Supercomputing Clusters with RapidIO Interconnect Fabric

Devashish Paul, Director Strategic Marketing, Systems Solutions
devashish.paul@idt.com

Ethernet Summit 2015
April 14-16, 2015 – Santa Clara, CA
Agenda

- RapidIO Interconnect Technology compliments Ethernet
- RapidIO Attributes
- Heterogeneous Accelerators
- Open Compute HPC
- Summary and Wrap Up

Low Latency | Reliable | Scalable | Fault-tolerant | Energy Efficient
Supercomputing Needs Heterogeneous Accelerators

- Chip to Chip
- Board to Board across backplanes
- Chassis to Chassis
- Over Cable
- Top of Rack
- Heterogeneous Computing

Rack Scale Fabric
For any to any compute
### HPC/Supercomputing Interconnect ‘Check In’

<table>
<thead>
<tr>
<th>Interconnect Requirements</th>
<th>RapidIO</th>
<th>Infiniband</th>
<th>Ethernet</th>
<th>PCIe</th>
<th>The Meaning of</th>
</tr>
</thead>
</table>
| **Low Latency**           | ![Circle](#) | ![Circle](#) | ![Cross](#) | ![Circle](#) | • Switch silicon: ~100 nsec  
• Memory to memory: < 1 usec |
| **Scalability**            | ![Circle](#) | ![Circle](#) | ![Circle](#) | ![Cross](#) | Ecosystem supports any topology, 1000’s of nodes |
| **Integrated HW Termination** | ![Circle](#) | ![Cross](#) | ![Cross](#) | ![Circle](#) | Available integrated into SoCs AND Implement guaranteed, in order delivery without software |
| **Power Efficient**        | ![Circle](#) | ![Cross](#) | ![Cross](#) | ![Circle](#) | 3 layers terminated in hardware, Integrated into SoC’s |
| **Fault Tolerant**         | ![Circle](#) | ![Circle](#) | ![Circle](#) | ![Cross](#) | Ecosystem supports hot swap  
Ecosystem supports fault tolerance |
| **Deterministic**          | ![Circle](#) | ![Circle](#) | ![Cross](#) | ![Circle](#) | Guaranteed, in order delivery  
Deterministic flow control |
| **Top Line Bandwidth**     | ![Cross](#) | ![Circle](#) | ![Circle](#) | ![Cross](#) | Ecosystem supports > 8 Gbps/lane |

- Lowest Deterministic System Latency
- Scalability
- Peer to Peer / Any Topology
- Embedded Endpoints
- Energy Efficiency
- Cost per performance
- HW Reliability and Determinism

Clustering Fabric Needs

- Scalability
- Low Latency
- Hardware Terminated
- Guaranteed Delivery

Ethernet
- Lowest Deterministic System Latency
- Scalability
- Peer to Peer / Any Topology
- Embedded Endpoints
- Energy Efficiency
- Cost per performance
- HW Reliability and Determinism

PCle
- Lowest Deterministic System Latency
- Scalability
- Peer to Peer / Any Topology
- Embedded Endpoints
- Energy Efficiency
- Cost per performance
- HW Reliability and Determinism

Ethernet in s/w only
RapidIO on Ethernet Electricals

- Ethernet for external networking
- RapidIO for internal fabric on ethernet electricals

Ethernet

RapidIO Interconnect
Open Compute Project HPC and Supercomputing

Project Charter Items

Fully open **heterogeneous computing**, networking and fabric platform

Optimized for multi-node **processor agnostic** any to any computing using x86, ARM, PowerPC, FPGA, ASICs, DSP, and GPU silicon on hardware platform

Enables rapid innovation in low latency high Performance Computing and Big Data analytics through open non-lock-in computing, interconnect, and software stack.

**Energy** efficient compute density

Distributed and central **storage** for large data manipulation (non spinning disk) with low latency

Operating System – Linux based operating systems and developer tools and **open APIs**

Path to **Open Silicon** and Open APIs, initially leveraging existing industry standards, later developing its own silicon

**Reuse** developments from OCP Server group and Open Rack where appropriate

Leverage **industry standard interconnects**, no proprietary interconnects for main fabric and networking

---

Devashish Paul, co lead HPC Project

OCP HPC Project San Jose Summit – Mar 2015

---

**Vendor Agnostic Low Latency Multi Processor Computing**
• 20 Gbps per port / 6.25Gbps/lane in productions
• 40Gbps per port / 10 Gbps lane in development
  Embedded RapidIO NIC on processors, DSPs, FPGA and ASICs.
• Hardware termination at PHY layer: 3 layer protocol
• Lowest Latency Interconnect ~ 100 ns
• Inherently scales to large system with 1000’s of nodes

• Over 13 million RapidIO switches shipped
• > 2xEthernet (10GbE)
  Over 70 million 10-20 Gbps ports shipped
• 100% 4G interconnect market share
• 60% 3G, 100% China 3G market share
Peer to Peer & Independent Memory System

- Routing is easy: Target ID based
- Every endpoint has a separate memory system
- All layers terminated in hardware

