ODSA: Technical Introduction

Bapi Vinnakota, Netronome
ODSA Project Workshop
March 28, 2019
ODSA: A New Server Subgroup (Incubation)

- Extending Moore’s Law
  - Domain-Specific Architectures: Programmable ASICs to accelerate high-intensity workloads (e.g. Tensorflow, Network Flow Processor, Antminer...)

- Chiplets: Build complex ASICs from multiple die, instead of as monolithic devices, to reduce development time/costs and manufacturing costs.

- Open Domain-Specific Architecture: An architecture to build domain-specific products
  - Today: All multi-chiplet products are based on proprietary interfaces
  - Tomorrow: Select best-of-breed chiplets from multiple vendors
  - Incubating a new group, to define a new open interface, build a PoC
  - Today is our first workshop as an OCP project!

Thanks to:
Achronix: Quinn Jacobson, Manoj Roge; Aquantia: Ramin Farjad; Avera Semi: Dan Greenberg, Mark Kuemmerle, Wolfgang Sauter; Ayar Labs: Shahab Ardalan; ESNet: Yatish Kumar; Kandou: Brian Holden, Jeff McGuire; Netronome: Sujal Das, Jim Finnegan, Jennifer Mendola, Brian Sparks, Niel Viljoen; NXP: Sam Fuller; OCP: Bill Carter, Archna Haylock, Dharmesh Jani, Steve Roberts, Seth Sethapong, John Stuewe, Aaron Sullivan, Siamak Tavallaei; Samtec: Marc Verdiell; Sarcina: Larry Zu; zGlue: Jawad Nasrullah.
Tailor architecture to a domain*

- Server-attached devices — programmable, not hardwired
- Integrated application and deployment-aware development of devices, firmware, systems, software
- 5-10X power performance improvement

- Big - more of a processor to I/O mismatch => more memory
- Each serves a smaller market

*A New Golden Age for Computer Architecture
John L. Hennessy, David A. Patterson

Google TPU vs. CPU and GPU

Source: "An in-depth look at Google’s first Tensor Processing Unit (TPU)," Google Cloud, May 2017

Netronome NFP vs. CPU and FPGA

Source: Netronome, based on internal benchmarks and industry reports related to Xeon CPUs and other FPGAs
Monolithic vs Chiplets

Shrink: Monolithic process shrink
Integration: Multi-chip on same process

Integration provides nearly all the benefits of a shrink at a fraction of the cost, because of efficient inter-chiplet interconnect

AMD Data
4 Die are ~30% cheaper than a single large die

https://www.netronome.com/media/documents/WP_ODSA_Open_Accelerator_Architecture.pdf
COST & PERFORMANCE DISPARITY IN SCALING

High Density Multi Chip Packaging interconnect
- 0.1-10 mm; 500 IO/mm shoreline; 0.1-1.0 pJ/bit
- \( FOM_{\text{LongReach}} = \frac{1 \text{Tb/mm}}{1 \text{pJ/bit}} = 1000 \)
- \( FOM_{\text{ShortReach}} = \frac{1 \text{Tb/mm}}{0.1 \text{pJ/bit}} = 10,000 \)
- Latency = 2.5 ns

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)
PHY Layer Options

Interface Comparisons

<table>
<thead>
<tr>
<th>Interface</th>
<th>BoW</th>
<th>USR</th>
<th>XSR (CDR)</th>
<th>MR</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interop</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Legacy Tech</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Si FOM (Density/p)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lam FOM (Density/P)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BW Density (Si)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BW Density (Laminate)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Power FOM (larger is better)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Common Logical Layer

- Bandwidth at all Cost
  - High Data Rate DoW
  - Interposer
  - Si Bridge

- Intercopper
- Assembly Yield/Rework

Cost-Optimized Interfaces

- PCIe
- VSR
- BoW MCM
- XRS
- USR

Technology Choices
- Better Materials

https://www.netronome.com/media/documents/WP_ODSA_Open_Accelerator_Architecture.pdf
Domain-specific accelerators

- Host-attached programmable logic optimized for an application domain
  - Tensorflow, Netronome NFP, Crypto, IoT,...
- Domain-specific accelerators contain lots of generic logic ~35-45% of silicon area, development time
  - Network, Host, Memory Interfaces
  - General-purpose CPUs
  - SRAM, interconnect
  - Domain-specific logic works in coordination with host and/or CPU SW
- Ideally
  - Investment in a DSA should be limited to the domain-acceleration logic
- In reality
  - Buy IP for the “non-core” parts, spend $$'s test and integration
Multi-Chiplet Reference Architecture for DSA

<table>
<thead>
<tr>
<th>Design Function</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP Qualification</td>
<td>Verified IP for inter-chiplet communication</td>
</tr>
<tr>
<td>Architecture</td>
<td>Leverage reference architecture.</td>
</tr>
<tr>
<td>Verification</td>
<td>Focus investment on domain-specific logic.</td>
</tr>
<tr>
<td>Physical</td>
<td>Reuse chiplets instead of IP for 40% of the functions in a monolithic design</td>
</tr>
<tr>
<td>Software</td>
<td>Open source firmware and software for host-attached operation</td>
</tr>
<tr>
<td>Prototype</td>
<td>Aim for reference package design with area, power budgets and pinouts for components</td>
</tr>
<tr>
<td>Test and Validation</td>
<td>Develop workflow for chiplets</td>
</tr>
</tbody>
</table>
Multiple chiplets need to function as though they are on one die.

