CPU Agnostic Motherboard design with RapidIO Interconnect in Data Center

Multi-Processor Embedded Interconnect

Switched | Scalable | Low Latency | Reliable

10 Gbps  20 Gbps  40 Gbps  100+ Gbps

WIRELESS INFRASTRUCTURE  |  SERVER  |  HPEC  |  IMAGING  |  AEROSPACE  |  INDUSTRIAL

Devashish Paul | Senior Product Manager IDT | Chairman
RapidIO Trade Association: Marketing Council

© 2013 RapidIO Trade Association
• Introductions
• An Open Approach = RapidIO Overview
• Enabling Data Centers with RapidIO Interconnect
• RTA Data Center Initiative
• Motherboard Phase 1 and Phase 2
• Scale Out (networking overview)
RapidIO’s Value Proposition

- An important enabling technology offering leading edge performance and features
- A Multi-Processor Fabric
- A Disruptive Architecture, that changes the basic economics of deployment in all multi-processor applications
- A Mature technology – 10 years of market deployments
- Designed for the unique requirements of multi processor systems
- A scalable solution for board, backplane and inter-chassis communication
- Offers lowest overall system power
- Provides superior end to end packet latency, throughput and fault tolerance
- Offers flexibility to support evolving system configurations, even in the field
RapidIO Interconnect combines the best attributes of PCIe and Ethernet in a multi-processor fabric.
RapidIO Overview

- Over 6 million RapidIO switches shipped
- Over 30 million 10-20 Gbps ports shipped > Ethernet (10GbE)
  - 100% 4G interconnect market share
  - 60% 3G, 100% China 3G market share

Today: 6.25Gbps/lane - 20 Gbps/port embedded RapidIO i/f on processors, DSPs, FPGA and ASICs.

- 10Gbps/lane in development (10xN)
- 25Gbps/lane (25xN) next generation
- No NIC needed with embedded fabric interface
- Hardware termination at PHY layer: 3 layer protocol
- Lowest Latency Interconnect ~ 100 ns
- Inherently scales to large system with 1000’s of nodes
RapidIO Ecosystem

- Axxia Communications Processor
- DSP, PowerQUICC & QorIQ multicore
- FPGA: Virtex 4/5/6 families
- FPGA: Arria and Stratix Family
- FPGA: Virtex 4/5/6 families
- FPGA: Virtex 4/5/6 families
- DSP Oct22xx
- Network Processor Octeon 2 family
- PowerPC based processors 460GT
- Network Processor WinPath3
- DSP: several products in TCI64xx family
- Multicore Processor
- Switches, Bridges & IP CPS and Tsi Family
- Wireless Baseband Processor
- ARM
- HUAWEI
- HISILICON
- AMD
- Cavium Networks
- Applied Micro
- WINTEGRA
- Intel
- MOBI VEIL
- Texas Instruments
- LSI
- FREESCALE
- ALTERA
- IDT
- LATTICE
- BROADCOM
Wireless topologies: moving to Data Center and HPC

Embedded Topologies Up to 20 Gbps
- Wireless Base Station
- Video Conferencing
- Imaging
- Mil / Aero

Computing Topologies: 20 – 100 Gbps
- Micro Server
- Super computers
- Blade server
- Storage

Wireless, Video, Military, Imaging
2x Energy Efficiency of top Green 500 HPC

Features

- Compute = 6.4 Gflops/W
- Switching = 2.2 Tbps/Chassis

“Green 500” #1: 3.2 GFlops/Watt (June 2013)

Successful Industry Collaboration

- **Open Standard:** Interoperable
• Today: 80% North-South traffic in bulk server
• Compute and Analytics = East West “any node to any node traffic”
• Desire for high bandwidth fat pipe interconnect fabric inside the box (already done in wireless)
• Reduce energy, power and latency associated with NIC’s and adaptors
• Wins: Supercomputer, Server, Storage, Network access

