Introduction

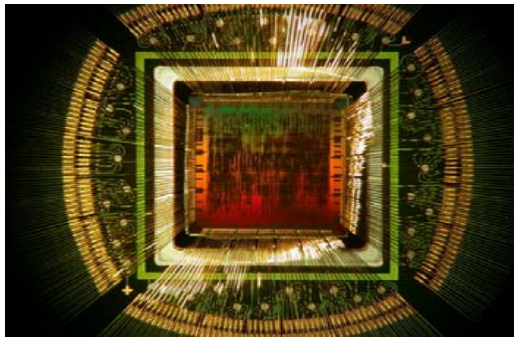
Dr. Holger Fröning
October 2010
info@extoll.de

History

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

ATOLL cluster

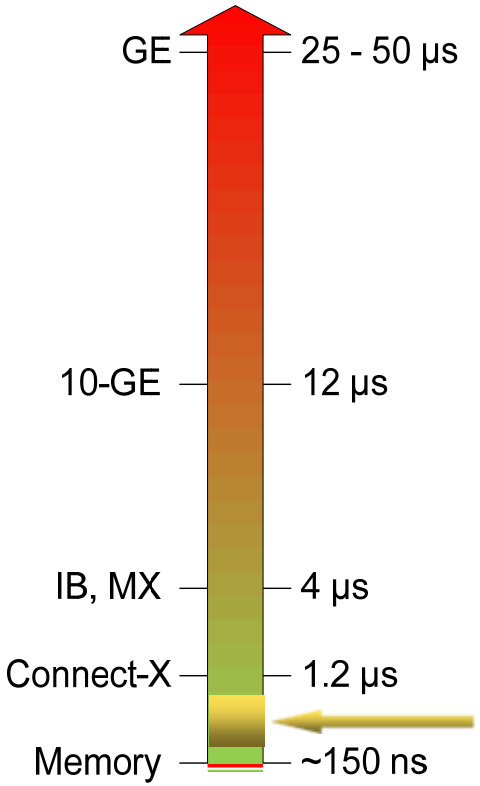
- EXTOLL project started in 2005
  - FPGA (Xilinx Virtex4 based) prototype since 2008



ATOLL-Die (bonded)

- Start-up company as a spin-off in 2010

Introduction I



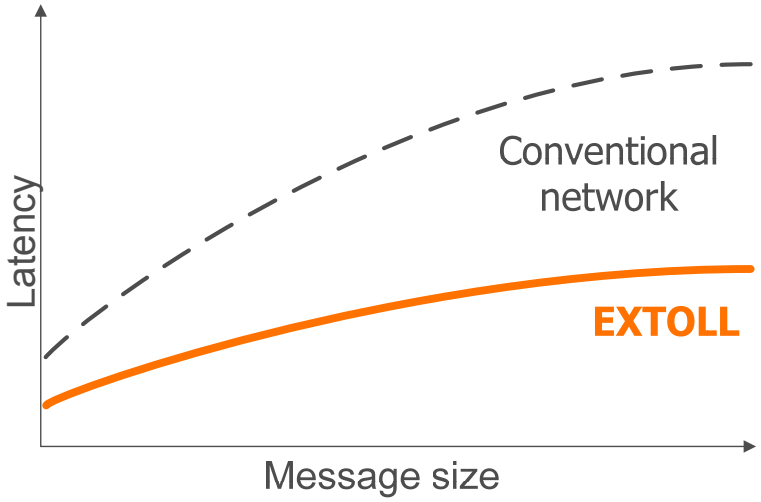
- Interconnection networks are the **key component** of parallel systems
- Prof. Patterson stated: "*Latency lags Bandwidth*"

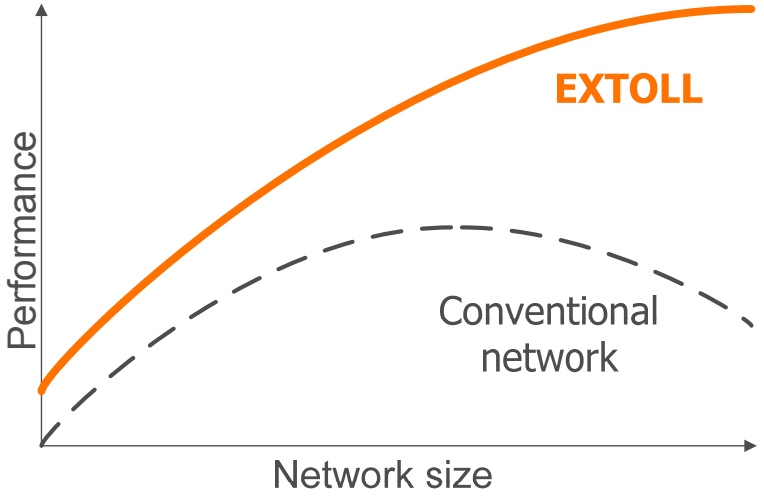
- Need to significantly **lower communication latency** and improve communication in parallel systems
  - Finer grain parallelism
  - PGAS systems
  - Improve scaling
- **EXTOLL project at the CAG**

