VAULT 2019:

Optimizing Storage Performance for 4-5 Million IOPs

James Smart, Emulex Connectivity Division

February 26, 2019

Introduction

- Emulex Connectivity Division – ECD
  - Roots to Emulex Corporation, started in 1979
  - Acquired by Avago May 2015

- Fibre Channel
  - Leading Protocol for SANs and Enterprise Storage
  - Best known for SCSI protocol, now supporting NVMe.

Today's Discussion

- An overview of the performance enhancements in the Emulex LPFC Fibre Channel Driver rev 12.2.0.0

- Motivation:
  - New generation of ASIC supporting many millions of IOPs

- Goals:
  - Making the driver as fast as the hardware
  - Identifying and resolving OS Issues
  - Optimizing the sharing of critical resources
  - Simplifying out-of-box performance

Emulex Fibre Channel Adapter Architectures

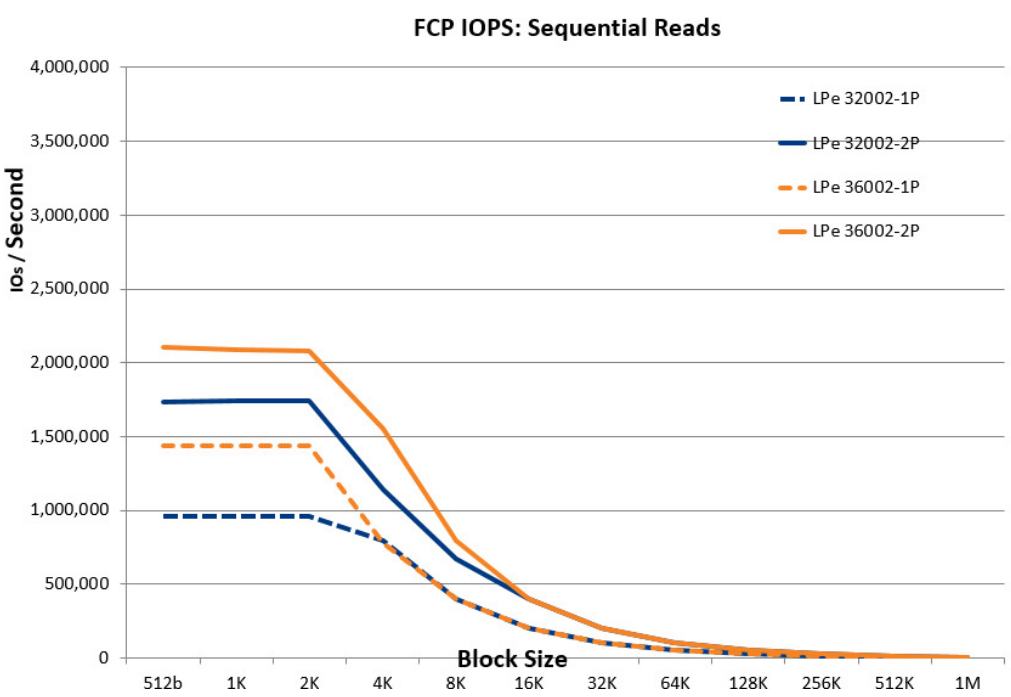
- SLI-3 Interface (~2004)
  - Interface for 8G and older adapters
  - One primary processing engine per port
  - 3 Cmd/Rsp Rings for I/O traffic
  - INTx/MSI/MSIX – but only a couple of vectors
  - ~250K IOPs
- SLI-4 Interface (~2010)
  - Interface for 16G and newer adapters
  - Dynamic Parallel Processing Engines and HW Offloads
  - 1000s of WQ/CQ pairs for I/O traffic
  - INTx/MSIX – up to 1000 vectors (EQs)
  - Gen 5 (16G) Multiprotocol:
    - 1.2 M IOPS
  - Gen 6 (32G) FC:
    - 1.5-2M IOPs, 25-30us hw latency
  - Gen 7 (64G) FC:
    - 5+M IOPs, <10us hw latency

