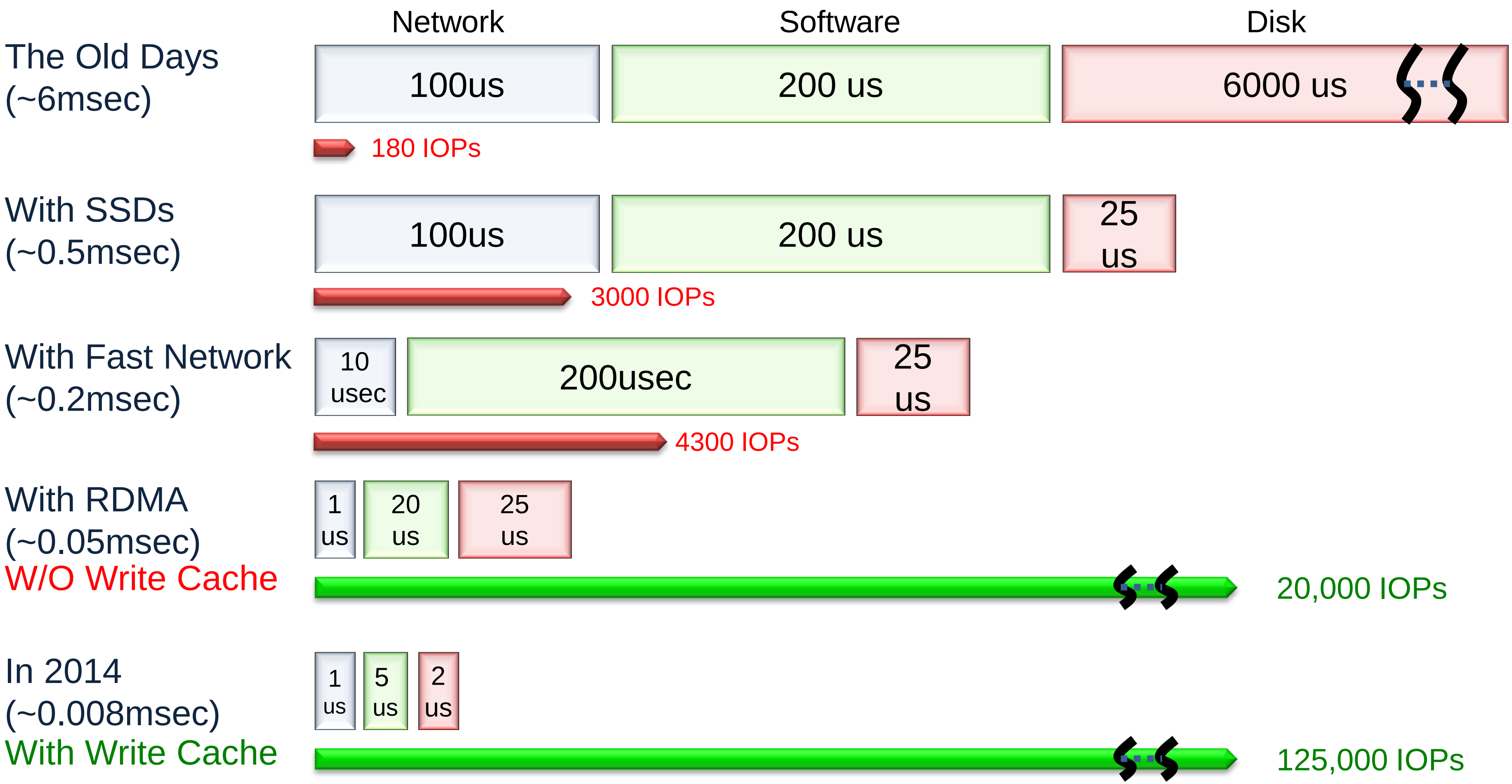
Achieving100Gb/s Flash Connectivity

Why and How

Kevin Deierling

Vice President Mellanox Technologies

Flash is Fast!



