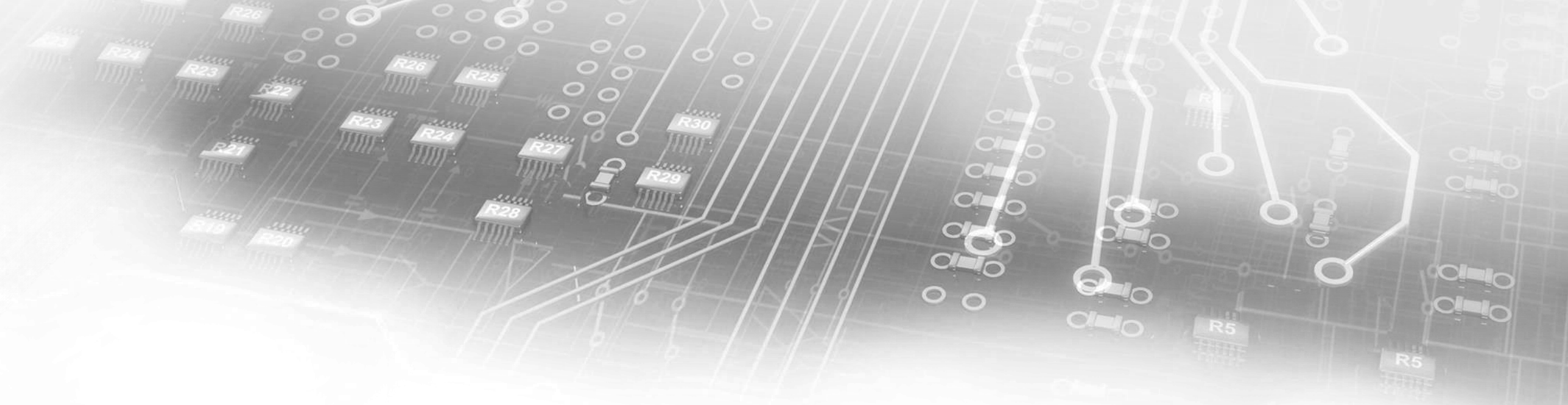


Accelerating the Data Center

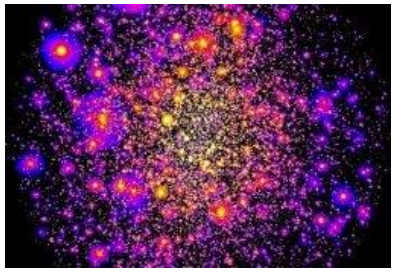
Multicore World 2019

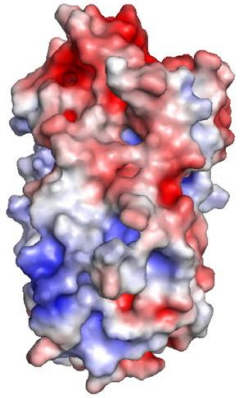
Karen Schramm

VP Technology, Compute and Connectivity Division (CCX)

Processing Demand in the Data Center

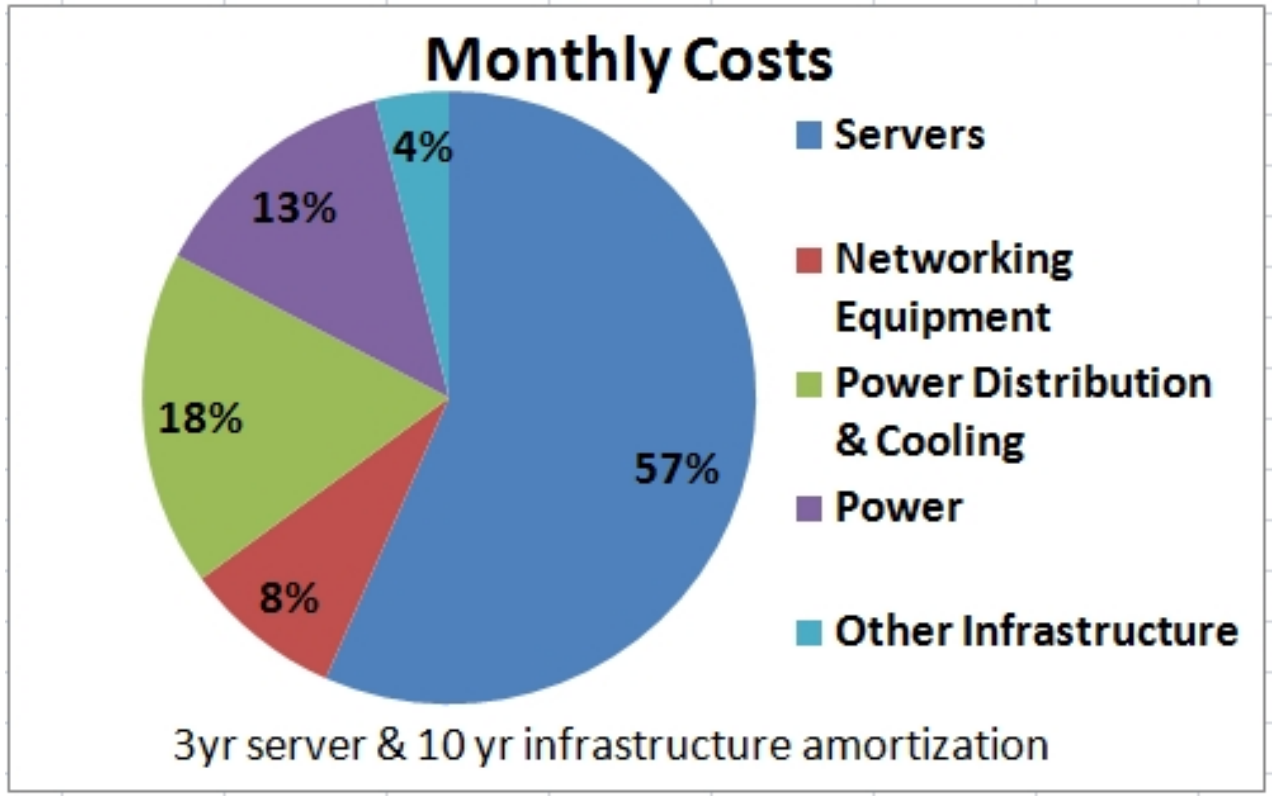
- **Cloud computing continues to grow rapidly**
  - **7M** new servers will be deployed in cloud DCs 2019
  - **7M** additional will be deployed in Enterprise DCs
- **Demand for processing power is increasing**





BROADCOM

Hyper-Scaling Data Center (HSDC) Performance



source: James Hamilton, AWS

- **Adding more CPUs challenging**
  - HSDC's power constrained
  - Servers dominate HSDC costs
- **Servers expensive**
  - High performance CPUs
  - Memory: DRAM & HDD/SSD
  - Mechanical systems and power

BROADCOM

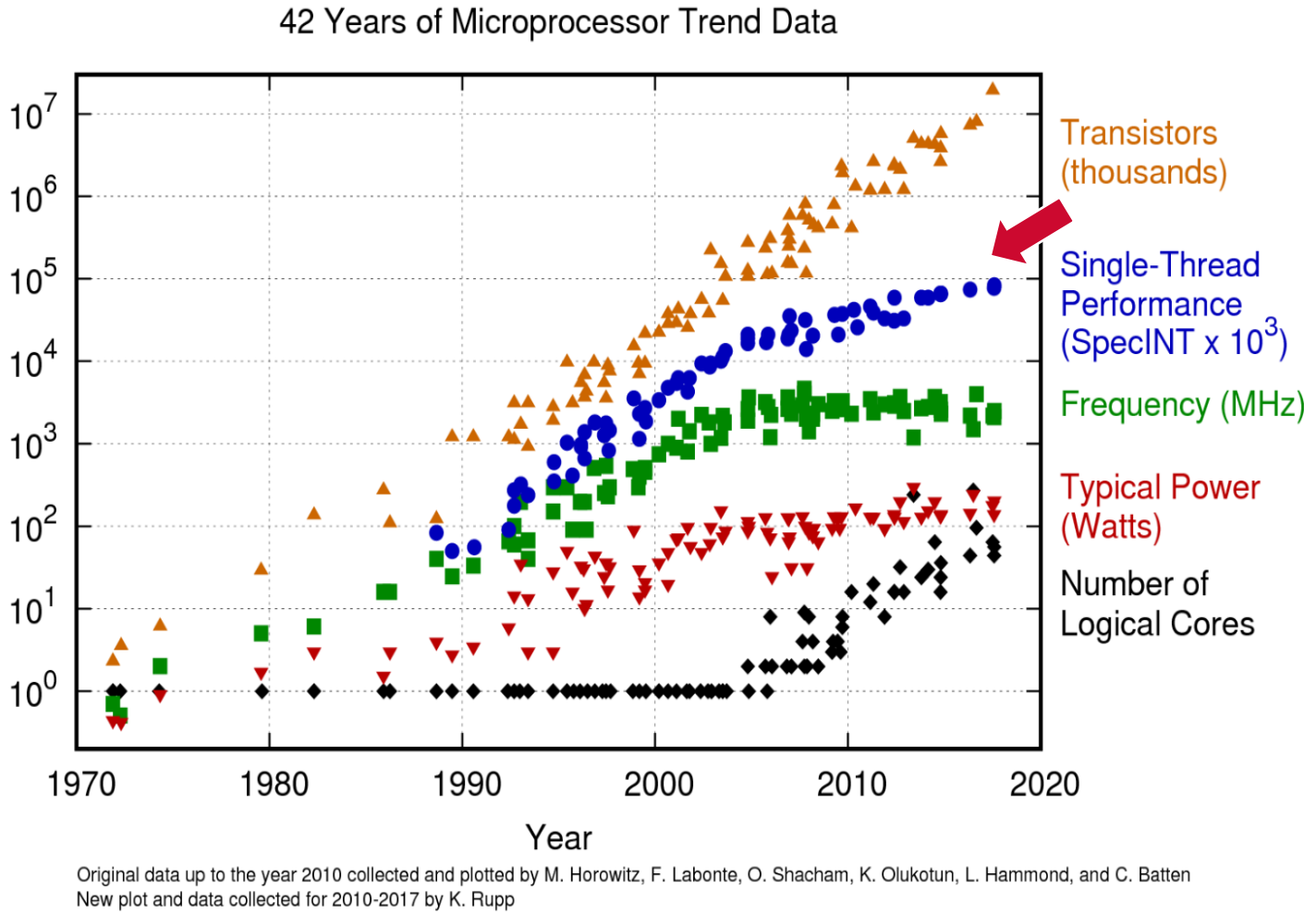
Compute Server Processing Power

- **Moore's law slowing down**

  As semiconductors move to 10nm, 7nm, 5nm …

  - ➢ new process steps taking longer, costing more

- **Single core performance slowing for years**
  - – Clock rate increases → slowed
  - – Performance advances based on
    - ➢ adding more cores
    - ➢ specialized instructions
    - ➢ other architectural improvements



