An Analysis of Three Gigabit Networking Protocols for Storage Area Networks

Kaladhar Voruganti Prasenjit Sarkar
IBM Almaden Research Lab
San Jose, CA, USA

Abstract

With the steady increase in the storage needs of most organizations, block storage management is becoming an important storage management problem. Both databases as well as file systems ultimately rely on the presence of an efficient and scalable block storage management system. SCSI protocol is the protocol of choice for block storage management. Different transport mechanisms such as parallel SCSI, Fibre Channel, iSCSI and Infiniband can be used to transfer SCSI protocol data. Distance and connectivity limitations are the key drawbacks of parallel SCSI. Therefore, Fibre Channel, Infiniband and iSCSI (SCSI over TCP) are competing to emerge as the dominant next generation SCSI transport mechanism. By analyzing the different components of a network protocol, this paper evaluates whether Fibre Channel, Infiniband and iSCSI are a) suitable for gigabit wire speeds b) scalable across thousands of interconnected devices c) satisfy the needs of storage environments and storage protocols.

1 Introduction

Fibre Channel, Infiniband and iSCSI (SCSI over TCP/IP) are three popular gigabit speed technologies that are competing to become the next dominant networking technology for storage area networks (SANs). Though native ultra SCSI technology supports gigabit networking speeds (160Mbytes/Sec and 320Mbytes/Sec), the distance (few meters) and connectivity limitations (16 devices to a channel) are hampering its acceptance as the gigabit networking technology of choice for the emerging storage area networks.

1.0.1 Protocol Evaluation Criteria

The goal of this paper is to analyze the three protocols according to the following criteria:

- **Protocol efficiency for gigabit speeds:** It is a prerequisite for the successful networking protocol to provide high throughput and a high I/O rate because native storage transport protocols like ultra SCSI already satisfy this criteria. In the past, networking protocols were not used for storage environments because the conventional networking protocols could not match the bandwidths provided by the native storage protocols (such as SCSI). Network protocol components such as flow control, congestion control, support for zero-copy, packet size, fragmentation/assembly, service class support, and CRC calculation are used to evaluate this point.

- **Protocol scalability:** It is necessary for the successful protocol to scale with respect to the number of storage devices because the storage needs of most large organizations are in the order of petabytes and the storage device capacities are in the order of gigabytes. It is also necessary to evaluate how easily these protocols can be mapped on to the IP network because the majority of the wide area networks use the Internet protocol. Network protocol components such as address length, discovery, timeout mechanism, and flow control are used to evaluate this point.

- **Compatibility with SAN protocols and environments**: The successful protocol must be able to efficiently handle 4K and 8K block sizes (database and file system page sizes). It must be efficient for LAN distances since most of the current SAN deployments are LAN based. Furthermore, it should be able to efficiently handle the SCSI protocol (most of the server disks and tapes are SCSI based). Finally it should be reliable and secure otherwise organizations will not trust general networking protocols for storing their confidential and critical data. Network protocol components such as CRC calculation, security, flow control and packet size are used to evaluate this point.

Section 2 gives a brief overview of the three protocols. Section 3 analyzes how the three protocols implement the different network components, and it evaluates each of the protocols according to the above mentioned criteria. Finally, Section 4 summarizes the key findings and it contains our conclusions.

2 Protocol Overview

Fibre Channel, Infiniband, and iSCSI are the three protocols being evaluated in this paper. This section will now provide a very brief overview of these three protocols.

2.1 Fibre Channel

Fibre Channel (FC) protocol [Ben96] covers the physical, link, network and transport layers of the OSI network stack. Fibre Channel provides support for many different service classes (acknowledged connection oriented, acknowledged connectionless and unacknowledged connectionless). Fibre Channel protocol contains a definition for SCSI over Fibre Channel called FCP. Fibre Channel is a secondary network protocol (that is does not compete with in-box protocols such as PCI). Frames, sequences and exchanges are the three major data transfer constructs in FC. Frames are the basic unit of data transfer. The transfer of a group of frames in one direction is known as a sequence. A group of sequences in both directions are known as an exchange. Fibre Channel currently supports link speeds of 2Gbps (10Gbps coming soon) and it can operate in a switched network environment.

