NEW WORKLOADS AND THE EVOLVING NETWORK

Uri Cummings
Connectivity Group Chief Technical Officer, Data Center Group
Intel Corporation

uri.cummings@intel.com
By 2021, 95% of all data center traffic will be based in the cloud.

Source: Cisco Global Cloud Index, 2016-2021
Data center network inflections

Critical workloads redefine the network

Cloud Scale: the big get bigger

Pervasive offloads Distributed intelligence
Emerging workloads redefining the network

*Networks reconfigured for east-west, jitter reduction, and accelerators*

**Traditional Networks and Datacenter Fabrics**

**Network:** Flexibly-connected topology that enables endpoints to share data (e.g., arbitrary non-cyclic topologies)

**Scale-Out Fabric:** Specialized form of a network, with a fully-connected topology, that is engineered and tuned to sustain performance & latency at high scale while reducing jitter (e.g., Cloud Ethernet and HPC Fabric)

**Data and the Emergence of Accelerator Networks**

**Node-focused:**
- Light-weight network that interconnects a cluster of local accelerators with high-bandwidth, low-latency links
- Lowest power/bit
- Code & connectivity optimized for locality
- Emerging Need: Data-lakes

**Accelerator scaling fabric:**
- Used for high-bandwidth memory sharing/coordination at larger scale
- Allows separation of accelerator vs. general communications, minimizing interference in the scaling fabric.
Objective: Grow acceleration capabilities for network-related workloads

Agility
Continuous Integration

Cloud optimizes at the speed of Software

Infrastructure Acceleration
Lower Datacenter Tax
Reduce utility cycles in CPU

Application Acceleration
Enhance Growing Applications
Specialized infrastructure

Xeon Cores
Monetized Work
Infra.

Cloud
TOR
NIC
CPU
Infrastructure
Monetized work

Objective: Grow acceleration capabilities for network-related workloads
Emerging use case: bare-metal server

Bare metal servers

- Offer new usage models—full server rentals
- Improve infrastructure security in cloud
- Bolster performance, reduce cost/bit by removing bottlenecks
Cloud-era data center network

**Core (Spine)**
- Control CPU
- Ethernet Switch

**TOR (Leaf)**
- Control CPU
- Ethernet Switch

**Node (Shelf)**
- NIC
- Chipset
- CPU

**Accelerator Pool (Appliance)**
- NIC
- Chipset
- CPU
- LAN SoC

**Switch defines network architecture**
- Bandwidth and radix
- Power efficiency
- Reliability

**Bandwidth to the node is exploding**
- East-West communication
- ML, DL, application & network acceleration

**Single mode becoming ubiquitous**
- Copper $\rightarrow$ Optics $\rightarrow$ >100G

**NIC is innovation point for platform offloads**

Pod: Ethernet, PCIe, Inner Loop
Motivating smart NIC

*Integrating end-to-end capabilities in the platform*

<table>
<thead>
<tr>
<th>CORES/COMPUTE</th>
<th>FPGA</th>
<th>HOST INTERFACES</th>
<th>INTERFACE TECH.</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Xeon/Core</td>
<td></td>
<td>• PCIe Gen3, 4 and 5 IAL (I/O, Memory, Cache)</td>
<td>• DMA</td>
</tr>
<tr>
<td>• Atom</td>
<td></td>
<td>• Stratix10 Flexible tile architecture</td>
<td>• SVM/Coherent</td>
</tr>
<tr>
<td>• Graphics/Media</td>
<td>• Flexible state machines</td>
<td></td>
<td>• Hardware queue engines</td>
</tr>
<tr>
<td>• QAT/Crypto/Comp.</td>
<td></td>
<td>• CPU/Coherent</td>
<td>• Virtual device manager</td>
</tr>
<tr>
<td>• Special purpose (accelerators)</td>
<td></td>
<td>• On-package (RLink)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NETWORK</th>
<th>SOFTWARE</th>
<th>PACKAGE</th>
<th>STORAGE MEMORY</th>
</tr>
</thead>
<tbody>
<tr>
<td>• MAC/PHY/SerDes</td>
<td>• Compilers, tools WOS OTC/Intel Clear Linux P4 API working group Libraries/Standards bodies</td>
<td>• BGA: Low-cost to large footprint Package MCP &amp; Interfaces (RLink)</td>
<td>• NVMe</td>
</tr>
<tr>
<td>• RDMA (RoCE, iWARP)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• OPA for HPC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Packer parsers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• P4 engines</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Schedulers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Switches</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Programmable assets
- FPGA
- CPUs
- Flexible state machines

Networking assets
- Broad market NIC business
- Open source contribs
- DPDK
OCP NIC v3.0 enables the smart NIC

Joint effort across system/NIC/Connector suppliers and end users

Community development since Feb 2017 solved major challenges

- Improved serviceability
- Manage NIC / platform transition
- Ready for PCIe Gen4 and Gen5
- Enlarge PCB space for smart NIC use case
- Allow higher TDP ASICs
- Improve mechanical interface

<table>
<thead>
<tr>
<th></th>
<th>Single connector</th>
<th>Dual connector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>Low profile—similar profile to OCP NIC 2.0 card</td>
<td>Larger PCB width to support additional NIC</td>
</tr>
<tr>
<td>Large</td>
<td>Up to 16 PCIe lanes</td>
<td>Up to 32 PCIe lanes</td>
</tr>
<tr>
<td></td>
<td>80W max</td>
<td>150W max</td>
</tr>
</tbody>
</table>
Significant OCP contributions in...

OCP Ethernet Network Adapters

>25 OCP-compliant NIC and PHY SKUs, covering 1GbE, 10GbE, 25GbE, and 40GbE

>1M Intel-branded ports shipped in OCP form factor (excluding partner shipments)

100G Silicon Photonics

Fully compliant with 100G CWDM4-OCP, CWDM4 MSA, QSFP, and CAUI-4 specs

500m, 2km, or 10km reach on duplex single mode fiber
Intel 400G CWDM8 QSFP-DD optical module

Technology for 400G is working in the lab

400G Transmitter: Optical Output

400G Receiver: Host Side Electrical Output