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

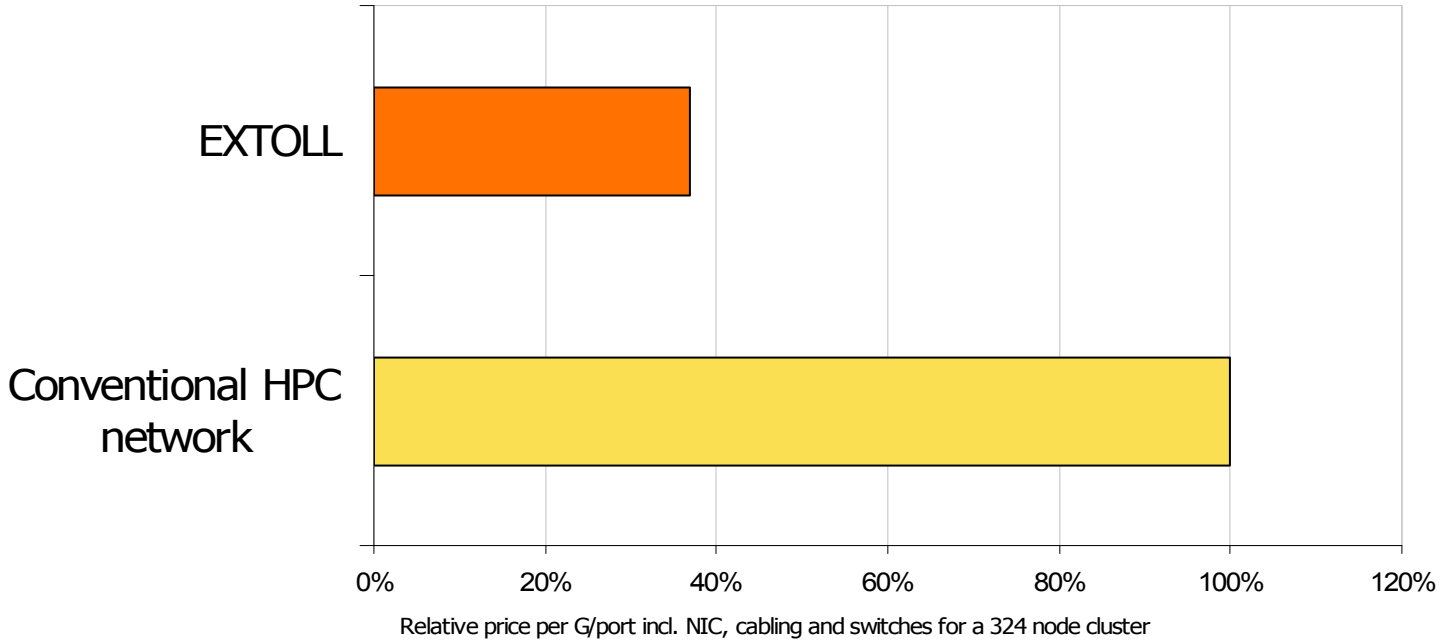




More Performance for HPC customer…

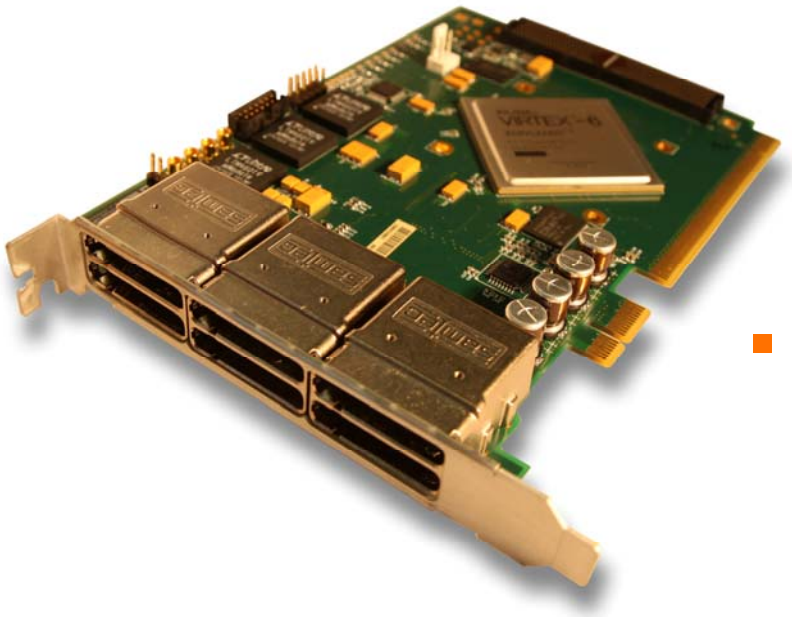
Introduction III

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



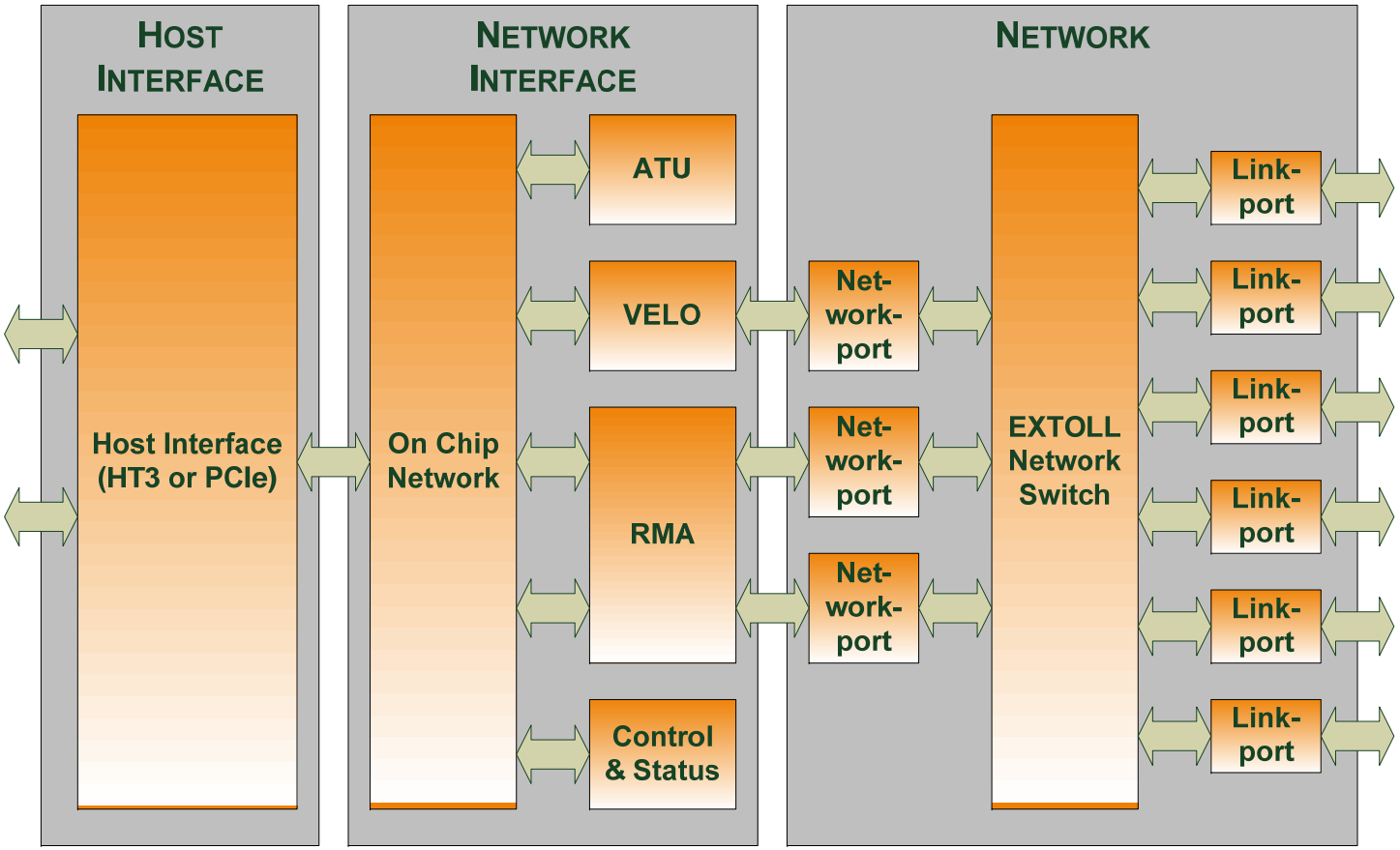
...lower cost

Architecture - Ideas

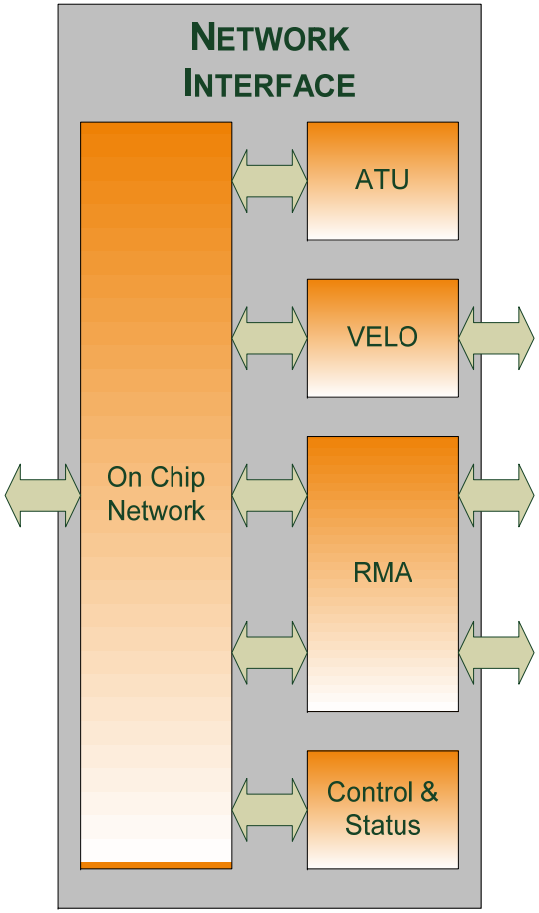


- Lean network interface
  - Ultra low latency message exchange
  - Extremely high **hardware** message rate
  - Small memory footprint
- **Switchless-Design** - 3D Torus Direct Network
