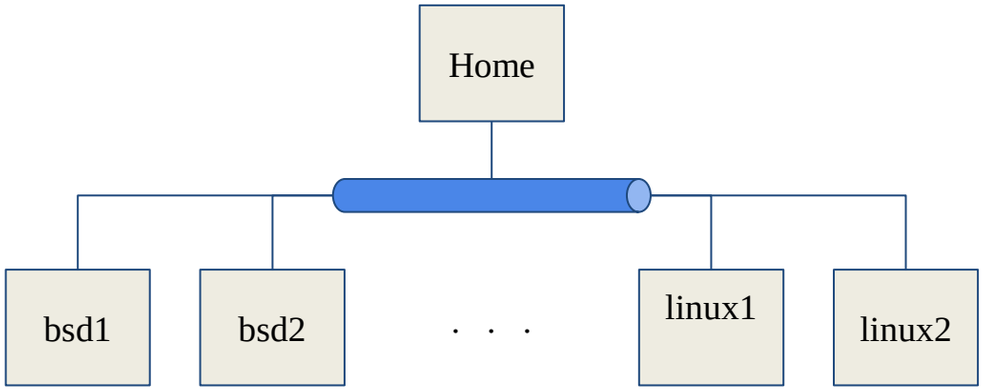
The Network File System

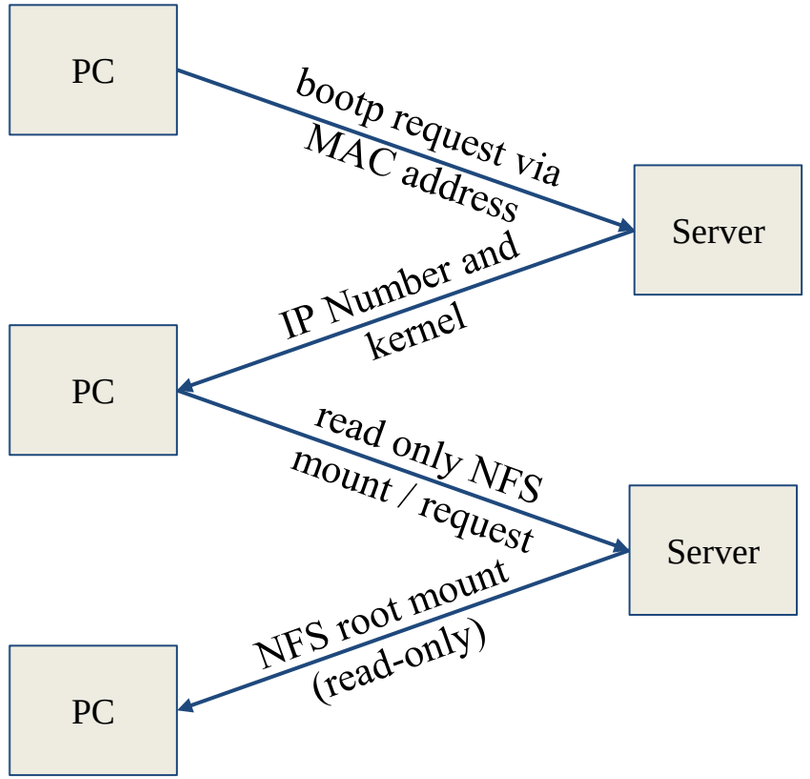
lwhsu (2019-2020, CC BY)
? (?-2018)

交大資工系資訊中心
Computer Center of Department of Computer Science, NCTU

NFS

- Share filesystem(s) to other hosts via network
- NFS History
  - Introduced by Sun Microsystems in 1985
  - Originally designed for diskless client-server architecture





The PC then starts the appropriate X-Server using the MAC address as a key

Components of NFS – mounting protocol (1)

- NFSv1
  - In-house experiments in Sun
- NFSv2
  - Synchronous write
  - V2 NFS server must commit each modified block to disk before replying to NFS client
  - Cause long delay when there is a NFS write operation
  - UDP
- NFSv3 in 1990s
  - Asynchronous write
  - Provide increase performance and better support for large files
  - TCP support

Components of NFS – mounting protocol (2)

- NFSv4 in 2003s
  - Influenced by AFS and SMB/CIFS
  - NFSv4 ACL
  - Stateful protocol
  - Unicode support
  - Only port 2049 is used

