

Network based Storage Solutions

G. Sivakumar

Indian Institute of Technology, Bombay

<http://www.cse.iitb.ac.in/~siva>

Acknowledgement: Some slides/information are adapted from the tutorials provided by Storage Network Industry Association at www.snia.org



Goals of Shared Storage

- Increased flexibility in deploying and managing storage,
- Improved quality of service, and
- Increased operational efficiency.

using the following elements

- Physical Storage Devices (disk drives, disk arrays, controllers, tape libraries,...)
- Logical Storage Resources (volumes, files)
- Interconnection Network (many choices)
- Host Computers (equipped with Host Bus Adaptors, Network Cards)



A Sun Storedge NAS server (2 Terabytes disk) at IIT.

- 300 MB for students, 500 for faculty.
- Mountable from anywhere in campus
- Across platforms (Linux/Windows)
- **smbd** server
- LDAP authentication
- Backed-up using AMANDA
- Also Provides homepages for all
(<http://homepages.iitb.ac.in/~user>)



Logical Volumes/Partitions

Internally *bighome* has a *fibre channel* connected set of disks (Hitachi and Seagate)

```
bash-2.03# df -k
Filesystem            kbytes    used   avail capacity  Mounted on
/dev/vx/dsk/rootvol   4032504 2873271 1118908    72%      /
/dev/vx/dsk/usr       4032504 1633010 2359169    41%     /usr
/proc                  0         0         0     0%     /proc
fd                     0         0         0     0%   /dev/fd
mnttab                 0         0         0     0%  /etc/mnttab
/dev/vx/dsk/var       4032504  944525 3047654    24%     /var
swap                   21775512      32 21775480     1%   /var/run
swap                   21838944  63464 21775480     1%    /tmp
/dev/dsk/c5t2d0s4     1016122 148331  806824    16%    /nsr
/dev/vx/dsk/opt       2734129 1025099 1654348    39%    /opt
/dev/vx/dsk/data3/vol03
                        571785216 92363100 449516693    18%   /data3
/dev/vx/dsk/data1/vol01
                        571785216 176296565 370778827    33%   /data1
/dev/vx/dsk/data4/vol04
                        500312064 181497552 298889755    38%   /data4
/dev/vx/dsk/data2/vol02
                        500312064 162779319 316436950    34%   /data2
```



User Distribution

```

bash-2.03# ls /data1/
1994                2000                2004                alumni.old_batches  misc                quotas
1997                2001                TT_DB              deleted             other              quotas
1998                2002                acre               fac                 prjstf            rs
1999                2003                admin             lost+found          quotadir          stf

bash-2.03# ls /data1/2004/
btech dd      im      mdes  mm      msc      mtech  rs

bash-2.03# ls /data1/2004/mtech/
aero      cese      civil      cse      ese      ieor      me      re
bioschool che      cor      ee      geos      it      met      sce

bash-2.03# ls /data1/2004/mtech/cse/z4305
z4305001 z4305015 z4305024 z4305028 z4305802 z4305807 z4305905
z4305011 z4305020 z4305025 z4305029 z4305804 z4305813

bash-2.03# ls -ld /data1/2004/mtech/cse/z4305*
drwxr-xr-x 2 aniketpkate 104305      96 Jul 21 18:18 /data1/2004/mtech/cse/z4305001
drwxr-xr-x 2 dheren 104305      96 Jul 21 18:19 /data1/2004/mtech/cse/z4305011
drwxr-xr-x 2 vamshi 104305      96 Jul 21 18:19 /data1/2004/mtech/cse/z4305015
drwxr-xr-x 5 z4305020 104305      1024 Sep 30 01:53 /data1/2004/mtech/cse/z4305020
drwxr-xr-x 2 z4305024 104305      96 Jul 21 18:19 /data1/2004/mtech/cse/z4305024
drwxr-xr-x 2 sheetal_sonare 104305      96 Jul 21 18:19 /data1/2004/mtech/cse/z4305025
drwxr-xr-x 2 sr_anilk 104305      96 Jul 21 18:19 /data1/2004/mtech/cse/z4305028
drwxr-xr-x 2 sreekanth 104305      96 Jul 21 18:19 /data1/2004/mtech/cse/z4305029
drwxr-xr-x 2 nadesai 104305      96 Jul 21 18:19 /data1/2004/mtech/cse/z4305802
drwxr-xr-x 2 sjagdhesh 104305      96 Jul 21 18:19 /data1/2004/mtech/cse/z4305804
drwxr-xr-x 2 amruta 104305      96 Jul 21 18:19 /data1/2004/mtech/cse/z4305807
drwxr-xr-x 2 vselot 104305      96 Jul 21 18:19 /data1/2004/mtech/cse/z4305813
drwxr-xr-x 2 samveen 104305      96 Jul 21 18:19 /data1/2004/mtech/cse/z4305905

```



Administering Bighome

- *Automagically* when LDAP accounts are added.