2.2 Infiniband

Infiniband (IBA) protocol [IBA00] covers the physical, link, network and transport layers of the OSI network stack. IBA can act as both a primary (within the box) and secondary network protocol. IBA provides support for many different service classes (acknowledged connection oriented, acknowledged connectionless, and unacknowledged connectionless). IBA provides the QueuePair programming abstraction that allows application programs to transfer data directly from the network card into the application memory (remote direct memory access or RDMA) and vice-versa. IBA provides the notion of verbs (programming APIs) which allows application programs to send and receive data. Frames are the basic unit of data transfer in IBA. A group of IBA devices are managed as part of a subnet. There is one subnet manager for each subnet. SVP (SCSI over VIA protocol) is the current industry proposal that tries to map SCSI over IBA. IBA currently supports link speeds of 2.5 Gbps and it can operate in a switched network environment.

2.3 iSCSI

iSCSI protocol [Sea00] defines the operation of SCSI over TCP. iSCSI protocol tries to leverage the existing TCP over IP over gigabit Ethernet infrastructure. Since TCP is used across both LANs and WANs, iSCSI is also meant to be a secondary (outside the box) protocol. iSCSI uses TCP flow control, congestion control, segmenting, and it builds upon the IP addressing and discovery mechanisms. iSCSI can be implemented as either a combination of commodity gigabit Ethernet card with the TCP/IP and iSCSI layers in software or it can be implemented as a specialized network card which implements Ethernet, IP, TCP and iSCSI layers. Gigabit Ethernet currently supports 1 Gbps link speeds (10 Gbps speeds coming soon). It can operate in a switched network environment.

3 Gigabit Wire Speed Efficiency

Data communication cost consists of software transmission overhead (at the message sender and receiver sides) and on-wire propagation overhead. With the emergence of gigabit fibre-optic networks the propagation overhead has been drastically reduced. Therefore, in order to realize the benefits of gigabit wire speeds the transmission overheads need to be also reduced. This section describes each of the transmission overheads within the context of each of the protocols. The following transmission overheads are analyzed in this section:

3.1 I/O Architecture

As shown in Figure 1, data flows from either user memory or kernel memory through first the system local bus and it then flows via the PCI bus to the network card. The network card, in turn, puts the data on the wire. Both Fibre Channel and iSCSI are second order network protocols which must use the PCI bus protocol (or a substitute) in order to transfer the data into memory. However, Infiniband (IBA) can act as both a second order and a first order network. Therefore, data can flow straight from the peripherals to the system local bus. More importantly, IBA utilizes a switched interconnect architecture which is more scalable (with respect to bandwidth) than the bus architecture provided by PCI or PCI-X. Furthermore, FC and iSCSI can also use IBA as the first order network, but they will incur protocol conversion overhead that is not present if IBA is used as both the first and second order network.

3.2 Zero-Copy

As shown in Figure 1, in wide-area network protocol stacks (TCP/IP) the data is copied from the network

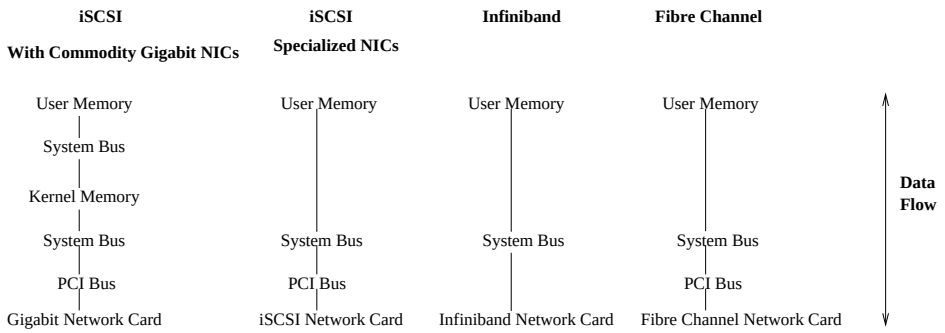

Figure 1: Data Flow in the Different Protocols

card first into the kernel memory, and it is then subsequently copied from the kernel memory into the application memory. Ideally, it is desirable to copy the data straight from the network card to the user application space and the avoidance of passing the data through the kernel memory is known as *zero-copy*. The lack of zero-copy not only increases the path length of the data flow but it also increases the CPU utilization of the host machine. The host CPU becomes a bottleneck when the user application is dealing with large data block sizes and this, in turn, bounds the maximum throughput that can be achieved by protocol.