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

BROADCOM

Alternative Computing Platforms for DC Acceleration

GPUs

Orders of magnitude increase in performance for key applications → higher cost & power but better TCO

FPGAs

More performance than software for evolving applications → higher cost & power with flexibility

SmartNICs

Increased performance for network and storage applications → lower cost & power with flexibility

Make Xeon server more efficient at running applications

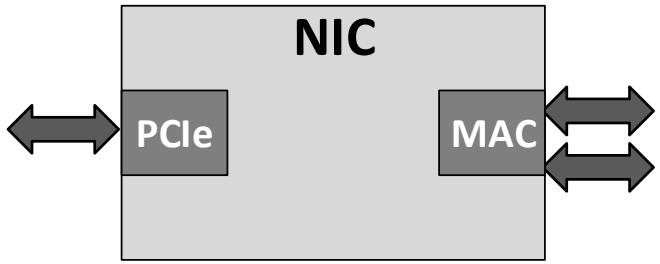
- Offload non-application processing from server CPU
- Lower data access time for applications
- Take network traffic off UPI to improve application efficiency

BROADCOM

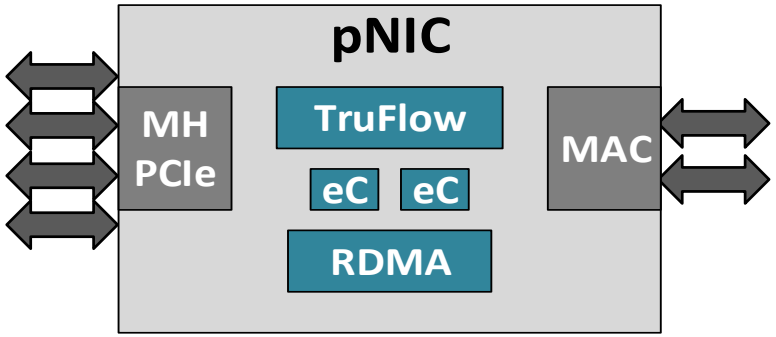
What's a SmartNIC?

NICs are evolving to meet Data Center demands

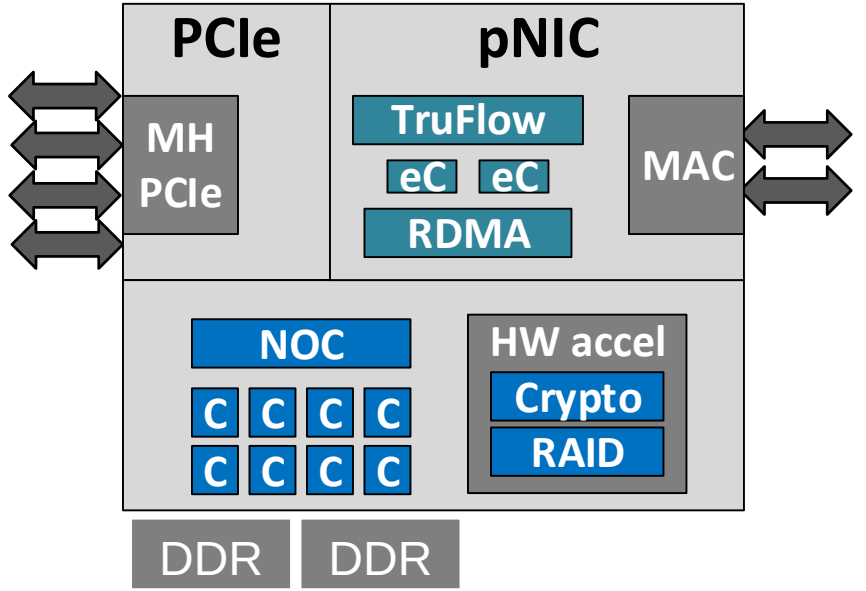
- **Traditional NIC**
  - PCIe to Ethernet
  - Stateless offloads: checksums, RSS, LSO



- **Performance NIC**
  - Offloads for complex, stateful features → e.g. RDMA, packet processing
  - Embedded CPUs assist complex features



- **SmartNIC**
  - Multicore processor & DDR
  - Integrated pNIC with advanced PCIe
  - Hardware acceleration → crypto, RAID, data integrity …



BROADCOM

SmartNIC Opportunity

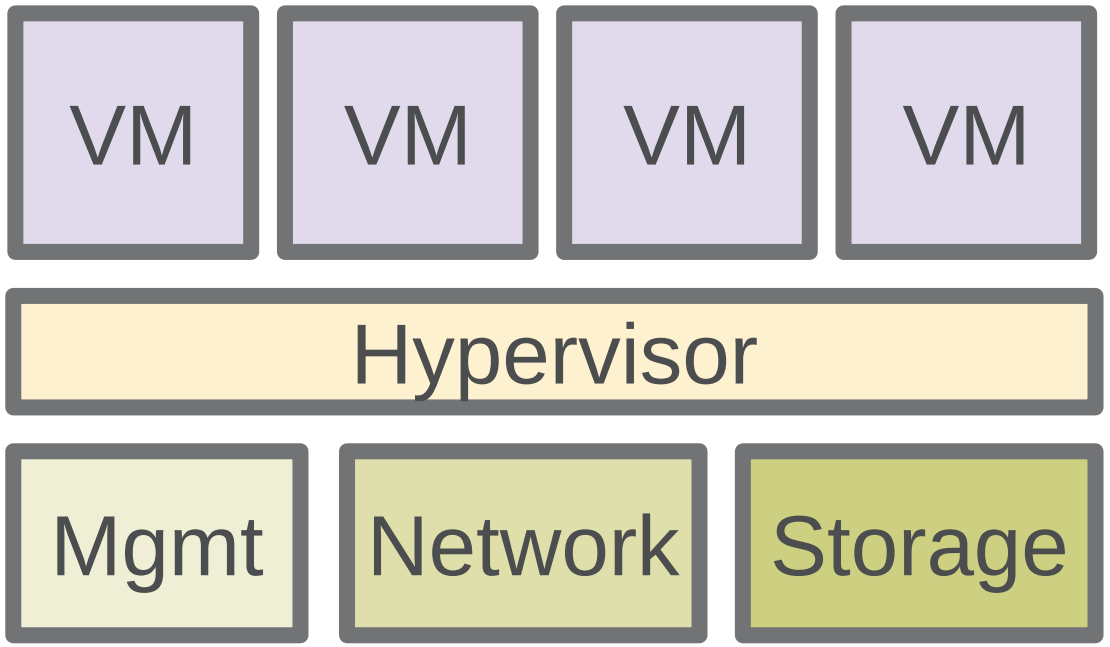
Make DC compute server more efficient at running applications

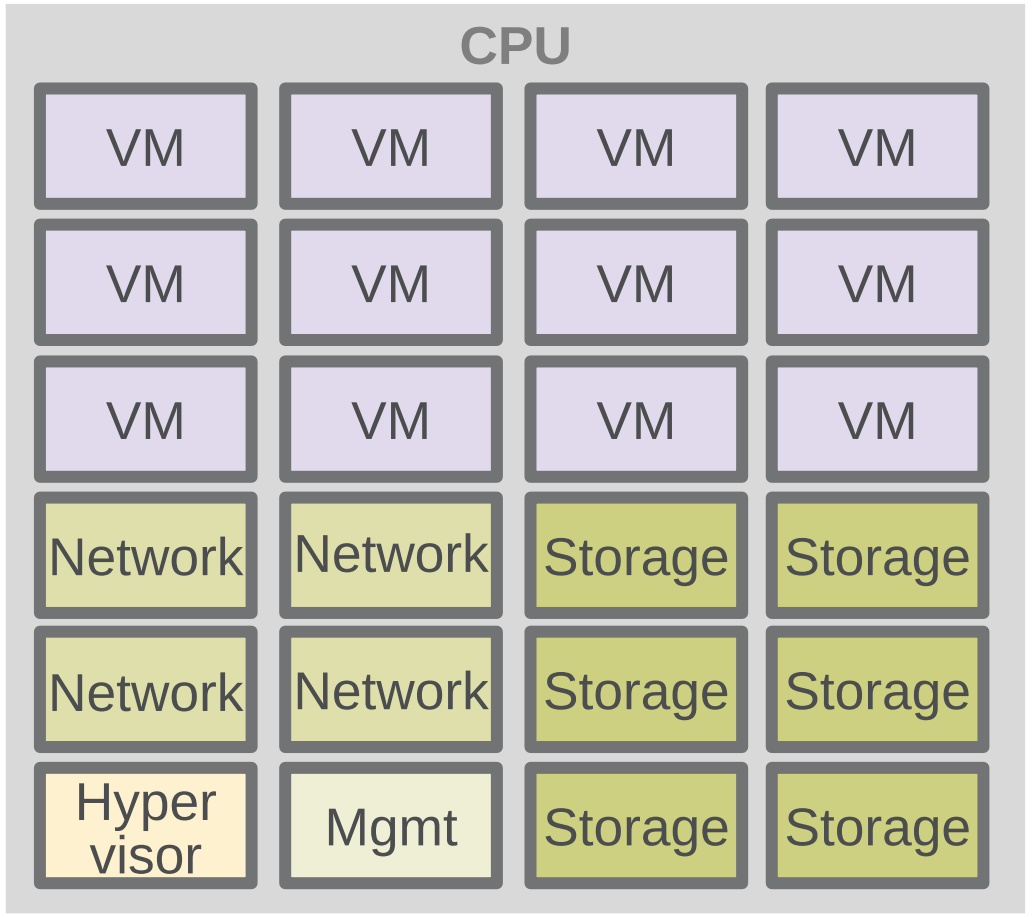
- Offload non-application processing
  - Move networking and storage services: Xeon → SmartNIC
- Reduce time applications wait for data
  - Disaggregate storage to cost effectively move to SSDs
- Move network traffic off CPU coherent interconnect
  - Distribute network & storage traffic directly to each core

