The Fungible DPU™:
A New Category of Microprocessor for the Data-Centric Era

Hot Chips 2020
Our Mission

Revolutionize the **performance, economics, reliability** and **security** of all scale-out data centers

Core Technology: a new category of microprocessor called the **Fungible DPU™**, associated software, and systems
Context
Problems Facing Data Centers

Large footprint (power and space)
  • Inability to pool expensive resources
  • Inefficient execution of Data-Centric Computations

Scaling challenges (very small & very large)

Increasing complexity (technology limits)

Security vulnerabilities
Headwinds

Network and storage speeds are increasing faster than compute
Applications need to access ever larger data sets
Moore’s Law slowing and will likely plateau
Security attacks increasing in frequency and sophistication
Data-Centricity Will Drive the Architecture

**EARLY-MID COMPUTE-CENTRIC ERA**
- Multiple CPU Types to Single CPU Type

**LATE COMPUTE-CENTRIC ERA**
- X86
- GPU
- SSD
- HDD

**DATA-CENTRIC ERA**
- Scale-out of Hyperconverged X86 Servers
- Scale-out of Hyperdisaggregated Servers Powered by the Fungible DPU™

Timeline:
- 1957
- 1969
- 1981
- 1993
- 2005
- 2017
- 2020
Our Approach:
Clean Sheet, Fundamentals Based Design
Confront the Root Causes

Inefficient data interchange between nodes

Inefficient execution of data-centric\(^1\) computations inside nodes

Unreliability

Inflexibility

Insecurity

---

\(^1\) Data-Centric Computations:
- All work arrives as packets.
- Require frequent context switching.
- Involve modification of state.
- I/O dominates arithmetic and logic.
Fungible DPU™ Addresses all Five Root Causes

- **Efficient interactions between nodes**
- **Full data-path programmability**
- **Efficient execution of data-centric computations**
- **Strong security with no compromise to performance**
- **Fundamental improvements to infrastructure reliability**
Where does the Fungible DPU™ sit?

Fungible DPU™ in each node
• Connects node to network
• Functions as intra-node hub

Programmable data and control planes

Handles data-centric computations

End point of TrueFabric™

Provides:
✓ Full flexibility with high performance
✓ Strong end-to-end security
✓ Pooling of resources
✓ Reliable low-jitter, low latency fabric
✓ Observability and telemetry
The Fungible F1 DPU™
Fungible F1 DPU™ Architecture

8 Data Clusters
- 192 processor threads
- Full cache coherency
- Tightly integrated accelerators

Control Cluster
- 8 processor threads
- Secure complex
  - HSM
  - Root-of-trust
- Work scheduler

Memory & I/O Interfaces
- 800 Gbps network unit
- 4x16 G3/G4 PCIe unit
- DDR4
- HBM

On Chip Network
- High performance
- Low latency
- Any unit to any unit
Data Cluster

Runs the Data Plane
6 Cores * 4 Threads

Multi-Threaded Accelerators
- Data movement
- Data lookup
- Data security
- Data reduction
- Data protection
- Data analytics
Control Cluster

- Runs the Control Plane on Linux
- 4 cores * 2 threads

Secure Enclave
  - Secure Boot
  - Secure Key Vault
  - Binary Signing and Authentication

Public Key Crypto Engines
  - RSA
  - Elliptic Curve

True Random Number Generator

Physically Unclonable Function
CPUs, Caches, External Memory

**CPUs**
- MIPS-64, 9-stage, dual-issue, 4xSMT, FPU/SIMD unit
- IPC on data-centric workload close to CPU-max
- Full hardware virtualization
- Large I+D L1\$, shared L2\$, full system-wide coherency

**High Bandwidth HBM2 Memory**
- 8GB, 4Tbits/sec
- Integrated in the package

**High Capacity DDR4 Memory**
- 2xDDR4 controllers, ECC enabled, up to 2666 MHz
- Up to 512GB
- Support of RDIMM, NVDIMM-N

- Fully general programmability
- All code in ANSI-C
- Fast thread switching
- Tight coupling with accelerators
- No performance compromises
High-Performance (800G) Flexible Network Engine

- Low, deterministic latency
- Full cross-section bandwidth
- End-to-end congestion control
- End-to-end error control
- End-to-end encryption
- Network virtualization
- Granular QoS
- Enables disaggregation at scale

- Implements TrueFabric™ end point
- Low latency Ethernet MAC with FEC
- Integrated L2/L3/L4 forwarding
- Low latency transit switching
- Support of general virtualization protocols
- Tight integration with data clusters
- P4-like language controls
  - Parsing, encapsulation, decapsulation
  - Rx/Tx acceleration
  - Lookup acceleration
- All packets are AES-GCM encrypted
- Precision time protocol

- High-Performance (800G) Flexible Network Engine
- Implements TrueFabric™ end point
- Low latency Ethernet MAC with FEC
- Integrated L2/L3/L4 forwarding
- Low latency transit switching
- Support of general virtualization protocols
- Tight integration with data clusters
- P4-like language controls
  - Parsing, encapsulation, decapsulation
  - Rx/Tx acceleration
  - Lookup acceleration
- All packets are AES-GCM encrypted
- Precision time protocol
High-Performance (512G) Flexible Host Engine

Includes 16 independent dual-mode controllers
  • EP/RC in any combination (4x16 to 16x4)

End point for X86 or ARM CPUs
  • Hardware virtualization
    • SR-IOV with 64 PFs, 1024 VFs
    • Fine-grain QoS support
  • Software flexibility
    • Full network, storage, and security virtualization

Root complex
  • High performance abstraction layer
  • Fully flexible data and control planes
  • Connects to and abstracts SSDs, GPUs, FPGAs
Data Path Programming Model