Network Software Disk

The Old Days
(~6msec)
100us 200 us 6000 us
180 IOPs

With SSDs
(~0.5msec)
100us 200 us 25 us
3000 IOPs

With Fast Network
(~0.2msec)
10 usec 200usec 25 us
4300 IOPs

With RDMA
(~0.05msec)
W/O Write Cache
1 us 20 us 25 us
20,000 IOPs

In 2014
(~0.008msec)
With Write Cache
1 us 5 us 2 us
125,000 IOPs

The Storage Delivery Bottleneck

Server

+

24 x 2.5" SATA 3 SSDs
(each is 500MB/s)

= 12GB/s =

15 x 8Gb/s Fibre Channel Ports

OR

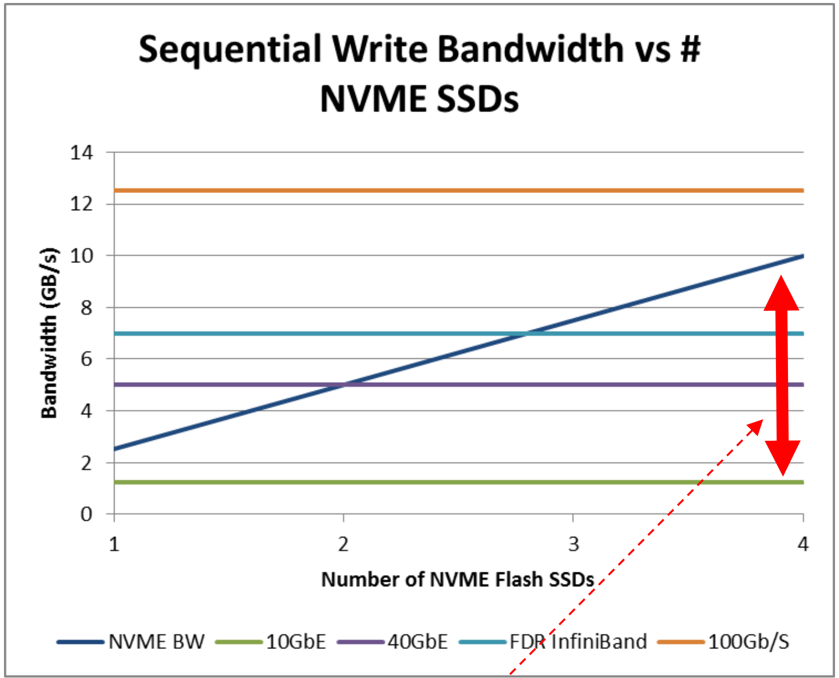
10 x 10Gb/s iSCSI Ports (with offload)

OR

2 x 40-56Gb/s IB/Eth port (with RDMA)

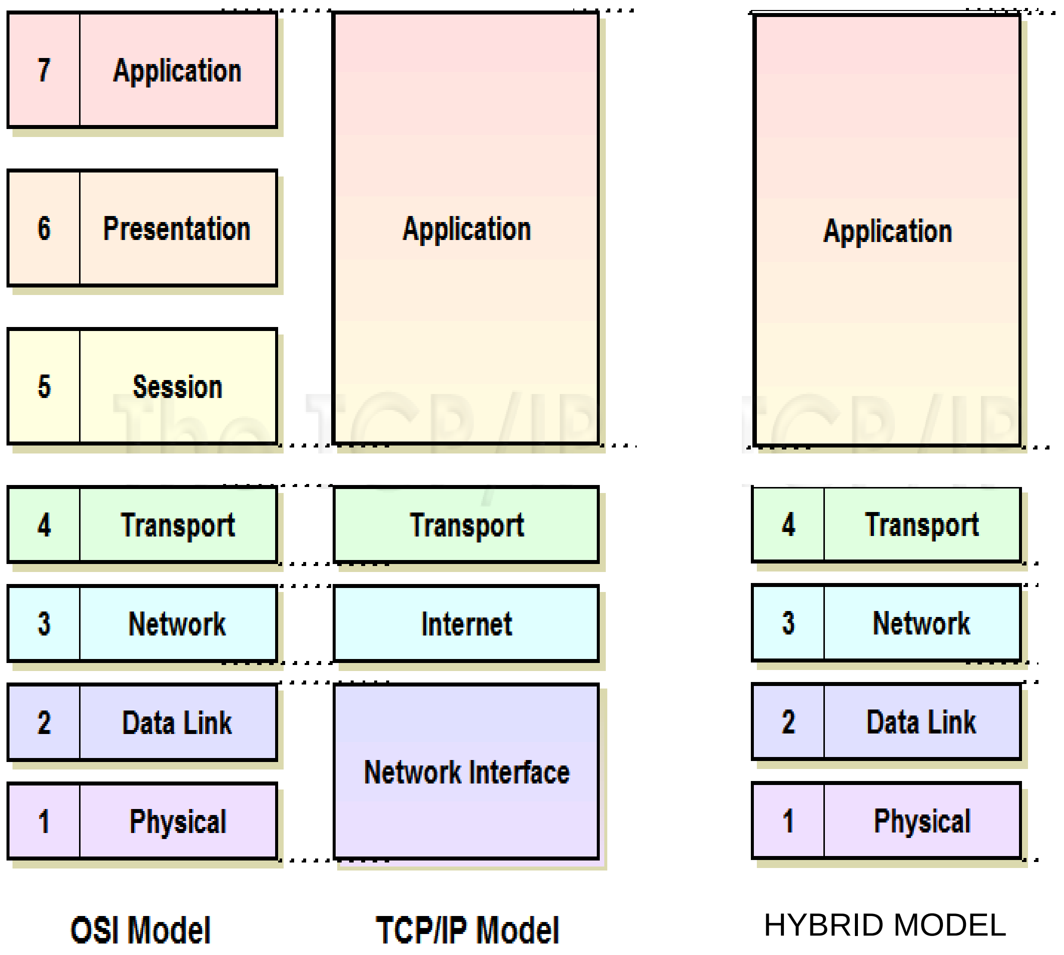
NVMe Flash is Even Faster!

