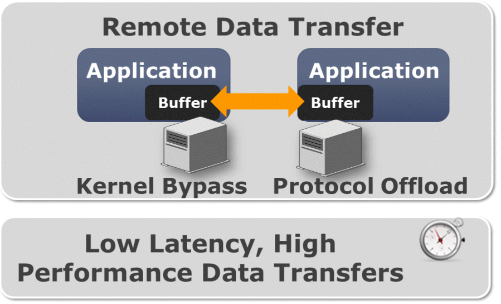
# OCP: Rack Level Optimization in a Post Moore Era

OCP Summit, October 2014

Kevin Deierling, Mellanox Technologies – kevind at mellanox.com

Mellanox®
TECHNOLOGIES
Connect. Accelerate. Outperform.™

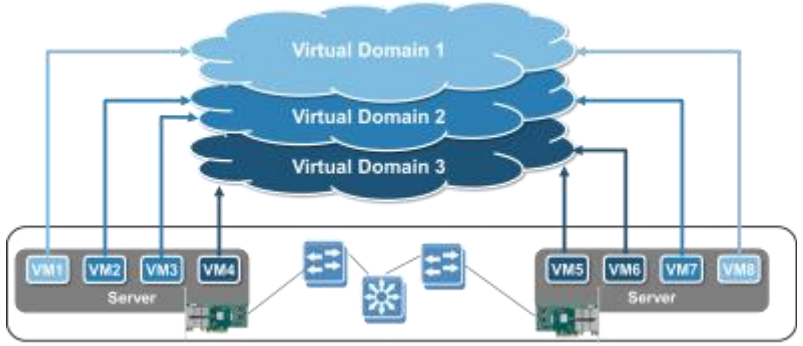
# End to End Provider of Smart, Fast, Open Interconnects



