The Evolution of iSCSI

Fred Knight, NetApp
Andy Banta, SolidFire, Now part of NetApp
@andybanta

May 24th, 2016

Today's Presenters

SNIA Ethernet Storage Forum

Andy Banta
Storage Janitor
SolidFire/NetApp

Fred Knight
Standards Technologist
NetApp

David Fair
Chair, SNIA-ESF
Intel Corp.

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

The Journey to Modern iSCSI

- A brief history of iSCSI
- How iSCSI works
  - The basics
  - What makes iSCSI attractive today
- IETF enhancements to iSCSI
- Enhancing iSCSI performance with iSER and RDMA Ethernet

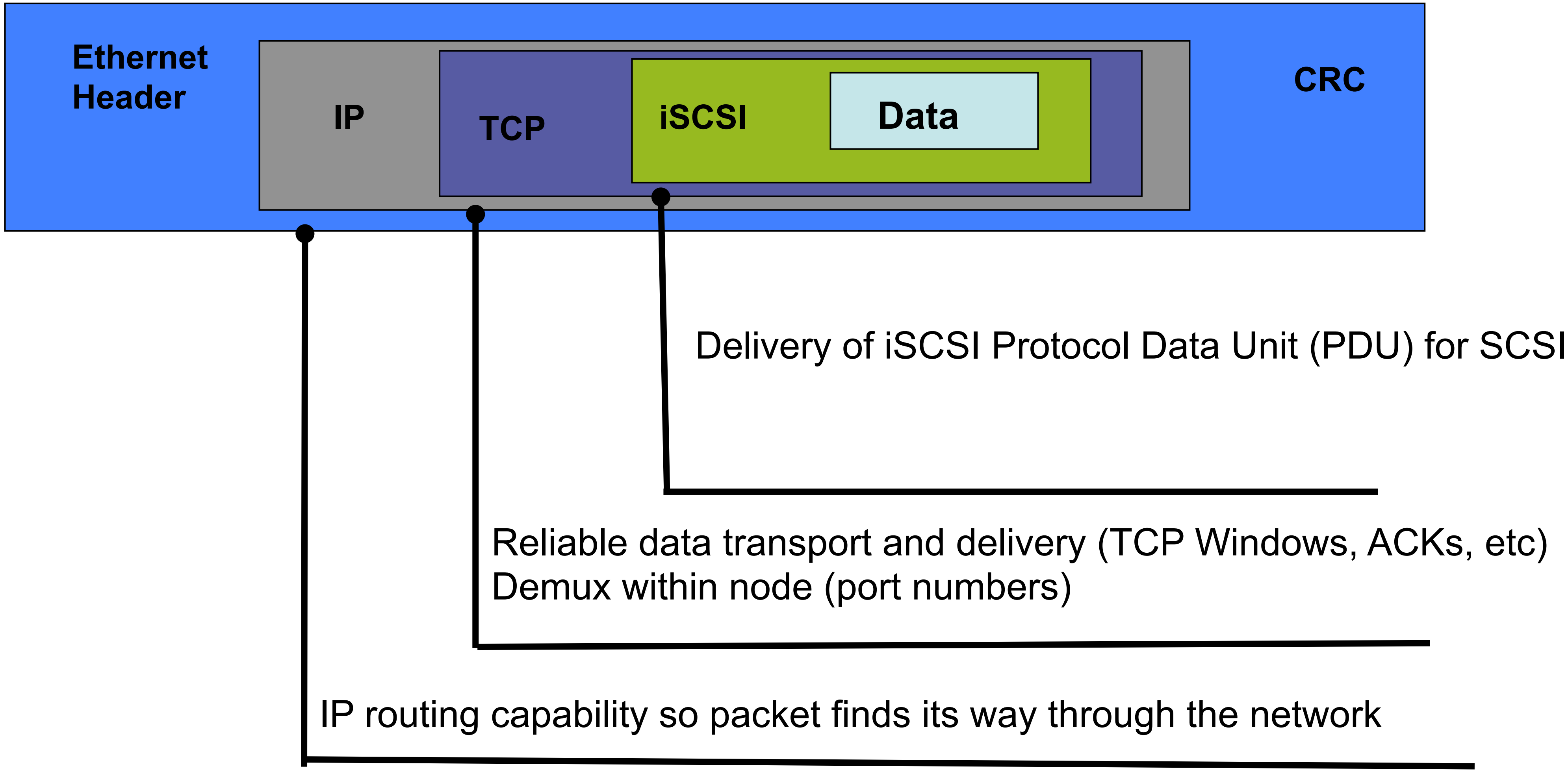
iSCSI History

SNIA ESF | ETHERNET STORAGE

- SCSI originally designed for communication with storage
- Fibre Channel drove separation of device vs. transport
  - Commands for devices vs. how the commands are transmitted
- iSCSI approach: Use TCP/IP as basis for transport
  - Only mainstream SCSI transport that does no hardware definition
- Defined in IETF RFC 3720 in 2002-2004 (& RFC 7143 in 2014)
  - By IBM, Cisco, HP and others
- Defined block storage over a standard (TCP/IP) network

How iSCSI works

- Encapsulates SCSI commands in TCP/IP packet



SCSI Concepts

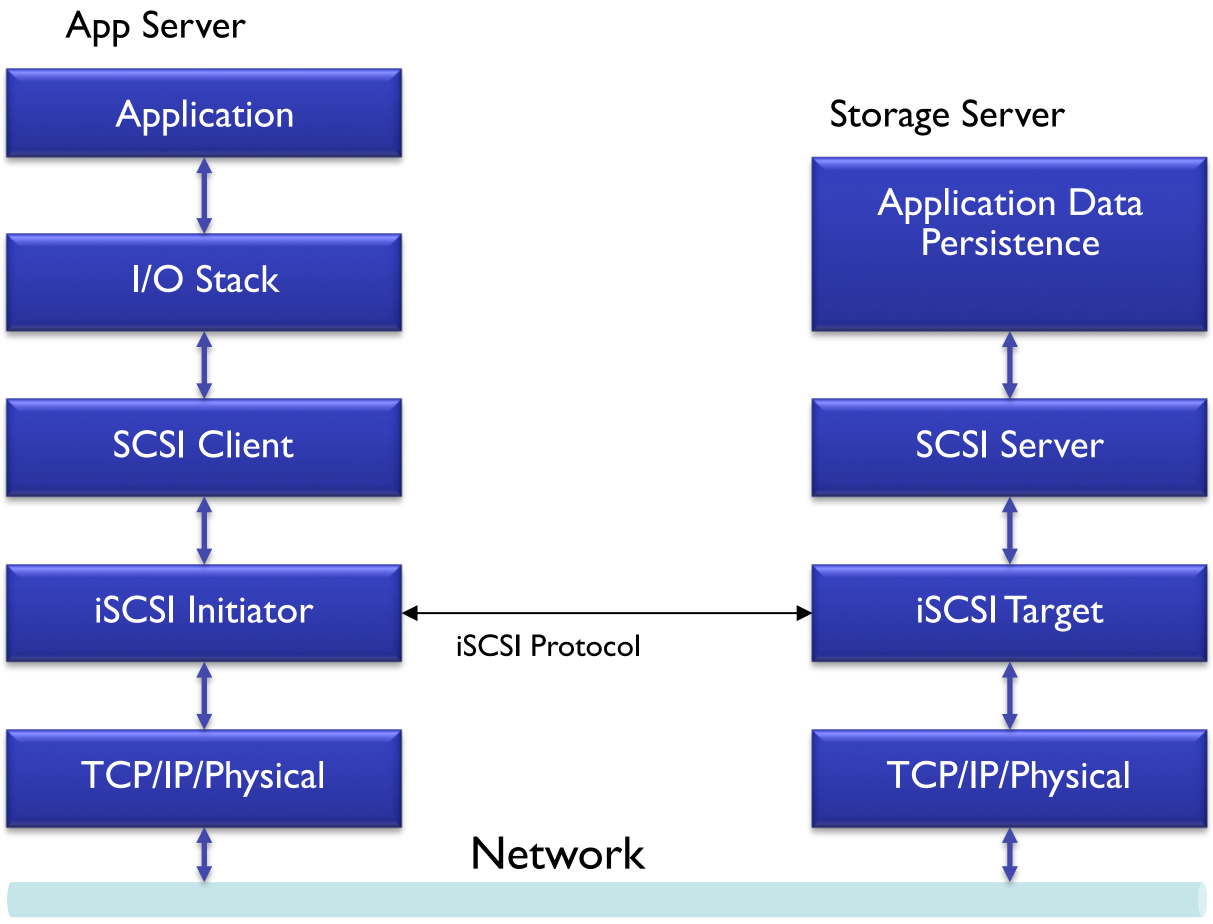
- Request / Response protocol
  - There can be no response until there is a request
  - INITIATORS are where requests are created
  - TARGETS are where requests are serviced and responses created
- SCSI INITIATORS are usually hosts
  - Compute equipment, like servers or workstations
  - But hosts can also be targets
  - iSCSI initiators can maintain multiple parallel connections to multiple targets

iSCSI Concepts

- SCSI TARGETS are usually storage devices
  - But storage devices can also be initiators
  - iSCSI Targets can maintain multiple parallel connections to multiple targets
- iSCSI defines several identifiers to enable this:
  - iSCSI names
  - iSCSI initiator session identifiers (ISID)
  - iSCSI connection identifiers (CID)
  - iSCSI target portals (TPGT)

iSCSI?