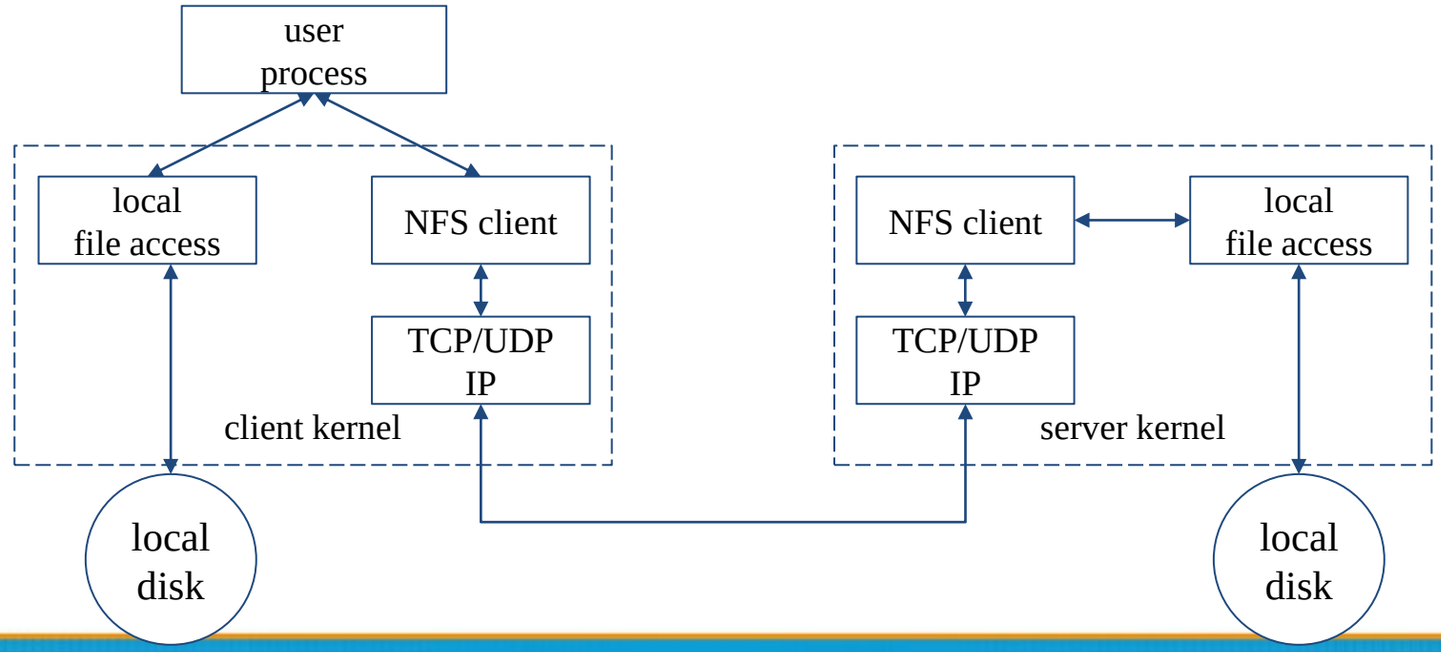
- NFSv4.1 in 2010
  - pNFS, parallel access, distributed servers
  - Multipathing

Components of NFS – mounting protocol (3)

- NFSv4.2 in 2016
  - Minor revision to NFSv4.1, adds some optional features
    - lseek (SEEK_DATA/SEEK_HOLE)
    - posix_fallocate()
    - posix_fadvise (POSIX_FADV_WILLNEED/POSIX_FADV_DONTNEED)
    - Server side copy of byte ranges between two files on the same NFS mount
    - point when the copy_file_range(2) syscall is used.
    - Extended attribute support as specified by RFC-8276.

Components of NFS – mounting protocol (4)

- Sun's Open Network Computing (ONC) Remote Procedure Call (RPC) distributed computing standards
  - NFS client → RPC → Transport Layer → …
  - Transport Layer
    - UDP: Lack congestion control
    - TCP: become more suitable



Components of NFS

- Including
  - Mounting Protocol
  - Mount Server
  - Daemons that coordinate basic file service
  - Diagnostic utilities

Components of NFS – Server-side NFS (1)

- NFS Server
  - Export sharing filesystem
    - System dependent
  - Waiting for “mount request”
    - mountd (rpc.mountd) daemon
  - Waiting for “file access request”
    - nfsd (rpc.nfsd) daemon
  - Lock the files being accessed (optional)
    - lockd (rpc.lockd) daemon
  - Check the correctness of the files (optional)
    - statd (rpc.statd) daemon

