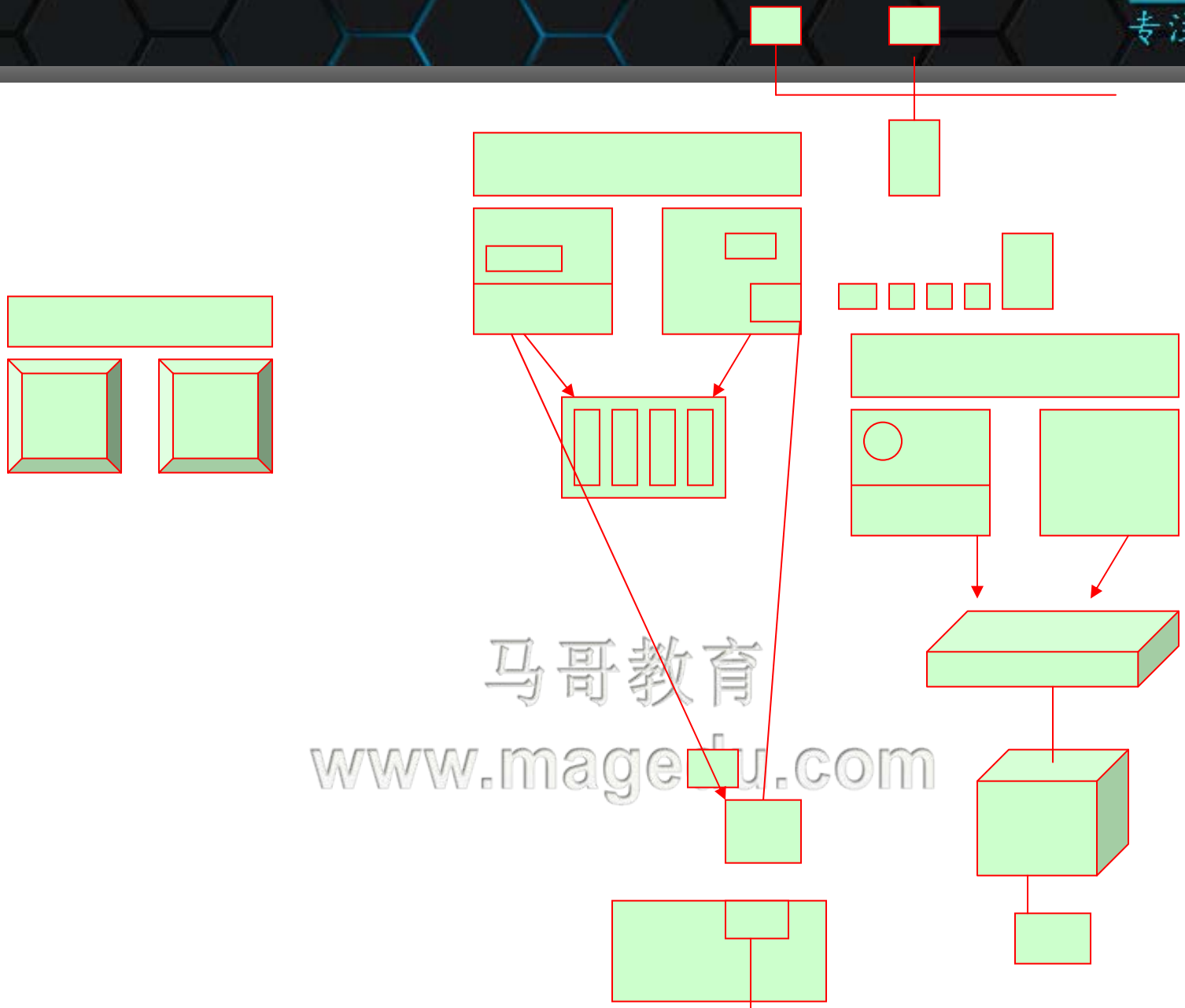
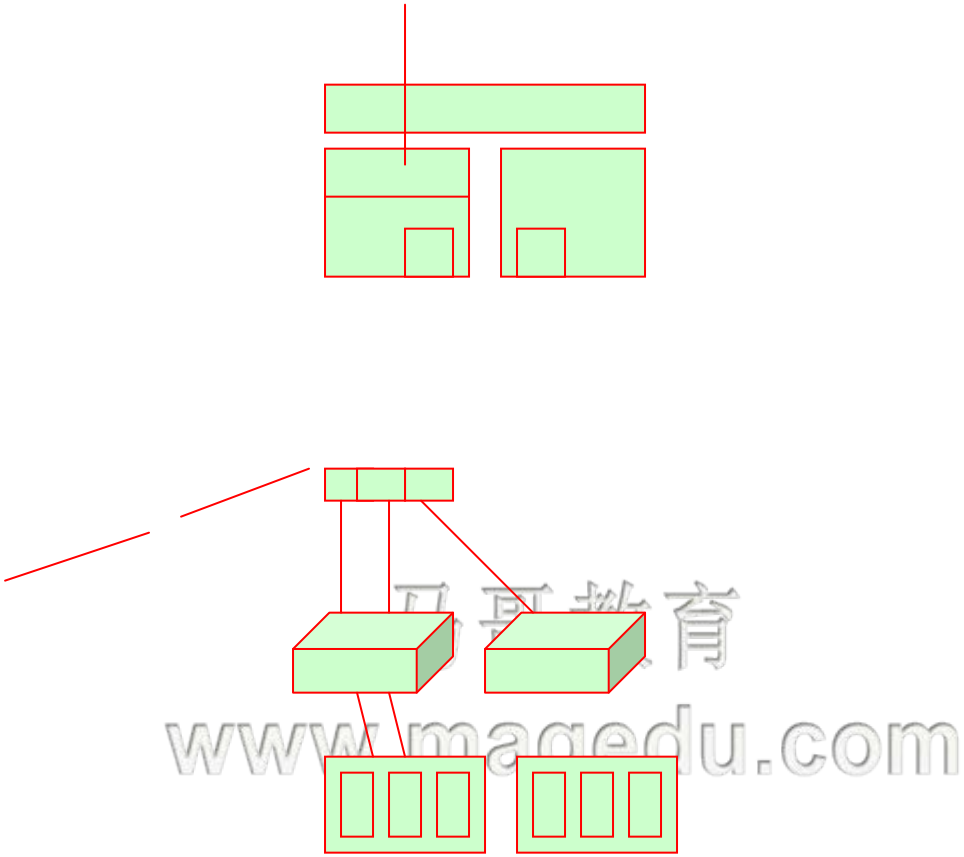