- iSCSI is a client-server SCSI transport protocol
  - iSCSI can run on any physical network that TCP/IP can run on – Ethernet, InfiniBand,..
- Any type of SCSI device can be accessed over iSCSI
  - Block Storage is the most typical (and the only supported on Windows Server)
- Original protocol spec is RFC 3720
  - RFC 5048 corrects/clarifies the original
  - RFC 7143 replaces the original

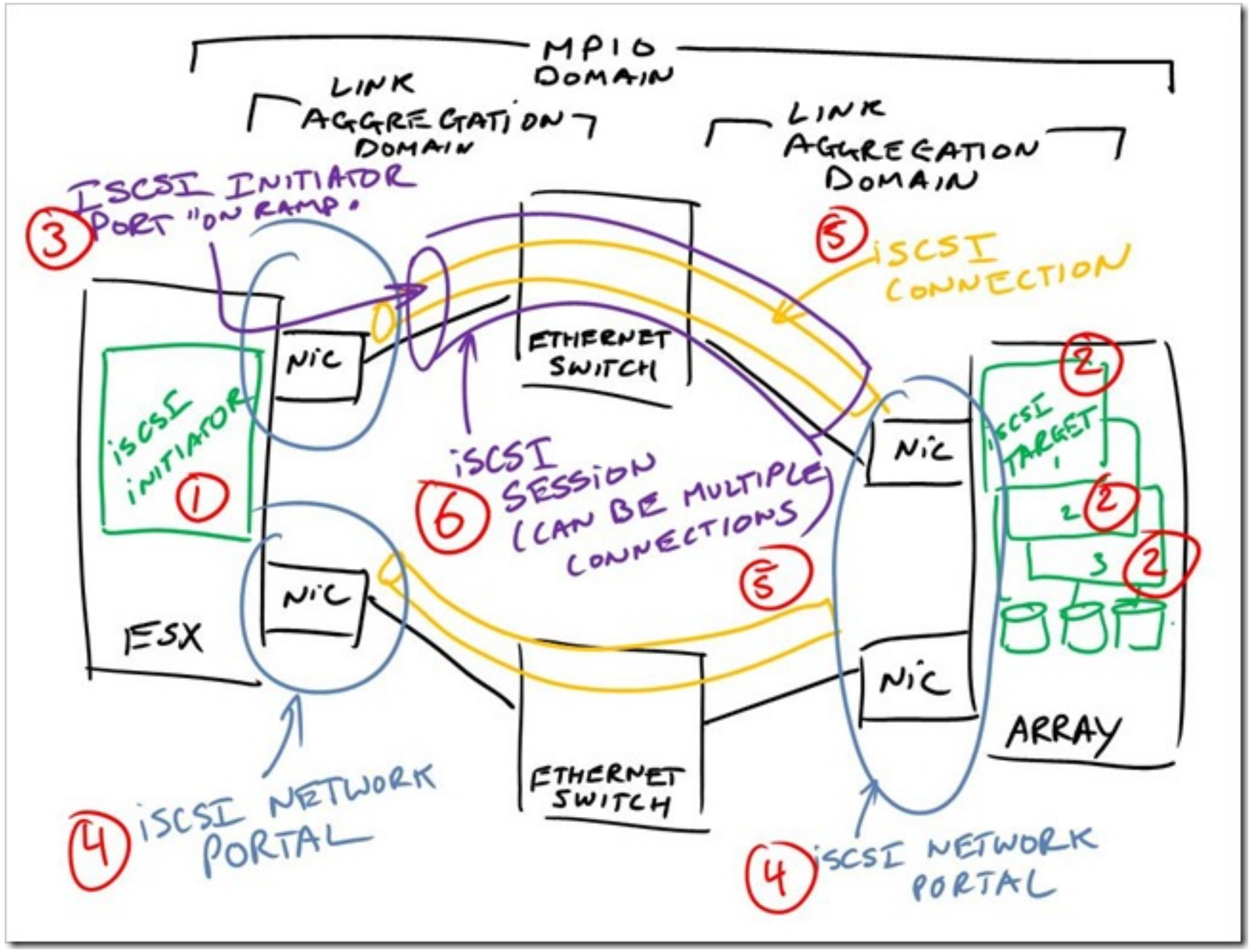


iSCSI Names

- iSCSI names are globally unique
- iSCSI nodes have names
  - Similar to Node World Wide Names (WWN) in Fibre Channel
- iSCSI Qualified Name
  - Example name: iqn.1997-11.org.snia:hedgetrimmer-1926184
  - Used everywhere
- EUI iSCSI Name
  - EUI is the IEEE EUI-64 format (Extended Unique Identifier)
  - EUI-48 and EUI-64 define first 24 bits as the Company ID
  - Example name: eui.02004567A425678D
  - Used few places
- iSCSI Alias
  - “Friendly,” internally viewable, only

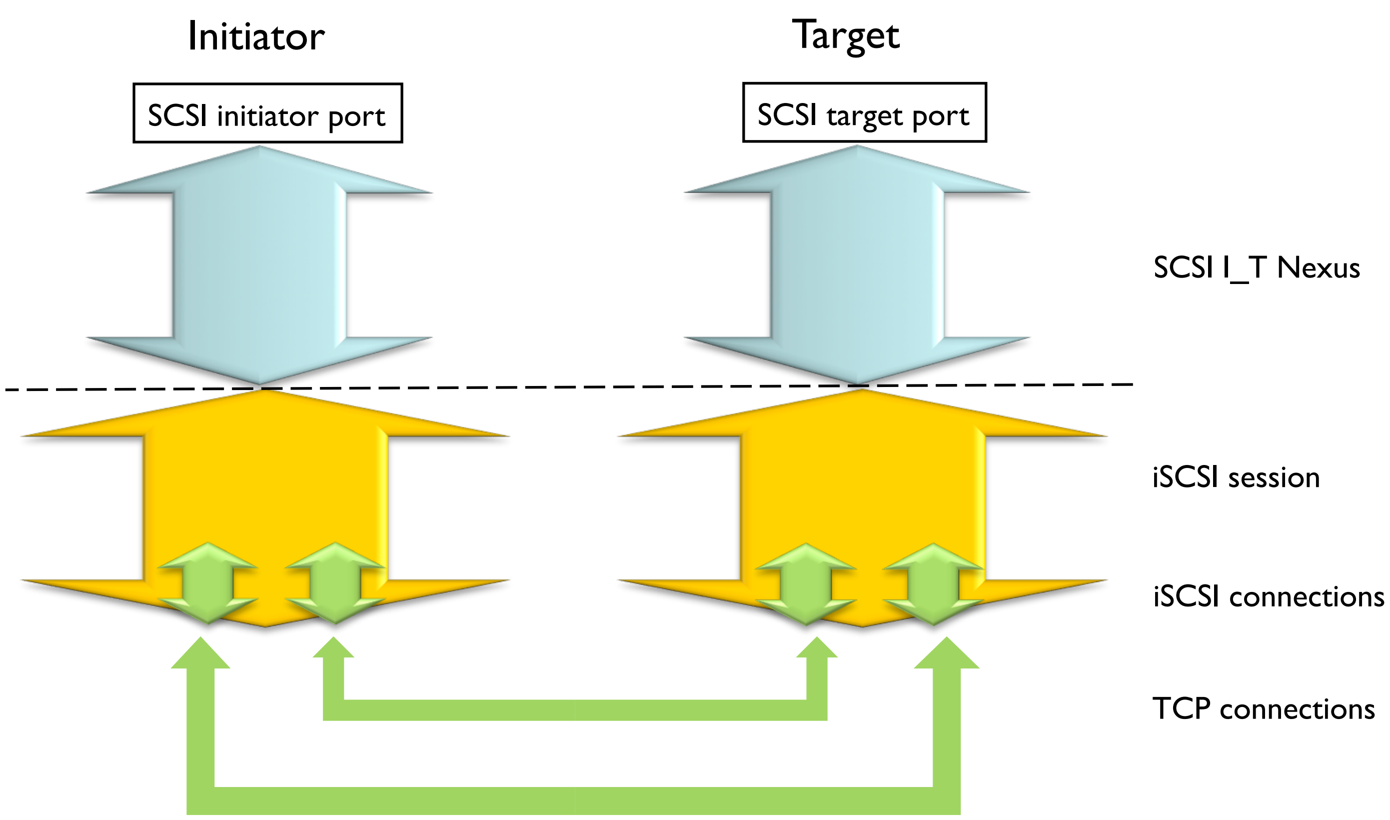
Sessions, Connections and Target Portals