- Flash based SSDs are fast!
  - NVMe: @2.5 GBytes/s
  - DIMM: @10 GByte/s
- Peak throughput is key
  - Particularly for certain workloads
    - Ingest, mirroring, journaling, messaging
- Performance Saves $$’s
  - BW=>Latency=>Performance
  - Performance=>Efficiency
  - Efficiency=>TCO



The Networking Flash Gap!!

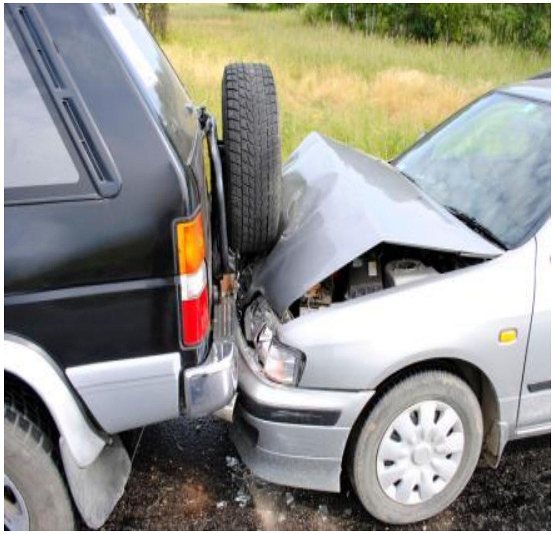
100Gb/s Needs Innovation @ Every Layer



- Application Layer
  - Message format
- Presentation Layer
  - Coding 1's and 0's
- Session Layer
  - Authentication, Permissions, Persistence
- Transport Layer
  - End-to-end error control
- Network Layer
  - Addressing, routing
- Link Layer
  - Error detection, flow control
- Physical Layer
  - Bit stream, physical medium, analog symbol mapping bits

Innovation Required @ 100Gb/s

- Transport Layer Innovation Required
  - TCP/IP dropped packets a non-starter.
  - Rear-ending someone is not the best way to figure out there is congestion
  - Explicit notification required
  - RDMA, virtual nics, virtual traffic steering, affinity
- Network Layer
  - Virtual as well as physical routing (Easy VM migration)
- Link Layer
  - Lossless Networks using Flow control
    - PFC (on/off) flow control is a blunt instrument
    - IETF considering credit based flow control modeled after InfiniBand
- Physical Layer
  - 100Gb/s signaling means 10ps symbol period!!
    - 3 mm pulse of light in free space!
    - Less <<1cm on FR4 … Not feasible at this rate
  - Lower symbol rate required through either:
    - Parallel streams: ex: 4x25Gb/s
    - Multi-bit/symbol: ex: PAM4, WDM