MIPS-64 Hardware Threads Execute Run-To-Completion C-Code

Host Engine

High-Performance Any-to-Any “Call-Continue” Fabric

Network Engine

Heterogeneous Accelerator Threads
Fungible DPU™ Software

**Cluster Services**
- Topology, Configuration, Monitoring, Telemetry, Policy

**Control Plane**
- Northbound Rest APIs
- Southbound APIs
- FunVisor
- Linux
- VM OS
- Hypervisor / OS
- PCIe VFs
- Agents

**Data Plane (FunOS)**
- Network
  - *PoE
  - *Fabric
  - *Full TCP
  - *Full RDMA
  - *Full TLS
  - *VXLAN
  - Stateless Offload
  - Encryption
  - Granular QoS
- Storage
  - *NVME-o-FCP
  - *NVME-o-TCP
  - *Initiator + Target for Block Store
  - KV Store
  - File Store
  - Erasure Coding
  - Compression
  - Encryption
  - *Pooling
- Virtualization
  - *VXLAN
  - *V-Router
  - *V-Switches
  - *Per VF QoS
  - *Full BMV
  - *Full RPC Offload
  - All Accelerators exposed via PCIE VF’s
- Security
  - *PUF
  - *Secure Enclave
  - *Key-Gen
  - *Key-Store
  - *TLS
  - *Regex
  - *Hash
  - *Stateful Firewall
- Analytics
  - *Sort
  - *Filter
  - *Join
  - *Select
  - *Map
  - *Reduce
  - *Lambda-o-FCP
  - *Lambda-o-TCP

**FunOS Nucleus** (scheduling, timer, memory management)

**DPU**
- CPU (X86)
- PCIe VFs
- Southbound APIs
Multiple Levels of Programmability

Software running on the DPU
- DPU control plane
- DPU data plane

Software running on a PCIe connected X86 Host
- OS drivers and Agents
- Data path code execution via eBPF

Cluster Services for management and control of multiple DPU systems
- Northbound APIs for orchestration systems
## Infrastructure Services Performance

<table>
<thead>
<tr>
<th>Service</th>
<th>Measured Performance</th>
<th>Estimated(^1) Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCP(^2) (Single Flow, Multi Flow)</td>
<td>50Gbps, 250Gbps</td>
<td>70Gbps, 400Gbps</td>
</tr>
<tr>
<td>TLS(^2) Session Setup Rate</td>
<td>32,000/sec</td>
<td>100,000/sec</td>
</tr>
<tr>
<td>IPSEC(^2) (Single Flow, Multi Flow)</td>
<td>-</td>
<td>10Gbps, 250Gbps</td>
</tr>
<tr>
<td>Stateful Firewall(^2)</td>
<td>-</td>
<td>370Gbps</td>
</tr>
<tr>
<td>OVS</td>
<td>-</td>
<td>400Gbps</td>
</tr>
<tr>
<td>Load Balancer</td>
<td>256Gbps</td>
<td>300Gbps</td>
</tr>
<tr>
<td>Block Store (4K IOPS)</td>
<td>8M</td>
<td>10M</td>
</tr>
<tr>
<td>Video Streaming</td>
<td>256Gbps</td>
<td>300Gbps</td>
</tr>
<tr>
<td>TPC-H Benchmark (relative to X86)</td>
<td>3X-100X</td>
<td>-</td>
</tr>
</tbody>
</table>

\(^1\) Full chip dedicated to service  
\(^2\) All measurements are full-duplex
## TrueFabric™ Performance

<table>
<thead>
<tr>
<th>Traffic Pattern</th>
<th>1024 * (Node to Node)</th>
<th>1024 Node to 1024 Node</th>
<th>1024 Nodes to 1 Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fabric Utilization</td>
<td>90.7%</td>
<td>93%</td>
<td>90%</td>
</tr>
<tr>
<td>Latency Mean</td>
<td>1.84µs</td>
<td>2.10µs</td>
<td>1.71µs</td>
</tr>
<tr>
<td>Latency Variance</td>
<td>0.13µs</td>
<td>0.32µs</td>
<td>0.12µs</td>
</tr>
<tr>
<td>Latency P99</td>
<td>2.14µs</td>
<td>3.30µs</td>
<td>1.75µs</td>
</tr>
</tbody>
</table>

Network Configuration:
- 1024 Nodes
- 200Gbps/Node
- Two-tier leaf-spine
- Leaf ZLL: 500ns
- Spine ZLL: 500ns
- iMix packet profile

![Latency Distribution](image1)
![Latency Distribution](image2)
![Latency Distribution](image3)
A New Category of Microprocessor
Purpose-built for the data-centric era

**CPU**
- General-purpose
  - Multi-core, MIMD
  - High IPC for single threads
  - Fine-grain memory sharing
  - Classical cache coherency
  - Based on locality of reference
  - Ideal for low to medium I/O

**GPU**
- Vector floating point
  - Multi-core, SIMD
  - High throughput for vector processing
  - Coarse-grain memory sharing
  - Relaxed coherency
  - Based on data >> instructions
  - Ideal for graphics, ML training

**Fungible DPU™**
- Data-centric
  - Multi-core, MIMD + tightly-coupled accelerators
  - High throughput for multiplexed workloads
  - TrueFabric™ enables disaggregation and pooling
  - Specialized memory system and on-chip fabric
  - Ideal for network, storage, security, virtualization
  - Data-centric computations run >10X more efficiently
Announcing the Fungible F1 and S1 DPUs
Common Architecture and Programming Model

- Storage target
- AI Server
- Security appliance
- Analytics

- Bare-metal virtualization
- Storage initiator, local instance storage
- NFV applications
- Node security
THANK YOU