

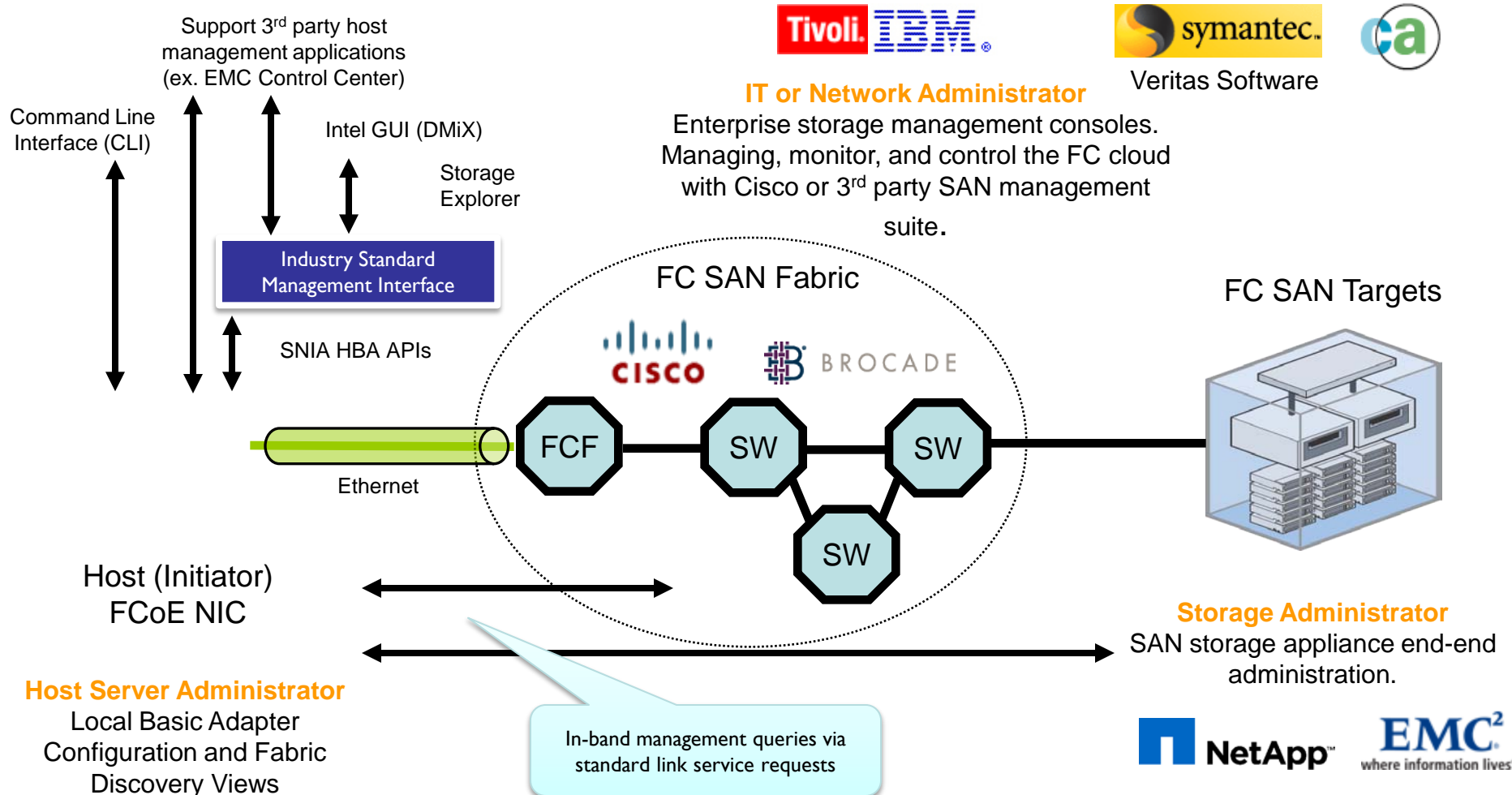
Open-FCoE: Architecture, Management and Performance

Prafulla Deuskar
Intel Corporation

What is Open-FCoE?

- ❑ Open-FCoE project is an implementation of an FCoE initiator. The goal of the project is to have a native FCoE initiator in different operating systems. Not all OS initiators are open-source.
- ❑ Hosted at www.open-fcoe.org
 - ❑ Git repository
 - ❑ Developers mailing list

Open-FCoE: Management



Operating Systems & Drivers

- ✓ Windows: Released Intel initiator w/FCoE Logo
- ✓ Linux: Native in kernel, RHEL6, OEL6 & SLES-SPI
- ✓ ESX: Native in ESX 5.0, passed VMware certs, EMC/NetApp certification ongoing
- ✓ MAC OS & Solaris 11: 3rd party development

Testing, Validation and Reliability

- ✓ Intel validation² complete for Windows, Linux and ESX
 - Application testing ongoing: Exchange, Databases, etc
- ✓ In testing at >14 storage, server, OS & testing partners + Open-FCoE community contributions
- ✓ Demo'ed interop at UNH and SNW on par w/peers
- ✓ 20+ FCoE POC's in process
- ✓ Active in TI0/11, IEEE, EA, FCIA, SNIA, SNW

Storage Systems

- ✓ Over 40 systems ship with 1 & 10G Intel Ethernet
- ✓ FCoE Target: Linux target code completed
 - SanBlaze launched 2 FCoE products with the Intel X520

¹ Denotes native OS FCoE and DCB/DCBx support

² Executed in Intel LAD's world-class Storage Networking Validation Lab

OS Support

OS	Completion
Windows 2008+	Intel Released
SLES-SPI ¹	OSV Released
RHEL6, OEL ¹	OSV Released
ESX 5.0 ¹	OSV Released (vSphere 5)