*High Performance, Efficient Virtual Networks*

OPEN Compute Project

*Open Platform*



*SDN & Network Virtualization*

## End-to-End InfiniBand and Ethernet Portfolio

ICs

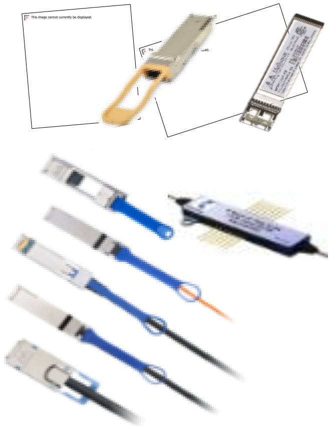
Adapter Cards

Switches/Gateways

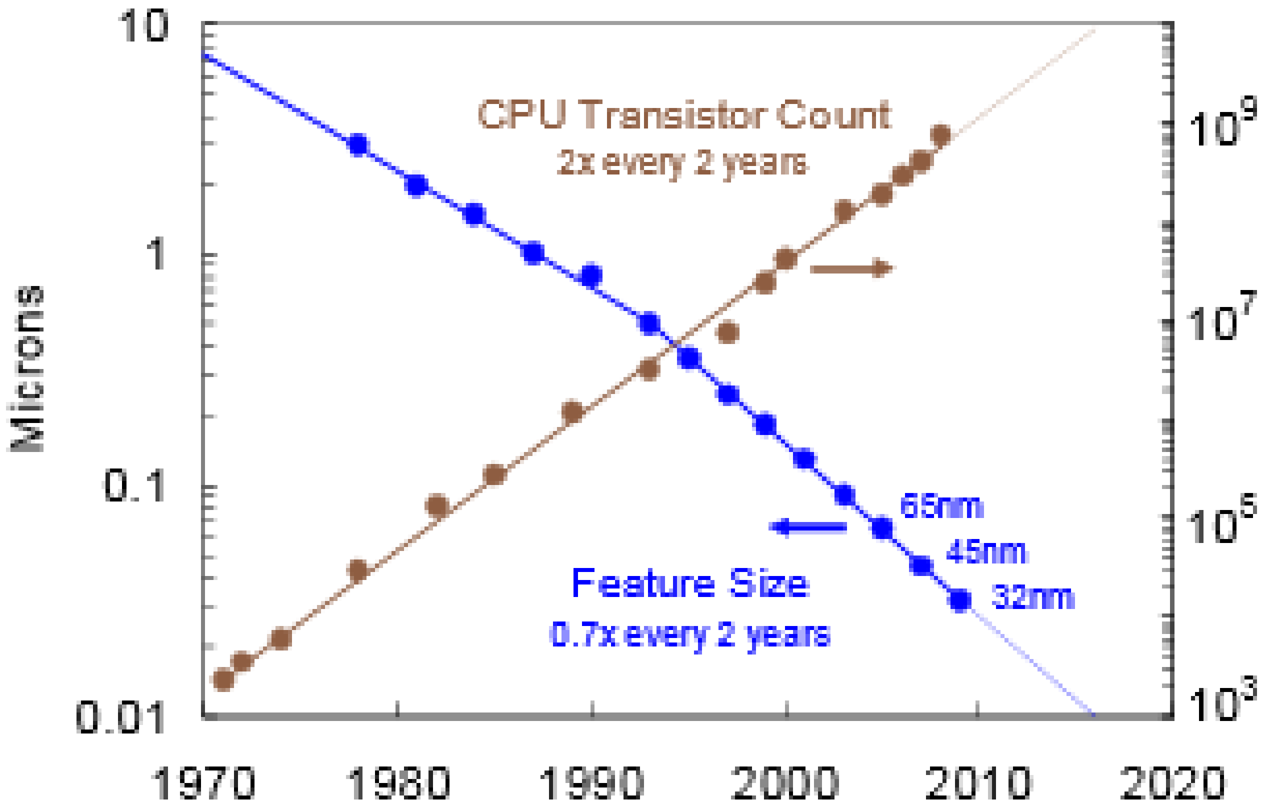
Host/Fabric Software

Metro / WAN

Cables/Modules

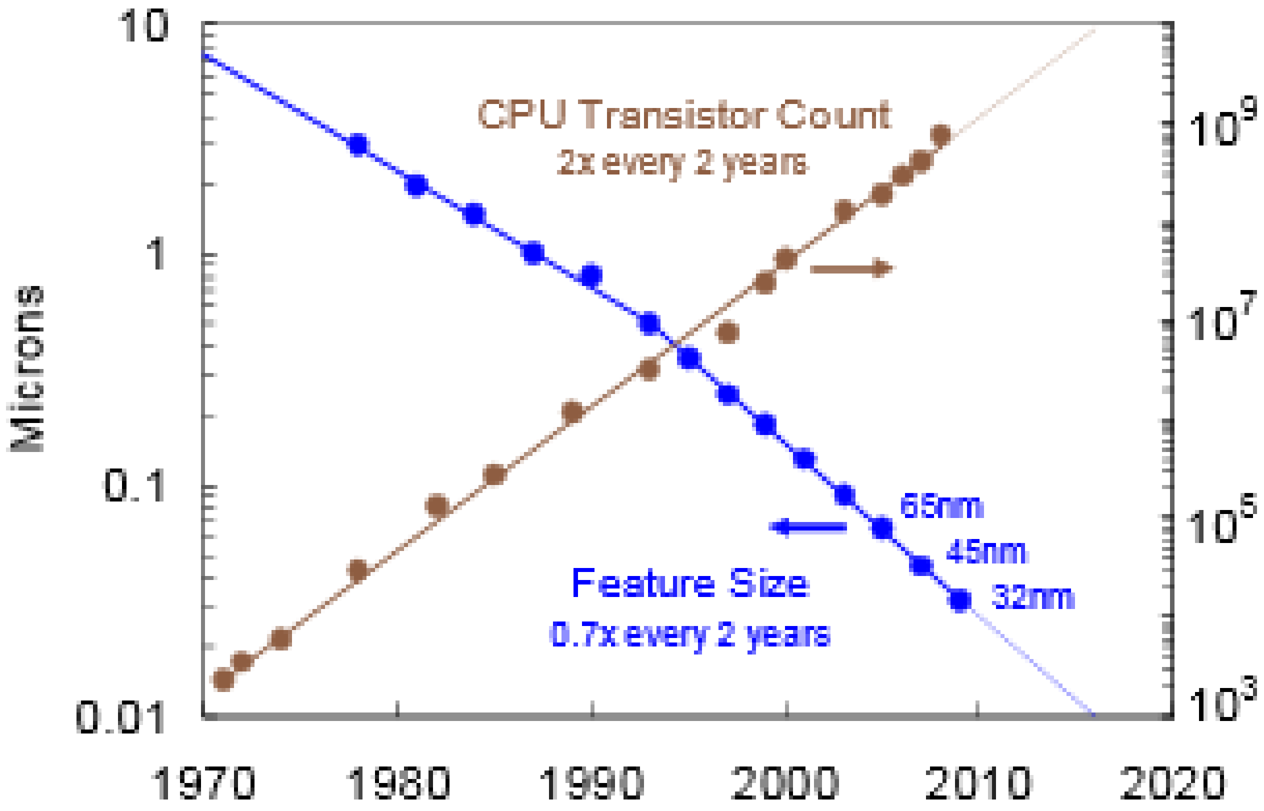


# Moore’s Law: Driving Computer Architectures for 45 Years

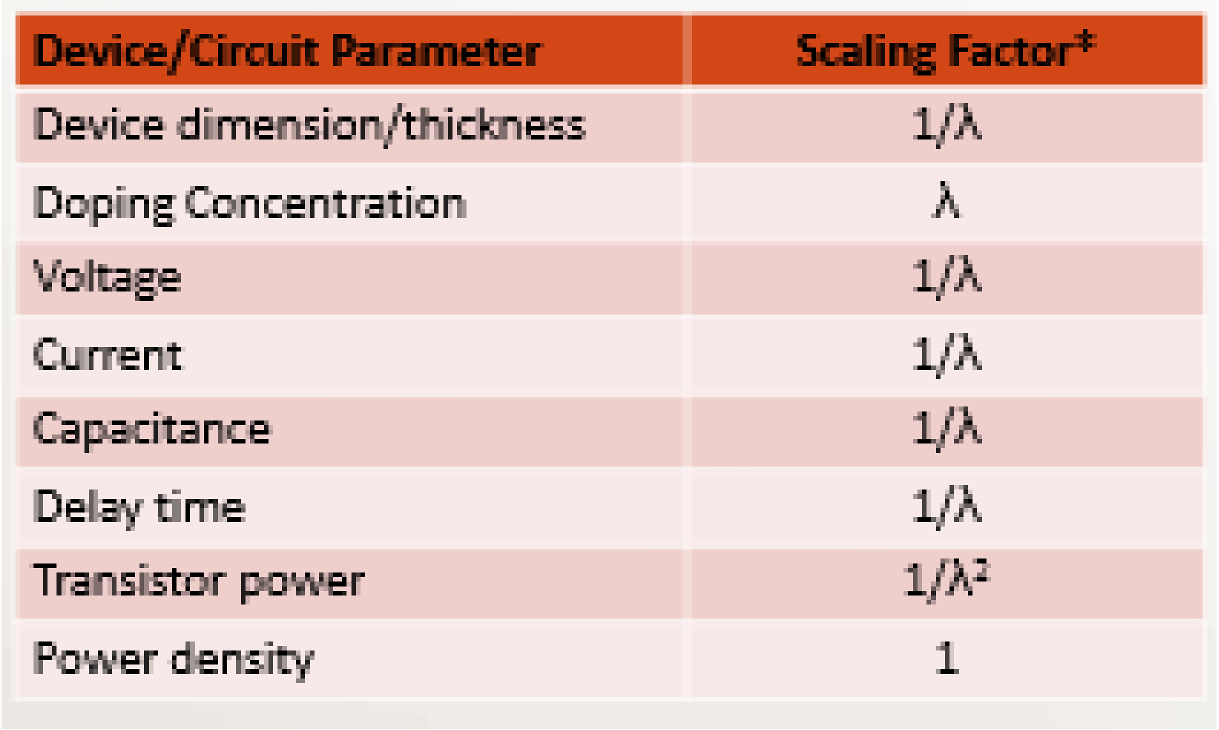


- Moore’s Law: Chip transistor count doubles roughly every two years since 1970