  - Reliable network
  - High Scalability
  - Extremely efficient, pipelined hardware architecture

Architecture - Overview

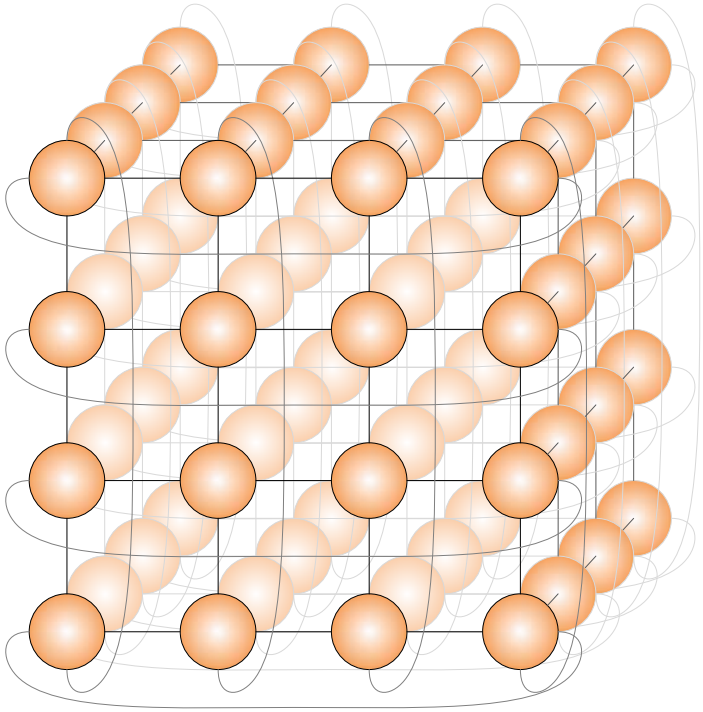


NIC Features



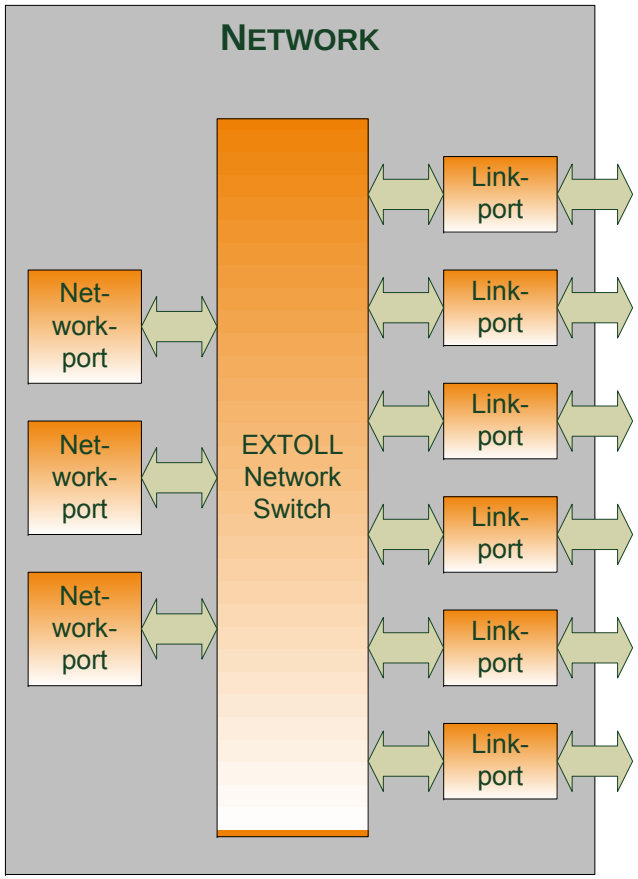
- VELO: Very fast two sided messaging
- RMA: Optimized access to remote memory
  - Local and remote completion notifications for all RMA operations
- Fully virtualized
  - user space and/or different virtual machines
  - secure
- Hardware address translation unit (ATU) including TLB