Storage Qualification

Co.	Windows	Linux	ESX
Netapp	Done	Done	In Progress
EMC	Done	Done	In Progress

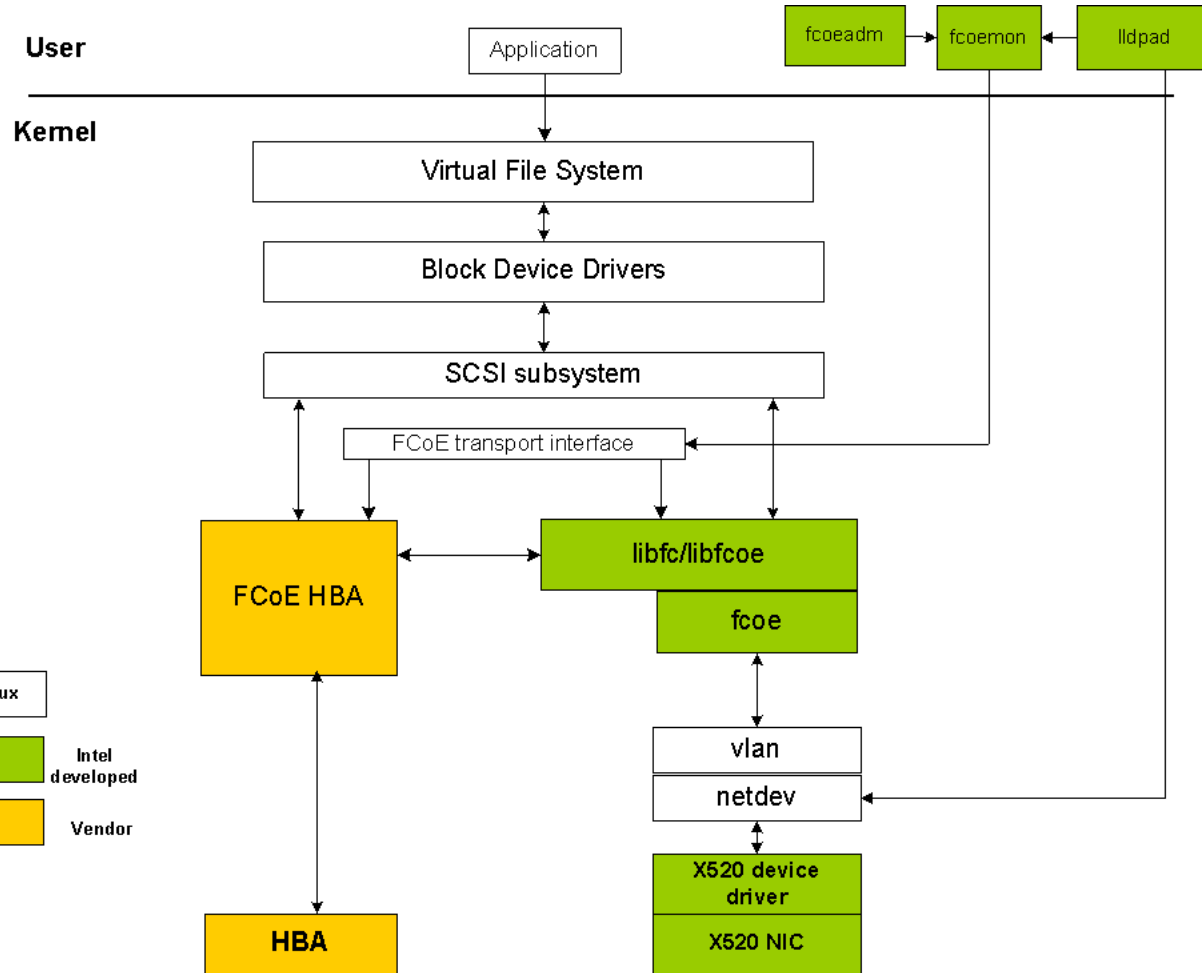
Validated Switches

Vendor	Model
FCoE Switch	
Cisco	Nexus 5010/20
Brocade	8000
Fibre Channel Switch	
Cisco	MDS 9216A, 9124, 9222i
Brocade	Brocade 48/4900

X520 (82599) FCoE Offloads

- ❑ **FC CRC Offload**
 - ❑ On Receive calculates the FC CRC and validates that it matches the FC CRC in the FCoE frame. If the CRC is correct, good CRC status is indicated to Software. Automatically calculates and inserts Ethernet and FC CRC for all transmitted FCoE frames
- ❑ **FC Large Receive Offload**
 - ❑ SW programs a DDP context before it issues FCP read command
 - ❑ DDP context contains – SG list, offset into first buffer, number of buffers, size of each buffer (fixed size)
 - ❑ DDP context is identified by XID, HW supports 512 DDP contexts
 - ❑ HW identifies FC frame sequence and copies data into DDP buffers if frames are in order
- ❑ **FC Large Sequence Offload**
 - ❑ Similar in concept to TCP LSO
 - ❑ SW provides a context and prototype header
 - ❑ SW provides data descriptors to describe data buffers
 - ❑ HW performs segmentation
- ❑ **FCoE Redirection Table**
 - ❑ Indicates ingress FCoE frames on different RX queues based on OXID or RXID (based on whether it is initiator or target mode)
 - ❑ Allows for traffic to be balanced across different cores

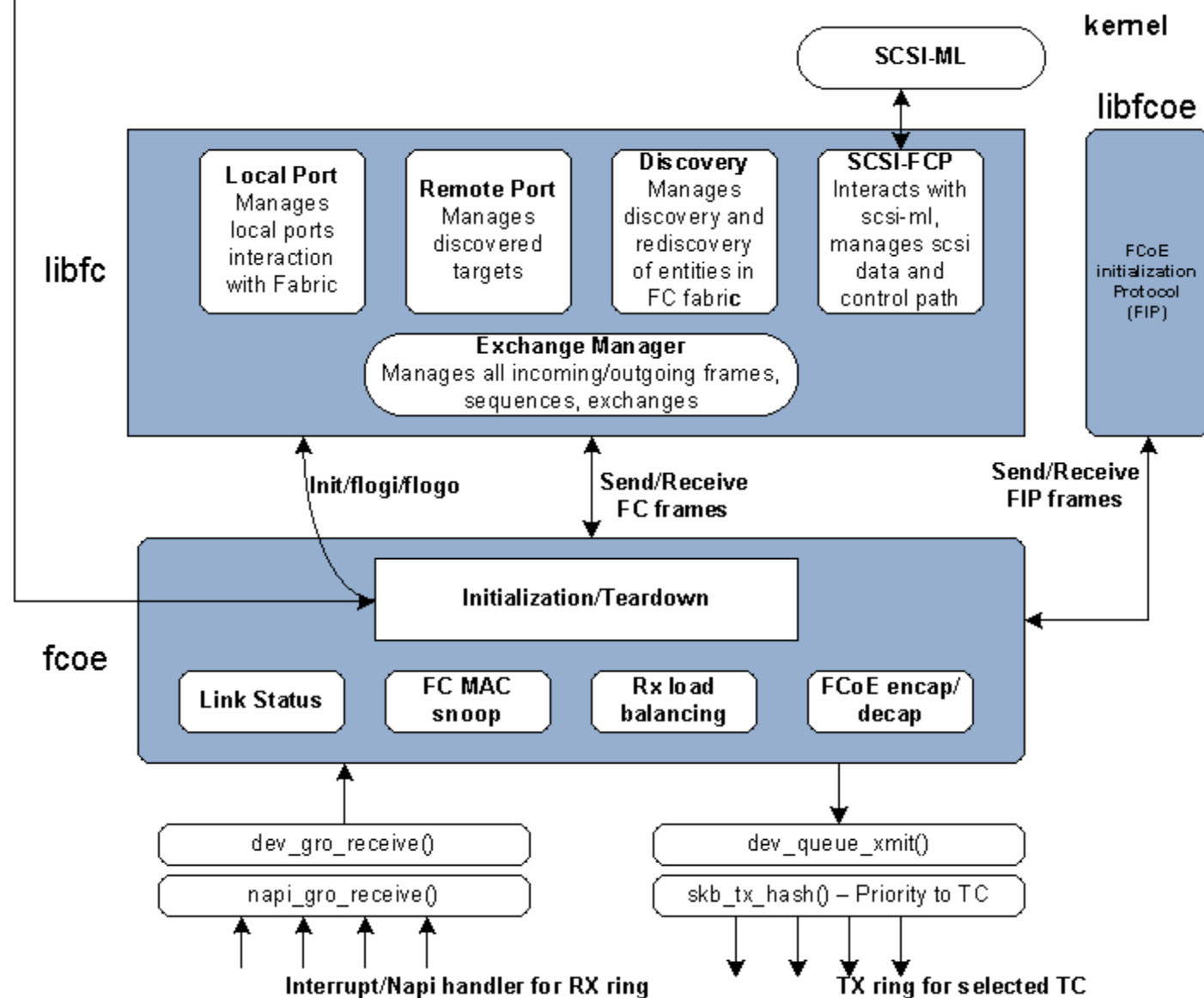
Linux Open-FCoE stack



- libfc – Implements the FC protocol as defined in FC specs - FC-FS, FC-LS, FC-GS, SCSI-FCP
- libfcoe – Implements the FIP as defined in FC-BB-5. Also has support for VN2VN port (point-point, multi-point) which is being defined in FC-BB-6.
- Fcoe – Implements the protocol driver which encaps/decaps FC frame into FCoE and vice versa. Also has handler for FIP frames.
- lldpad – A generic lldp agent daemon for Linux which supports LLDP TLVs for different protocols that run on top of LLDP (DCBX, EVB, EEE etc). Maintained by Intel at www.open-lldp.org.
- fcoemon – Agent which manages FCoE instances in the system – triggered via configuration files, CLI or runtime events
- fcoeadm – CLI for managing (create, destroy, query) FCoE instances in the system