  - Linear shrink of 30% results in half the area
  - Keep the cost/area about constant while shrinking

# Moore’s Law & Dennard Scaling



No compromises: Everything gets better as transistors shrink



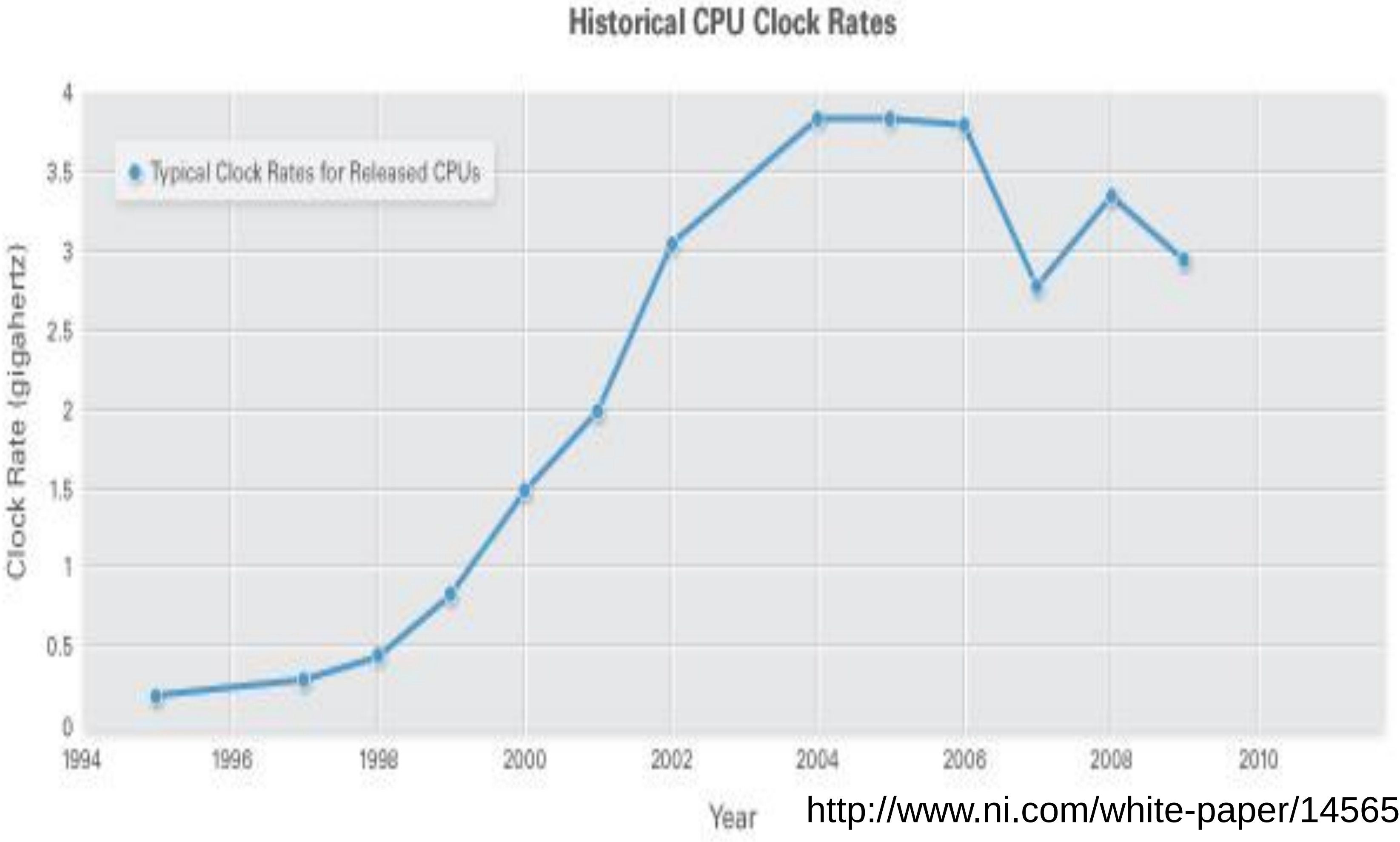
| Device/Circuit Parameter | Scaling Factor* |
|---|---|
| Device dimension/thickness | 1/λ |
| Doping Concentration | λ |
| Voltage | 1/λ |
| Current | 1/λ |
| Capacitance | 1/λ |
| Delay time | 1/λ |
| Transistor power | $1/\lambda^2$ |
| Power density | 1 |