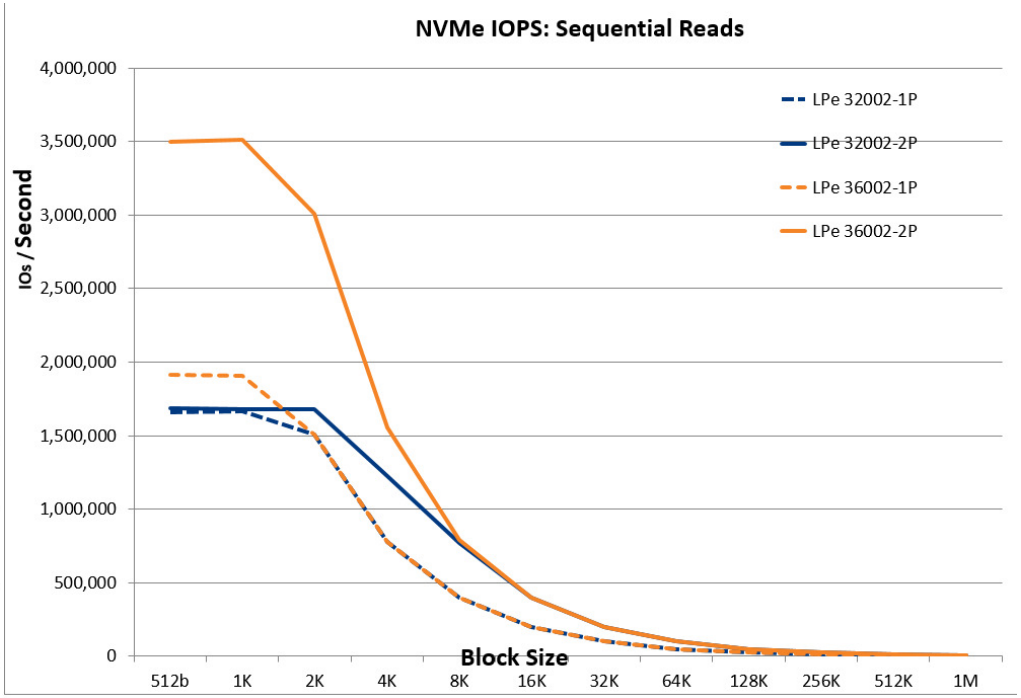
- LPFC Driver supports both architectures and has “evolved” over time…

BROADCOM

The Hardware is Faster – Pre-Driver Enhancements

- 32G; HT off; nomerges; noop scheduler; 2x7 CPU system, 2.1GHz, IOChannel=8, imax=0