Network Features I



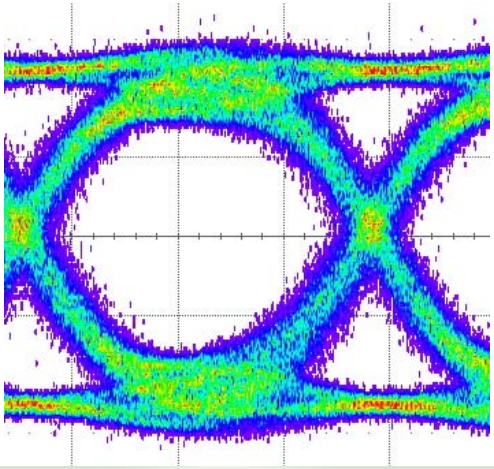
- 64k nodes
- hundreds of endpoints per node
- Efficient network protocol
  - low overhead even for very small packets
- Support for arbitrary direct topologies
- Choice for implementation: 6 links
- Natural topology 3-d torus

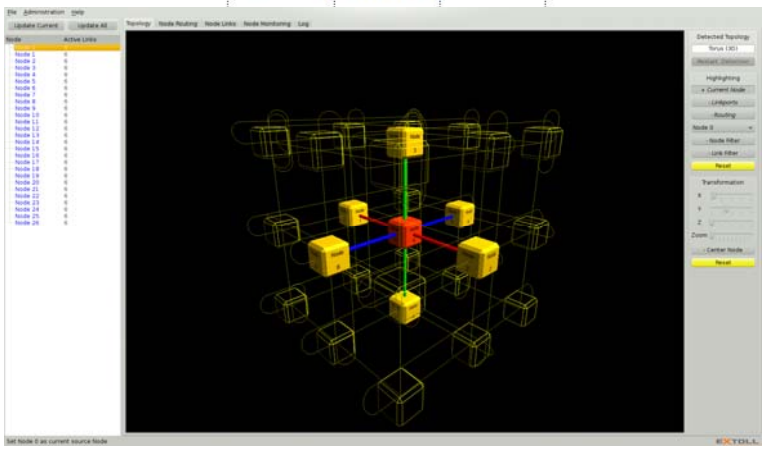
Network Features II



- 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

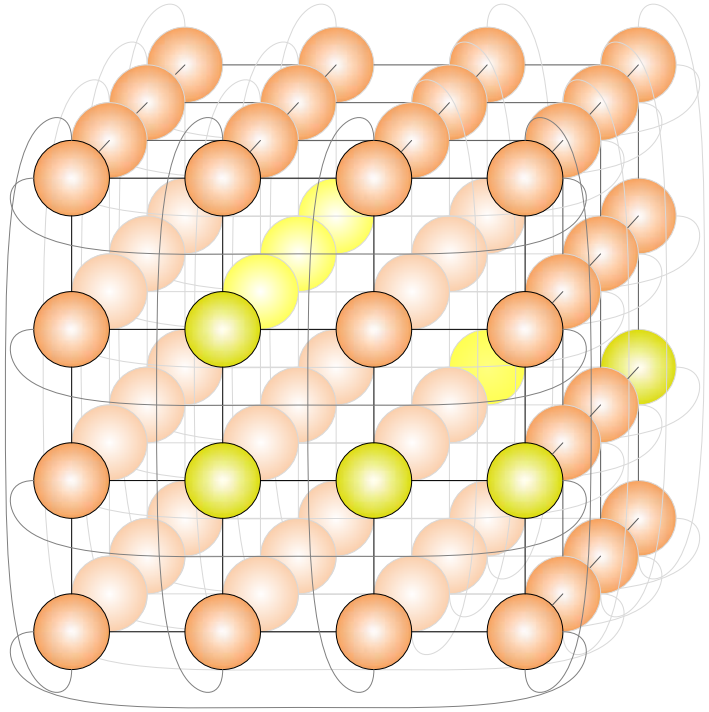




- Reliable message transport
- Link level retransmission protocol for reliable data transmission
- All network protocol elements are secured by strong CRCs
- All internal memory protected by ECC for high reliability

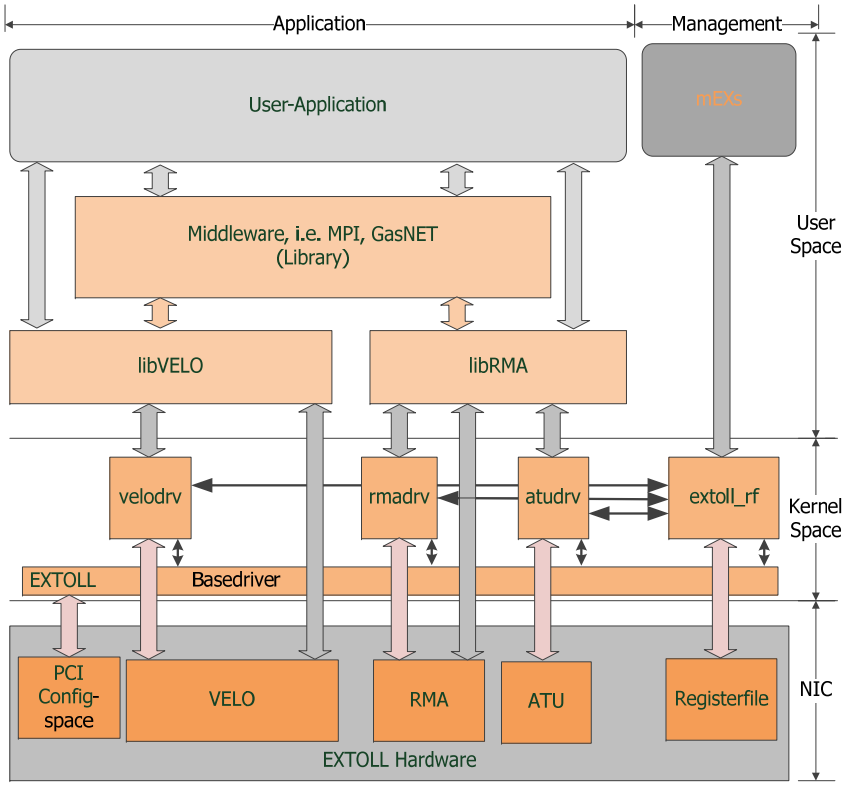
- Isolation of communication groups in the network
- Process isolation on the node

