

Storage Over IP II

Implementace iSCSI v komerčních zařízeních

(Nishan IPS3300)

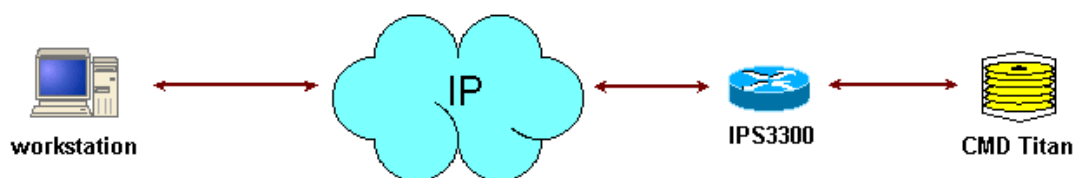
Jan Haluza, Filip Staněk
Technická zpráva CESNETu č. 12/2002
28. října 2002

Úvod

Technologie iSCSI, popsána v příslušném IETF [1] dokumentu, zapouzdřuje SCSI komunikaci do IP protokolu, čímž umožňuje přistupovat přes existující IP síť k SCSI zařízením a tímto implementovat tzv. vzdálený datový sklad. V dnešní době je již tato technologie standardizována několika komerčními výrobci. Mezi ně patří i firma Nishan Systems se svým výrobkem Nishan IPS3300 [2]. Tento model byl testován z pohledu iSCSI v roli target (server). Toto zařízení překládá iSCSI požadavky iniciátorů (klientů) na SCSI požadavky, které dále předává připojenému SCSI zařízení např. diskovému poli a zpětně výsledná data (opět zaobalena v iSCSI) zasílá iniciátoru.

Tato zpráva popisuje zkušenosti s prací se zařízením Nishan IPS3300 s připojeným diskovým polem v roli target spolu s iniciátorem postaveném na standardním PC s operačním systémem třídy Linux.

Konfigurace systému



Veškerá zde popsaná měření byla prováděna na následujících zařízeních.

HW konfigurace:

target:

- Nishan Systems IPS3300, IP Storage Switch
- CMD Titan CRD-7220, diskové pole
- SCSI disk, Compaq HD00431731, rev. 3208, 4 GB

initiator:

- PC Intel Pentium III 500 MHz s 512 kB cache, 384 MB paměti RAM, NetGear GA620 Gigabit Ethernet, SCSI řadič Adaptec 2940U2/W, hard disk Seagate ST 136403LW (Typ Direct-Access, ANSI SCSI rev. 02, 80 MB/s transfers)

SW konfigurace:

target:

- systém
Software Version: 3.90.0.20020920
Firmware Version: 0.3.4
JUMBOFRAME: DISABLE

- SNS
SNS Role(current): SERVER
Version: 10
Priority: 0
SNS Communication Type: BROADCAST
Broadcast Port Number: 50000

initiator:

- sestava označována jako "rave":
OS Linux Debian 3.0 (Woody), jádro 2.4.19-pre1-ac1
- sestava označována jako "ant":
OS Linux Redhat 7.3 (Valhala), jádro 2.4.18-3 (Redhat)

iSCSI switch byl s diskovým polem propojen pomocí Fibre Channelu. Disk byl pro potřeby testu rozdělen na čtyři partice o velikosti 1 GB. Klientská PC byla k iSCSI switchi připojena přímo přes gigabitový ethernet switch bez dalších mezičlánků, které by mohly mít vliv na výkon systému.

Na straně klientských PC v roli iniciatora byla použita software implementace od firmy Cisco [3] linux-iscsi. Ostatní implementace se ukázaly jako nepoužitelné, kvůli neshodné verzi iSCSI protokolu.

Pro potřeby měření výkonu systému byl použit benchmark IOzone [4].

Initiator

Software pro provoz funkce iniciatora byl balík *linux-iscsi*. Zdrojové texty jsou k dispozici na webové adrese projektu Cisco [3], binární verze jak modulů linuxového jádra, tak i iSCSI démona (iscsid) a podpůrných skriptů jsou k dispozici v dnešních moderních linuxových distribucích firem RedHat, SuSE. Konfigurace této implementace byla popsána v předchozí zprávě Cesnetu, proto zde již nebude opětovně uváděna. Více viz. Cesnet [5].