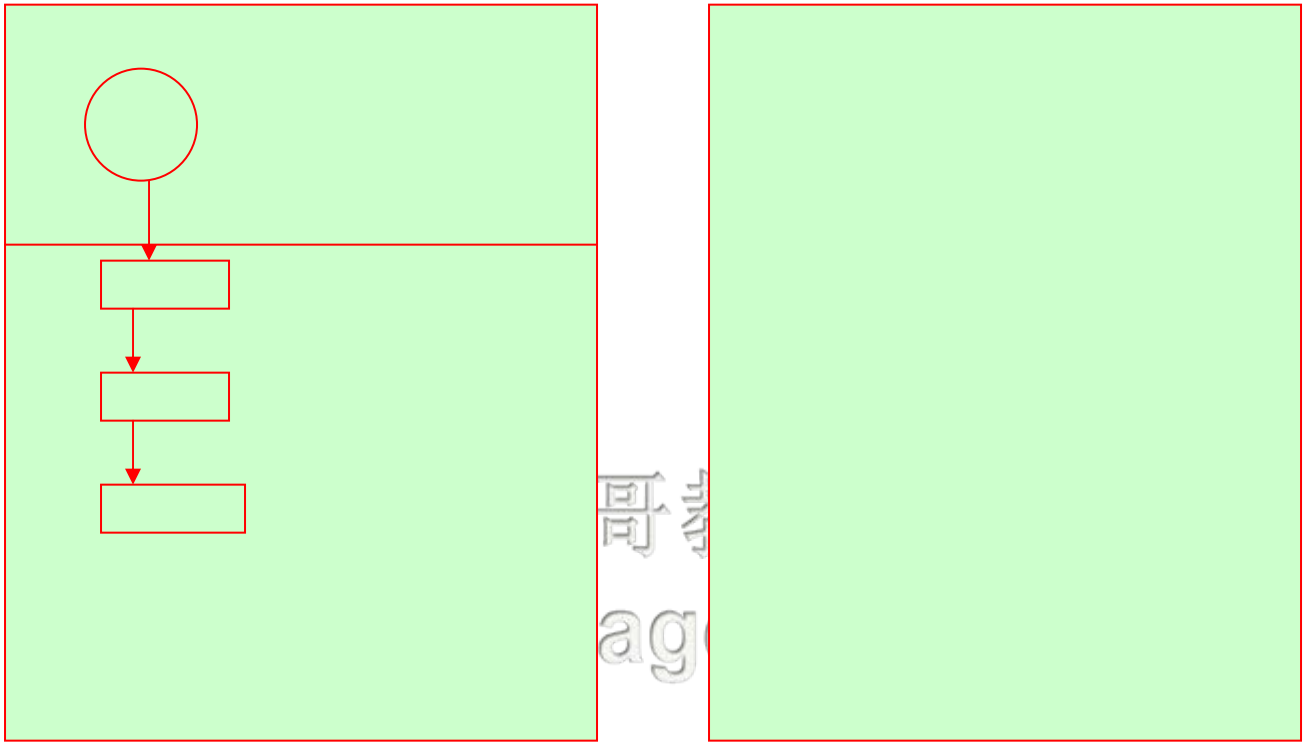
马哥教育