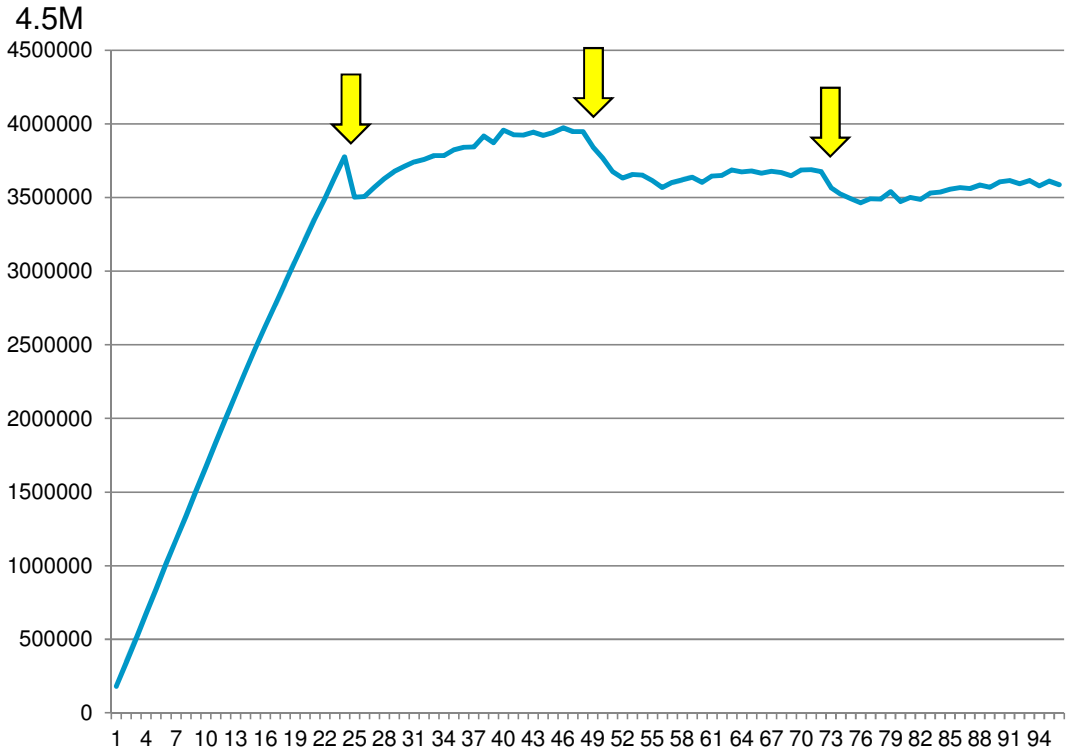




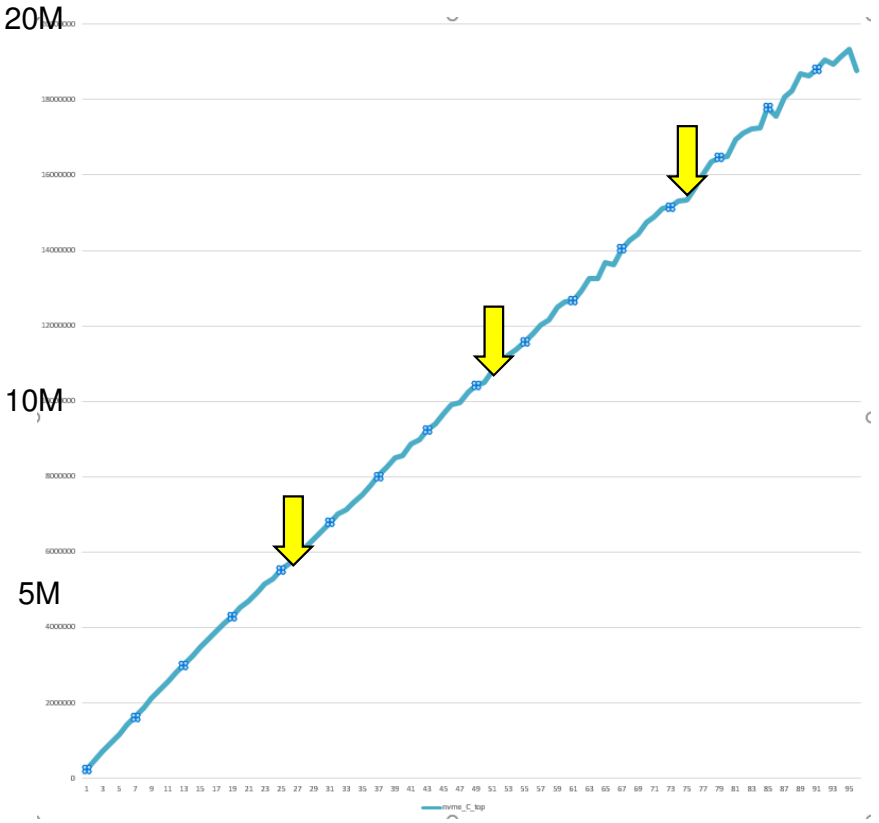
- Driver and/or I/O stack holding back both Gen6 and Gen7 ASICs

BROADCOM

I/O Stacks – Scaling Across Sockets



SCSI



NVME

- Custom “loopback” driver to measure stack
- 4x24 (96) CPU system, 2.7GHz; 4.18ish kernel