Je třeba poznamenat, že tento iniciator pracoval oproti jiným implementacím naprosto uspokojivě.

Praktické zkušenosti:

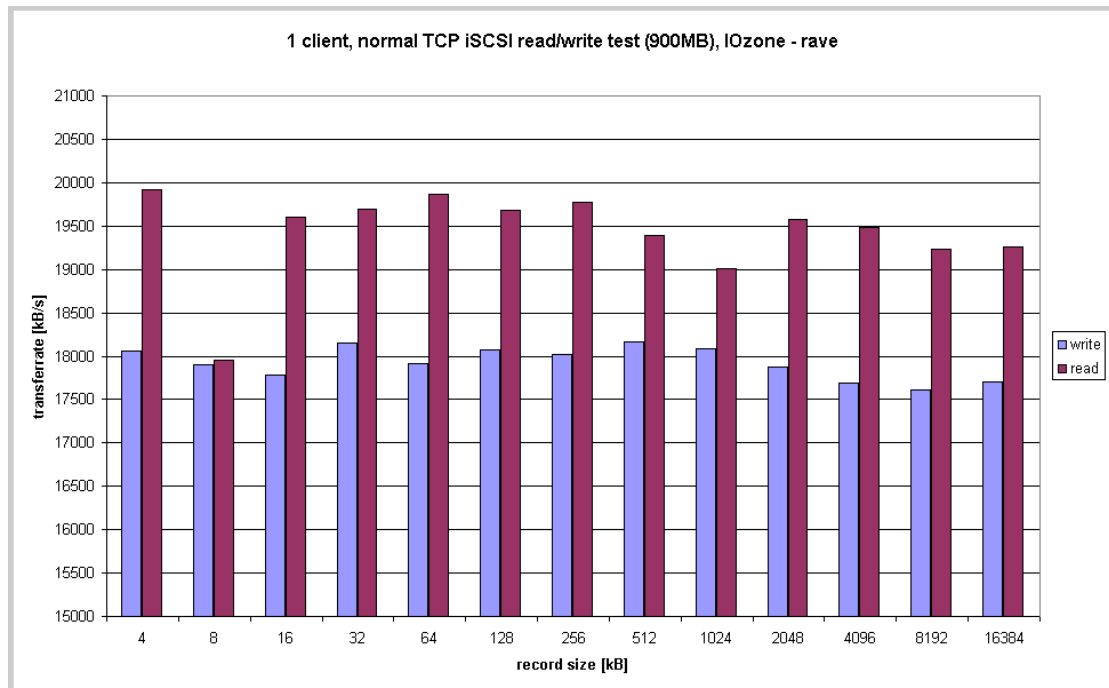
IPS3300 se dá řídit přes sériovou konzoli, pomocí webového Java appletu (verze 1.3.1), nebo pomocí přiloženého software SANvergence Manger. Přístup je tudíž platformně nezávislý. IPS3300 obsahuje osm portů dual mode Fibre Channel/Gigabit Ethernet. Porty 7 a 8 se dají nakonfigurovat pro použití s protokolem iSCSI. Zbylé porty slouží pro připojení zařízení přes FPC (Fibre Channel Protocol).

Výkon:

Měření probíhala na sestavě *rave*, pomocí benchmarkovacího programu IOzone [4] s následujícími parametry:

```
iozone -Rb iscsi.wks -n 900m -g 900m -z -c -a
```

Tedy s výstupem do binárního souboru ve formátu MS Excel, s minimální a maximální velikostí souboru 900 MB, s použitím funkce close() a automatickým měřením, s velikostmi přenášených bloků od 4 kB do 16 MB.



| record size | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 | 2048 | 4096 | 8192 | 16384 |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| write | 18060 | 17906 | 17780 | 18156 | 17909 | 18072 | 18020 | 18161 | 18084 | 17874 | 17693 | 17612 | 17708 |
| read | 19916 | 17956 | 19602 | 19700 | 19865 | 19687 | 19774 | 19390 | 19011 | 19579 | 19480 | 19228 | 19260 |

Změna velikosti TCP okna

Změna velikosti TCP okna jak pro čtení, tak i zápis oproti standardním hodnotám linuxového jádra (verze 2.4.18 a 2.4.19) neměla na výkon čtení / zápisu podstatný vliv. Změna byla provedena u default velikosti bufferu TCP socketů pro zápis (z 16 kB na 64kB) a čtení (z 85 kB na 1MB).

