OSL
UNIX Pfadfinder

iSCSI Einführung

Copyright und Handelsmarken

Beschränkungen

Versionen dieses Dokumentes

Version	Datum	Author	e–mail	Inhalt / Änderungen
1.0	11.11.2004	HO	pathfinder@osl–it.de	Erste Fassung

1.Einleitung

Das iSCSI Protokoll beschreibt den Transport von SCSI Paketen über TCP/IP Protokollen. In einem iSCSI Netz werden Initiators (Disk, Tape) and Targets (Host) über Gigabit Ethernet Switches zu einem Storage Netzwerk verbunden. Ziel ist es, Storage Funktionalitäten auf herkömmlicher Kommunikations (Internet) Netzwerkinfrastruktur abzubilden.

Im Gegensatz zu bereits existierenden Storage–optimierten Netzwerktechnologien (Fibre Channel), versprechen sich die Akteure große Preisvorteile gegenüber Fibre Channel und somit eine Herabsetzung der Hemmschwelle auf Seiten der Kunden. Mit dieser Technologie, so die Hoffnung, hält neue Funktionalität unter Einsatz bereits vorhandener Infrastruktur (Kabel, Switche, NIC) Einzug in die Rechenzentren kleinerer, mittelständischer Unternehmen, welche bisher den Einsatz eines Speichernetzes scheuten.

Der vorliegende Artikel setzt sich das Ziel, eine Einführung in die iSCSI Technologie mit Hinblick auf einen möglichen Einsatz im Rechenzentrum zu bieten. Gleichfalls soll durch den Vergleich mit Fibre Channel der aktuelle Stand der iSCSI Entwicklung hinterfragt werden.

2. iSCSI im Vergleich

Technologie	Übertragungsgeschwindigkeit	Nutzdatenrate
Fibre Channel	2 Gbit/s (256 MByte/s)	180 MByte
1 GBit Ethernet	1 Gbit/s (128 MByte/s)	40 MByte/s
10 GBit Ethernet	1 Gbit/s (128 MByte/s) * 10	40 MByte/s * 10

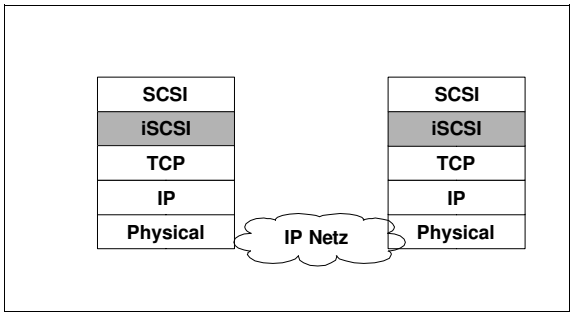
Rechenbeispiel: 1GBit/s Fibre Channel 1024MBit/s –> Transmission Encoding 10/8 Bit 102,4MByte/s –> davon ca. 90% Payload = 90 MByte/s.

Die Motivation für iSCSI liegt auf der Hand, Storage Netzwerke sollen endlich für mittelständische Kunden erschwinglich werden. Mit 10 GBit Ethernet hat man die technologische Basis, auf welcher man mit Fibre Channel Übertragungsraten konkurrieren kann. Darüber hinaus versprechen einfache Ethernet Netzwerkkomponenten einen immensen Preisvorteil. Billiger und schneller ! Was will man mehr ?

3.Technologie

Das iSCSI Protokoll übernimmt als zusätzlicher Layer die Aufbereitung von SCSI Befehlen des Host Systems und bereitet diese für den Transport durch TCP vor. Für iSCSI bot sich der Transport über TCP an, da dieses Protokoll

- die Reihenfolge der IP Pakete gewährleistet
- verlorengegangene Pakete neu anfordert



d.h. Transportsicherheit für iSCSI gewährleistet. Allerdings bietet TCP keine Möglichkeiten zur Kanalbündelung. Ebenso ist es nicht in der Lage ein Failover durchzuführen. Eine iSCSI Session ist deshalb als eine oder mehrere TCP Verbindungen definiert, welche ihrerseits über mehrere physikalische Verbindungen ermöglicht werden kann.

Asymmetrisch:

Trennung der SCSI Kommandos von SCSI Data, welche über getrennte TCP Verbindungen übertragen werden. Dies hat den Vorteil, dass wichtige Management Befehle nicht vom Datenverkehr behindert werden.