100 ns 1 hop switching
## Why RapidIO for Low Latency

### Bandwidth and Latency Summary

<table>
<thead>
<tr>
<th>System Requirement</th>
<th>RapidIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Switch per-port performance raw data rate</td>
<td>20 Gbps – 40 Gbps</td>
</tr>
<tr>
<td>Switch latency</td>
<td>100 ns</td>
</tr>
<tr>
<td>End to end packet termination</td>
<td>Sub 1 us</td>
</tr>
<tr>
<td>Hardware Failure Recovery</td>
<td>2 us</td>
</tr>
<tr>
<td>NIC Latency (Tsi721 PCIe2 to S-RIO)</td>
<td>300 ns</td>
</tr>
<tr>
<td>Messaging performance</td>
<td>Excellent</td>
</tr>
</tbody>
</table>
Why RapidIO for Data Center

Comparing Embedded Interconnects:
S-RIO 10xN = 2-3x Embedded Ethernet

Throughput, small form factor, low latency and dense multi processor computing required for micro server
Scaling PCIe with RapidIO

- X86 processors lead market for performance in terms of Gflops, but lack easy scale out for clusters
- Performance is key in: High Performance Computing, Server, Imaging, Wireless, Aerospace and other embedded applications
- The same applications need the performance of RapidIO Interconnect for:
  - Peer to peer networks with scalability
  - Lowest system power with protocol terminated in Hardware
  - Lowest end to end packet latency

PCIe to RapidIO NIC Attributes
13x13mm, 2W, hard terminated, 20 Gbaud per port, $49

6U Compute Node with 2x Intel I7 in production
Scaling PCIe with RapidIO: Atom based x86 Server

- Easily scale PCIe by using S-RIO switching and PCIe to S-RIO NIC devices
- Small form factor, low power NIC. 13x13 mm with 2 Watts
- Total power for 8 nodes = 23W (typ.)
- NIC latency 600 ns, switch latency 100 ns, superior to Enet and Infiniband NIC based scaling

Low latency, High Density, x86 Computer nodes
Scale PCIe with RapidIO: Two x86 Compute

Computer Node with PCIe to RapidIO NIC and switching

- 13x13 NIC, 25x25 32 Lane Switch
- 2W per NIC, 4W per Switch
- Total interconnect power 8W (typ.)
- $49/NIC, $80/Switch
- Total Interconnect per Compute Node $65

Low Energy, High Density 20 Gbps Infinite Scaling
### 48 card/96 processing nodes (Two x86 per card): Power, Scaling, Cost

<table>
<thead>
<tr>
<th></th>
<th>RapidIO (x86) (compute switch + NIC+ central)</th>
<th>10G Ethernet (central +NIC)</th>
<th>PCIe (compute node switch + HBA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Switch Only Power</td>
<td>0.33W / 10G</td>
<td>1.25W / 10G</td>
<td>1.25W / 10G</td>
</tr>
<tr>
<td>Aggregate Power 96 Nodes</td>
<td>0.384KW</td>
<td>0.424KW</td>
<td>0.834 KW</td>
</tr>
<tr>
<td>Interconnect Bandwidth per card</td>
<td>160 Gbps</td>
<td>20 Gbps</td>
<td>84 Gbps</td>
</tr>
<tr>
<td>Interconnect Gbps per Watt</td>
<td>20 Gbps/W</td>
<td>2.26 Gbps/W</td>
<td>4.84 Gbps/W</td>
</tr>
<tr>
<td>Cost per node</td>
<td>$39</td>
<td>$257</td>
<td>$122</td>
</tr>
<tr>
<td>$ per Gbps</td>
<td>$0.65/Gbps</td>
<td>$25.62/Gbps</td>
<td>$2.88/Gbps</td>
</tr>
<tr>
<td>Micro Server Any to Any</td>
<td>Yes</td>
<td>Partial (limited to 10 Gbps)</td>
<td>Limited</td>
</tr>
<tr>
<td>Scaling @ 20Gbps</td>
<td>Sub 1 us</td>
<td>&gt;10-30 us</td>
<td>&gt;5-10 us</td>
</tr>
</tbody>
</table>

### Power <25%  Interconnect 4-8x  Cost <75%
RapidIO Data Center Initiatives

- **Data Center Compute and Networking task group** = DCCN
- **Task group** being set up inside the RTA to drive collaborative development of RapidIO reference designs to target data center organizations:
  - OCP,
  - Scorpio,
  - Financial Data Center,
  - Super computer
- Reference designs for Compute/Motherboard and Networking/Switching
- Open Industry wide collaboration with several semiconductor, software and board/systems vendors participating
- **Goal:**
  - Release Phase 1 spec’s in 2H 2013
  - Phase 1 = functionality optimized CPU agnostic compute/motherboards
  - Phase 2 = performance, density, power optimized CPU agnostic compute/motherboards
CPU Agnostic RapidIO 20 Gbps Motherboards