= a socket boundary

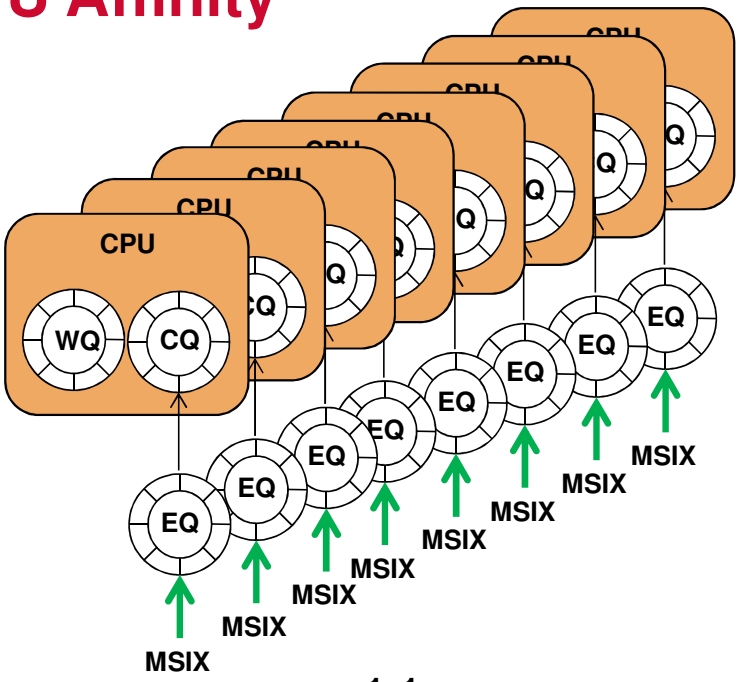
BROADCOM

LPFC Work Areas Identified

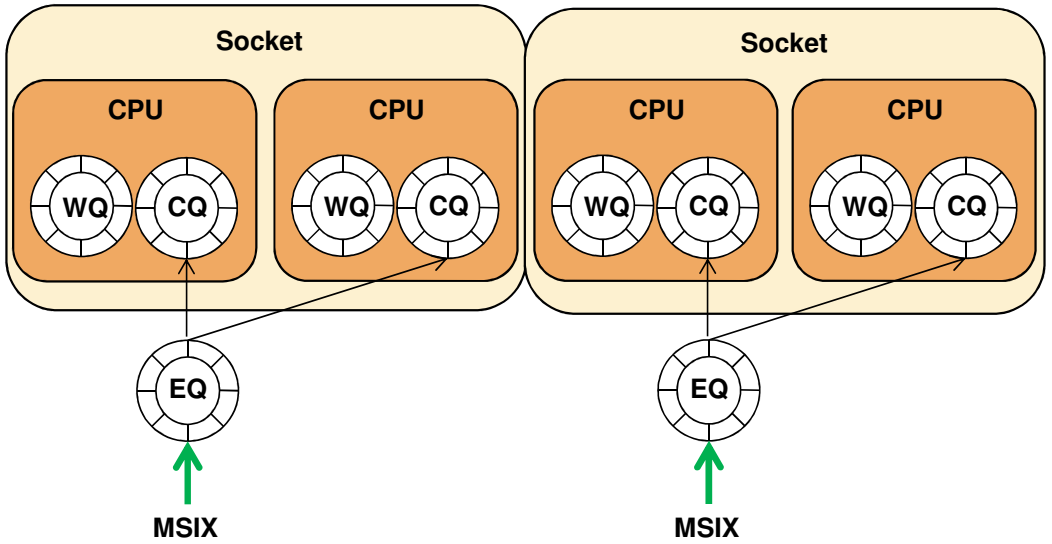
- **CPU Affinity**
- **Shared Adapter Resources**
- **Interrupt Handling**

BROADCOM

CPU Affinity



1:1



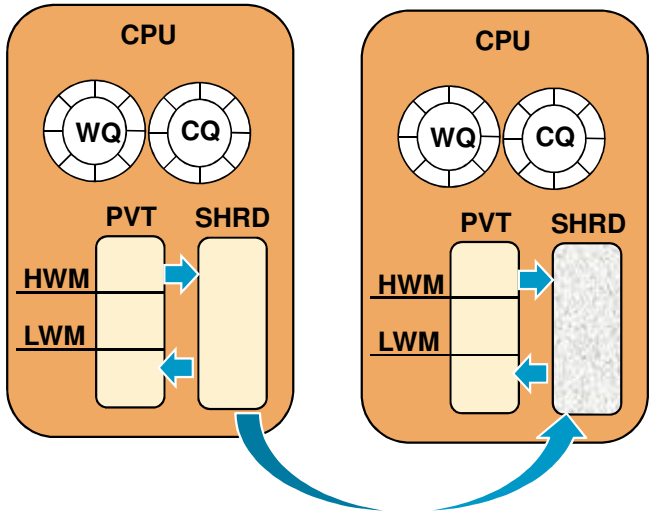
Balanced

- Per-CPU WQ/CQ (a “Hardware Queue”)
- Interrupt vector/EQ per CPU
- When fewer, balance across sockets and CPUs within socket
- Per hardware queue statistics