Storage Server Administration

LDAP Administration

EMail Administration

Storage Server Administration : (bighome.iitb.ac.in)

- When you add / delete a user home-directories are automatically created / deleted
- Even the **changeuid** script takes care of renaming the home directories.
- If one has to do something manually take care of the following :

1. Attribute "homeDirectory" on ldap has the correct current value.
2. It has the format "/home/dept/group/userid" where dept=(cse, aero, cc, etc ...) Note : dept name in lower case group = (fac, stf, prjstf, rs, misc) and for students : group = (bt01, mt02, dd01 ...etc ...) For exact details on this read the scripts "addstudent.cgi" "addstaff.cgi" etc ...on web.iitb.ac.in
3. Administration Scripts on bighome.iitb.ac.in are : 1. [delhomedir.pl](#) 2. [mkhomedir.pl](#) 3. [renamehomedir.pl](#) in "/usr/local/bin" . Usage is shown when you run the script. For detailed logic you can read the scripts.
4. The directories are actually created randomly on the four partitions data1 ...data4 and appropriate link is crated in /home/.... as given above.
5. Quotas are specified for users ugqta, pgqta, rsqta, facqta, stfqta, prjstfqta, miscqta. Depending on the categories, the appropriate quota is copied for a user.



How Home Directories Created

```
if ($status eq "misc" || $status eq "stf" || $status eq "prjstf" || $status eq "fac") {
    $data = "/data"."$data/$dirs[2]/$dirs[3]/$user";} else {
    $data = "/data"."$data/$dirs[2]/$dirs[3]/$dirs[4]/$user";
}

#print LOG "mkhomedir.pl @dirs[2] @dirs[3]\n";
print LOG "mkhomedir.pl link_name=$dir real_homedir=$data\n";
my $basedir = join '/', @dirs;
if (!chdir ($basedir)) {
    system ("/usr/bin/mkdir -m u=rwx,og=rx -p $basedir") && die "Couldnt create $basedir";
}

system ("/usr/bin/mkdir -m u=rwx,og= -p $data") && die "Couldnt create $data";
system ("/usr/bin/chown $uid:$gid $data") && die "couldnt set permissions for $data";
system ("/usr/bin/ln -fs $data $dir") && die "Cant set symlink to $dir";
system ("/opt/VRTSvxfs/sbin/vxedquota -p $qta $user") && die "couldn't set quotafor $user";
```



Configuring SMBD

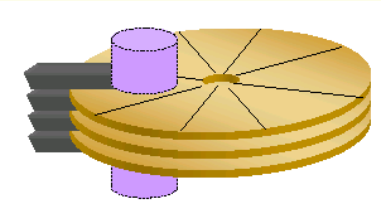
```
bash-2.03# less /usr/local/lib/smb.conf
# Samba config file created using SWAT
# from pitta.cc.iitb.ac.in (144.16.106.14)
# Date: 2002/09/30 16:42:17

# Global parameters
[global]
    workgroup = COMPUTER-CENTRE
    netbios name = BIGHOME
    interfaces = ge0
    encrypt passwords = Yes
    os level = 65
    preferred master = True
    domain master = True
    wins server = 10.200.1.21
    ldap server = ldap.iitb.ac.in
    ldap port = 389
    ldap suffix = ou=people,dc=iitb,dc=ac,dc=in
    ldap admin dn = cn=sambaproxy,ou=people,dc=iitb,dc=ac,dc=in
    ldap ssl = no

[homes]
    read only = No
```



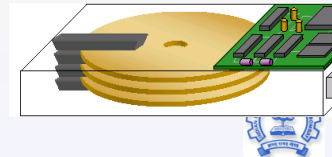
Physical Disks



- Head
- Cylinder
- Sector

Data Transfer speed depends on

- Speed of I/O technology (SCSI, Fibre Channel, ...)
- Seek time (time for heads to move to a new track)
- Latency (time for sector to rotate under heads)
- Speed and size of disk drive



Drive Interface

- Parallel IDE/ATA or Serial ATA (SATA)
 - Integrated Drive Electronics (IDE), AT Attachment (ATA)
 - lower-cost devices than SCSI,
 - controller/drive electronics wedded to the drive itself.
 - IDE standardizes how disk drives are connected to servers
 - via a ribbon cable and an interface connector on a PC or server motherboard.
 - IDE/ATA protocol five+ revisions,
 - For 10 years now speeds and feeds keeping pace with speed- and capacity-hungry applications.
 - Little improvement in transfer rate expected in future
 - Difficult to create arrays with Parallel
- Parallel SCSI or SAS (Serial Attached SCSI)
- Fibre Channel