1. Initiator
2. Target
3. Initiator Port
4. iSCSI Network Portal
5. iSCSI Session
6. iSCSI Connection



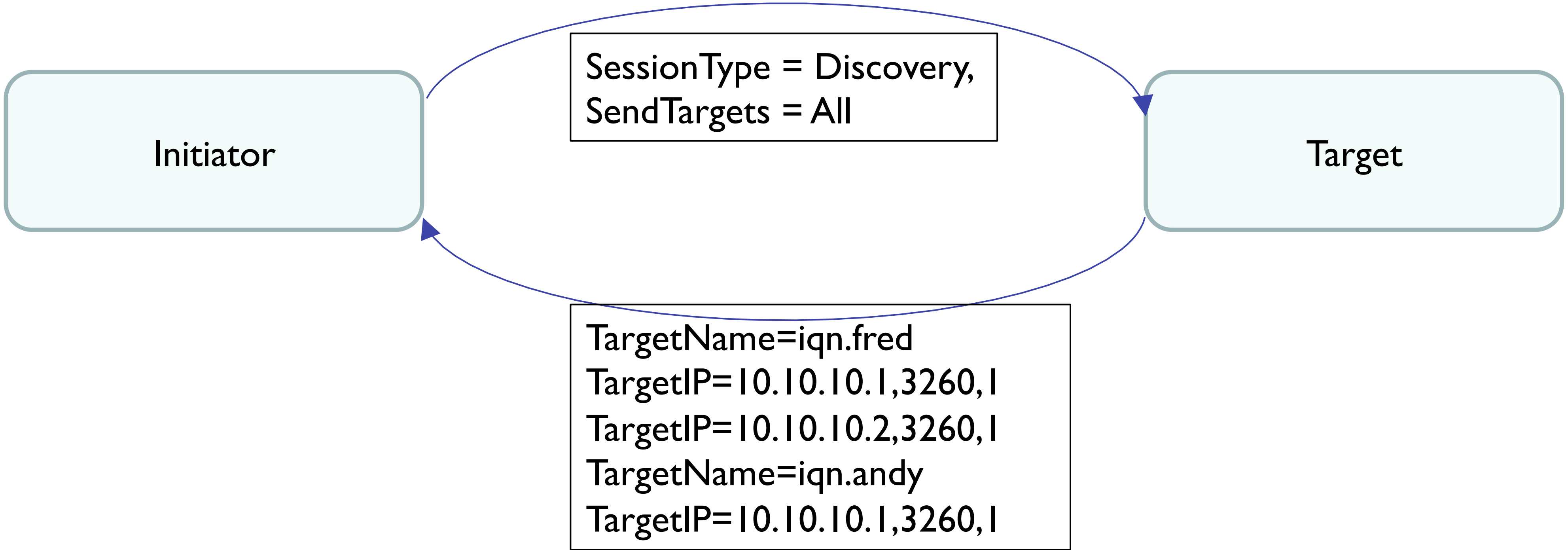
I_T nexus & multi-connection sessions

- iSCSI has native protocol support for combining multiple reliable transport connections into a single iSCSI session
  - "Connection allegiance" for each I/O
  - Scaling throughput with multiple NICs
  - Load balancing and connection failure resiliency for I/Os in progress
- iSCSI is a "SCSI transport protocol"
  - iSCSI in turn relies on a *different transport protocol* unrelated to SCSI semantics



iSCSI Discovery

- **Static Discovery**
  - Connect using IP address and iSCSI name
- **Dynamic Discovery via SendTargets**
  - Requests all iSCSI names and IP address

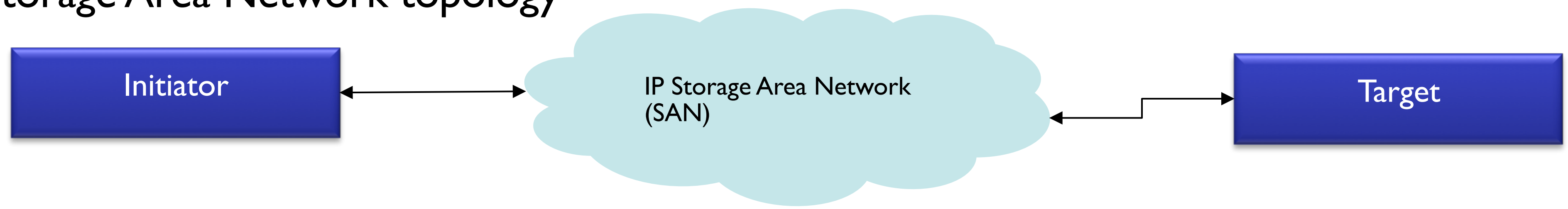


iSCSI Discovery, Continued

- iSNS (Internet Storage Name Service)
  - Managed by an outside server
  - Similar to DNS
  - Can provide callbacks for changes
  - Rarely used, because storage normally isn't that dynamic
- Authentication
  - IP masking
  - iSCSI name masking
  - Challenge-Handshake Authentication Protocol (CHAP)
  - Target authenticates the initiator
  - Initiator can authenticate the target

Terms: Initiator, SAN, Target, Session

Storage Area Network topology



How communication occurs between a target and an initiator

Discovery of target
- Static Configuration
- Internet Storage Name Service (iSNS)-based discovery

TCP connection open
- To the discovered iSCSI Target portal

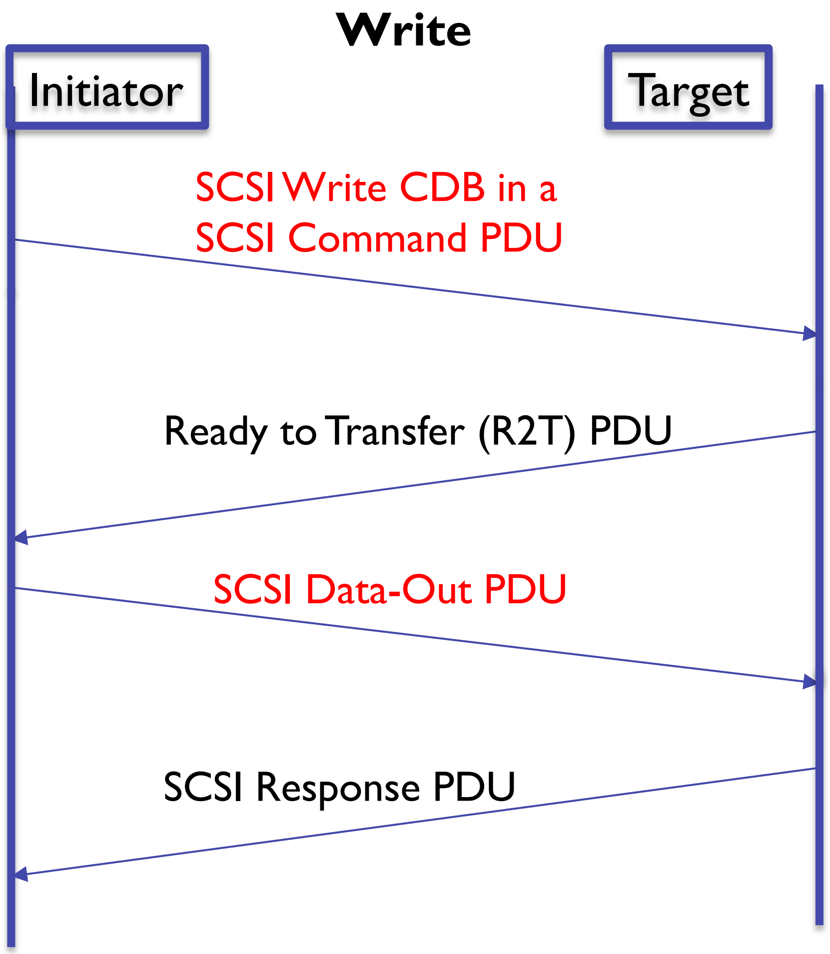
iSCSI Session login
- Security Negotiation
- Operational Negotiation
- Full Feature Phase
- "I_T nexus"