BROADCOM

Shared Resources

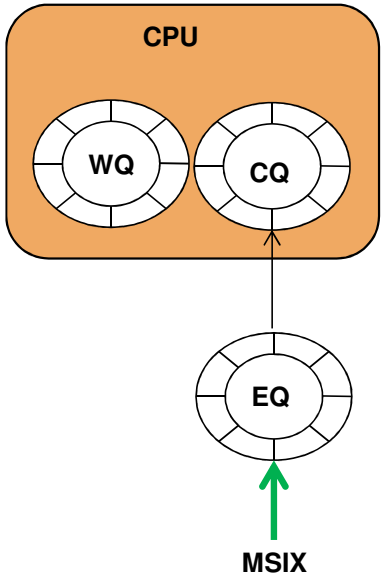
- FC exchanges
  - Adapter has a fixed number
  - Needed for SCSI and NVMe
  - Exchange assigned to each IO for the duration of the IO
  - Partitioning per CPU resulted in few resources per CPU, thus lots of IO “busying”
  - Solve by pools per Hardware Queue with resources migrating between Hardware Queues on as-needed basis



BROADCOM

Interrupt Handling

- Interrupt Handling:
  - Disassociate EQ from CQ
    - EQ must be serviced by ISR
    - CQ serviced by Independent Thread
      - Thread may be scheduled to different CPU than ISR
- CQ Processing Tenancy
  - Aka: How much work you do while in the thread
  - Large limits put in. If limit reached and work remains, re-schedule
- Periodic Queue Pointer Updates to Hardware
- Interrupt Rate Management
  - Interrupt re-enablement
    - Use architecture-specific re-arming to reduce interrupt rate
  - Interrupt delay largely left “immediate”
  - Exception: CPU shared by Interrupt Vectors or HWQs



BROADCOM

LPFC 12.2.0.0 Performance Levels

Reads:	1 Port	2 Port
SCSI	3.73M	4.56M
NVME	5.2M	5.3M
Writes:	**1 Port**	**2 Port**
SCSI	3.69M	4.59M
NVME	4.47M	5.2M
	↑ 2.5x	↑ 2x

- Average Latency Benchmark (fio clat): 10.5us

- Configuration:
  - Initiator:
    - 4x24 CPU @ 2.7GHz
    - Emulex LPe35002 adapter (2x32G)
  - Targets (6)
    - 16,24,32,48,64,96 CPUs @ 2.2-3.6GHz
    - 3 2-port Emulex LPe35002 adapters (6x32G)
    - Ramdisk-based targets
    - SCSI: 2 luns per port
    - NVMe: 2 subsystems per port; 1 NS per Subsystem
  - Switch interconnected
  - FIO and MAIM tools

BROADCOM

OS Issue: I/O Busy/Retry Affects

	1-Port Read IOps (LUNs=28 and CPUs=28)								
	OS1			OS2			OS3		
Total # of out standing IO/LUN	SCSI Block QD=64	SCSI Block QD=128	SCSI Block QD=256	SCSI Block QD=64	SCSI Block QD=128	SCSI Block QD=256	SCSI Block QD=64	SCSI Block QD=128	SCSI Block QD=256
32	3.98 Millions	3.98 Millions	3.98 Millions	3.35 Millions	3.35 Millions	3.35 Millions	2.49 Millions	2.49 Millions	2.49 Millions
64	4.07 Millions	4.07 Millions	4.18 Millions	3.38 Millions	3.38 Millions	3.38 Millions	2.54 Millions	2.54 Millions	2.54 Millions
128	162k	3.82 Millions	3.82 Millions	2.00 Millions	3.10 Millions	3.10 Millions	2.41 Millions	2.41 Millions	2.41 Millions
256	137K	179K	3.56 Millions	1.55 Millions	1.99 Millions	2.98 Millions	2.41 Millions	2.41 Millions	2.41 Millions
512	66K	71K	174K	1.21 Millions	1.34 Millions	972K	2.37 Millions	2.37 Millions	2.37 Millions
1024	70K	70K	181K	1.15 Millions	1.16 Millions	906K	2.37 Millions	2.37 Millions	2.37 Millions

BROADCOM

OS Study and Aid Needed

- Cross-Socket SCSI MQ
- I/O Retry Policies
- Cross-CPU Submission vs Completion

Q & A ?

BROADCOM

Thank You