BROADCOM

Step #1: Offload non-application processing

- **Virtualized compute server**





Functions running on CPU

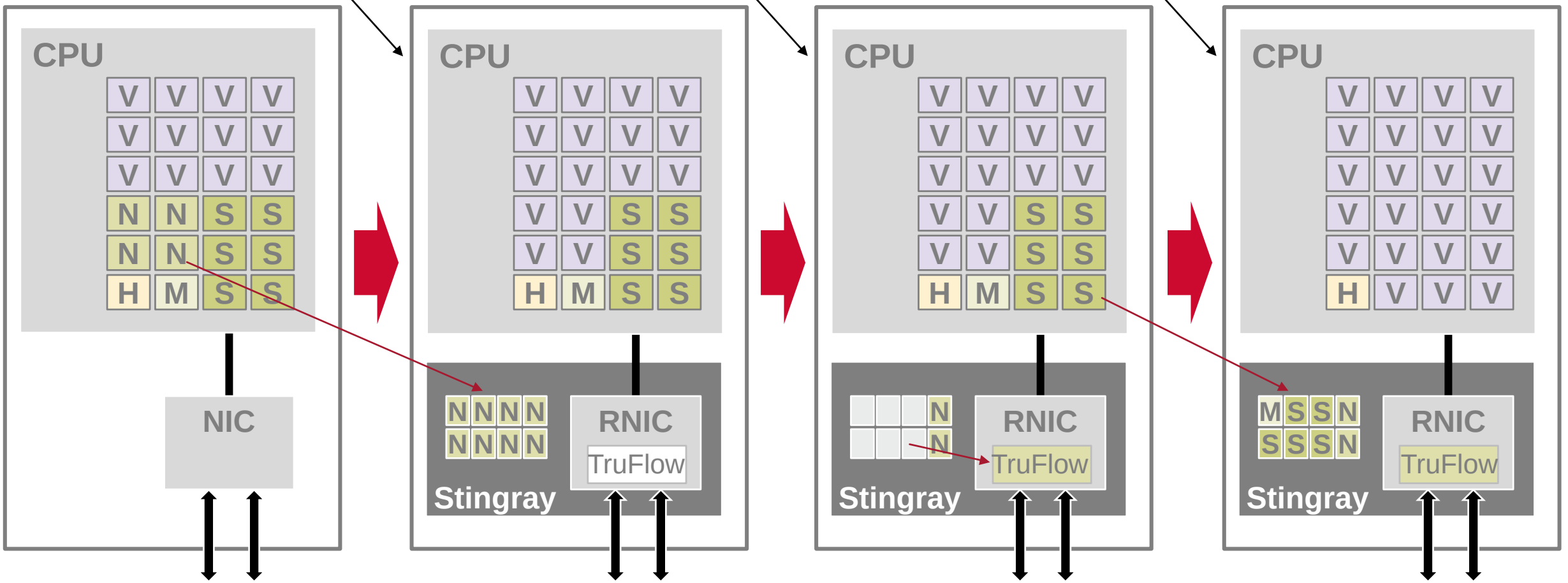
Functions mapped to 24 CPU cores

- E.g. Skylake Xeon Platinum 8168

BROADCOM

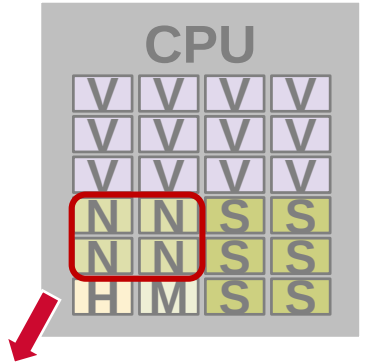
Migrate Services off Compute Server

- **Move networking services to SmartNIC → free up CPU cores for applications**
- **Hardware acceleration in SmartNIC provides performance gains**
- **Maximize available CPUs by moving management & storage services**



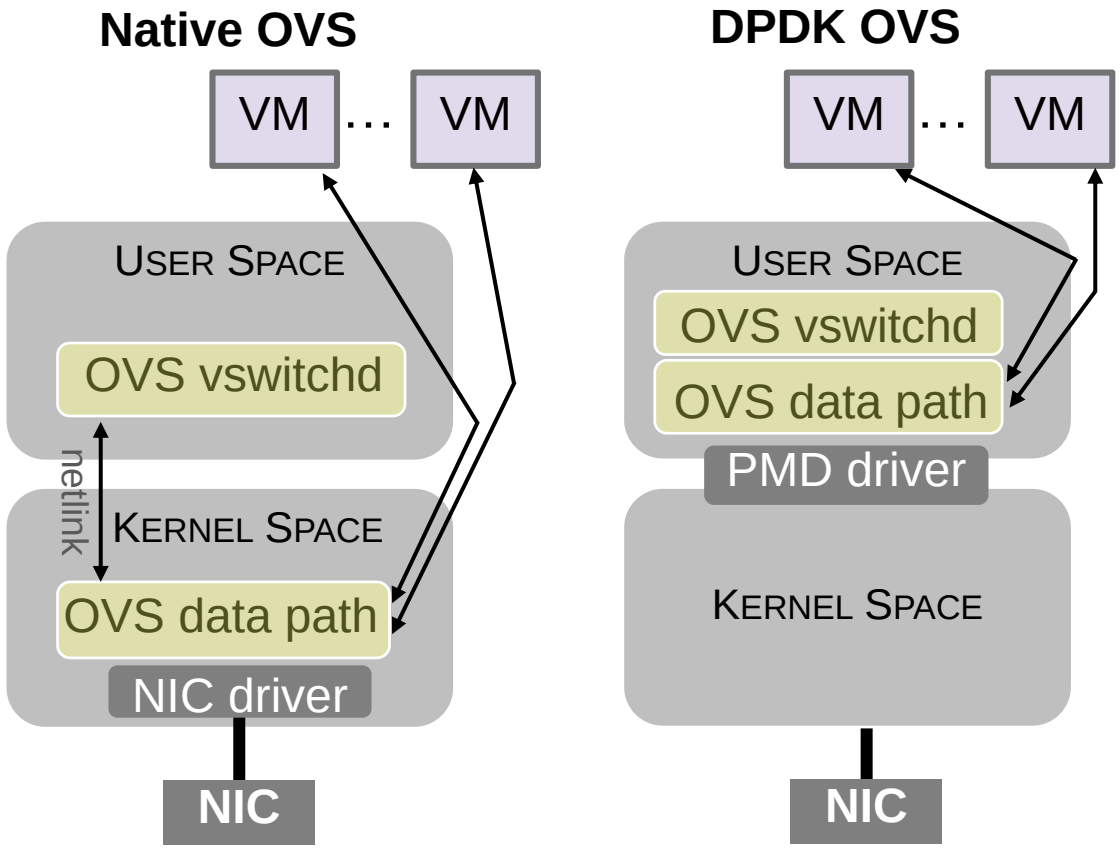
BROADCOM

Network Virtualization using Open vSwitch



- **Open vSwitch (OVS) provides network virtualization services**
  - Layer between NIC and VMs, processing all transmit & receive packets
  - Parses packets and applies forwarding policy
  - Provides fair access to all VMs

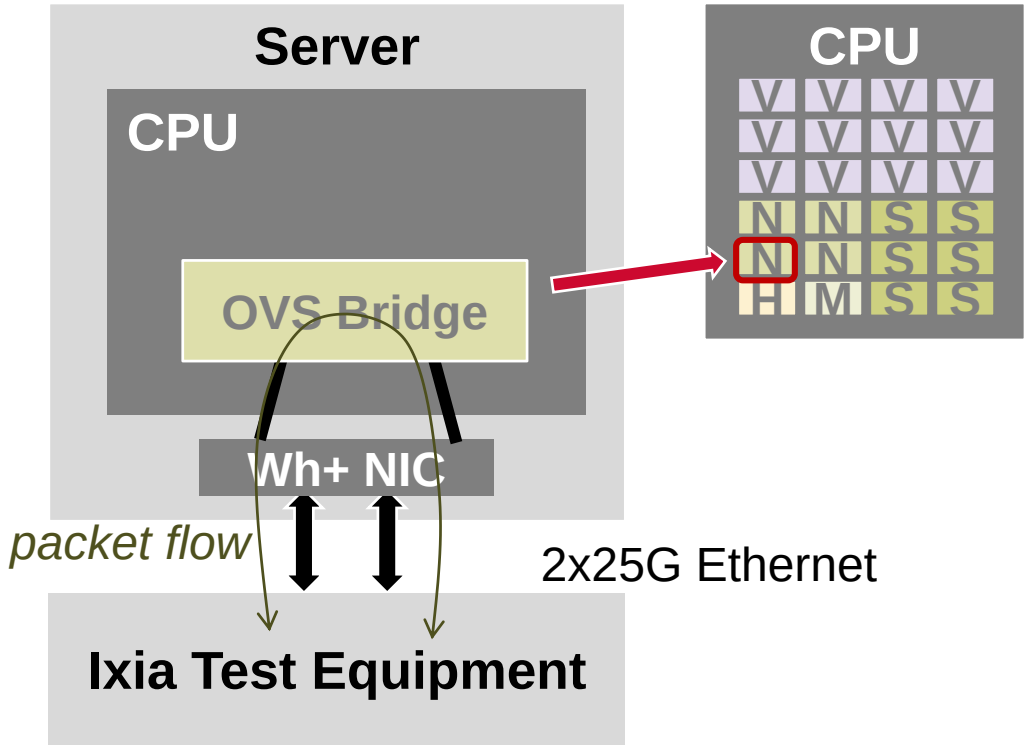
- **OVS performance measured in PPS**
  - Native OVS uses kernel space fast data path and user space control plane
  - OVS DPDK uses poll mode driver (PMD) to run fast path in user space

