Using SmartNICs as New Platform for Storage Services

Fazil Osman
Broadcom Corporation
Why a SmartNIC

- Moore’s Law diminishing returns
  - Vertical scaling power & cost model no longer viable
- CPU costs increasing
  - Economic benefits to limiting core count
- Multi-socket interconnect bottleneck
  - I/O, memory transactions across interfaces add latency
  - 2nd socket often used to get more memory and I/O
  - TCO penalty for 2nd socket
- Distributed cloud architecture
  - Smaller fault domains

![42 Years of Microprocessor Trend Data](image)

Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten.


![Monthly Costs](image)

3yr server & 10yr infrastructure amortization

Source: James Hamilton, AWS
What is SmartNIC

Architectural flexibility to **quickly offload** multiple overhead IaaS services

---

- **Onload** Hardware Appliances...
  - Firewall
  - IDS/IPS
  - SD-WAN
  - Router
  - ADC
  - vTAP
  - Packet Broker

- **...Offload** SDS, SDN, NFV Services
  - NVMe-oF
  - RAID/EC
  - KV Store
  - IPSec/SSL/TLS
  - vSwitch
  - vRouter
  - NFV VNFs
Evolution of SmartNIC…

**FPGA + NIC**

**Pros**
- Typical single-function offload
- Good performance

**Cons**
- Hard to design for performance
- Slow feature velocity (RTL)
- High power
- Large devices are expensive

**Network Function Processor**

**Pros**
- More than single function

**Cons**
- Non-standard programming
- Can be expensive
- High power

**SmartNIC**

**Pros**
- Performance/Watt
- General-purpose with standard programming
- Great feature velocity

**Cons**
- Performance varies based on CPUs, DDR, and availability of integrated accelerators

**HFT, HPC, Telco I/O**

**Telco I/O**

**Cloud DC & Telco**
## Platform Economics: CPU Workload Partitioning

<table>
<thead>
<tr>
<th>Service</th>
<th>Typical Utilization</th>
<th>Example (165W, 18C)</th>
<th>Example (165W, 18C)</th>
<th>~18 Cores Remaining</th>
<th>~4-6 Cores Consumed</th>
<th>~2-4 Cores Consumed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage Services</td>
<td>&lt;50% Utilization Typical with Virtualization*</td>
<td>x86 36 Cores</td>
<td>x86</td>
<td>18 Cores Remaining</td>
<td>4-6 Cores Consumed</td>
<td>2-4 Cores Consumed</td>
</tr>
<tr>
<td>Networking Services</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Technologies</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AES Encryption</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVMe-oF™</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OVS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>contrail</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VNF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Only 8-12 Cores Available for Applications

- Services can consume most of the remaining cores

* Gartner and Moor Insights & Strategy

~ $8,000 Platform
Including Southbridge and High-Performance NIC

~ 380W Platform
Including Southbridge and High-Performance NIC

Typical with Virtualization* ~18 Cores
Often oversubscribed for memory

18 Cores Remaining

~4-6 Cores Consumed

~2-4 Cores Consumed

~36 Cores
x86

~18 Cores Remaining

~4-6 Cores Consumed

~2-4 Cores Consumed

~380W Platform
Including Southbridge and High-Performance NIC

~8,000 Platform
Including Southbridge and High-Performance NIC
Platform Economics: SmartNIC Workload Partitioning

~ $4,000 Platform
Including Southbridge and High-Performance NIC Built Into SmartNIC

<50% Utilization
Typical with Virtualization*

Storage Services

Minimal Virtualization Overhead

AES Encryption
NVMe-oF™

Networking Services

16-18 Cores Remaining

16-18 Cores Available for Applications

16-18 Cores Available for Applications

Run on SmartNIC

Run on SmartNIC

Offloading services to SmartNICs frees up cores for applications

* Gartner and Moor Insights & Strategy
SmartNIC Storage Use Cases

- Server CPU
- Datacenter Fabric
- SmartBOF

**Clients**
- x86
- SmartNIC 10/25/50G
- Eliminates x86 Socket
- Simplifies Operational Model
- Greater Security Isolation

