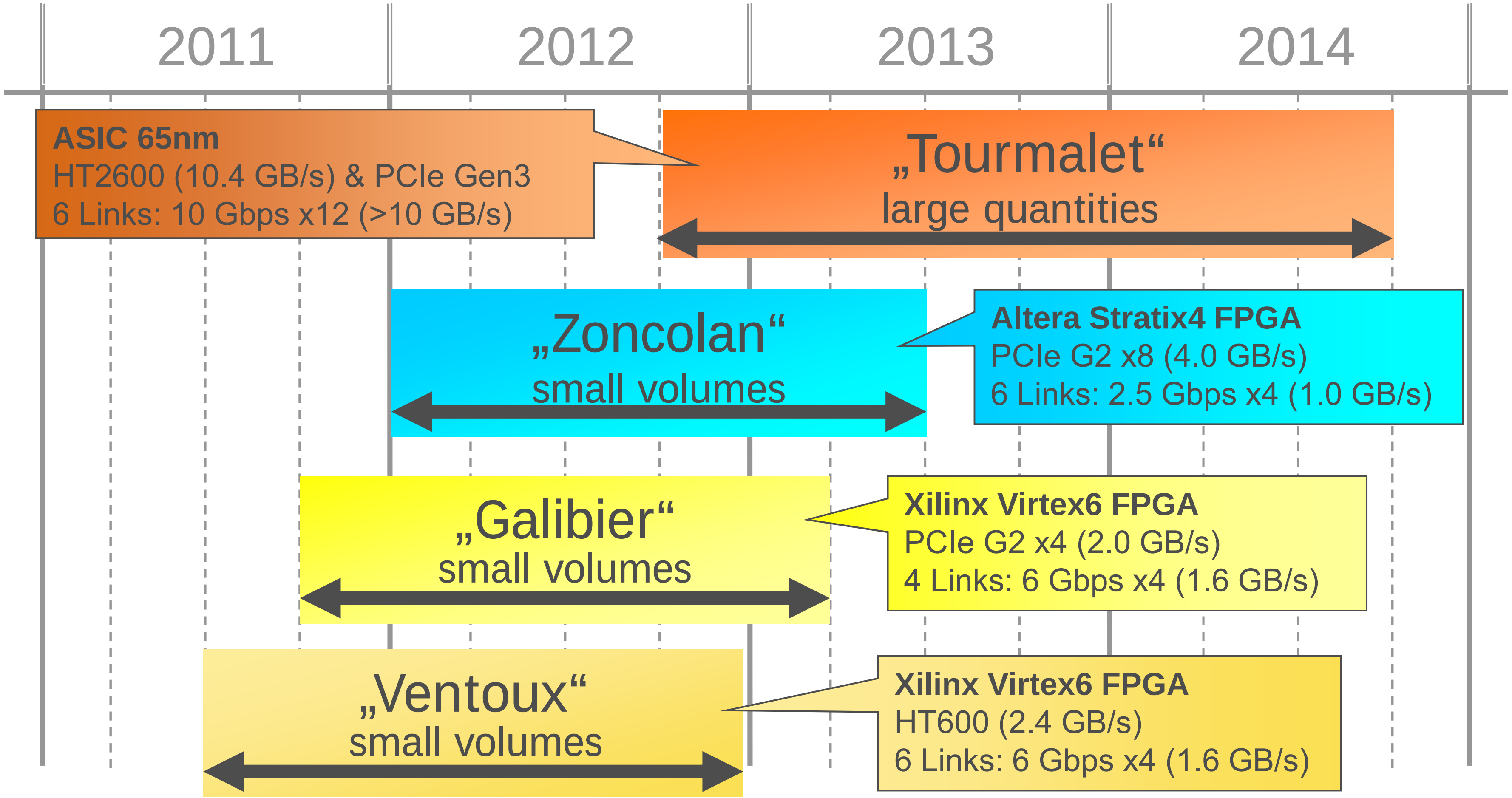
IB Cluster: Opteron Shanghai 2358 2.6GHz

EXTOLL Cluster: Opteron Shanghai 2380 2.5GHz

EXTOLL

What‘s next?

Hardware Roadmap



ASIC 65nm
HT2600 (10.4 GB/s) & PCIe Gen3
6 Links: 10 Gbps x12 (>10 GB/s)

„Tourmalet“
large quantities

„Zoncolan“
small volumes

Altera Stratix4 FPGA
PCIe G2 x8 (4.0 GB/s)
6 Links: 2.5 Gbps x4 (1.0 GB/s)

„Galibier“
small volumes

Xilinx Virtex6 FPGA
PCIe G2 x4 (2.0 GB/s)
4 Links: 6 Gbps x4 (1.6 GB/s)