SCSI I/Os
- Starting with INQUIRY and REPORT LUNS
- Read, Write, Data transfer, Response

Session Logout
- Ends the "I_T nexus"
- None of the Logical Units (LUs) can be accessed

TCP Connection close
- Gracefully (TCP FIN) or ungracefully (TCP RST)

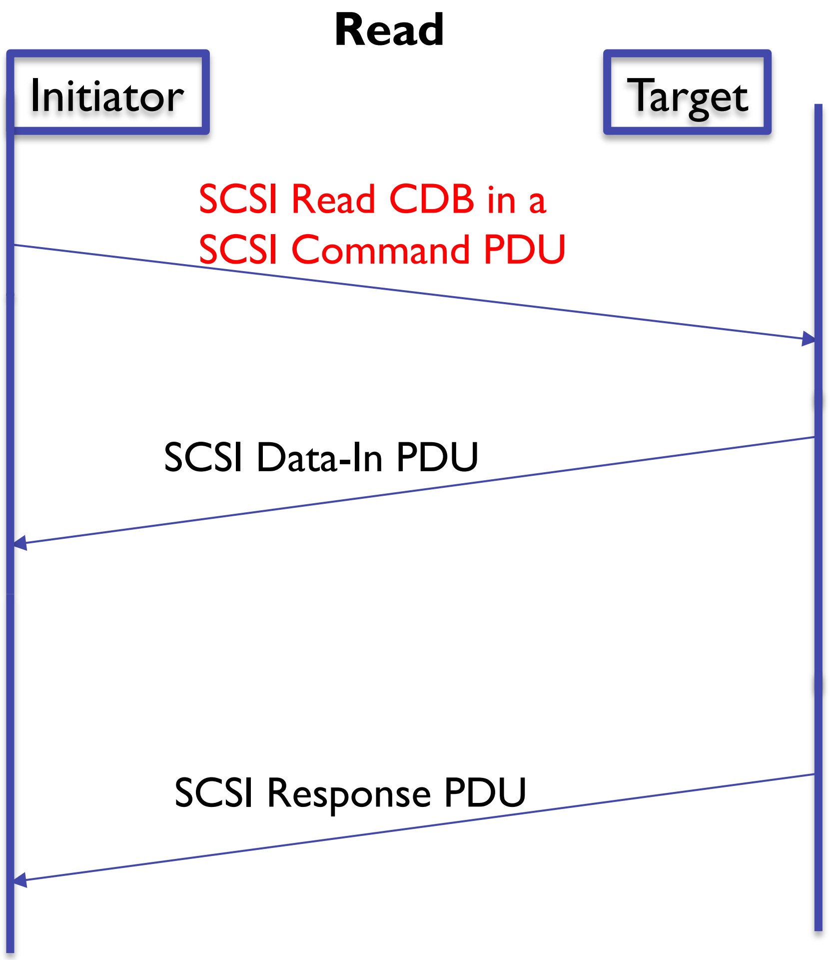
SCSI I/O Mapping onto iSCSI transport



Don't forget async events that can happen any time….

- Connection drops
- Session drops
- PDU drops
- Application timeouts
- Initiator cluster failovers
- SCSI errors
- Implementation code bugs
- …

iSCSI spec includes prescriptive Error Recovery semantics – including formal negotiation support for inter-operation between initiator and target implementations with asymmetric capabilities



How iSCSI Changed the Storage Game

- Transport over commodity, established protocols
- Login redirection
- Data readiness modifications
- Software Initiators rule the day

Commodity Transport

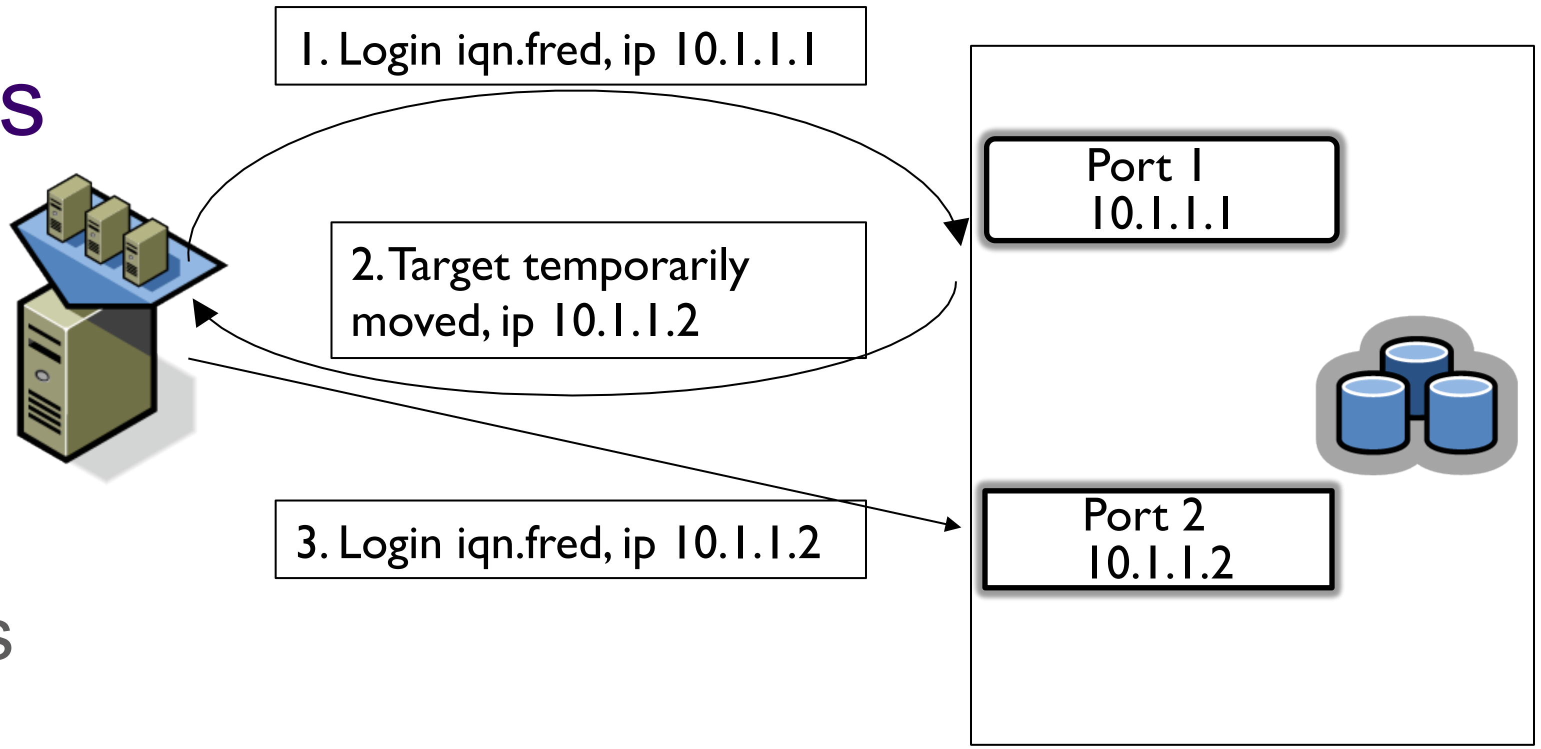
- iSCSI defined encapsulation method, nothing else
  - Fibre Channel, SAS define Layers 0-2
- Every win for Ethernet is a win for iSCSI
  - 10Mbit to 100Gbit and beyond with the same standard
  - VLANs provide ZONE fencing that Fibre Channel wrestled with
- Ethernet NIC technology helps out seamlessly
  - LRO, TSO, TOE apply
  - Data Center Bridging added a class for iSCSI, iSCSI didn't need to add it
- TCP/IP FTW!
  - Resiliency out of the box, with multi-layer flow-control
  - Security (firewalls), In-flight encryption (IPSEC), Routing (you name it)