Henry Dennard, IEEE JSSC Oct 1974

- Moore’s Law: Chip transistor count doubles roughly every two years
  - Linear shrink of 30% results in half the area
  - Keep the cost/area about constant while shrinking

- Dennard Scaling: As a MOSFET transistor shrinks it gets:
  - Faster
  - Lower power (constant power density)
  - Smaller/lighter

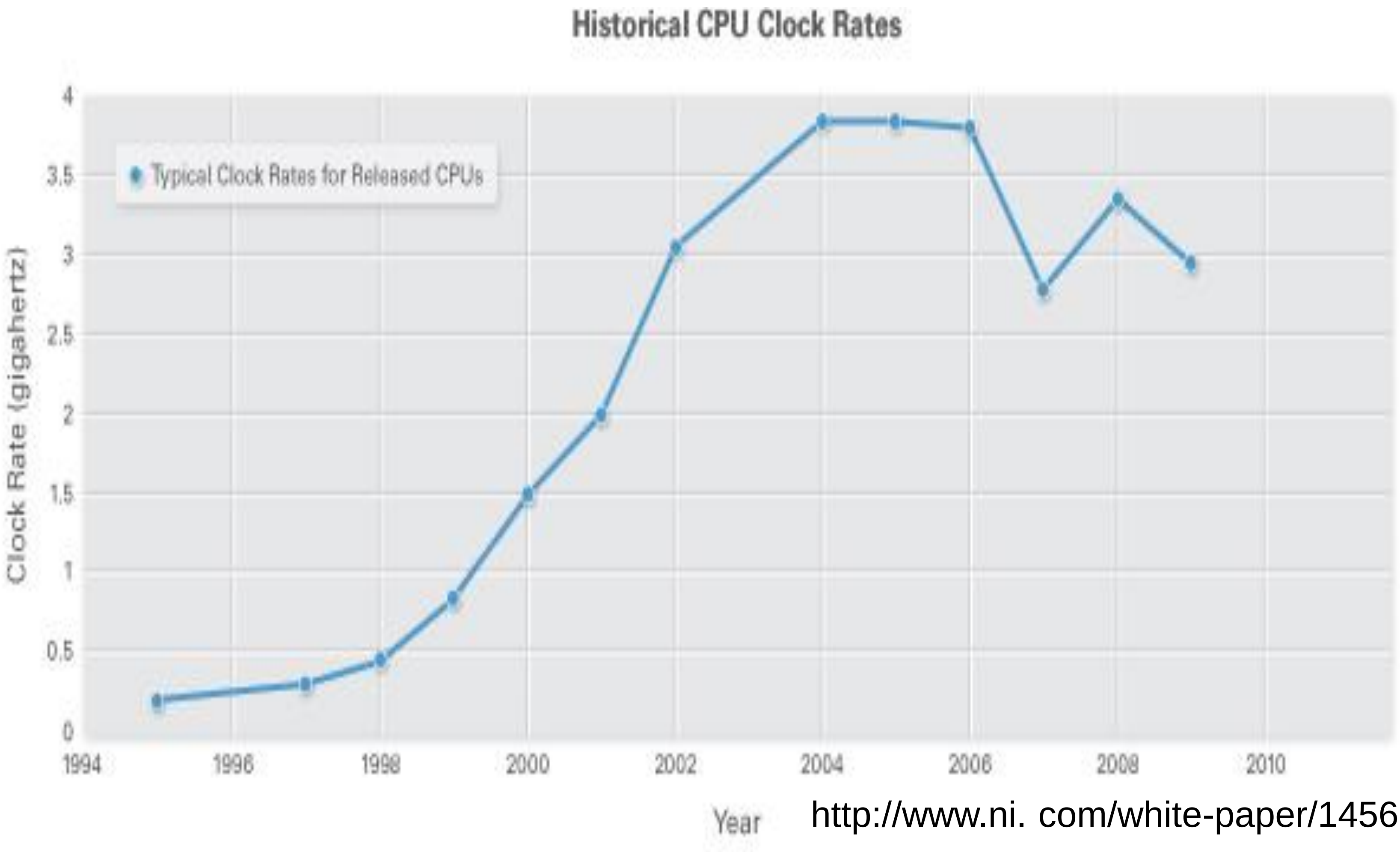
# Now for the Bad News …



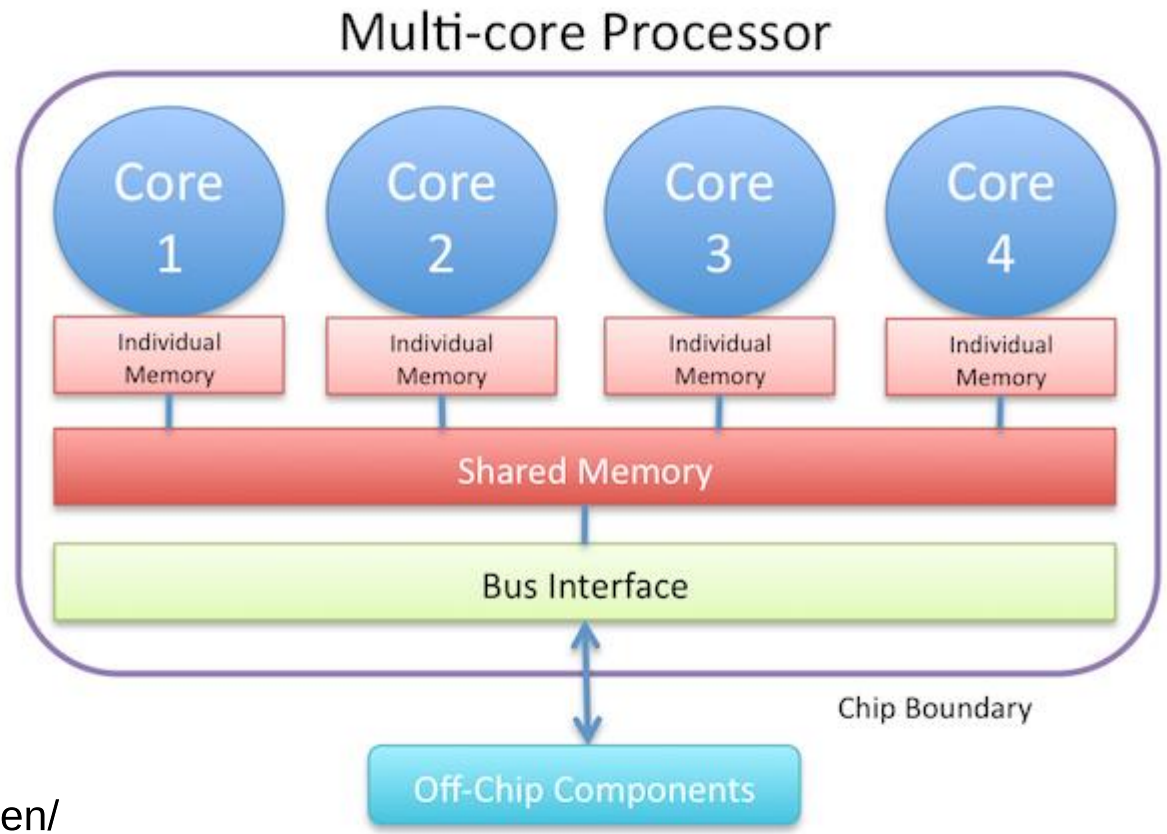
http://www.ni.com/white-paper/14565/en/

- Dennard scaling broke about a decade ago
  - Both power density & performance stopped scaling
- Higher power & new process/fabs = higher costs
- Economic half of Moore’s law has crumbled too

# Only Half of Moore's Law Still Going - Performance thru Parallelism



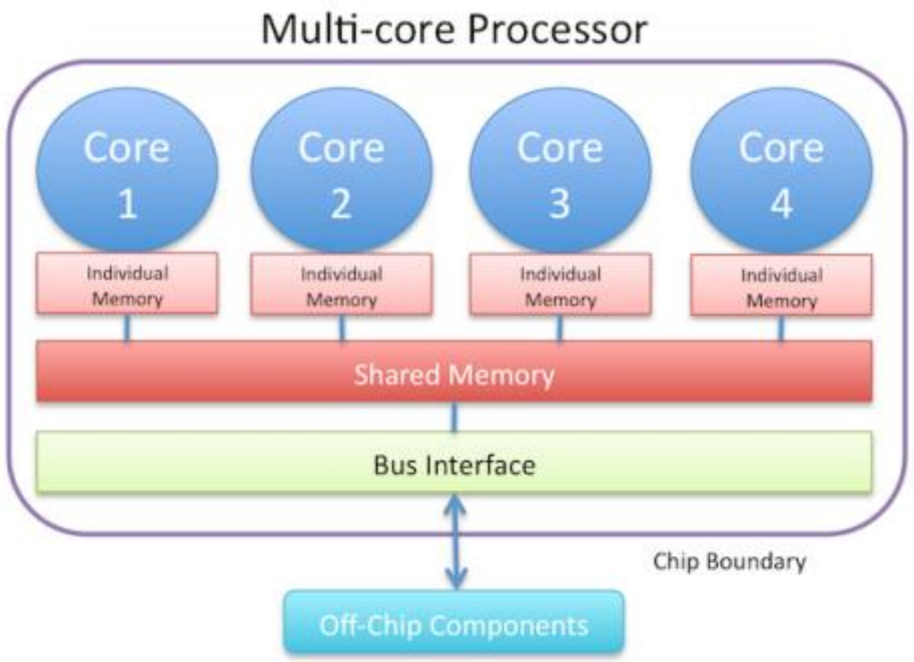
http://www.ni. com/white-paper/14565/en/

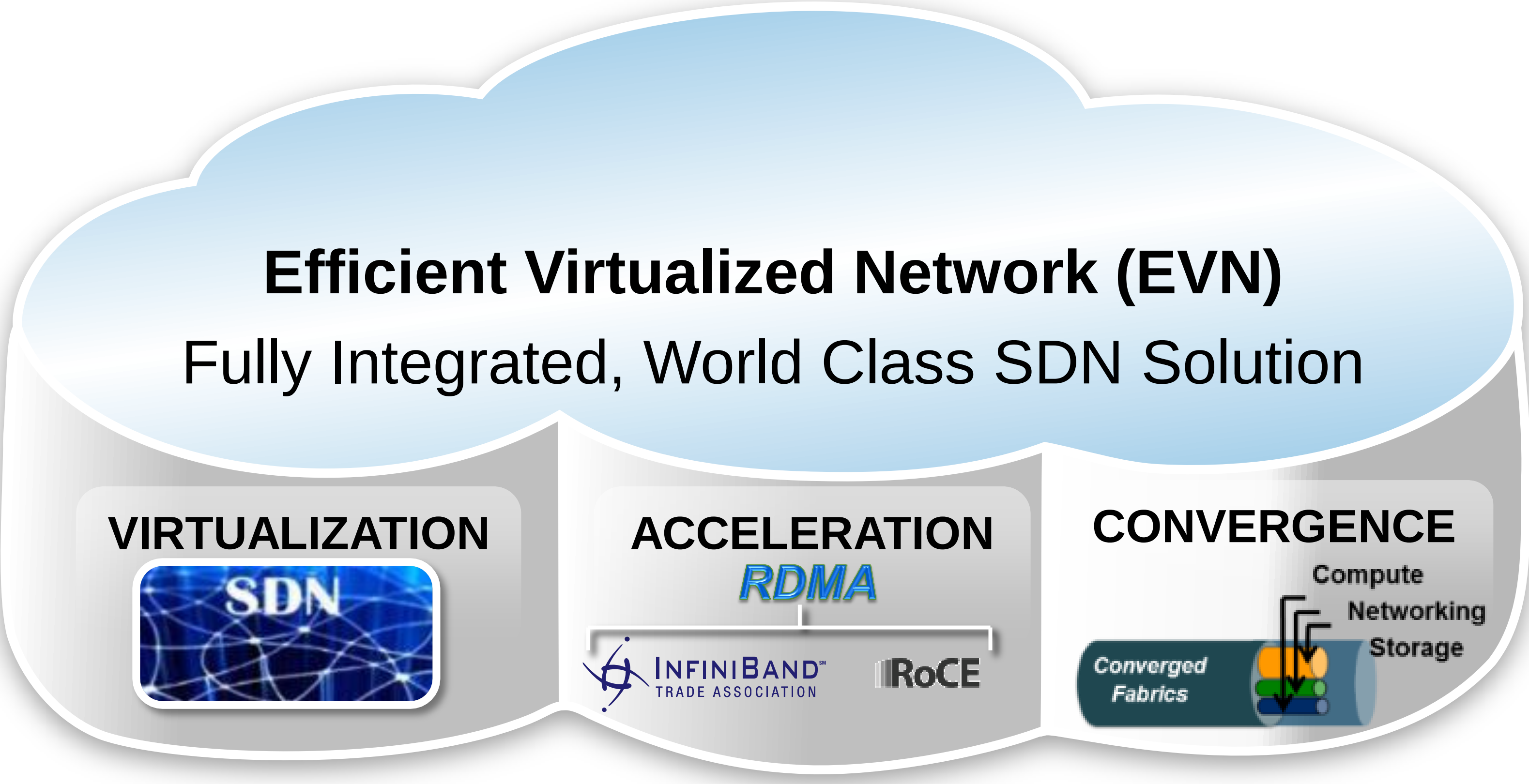


- Dennard scaling broke about a decade ago
  - Both power density & performance stopped scaling
- Higher power & new process/fabs = higher costs
- Economic half of Moore's law has crumbled too