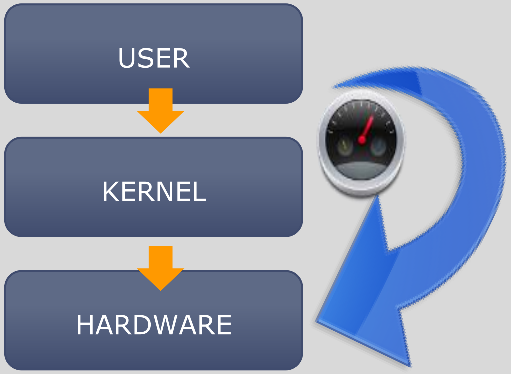


TCP/IP Implicit Congestion Notification aka dropped packets and timeouts

PFC: Priority Flow Control

RDMA: Critical for 100Gb/s

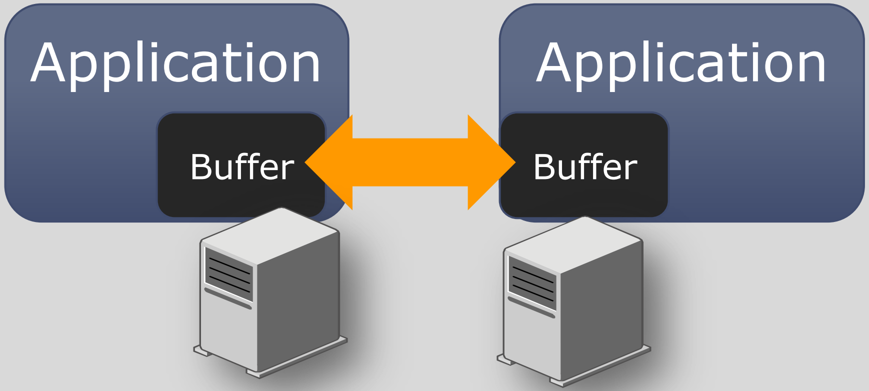
ZERO Copy


USER

KERNEL

HARDWARE

Remote Data Transfer


Application

Buffer

Application

Buffer

Kernel Bypass

Protocol Offload

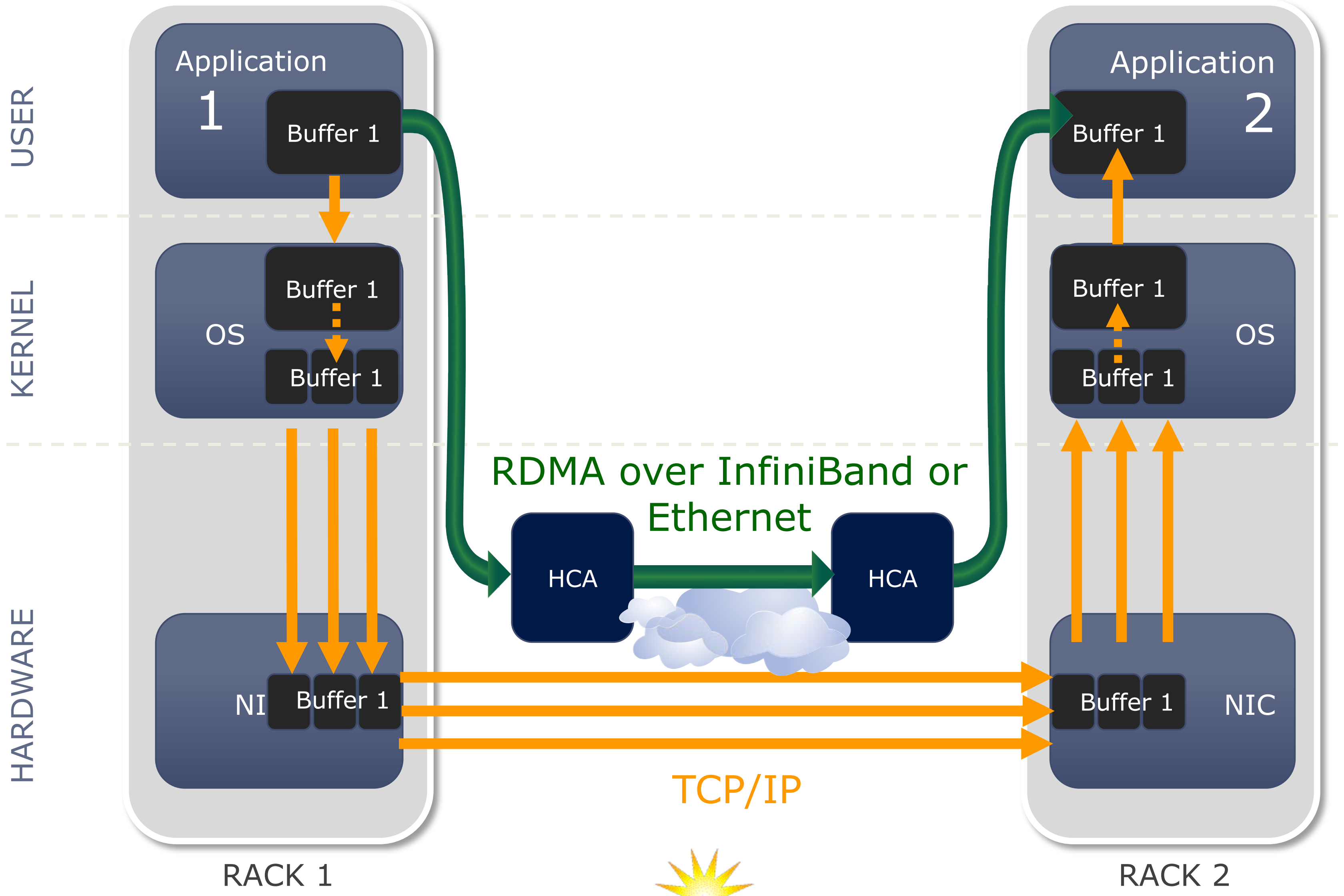
Low Latency, High Performance Data Transfers