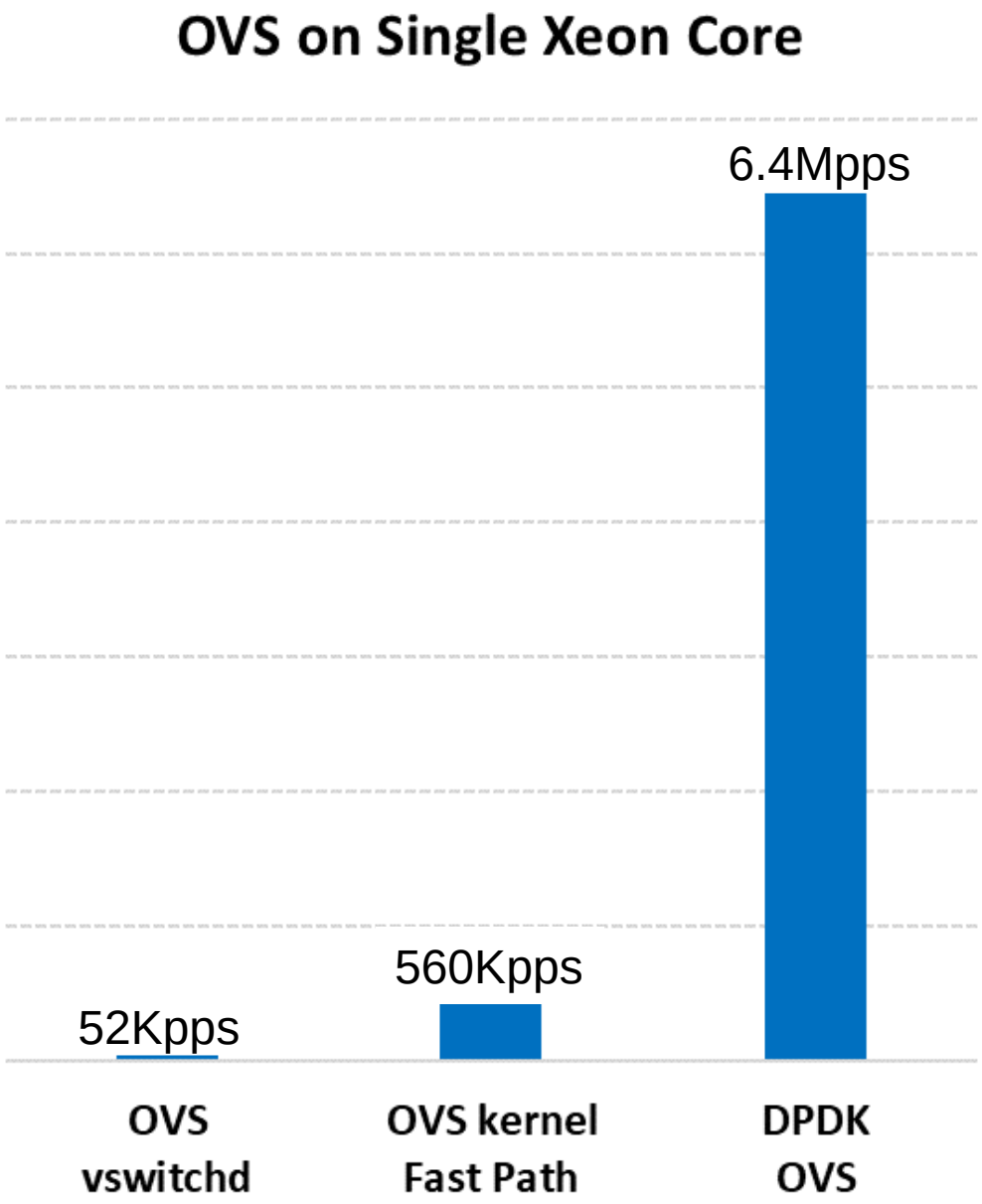


BROADCOM

OVS Performance on Single Xeon

- **Data collected using Dell R730**
  - Dual socket Intel Xeon E5-2697v4 @2.6GHz
  - RHEL 7.6 (3.10 kernel), no HT
- **Layer 3 forwarding, no network overlay**

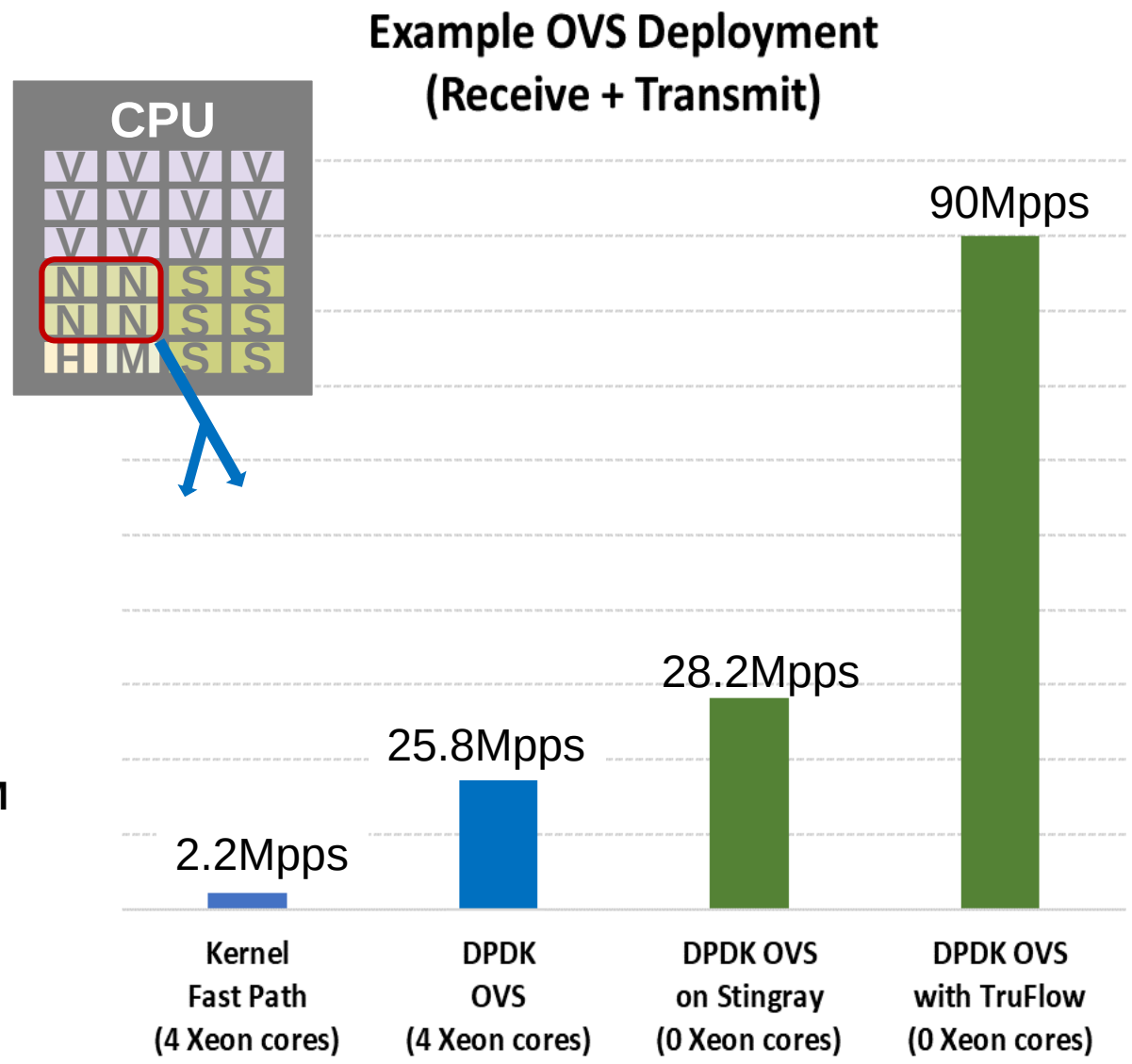




BROADCOM

OVS Performance: 4x Xeon cores vs. Stingray

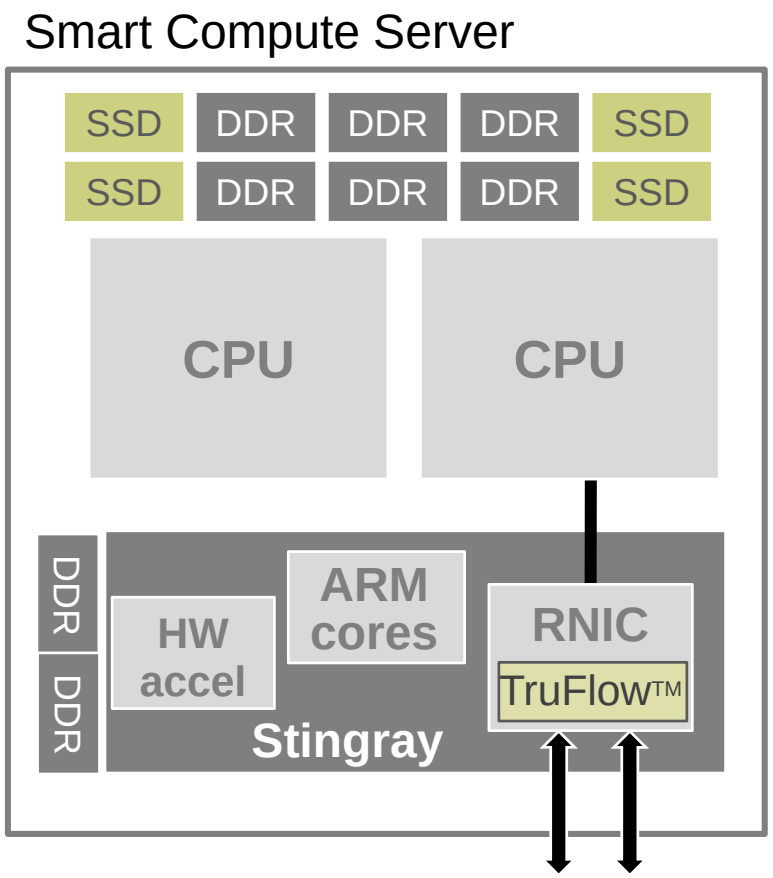
- **OVS performance on 4 Xeon cores**
  - Performance scales roughly linearly per core
  - Stingray 8x A72 cores is similar performance
- **SmartNIC frees 4 Xeon cores for revenue generating applications**
  - Generate up to $18,000 additional revenue over server life cycle
- **Dedicated network processing HW accelerates performance → TruFlow™**
  - Increases PPS
  - Reduces power



BROADCOM

Accelerate Applications using Offload to SmartNIC

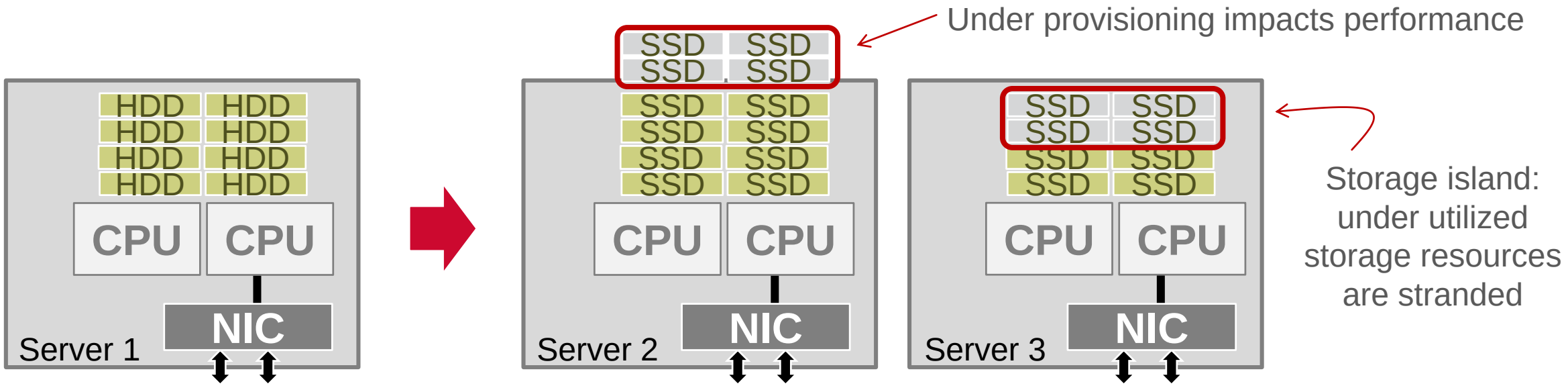
- **SmartNICs optimized to accelerate network and storage services**
  - Hardware provides high performance, low power
  - Software provides flexibility and customization
- **Moving these services off more expensive, higher power CPUs frees up cores for applications**
  - Improves total cost of ownership (TCO) for server
- **Ideal for bare metal services**
  - Leaves CPUs totally available for customer use
  - Isolates DC management and services from customer code