Open-FCoE details

/sys/module/libfc/destroy
/sys/module/libfc/destroy



```
[root@F19 ~]# fipvlan --help
Usage: fipvlan [ options ] [ network interfaces ]
Options:
-a, --auto           Auto select Ethernet interfa
-c, --create         Create system VLAN devices
-s, --start         Start FCoE login automatic
-h, --help          Display this help and exit
-v, --version       Display version information

[root@F19 ~]#
[root@F19 ~]# fipvlan -a
Fibre Channel Forwarders Discovered
interface      | VLAN | FCF MAC
-----
eth2           | 152  | 00:0d:ec:6d:8f:00
eth3           | 152  | 00:0d:ec:6d:8f:00
```

```
[root@F11 ~]# fcoeadm -i eth2.152-fcoe
Description:      82599EB 10-Gigabit SFI/SFP+ Network Connection
Revision:        01
Manufacturer:    Intel Corporation
Serial Number:   001B21909858
Driver:          ixgbe 3.2.9-k2
Number of Ports: 2

Symbolic Name:   fcoe v0.1 over eth2.152-fcoe
OS Device Name:  host4
Node Name:       0x1000001B2190985A
Port Name:       0x2000001B2190985A
FabricName:      0x20980000DEC6D8F01
Speed:           10 Gbit
Supported Speed: 10 Gbit
MaxFrameSize:   2112
FC-ID (Port ID): 0x270027
State:           Online
```

```
[root@F11 ~]# fcoeadm --help
Version 1.0.18
Usage: fcoeadm
      [-c|--create] <ethX>
      [-d|--destroy] <ethX>
      [-r|--reset] <ethX>
      [-S|--Scan] <ethX>
      [-i|--interface] [<ethX>]
      [-t|--target] [<ethX>]
      [-l|--lun] [<ethX>]
      [-s|--stats] <ethX> [<interval>]
      [-v|--version]
      [-h|--help]
```

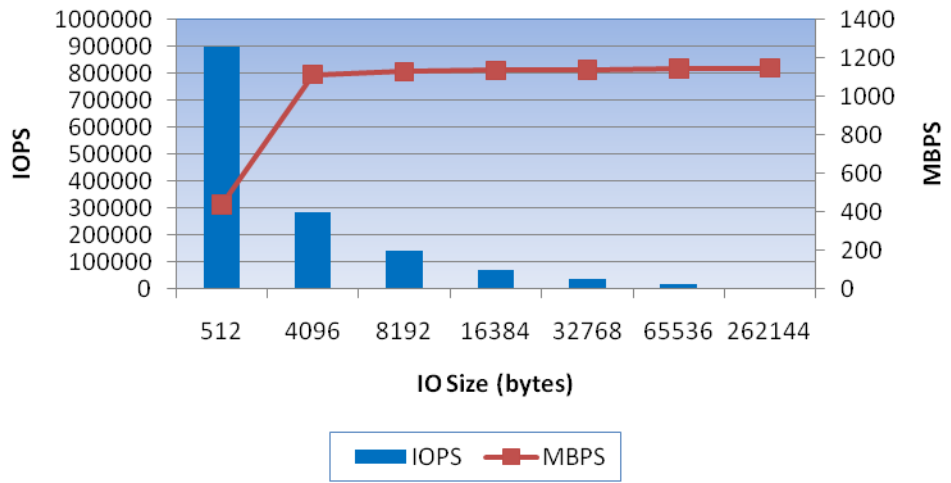
```
[root@F19 ~]# fcoeadm -t | less
Interface:      eth3.152-fcoe
Roles:          FCP Target
Node Name:      0x50060160BB202280
Port Name:      0x500601623B202280
Target ID:      0
MaxFrameSize:  2048
OS Device Name: rport-9:0-1
FC-ID (Port ID): 0x2701EF
State:          Online
```

LUN ID	Device Name	Capacity	Block Size	Description
0	/dev/sdb	8.00 GB	512	DGC RAID 5 (rev 0430)
1	/dev/sdc	8.00 GB	512	DGC RAID 5 (rev 0430)
2	/dev/sdd	8.00 GB	512	DGC RAID 5 (rev 0430)
3	/dev/sde	8.00 GB	512	DGC RAID 5 (rev 0430)

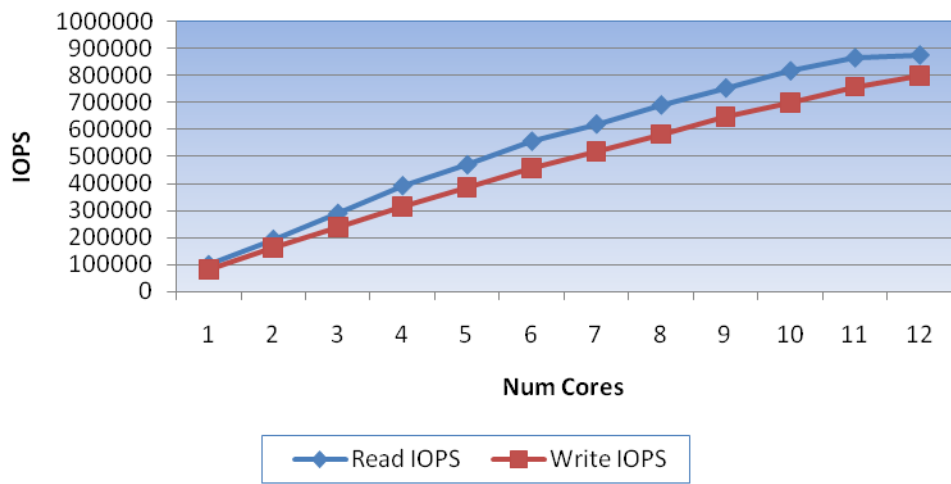
Performance..

- ❑ Eliminate global locks
 - ❑ SCSI mid layer
 - ❑ VLAN driver
 - ❑ Allocate per-cpu data structures and locks in the FCoE stack
- ❑ Make data structures cache aligned
 - ❑ Issue prefetch instructions
- ❑ Numa aware memory allocation(s)
- ❑ Multiple TX and RX queues
 - ❑ Allocate MSI-X vector per queue pair
 - ❑ Assign MSI-X vector to unique core
- ❑ Align command queue/completion to same core
- ❑ Load balance traffic

Read Performance



Core Scaling



Test Configuration

HT Disabled
 SMP fixed IRQ affinity
 DDP Threshold = 4096
 fio iodepth = 64
 noop IO scheduler
 Nomerges = 2 (IO merging completely disabled)
 Sequential IO

SUT: Intel® Thurley

Intel® Xeon® Processor X5680 (12M Cache, 3.33 GHz, 6.40 GT/s Intel® QPI)
 24GB DDR3
 Upstream 2.6.38 + fcoe stack cache alignments + skip vlan layer on tx + per CPU pci pool for DDP + numa mem allocs for tx/rx rings + no low latency interrupt