- So with only half of Moore's law intact what is a multi-Billion dollar Fab to do ?
- Not faster cores but more and more of them …

# More Cores Requires More Network!




Efficient Virtualized Network (EVN)
Fully Integrated, World Class SDN Solution
VIRTUALIZATION
SDN
ACCELERATION
RDMA
INFINIBAND TRADE ASSOCIATION
RoCE
CONVERGENCE
Compute
Networking
Storage
Converged Fabrics

Multicore needs higher bandwidth … but also:

- Support for many cores: Virtualization
- Efficient transport, low overhead data movement
- Converged: Compute, networking, storage

Efficient Virtualized Network (EVN)

- Network Virtualization
- RDMA
- Convergence

# Open Compute Project: Efficient Platform in a Post Moore Era

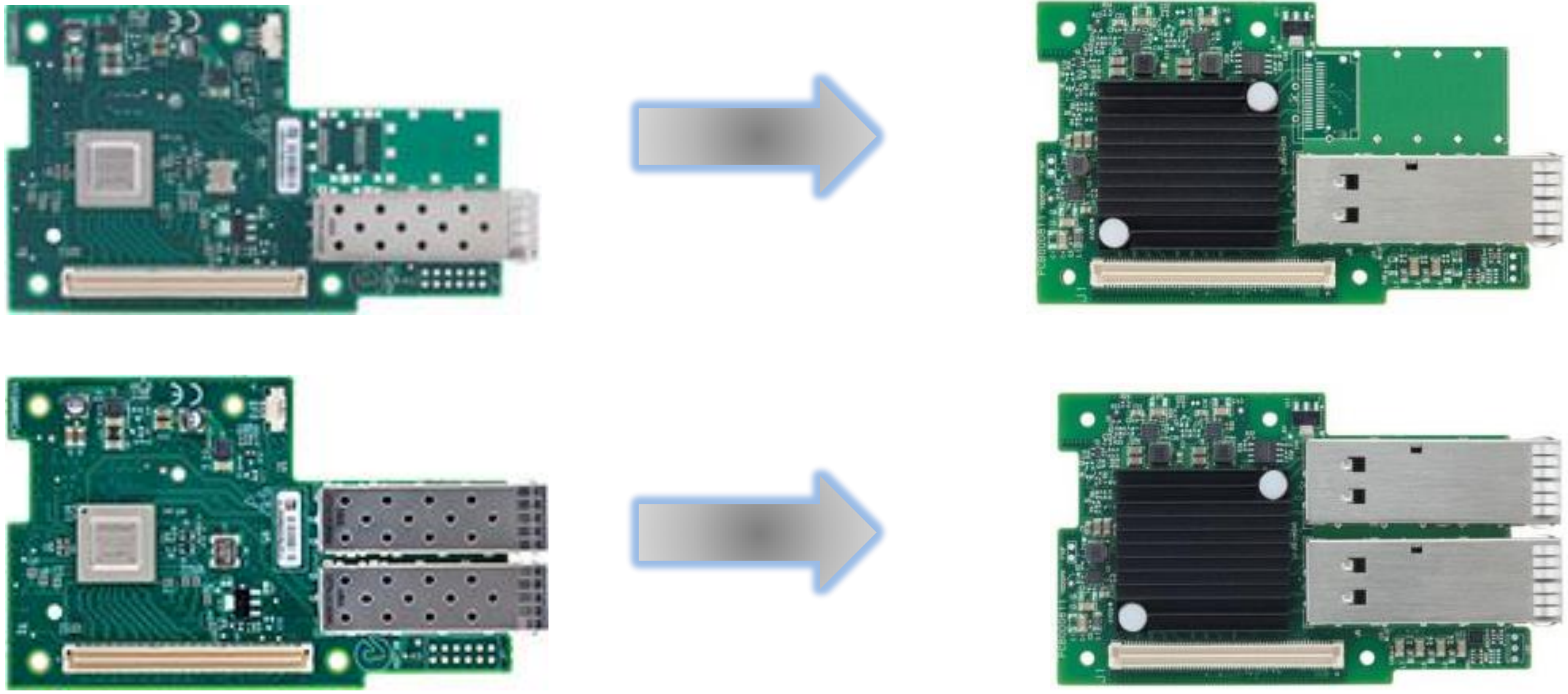
- Rack Level Optimization
  - Tightly integration of high performance servers and networking
  - Shared power, cooling, and rack resources
  - Open platform drives high volumes
- High Performance Networking
  - Efficient rack requires an efficient, high performance network
  - Single and dual port 10GbE and 40GbE OCP Adapters

# OCP Ethernet Mezzanine Cards for 10 & 40GbE

OPEN
Compute Project

10GbE SFP+

40GbE QSFP



Single Port

Dual Port

OCP Mezz 1.0

OCP Mezz 2.0

# 40GbE OCP 2.0 Adapters are Here!

## Single / Dual Port OCP Adapter

- PCIe Gen3 x8
- 10Gb/s / 40Gb/s / 56Gb/s data rates support
- <1usec latency

## Industry Leading Performance

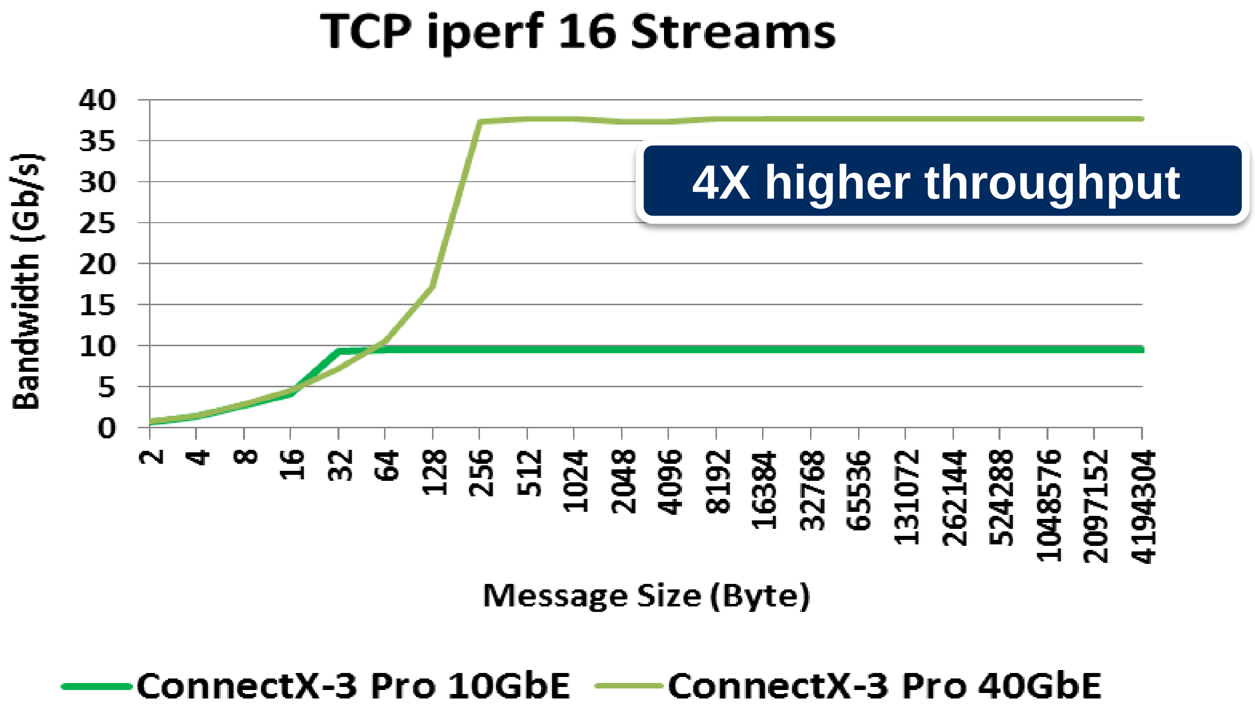
- Up to 4 times higher message rate & throughput than 10GbE