BROADCOM

Step #2: Disaggregation of Storage

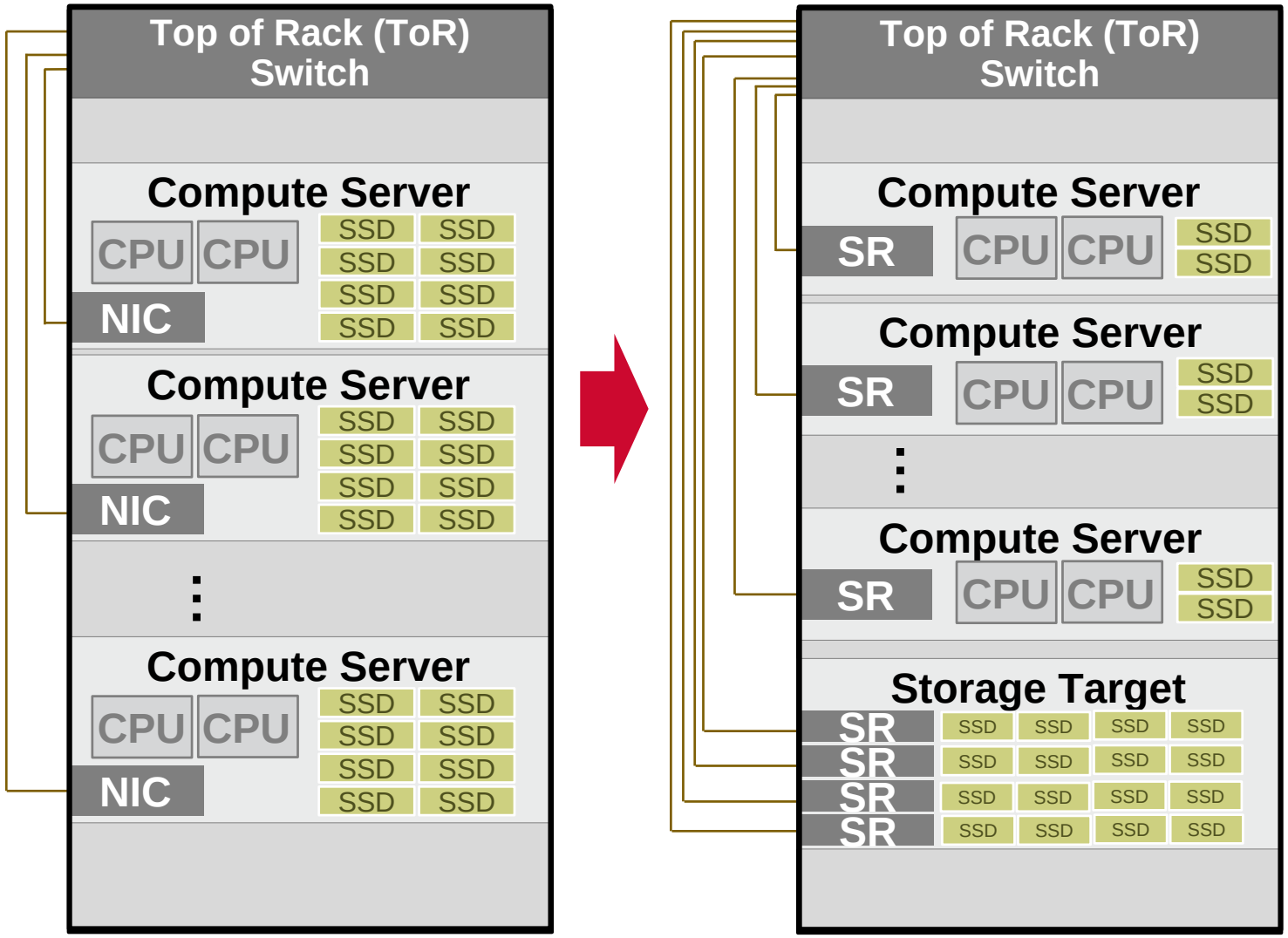
- **IO access time has major performance impact on applications**
  - Data centers moving to SSDs to accelerate performance
- **SSD cost per terabyte 2x – 4x more expensive than HDD**
  - Efficient deployment of SSDs can be challenging
    - e.g. Application on server 3 needs 4x SSDs → 4 SSDs idle, wasted $$ → 'storage island'
    - e.g. Application on server 2 needs 12x SSDs → insufficient high speed storage, slows application



BROADCOM

Scalable High Performance Storage → Disaggregation

- **Disaggregated storage using SmartNIC**
  - Low power, network connected
  - High performance 100G Ethernet
- **Allocate, upgrade and scale storage without affecting compute servers**
- **Questions:**
  - Latency for network attached SSDs?



SR = Stingray

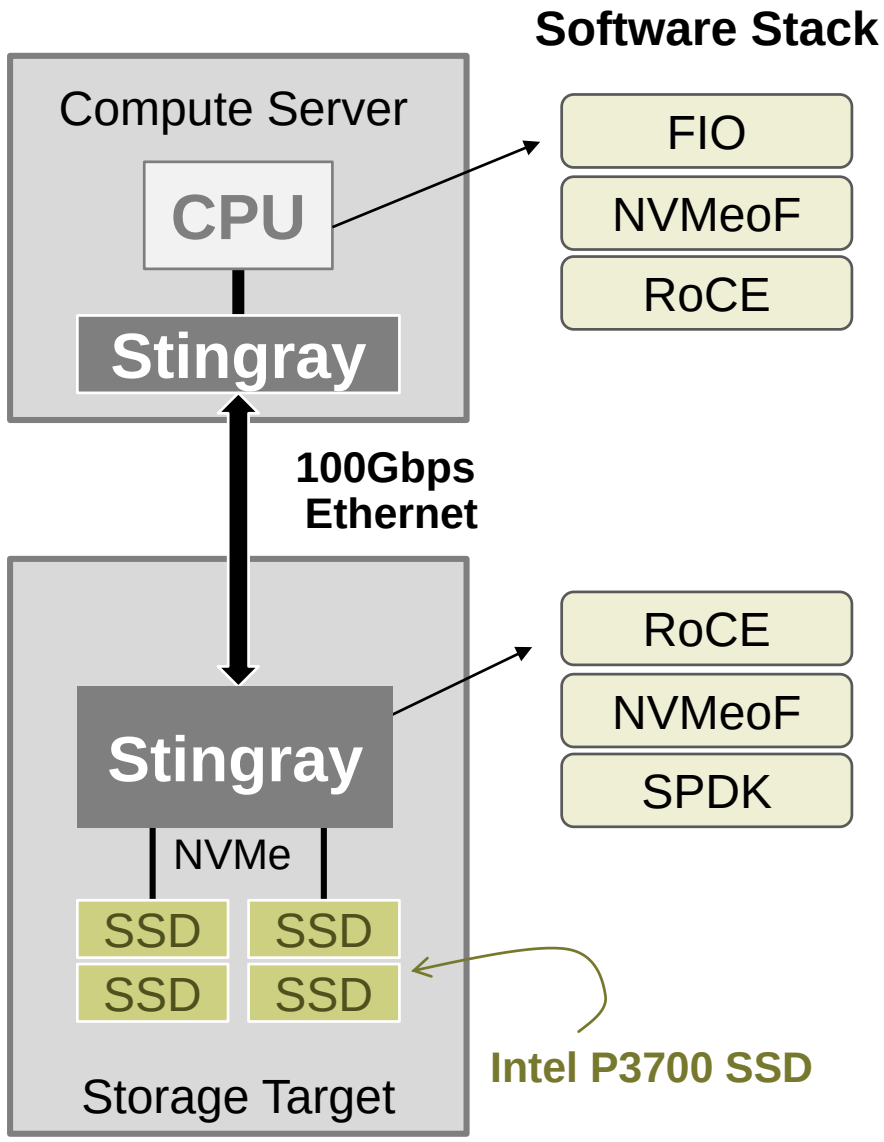
BROADCOM

Latency of Network Attached SSD

- **Nominal access time for SSD: < 0.1ms**
  - Access time varies by operation: read, write, random, sequential, etc.
  - HDD access time is ~ 5.5ms to 8ms

- **Latency data collected using system shown**

Operation – single block	Latency to Null device on Storage Target	Latency to SSD on Storage Target
4KB Random Read	20.67 usec	66.38 usec
4KB Random Write	21.66 usec	51.23 usec

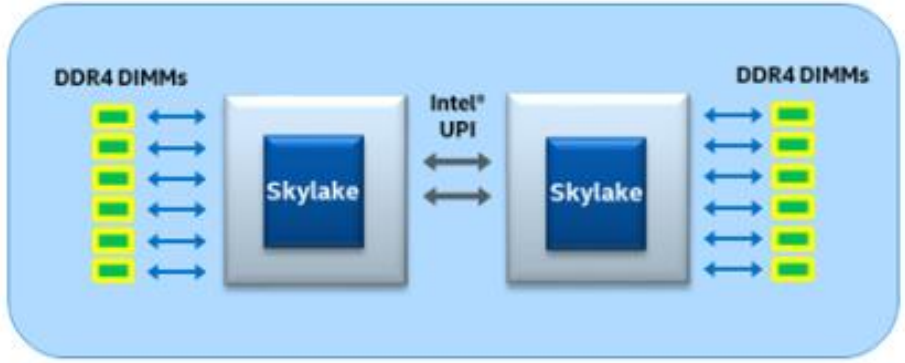


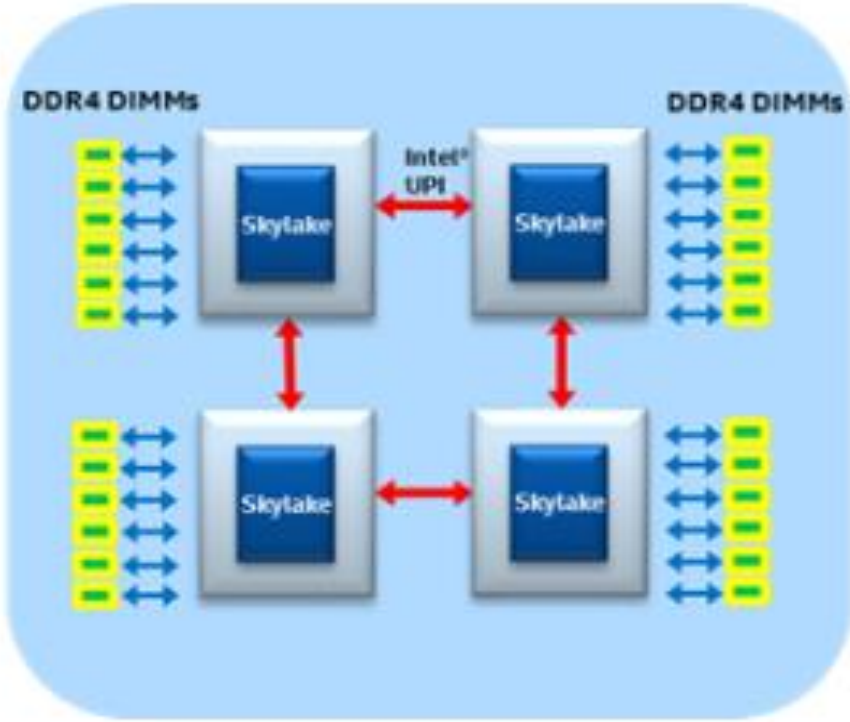
Disaggregated storage provides access times < 0.1ms