In FC, the FC network cards perform the necessary zero-copy processing. The message from the sender to the receiver contains the Exchange ID and the Source ID. Furthermore, the message sender network card keeps track of the memory location into which the response from the target will be stored. The response message from the target contains the same Exchange Id and it also contains the Destination ID (which was the Source ID in the request message). The host FC card then uses the Exchange Id along with the frame number within the sequence to determine the target memory location for the request payload. Thus, the host FC card keeps the association between the destination memory location and the Exchange ID. This, in turn, allows the FC network card to directly place the contents of the message into the application memory space. Since each FC frame contains the Exchange ID, Sequence ID, and Frame Number (within the sequence) information, loss of a frame or out-of-order arrival of a frame does not prevent the zero-copy operation from taking place.

The combination of queue pair and RDMA concepts in IBA allow for the implementation of zero-copy semantics. In IBA, when the sender sends a read request to the receiver, the sender also sends the RKey and the starting address of the memory region into which it wants the target data to be placed. The message receier does a RDMA write operation to put the data directly into the message sender's application memory. Based upon the starting memory address provided by the initiator, the target has to calculate the memory address for all of the frames which together constitute the response. For example, if the initiator wants to perform a 16K read operation, then the target sends the response data via four 4K IBA frames. The RDMA address for each of the frames is calculated by the target based upon the starting address sent initially by the initiator. The RKey is a authorization key which gives the read request receiver the permission to place the data into the message sender's address space. Since each frame contains the address of their corresponding host memory location, loss of frames or out-of-order arrival of frames do not prevent zero-copy operations.

The TCP protocol is a stream-based protocol, and it is difficult to implement zero-copy semantics for stream-based protocols because the TCP segments could be spread across multiple Ethernet frames. Therefore, the iSCSI header and data could be spread across multiple Ethernet frames. Also, data belonging to multiple unrelated iSCSI requests could arrive as part of a single TCP segment. Therefore, if the Ethernet frame containing the iSCSI header information is dropped or it arrives out of sequence, then the network cards need to buffer the trailing Ethernet frames until the Frame containing the iSCSI header information is retransmitted. If iSCSI protocol is operating at link speeds then the network card would require a lot of memory in order to buffer the trailing Ethernet frames until the Frame containing the iSCSI header arrives at the network card. Thus, in order to reduce the memory requirements at the network cards

and efficiently implement zero-copy, it is necessary for iSCSI to adopt a framing mechanism similar to the one present in either FC or IBA.

It is desirable for iSCSI to use commodity gigabit Ethernet network cards to implement zero-copy semantics rather than customized iSCSI network cards because the price of commodity Ethernet cards will be lower than the cost of customized iSCSI cards. If commodity gigabit Ethernet cards are used to implement iSCSI, then it is also necessary to avoid copying data into the kernel memory and then subsequently into the application buffer. Remapping kernel memory as application memory, and introducing RDMA tag fields in TCP segment headers (as optional fields) are some of the solutions being proposed to realize zero-copy semantics in iSCSI. It is important to note that all iSCSI zero-copy semantics require a change to the TCP stack code (which is under the control of operating system vendors). Furthermore, the socket interface layer to the TCP stack has to be changed to allow for the passing of addresses of the user data buffers to the network cards.

3.3 Fragmentation and Assembly

The storage application program can issue block I/O requests that typically vary in size between 4K and 64K. If the storage application has issued a block I/O request for a 8K block, then this request needs to be mapped on to 1.5 K sized Ethernet frames in iSCSI, 2K sized FC-2 frames in Fibre Channel, and 4K sized frames in Infiniband. Hence, there is a cost associated with fragmenting and assembling the larger sized block I/O requests into smaller sized network frames. The protocols with the larger frame sizes have some advantage over the protocols with the smaller frame sizes due to lower fragmentation and re-assembly costs.

The FC and IBA network cards perform the fragmentation and reassembly operations and thus, they offload these operations from the host CPU. If iSCSI is implemented using specialized iSCSI network cards, then iSCSI network card can also offload this operation from the host CPU. However, if iSCSI is implemented using commodity gigabit Ethernet cards, then it is necessary for the network card to minimize the number of times it interrupts the host for placing the Ethernet frames into the host memory. The gigabit Ethernet cards provide interrupt coalescing support which reduces the number of times the network card interrupts the host. With commodity Ethernet network cards, the TCP segmenting and assembly work is performed by host, and this adds to the CPU utilization.