## Advanced Features

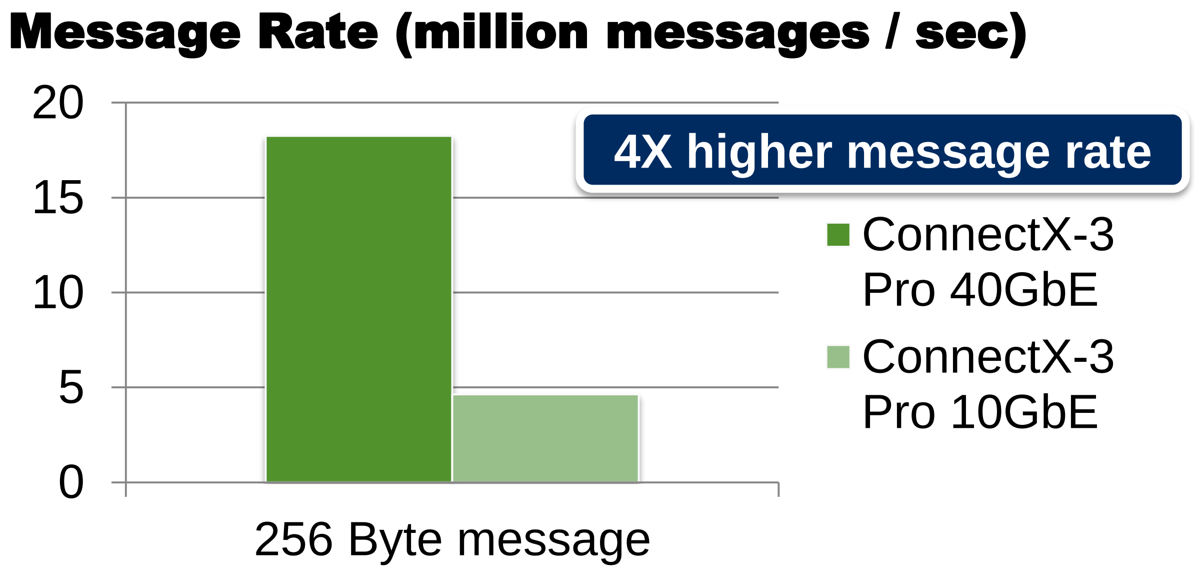
- RoCE (RDMA over Converged Ethernet)
- SR-IOV support with embedded switch
- NVGRE / VXLAN Stateless Offloads

## Host Management

- Baseband Management Controller Interface
  - IPMI over SMBus
  - NC-SI over RMII
- PXE and UEFI
- IPv6 Support



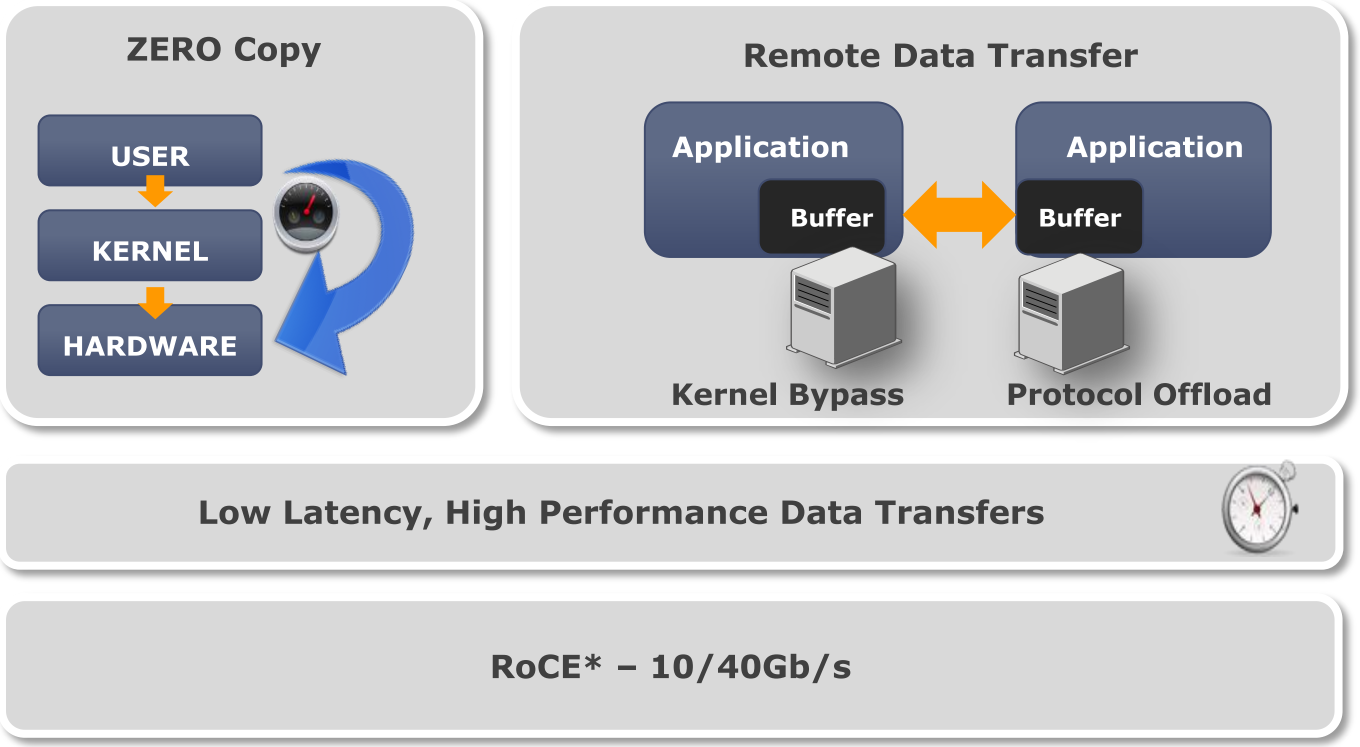
4X higher throughput



4X higher message rate

# RDMA over Converged Ethernet (RoCE)

- Highest performance in the industry
  - Latency of ~1us
  - Reliable/unreliable, connected/datagram
  - Unicast and multicast
- Fast storage access
  - RoCE iSER block storage
  - New file and object storage is RoCE enabled
  - RoCE v2 enables routability
- RoCE for virtualized environments
  - RoCE for SR-IOV connected VMs
  - Near 40GbE throughput in virtualized environments