- Compute Nodes with x86 use PCIe to S-RIO NIC
- Compute Nodes with ARM have native RapidIO endpoints
- DSP for compute intensive applications
- Up to 20 Gbps per link
- Ultra low latency
- Scales to 64k nodes
- 100 ns switch latency
Phase 1: CPU Agnostic RapidIO 20 Gbps Motherboards

PHASE 1:
- Mechanical and Electrical form factors ideally useable by OCP
- Re use existing RapidIO Processor ecosystem of AMC/Daughtercards
- Base Motherboard with connectors for AMC/daugthercards
- Compute Nodes with x86 use PCIe to S-RIO NIC on Daughtercard
- Compute Node with ARM/PPC/DSP/FPGA are native RapidIO connected with small switching option on card
- DSP for compute intensive applications
- Up to 20 Gbps per link
- Small RapidIO switch per base card
- 20-40 Gbps RapidIO links to backplane and front panel for cabling
- Local switching card/s will have similar form factor to Computer/Motherboards
Phase 2: CPU Agnostic RapidIO 20 Gbps Motherboards

- Mechanical and Electrical form factors ideally useable by OCP
- Cost, Density, and Power Optimized
- More processing capacity per motherboard
- Remove daughtercards and connectors from phase 1
- Compute Nodes with x86 use PCIe to S-RIO NIC on motherboard
- Compute Node with ARM/PPC/DSP are native RapidIO connected with small switching option on motherboard
- Up to 20 Gbps per link
- Small RapidIO switch per base card
- 20-40 Gbps RapidIO links to backplane and front panel for cabling
- Local switching card/s will have similar form factor to Computer/Motherboards
- Interoperable with Phase 1 cards
• Switching at board, chassis, rack and top of rack level
• Scalability to 64K nodes, roadmap to 4 billion
• Application focus
  – Latency sensitive analytics
  – Financial
  – High-Performance Computing (HPC)
• Proven in Telecom, Medical, Industrial
Small form factor RapidIO Embedded Boards with Intel and TI Processors

- Intel I7 with RapidIO to PCIe Bridging
  - 181 x 74 mm
  - 1x Intel Core 7 quad core processor
  - 40 W typical
  - 16-32 Gbps backplane connectivity with Tsi721
  - PCIe to RapidIO NIC only 2W, 300 ns latency

- AMC .4 with Native RapidIO
  - 181 x 74 mm
  - 2x TI C6678 Processors
  - 640 GMAC + 320 GFLOPS (2 CPUs each with 8 cores running fixed and floating point
  - 40W typical
  - 40 Gbps backplane connectivity
  - No NIC zero processor latency

Small, low power
Low latency

Small, higher compute
Lowest latency
Small form factor RapidIO Embedded Boards
Freescale and TI Processors

- ARM + DSP AMC with Native RapidIO
  - 1x TI 66AK2H12: 4 ARM Cortex-A15 cores and 8 C66x DSP cores
  - 2x TI C6678: 16 C66x DSP cores
  - 518 GLOPS at < 50W typical
  - Up to 26GB DDR3 + ECC
  - 3x RapidIO quad lane to AMC

- P4080 and P5020 AMC
  - Footprint compatible octal e500 or dual 64-bit e5500 Power Architecture cores
  - 40W typical
  - Up to 16GB DDR3 memory
  - 2x RapidIO quad lane to AMC
New Spec: RTA S-RIO 10xN Overview

3rd Generation
Scalable embedded
peer to peer
Multi processor
Interconnect
On board, board-to-board and
Chassis to Chassis

- S-RIO 10xN: data rate of 40-160 Gbps per port
- 10.3125 Gbaud per serial lane with option of going to 12.5 Gbaud in future
- Long-reach support (100 cm through two connectors), Short Reach 20 cm 1 connector, 30 cm no connector
- Backward compatibility with RapidIO Gen2 switches (5 & 6.25 Gbps) and endpoints
- Lane widths of x1, x2, x4, x8, x16
- Speed granularity from 1.25, 2.5, 3.125, 5, 6.25, 10.3125 Gbaud