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prev Packet</td>
<td>1 3</td>
</tr>
<tr>
<td>Ack D</td>
<td>1</td>
</tr>
<tr>
<td>Rsvr</td>
<td>2</td>
</tr>
<tr>
<td>Prio</td>
<td>2</td>
</tr>
<tr>
<td>TT</td>
<td>2</td>
</tr>
<tr>
<td>Target Address</td>
<td>8 or 16</td>
</tr>
<tr>
<td>Source Address</td>
<td>8 or 16</td>
</tr>
<tr>
<td>Transaction</td>
<td>4</td>
</tr>
<tr>
<td>Ftype</td>
<td>4</td>
</tr>
<tr>
<td>Size</td>
<td>4</td>
</tr>
<tr>
<td>Source TID</td>
<td>8</td>
</tr>
<tr>
<td>Device Offset Address</td>
<td>32 or 48 or 64</td>
</tr>
<tr>
<td>8 to 256 Bytes</td>
<td></td>
</tr>
<tr>
<td>Optional Data Payload</td>
<td>16</td>
</tr>
<tr>
<td>CRC</td>
<td></td>
</tr>
<tr>
<td>Next Packet</td>
<td></td>
</tr>
</tbody>
</table>

Why RapidIO for Low Latency

<table>
<thead>
<tr>
<th>Bandwidth and Latency Summary</th>
<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>
<td></td>
</tr>
<tr>
<td>Switch latency</td>
<td>100 ns</td>
<td></td>
</tr>
<tr>
<td>End to end packet termination</td>
<td>~1-2 us</td>
<td></td>
</tr>
<tr>
<td>Fault Recovery</td>
<td>2 us</td>
<td></td>
</tr>
<tr>
<td>NIC Latency (Tsi721 PCIe2 to S-RIO)</td>
<td>300 ns</td>
<td></td>
</tr>
<tr>
<td>Messaging performance</td>
<td>Excellent</td>
<td></td>
</tr>
</tbody>
</table>
RapidIO.org ARM64 bit Scale Out Group

- 10s to 100s cores & Sockets
- ARM AMBA® protocol mapping to RapidIO protocols
  - AMBA 4 AXI4/ACE mapping to RapidIO protocols
  - AMBA 5 CHI mapping to RapidIO protocols
- Migration path from AXI4/ACE to CHI and future ARM protocols
- Supports Heterogeneous computing
- Support Wireless/HPC/Data Center Applications

Source – [www.rapidio.org](http://www.rapidio.org), Linley Processor conf
NASA Space Interconnect Standard

Next Generation Spacecraft Interconnect Standard

Key Driving Differentiators

- Serial RapidIO has the following salient features among four protocols:
  - Transparent compatibility with wired and fiber-optic
  - Applicable to chip-to-chip, board-to-board, and box-to-box
  - Light-weight and modular (features are configurable)
  - Low power with less than 192 mW per node
  - Scalable fault tolerance with link-level error detection
  - Scalable bandwidth up to 3.125 Gbps per lane
  - Real-time with sub-microsecond latency and jitter
  - Switch-based flexible topology
  - Built-in shared-memory support with low S/W overhead
  - Embedded provisions allow backward-compatible protocol extension

RapidIO selected from Infiniband / Ethernet /FiberChannel / PCIe

NGSIS members: BAE, Honeywell, Boeing, Lockheed-Martin, Sandia Cisco, Northrup-Grumman, Loral, LGS, Orbital Sciences, JPL, Raytheon, AFRL
RapidIO in Open Compute Interconnect

Supports Opencompute.org HPC Initiative
Computing: Open Compute Project
HPC and RapidIO

- **Low latency analytics** becoming more important not just in HPC and Supercomputing but in other computing applications
- mandate to create open latency sensitive, energy efficient board and silicon level solutions
- **Low Latency Interconnect RapidIO submission**

```
RapidIO = 2x 10 GigE Port Shipment
```

```
Input Reqts
```

```
RapidIO Interconnect Development
```

```
Laser Connect 2.0
```

```
4x 25 Gbps Multi Vendor Collaboration
```

Latency
Scalability
Hardware Termination
Energy Efficiency
RapidIO at CERN LHC and Data Center

• RapidIO Low latency interconnect fabric
• Heterogeneous computing
• Large scalable multi processor systems
• Desire to leverage multi core x86, ARM, GPU, FPGA, DSP with uniform fabric
• Desire programmable upgrades during operations before shut downs
Heterogeneous HPC with RapidIO based Accelerators
RapidIO Heterogenous switch + server

- 4x 20 Gbps RapidIO external ports
- 4x 10 GigE external Ports
- 4 processing mezzanine cards
- In chassis 320 Gbps of switching with 3 ports to each processing mezzanine
- Compute Nodes with x86 use PCIe to S-RIO NIC
- Compute Node with ARM/PPC/DSP/FPGA are native RapidIO connected with small switching option on card
- 20Gbps RapidIO links to backplane and front panel for cabling
- Co located Storage over SATA
- 10 GbE added for ease of migration
X86 + GPU Analytics Server + Switch