InfiniBand - 56Gb/s

RoCE* – 40Gb/s

* RDMA over Converged Ethernet

RDMA: How it Works



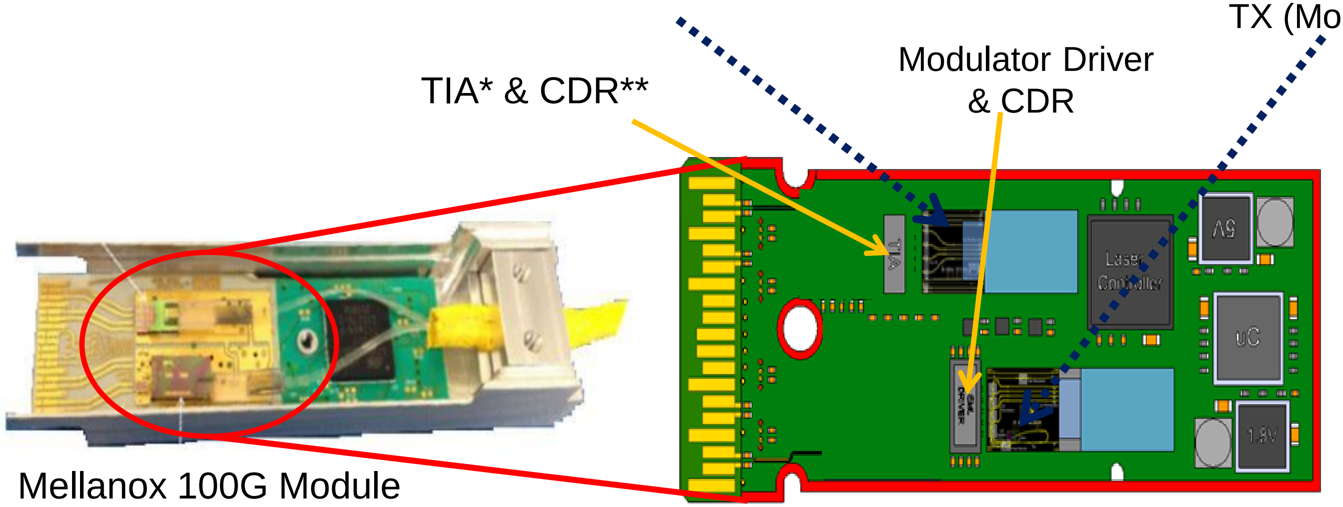
Phy Layer: 100Gb/s in QSFP28 Package

RX (Photo Detector)

TX (Modulator)