**ZERO Copy**

**USER**

**KERNEL**

**HARDWARE**

**Remote Data Transfer**

**Application**

**Buffer**

**Application**

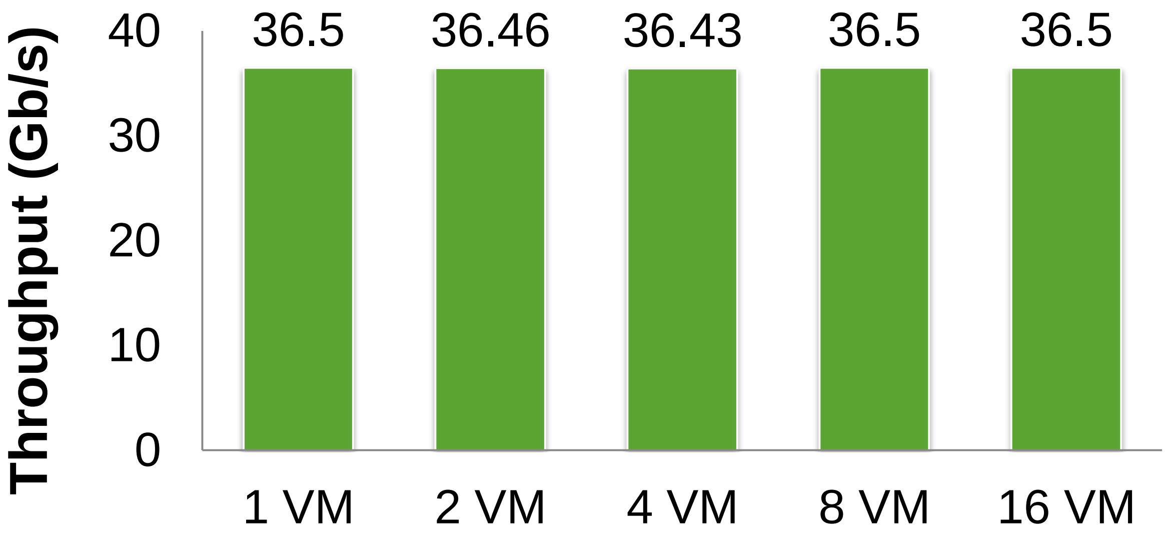
**Buffer**

**Kernel Bypass**

**Protocol Offload**

**Low Latency, High Performance Data Transfers**

**RoCE* – 10/40Gb/s**



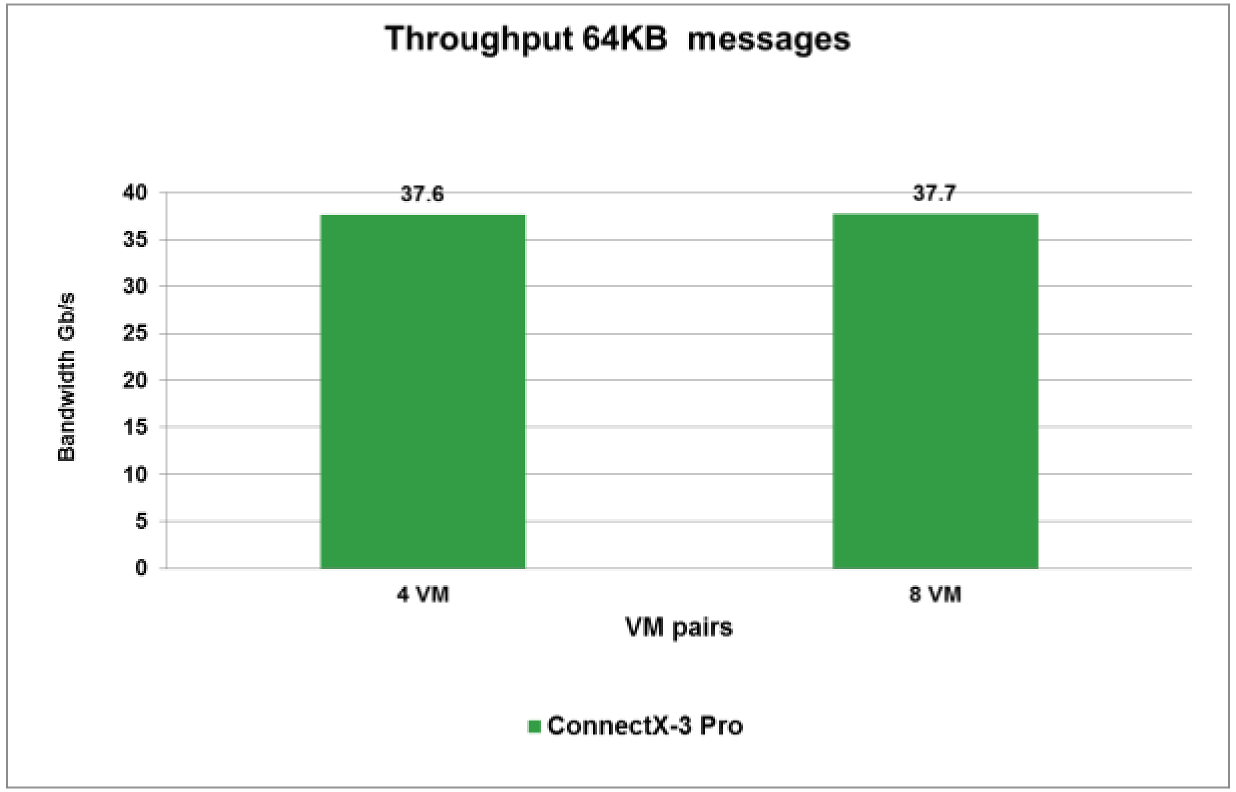
# Leading Solution for Virtualized Environments

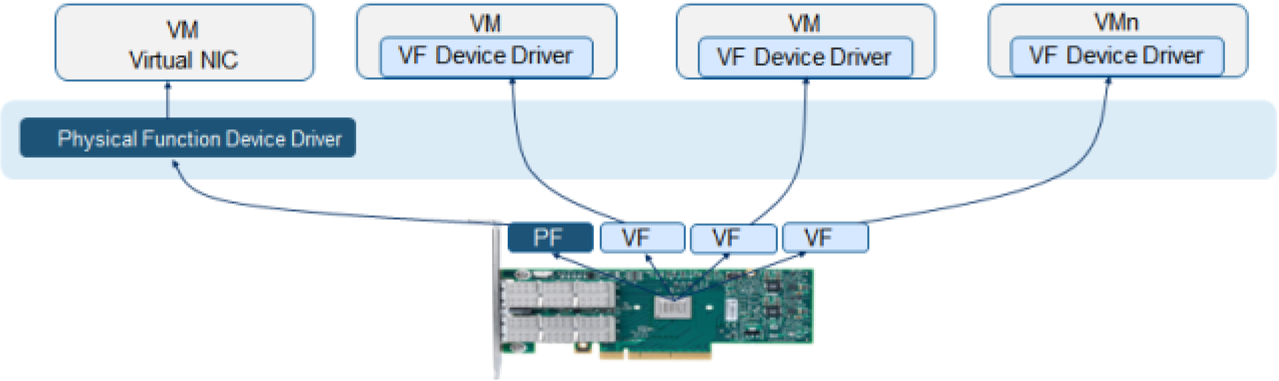
- Single Root I/O Virtualization (SR-IOV) support

- Highest throughput
  - Enabling more VMs on single machine
  - Highest traffic rate for each VM

- Overlay networks offloads
  - NVGRE and VXLAN
  - Breaking the 10GbE throughput barrier