BROADCOM

Step #3: Disaggregation - CPU

- **Coherent, high speed interconnect used to scale processing power**
  - e.g. Intel Ultra Path Interconnect* (UPI) in Skylake Scalable Products platforms
    - UPI operates up to 10.4GT/s
  - Enables single, shared address space
- **Enable ccNUMA architecture**
  - SMP (Symmetric Multi-Processing) uses shared memory bus
  - ccNUMA (Non-Uniform Memory Access) uses physically different busses with cache coherency
- **Or just connect multiple CPU sockets**
  - Shared network connectivity

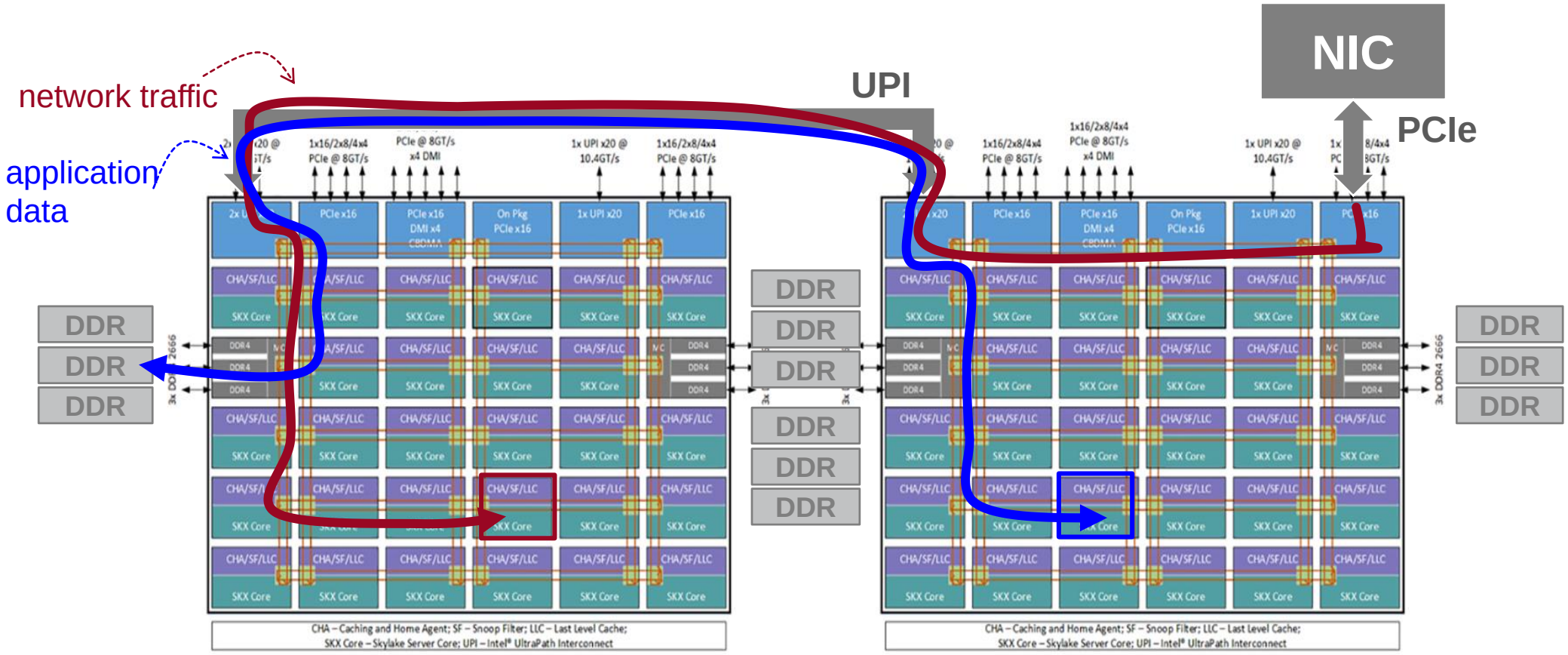




BROADCOM

Network Connectivity using UPI

- **Network packets may traverse UPI to reach destination core**
  - Network traffic competes with application data for UPI bandwidth

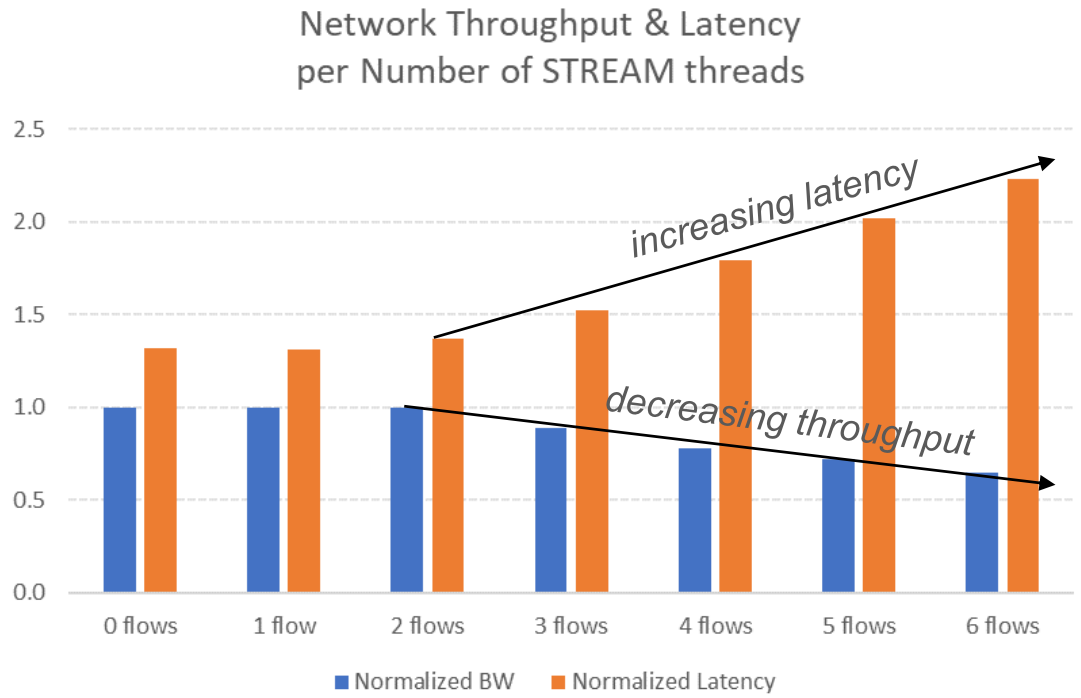


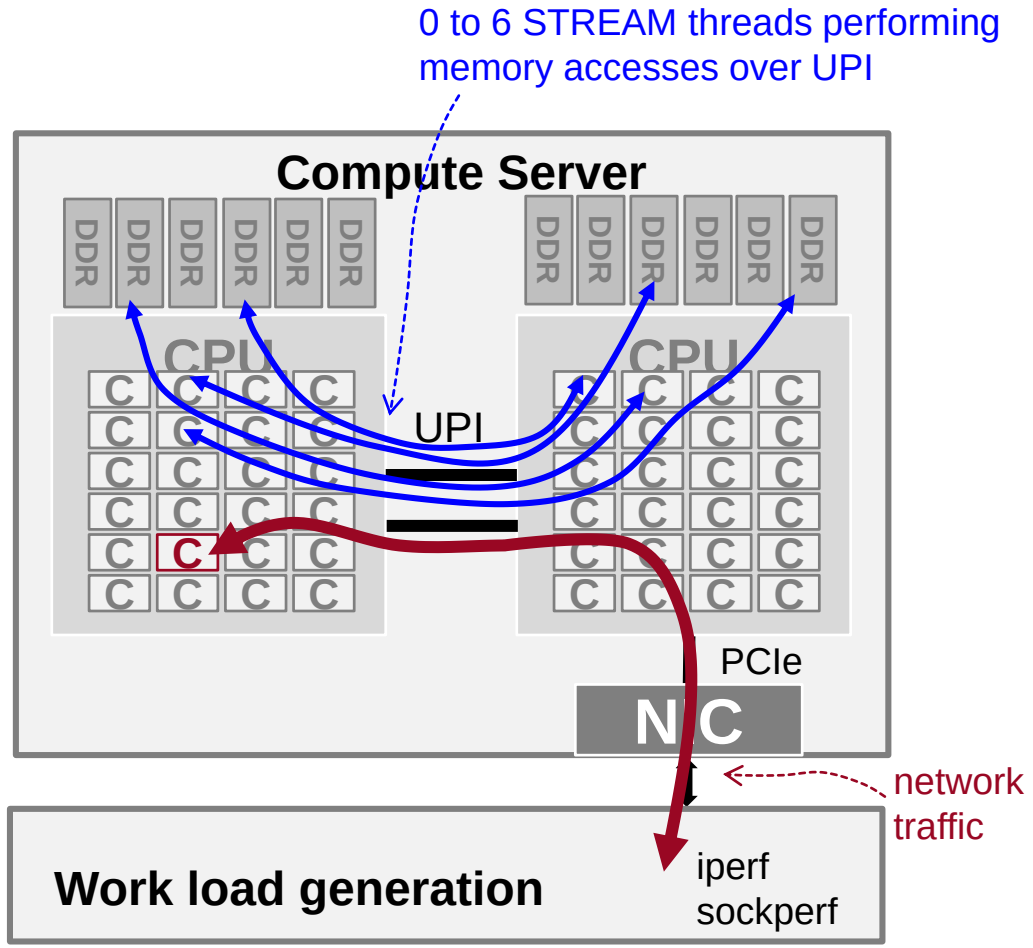
Purley Skylake mesh architecture

BROADCOM

Network Performance over UPI

- **Test scenario**
  - Generate network traffic to core reached via UPI
  - Measure network traffic throughput and latency while adding STREAM threads
    - STREAM generates memory reads to stress UPI
    - Results normalized to MH performance, no streams