Data Center Bridging

- Make use of FCoE features for iSCSI
  - iSCSI given its own class
  - Slight performance gain, provide consistent bandwidth
- Priority-based Flow Control – IEEE 802.1Qbb
  - “Lossless” – really less loss
- Enhanced Transmission Selection IEEE 802.1Qaz
  - Assignment of bandwidth for classes
- Congestion Notification – IEEE 802.1Qau
  - End-to-end congestion management
- DCBX IEEE 802.1AB
  - Negotiation of DCB parameters

iSCSI Login Redirection

- iSCSI can temporarily reroute sessions
  - Load balancing
  - Ability to scale out
  - Self healing
  - Rolling upgrades



Data and Status Readiness Improvements

- Eliminates some of the end-to-end traffic for coordination
- Immediate data
  - Allows data to be sent with write commands
  - Size negotiated at login
  - Especially valuable with small writes
- Initial ready to transmit (IntialR2T)
  - First write data PDU to be sent immediately after the command
- Read response included with data
  - An OK status can be returned with the last data PDU for a read

Software Defined iSCSI

- The final benefit of being commodity-based
- Initial push for hardware iSCSI HBAs was short-lived
- Multicore CPUs meant the host could be the initiator
  - Additional cores far cheaper than HBAs
- Scalability comes with core speed and count
- Provides virtual systems with SAN support
  - No need to hand off complete HBA
  - No need to emulate iSCSI HBA
- Supported in every mainstream host base-OS
  - MS, Linux, VMware, Solaris

IETF Enhancements for iSCSI

SNIA ESF | ETHERNET STORAGE

- New RFCs
- PDU and SCSI Enhancements
- Task Management Updates
- Key Changes

RFC 7143 & 7144: Goals

- 7143: iSCSI spec consolidation
  - Goal: pulling together about half a dozen older RFCs into one coherent spec, making "minor" modifications to improve interop, and obsoleting a few specific unimplemented features
- 7144: SAM-5 compliance of iSCSI
  - Goal: Extending iSCSI protocol to be a SAM-5-compliant storage transport protocol, negotiable at a session granularity, fully compatible

SCSI PDU Updates

- Command Priority
  - An IN argument to the SAM-5 Execute Command () procedure call model
  - Indicates the relative scheduling importance of this task in comparison to other SIMPLE tasks
  - SCSI Command PDU addition (4-bits)
- Status Qualifier
  - An OUT argument to the SAM-5 Execute Command () procedure call model
  - Status qualifier provides additional information about the reason for the status code
  - SCSI Response PDU addition (2 bytes)

Sense Data

- Allowance for sense data
  - Typically, Sense Data is in DataSegment if the status is CHECK CONDITION
  - New draft explicitly allows Sense Data to be present anytime, independent of status

Task Management Function Additions

SNIA ESF | ETHERNET STORAGE

- Following new Reason Codes are now allowed in an iSCSI TMF Request PDU
  - `QUERY TASK (9)`: is the Referenced Task Tag present in the task set?
  - `QUERY TASK SET (10)`: is there a task from "my" I_T_L nexus in the task set?
  - `I_T NEXUS RESET (11)`: perform an I_T nexus loss function for all LUs accessible via "my" I_T nexus
  - `QUERY ASYNCHRONOUS EVENT(12)`: is there a unit attention condition or a deferred error pending for "my" I_T_L nexus?

- New TMF Response "Function succeeded" (equivalent to the FUNCTION SUCCEEDED SAM-4 service response)

iSCSIProtocolLevel

- New session-scoped (LO) text key
- iSCSIProtocolLevel negotiation decides the iSCSI protocol features that may be used on the session
- Plan is that each new standards-track RFC with protocol features will “claim” a new value
- Higher negotiated value → implicit support for lower numbered values
- Current legal values
  - 0: no version claimed
  - 1: iSCSI Consolidated RFC compliance (7143)
  - 2: iSCSI SAM-4 RFC compliance (7144)

Key Change List…

SNIA ESF | ETHERNET STORAGE

1. Consolidates RFCs 3720, 3980, 4850 and 5048, and made the necessary editorial changes
2. Claims a value for the new iSCSIProtocolLevel
3. Removes Markers and related keys
4. Removes SPKM authentication and related keys
5. Explicitly allows initiator+target implementations, including the composite device naming
6. Clarifies that SLP-based discovery cannot be relied on for interoperability
7. Specifies formal protocol artifact relationships via UML diagrams

Key Change List… (contd)

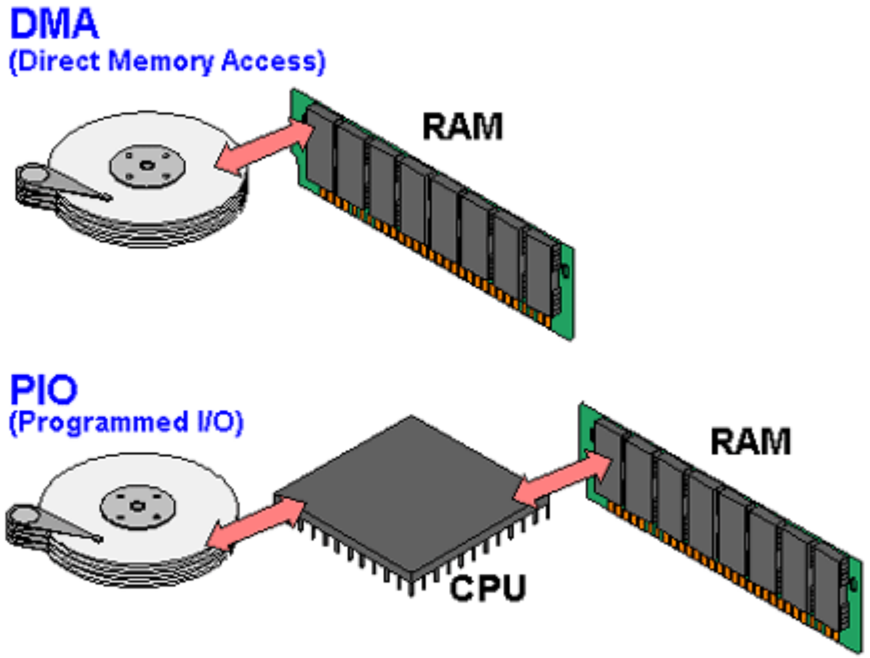
8. Makes FastAbort implementation a "SHOULD" from the previous "MUST"
9. Requires implementing IPsec, 2400-series RFCs (IPsec v2, IKEv1); and SHOULD implement IPsec, 4300-series RFCs (IPsec v3, IKEv2).
10. Restricts the usage of X#, Y# and Z# name prefixes
11. Provides guidance on minimal number of text negotiation responses
12. Provides guidance on Kerberos authentication, OCSP usage, and extended sequence numbers (ESNs)

iSCSI Extensions for RDMA

- What is iSCSI with RDMA
- How it's defined
- Read and Write Examples
- Performance comparisons

RDMA. Another 4-Letter Abbreviation?