Symmetrisch:

SCSI Kommandos und SCSI Payload werden über einen Kanal übertragen. TCP stellt dabei sicher, dass nach dem Senden/Empfangen von Daten dies quittiert wird. Nachteil ist hier, dass Datenverkehr wichtige Management Befehle blockieren kann.

Schlussendlich wurde dieser Nachteil in Kauf genommen, weil die asymmetrische Herangehensweise nicht gewährleisten kann, dass SCSI Daten zeitlich vor SCSI Kommandos empfangen werden (SMP Systeme). Dieses Problem hätte eine Interprozess Kommunikation auf Hostebene notwendig gemacht, welche dem Anspruch eine eigenständige, gekapselte Lösung zu schaffen, entgegensteht.

3.1. Direct Data Placement

Bei steigenden Datenraten macht sich ein ebenso steigendes Memory–zu–Memory (zwischen einzelnen Protokoll Schichten) Kopieren auf der Empfängerseite bemerkbar. Mit Direct Data Placement Technologien wird das Ziel verfolgt, CPU Zeit (Zyklen) im Netzwerk Verkehr zu reduzieren.

Unter Direct Data Placement versteht man das sofortige Verschieben ankommender Daten vom TCP Layer in den für diese Verbindung vorgesehenen SCSI Buffer. Dies ist insbesondere für SCSI notwendig, da hier mehrere Empfangsbuffer auf IO warten (pending Requests).

Bei iSCSI über Gigabit Ethernet wird der TCP/IP Stack die CPU und Speicherresourcen stark beanspruchen. Dieses Problem wird sich nicht mit dem iSCSI eigenen Data Placement lösen lassen, sondern muss für die IP Protokoll Familie insgesamt gelöst werden. Dafür stehen mit RDMA (Remote Direct Memory Access) der IETF Vorschläge zur Beschlussnahme vor. RDMA sieht sogenannte RNIC – RDMA aware NIC – vor, welche in der Lage sind, höherliegende Protokolle (iSCSI, NFS) zu erkennen und Pakete direkt in deren Empfangsbuffer zu kopieren.

Für iSCSI kam diese Entwicklung zu spät. Mittlerweile liegt mit iSER die iSCSI Extension für RDMA vor. Zum Zeitpunkt (Nov. 2004) liegt iSER die RDMA Extension für iSCSI als Draft vor.

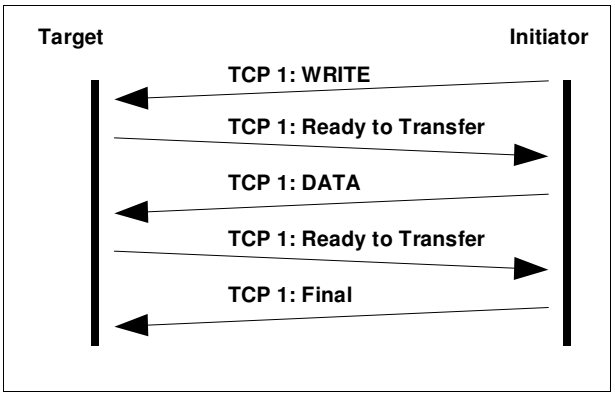
3.2. Session Management

Bevor Daten zwischen Target und Initiator ausgetauscht werden können, wird vom iSCSI Layer eine Verbindung initiiert. Dabei wird über bekannte TCP Ports (IANA 3260) eine Verbindung vom Initiator aufgebaut. Das Target wiederum lauscht auf ankommende Verbindungswünsche. Nachdem die TCP Verbindung steht, beginnt die iSCSI **Login Phase** (3.2.3. iSCSI Login). Da iSCSI Namen mehrere IP Adressen (bzw. IP Adressen mehrere iSCSI Namen) haben können, werden zur eindeutigen Identifizierung Session IDs verwendet. Dazu erzeugen Initiator und Target jeweils eine ID (ISID/TSID). Dieses Session ID Paar wird zur eindeutigen Identifizierung der entstandenen TCP Verbindung verwendet. Gleichfalls kann während der **Login Phase** die Art der Authentifizierung vereinbart werden, dies ist möglich aber nicht notwendig.

Login bezeichnet den Aufbau einer TCP Verbindung zwischen zwei iSCSi Geräten welche entweder einer neuen Session zugeordent oder einer bereits existierenden Session hinzugefügt werden.