drbd基础应用

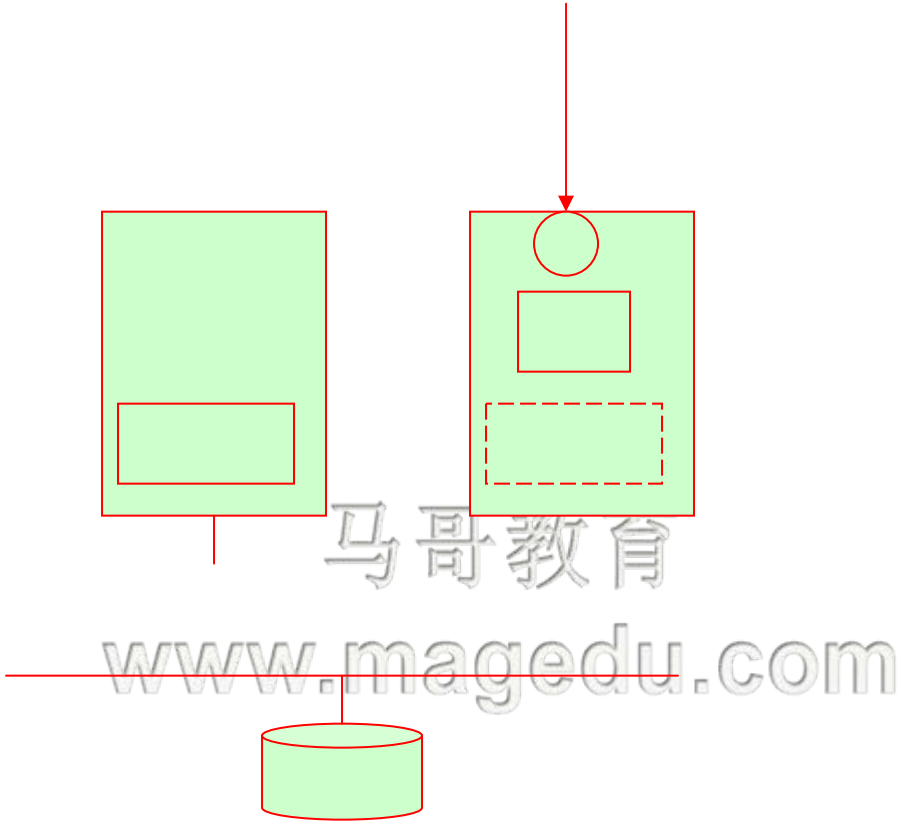
主讲：马永亮(马哥)
QQ群:169777636
客服QQ：2813150558，1661815153
http://www.magedu.com
http://mageedu.blog.51cto.com