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

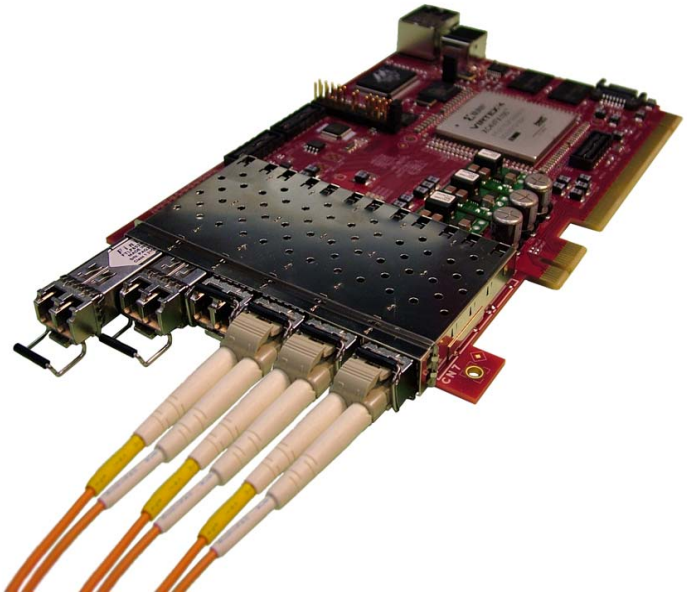


- Optimized for HPC-Users
- OS bypass
- Linux kernel drivers manage resources
- Low-level API libraries
- MPI support
  - OpenMPI
- PGAS support
  - GASNet (prototype)
- Management software

Prototype Performance Results

EXTOLL FPGA Prototyp

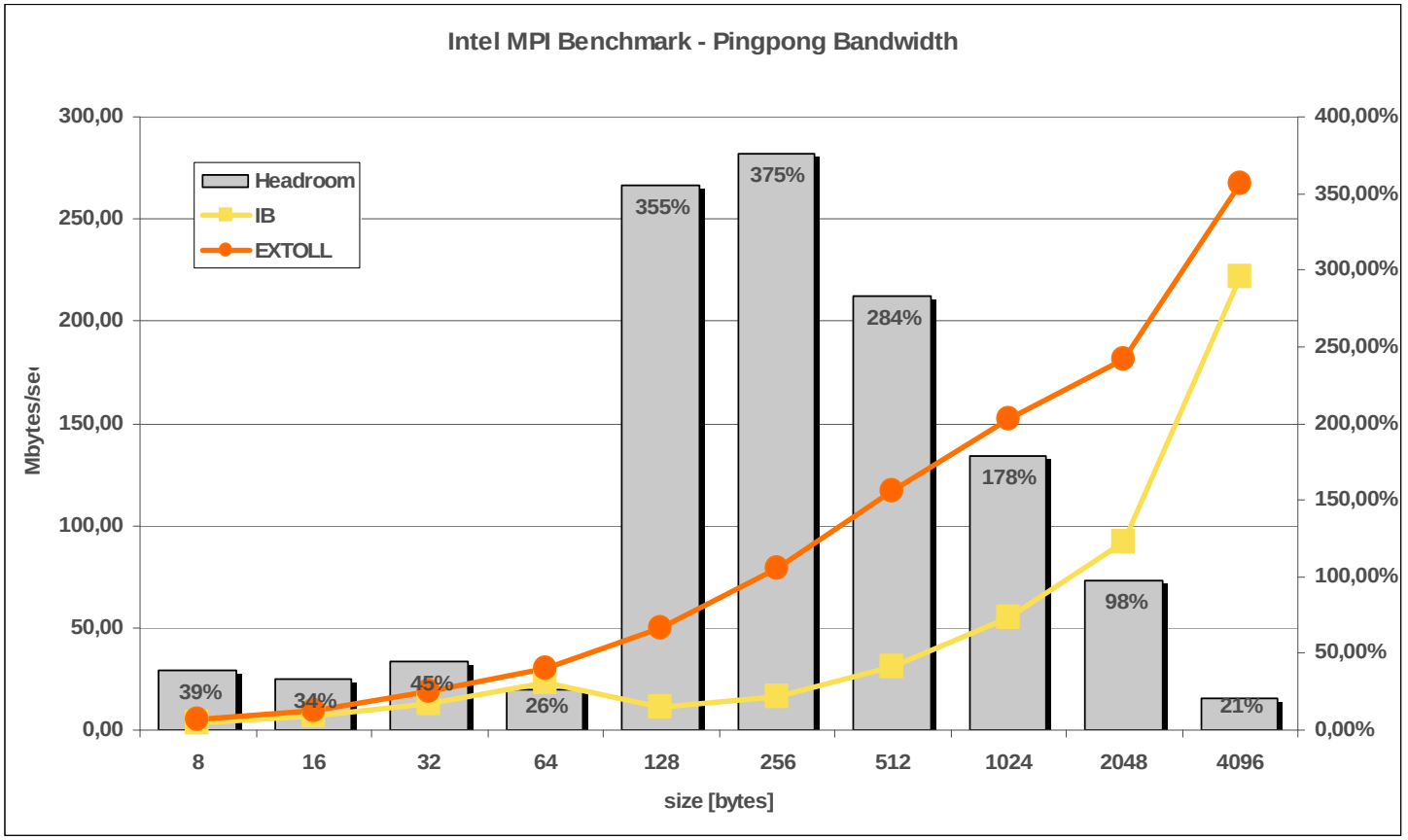
- Xilinx Virtex 4 based
- HT400 16-bit
- 150 MHz core frequency
- 6 optical links
- 6 Gbit/s link bandwidth
- ~ 1 µs end-to-end latency
- FPGA 90% filled