Key Additional Features
- 10 Gbps per lane
- 10 to 160 Gbps per port
- 64/67 encoding
- Power management
RapidIO 10xN to 25xN Highlights

- Builds on deployment of over 30 M 10-20 Gbps RapidIO Ports
- > 6M switches shipped
- exceeds 10 GbE port shipments in 2012

10Gbps/lane switches in development
25Gbps/lane next gen backplane switches

- PCIe3 NIC to RapidIO 10xN
- RapidIO Endpoint IP (10xN) for ASICs

- Large Ecosystem of software support
  - Linux, Windows, VxWorks
  - Boards and Debug Software
RapidIO Data Center Summary

- 30 million RapidIO Ports @ 10-20 Gbps deployed
- Ideal for power-latency-performance optimized applications
- Highest Gflops per watts with Green 500 ready product launch (6.4 Gflops per watt)
- **Task group** set up inside the RTA to drive collaborative development of RapidIO reference designs to target data center organizations:
  - OCP,
  - Scorpio,
  - Financial Data Center,
  - Super computer
- Reference designs for Compute/Motherboard and Networking/Switching
- Open Industry wide collaboration with several semiconductor, software and board/systems vendors participating
• RIONET transports Ethernet frames over RapidIO
  – Kernel mode interface
  – Low performance

• DMA support for Tsi721 PCIe to RapidIO bridge
  – Part of Linux DMA subsystem
• IDT is working with TI, nCore and Prodrive on an OpenMPI implementation for RapidIO, to be used in Brown Dwarf.

• The OpenMPI BTL uses a RapidIO “User Mode Interface” for low latency.
nCore Lithium Suite

- nCore Lithium Suite is the fastest way to performance and productivity on Keystone II
  - Advanced support for compute offload using OpenMP with accelerator model
  - Optimized IPP replacement library for C66x, Advanced DMA library for C66x
  - Performance Optimization Tool Layers (PAPI for A15 and C66x DSP)
  - OS distribution with Ubuntu’s update system for access to desktop and server repositories
  - Commercial Support
  - nCore is the worldwide leader in TI Keystone software technologies

- More information: [http://www.ncorehpc.com](http://www.ncorehpc.com) or [info@ncorehpc.com](mailto:info@ncorehpc.com)

Currently Supported Platforms:
- nCore BrownDwarf YCNODE
- TI’s XTCIEVMK2X EVM
- CSPI’s KeyDSP PCIe Cards
- Others to follow
RapidIO User Mode API for OpenMPI

User application

MPI API

PML
COLL
OSC
BTL
MTL

Support Interfaces

OB1
CM
TUNED
FC
RDMA
PORTAL54
TCP
SM
MXM
PSM
PORTAL54

RapidIO User Mode API
High Performance Applications

• High performance Applications require a “User Mode Interface”

• Applications are tied to the User Mode Interface

• IDT is working toward a User Mode Interface which supports TI, Freescale, and Tsi721
RIONET vs User Mode API

- Application is completely portable
- Fault tolerant
- Low performance, high latency due to application/kernel interactions
  - Memory copies
  - Interrupt latency
- Application written specifically for User Mode API
  - Fault tolerant
  - High performance, low latency
  - Messages sent directly to hardware
  - Messages received directly from hardware
**RapidIO vs Ethernet Latency Path**

**Long Latency Path**
- Memory
- Processor
- Ethernet NIC
- Ethernet Switch
- PCI Express (I/O Interconnect)
- Ethernet LAN Interconnect With NIC

**Short Latency Path**
- Memory
- Processor
- Ethernet NIC
- Ethernet Switch

**ETHERNET**
- Long Latency
- NIC Required High Footprint
- Higher Power
- Software Intervention

**RAPIDIO**
- Short Latency
- NO NIC embedded Fabric
- Low Power
- Native Hardware Termination
Compare to PCIe and Ethernet 2 Node Compute

Micro Server Building Block with PCIe only

Micro Server Building Block with Ethernet only

Scale out limitation with PCIe, Bandwidth and latency Limits with Ethernet