**Targets**
- Smaller Fault Domain
- Lower TCO
- SmartNIC 100G

**Compression**
- NVMe

**Storage Services**
- iSCSI to NVMe-oF bridge
- RAID/EC
- End-to-end software-defined storage
- End-to-end NVMe/TCP
- NVMe-oF target stack
- In-memory database
- Virtual SAN
- Key-value store

2019 Storage Developer Conference. © Broadcom Corporation. All Rights Reserved.
### Example: Small vs. Large Fault Domains

#### Test Summary

<table>
<thead>
<tr>
<th>Parameter</th>
<th>4x Stingray Targets</th>
<th>2 Socket-x86 Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network Link</td>
<td>4x 25G</td>
<td>1x 100G</td>
</tr>
<tr>
<td>NVMe SSDs (x2 Gen3)</td>
<td>32</td>
<td>30</td>
</tr>
<tr>
<td>4K Random Read</td>
<td>2.0M IOPS</td>
<td>1.8M IOPS</td>
</tr>
<tr>
<td>512K Sequential Write</td>
<td>37K IOPS</td>
<td>18K IOPS</td>
</tr>
<tr>
<td>Tail Latency (mean – P90% – P99.9%)</td>
<td>2 ms – 6.2 ms – 11 ms</td>
<td>2.3 ms – 12.9 ms – 23.5 ms</td>
</tr>
<tr>
<td>CPU+DRAM Power (estimated)</td>
<td>160W</td>
<td>300W</td>
</tr>
</tbody>
</table>

#### SmartNIC Disaggregated Storage Advantages
- Better performance
- Lower power
- Smaller fault domain reduces blast radius exposure (16TB vs 60TB)
Stingray-Based Storage Platforms

WDC F3100 and Stingray
- Low power
- NVMe-oF Enabled
- Software-programmable
- Low latency
- Up to 10 Modules and >20M IOPS

2U-24 Drive Systems
- Full HA Support
- NVMe-oF Enabled
- >10M IOPS

Scale Out

Enterprise Class

Celestica Euclid

AIC Manta

Winstron Lymma

Copyright © 2019 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.
RedisEdge on SmartNIC Use Case

Deployed within hours using standard container
SmartNIC in NVMe-oF™ – We Have Come A Long Way but…

OS Support for NVMe-oF
- Limited to recent versions of Linux
- No announced support for other operating systems

Ecosystem is maturing but broad adoption requires solution to OS support problem
**Pure Storage Use Case**

Bringing NVMe over Fabrics to Windows client solutions

---

**Problem:**

Windows does not support NVMe-oF natively limiting FlashArray scalability

**Solution:**

Co-developed Windows iSCSI to NVMe-oF on Stingray SmartNIC

---

SQL SERVER DB SNAPSHOT PORTABILITY WITH CLOUdsnAP

STINGRAY ENABLING NVME-OF™ FOR PURE FLASHARRAY//X

Windows SQL server is largest application with Pure customers
Broadcom Glass Creek Adapter

Introducing Glass Creek

Applications

- Storage disaggregation for any OS
- Works with standard NMVe drivers
- Storage virtualization
  - Bare metal and virtualized servers
- Storage services offload
  - Logical Volume Management
  - RAID/EC, De-dupe, Crypto

- NVMe-oF Initiator
- Networking Services vSwitch
- Storage Services RAID/Crypto/LVM
- Accelerators
  - TruFlow™
  - Crypto
  - RAID/EC
  - De-dupe
- L2
- RoCE
- NVMe
- SlimSAS
- 2x25G
- 50G NIC
- 2x25G

SSD...SSD

2019 Storage Developer Conference. © Broadcom Corporation. All Rights Reserved.
Stingray Solutions

- **NVMe-oF SmartBOFs**
  - High performance fabric storage appliances
  - Enterprise-capable high-availability solutions
  - Cloud scale with low blast radius

- **Expanding NVMe-oF™ ecosystem**
  - Multiple vendors
  - NVMe-oF and **now** NVMe virtualization

- **Highest performance SmartNIC**
  - System architecture, cost and performance
  - End of Moore’s Law
  - Dual socket architectures are inefficient

- **SmartNIC Use cases**
  - General-purpose programmability
  - Offloading storage and networking services
  - Bare metal and virtualization servers
  - Security