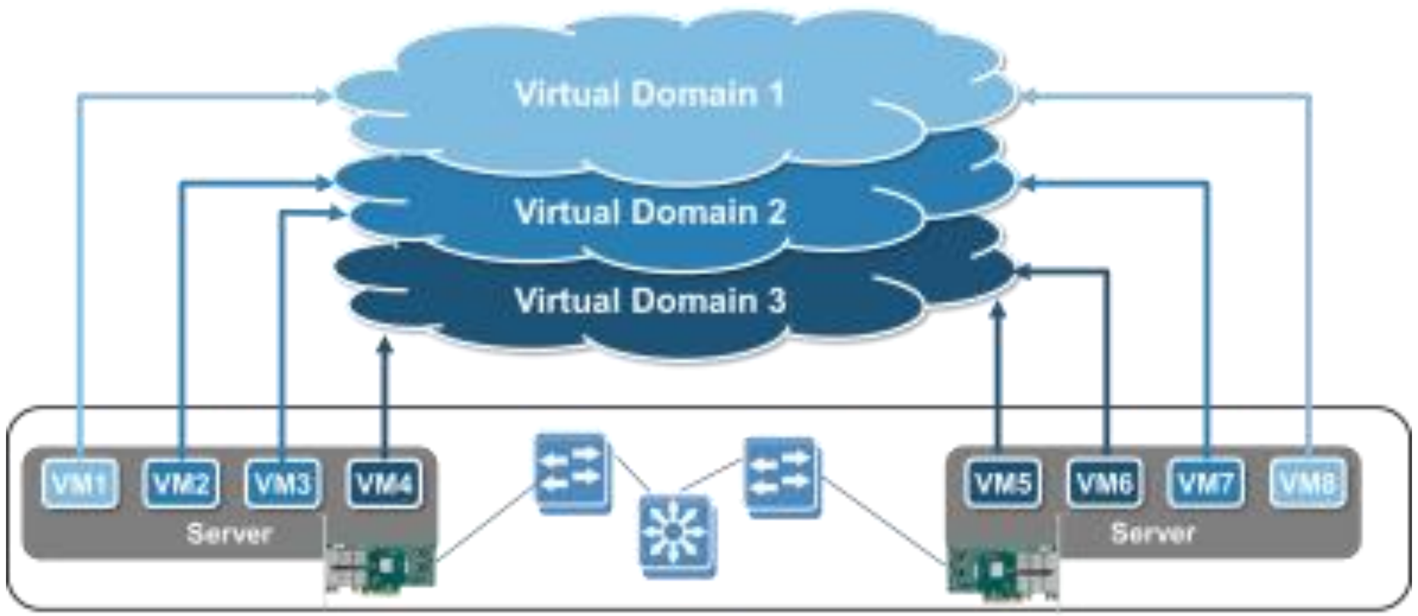
- CPU offloads
  - SR-IOV enables application-direct access
  - Improving CPU utilization



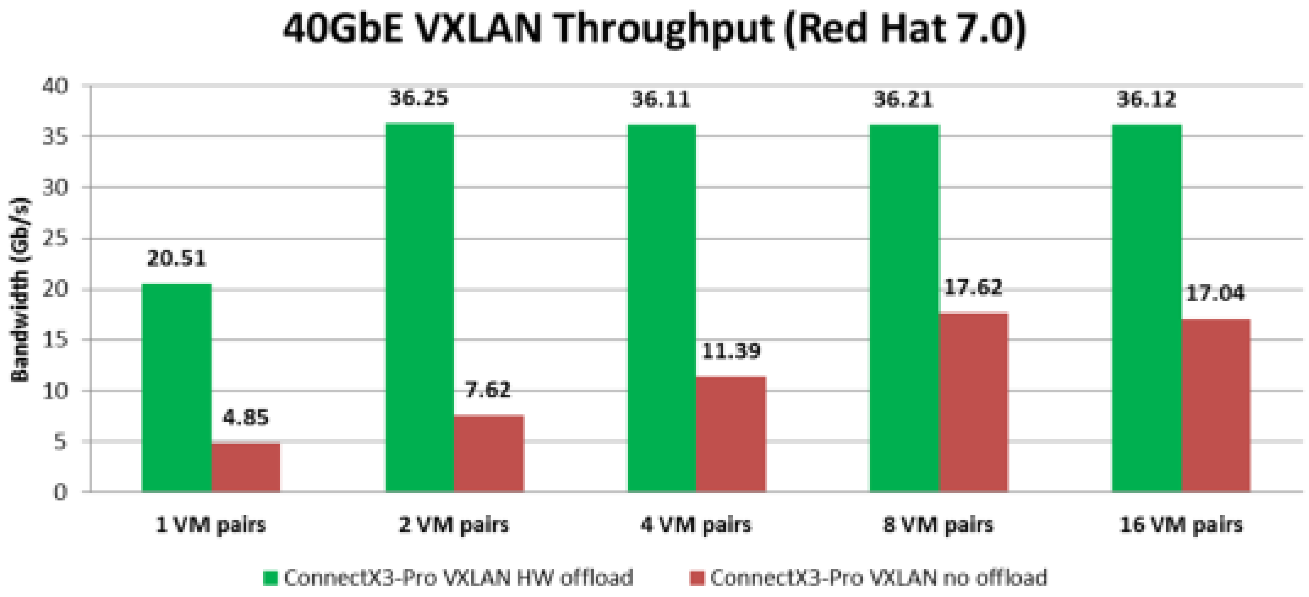


# OCP with Offload Engines for Overlay Network Protocols

- Overlay Network Acceleration
  - Unmatched throughput for virtualized environments
  - VXLAN and NVGRE supported
- Powerful Overlay Acceleration Engines
  - Checksums, LSO, Flow ID calculation,
  - VLAN encapsulation
  - Advanced steering mechanisms: RSS, VMQ
- Acceleration Increases Application Performance
  - Near line rate throughtput
  - 75% improvement in CPU usage



*SDN & Overlay Network Virtualization Acceleration*



# OCP Cards – Product Family

Open Compute Project

| | Ethernet | |
|---|---|---|
| | MCX341A-X (single port)<br>MCX342A-X (dual port) | MCX345A-B (single port)<br>MCX346A-B (dual port) |
| **Connector and Port Speed** | SFP+<br>10GbE | QSFP<br>40/56GbE |
| **PCIe 3.0 Speed** | 8.0GT/s (52Gb/s) | |
| **Features** | ConnectX-3 and ConnectX-3 Pro Flavors | |
| **OS Support** | RHEL, CentOS, SLES, OEL, Windows, ESX/vSphere, Ubuntu, Citrix, Fedora | |

# OCP Boards

| Silicon | Port Speed (cage) | Host Management | | | OPN |
|---|---|---|---|---|---|
| | | IPMI | NC-SI | LACP | |
| ConnectX-3 | 10GbE (SFP+)<br>MCX341A / MCX342A | – | – | – | -XCCN |
| | | √ | – | – | -XCDN (IPv4) |
| | | √ | – | – | -XCEN (IPv6) |
| | | √ | – | √ | -XCFN |
| | | √ | √ | – | -XCGN |
| ConnectX-3 Pro | 10GbE (SFP+)<br>MCX341A / MCX342A | – | – | – | -XCPN |
| | | √ | √ | – | -XCQN |
| ConnectX-3 Pro | 40GbE (QSFP)<br>MCX345A / MCX346A | – | – | – | -BCPN |
| | | √ | √ | – | -BCQN |

# Summary

- Moore’s Law is Only Half Alive
  - Dennard Scaling has cracked
  - Moore’s Law will break. Not because of the laws of physics but rather the laws of economics
- So Scaling will be at the Rack & Data Center Level
  - Drives requirement for high performance Efficient Virtual Networks
    - Network Virtualization
    - RDMA: Low latency data movement
    - Convergence: Compute, Networking, Storage
  - RoCE Ready Racks
- OCP Platform is About Rack Level Optimization
  - Open! Really!
  - OCP 2.0 is here with 40Gb/s Adapters
- Efficiency and performance for HyperScale and Enterprise Workloads

# Thank You!

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™