- Energy Efficient Nvidia K1 based Mobile GPU cluster acceleration
  - 300 Gb/s RapidIO Switching
  - 19 Inch Server
  - 8 – 12 nodes per 1U
  - 12 - 18 Teraflops per 1U

DCCN PCB fitted in 19” rack-mount enclosure & OCP 21” shelf
38x 20 Gbps Low Latency ToR Switching

- Switching at board, chassis, rack and top of rack level with Scalability to 64K nodes, roadmap to 4 billion
- 760Gbps full-duplex bandwidth with 100ns - 300ns typical latency
- 32x (QSFP+ front) RapidIO 20Gbps full-duplex ports downlink
- 6x (CXP front) RapidIO 20Gbps ports downlink
- 2x (RJ .5 front) management I²C
- 1x (RJ .5 front) 10/100/1000 BASE-T
Analytics Acceleration with GPU Clusters

- Announced at SC14 New Orleans
- Can be used in Data Center or edge of wireless network for analytics acceleration
- Proto cluster based on Jetson K1 PCIe boards
- NVIDIA K1 + RapidIO cluster using Tsi721 PCIe to S-RIO

Low Energy Mobile GPU + RapidIO for Analytics
16 Gbps per Node, 24 flops per bit I/O “Flux”
RapidIO with Mobile GPU Compute Node

- 4 x Tegra K1 GPU
- RapidIO network 140 Gbps embedded RapidIO switching
- 4x PCIe2 to RapidIO NIC silicon
- 384 Gigaflops per GPU
- >1.5 Tflops per AMC
- 12 Teraflops per 1U
- 0.5 Petaflops per rack
Social Media Real Time Analytics – FIFA World Cup 2014

Analyze User Impressions on World Cup 2014

Future Analytics acceleration using GPU
RapidIO 10xN Development and Futures

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

- 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, 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, 2, 4, 5, 10, 20, 40 Gbps

• Released: 40 Gbps IP core
  • Development: Switches
• Definition: Bridging, Embedded 40 Gbps NIC and NIC
  • Future: ARM64 bit Cache Coherency IP
Summary Low Latency Data Center Acceleration with RapidIO + GPU

- 100% 4G market share,
- all 4G calls worldwide go through RapidIO switches
- 2x market size of 10 GbE (70 million ports S-RIO)
- 20 Gbps per port in production
- 40 Gbps per port silicon in development now
- Low 100 ns latency, scalable, energy efficient interconnect
- Supports dense GPU based compute up to 600 Gflops per rack, based on 386 Gflops per GPU single precision
- 12-18 Teraflops per 1 U server
- Ideal for analytics, deep learning and pattern recognition

Low Latency | Reliable | Scalable | Fault-tolerant | Energy Efficient
Backup: RapidIO Protocol Background
RapidIO with x86 and/or and GPU

- RapidIO based x86 + GPU heterogenous computing
  - X86 moves data
  - GPU provides compute/analytics
  - RapidIO NIC silicon 13x13, 2W
- Best-in-class end-to-end latency < 1.2 microseconds any to any
- Supports any kind of topology
- Ideal for analytics, deep learning and pattern recognitions

Lowest latency scalable acceleration
RapidIO Interoperable Eco-system

Open Ecosystem – No Vendor Lock-in

Source - http://www.rapidio.org/files/RapidIO_Asia_Summit_Intro.pdf

Low Latency | Reliable | Scalable | Fault-tolerant | Energy Efficient
Requirements: Low Latency

- RapidIO networks are built around two “Basic Blocks”
  - Endpoints
  - Switches
- End points source and sink packets
- Switches pass packets between ports without interpreting them
- Simple routing through Look Up Tables
- Typical Switch Latency is 100-150ns
Messaging

- Messaging uses a push architecture
- Receiver is responsible for storing the message in its memory system
- Overall system latency per message transfer is reduced significantly
RDMA = Remote Direct Memory Access

One process writes directly into another process.

The fabric operates as a bus extension.

RDMA is the most efficient, lowest latency, interprocess communication (IPC) mechanism.

Hardware:
- Messaging: RDMA connection management
- Read/Write/Atomic: RDMA transfers
- Doorbells: Events
- All terminated in hardware

Software:
- The OS connects processes for RDMA
- The OS is bypassed for RDMA transfers

Low Latency * Scalable * Integrated HW Termination * Power Efficient * Fault Tolerant
Low Latency Open Fabric for Heterogeneous Computing

OCP HPC Fabric Interconnect Silicon

- HPC needs huge scale of any to any processing nodes
- Latency is a primary concern for this market as well as those that need analytics
- Energy footprint is an issue
- Low hanging fruit is to eliminate latency and power from NIC’s and other interconnect devices
- Need native protocol termination on processing endpoints, like processors, DSP, GPU, FPGA
- Diverse industry initiatives to create proprietary clustering fabrics at many startups and large processor vendors
- Prefer to start with some industry standard options that scale, have low latency, multi vendor collaboration etc
- Take best attributes of PCIe, Infiniband, RapidIO, Ethernet technologies to reach exascale computing

No proprietary vendor lock in, open standard, interoperable silicon

OCP HPC Project – Sep 2014