Modulator Driver & CDR

TIA* & CDR**

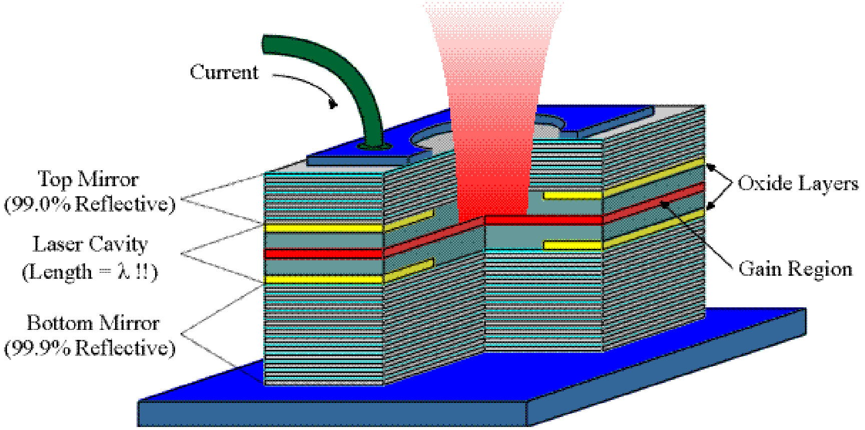


Mellanox 100G Module

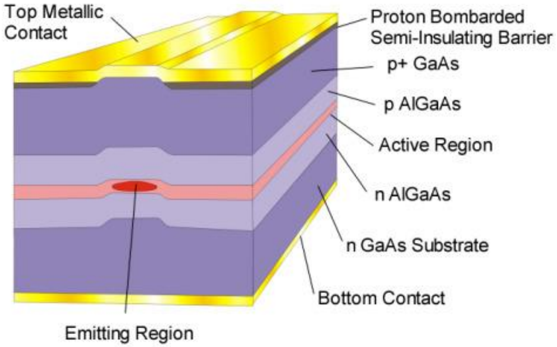
- To fit 100Gb/s in QSFP package requires:
  - – Low power electronics
  - – 4x25+ Gb/s modulators and detectors
- Silicon photonics integration:
  - no lenses for the laser
  - no isolators
  - no TEC

* TIA – Transimpedance Amplifier
** CDR – Clock Data Recovery

Two Basic Technology Options



VCSEL Based

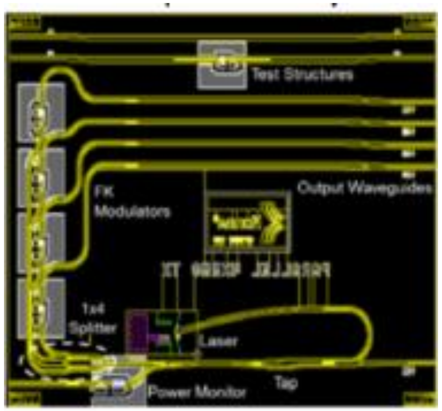


Silicon Photonics Based

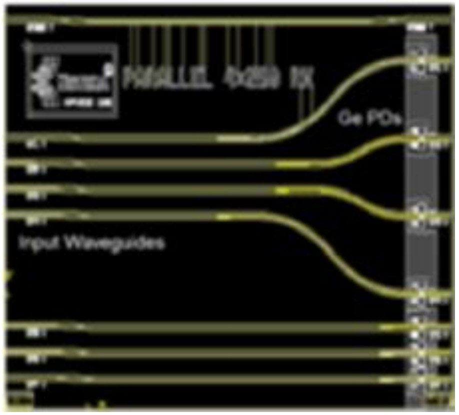
- Direct laser modulation
  - VCSEL
  - 850nm
  - Multi-mode fiber

- Silicon Photonics
  - Fabry Perot or DFB
  - 1550nm
  - Single-mode fiber

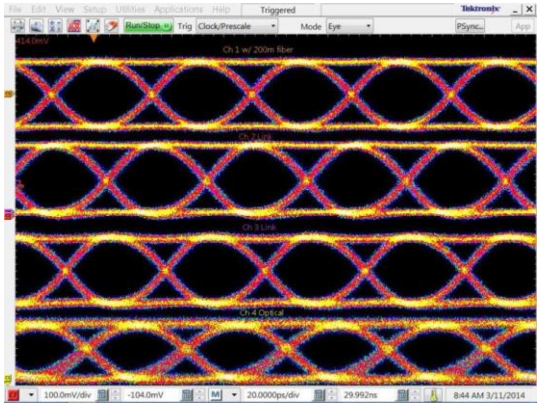
Silicon Photonics



TX (Modulator)



RX (Detector)



Trig Clock/Prescale

Mode Eye

100.0mV/div

-104.0mV

20.0000ps/div

29.992ns

Electrical & Optical
Eye Diagram

- Electro-Optical Modulation
  - Franz-Keldysh optical absorption modulation

Two Technologies, Same QSFP

VCSEL Based QSFP

Silicon Photonics Based QSFP

- Quad Small Form Factor Pluggable (QSFP)
  - Flexibility: Copper, Single Mode, Multi Mode

Thanks!

Questions