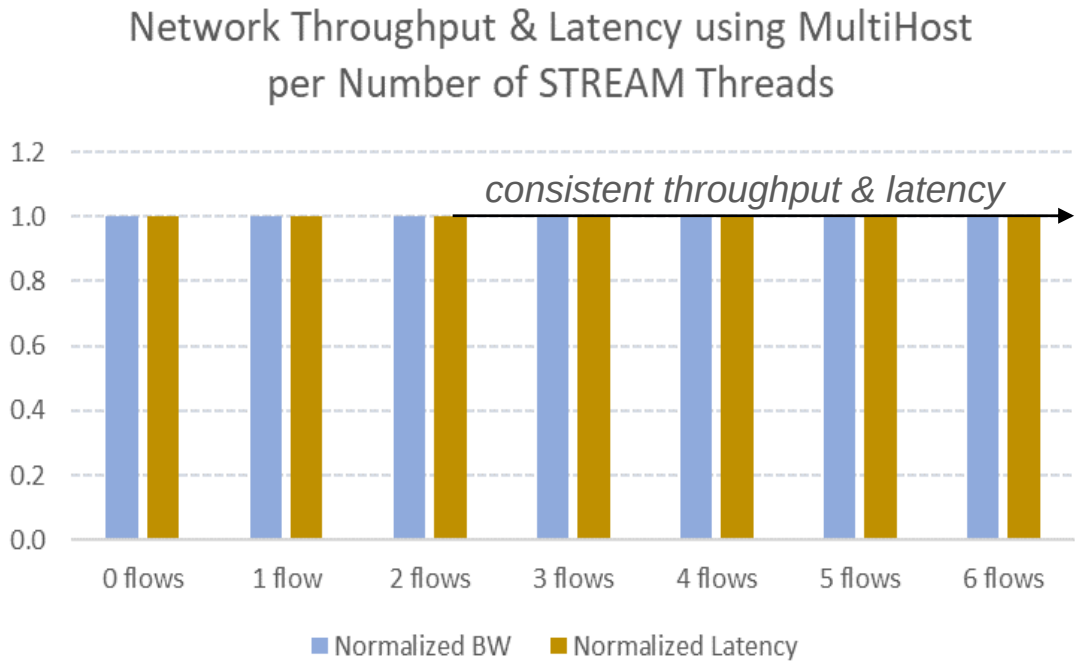


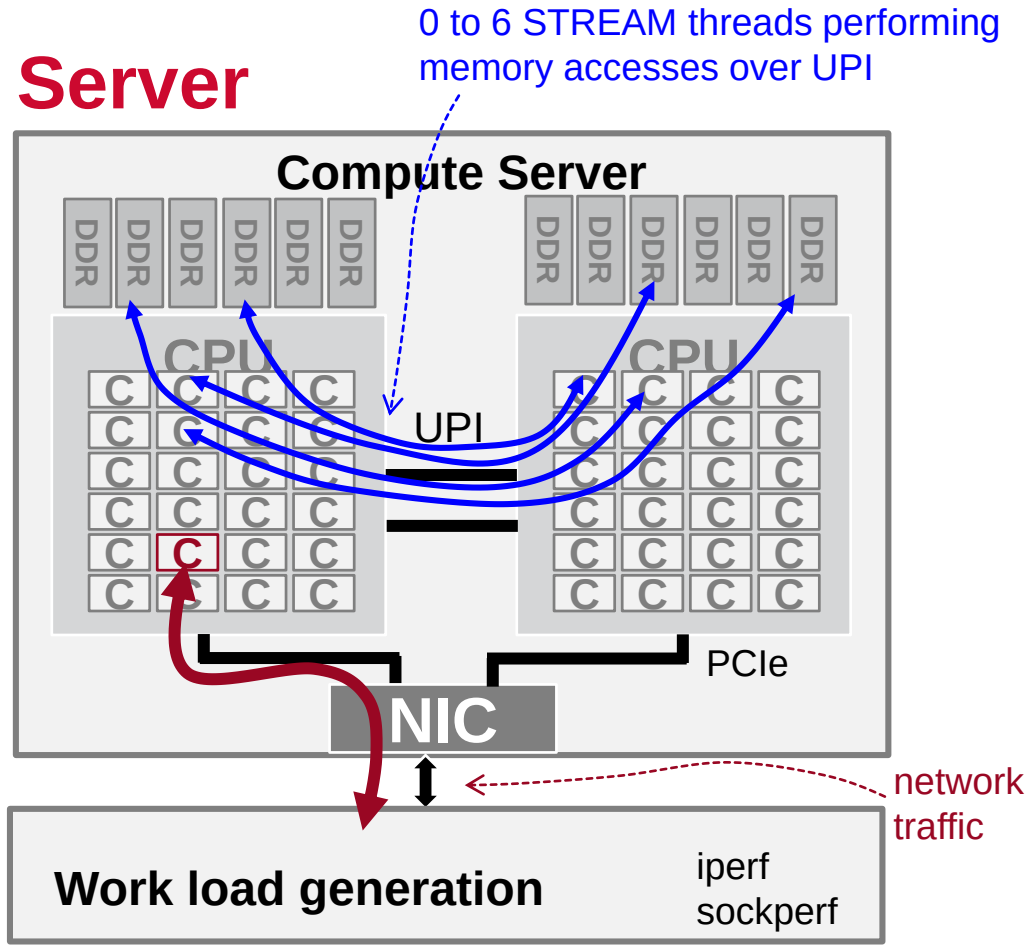
Under increasing load on UPI, throughput drops and latency grows

BROADCOM

Multi-Host NIC for Efficient Multi-Socket Server

- **Multi-host provides independent PCIe connections to each CPU**
  - Take network traffic off UPI
  - Supported by OCP in 2x and 4x host configurations





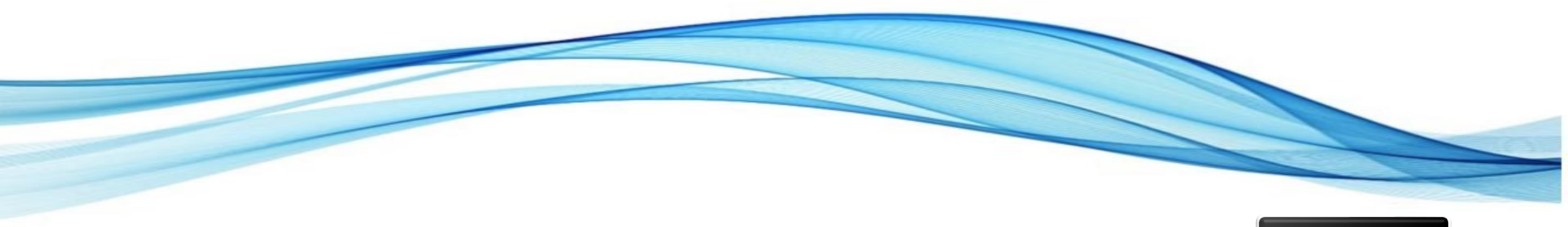
- **Frees UPI bandwidth for NUMA applications**
- **Enables using lower cost CPUs without UPI for multi-socket servers not requiring NUMA**

BROADCOM

SmartNICs Accelerate Data Centers

Accelerate by maximizing CPU processing power

– Focus high performance CPUs on running high performance applications



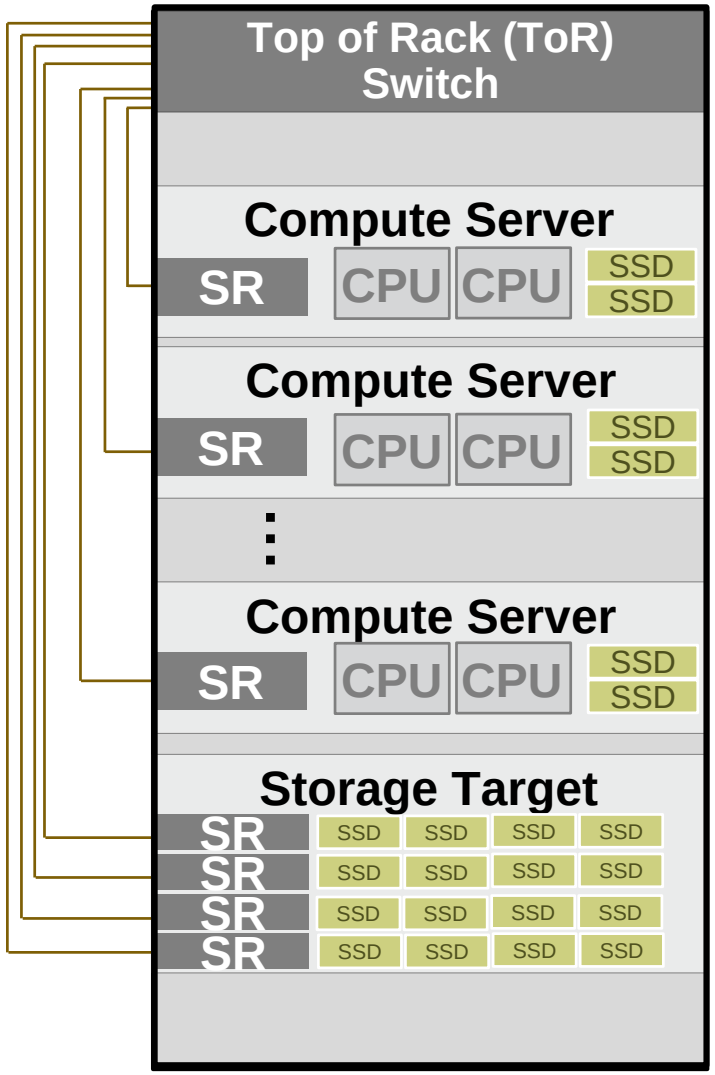
- **Offload network & storage services**
- **Deploy more high speed storage using disaggregation**
- **Move network traffic off CPU coherent interconnect**