Nach erfolgtem Login wird in die **Full Feature Phase** gewechselt. In dieser Phase ist das Senden von SCSI Kommandos und SCSI Read/Write Operationen erlaubt.

Connection Allegiance – Verbindungszugehörigkeit beschreibt die Notwendigkeit, dass SCSI Responses von dem TCP Kanal übertragen werden, welcher zuvor das entsprechende SCSI Request übertrug. D.h. SCSI Request/Response dürfen nicht über verschiedene TCP Kanäle übertragen werden. Diese Verbindungszugehörigkeit bezieht sich auf die Kommandoebene, d.h. nachfolgende SCSI Request/Response Paare können auf anderen TCP Verbindungen abgewickelt werden.



3.3. iSCSI Sicherheit

Wie bereits unter "Session Management" beschrieben, ist iSCSI in der Lage die Sicherheitsmöglichkeiten der IP Protokoll Familie zu nutzen. IPSec stellt sich dem höhergelegenem iSCSI Protokoll transparent dar und kann ohne weiteres zur Sicherung des Transportweges verwendet werden. Zur Authentifizierung wiederum kommen Techniken, wie

- CHAP (Challenge Handshake Authentication Protocol)
- SRP (Secure Remote Password).
- Kerberos
- Simple Public–Key Mechanism (SPKM)

zum Einsatz. Da sich iSCSI Geräte in einem IP Netz befinden, empfiehlt sich der Einsatz von Authentifizierungs Mechanismen.

3.4. iSCSI Fehler Behandlung

3.4.1. Verlust aller TCP Verbindungen

SCSI typisch bleiben iSCSI Sessions und deren Verbindungen aktiv bis diese explizit geschlossen oder fehlerhaft werden (TCP keep–alive). Ein Verlust einzelner Verbindungen zieht ein Beenden und ein erneutes Initieren dieser Verbindungen nach sich. Bei Verlust aller Verbindungen bedeutet dies ein Beenden der iSCSI Session. Hierbei werden alle Verbindungen freigegeben und SCSI Pakete der Warteschlange verworfen, bevor ein neuer Login Versuch gestartet wird.

3.4.2. Fehler auf dem Übertragungsweg

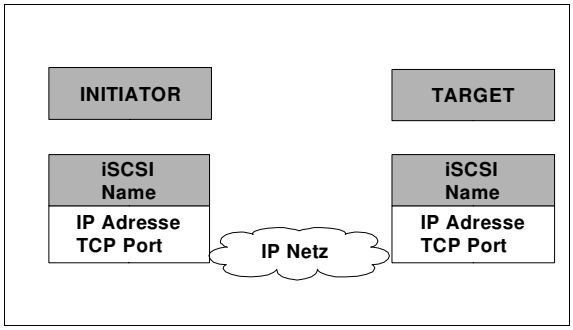
Initiator und Target sind in der Lage SCSI Befehle und Daten solange zwischenzuspeichern, bis deren Empfang (bzw. Weiterleitung) bestätigt wird. Weiterhin sind Initiatoren in der Lage "ping Packete" an Targets zu senden um Erreichbarkeit und Status zu testen. Fehlerhafte Verbindungen werden geschlossen und reinitiiert.

TCP versieht seine Payload mit einem 16 Bit CRC, welcher für Storage als zu schwach angesehen wurde. Ebenso ist die TCP Checksumme nicht in der Lage iSCSI Daten zu schützen falls diese im Speicher des Hostssystems oder auf dem PCI Bus korrumptiert werden. Deshalb werden iSCSI Daten mit einer 32 Bit großen Prüfsumme versehen. Da iSCSI Pakete mit einer fortlaufenden Sequence Nummer versehen werden, können fehlerhafte bzw. nicht empfangene Pakete neu angefordert werden.

3.5. iSCSI Adressen und Namen

Jedem iSCSI Gerät muss ein Name zugeordnet werden. iSCSI Geräten werden über Namen und nicht über Adressen angesprochen, was Voraussetzung für ein problemloses Verschieben von iSCSI Geräten in andere Netzwerksegmente ist.

Nur beim Versenden von Discovery Paketen darf der Target Name entfallen.



1 iSCSI Namen sind global eindeutig.

2 iSCSI Name gelten für die gesamte Lebensdauer.

