PCI Express switch over Ethernet or Distributed IO Systems for Ubiquitous Computing and IoT Solutions

02, March, 2017
Deepak Pathania, NEC Corporation
"There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days“, Google, 2010
Challenge faced in Real-Time Data Analytics

**Accelerators/AI Engines**
- Xeon Phi
- GPU
- FPGA

**Situation**
Big Data of varying characteristics, such as Live feeds, graphics, video, text, etc., comes into cloud computers.

**Demand**
This data is to be processed and analyzed in real-time.

**Valid Solution**
To accelerate such processing, a large number of accelerators such as GPUs and FPGAs, along with high speed storage are required.

**Issue**
However, instead of building servers with such accelerators, Cloud vendors still prefer building homogeneous servers due to TCO and efficiency considerations.

**NEC’s Solution**
ExpEther technology that allows building dynamic accelerator deployment system.
What is ExpEther?

A technology that can extend PCI Express bus beyond the confines of a computer chassis via Ethernet without any modification of existing hardware and software.
ExpEther can build new type of computing environment without physical constraints.
ExpEther Engine is seen as PCIe Switch from CPU

- Ethernet region is invisible from the CPU

ExpEther is one example of implementation of PCIe Switch
ExpEther Architecture

- Achieve the “System on Network”
  - Merge the PCI Express technology into Ethernet technology

- Connect logically in MAC layer
  - No impact for upper or lower layer of the PCIe and Ethernet standard for future expansion

### Application
- OS
- PCI Driver
- EFI/PCI BIOS

### Ethernet
- Application
- OS
- NDIS Driver

---

**ExpEther Logic**
- MAC
- PHY
  - 40G
  - 10G
  - 1G

**Ethernet Logic**
- MAC
- PHY
  - 10M
  - 100M
  - 1G
  - 10G
  - 40G

---

**Software**
- Upper Compatible
- No modification for future expansion of ExpEther or Ethernet

---

© NEC Corporation 2017
Features of ExpEther

ExpEther Engine is compliant with PCIe and Ethernet Standard

- PCI-SIG PCI Express Certified
- Can use off-the-shelf L2 Ethernet Switch

1. Equivalent to direct connection (Ethernet is invisible from CPU/I/O)
2. Low Latency (L2 Ether w/o SW stack)
3. No packet loss (Adding reliability to Ethernet)
4. I/O Dynamic Reconfiguration (Hot-Plug Scheme)
It is difficult to extend the PCIe to long reach by cable because of DLLP timeout rule

- DLLP timeout is less than 200 usec (depending on chipset)
- TLP timeout is 50 msec, but can be extended to 64 seconds by configuration

ExpEther is unaffected by DLLP timeout

- It is possible to extend the IO devices to long reach
Two Ethernet connections are established between the Host Chip and I/O Chip

- Load balancing for performance
- Path redundancy for failure recovery

**Load-balancing**
Round-robin data packet transmission between the two redundant connections

**Failure Recovery**
Quickly detects path failures and switches paths

10G ExpEther NIC

Dual Port

10G ExpEther NIC
TCP/IP: Rate control is triggered by packet loss (TCP Reno)

Packet loss causes significant performance degradation because of retransmission.

ExpEther: Rate control is always done by measuring network latency

Packet loss does not occur basically in ExpEther.

ExpEther engine always measures the frame arrival time of receive side and minutely controls the frame rate to avoid packet loss.
Loss-less ExpEther Frame

- Ethernet may lose packets, but PCIe does not allow losing any TLP.
- ExpEther ensures that the packets certainly arrive at end by Ack/Nack scheme in Ethernet.

Ether network

**Retransmission**

- ExpEther packet
- ACK packet

Sender

1. Timer reset
2. Timer reset
3. Timer reset
4. Buffer Release
   - Seq1~Seq 5
5. Timer reset

Receiver

1. ACK Timer set
2. ACK Timer set
3. ACK Timer expire
4. ACK Timer expire

Timer expire
But there is no frame in buffer.
Re-transmission is not started.
ExpEther Reliability ~ Multi-Path

<table>
<thead>
<tr>
<th>Multi-Path IO (MPIO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• MPIO is one of the technic for achieving high-reliability. If the target IO device supports MPIO, it can support MPIO even under ExpEther.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Multi-Path Ethernet</th>
</tr>
</thead>
<tbody>
<tr>
<td>• ExpEther supports the high-speed network path failover.</td>
</tr>
</tbody>
</table>
Sequence of Network Path Failover (1/2)

Both network paths are used as ACT-ACT

If a path is failed, ExpEther resends lost packets. This failover time is about 10 RTT (several microseconds).

ExpEther Packet

EE NIC (Tx side)

Retransfer Buffer

Arbiter

16

18

17

Lost Packet

Ordering Buffer

Sequence Number Check

EE (Rx side)

Rcv. Buf.

4

3

2

1

Resending

13

EE NIC (Tx side)

Retransfer Buffer

Arbiter

17

16

15

14

13

Lost Packet

Re-send packets after several microseconds

EE (Rx side)

Rcv. Buf.

Rcv. Buf.

Ordering Buffer

Orchestrating a brighter world

© NEC Corporation 2017
Network path is recovered by some Ethernet recovering scheme like P-Flow linked with EE manager.

