The SmartNIC Revolution for Efficient Cloud Infrastructures

Ash Bhalgat, Sr. Director, Cloud Marketing, Mellanox
MWCA-LA, October, 2019
Mellanox Overview

$1.09B  
2018 Revenue

$5.01  
2018 Non-GAAP EPS

$265M  
2018 Cash From Operations

$439M  
2018 Ending Cash & Investments

~2,800  
Employees worldwide

Company Headquarters:
- Yokneam, Israel
- Sunnyvale, California
- Worldwide Offices

Ticker: MLNX
Comprehensive End-to-End Portfolio

InfiniBand and Ethernet

- Software
- NPU & Multicore
- Metro / WAN
- Switches/Gateways
- Adapter Cards
- ICs
- Cables/Modules
Mellanox Cloud Scale Architecture

Use Smart Interconnects to Disaggregate and Virtualize Data Center Infrastructure
Intelligent Offloads – Higher Performance, Efficiency, Scalability
Software Defined Everything Creates Bottlenecks

- **Software Defined Everything (SDX)** Consumes CPU cores for Packet/Flow Processing and Creates Bottlenecks
  - Virtualization, Storage, Switching, Routing, Load Balancing

- **Security**: Consumes CPU cores for Security Processing
  - Layer 4 Firewall, encryption, host introspection
  - Intrusion detection & prevention

Available for Application Processing

**Bare Metal**

**Software Virtualization**

**Accelerated Virtualization**

**SDX in SmartNIC**

**Security in SmartNIC**
Software Defined Network, Storage, Security Transition

VM/Container

Hypervisor
- Network Virtualization
- Storage Virtualization

NIC Hardware

VM/Container

Hypervisor
- Network Ctrl Pane
- Storage Virtualization

NIC Hardware

VM/Container

Hypervisor
- Network Virtualization
- Storage Virtualization
- Security

Smart NIC HW

Bare Metal Server

Hypervisor
- Network Virtualization
- Storage Virtualization
- Security

Smart NIC HW
Mellanox SmartNICs – an Acceleration Strategy

**Commodity NICs**
- Not programmable
- Basic segmentation offloads
- 1G/10G NICs with CPU doing heavy lifting
- Priced as per the value

**No Mellanox NIC for this market**

**ConnectX-5/6/6-Dx**
- Best performance for price
- Built-in hardware offloads
- Extra flexibility, efficiency and performance

**BlueField 1 and 2**
- Highly customizable
- Leverage hardware accelerations
- Full programmability
ConnectX-6 Dx - Datacenter Security Adapters

Leading NIC Capabilities and Performance
- Single and Dual port Ethernet 10 / 25 / 50 / 100Gb/s
- Single port 200GbE
- 50Gb/s PAM4 SerDes
- PCIe Gen3/Gen4 x16
- Total bandwidth 200Gb/s, message rate of up to 215Mpps

Advanced Security Capabilities
- Inline IPsec (100Gb/s) and TLS (200Gb/s) crypto for data in motion
- Storage encryption for data at rest (as in ConnectX-6)
- Key management
- Hardware Root-of-Trust for secure boot
- Stateful rules checking for Connection Tracking

Advanced Networking Capabilities
- Advanced Network Virtualization Offload Engine (ASAP²)
  - SR-IOV and VirtIO Accelerations
  - RoCE enhancements (for Lossy networks)
  - Selective Repeat
  - Congestion control enhancements
BlueField-2 – New Range of Capabilities

High Performance SoC
- Embedded ConnectX-6 Dx adapter
- Single and Dual port Ethernet & InfiniBand 10/25/50/100Gb/s, single 200Gb/s
- PCIe Gen3/Gen4 x16, total throughput 200Gb/s
- 8 Arm® A72 CPUs @1.5GHz-2.5GHz
- One channel of DDR4 @3200MT/s

Advanced Hardware Accelerations
- Networking and virtualization accelerations – RDMA, ASAP2, VirtIO , SR-IOV
- Security: Crypto (IPsec, TLS, AES-XTS), Isolation, PKI, SHA-2, Regular Expression & DPI
- Storage Accelerations – NVMe, (De)Compression, Dedup, RAID, CRC64, NMVe SNAP
- Host agnostic network solution

Scalability and Programmability
- For smartNIC and as a controller in system
- User specific application to run over Arm cores
- Integrated control and data planes
- For smartNIC: Security application isolated from main host
Unmatched Application Performance & Efficiency

**NFV**
Network Function Virtualization

**DPDK**
- Data Plane Development Kit
- CPU does the work
- Highest DPDK performance
- Line rate packet forwarding
- Linux project maintainer

**ASAP2**
- Accelerated Switching & Packet Processing
- NIC does the work
- Hardware OVS acceleration
- 8X-10X performance gain
- Zero CPU utilization

Both BlueField & ConnectX
Fully Support both DPDK & ASAP2
Delivering Unmatched Performance and Efficiency
Open vSwitch (OVS) is a powerful open source, multi-layer virtual switch
- Flow lookups and load-balancing across VM’s
- Protocol tunneling (GRE, VXLAN, etc.)
- QoS and traffic policing
- VM state mobility (migration)

However, OVS creates performance problems
- Breaks offload capabilities in some network adapters
- Open vSwitch in hypervisor prohibits SR-IOV between NIC and host VM’s
  - Soft OVS routes all traffic through HV and uses expensive memcopy

http://openvswitch.org
Accelerating OVS with ASAP²

- "Accelerated Switching and Packet Processing" (ASAP²) restores performance
  - Allows NIC to perform programmable switching on flows
  - Preserves control plane and all management interfaces of OVS