Components of NFS – Server-side NFS (2)

- Exporting filesystem
  1. Edit export configuration file
     - Each line is “what to export and how”
  2. Reload related daemons

System	Exports info file	How to reload
FreeBSD	/etc/exports	/etc/rc.d/mountd reload
Linux	/etc/exports	/usr/sbin/exportfs -a
Solaris	/etc/dfs/dfstab	/usr/sbin/shareall
SunOS	/etc/exports	/usr/sbin/exportfs -a

Components of NFS – Server-side NFS (FreeBSD.1)

- Exporting filesystem
  - /etc/exports
    - White-space separated
    - Format: *directory-list options-list client-list*

Option	Description
-ro	Exports read-only, default is (read-write)
-alldirs	Allow any subdirectory to be mounted
-maproot=user	Maps root to the specified user.
-mapall=user	Maps all UIDs to the specified user.

Client	Description
hostname	Host name (ex: mailgate ccserv)
netgroup	NIS netgroups
-network -mask	-network 140.113.235.0 -mask 255.255.255.0

Components of NFS – Server-side NFS

(FreeBSD.2)

- Example of /etc/exports

```
/raid -alldirs -maproot=root mailgate ccserv backup
/raid -alldirs -maproot=65534 -network 140.113.209 -mask 255.255.255.0
/home -ro -mapall=nobody -network 140.113.235.0 -mask 255.255.255.0
/usr/src  /usr/obj -maproot=0 bsd_cc_csie
```

  - Network and mask cannot be in the same line with hosts and netgroups

- Reload daemons
  - % kill -1 `cat /var/run/mountd.pid`
  - /etc/rc.d/mountd restart
  - /usr/sbin/service mountd restart

Components of NFS – Server-side NFS (Linux.1)

- Exporting filesystem
  - /etc/exports
    - Format: *directory client-list-with-option*
    - E.g.: /home1 bsd1(ro)

Client	Description
hostname	Host name (ex: mailgate ccserv)
@netgroup	NIS netgroups
ipaddr/mask	CIDR-style specification (ex: 140.113.235.2/24)
Wild cards * ?	FQDN with wildcards (ex: bsd*.cs.nctu.edu.tw)

Components of NFS – Server-side NFS (Linux.2)

Option	Description
ro,rw	Read-only, Read-write (default)
rw=list	Hosts in the list can do rw, others ro only
root_squash	Maps UID 0 and GID 0 to the value of anonuid and anongid (default)
no_root_squash	Allow root access
all_squash	Maps all UID and GID to anonymous one
subtree_check	Check that the accessed file is in the appropriate filesystem and in the exported tree.
no_subtree_check	Disables subtree checking
anonuid=xxx	Related to root_squash
anongid=xxx	Related to root_squash
secure	Require remote access from privileged port
insecure	Allow remote access from any port
noaccess	Prevent access to this dir and it's subdir

Components of NFS – Server-side NFS

(Linux.3)

- Example of /etc/exports

```
/home1      ccsun*.csie.nctu.eud.tw(rw)
/home2      @sun_cc_csie(ro)  dragon(rw,no_root_squash)
/home       ccpc1(rw,all_squash,anonuid=150,anongid=100)
/ftp/pub    (ro,insecure,all_squash)
/users      *.xor.com(rw)
/users/evi  (noaccess)
```

- Run /usr/sbin/exportfs
  - % /usr/sbin/exportfs -a
  - Maintain /var/lib/nfs/xtab table which is read by mountd

Components of NFS – Server-side NFS (Solaris.1)

- Exporting filesystem
  - /etc/dfs/dfstab
  - Each line will execute “share” command to export one NFS
    - Format: *share –F nfs -o option-list directory*
    - E.g.: /home1 bsd1(ro)
- Run shareall command
  - % /usr/sbin/shareall

Client	Description
hostname	Host name (ex: mailgate ccserv)
netgroup	NIS netgroups
IP networks	@CIDR-style specification (ex: @140.113.235.2/24)
DNS domains	.xxx.yyy any host within the domain (ex: .nctu.edu.tw)

Components of NFS – Server-side NFS

(Solaris.2)

Option	Description
ro,rw	Read-only to all, Read-write to all
ro=list, rw=list	Hosts in the list can do ro/rw
root=list	Lists hosts permitted to access this filesystem as root. Otherwise, root access from a client is equivalent to by “nobody”
anon=xxx	Specify the UID to which root is remapped. Default is “nobody”
anongid=xxx	Related to root_squash
nosub	Forbids clients to mount subdirectories
nosuid	Prevents setuid and setgid from being created