3.4 Flow Control

Flow control is the process which controls the rate at which the sender sends frames into the network. The senders data injection rate is usually a function of the amount of available buffer space available at the receiver and at the intermediate network switches and routers.

FC uses a credit based flow control mechanism in which the data senders are allotted credits by the receiver of the data. The sender can only send data as long as it has not used up its quota of credits. The receiver returns credits back to the sender when it sends an acknowledgement to the sender. The receiver allocates credits when it has the necessary buffer space to store the sender's data. The credit based flow control mechanism ensures that there never is dropping of packets due to data congestion at the intermediate switches and routers.

IBA also uses a credit based flow control mechanism that is similar to FC flow control mechanism. Thus, IBA switches also do not drop packets due to congestion. However, in addition to the credit based flow control mechanism, IBA also provides a static rate control mechanism which tries to minimize link transfer rate mismatch. For example, as shown in Figure 2, there are two source nodes, an intermediate switch node and two destination nodes. The speed of one of the source nodes is 12X and the speeds of the other source node is 3X and the speed of one of the destination node is 2X and the other destination node is 1X. Both the sources have two credits available to them from the switch node. The 12X link rate source can unfairly take up more of the buffer space at the switch because its destination's link rate is only 1X and therefore, the destination cannot accept the data at 12X rate. Even though the second destination can accept data at a faster rate, its sender (sender 2) cannot send the data at that rate because there are no buffer spaces available at the intermediate switch. Thus, the static rate control mechanism in IBA allows source 1 to inject data into the network at 1X speed, and thus, increase the overall throughput of the network. Fibre Channel's credit based flow control mechanism currently does not have a static rate control mechanism to handle heterogeneous link speeds.

The flow control mechanism used in iSCSI is dependent on the flow control mechanism that is used by TCP/IP. There is only end to end flow control mechanism in TCP/IP. That is, the two end-points of a connection negotiate a window size, which is based upon the buffer space available at their respective ends. The

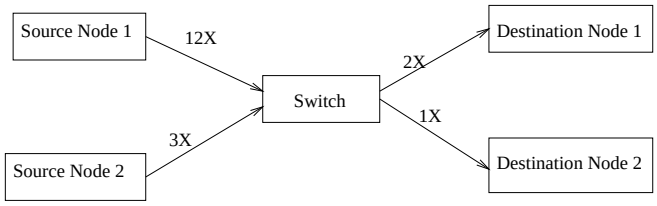

Figure 2: Static Rate Control in IBA

window represents the number of messages that can be sent without receiving an acknowledgment from the receiving side. At gigabit wire speeds, it is necessary for the sender and receiver to increase the size of their windows because the data arrives at the destination as fast as the sender can inject it into the network. Previously, in slow speed networks, having a small window size was not a problem because the data propagation delay gave the destination node enough time to process the data present in its small receive buffers. Furthermore, TCP does not use a credit-based flow control mechanism. Therefore, congestion can occur at the intermediate switches as well at the destination nodes. TCP reacts to congestion by dropping packets.

4 Protocol Scalability

Address space, discovery, flow control and timeout mechanisms are the four protocol scalability issues being evaluated in this section.

4.1 Address Space

FC uses 24 bit addresses whereas IBA and iSCSI can use 128 bit addresses. Thus, FC addressing is less scalable than the other two protocols. In FC, the network cards usually have a restriction on the number of simultaneous LOGINs they can manage. This a further scalability barrier present in FC that is not present in the other two protocols.

4.2 Discovery

In FC, when a new device comes on-line, it contacts its fabric manager (a switch). The fabric manager, in turn, informs all the devices have registered with the fabric manager and which want to be informed about this event. Furthermore, in FC, when a device comes on-line, it performs a login with all the other devices that are present in the same zone. A zone is an access control mechanism which allows communication only between its members. Moreover, the switch to which this device connects, informs all the other switches in the fabric about this event. This mechanism is not scalable in environments with thousands of switches and devices.

The subnet manager plays a key role in the IBA discovery mechanism. Upon being made active the subnet manager queries all the devices in the subnet and updates its database with the active port and route information. The subnet manager periodically sweeps the subnet to get the up to date status of active and inactive ports. Even though the sweeping mechanism is limited to a subnet, it can add a lot of extra traffic in large subnets (a subnet has an address space of 16 bits).

