A Resource disaggregated Platform for realizing Ultra-Fast Failover Recovery High Availability Systems

Deepak Pathania, Senior Technical Leader
NEC Corporation and NEC Technologies India Private Limited
So, What can be a Resource Disaggregated Platform?

A technology that extends PCI Express beyond the confines of a computer chassis via Ethernet, WITHOUT any modification of existing hardware and software or PCIe switch over Ethernet (ExpEther or EE)
ExpEther Engine is seen as PCIe Switch from CPU
- Ethernet region is invisible from the CPU
ExpEther can build new type of computing environment without physical constraints.
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
Resource Disaggregated Platform or ExpEther features

1. Equivalent to direct connection (Ethernet is invisible from CPU/IO)
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)
Dual Path for Throughput and Reliability

• Two Ethernet connections are established between the Host Chip and I/O Chip
  • Load balancing for performance
  • Path redundancy for failure recovery
Frame Rate Control

TCP/IP : Rate control is triggered by packet loss (TCP Reno)

Packet loss causes significant performance degradation because of retransmission.

Slow Start
Avoid Congestion
Avoid Congestion
Avoid Congestion

Network Bandwidth

Congestion (Packet Loss)

Time

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.

Network Bandwidth

Avoid Congestion

Time
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).
• 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.
• Multi-Path IO (MPIO)
  • MPIO is one of the technic for achieving high-reliability. If the target IO device supports MPIO, it can support MPIO even under ExpEther.

• Multi-Path Ethernet
  • It supports the high-speed network path failover.
Dynamic Reconfiguration and Hot-Plug Capability

### Logical View

- **EE Manager**
- **Hosts**: A, B, C, D, E, F, G, H, I, J
- **Groups**: Group#1 (B, D, G, I), Group#2 (A, J), Group#3 (C, E, H), Group#4 (F)
- **Ethernet Fabric**
- **PCIe Switches**: Switches connected to each group of hosts.
Dynamic Reconfiguration and Hot-Plug Capability

- 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.
 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 I/O 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 Lineup

- **1G/10G ExpEther**
  - **ExpEther HBA**
    - 1G
    - 1G
  - **ExpEther Client**
    - 2x 1000BASE-T
    - DVIx1, HDMI x1
    - USB3.0 x1
    - USB2.0 x3
    - Headphone x1
    - Microphone x1
  - **ExpEther IO Expansion Unit**
    - x16 PCIe x1 slot
    - Dual 1000BASE-T
    - x16 PCIe2 x2 slots (full height/full length)
    - Dual 10G SFP+ per slot

- **40G ExpEther**
  - **ExpEther HBA**
  - Network I/F: QSFP+ x 2
  - Form Factor: PCI Low Profile
  - 40G
  - 40G
  - 3U
  - 400mm
  - IO Expansion Unit
  - IO Interface: x8 PCI Express 3.0
  - Slots: x16 Slot x 4
  - Network I/F: QSFP+ x 4
  - Support IO: GPGPU (K80, P100, etc)
  - 19" Rack Size
  - 1,000W PSU for 2-Slot IO Expansion Unit
  - 800W PSU for 4-Slot IO Expansion Unit
  - 1G
  - 1G
  - 1G
  - 10G
  - 10G
  - 10G
  - 40G
  - 40G
Performance of EE vs Local with PCIe based SSD’s

There is no impact on bandwidth in ExpEther that can fully support PCIe x8 gen3 (64Gbps)

<table>
<thead>
<tr>
<th>name/depth</th>
<th>1</th>
<th>4</th>
<th>16</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>local</td>
<td>2363699.2</td>
<td>2266555.7</td>
<td>2264849.0</td>
<td>2197777.0</td>
</tr>
<tr>
<td>ExpEther</td>
<td>2657348.27</td>
<td>2490436.2</td>
<td>2491665.0</td>
<td>2462105.6</td>
</tr>
</tbody>
</table>

ExpEther/local (%) = 112.42% for read and 101.36% for write

There is no impact on bandwidth in ExpEther that can fully support PCIe x8 gen3 (64Gbps)
Performance of EE vs Local with PCIe based SSD’s

ExpEther can achieve the same IOPS as local by increasing the IO depth parameter to hide the latency of Ethernet.
ExpEther able to achieve almost the same performance as local case. In DeepLearning applications, the data exchange between host and GPU is very small and most process time is consumed in GPU, so there is no performance impact in ExpEther.
Ultra-Fast Failover Recovery for Database system with EE

High-speed data restore to Standby server for In-memory DB.

- NVMe SSD is faster than Fiber Channel.
- Use NVMe SSD as Journal for DB.
- When Active Server fails, NVMe SSDs’ connection is switched, allowing for DB journal restore on Standby Server.
Ultra-Fast Failover Recovery for Database system with EE and ExpressCluster X

- <Real pictures of the setup to be added here on left hand side>
•<Experiments results, observations and conclusions to be added here>
Ultra-Fast Failover Recovery for Database system with EE and ExpressCluster

• <Comparisons with existing Failover system architectures to be added here>
Service Acceleration Platform with ExpEther

Remote IO Devices

Compute Node

Accelerator Node

Hi-speed Storage Node

Accelerator Resource Pool

IO devices can be dynamically allocated to appropriate host according to workload
Future Roadmap of ExpEther or Universal Interconnect

- **Soft EE**
- **4G**
- **Main IoT Controller Server (EE HBA Cards) at Home**
- **WiFi**
- **WiFi HW EE**
- **NEC EE Chip**
- **HW EE**

- **Smartphones and watches**
- **Car Infotainment Systems**

**Home Automation, Entertainment and Security Appliances**

**Scalable/On-demand Remote IO Computing and Storage**
- **GPGPU's Stacked EE IO-Boxes**
- **NVMe's Storage Stacked EE IO-Boxes**

**NEC**

*Orchestrating a brighter world*
Summary

• <To be Added>
Thank you