Components of NFS – Server-side NFS (3)

- nfsd daemon
  - Handle NFS file access request from NFS clients
  - Number of nfsd's thread is important
    - Too small, some NFS requests' response will be delayed
    - Too large, load will be high
    - nfsd(8)
      - -n thread
      - --maxthreads --minthreads
- In FreeBSD
  - Specify nfsd options in /etc/rc.conf
    - nfs_server_enable="YES"
    - nfs_server_flags="-u –t –n 4"

Components of NFS – Client-side NFS (1)

- NFS Client
  - Mount NFS filesystem first
  - Access file under NFS filesystem
- mount command (mount_nfs(8))
  - [format]
    - mount [-o options] host:directory mount-point
  - E.g.,
    - % mount –t nfs ccbsd4:/home/www /home/nfs/www
- /etc/fstab (/etc/vfstab in Solaris)
  - % mount -a -t nfs (FreeBSD, Linux)
  - % mount -a -F nfs (Solaris)

```
# Device            Mountpoint  Fstype  Options      Dump  Pass#
dragon:/usr/man     /usr/man    nfs     ro,bg,soft   0     0
ccserv:/spool/mail  /var/mail   nfs     rw,bg,intr   0     0
```

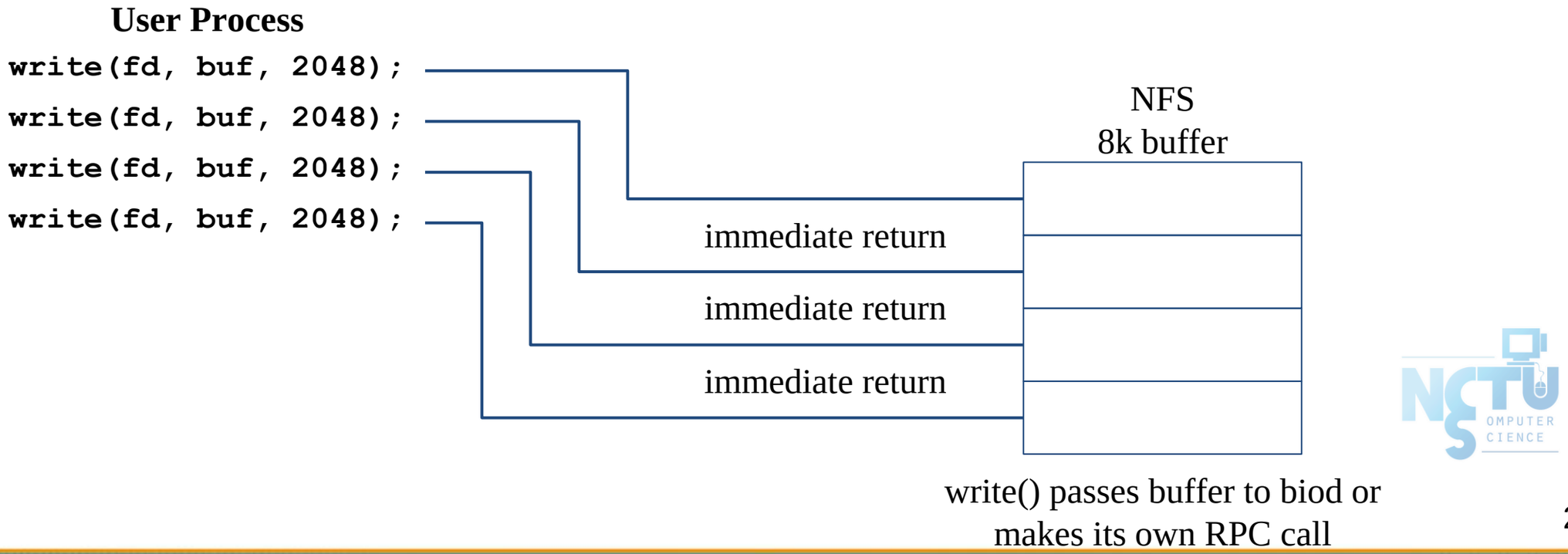
Components of NFS – Client-side NFS (2)