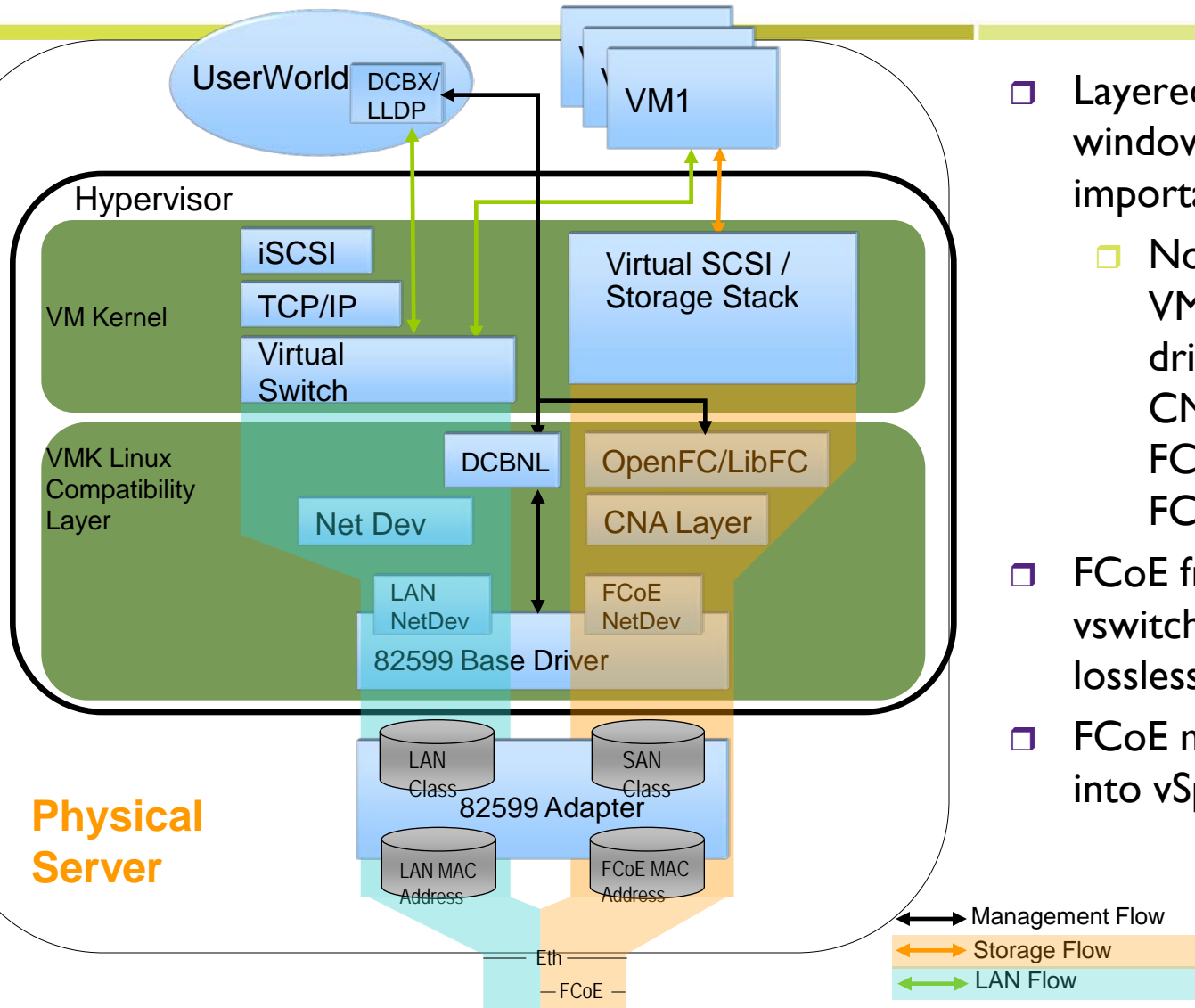
Network Configuration

Brocade* Connectrix* M8000
 X520 connected @ 10Gbps

SanBlaze FCoE soft target:

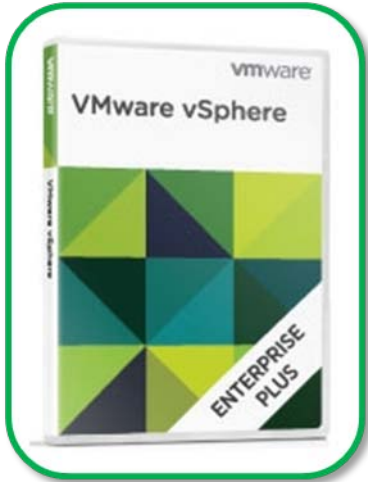
Intel® Xeon® Processor X5680 (12M Cache, 3.33 GHz, 6.40 GT/s Intel® QPI)
 24GB DDR3
 Four 10 Gig X520 ports with each having 16 SANBlaze luns

ESX 5.0 – Open-FCoE Architecture



- Layered model just like windows and linux with important differences
 - No protocol driver in VMKLinux, so the base driver has to assist the CNA layer into forwarding FCoE frames to Open-FCoE stack
- FCoE frames don't go through vswitch, as vswitch is not lossless
- FCoE management integrated into vSphere 5.0 and esxcli

VMware vSphere® 5 with Intel® Corporation FCoE Adapter



The screenshot displays the VMware vSphere Client interface for a virtual machine named "WIN2K364VM". The "Configuration" tab is selected, showing the "Storage Adapters" section. A green box highlights the "Intel Corporation FCoE Adapter" section, which contains two entries:

Device	Type	WWN
vmhba37	Fibre Channel o...	10:00:00:1b:21:90:8f:43 20:00:00:1b:21:90:8f:43
vmhba35	Fibre Channel o...	10:00:00:1b:21:90:8f:42 20:00:00:1b:21:90:8f:42

Below this, the "iSCSI Software Adapter" section shows the "vmhba36" entry:

Device	Type	WWN
vmhba36	iSCSI	iqn.2011-08.com:r7-svr10:

The "Details" section for "vmhba36" shows the following information:

- Model: iSCSI Software Adapter
- iSCSI Name: iqn.2011-08.com:r7-svr10
- iSCSI Alias:
- Connected Targets: 0
- Devices: 0
- Paths: 0

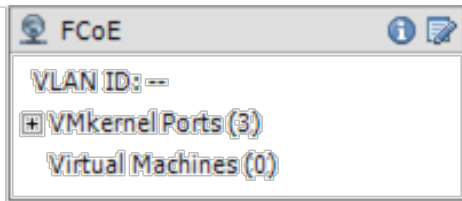
The "View" section shows "Devices" and "Paths" tabs. The "Recent Tasks" section at the bottom shows "License Period: 42 days remaining" and "Administrator".

Intel® FCoE Adapter Configuration**

Use vSphere Client to configure FCoE adapters on Intel® Ethernet Adapter X520

Setup Network switch ports for DCB and FCoE support -- The switch configures DCB on the FCoE adapter

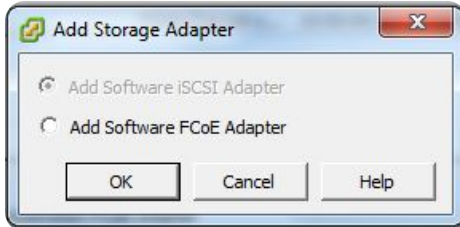
Set Up Networking for FCoE -- Connect the VMkernel to physical NICs that will be used for FCoE for FIP (FCoE initiation protocol) to communicate with the switch using LLDP to setup DCB on the Intel FCoE adapter



Procedure

1. Log in to the vSphere Client, and select a host from the inventory panel.
2. Click the **Configuration** tab and click **Networking**.
3. In the Virtual Switch view, click **Add Networking**.
4. Select **VMkernel** and click **Next**.
5. Select **Create a virtual switch** to create a new vSwitch or **add new port group** to existing
6. Select the network adapter (vmnic#) that supports FCoE and click **Next**.
7. Enter a network label and click **Next**.
8. Specify the IP settings and click **Next**.
9. Review the information and click **Finish**.

Activate Software FCoE Adapters -- You must activate software FCoE adapters



Procedure

1. Log in to the vSphere Client, and select same host from step 2 from the inventory panel.
2. Click the **Configuration** tab and click **Storage Adapters** in the Hardware panel.
3. Click **Add** and select **Software FCoE Adapter**.
4. From the drop-down list of physical network adapters, select the same **vmnic** from step 2.
5. Click **OK**.