When ExpEther device receives a management packet indicating the path recovered, it starts reusing both network paths.
Each ExpEther device has a Grouping ID to connect a Host and IO devices logically
- The ID is assigned by rotary switch or Manager software
- The ID can be set from 1 to 4,095 and it is used as VLAN tag
ExpEther Management Scheme Overview

Group ID (GID: 1~4,095)

- GID range from 1 to 15 is set by physical DIP switch residing on card.
- Setting GID to 0 allows Management Software to program a soft GID.

- Mng. Frame -
  - Special Ether Frame
    - ExpEther hard wired logic directly receives and sends the frames for configuration and management

- ExpEther Manager -
  - Configuration
    - Group ID Configuration
  - Monitoring
    - ExpEther network status
    - PCIe device status
    - New ExpEther detection
    - Failure detection
ExpEther Manager Library and SDK

### Library Structure

- **REST API**
  - CLI
  - Web Interface
  - OpenStack Combination
- **Java API**
  - Java GUI
- **C/C++ API**
  - Customer Original Application

---

**EEM Library / SDK**

- REST API
  - Java Servlet
  - Java API
    - ExpEther Manager Java Module
    - C/C++ API
      - ExpEther Manager C/C++ Library

---

**CLI**

- > EEM list
  - IO#0 Intel
  - IO#1 Broadcom
  - IO#2 Mellanox

**Web Browser**

- Java GUI App.

---

**Original Application**

- Host List
  - + Host 1
  - + Host 2
  - + Host 3

**IO List**

- - IO 1
  - + IO 2
  - + IO 3
  - - IO 4
  - - IO 5
  - - IO 6
  - - IO 7
  - - IO 8

---

* : Editable.

Input after pressing Enter.
### KEY Register/Delete

**TLP Encryption (TWINE)**
- 40G ExpEther supports TLP encryption
- The encryption key is configured by Management Software

**TWINE is developed by NEC**
- High-speed and quite small hardware implementation
ExpEther Technology Architectural Possibilities

- **Std-EE**: Standard PCIe-over-Ethernet
  - Foundation of ExpEther
- **MR-EE**: I/O sharing
  - Multi-hosts are able to share an IO device by using SR-IOV compliant device
- **P2P-EE**: I/O direct connection
  - Support for the Peer-to-Peer data transfer between I/O devices.
- **NTB-EE**: Remote direct memory access by NTB
  - Hi-speed data transfer between hosts
ExpEther Advantages

Dynamic Resource Reconfiguration & Sharing

Since ExpEther supports concept of GID/VLAN, IO resources can be dynamic allocated to different hosts based on need to application/workload.

No Change in OS or Driver

NVMeoE specification asks for changing drivers and OS, whereas no such change in required in ExpEther. NVMe are accessible with simple plug-and-play using ExpEther.

No space/length constraint

The length of the Ethernet fabric can be few meters to several kilometers with ExpEther. So servers can be somewhere else while IOs anywhere, which is especially useful for IoT

Reduced Costs

When expanding the systems by adding tens of hundreds of IO devices, no need to purchase expensive PCIe switches, ExpEther works on standard off-the-shelf Ethernet switches.
Service Acceleration Platform with ExpEther

**Compute Node**
- CPU/Chipset
- ExpEther HBA

**Accelerator Node**
- Ether Switch
- ExpEther Engines
- GPGPU
- Accelerator FPGA

**Hi-speed Storage Node**
- Ether Switch
- ExpEther Engines
- NVMe SSD

**Remote IO Devices**
- USB/VGA
- ExpEther Engine
- KVM

**Remote IO**
- USB Ctrl
- ExpEther Engine

**Sensors**

IO devices can be dynamically allocated to appropriate host according to workload.
Case: Resource Pool System for HPC (Osaka University)

- 64 servers and 70 IO devices for research in Osaka University
- There are GPUs, Flash storages and VDI accelerators as IO device
- The IO devices are dynamically connected to the servers through 10G ExpEther in accordance with server’s workload
Case: Easy Extension of Measurement Equipment (PXI)

Current PXI products are typically extended by PCIe cable. So the measurement system is fixed and the installation location is very limited.

If ExpEther engine is implemented into PXI chassis, the system can have a large number of PXI modules and dynamically configure the system.

PXI (PCI eXtensions for Instrumentation) is one of several modular electronic instrumentation platforms based on PCIe.
Living at the Edge for going Real-Time with ExpEther

**L5 Cloud ~ Analytics**
Rack-Scale or Resource pooling with dynamic reconfiguration allows low-cost, low-power and high performance computing data centers at the cloud level.

**L3 Edge ~ Abstraction/Real-Time Proc.**
ExpEther helps in bringing analytics to the edge. In combination with low-power and high-performance hardware like FPGA’s one can achieve an idealistic abstraction required for Real-time processing.

**L1 Device/Sensor ~ Smart Device**
ExpEther can connect devices directly to the edge and servers using simple everything in hardware approach or no complex software protocol stack for communication which is high-speed and low power. Making devices smarter.
ExpEther as a back-plane interconnect for Ubiquitous Computing and IoT solutions for Real-Time Analytics
You can see more detailed technical and product information in ExpEther web site.
Orchestrating a brighter world