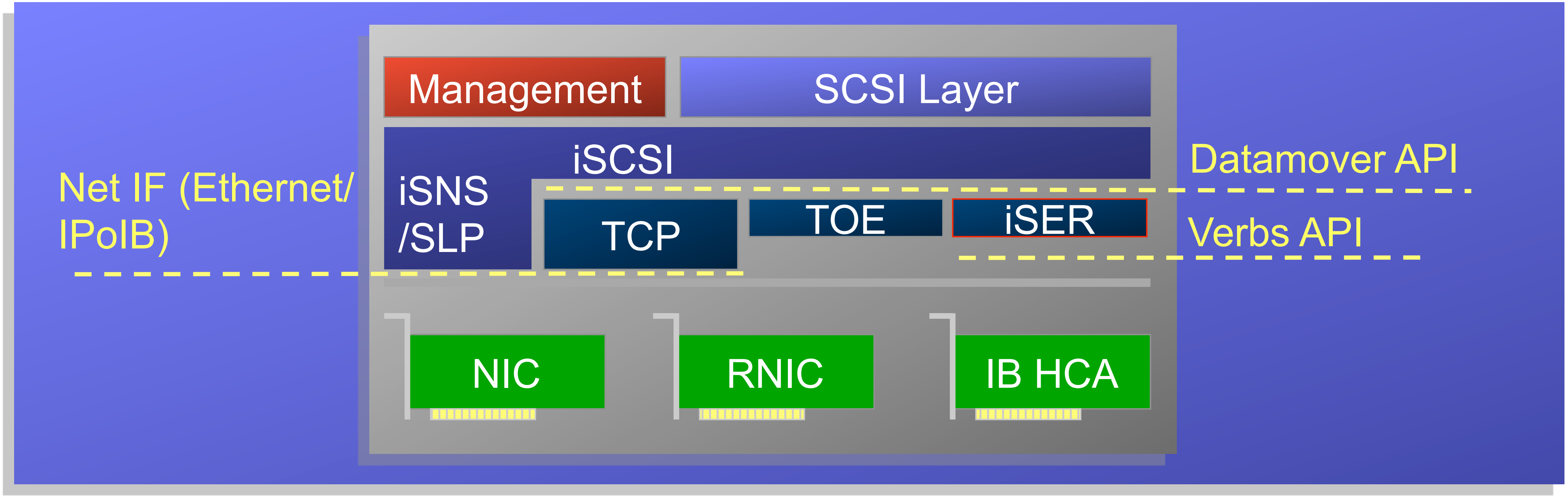
- Remote Direct Memory Access extends DMA across a network
- Refresher: Yeah, I've heard of DMA
- DMA allows a device to directly read or write host memory
  - No copy from device to host memory
  - No context switch
  - Offloads host CPU



Remote DMA

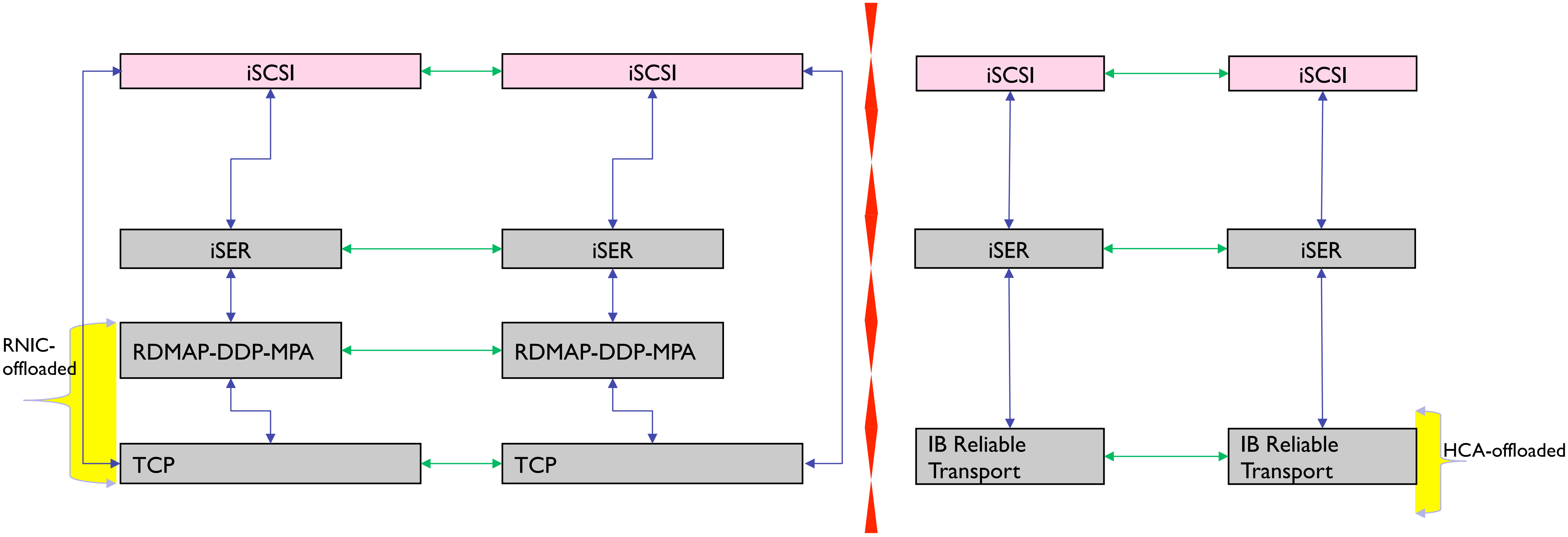
- Extends the concept of DMA to remote devices
- Eliminates the inevitable read/copy when receiving from a network
- Small but measurable win with traditional I/O
- New protocols added to traditional transports
  - iSCSI extensions to TCP/IP for RDMA (iWARP)
  - RDMA over converged Ethernet (RoCE, “rocky”)

iSER is an Open Standard



- IETF Standard RFC-7145
- Choice of iSER or iSCSI TCP transparent to the application and user
- Runs on top of InfiniBand, iWARP, and RoCE