- NFS mount flags

Flag	Systems	Description
ro or rw	S,L,F	Mount the NFS as ro or rw
bg	S,L,F	If failed, keep trying in background
hard	S,L	If server down, access will keep trying until server comes back
soft	S,L,F	If server down, let access fail and return error
intr, nointr	S,L,F	Allow/Disallow user to interrupt blocked access
retrans=n	S,L,F	# of times to repeat a request before error return
timeo=n	S,L,F	Timeout period of requests (tens of seconds)
rsize=n	S,L,F	Set read buffer size to n bytes
wsize=n	S,L,F	Set write buffer size to n bytes
vers=n	S	Selects NFS v2 or v3
nfsv3,nfsv2	F	Selects NFS v2 or v3
proto=prot	S	tcp or udp
tcp	L,F	Select TCP. UDP is default

Components of NFS – Client-side NFS (3)

- Client side daemons that enhance performance
  - biod (block I/O daemon, or called nfsiod)
  - Perform read-ahead and write-behind caching
  - A sysctl wrapper now (vfs.nfs.iodmin & vfs.nfs.iodmax)



Components of NFS – NFS Utilities (1)

- nfsstat
  - Display NFS statistics
    - % nfsstat -s (display statistics of NFS server)
    - % nfsstat -c (display statistics of NFS client)

```
$ sudo nfsstat -c
Client Info:
Rpc Counts:
  Getattr    Setattr     Lookup  Readlink       Read      Write     Create     Remove
  1065253      34196     379742      5187     111699     182603      18049      29803
   Rename       Link    Symlink      Mkdir      Rmdir    Readdir  RdirPlus     Access
    20838       4746          1         10       1003       4705          0     316560
    Mknod     Fsstat     Fsinfo  PathConf     Commit
        0      13742       3889          0      75747
Rpc Info:
 TimedOut    Invalid  X Replies    Retries  Requests
        0          0         69       3994   2267773
Cache Info:
Attr Hits     Misses Lkup Hits     Misses BioR Hits     Misses BioW Hits     Misses
  1920497    1259363    1256973     379714     352854     102015     521158     182603
BioRLHits     Misses BioD Hits     Misses DirE Hits     Misses
   347749       5187      14996       4685       6137          0
```

Components of NFS – NFS Utilities (2)

- showmount
  - % showmount -e [host]
    - show the hosts' export list (localhost if not specified)
  - % showmount -a
    - List all mount points

```
$ showmount -e magpie
Exports list on magpie:
/home              ccduty mailgate 140.113.209.0
/drongo            operator ccduty mailgate 140.113.209.0
$ showmount -a
All mount points on localhost:
bsd1:/home2
bsd1:/raid/home
csduty:/home2
csduty:/raid/home
linux1:/raid/home
linux2:/raid/home
nat235.dynamic:/raid/home
sun1:/raid/home
```

NFS in FreeBSD

- NFS server

/etc/rc.conf

```
…
nfs_server_enable="YES"
nfs_server_flags="-u -t -n 4"
rpcbind_enable="YES"
mount_enable="YES"
…
```

- NFS client

/etc/rc.conf

```
…
nfs_client_enable="YES"
…
```

NFS and ZFS

- No need to edit /etc/exports

/etc/rc.d/mountd

```
if checkyesno zfs_enable; then
        rc_flags="${rc_flags} /etc/exports
/etc/zfs/exports"
fi
```

- zfs(8)

```
sharenfs=on | off | opts
Controls whether the file system is shared via NFS, and what options
are used. A file system with a sharenfs property of off is managed the
traditional way via exports(5). Otherwise, the file system is
automatically shared and unshared with the "zfs share" and "zfs
unshare" commands. If the property is set to on no NFS export options
are used. Otherwise, NFS export options are equivalent to the contents
of this property. The export options may be comma-separated.
See exports(5) for a list of valid options.
When the sharenfs property is changed for a dataset, the mountd(8)
daemon is reloaded.
```

NFSv4

- Server
  - /etc/rc.conf
    - nfsv4_server_enable="YES"
  - /etc/exports
    - V4: / …
      - Specify the NFSv4 tree root.
      - Still need to specify files systems in other lines, as in v2 or v3
- Client
  - /etc/rc.conf
    - nfscbd_enable="YES"
      - Client side callback daemon
        - nfsv4(4), pnfs(4)
- nfsv4(4), pnfs(4)

Performance & Security

- Jumbo Frames
  - interface and switch/router both need to support
  - `ping -D -s <packetsize>`
  - `ping -D -g <sweepminsize> -G <sweepmaxsize>`
  - `ifconfig em0 mtu <size>`
- Firewall
- Storage Network