BROADCOM

On a Related Topic --- Network Attached Services …

Disaggregation in the Data Center

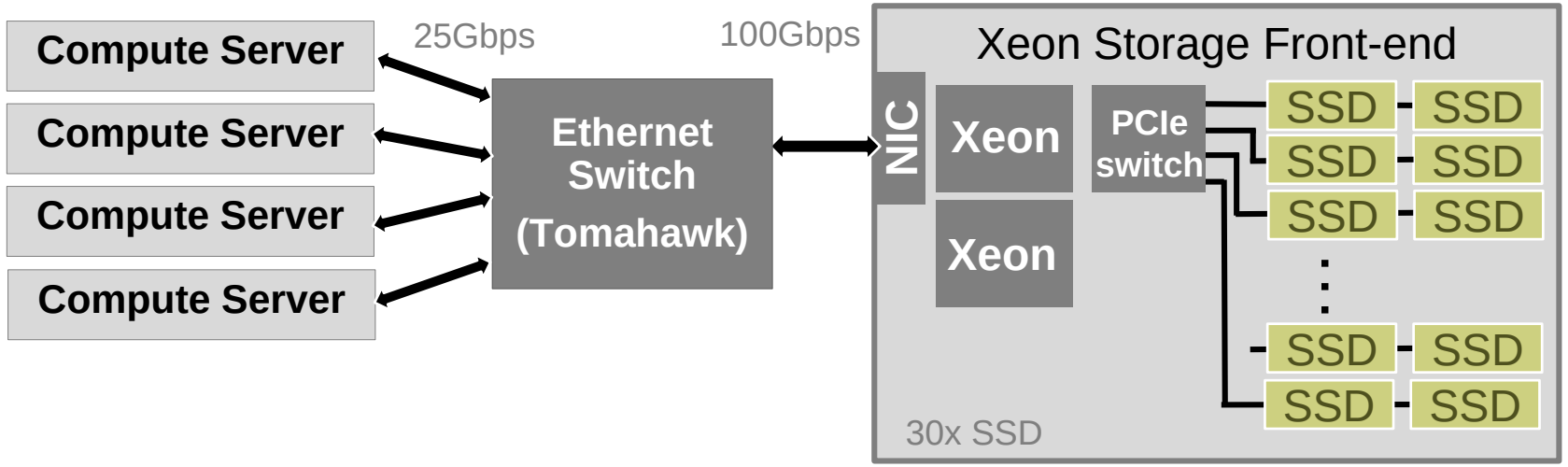
- **Disaggregating storage provides more flexibility in allocating resources**
- **This approach will be expanded to include other resources:**
  - GPUs
  - FPGAs … all types of hardware and software services
- **This trend will affect network traffic patterns and drive need for high utilization and lower latency**



SR = Stingray

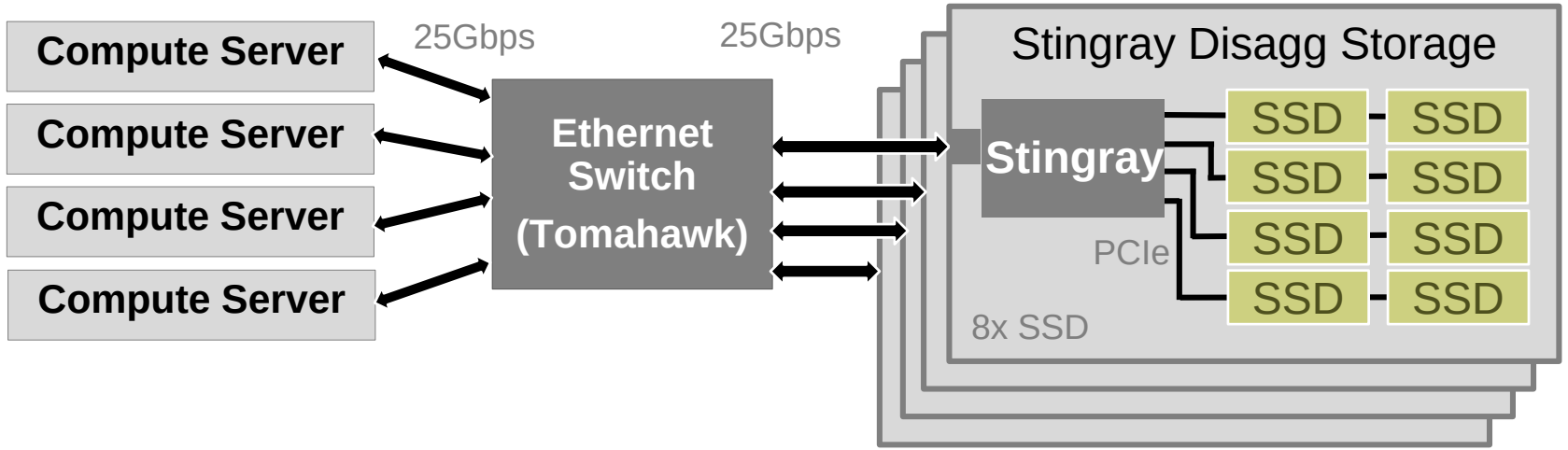
Performance of Stingray vs. Xeon in Storage Target

- **System #1**
  - 1 network attached storage target
  - 2x Xeon (Skylake)
  - 30x SSDs
  - 100G Ethernet



2x PCIe Gen3 to each SSD

- **System #2**
  - 4 network attached storage targets
  - 1x Stingray SmartNIC
  - 8x SSDs
  - 25G Ethernet



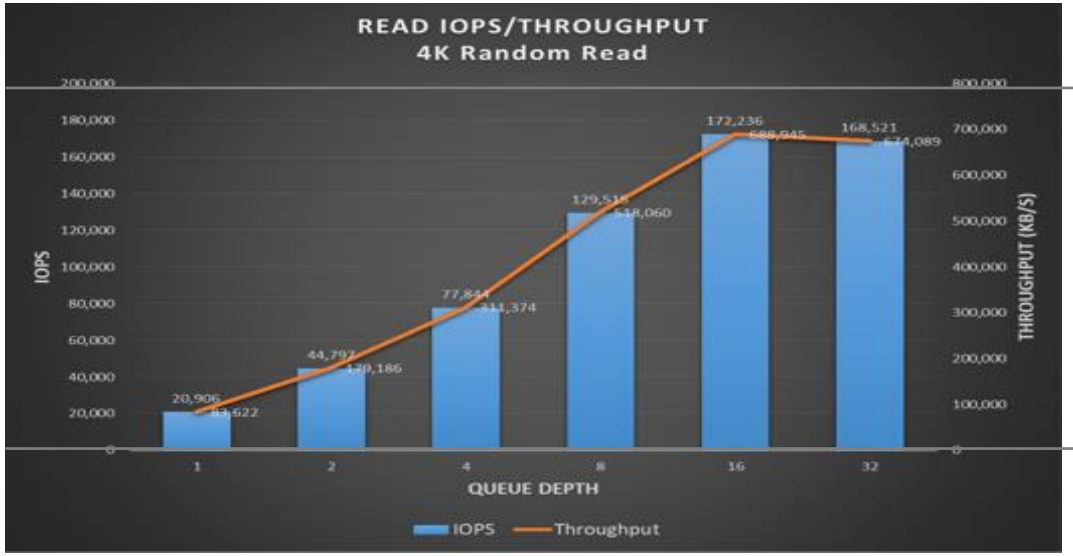
SSD = Samsung PM 983, 2TB

BROADCOM

Test Scenario

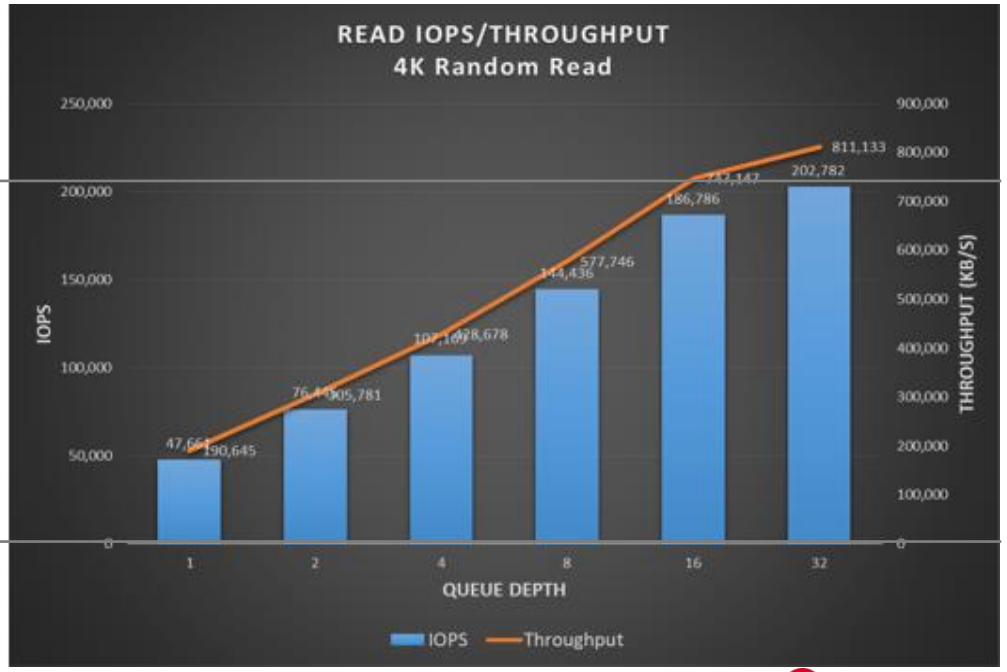
- **Stimulus: enhanced FIO**
  - 4K random read, 4K random write, 4K random read/write, sequential read & write
  - Drives conditioned prior to running tests
  - Network transport: TCP
- **Collected results for varied traffic patterns → 1 server to 1 SSD shown below**

System #1: Stingray



200K IOPS

System #2: Dual socket Xeon



BROADCOM

Test Result Summary

Parameter	4x Stingray Targets	1x Dual socket Xeon
Network Link	4x 25G	1x 100G
NVMe SSDs (x2 Gen3)	32	30
4K Random Read	2.0M IOPS	1.8M IOPS
512K Sequential Write	37K IOPS	18K IOPS
Tail latency (mean – P90 – P99.9)	2ms – 6.2ms – 11ms	2.3ms – 12.9ms – 23.5ms
CPU+DRAM Power (*estimated*)	160W	300W

- **SmartNIC disaggregated storage provided better performance at significantly lower power**
- **Smaller fault domain advantage of smaller disaggregation unit**
  - 16TB rather than 60TB of storage

Note: disaggregated storage affects network traffic patterns

BROADCOM

Hyper Scale Network Simulation

John Gage, 1984

- **Sun had it right**
  - The network IS the computer in today's Cloud Data Center
- **As Moore's law slows down and new technologies are deployed →**

  network performance == application performance
- **Meeting these emerging requirements and minimizing silicon re-spins requires understanding how applications and networks interact**

A CALL TO ACTION:
Our industry needs a framework for simulating data center networks with realistic models of application workloads

BROADCOM

Thank You

Karen Schramm

karen.schramm@broadcom.com