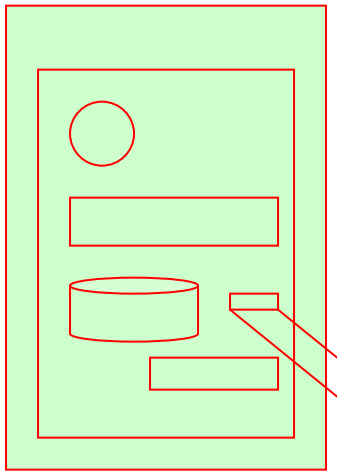




- HA Web
  - vip, httpd, webstore



DAS



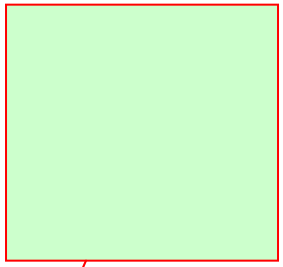
IDE(ATA)：133MB/s

SATA: 600MB/s

SCSI: 640MB/s

SAS: 6Gbps/

USB: 480MB/s



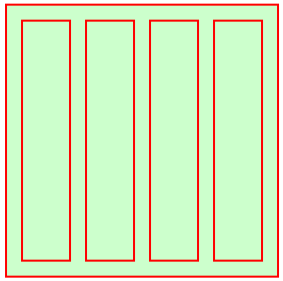
HBA: adapter

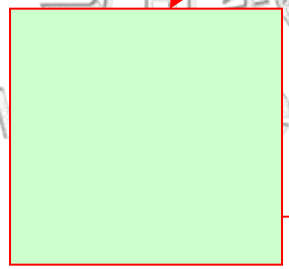
DAS: Direct Attached Storage

NAS: Network