RDMA with iSCSI



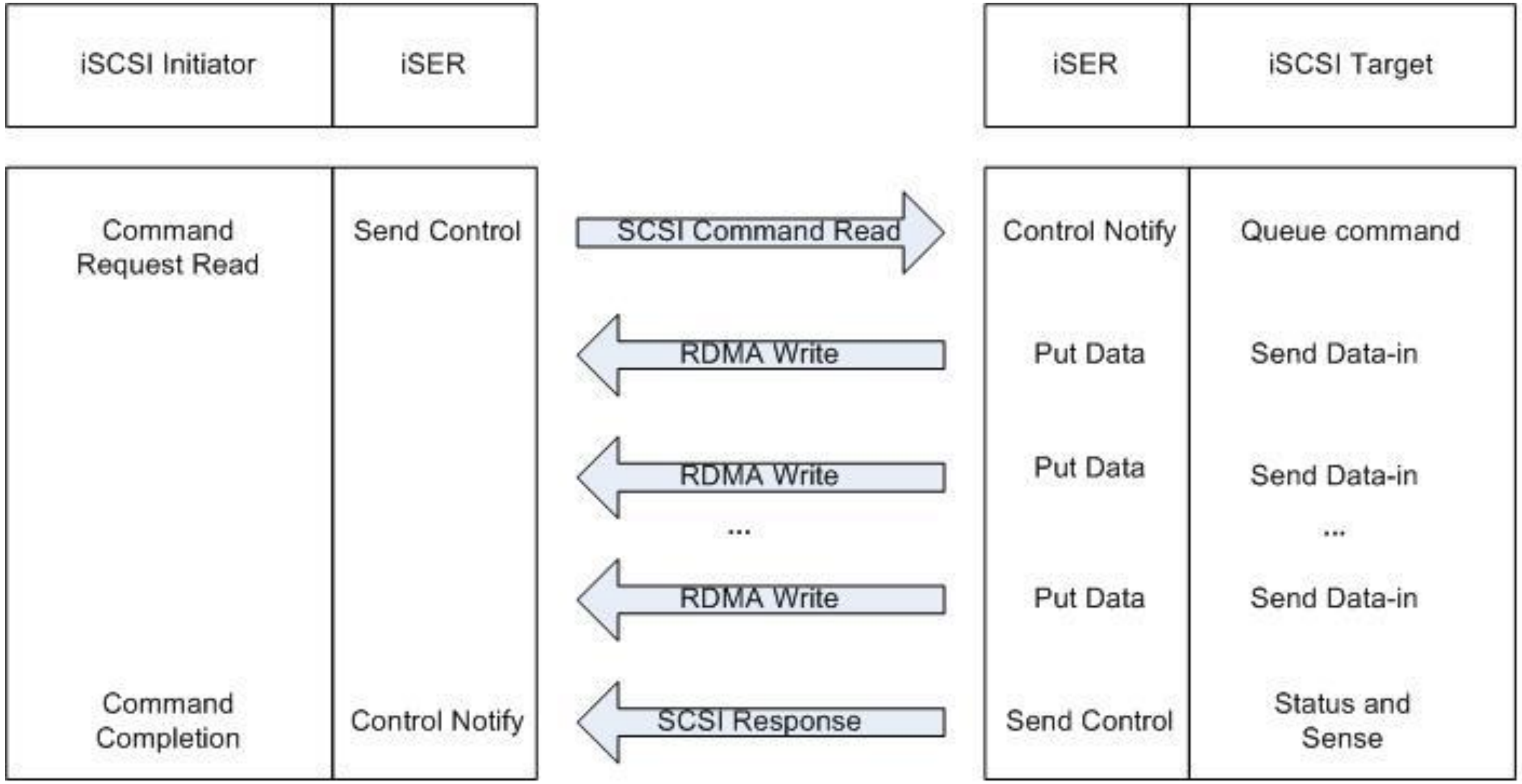
iSCSI with iWARP acceleration

iSCSI with IB acceleration

iSER: Coming of Age

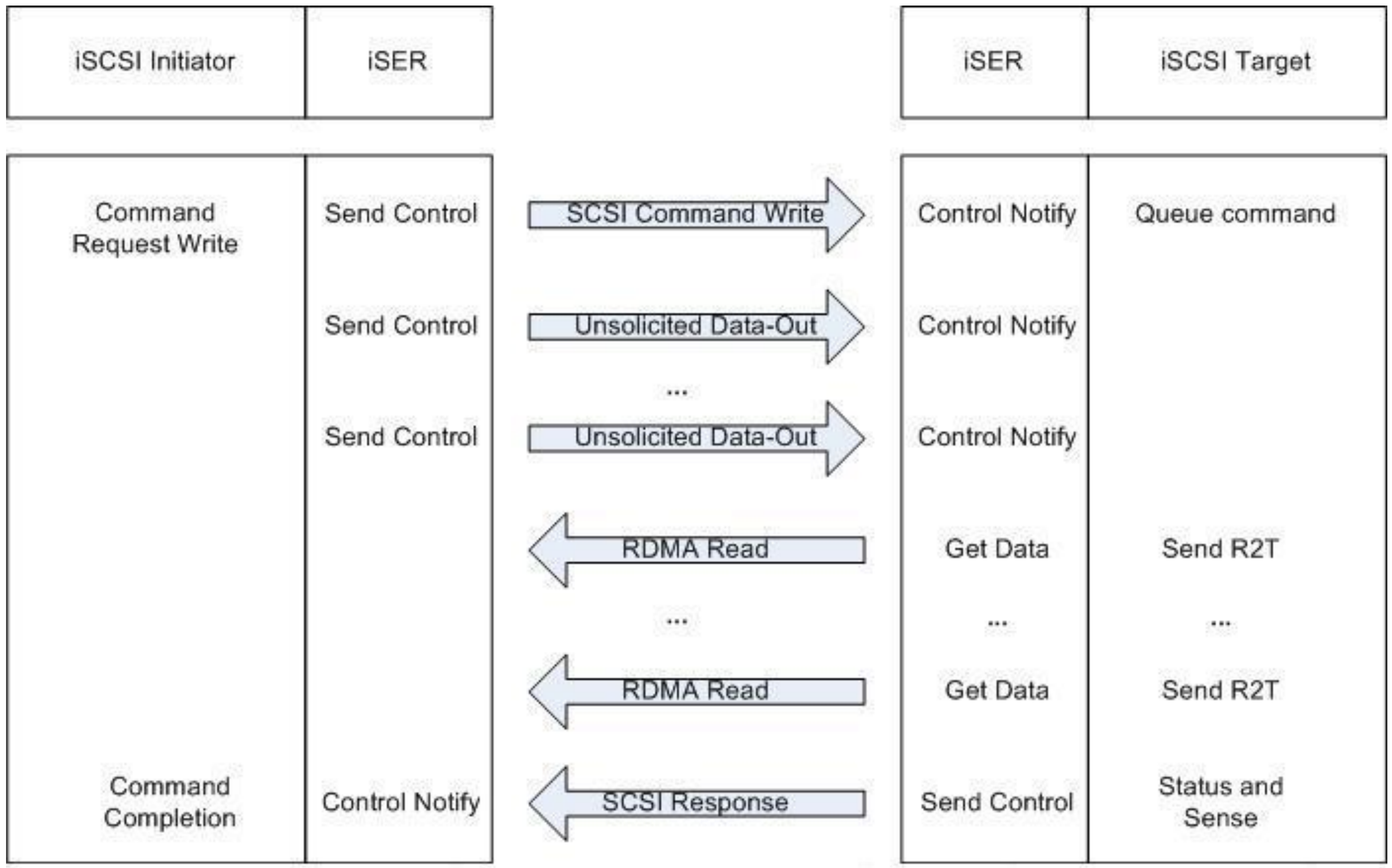
- iWARP continues iSCSI commodity approach
  - Layered on top of tradition TCP/IP, with all the benefits
- ROCe
  - Layered on UDP/IP
- No traditional reads or writes
- New Datamover Interface
  - Defines which end controls memory
- Initiator sends control to target
- Target gets or sends data
- Target returns control to initiator