Serial ATA

SATA is

- a standards-based interface that improves on parallel IDE/ATA
- interface transfer rate is 150Mb/s, (compared with parallel ATA's 100Mb/s)
- That's a 50% with large data volume requirements, such as video editing.
- First Generation SATA drives not yet more efficient than Parallel ATA,
- SATA-2 and SATA-3 - expected transfer rates of 300 and 600 Mb/s
- Still limited by max speed of PCI bus at 133 Mb/s

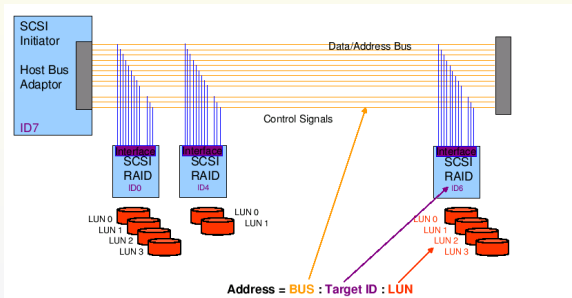


SCSI

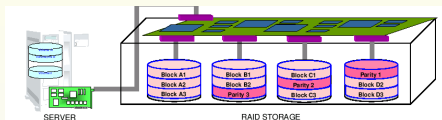
- Parallel interface I/O technology
 - Maximum cable run length is 25 meters
 - Speeds up to 320 MB/sec (Ultra SCSI-3 on 16 bit wide bus)
 - Maximum of 16 (I/O controller + devices) SCSI devices per bus
- Evolving standard: SCSI-1, SCSI-2, SCSI-3, SCSI-4 (1.2Gbps)
- Unaware of File or Operating Systems
- Block Data = SCSI Block I/O
- Scalability and distance limitations



SCSI Protocol



RAID

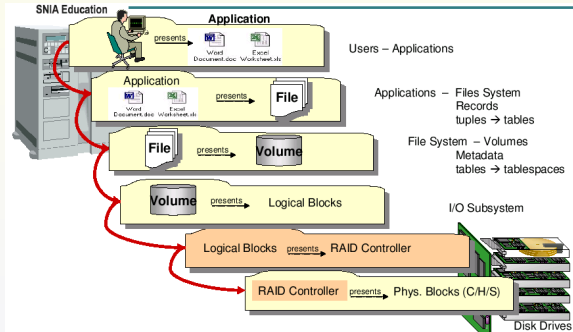


Redundant Array of Independent Disk

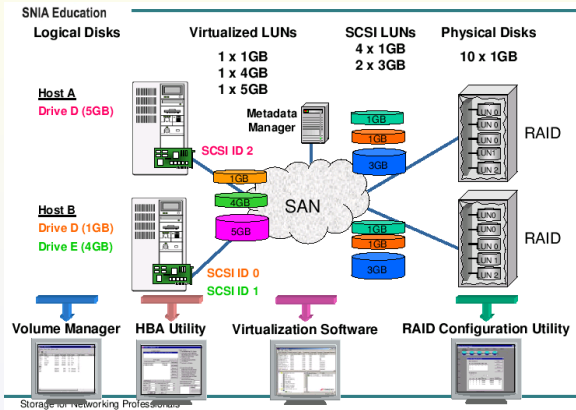
- RAID 1–Mirrored Volumes
- RAID 0+1–Mirrored Array
- RAID 4–Block-Level Striping with Parity Disk
- RAID 5–Striping with Distributed Parity
- RAID 10–Mirrored Striping Array



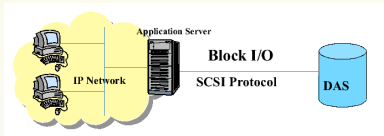
Logical View of Storage



Storage Virtualization



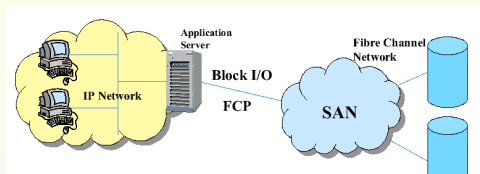
Direct Attached Storage (DAS)



- Storage is captive *behind* server CPU
- Data access is file system and platform dependant.
- Server CPU must handle user I/O requests also.
- Costly to scale; complex to manage



Storage Area Network (SAN)

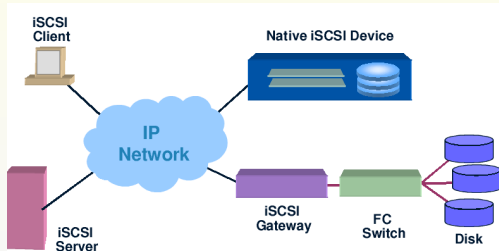


Scalability and Performance

- Storage Expansion No impact on servers
- Server Expansion No impact on storage
- Load Balancing & Failover Active parallel paths
- Bandwidth on Demand Robust topology
- Offload/Remote Backup



iSCSI



- Everything over IP
- IP over Ethernet



iSCSI (ctd.)

