DPU BOF

Presented by Dr. Jai Menon, Chief Scientist, Fungible
What is a DPU?

- A DPU or data processing unit is a specialized programmable processor tailored to efficiently execute data-centric tasks – they integrate general-purpose cores and hardware accelerators.

- Data-centric tasks involve stateful, multiplexed processing of high bandwidth streams of data.

- Storage, network and security data path processing are all data-centric.

DPUs complement CPUs & GPUs and will be a 3rd socket in data centers.
What is the Fungible DPU™

The Fungible DPU is a new class of programmable microprocessor that:

- Enables 10x more efficient execution of data-centric tasks; handle 10X events
- Implements an endpoint for a more efficient data center network that uses standard IPoE switches

F1 DPU runs 192 hardware CPU threads and dozens of accelerators in parallel

Tightly Coupled CPUs & Accelerators
Purpose-Built CPUs for data-centric tasks
Purpose-Built Memory Systems & Fabrics
Specialized multi-threaded Accelerators
Fully integrated with memory systems
Fungible’s Programming Model – Performance with Flexibility

Programmable Host Engine

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

Programmable Network Engine

CPU threads Execute Run-To-Completion C-Code

Heterogeneous Accelerator Threads

Traditional DPUs have loose coupling between cores and accelerated path
What can be Built Using DPUs?
Four Examples of DPU-Based Products

Storage Systems, Host-Based PCIe Cards, Disaggregated GPU servers

#1
High performance block, file, KV storage with line rate EC, compression, encryption

#2
PCle Host Cards (offload storage, network, security – initiator and target)

#3
Computational storage

#4
Disaggregated GPU servers
DPU-Based Next Gen Data Center

- **Standard x86 Servers**
  - Can be Diskless, GPU-less
  - DPU offloads network, security, storage

- **DPU-based GPU Servers**
  - Remote GPU pools with close to local performance

- **DPU-based Storage Servers**
  - (Flash, Disk)
  - Remote storage pools with close to local performance

- **TrueFabric (IPoE)**
  - Low zero load & tail latency
  - Congestion avoidance

- **Standard IPoE**
Tools, Libraries and Languages
DPU Programming can be Done at Various Levels

- All DPU configuration is exposed via REST API calls so that developers can configure & automate storage and network resources. Any REST-capable language can be used. New agents to expose new API calls can be built using standard Linux tools.

- The DPU exposes the various crypto/compression/EC accelerators via PCI and accessed via host device drivers. Any language that can exploit a kernel driver interface can use these features.

- Computational storage and network filtering can be programmed using eBPF, PCRE and domain specific languages such as SQL. This allows arbitrary code execution to be moved to where the data resides or is already moving in the network.

- Developers can use the FunSDK (coming soon) in C to build & extend with novel data paths.
Summary

- Next Generation Data Centers will need
  - DPUs as the third socket
    - To efficiently handle data-centric tasks
  - DPU-based storage and DPU-based host cards
    - To provide penalty-free disaggregation of resources (like storage and GPUs)
    - To enhance data-center networks
THANK YOU