- Integrated eSwitch within ConnectX handles L2/L3 flow switching
  - Encaps/Decaps of overlay headers
  - Header re-write (i.e. for NAT)

- ASAP² delivers over **66Mpps** with 64B packets
  - Zero hypervisor CPU utilization

- A good Smart NIC should accelerate OVS!
What is Mellanox PeerDirect™

- Purpose-built for Acceleration of Deep Learning
- Provides significant decrease in communication latency for acceleration devices
- Natively supported by Mellanox OFED
- Supports peer-to-peer communications between Mellanox adapters and third-party devices
- No unnecessary system memory copies & CPU overhead
- Enables GPUDirect™ RDMA, GPUDirect™ ASYNC, ROCm and others
- InfiniBand and RoCE

Designed for Deep Learning Acceleration
10X Higher Performance with GPUDirect™ RDMA Technology

- Purpose-built for Acceleration of Deep Learning
- Lowest communication latency for acceleration devices
- No unnecessary system memory copies and CPU overhead

GPUDirect™ RDMA Diagram

Graphs showing:
- GPU-InfiniBand-GPU Latency
- GPU-Mellanox-GPU Throughput (Uni-Dir)
- GPU-Mellanox-GPU Throughput (Bi-Dir)

Latency (usec) vs Message Size (Bytes)
SDN & Security Vulnerable to Internal Attacks

CPU-BASED SECURITY IS AN INHERENT PROBLEM, AS SOFTWARE BASED SECURITY CONTROLS ARE PLACED IN THE SAME TRUST DOMAIN AS A POTENTIAL ATTACKER

- Standard approach is to run security functions alongside the applications being protected
  - Implies complete trust in the security of the OS and the system hardware
  - Common security domain vulnerable to application level attacks (ex: Meltdown & Spectre)
- Application performance is destroyed by Distributed Denial of Service Attacks

Recent Examples

New “Zero-Day” Attack
Application Acceleration Delivers Data Center Efficiency

X86 SDN/Security Challenges/Vulnerabilities
- No isolation = Vulnerable Infrastructure
  - Application & security domain are identical
- Poor performance
- Poor scalability
- Vulnerable to DDOS attacks

Intellectual Networking

RoCE • InfiniBand RDMA

25, 50, 100 Gb/s
BlueField SmartNIC Brings Clouds to a New Level

**SmartNIC Benefits**
- Flexible Workload Acceleration
- SDN & security isolation & offload
- Scalable, distributed security
- Logically separated security domains
- Secure boot & firmware update
- Application control & visibility
- Fully programmable and extensible

**Application Acceleration Delivers Infrastructure Efficiency**

**Storage**

**Machine Learning**

**Big Data**

**Efficient Data Transport**

**SDN/Security**

- 25, 50, 100 Gb/s
- X86 SW Processing
Network Acceleration Use-Case

- Standard server environment
  - Intelligent flow-based routing/switching
  - Virtualization and provisioning
  - Crypto acceleration,
  - Offload other functions (DPI, Firewall...)

- Performance, Efficiency, Isolation
  - Offload x86 CPU
  - Accelerate small packet processing
  - Isolate networking mgmt

Wire-Speed networking / security services

10 / 25 / 50 Gb/s Interfaces
SmartNIC & Fabric Virtualizes Cloud Storage

What if you could have all the cloud storage appear as if it was in your server as local storage?

What if every server could have that?

- Very few apps are SAN aware
  - Databases & Distributed File Systems
  - NVMe-over-Fabrics: ideal for these

- Most apps use local storage
  - Results in poor resource utilization
  - Overprovisioning is expensive!
SmartNIC & Fabric Virtualizes Cloud Storage

- Very few apps are SAN aware
  - Databases & Distributed File Systems
  - NVMe-over-Fabrics: ideal for these

- Most apps use local storage
  - Results in poor resource utilization
  - Overprovisioning is expensive!

- BlueField SmartNIC Virtualization
  - Makes all cloud storage resources local
  - Better utilization & efficiency
  - No software disruption
Smart Cloud Architecture Virtualizes all Resources

- Each app accesses all cloud resources as if local
  - Relies on smart acceleration and virtualization capabilities

- Storage Virtualization
  - Massive “direct attached storage”
  - Every server sees entire cloud storage as local
    - vNVME is virtualized NVMe

- Every server can have 1 Petabyte of Storage
  - Smart Cloud Storage Resource Abstraction
  - Smart Virtualization abstracts away storage fabric
    - Provisioning
    - Network Addressing
    - Error correction
    - Thin provisioning
    - Allows all apps to benefit from storage fabric scalability

- Ultimately Virtualize all Cloud Resources
  - GPUs and Artificial Intelligence accelerators
  - Persistent memory
Bringing Hyperscale Efficiency to Private Cloud

Broad Hardware and Software Partnerships
Open Source Leadership

- Mellanox #5 corporate contributor to Linux 4.8 kernel
- https://lwn.net/Articles/701650/
- Multiple projects
  - SwitchDev – Open source kernel level switch interface
  - Security –
  - RDMA – RoCE the standard for
  - Tensorflow
  - Hadoop
  - Ceph
  - NVME-oF
  - Security - TLS offload
The Smart & Integrated Solution for the Cloud

- Machine Learning
- Acceleration
- Storage
- Security
- Big Data
- Offload
Thank You