Intel® FCoE Adapters are now enabled and ready to use

Storage Adapters			Add...	Remove	Refresh	Rescan All...
Device	Type	WWN				
iSCSI Software Adapter						
vmhba35	iSCSI	iqn.1998-01.com:vmware:R3-SVP5-572d009c				
Intel Corporation FCoE Adapter						
vmhba33	Fibre Channel o...	10:00:00:1b:21:55:1f:4a 20:00:00:1b:21:55:1f:4a				
vmhba34	Fibre Channel o...	10:00:00:1b:21:55:1f:4b 20:00:00:1b:21:55:1f:4b				

2011 Storage Developer Conference. © Intel Corporation All Rights Reserved.

Verify Intel® FCoE Adapter Configuration

Use esxcli in vSphere Management Assistant (vMA) to view Intel FCoE adapters

esxcli fcoe nic list

Port #1 Intel® Ethernet 10Gb X520

Intel FCoE Adapter Enabled

Priority, Source MAC, and VLAN are automatically configured

Port #2 Intel® Ethernet 10Gb X520

Intel FCoE Not Enabled

esxcli fcoe adapter list

Intel FCoE is on vmhba38

Intel FCoE adapter configuration details

```
vi-admin@localhost:~> esxcli -server= 10.0.2.65 fcoe nic list
```

```
Enter username: root
```

```
Enter password:
```

```
vmnic4
```

```
User Priority: 3
```

```
Source: 00:1b:21:69:9e:32
```

```
Active: true
```

```
Priority Settable: false
```

```
Source MAC Settable: false
```

```
VLAN Range Settable: false
```

```
vmnic5
```

```
User Priority: 3
```

```
Source: 00:1b:21:69:9e:33
```

```
Active: false
```

```
Priority Settable: false
```

```
Source MAC Settable: false
```

```
VLAN Range Settable: false
```

```
vi-admin@localhost:~> esxcli -server= 10.0.2.65 fcoe adapter list
```

```
Enter username: root
```

```
Enter password:
```

```
vmhba38
```

```
Source MAC: 00:1b:21:69:9e:32
```

```
FCF MAC:00: 0d:ec:fa:77:40
```

```
VNPort MAC: 0e:fc:00:16:00:00
```

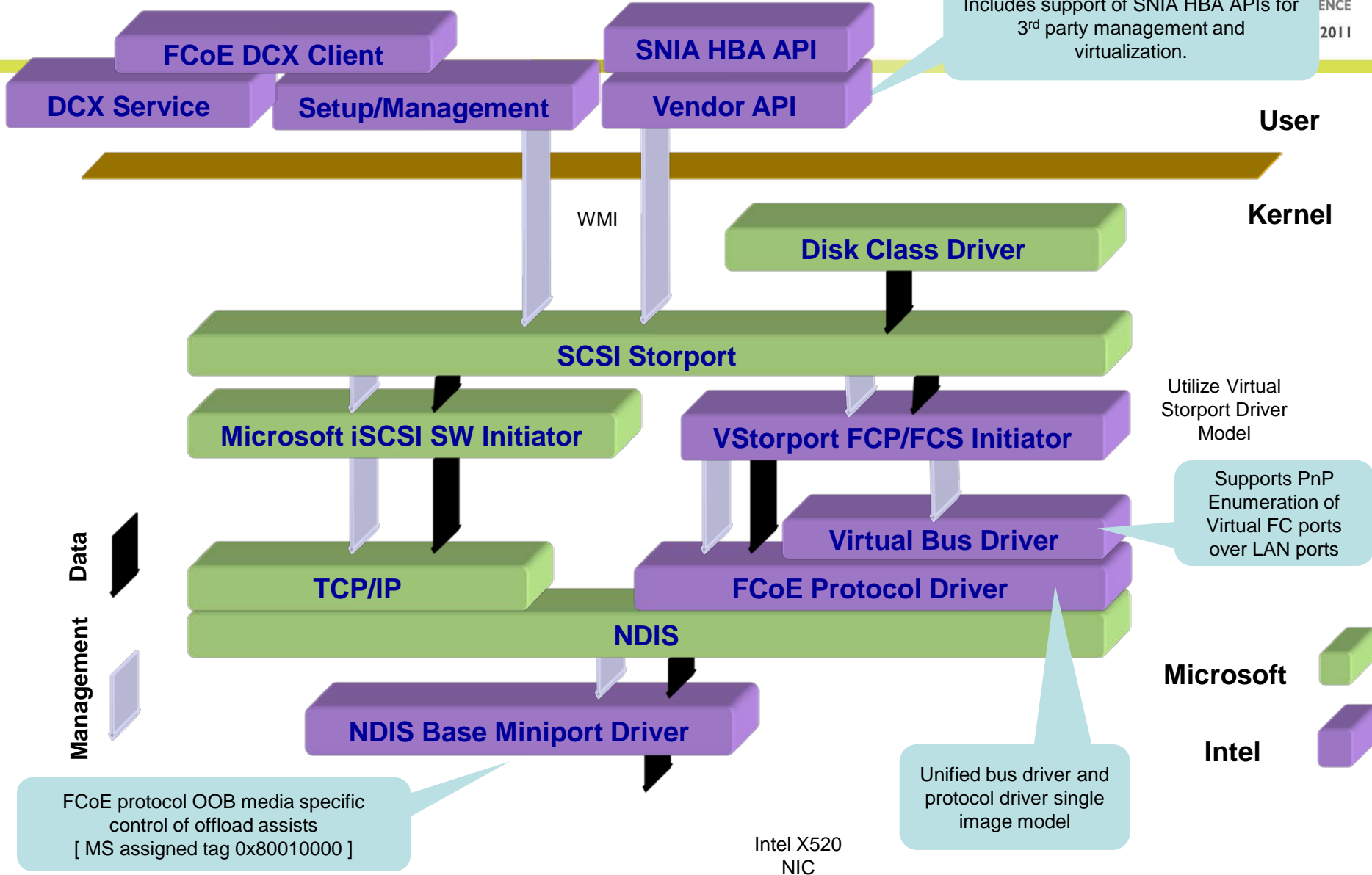
```
Underlying PNIC: vmnic4
```

```
Priority Class: 3
```

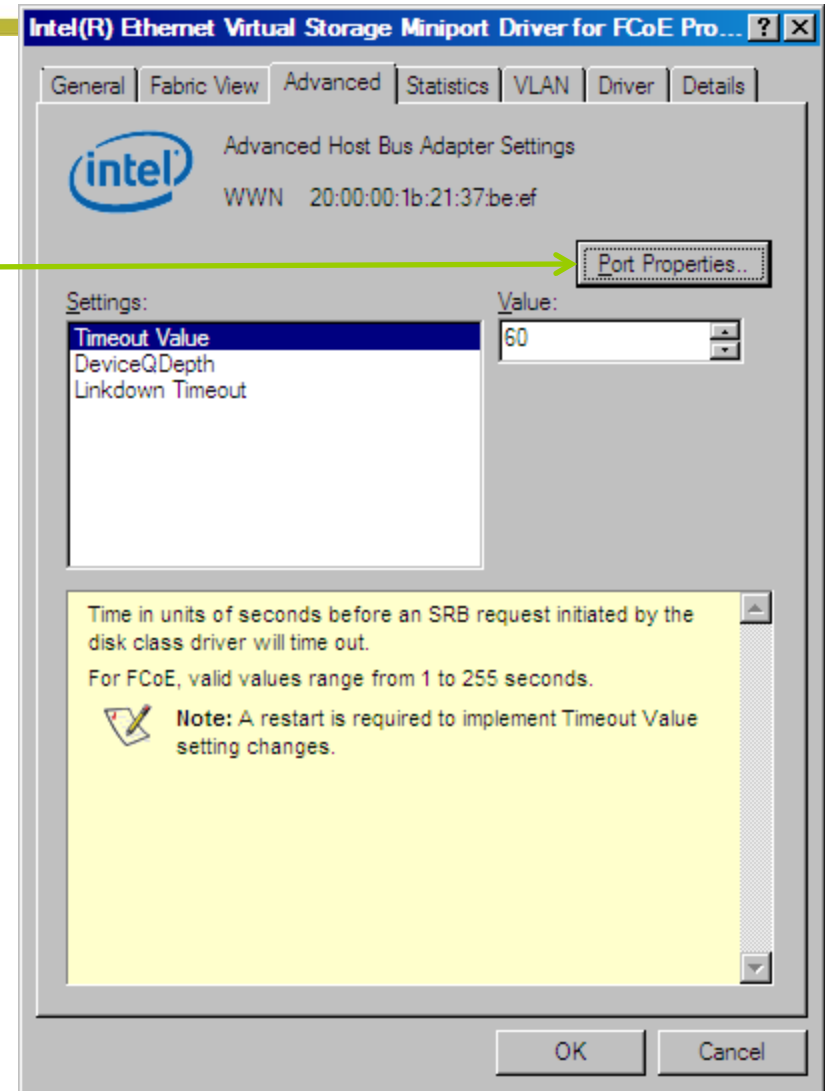
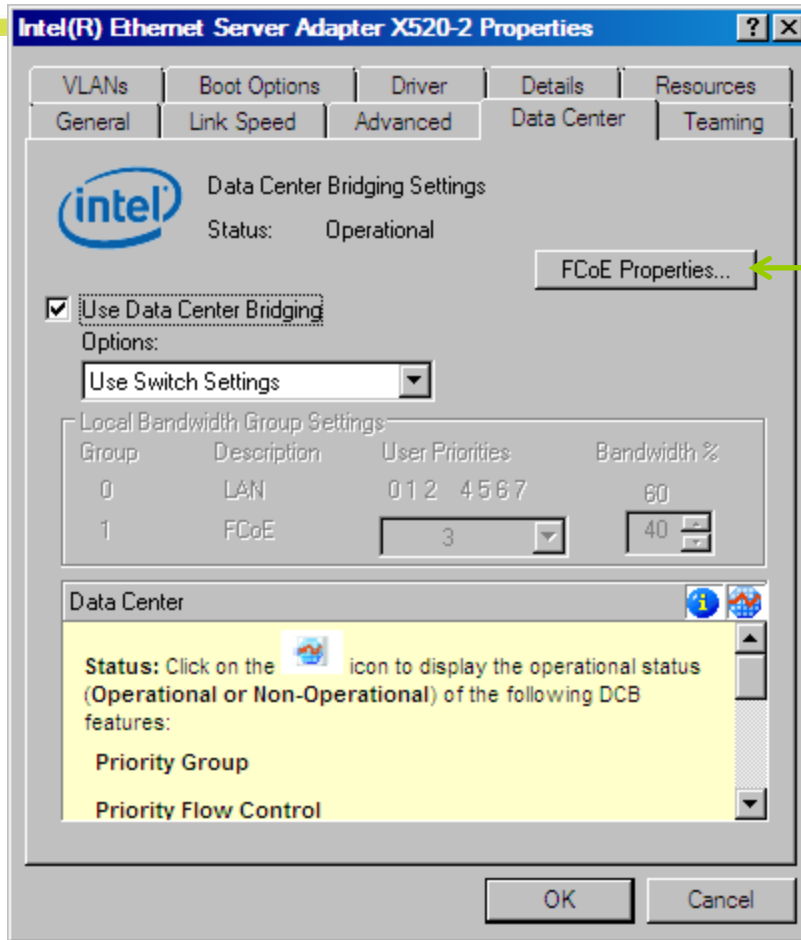
```
VLAN ID: 200
```

```
vi-admin@localhost:~>
```