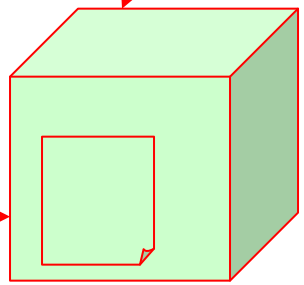
文件服务器: 文件系统级别

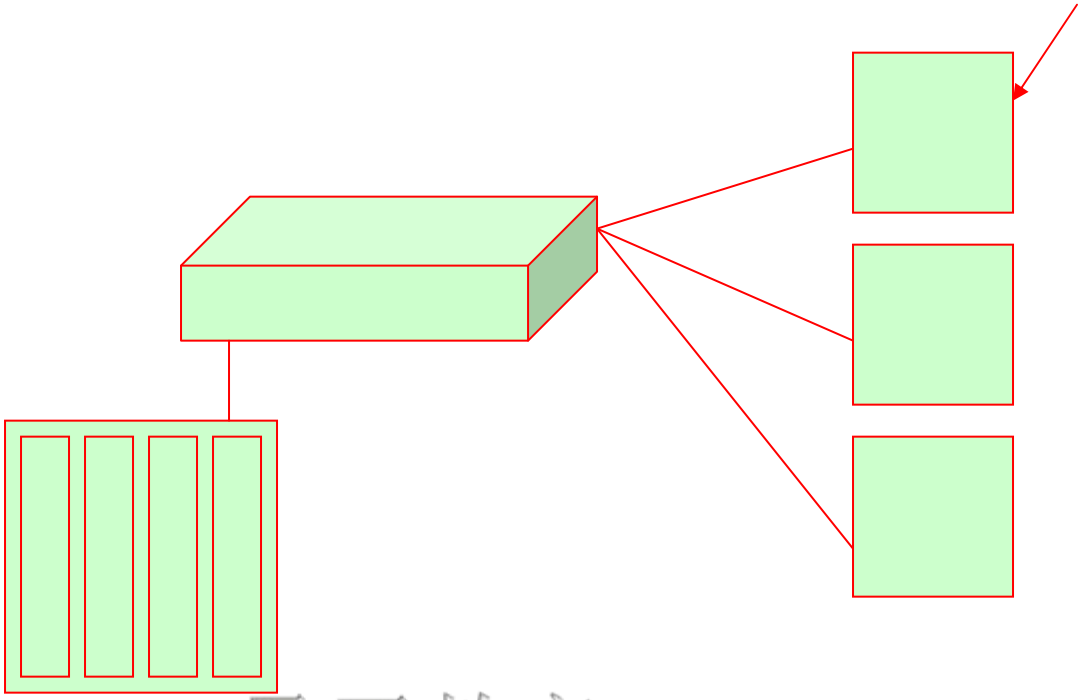
文件API





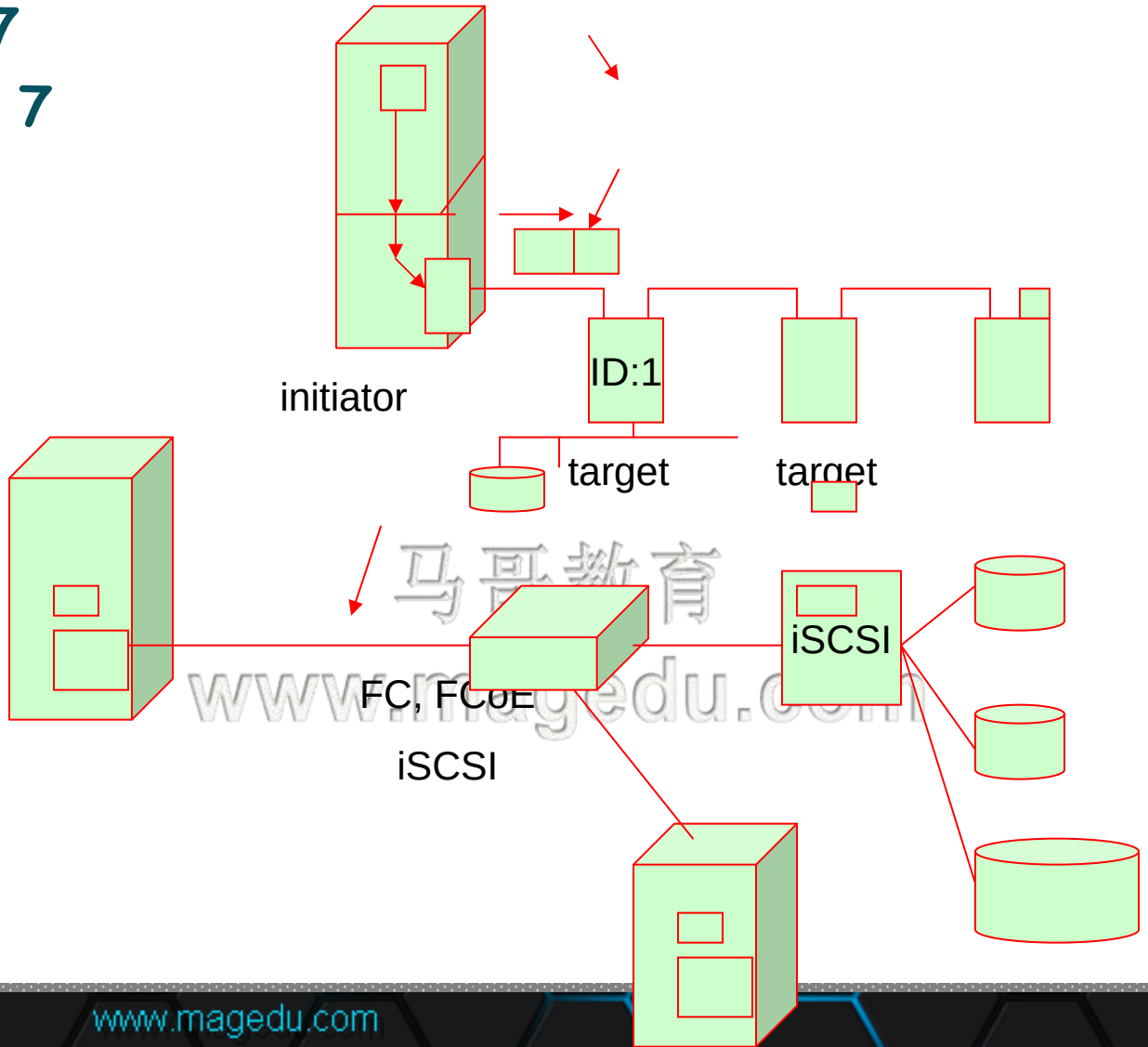
10000Mbps