Open Interface for Chiplet-Based Design

- **Application**: Firmware Infrastructure for: Accelerator Use, Host Integration
  - Coherent Protocol (CCIX^2 or TileLink^2 or...)
  - Instruction-Driven Transfer (ISF Transport Layer^2)
  - DMA

- **Network**: Routing (To Be Developed From ISF Routing Layer)

- **Link**: Inter-Chiplet (e.g. PCIe, TileLink^1 or new Link Layer^1)
  - Intra-Chiplet (From ISF Link Layer^1)

- **PCIe PIPE**: PIPE Adapter
- **PHY**: PCIe^2 PHY, XSR^2, USR SerDes, Bunch of Wires (BoW)

- **Substrate**: Organic Substrate Packaging^1 + Interconnect

- **Chiplet**: NFP, RISC^2, FPGA, ML ASIC, SerDes

1 New Open IP/Specification
2 Existing Open Standard

Source: ODSA
Multiple OCP projects use accelerators
Open architectural interface to support accelerator designs across multiple carrier cards
Power, management, reliability requirements vary across sockets
Enable a collection of ODSA-compliant chiplets, packages, sockets, in the OCP marketplace
## ODSA Landing Zones

<table>
<thead>
<tr>
<th></th>
<th><strong>Network I/O</strong></th>
<th><strong>Host I/O</strong></th>
<th><strong>Power</strong></th>
<th><strong>Size</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NIC 2.0</strong></td>
<td>Dual port x 25</td>
<td>X16 PCIe Gen 3</td>
<td>25w</td>
<td></td>
</tr>
<tr>
<td><strong>NIC 3.0</strong></td>
<td>Dual port x 200</td>
<td>SFF: x16 PCIe Gen 4/Gen 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>LFF: x32 PCIe Gen 4/Gen 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>M.2</strong></td>
<td>N/A</td>
<td>Single: x4 PCIe Gen 3/Gen 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>M.2 Dual</strong></td>
<td></td>
<td>Dual: x8 PCIe Gen 3/Gen 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>OAM</strong></td>
<td>8x16 SerDes Lanes</td>
<td>Typical: x16 PCIe</td>
<td>12V: 350w</td>
<td>102x165</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>48V: 700w</td>
<td></td>
</tr>
<tr>
<td><strong>Olympus</strong></td>
<td>Via x16 PCIe Cards</td>
<td>1x16 PCIe</td>
<td>75W-300W PCIe AIC</td>
<td>FHHL PCIe</td>
</tr>
<tr>
<td><strong>Tioga Pass</strong></td>
<td>Up to 100Gbps SH</td>
<td>x32 PCIe Gen3</td>
<td></td>
<td>6.5x20inch</td>
</tr>
</tbody>
</table>

Data from Ron Renwick, John Stuewe, Siamak Tavallaei, Whitney Zhao
Cross-chiplet ODSA fabric proposal

1 New Open IP/Specification
2 Existing Open Standard

Source: ODSA
Progress Since the Last Workshop

• Timeline:
  − ODSA Announced 10/1/18 7 companies
  − White Paper 12/5/18 10 companies
  − First Workshop 01/28/19 35 companies
  − Joined OCP 03/15/19
  − Today 03/28/19 53 companies

• PoC
  − Identified components, use cases

• Standards
  − Characterizing PHY, new interface proposal

• Business
  − Survey, business model
TIL in the last six months

- We’re solving the right problem, tbd on whether it’s the right solution.
- Analog (and cache coherence) engineers have lots of opinions, likely justified, but also confusing for mere mortals.
- How you do business drives chiplet economics and your technology choices.
  - Our interface definition must recognize this diversity while focusing our effort.
  - You need a new business/workflow model that make chiplets work across this diversity
How to Participate
Please Help! : Join a Workstream

Join the PoC, Build fast:
(Quinn Jacobson/Jawad Nasrullah)

Join Interface/Standards:
(Mark Kuemerle/Aaron Sullivan)

Join Business, IP and workflow:
(Sam Fuller/Jeff McGuire)

Provide ODSA chiplets
Provide FPGA IP
Develop Packaging + Socket, Dev Board
Provide PHY technology
Define Architectural Interface

Workstream contact information at the ODSA wiki