In the discovery mechanism used by iSCSI, a node can either hard code the address, or it can query a storage name server or it can send a multicast message inquiring which devices it can access. In large networks with thousands of devices, the storage nodes will use the mechanism of querying a storage name server rather than the multicast approach. Once the message sending node (initiator) acquires the IP address and TCP port number of the message receiving device (target) from the storage name server, the initiator establishes a connection only if it wants to communicate with the target.

4.3 Large Propagation Delays

FC and IBA credit-based flow control mechanism is not as scalable as the iSCSI (TCP) flow control mechanism. The credit-based flow control mechanism can lead to under utilization of the network in wide-area networks (with large propagation time delays) because the sender has to wait for a long time to get credits from the receiver for injecting new data into the network. Moreover, since the receiver could be simultaneously serving many sources, the receiver cannot give out more credits (to the sources) than the total amount the available buffer space in order to keep the network pipe full. Thus, the credit-based flow control mechanism is only adequate when the network propagation delays are small. However, the non-credit based flow control mechanism that is used by iSCSI is scalable, because the senders dynamically increase or decrease their data transfer rates at the expense of packet drops at the network nodes during periods of

high congestion.

4.4 Timeout Calculation Mechanism

In TCP/IP protocol, the nodes dynamically maintain the time it takes them to send a message to the other end of the connection. Therefore, as the traffic patterns change, the round trip time is adjusted accordingly to reflect the network characteristics. The round trip time is, in turn, used by the message senders to decide whether they should resend a message due to potential loss of the packet in the network. Both FC and IBA have a static timeout specification mechanism which does not adjust itself dynamically according to the network conditions. Thus, the message sending nodes in FC and IBA can timeout either too soon or late, and this can negatively impact the overall performance.

5 Storage Protocols/Environments

The three protocols being evaluated in this paper are general networking protocols. Therefore, it is necessary to evaluate whether they are compatible for storage protocols and configurations.

5.1 LAN based SANs

Since TCP flow control mechanism allows for packet drops during periods of congestion, the iSCSI (TCP) congestion control mechanism is inadequate for LAN based networks. LANs are usually oversubsribed and therefore, congestion is usually a very transient phenomenom. Moreover, the congestion control mechanism used by TCP during packet loss results in a drastic reduction of the sender's window size (slow start) and it takes the sender some amount of time before it can operate again at full speed. Thus, the TCP flow control and congestion control mechanisms are inadequate for LAN based SANs. Since the propagation delays are usually small in LANs, the credit-based flow control mechanisms used by FC and IBA are desirable because they do not allow for packet loss during periods of congestion.

5.2 Packet Size

Most of the storage applications typically deal with 4K and 8K block sizes. Therefore, the presence of Jumbo Frames (9K frames) will reduce the fragmentation and assembly overhead for these storage applications. Large frame sizes have not advocated in wide-area network environments because targets can potentially be not prepared for large unsolicited frame sizes. However, in SAN environments, the data receiving nodes are prepared for large sized block data transfers. If the write command initiating nodes (initiators) perform write operations without using the SCSI RTT mechanism, then the message receiving nodes (targets) could potentially be not prepared to receive large sized blocks. Thus, it is better to use large sized frames in over-subscribed LAN-based SAN environments. Currently, only Ethernet vendors provide proprietary Jumbo frame solutions. Therefore, LAN-based iSCSI SAN setups can potentially use Jumbo frames if all of the SAN devices use compatible Jumbo frame solutions.

5.3 Message Acknowledgements

The SCSI storage protocol is a query-response protocol and therefore, there is no need for explicit acknowledgements. That is, the response from the target for a query posed by the initiator is an acknowledgement itself, and therefore, no explicit acknowledgements need to be sent by the target to indicate that it has received the query from the initiator. SCSI level timeouts can be used by the initiator to recover from packet losses. Both FC and IBA support unreliable classes of service where the receiver of a message does not send an acknowledgement back to the sender. Since TCP (hence iSCSI) is a reliable transport protocol, the destination nodes send acknowledgement messages. Even though the acknowledgement messages can be piggybacked on other messages, the acknowledgement message processing overhead can unnecessarily increase the CPU utilization at the source and destination nodes if the iSCSI implementation utilizes commodity (non TCP offload) gigabit Ethernet cards.