„Ventoux“
small volumes

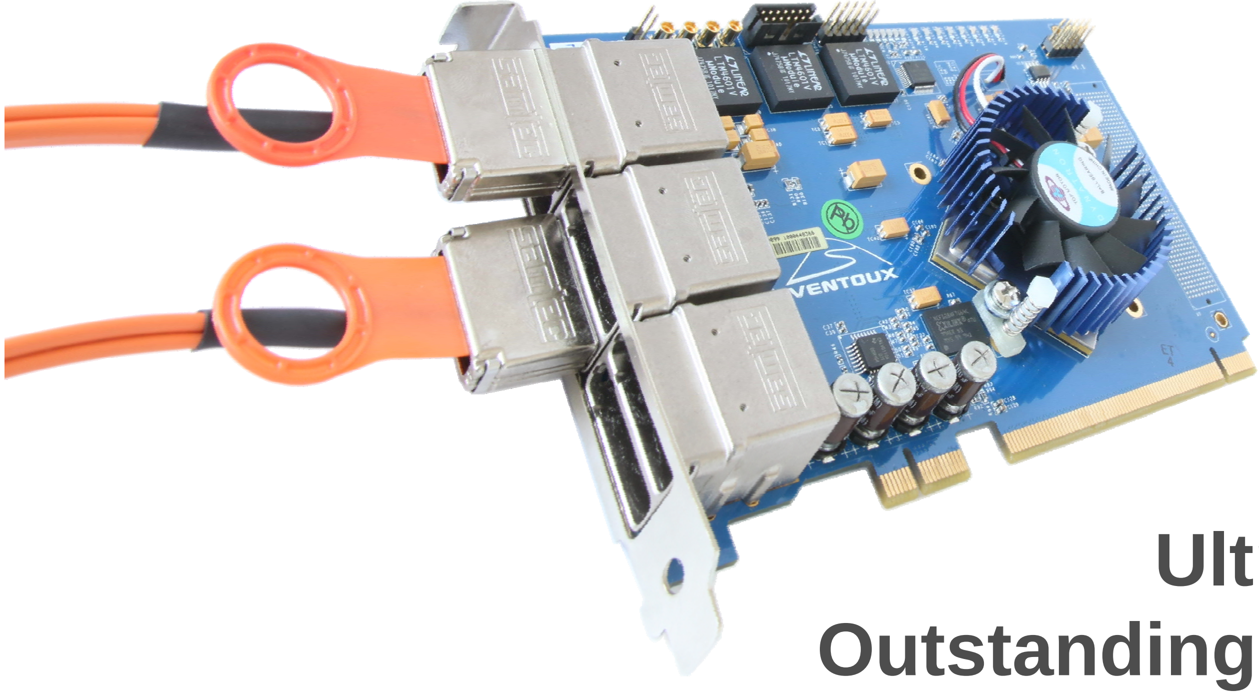
Xilinx Virtex6 FPGA
HT600 (2.4 GB/s)
6 Links: 6 Gbps x4 (1.6 GB/s)

Performance

Metric	EXTOLL Ventoux	EXTOLL Tourmalet	IB	Gigabit Ethernet	10G Ethernet
Start-up latency	<1.0us	<0.5us*	1.2us	30us	>10us
Hop latency	<150ns	<50ns*	100-450ns	>1000ns	>500ns
Peak BW	1.6GB/s	10GB/s*	4-8GB/s	125MB/s	1.25GB/s
Message Rate	>25M/s	~100M/s	7-23.7M/s		<2.5M/s

(*) values from simulation

Next Generation: „Ventoux“



VENTOUX

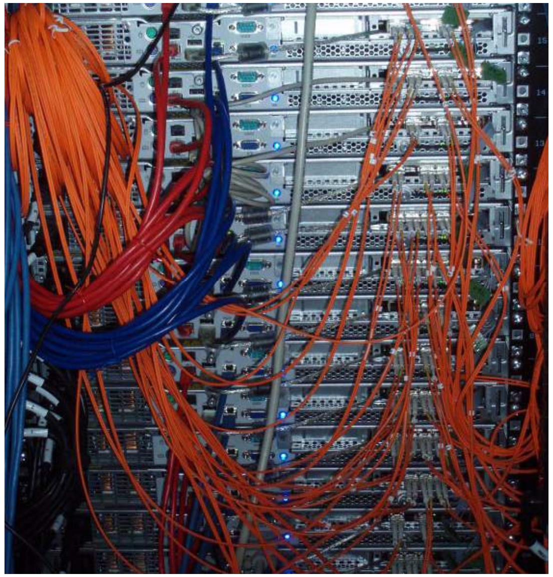
Ultra-low Latency!

Outstanding Message Rate!

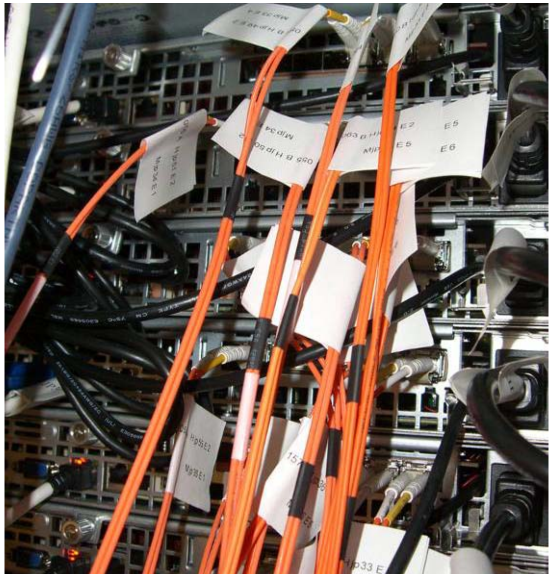
Presented at ISC 2011 in Hamburg

EXTOLL

Thank You!



Prototype cluster, Mannheim



Prototype cluster, Valencia, Detail

Prototype cluster, Valencia

Gefördert durch:

Bundesministerium
für Wirtschaft
und Technologie

aufgrund eines Beschlusses
des Deutschen Bundestages

Existenzgründungen aus der Wissenschaft
Ein Programm des Bundesministeriums für Wirtschaft und Technologie

eXIST

ziti

Computer Architecture Group,
University of Heidelberg