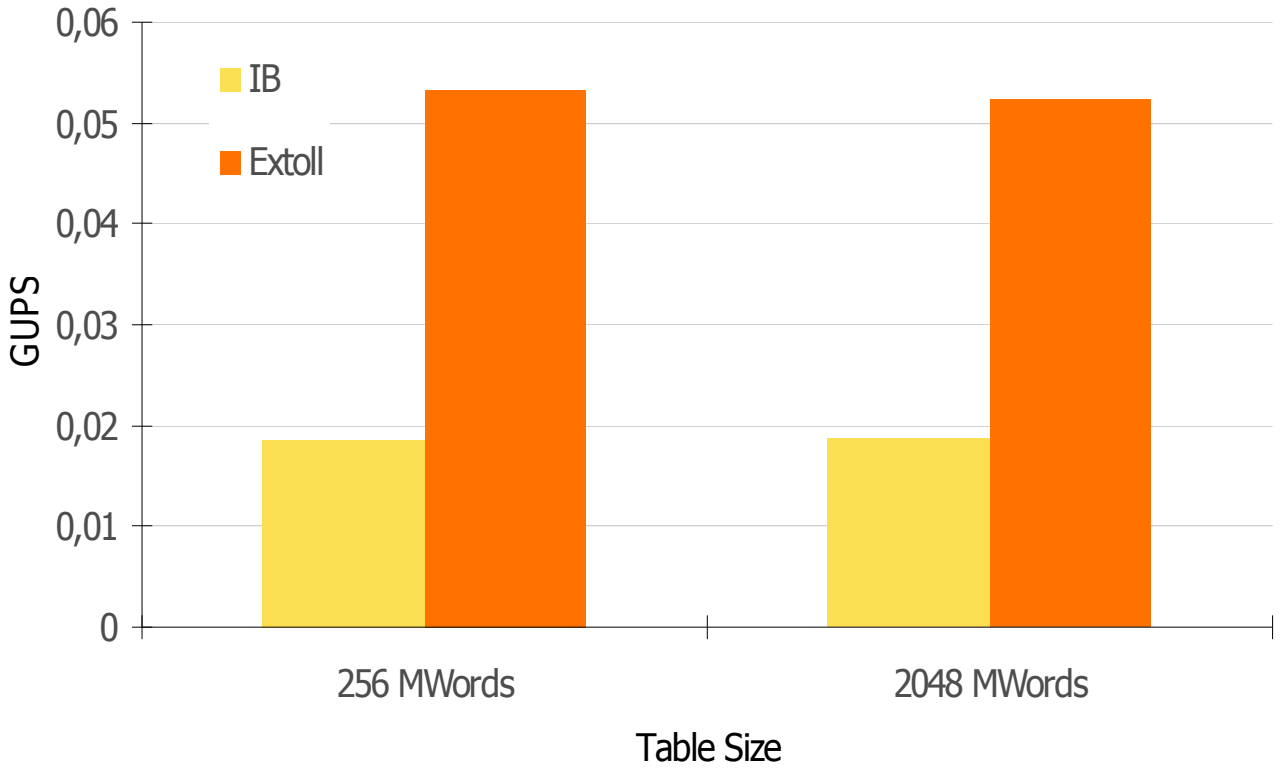
Test machines

- 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: Pingpong BW



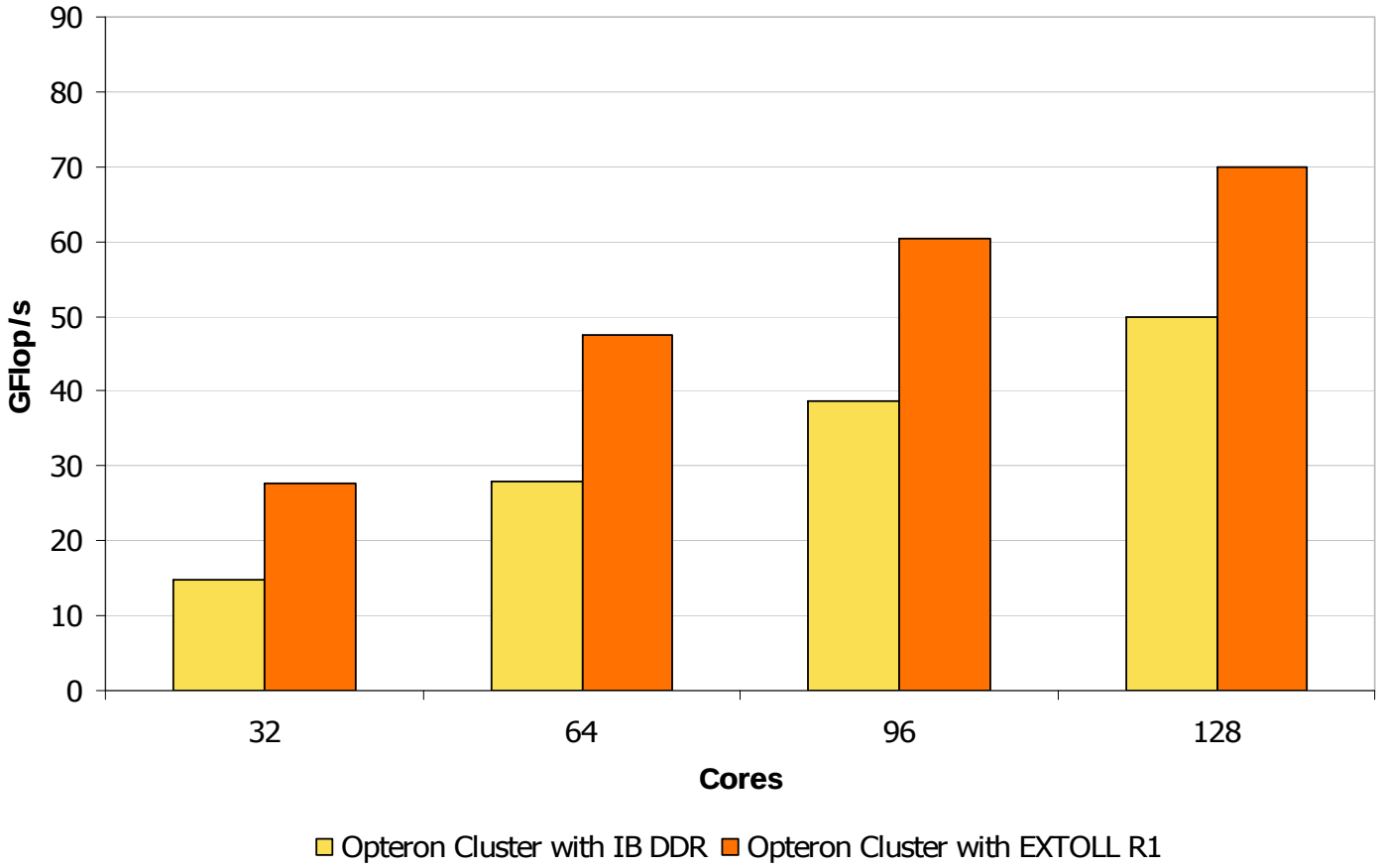
EXTOLL Virtex4: RandomAccess



GUPS = Giga updates per second

HPCC MPIRandomAccess with 32 processes on 2 nodes

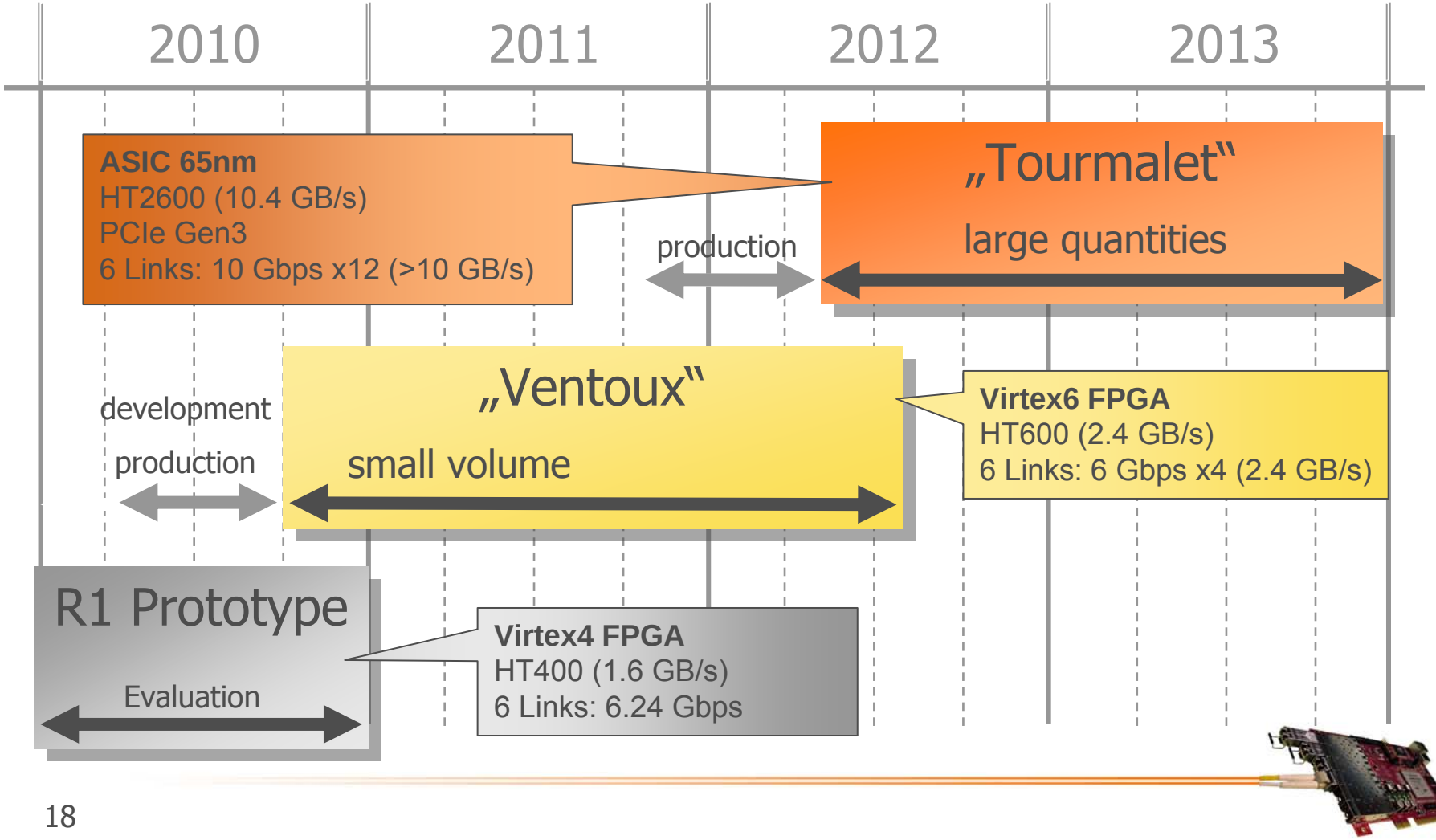
EXTOLL Virtex4: WRF V3



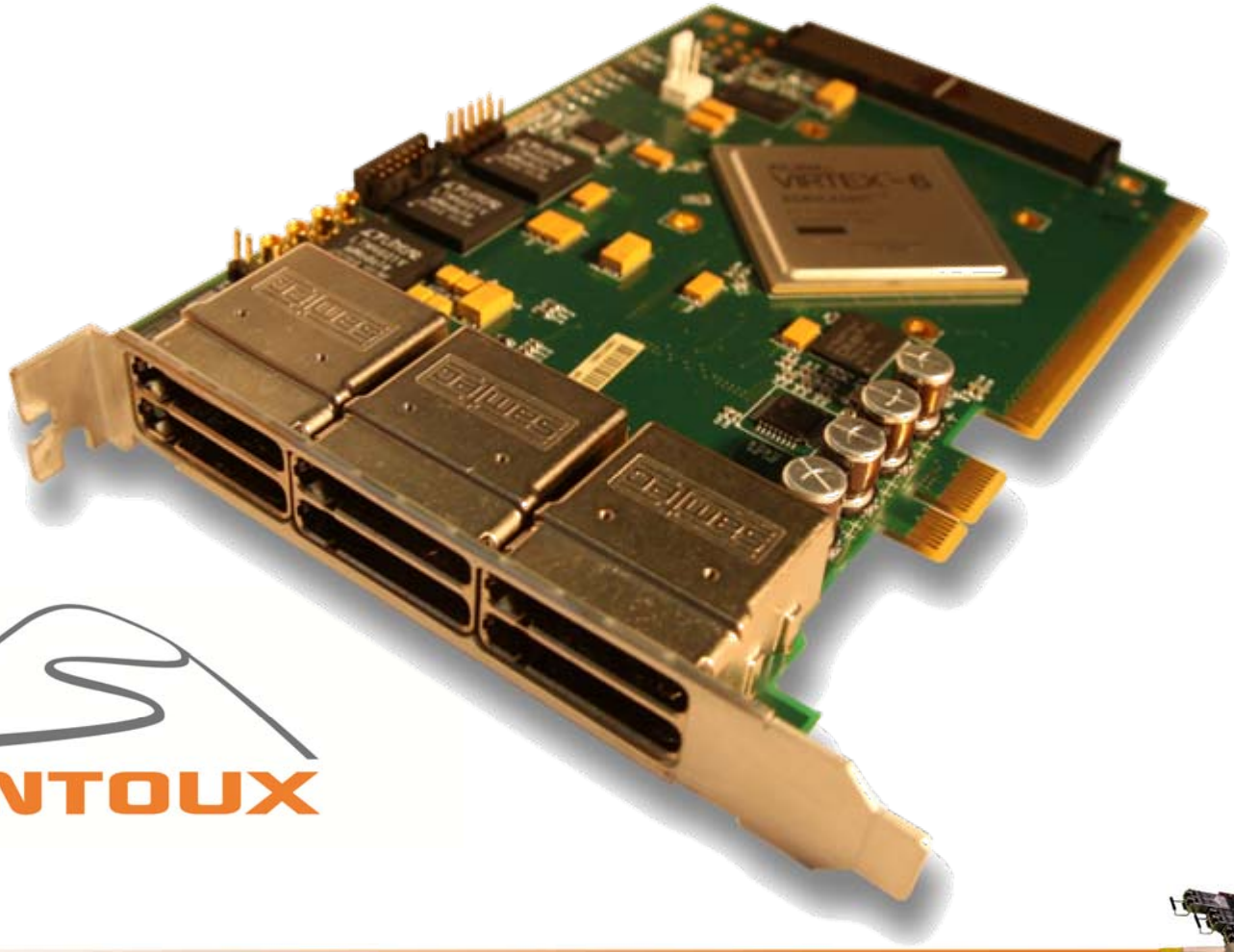
IB Cluster: Opteron Shanghai 2358 2.6GHz