5.4 CRCs

There is a CRC value associated with each FC frame and IBA frame, and there is a CRC associated with TCP and IP segments. The sender calculates a CRC value and places it in the packet, and the receiver re-calculates the CRC value and compares it with the value calculated by the sender to ensure that the data has not been corrupted. It is desirable to perform the CRC calculations at the network card and not at the host because the host can use its CPU cycles for other useful application related work. Therefore, most of the network cards provide checksumming offload support. If iSCSI is using commodity network cards with no TCP/IP offload support, then it is necessary to de-activate the software checksumming process in the TCP/IP. This can require kernel modifications in the TCP/IP stack.

The checksumming mechanism used by TCP/IP is much more error prone than the CRC mechanisms used by FC and IBA. TCP uses a 16 bit checksum whereas FC and IBA use 32 bit CRCs. 16 bit checksums used by TCP/IP are less robust than the 32-bit CRCs. Furthermore, TCP checksum is located before the payload in the data frame, whereas, FC and IBA

CRCs are located after the payload in the data frame. Therefore, the FC and IBA CRC calculations can be performed on the fly as the data is moved in and out of the network card without performing extra memory references (as is the case in TCP checksum calculation). Moreover, the frequency with which corrupted TCP packets are received in conventional WANs (due to the manner in which TCP checksum is calculated) is not an acceptable solution for storage applications that deal with mission critical corporate data. Thus, the iSCSI protocol is being supplemented with a 32 bit CRC mechanism in addition to the TCP checksum. Current commodity gigabit Ethernet network cards do not provide support for CRC calculations. This means that the host itself has to perform the CRC calculations and its CPU utilization will become high. This, in turn, makes implementing iSCSI using commodity network cards a less realistic solution.

5.5 Security

Corporations are careful about sending their confidential data on networks that span great geographical distances. IP based networks already have a well defined encryption and authentication mechanisms whereas, both IBA and FC do not have a security mechanism defined. This makes it much more safer to deploy iSCSI than IBA and FC. Both FC and IBA protocols have to be mapped on to IP (via gateways) in order for them to provide the level of security that is desired by corporations. Moreover, iSCSI would most likely have to be implemented using iSCSI network cards (that offload encryption functionality from the hosts) otherwise, the host CPUs can get saturated and thus, negatively impact the overall throughput.

6 Conclusion

IBA and FC have support for framing whereas, framing is not native to the stream-based TCP/IP and therefore, it has to be added at the iSCSI layer. iSCSI most likely needs to be implemented via specialized network cards in order to provide zero-copy, CRC and data encryption support. IBA and FC flow control mechanisms are more suitable for LAN environments whereas, iSCSI flow control is more suitable for WAN environments. It is desirable to increase the frame size in all of the three protocols to reduce the fragmentation/reassembly costs. TCP/IP congestion control (slow start) is not suitable for LAN based SANs. FC addressing capability (24 bits) needs to be increased in order for it to have the same addressing capability as IBA and iSCSI. iSCSI has its own CRC mechanism because TCP/IP checksum is not reliable enough for storage applications. iSCSI has a more scalable (when dealing with large networks with many millions of devices) discovery/Login mechanism than FC and IBA. iSCSI's dynamic timeout calculation mechanism is more suitable for WAN environments with unpredictable load changes. iSCSI can rely on the proven authentication and encryption mechanisms that are being used by the IP networks, whereas, security measures have not been well defined for IBA and FC. Finally, non-technical market forces will most likely determine which of the protocols will be successful. iSCSI has the advantage that organizations can use their existing IP network infra-structure (switches, routers and network administrators) to deploy iSCSI based SANs.

Acknowledgments

We want to thank our colleagues David Chambliss, Jim Hafner and Mike Ko for clarifying some Fibre Channel and Infiniband concepts.

References

[Ben96] A. Benner. Fibre Channel: Gigabit Communications and I/O For Computer Networks. McGraw-Hill, 1996.

[IBA00] Infiniband Architecture Specification Volume 1, Release 1.0. Infiniband Trade Association, 2000.

[Sea00] J. Satran and et al. iSCSI. In *http://search.ietf.org/internet-drafts/draft-ietf-ips-iscsi-04.txt*. IETF, 2000.