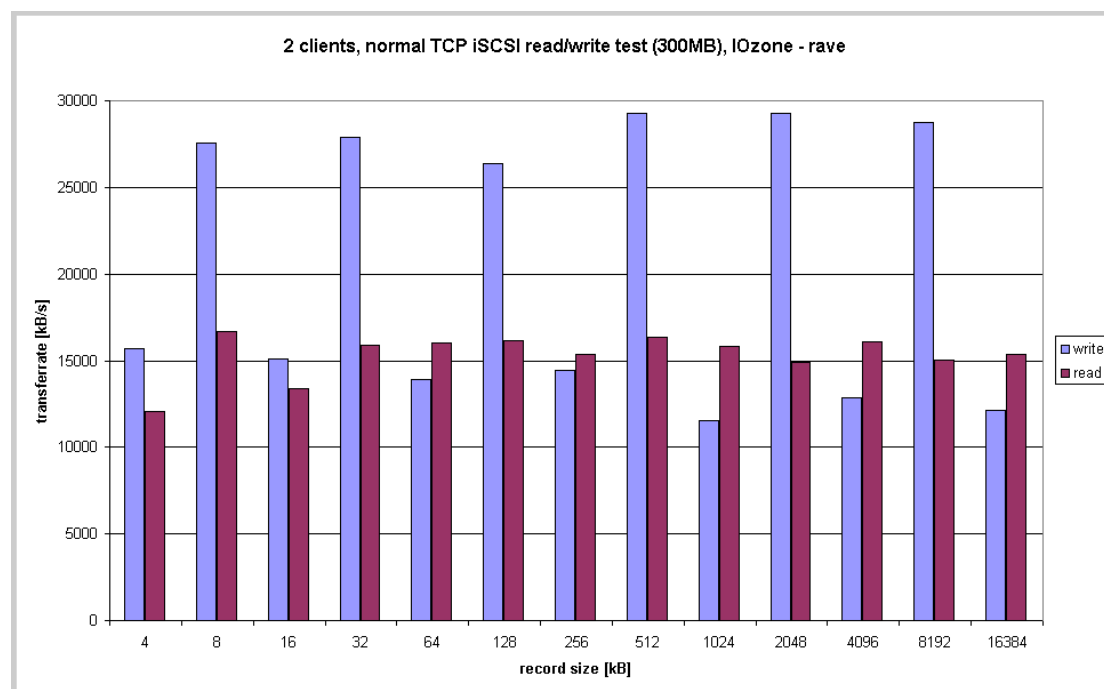
Dva klienti současně

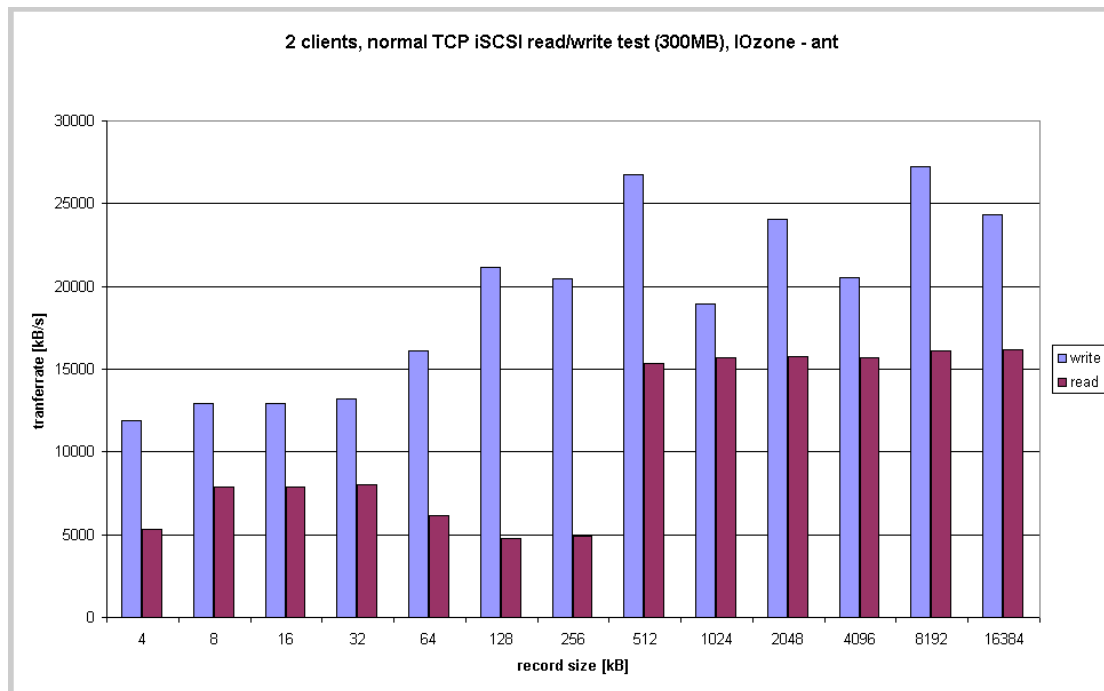
Pro měření přístupu dvou klientů současně, byla jako druhý klient použita sestava *ant*. U těchto měření byly z časových důvodů k testování použity tyto parametry:

```
iozone -Rb iscsi.wks -n 300m -g 300m -z -c -a -i 0 -i 1
```

Velikost souboru byla snížena na 300 MB a byly provedeny pouze testy na čtení a zápis. Je třeba si povšimnout zvýšení výkonu, které je ovšem způsobeno právě snížením velikosti přenášeného souboru, dále také jistého „poskakování“ výkonu, pravděpodobně způsobeného vynecháním některých mezitestů.

Zde se ukázalo, že díky nastavení Write/Back zápisu na diskovém poli, dopadlo podstatně lépe současné zapisování dvou klientů. Při Write/Back zápisu zasilá pole klientovi notifikaci o zápisu na disk v okamžiku kdy obdrží data i když v tomto momentu data ještě nemusejí být fyzicky uložena na disku. Čtení dvou klientů na jednom disku totiž způsobilo rozdělení výkonu.





rave

| record size | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 | 2048 | 4096 | 8192 | 16384 |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| write | 15665 | 27591 | 15087 | 27894 | 13896 | 26366 | 14433 | 29247 | 11556 | 29272 | 12848 | 28774 | 12127 |
| read | 12066 | 16710 | 13371 | 15917 | 16034 | 16184 | 15370 | 16369 | 15796 | 14876 | 16109 | 15010 | 15361 |

ant

| record size | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 | 2048 | 4096 | 8192 | 16384 |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| write | 11923 | 12933 | 12926 | 13194 | 16087 | 21167 | 20462 | 26774 | 18909 | 24051 | 20551 | 27216 | 24353 |
| read | 5337 | 7900 | 7867 | 8022 | 6183 | 4803 | 4913 | 15355 | 15720 | 15791 | 15692 | 16097 | 16207 |

Závěr:

Zařízení Nishan IPS3300 se jeví jako použitelné technické řešení pro realizaci vzdáleného skladu dat (spolu s patřičným diskovým polem) pomocí protokolu iSCSI.

Z testovaných verzí firmware byla pro iSCSI použitelná pouze poslední z uvedených verzí. U této verze nebyla však použitelná část týkající se SNS a tak nemohla být odzkoušena funkčnost tohoto protokolu a jeho implementace v zařízení.

Pro podrobnější analýzu nerovnoměrností při čtení z disku dvěma klienty současně by musela být provedena podrobnější analýza přesahující záruční dobu zařízení.