FCoE Windows SW Architecture



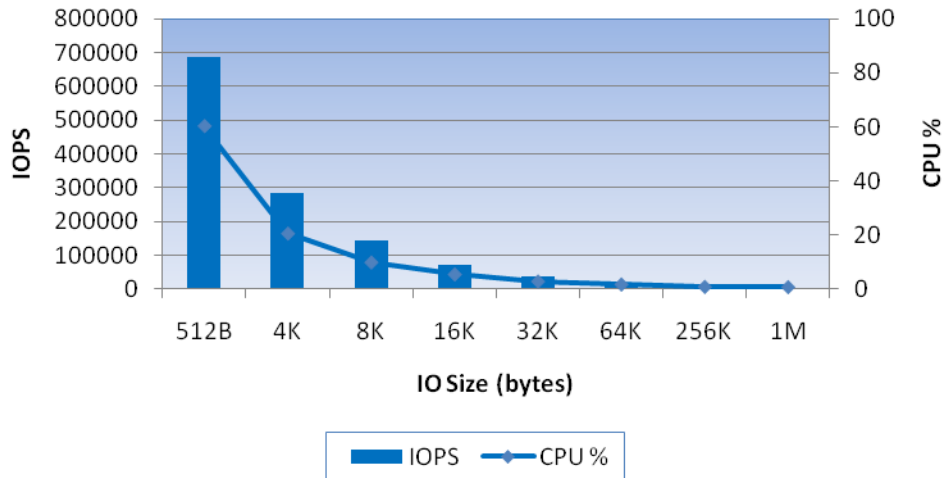
FCoE Management via DMiX



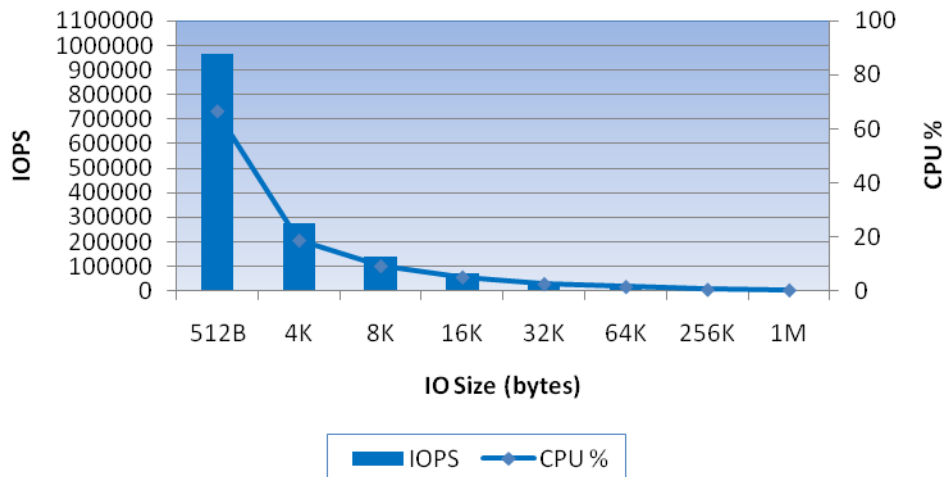
LAN device properties linked to FCoE storage device properties

Windows Open-FCoE Performance (IOMeter)

Write Performance



Read Performance



Test Configuration

Iometer v. 2006.7.27
 No. of Ports Tested = 1
 No. of Managers = 1
 No. of Workers = 16
 No. of luns = 96
 No. of Outstanding I/Os = 5
 Target Ports Used = 8

SUT: Intel® Thurley(CRB)

Intel® Xeon® Processor X5680 (12M Cache, 3.33 GHz, 6.40 GT/s
 Intel® QPI)
 24GB DDR3
 BIOS tx0093
 Windows® Server 2008 R2 x86_64

Network Configuration

Brocade® Connectrix® M8000
 82599EB Connected @ 10Gbps

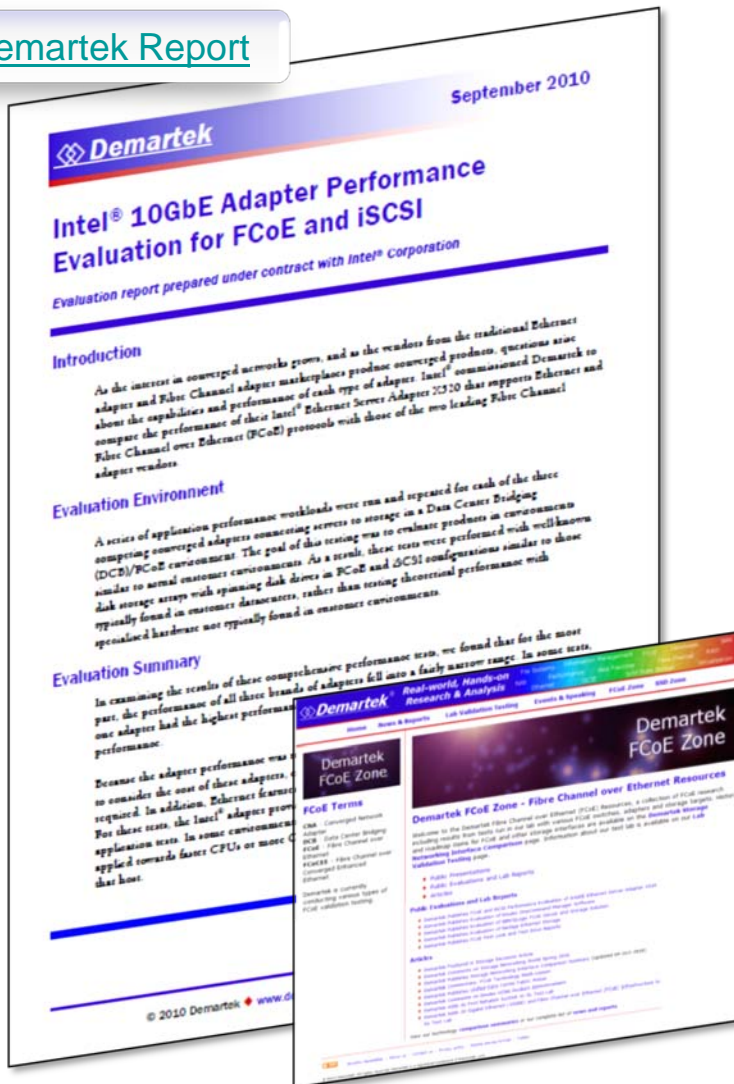
2 Ramsan® 400s:

4 4Gbps FC ports per Ramsan, total 8 ports
 12 luns per port
 Ram disks of 300 MB each

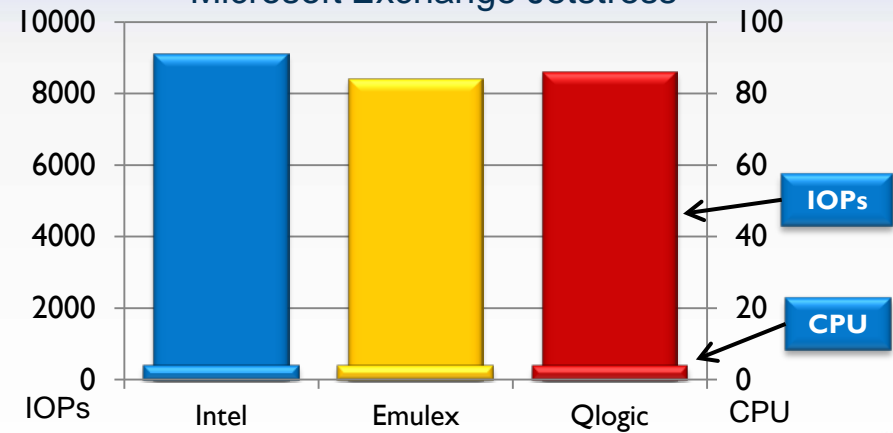
Unified Networking: Application Tests

No performance advantage for HBA, CNA, or Full Offloads

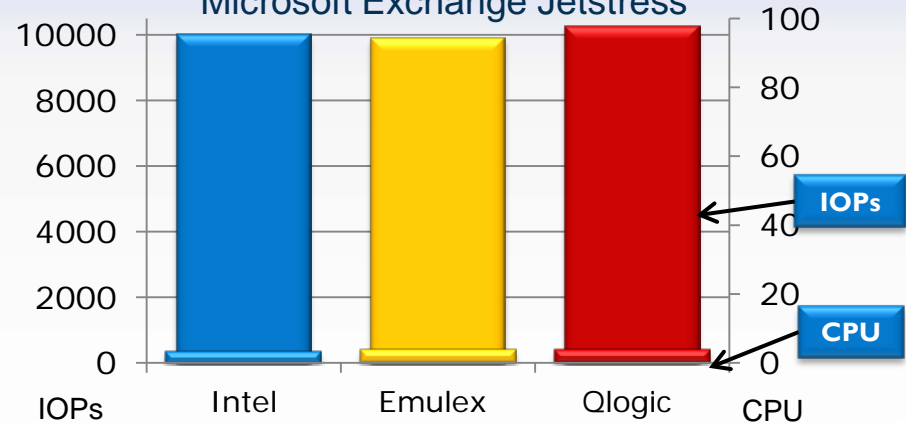
Demartek Report



IOPs and CPU Utilization for iSCSI** Microsoft Exchange Jetstress



IOPs and CPU Utilization for FCoE** Microsoft Exchange Jetstress



Yahoo* POC Results: VNX7500 Beta Report

Presented at EMC World by Ruiping Sun
Principal Database Architect, Yahoo

SDC 
STORAGE DEVELOPER CONFERENCE
SNIA ■ SANTA CLARA, 2011

YAHOO!

EMC²



Intel® Server Adapter X520-DA2



Open-FCoE



Provided by EMC Corp and Yahoo Corp. Originally presented at EMC World

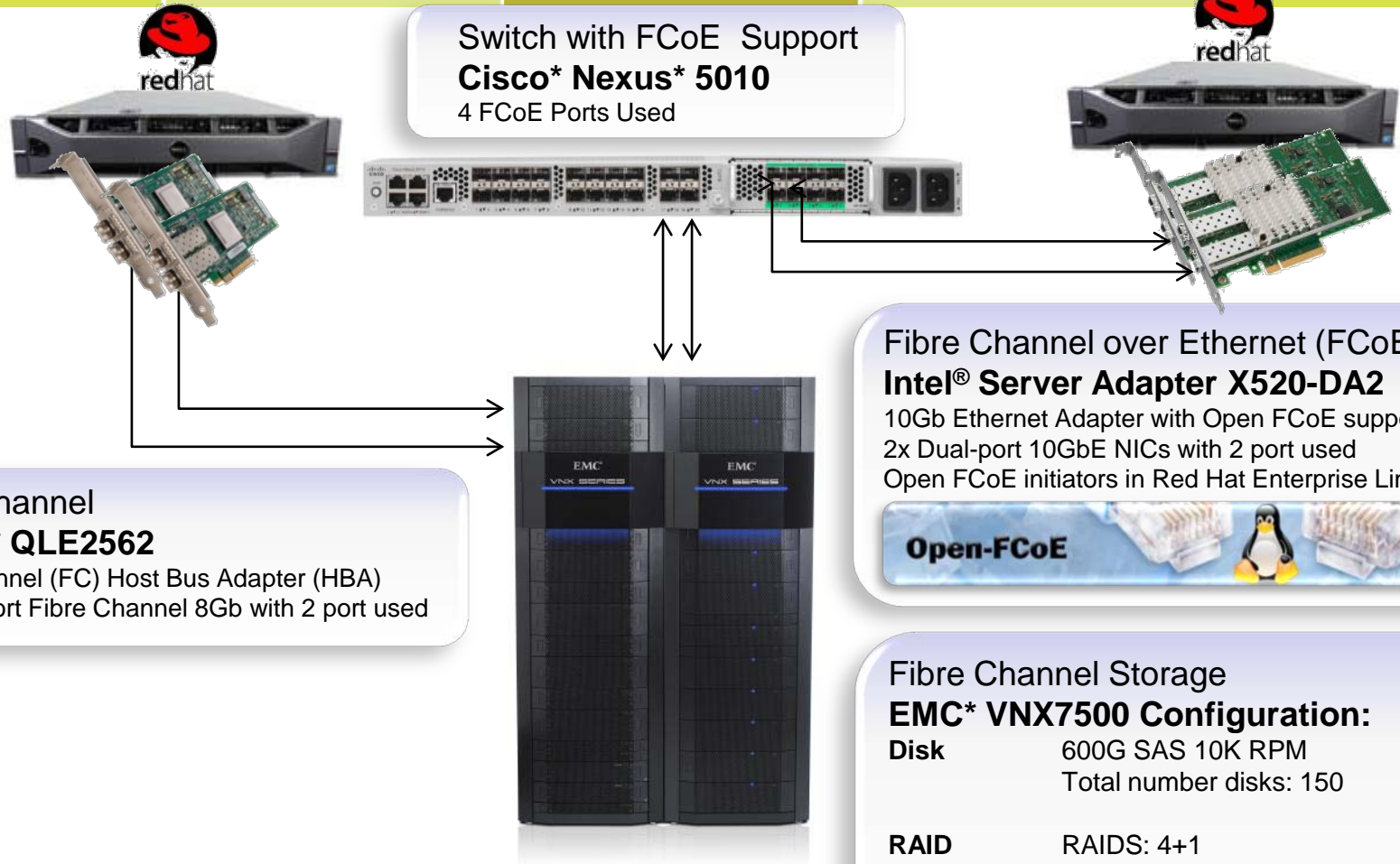
Data Warehouse/Mart IO Profile

	Production DB1	Production DB2
DB Type	Data Mart	Data Warehouse
Allocated Storage	100T	250T
DB Size	10T	1P (uncompressed)
Total Reading IO Per Day	24T	160T
Total Writing IO Per Day	6T	30T
Peak Reading	1GB/sec	4GB/sec
Peak Writing	500MB/sec	2GB/sec
ELT	Direct load and Merging	Direct load
Query	Parallel and Lookup	Mainly Parallel