3 Standort oder Adresse sind nicht ableitbar.

iSCSI Namen können bis zu 255 Byte lang sein und setzen sich aus 3 Teilen zusammen, dem Typ, dem Domain Inhaber und einem frei wählbaren, eindeutigen Namen. Als Typ Bezeichner sind **iqn** (für Firmen/Eigentümer einer eigenen Internet Domain) bzw. **eui** für von der IEEE als Naming Authority vergebene Namen. (siehe 3.2.6.3. iSCSI Name Structure)

Beispiele: iqn – iSCSI Qualified Name: iqn.2001–04.de.osl–it:storage.qa.symmetrix–sn–a8675309
eui – IEEE assigned globally unique identifiers: eui.02004567A425678D

Bei diesem Namensschema wird im Gegensatz zu Fibre Channel gewährleistet, dass redundante Pfade zu einem iSCSI Gerät auch als solche erkannt werden können. Soweit zur Theorie, inwieweit der iSCSI Namensraum vom SCSI Treiber mit seiner Ctrl–Target–Lun–Slice Metrik abgebildet werden kann, wird abzuwarten bleiben.

iSCSI Namen können über das iSNS (Storage Name Service) Protokoll aufgelöst werden. Dieses Protokoll bietet Storage Discovery und dient als zentrale Anlaufstelle zur Konfiguration eines iSCSI Netzes. Der Einsatz von iSNS ist optional, iSCSI Netze können alternativ über /etc/hosts verwaltet werden.

4. Ausblick

iSCSI ist trotz der bereits verabschiedeten Spezifikation Gegenstand weiterer Optimierung und Forschung. So wurde in aktuellen Untersuchungen eine Auftrennung in Sende– und Empfangsstrecke über eigene TCP Verbindungen vorgeschlagen. Diese Vorschläge und die gemachten Kompromisse bei der jetzigen Umsetzung des Protokolls (symmetrisches vs. asymmetrisches Modell) zeigen die Schwierigkeiten, Storage Anforderungen in IP Protokolle zu betten.

Ein weiteres nicht zu vernachlässigendes Problem ist die hohe Ressourcen Beanspruchung (CPU, Speicher) für den Netzwerkverkehr, welche durch sogenannte "TCP/IP Offload Engines" minimiert werden kann. Dieses extra Stück Hardware (1000 – 1500 EUR) speichert Netzdaten in einem Zwischenbuffer, was die CPU entlastet und die Latenz erhöht . Eine richtige Lösung für dieses Problem wird hoffentlich RDMA bringen.

Interessant: Fibre Channel und 1 Gb Ethernet teilen sich bereits heute eine optische Spezifikation für Kabel und Übertragungstechnik mit der Einführung von 10 GBit Technologie wird die Spezifikation für den physischen Layer auch auf InfiniBand ausgedehnt.

Bereits heute hat Fibre Channel mit der im August 2004 verabschiedeten 8 GBit/s zu 10 GBit Ethernet aufgeschlossen (höhere Nutzdatenraten). 2008 findet mit 10GFC auch bei Fibre Channel die 10 GBit/s Technik Einzug. Die versprochenen Geschwindigkeitsvorteile von iSCSI werden nicht von großer Dauer sein.

In naher Zukunft ist nicht zu erwarten, dass iSCSI Fibre Channel technologisch ablösen wird. Dazu sind noch zuviele Fragen ungelöst, nichtsdestotrotz kann iSCSi die Hemmschwelle für den Einsatz eines Speichernetzes herabsetzen und bei anhaltendem Trend zur Priorisierung von Storage (höherwertig/niederwertig) sogar dem Einsatz von Fibre Channel den Weg ebnen.

5. Quellen

[1] iSCSI RFC and Drafts
www.haifa.il.ibm.com/satrans/ips

[2] RFC3347 – Small Computer Systems Interface protocol over the Internet (iSCSI) Requirements and Design Considerations

[3] RFC 3720 – Internet Small Computer Systems Interface (iSCSI)

http://www.faqs.org/rfcs/rfc3720.html

[4] RDMA
http://www.rdmaconsortium.org

[5] iSer (iSCSI RDMA Extension)
http://www.rfc-editor.org/internet-drafts/draft-ietf-ips-iser-00.txt

[6] iSCSI: The Future of Network Storage:
http://www.extremetech.com/article2/0,1558,57869,00.asp

[7] Fibre Channel Industry Association
http://www.fibrechannel.org/