Literatura:

- [1] IETF „iSCSI, Internet Draft“ , <<http://www.ietf.org/internet-drafts/draft-ietf-ips-iscsi-12.txt>>, April 17, 2002
- [2] Nishan „IPS3000 Series: Multiprotocol IP Storage Switch“, <http://www.nishansystems.com/products/prods_3000.html>, November 21, 2001
- [3] Cisco „linux-iscsi“, <<http://linux-iscsi.sourceforge.net/>>, May 2002
- [4] IOzone „IOzone Filesystem Benchmark“, <<http://www.iozone.org/>>, June 2002
- [5] Cesnet „Storage over IP“, <<http://www.cesnet.cz/doc/techzpravy/2002/iscsi/iscsi.pdf>>, June 18, 2002

Příloha:

úplný výpis konfigurace Nishan IPS3300

Nishan Systems IPS 3300 CLI Login (SW Rev 3.90.0.20020920)
(Mgmt IP : 158.196.158.127, Build Date Sep 20 2002, 07:03:57)

```
IPS Switch(show)# mgmt
Mgmt port address current      : 158.196.158.127
Mgmt port mask current        : 255.255.255.0
Mgmt port gateway current     : 158.196.158.1
Mgmt port address on next reset : 158.196.158.127
Mgmt port mask on next reset  : 255.255.255.0
Mgmt port gateway on next reset : 158.196.158.1
Mgmt port MAC address         : 00:01:0F:01:00:C0
Mgmt Port LED                 : off
```

```
IPS Switch(show)#
Port Num                       : 7
Operational State              : UP
Port Type                      : GE iSCSI Layer 2
Port Name                      : Port 7
Port Speed                     : GIGABIT
Port MAC Address               : 00:01:0F:01:00:C7
Port Enabled                   : ENABLED
Port LED                       : OFF
Port Autonegotiation           : ENABLED
```

```
Layer 2 Properties
IP Address(TCP)                : 195.113.113.28
Mask(TCP)                     : 255.255.255.240
External Router Address(TCP)   : 195.113.113.17
Internal SAN Address(TCP)      : 158.196.33.2
```

```
Advanced TCP Port Properties
AutoReset                     : ENABLE
MtuSize                       : 1500
FastWrite                     : OFF
AssignMtu                     : AUTO
Compression                   : OFF
```

```
IPS Switch(show port)# config 1
```

```
Port Num                       : 1
Operational State              : UP
Port Type                      : FC Auto
Port Name                      : Port 1
Port Speed                     : GIGABIT
```

Port Enabled : ENABLED
Port LED : OFF
Port Autonegotiation : DISABLED

FC Properties

Type : FC_Auto
Port WWN : 20:00:00:01:0f:01:00:c1
Port ED_TOV(cur) : 2
Port ED_TOV(next) : 2
Port RA_TOV(cur) : 10
Port RA_TOV(next) : 10

IPS Switch(show)# sns

SNS Role(current) : SERVER
SNS Role(next reset) : SERVER
Version : 10
Priority : 0
SNS Communication Type : BROADCAST
MCAST Address : 225.1.1.20
Broadcast Port Number : 50000
Primary SNS Address : 158.196.33.1

IPS Switch(show)# system

System LED : OFF
System Date : Month: 10 Day: 2 Year: 2002
System Time : Hour: 15, Min: 13, Sec: 2
System Name :
System Contact :
System Location :
System Product Name : IPS 3000 Series
System Model Name : IPS 3300
System PCA Serial No : FL-0211-000015
System Product Serial No : USF21300313300-01
System Software Version : 3.90.0.20020920
System Firmware Version : 0.3.4
Switch IP Address(current) : 158.196.33.1
Switch Subnet Mask(current) : 255.255.255.0
Switch Gateway(current) : 158.196.158.1
Switch IP Address(Next Reset) : 158.196.33.1
Switch Mask(Next Reset) : 255.255.255.0
Switch Gateway(Next Reset) : 158.196.158.1
Switch MAC Address : 00:01:0F:01:00:C1
System Running Time : Hour: 45 Minute: 54 Sec: 15
System JUMBOFRAME : DISABLE