iSCSI is a SCSI transport protocol for mapping of storage block I/O over TCP/IP networks

- Allows universal access to storage over standard TCP/IP networks
- iSCSI ratified by IETF February 2003 (RFC TBD)
- Embedded Payload of TCP Segment
 - Transparent to TCP/IP
 - Transparent to Firewalls, VPN, Routers, etc.
- Topologies: Intranet, Internet, WAN, Wireless

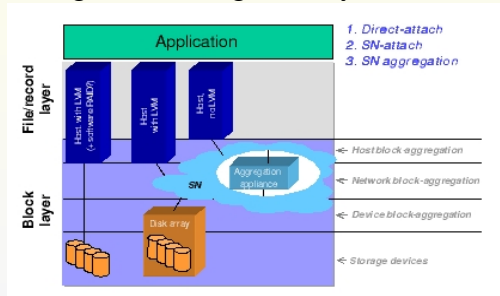


Why IP Access to Storage?

- Single access technology for both block and file
- Ubiquitous nature of IP
 - Access storage from LAN, MAN, WAN, Internet
 - Economy of scale
- Scalability
 - Distance, node count, performance (GE/10GE), expertise
- Manageable, secure and interoperable
 - (SNMP, MIBs, DNS, LDAP, tools)
 - (IPSEC, VLANs, ACLs, Firewalls)
- Substantial R&D investment in Ethernet/IP



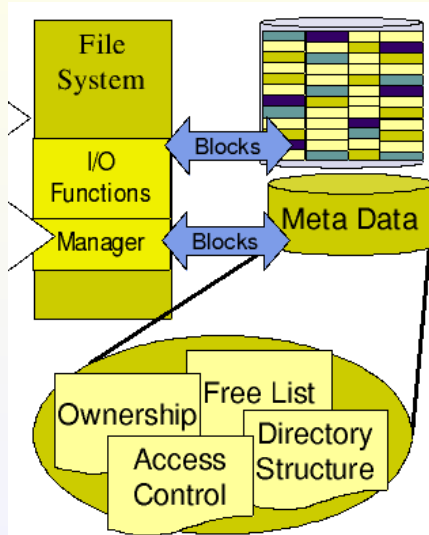
Storage Networking Industry Association reference model.



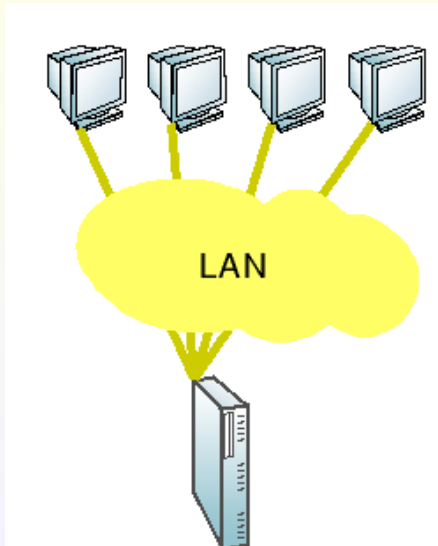
- Block I/O abstraction
- How to share across computers?



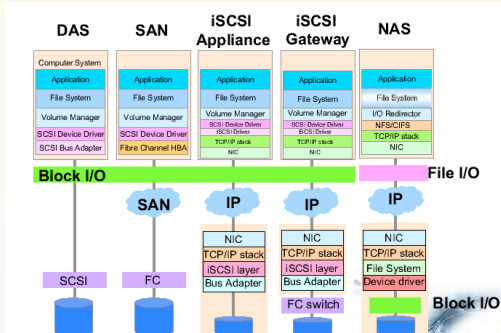
File System Abstraction



Network Attached Storage (NAS)



Comparison of DAS-SAN-iSCSI-NAS



Beyond SAN and NAS

- SAN offers high performance and scalability
- NAS is much better for cross-platform file sharing
- How to get the best of both?
 - SAN with NAS-head
 - NAS with SAN for internal storage

Recent research however feels that all **block** based storage interface methods have reached a plateau. So, what next?



Object Based Storage

- Variable size containers (can grow/shrink dynamically)
- Used to store any type of data (files, database records, images, multimedia)
- User-accessible attributes/methods.
- Device managed *metadata*

Allows cross-platform data sharing with application level security.
And scope for device *intelligence* (self-managed).

See August 2003 IEEE Communication Magazine for excellent survey.