iSER Read example



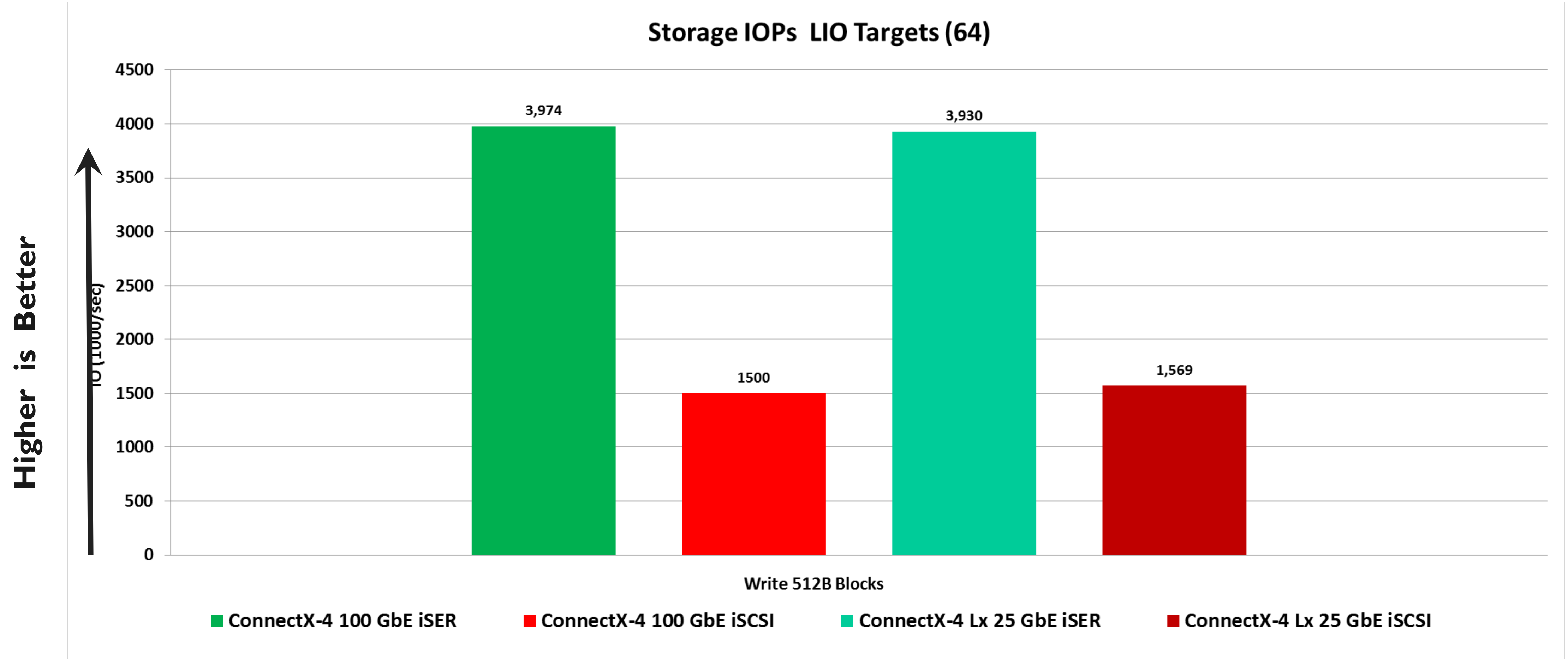
Source: Wikipedia

iSER Write example

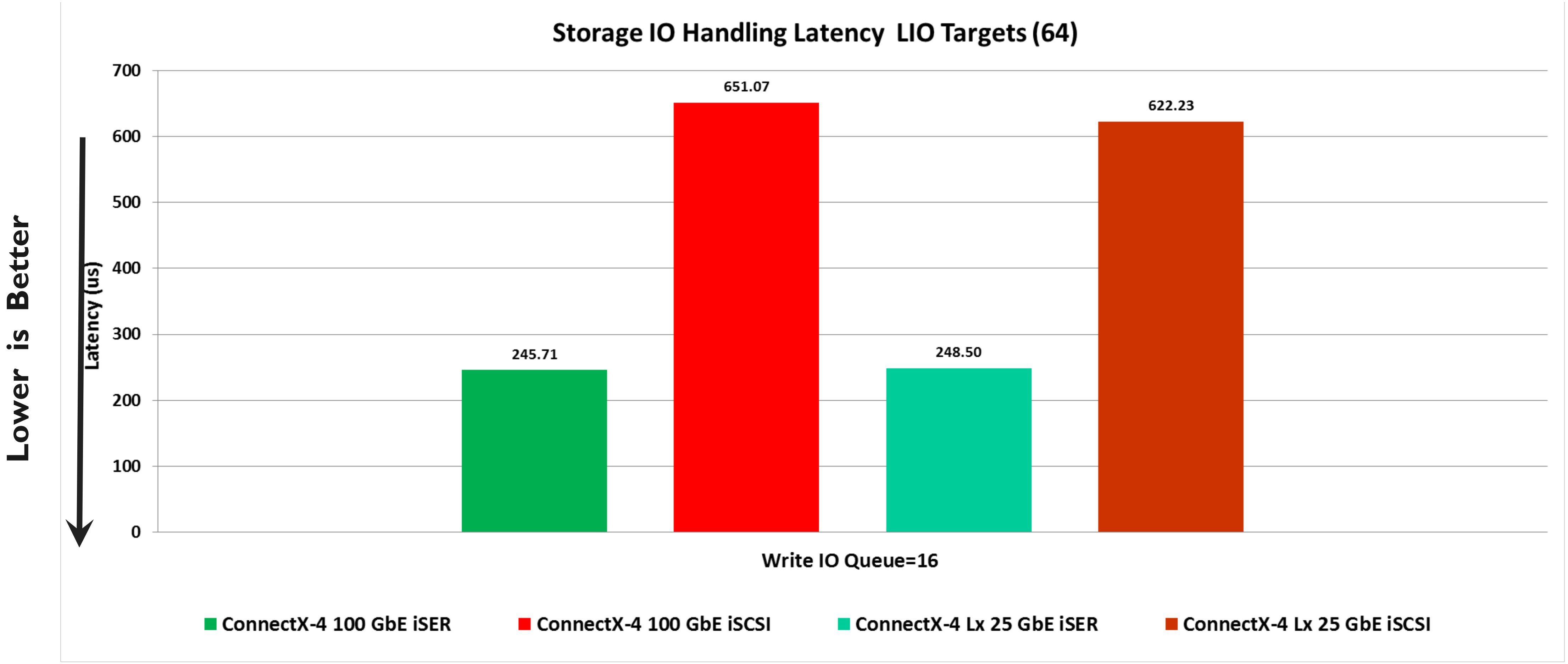


Source: Wikipedia

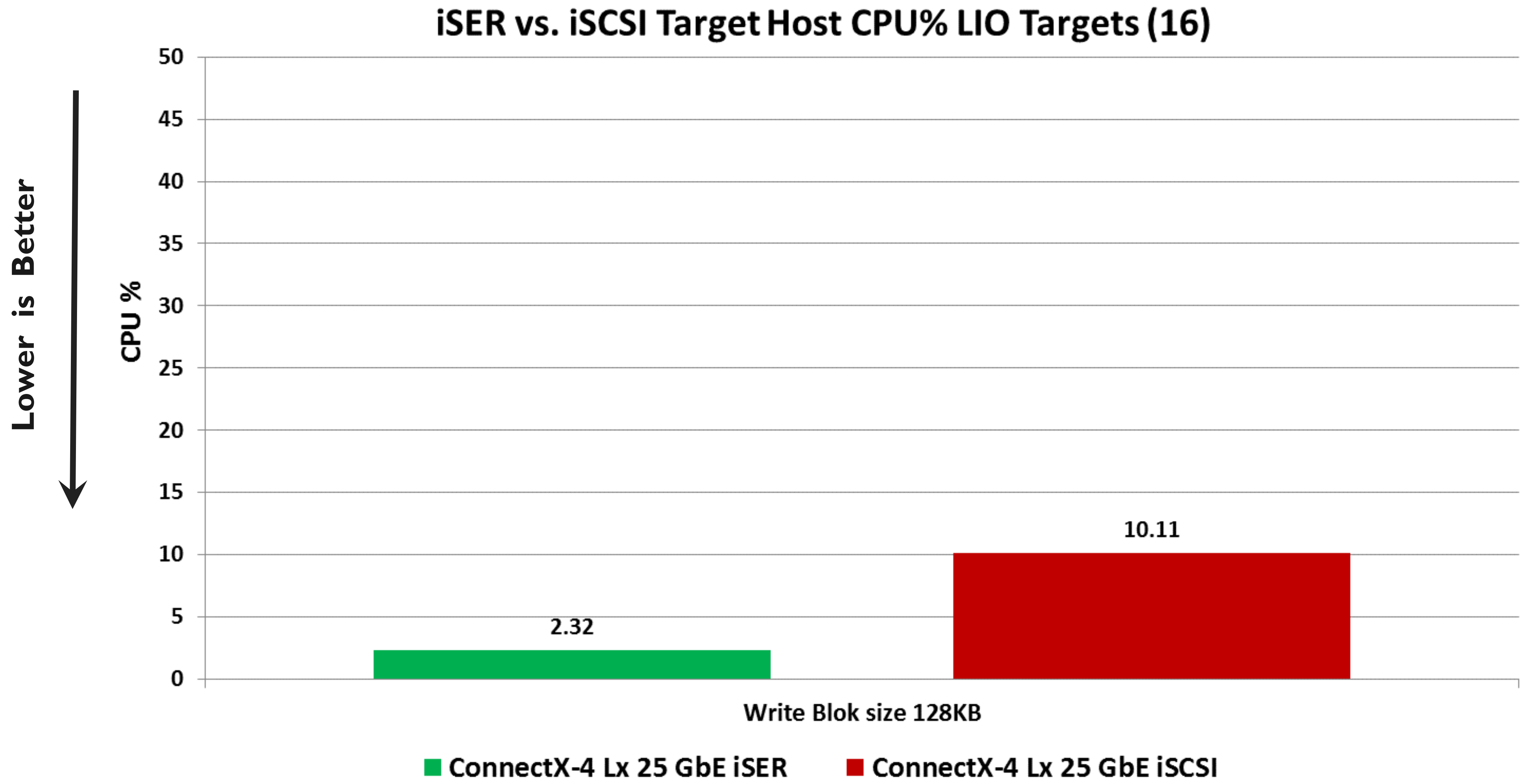
iSER Enables More IOPS



iSER Reduces Latency



iSER Frees Up The CPU



After This Webcast

SNIA Ethernet Storage Forum

- Please rate this Webcast and provide us with feedback
- This Webcast and a PDF of the slides will be posted to the SNIA Ethernet Storage Forum (ESF) website and available on-demand
- http://www.snia.org/forums/esf/knowledge/webcasts
- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-ESF blog
- http://sniaesfblog.org/
- Follow us on Twitter @SNIAESF

That's it!

Thanks

Fred Knight

Andy Banta @andybanta