EXTOLL Cluster: Opteron Shanghai 2380 2.5GHz

Hardware Roadmap



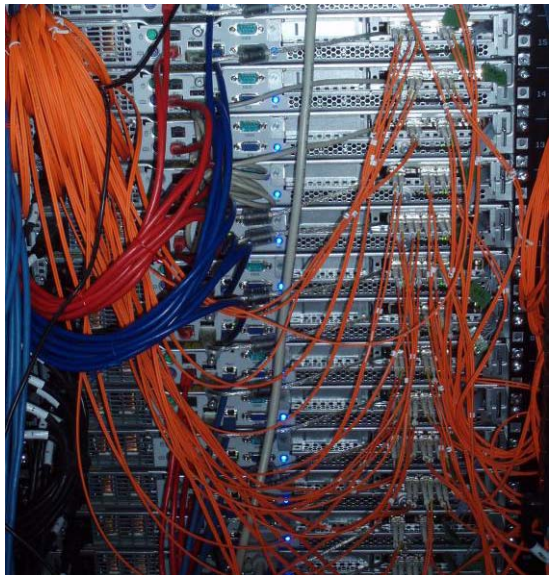
Ventoux add-in card



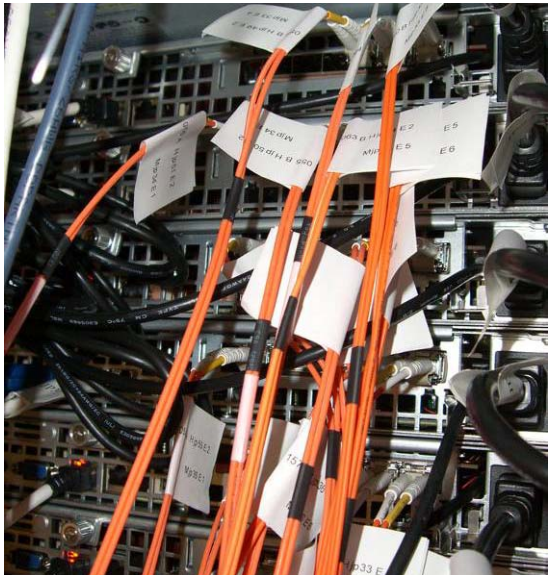
Performance

Performance Metric	EXTOLL R1 (Virtex4)	EXTOLL R2 (Virtex6)	EXTOLL R2 (ASIC)
	Measurement	Estimation	Estimation
Core Frequency	156 MHz	~300 MHz	>= 800 MHz
Internal datapath width	32 bit	64 bit	128 bit
Raw Link bandwidth	6.24 Gb/s	24 Gb/s	120 Gb/s
Half-roundtrip Latency (incl. software, single hop)	~1.2 µs	< 1.0 µs	400-600 ns
Message Rates (MPI)	~ 10 million	~ 40 million	> 100 million
Per hop latency	300 ns	<150 ns	<50 ns
Memory (de-)registration (one page)	~ 2 µs	~ 2 µs	~ 2 µs

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

eXIST

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

Computer Architecture Group,
University of Heidelberg