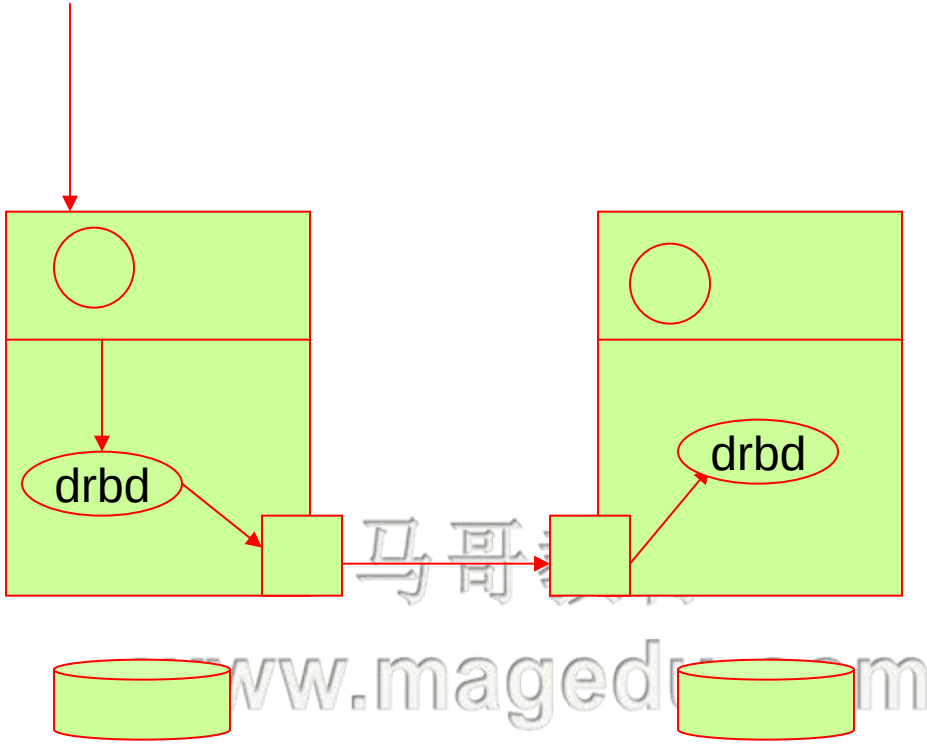




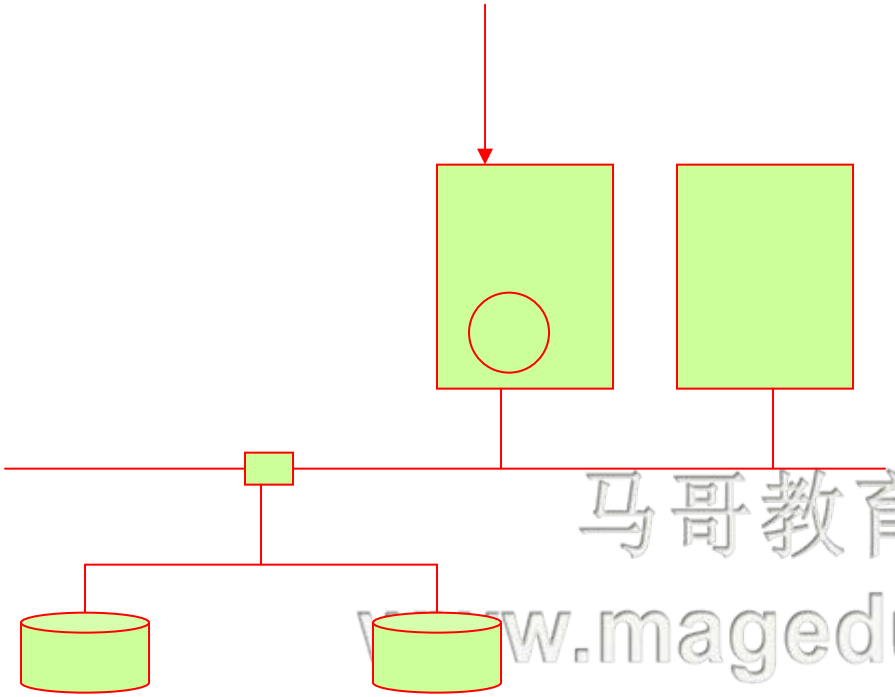
LUN: Logical Unit Number

- SCSI
  - 8, 7
  - 16, 7





- Server: rsync + inotify
- Client: rsync, sersync