Provided by EMC Corp and Yahoo Corp

The FCoE Infrastructure



Fibre Channel
Qlogic* QLE2562
 Fibre Channel (FC) Host Bus Adapter (HBA)
 2x Dual-port Fibre Channel 8Gb with 2 port used

Fibre Channel over Ethernet (FCoE)
Intel® Server Adapter X520-DA2
 10Gb Ethernet Adapter with Open FCoE support
 2x Dual-port 10GbE NICs with 2 port used
 Open FCoE initiators in Red Hat Enterprise Linux 6.0

Open-FCoE 

Fibre Channel Storage
EMC* VNX7500 Configuration:

Disk	600G SAS 10K RPM Total number disks: 150
RAID	RAIDS: 4+1 Total Raid Group: 30
LUN	1 LUN (2T) per Raid group Total LUN: 30 (2T/LUN)

Report: FCoE VS FC

	2 x 8Gb FC	10Gb FCOE
Sequential Read (120 Threads, 512k block)	1.56GB/sec	1.84GB/sec
Sequential Write (120 Threads, 512K block)	1.55GB/sec	1.73GB/sec
Random Read (240 Threads, 32K block)	20K IOPS	20K IOPS
Random Write (240 Threads, 32K block)	13K IOPS	13K IOPS

- ❑ Random IO is about the same
 - ❑ Transport capacity is not the bottom neck
- ❑ Sequential IO throughput is improved 15%
 - ❑ Transport capacity is the bottle neck



Provided by EMC Corp and Yahoo Corp

Report: FCoE VS FC

	2 x 8Gb FC	10Gb FCOE
MAX IO Throughput	800MB/sec Per port (per direction)	920MB/sec Per port (per direction)
Total CPU consumption (Server side, worst case)	30% of 1 core	45% of 1 core

- ❑ 15% IO transport capacity improvement over FC
- ❑ FCoE consume more CPU than FC
 - ❑ It is negligible



Provided by EMC Corp and Yahoo Corp

Conclusion

- ❑ FCoE SW initiators are here to stay!
- ❑ Performance on par with HW offloaded solutions in real world benchmarks

Backup

Intel® Unified Management

Turning on FCoE

Installation:

- ❑ FCoE Stack: Installed as part of the OS (if native) or Intel® 10GbE adapter
- ❑ MAC address: Dedicated for LAN and FCoE

Initialization:

- ❑ DCBx: detects DCBx switch (link partner) and exchange DCBx states
- ❑ Initiator: queries base driver if link partner supports FCoE

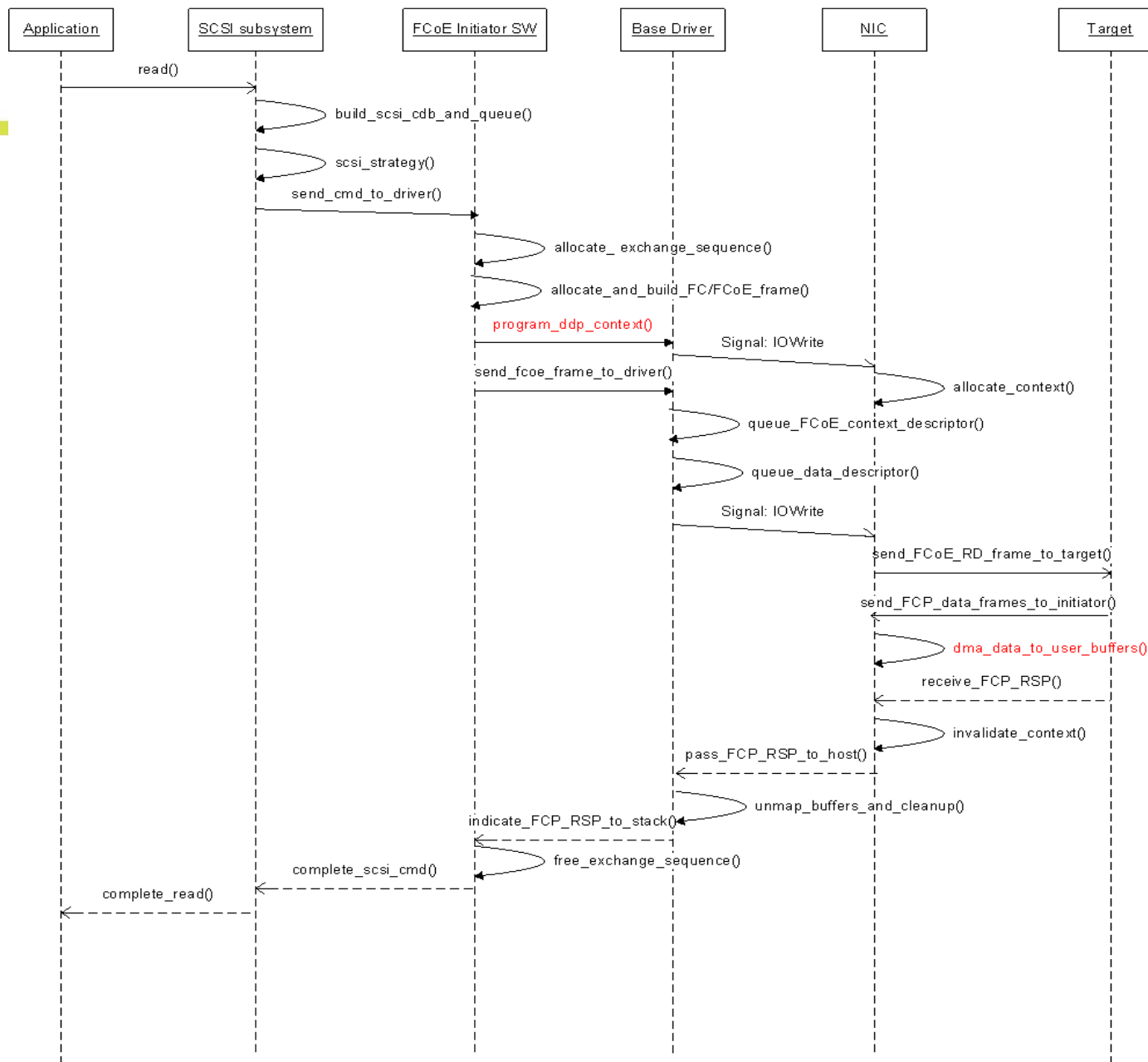
Discovery (FIP):

- ❑ Initiator: Automatically initializes and advertises as an FCoE initiator and seeks an FC switch and FCoE VLANs
- ❑ Switch: Switch advertises via FIP existence as an FC switch and FCoE VLANs
- ❑ Initiator: Collects the advertisements, selects correct VLAN and identifies master FCF switch for log in

Run time:

- ❑ VLANs: LAN and FCoE co-exist on separate VLANs:
 - ❑ FCF Switches support VLAN to VSAN mapping
- ❑ SAN: Will be seen as a FC node and respond to any in band FC mgmt request
- ❑ Teaming: Support host side teaming with MPIO.
 - ❑ Dependent on switches to support switch teaming with MPIO
- ❑ DCB: Provide no-drop behavior and allocate bandwidth

SCSI Read



SCSI Write