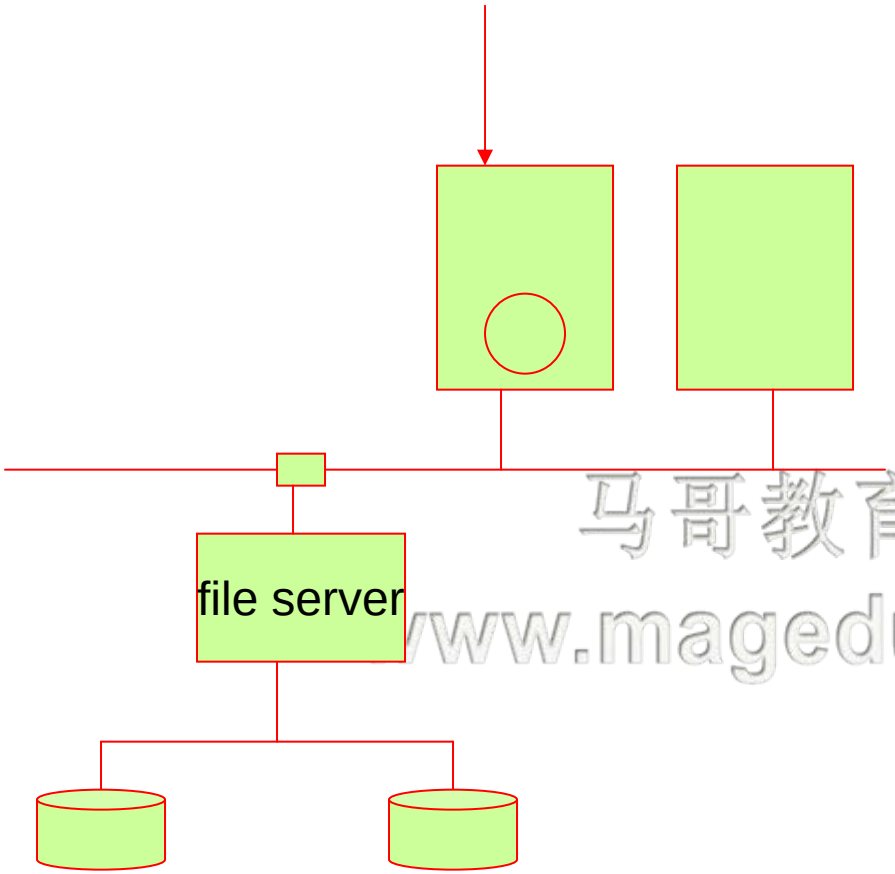
DAS

NAS

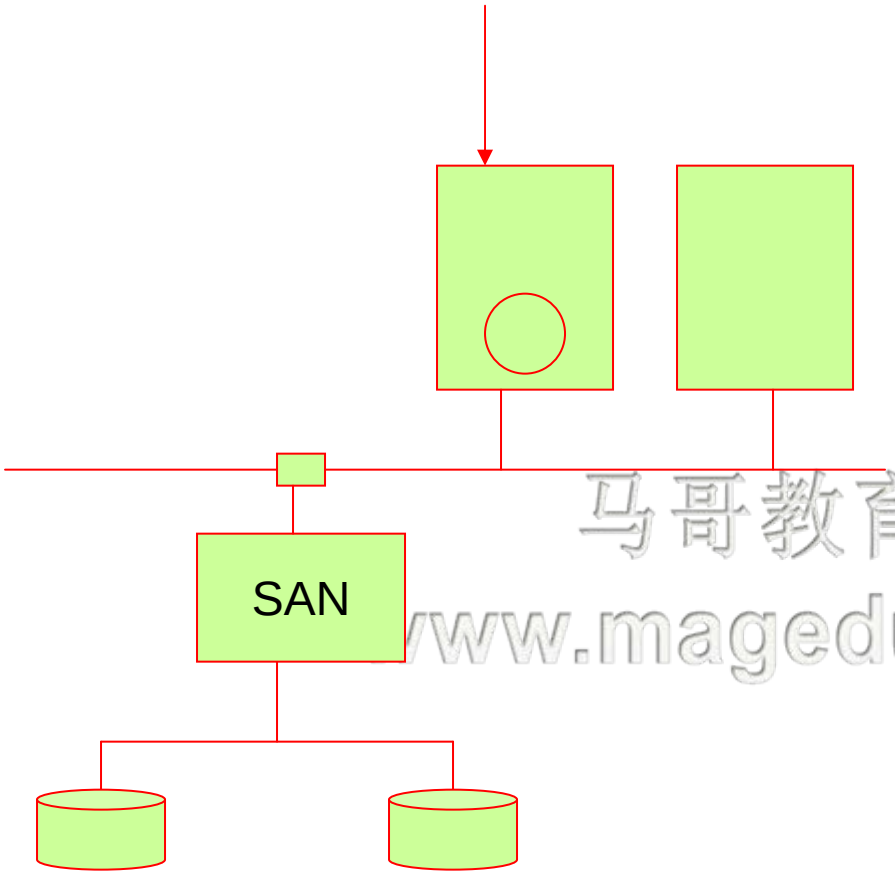
SAN

scsi: initiator, target

文件级别

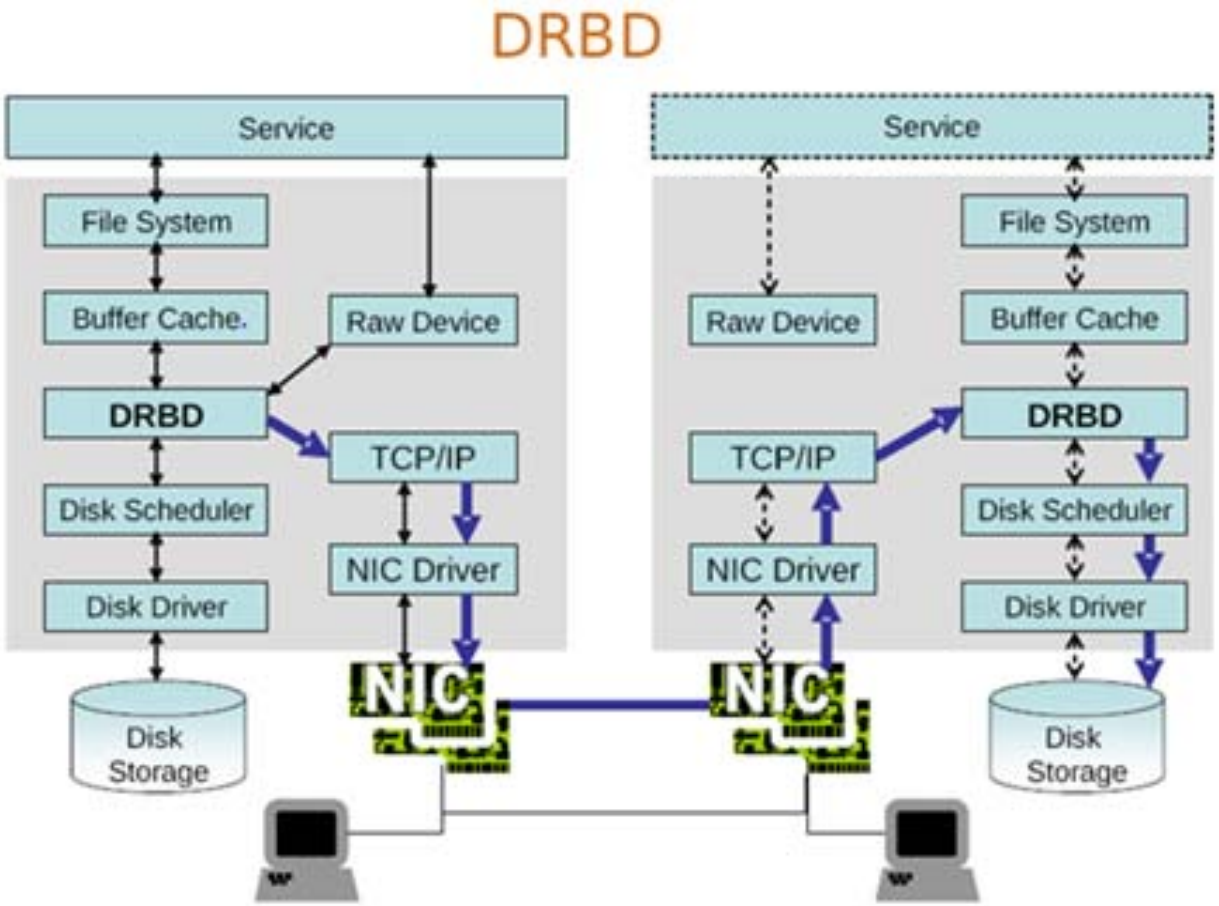


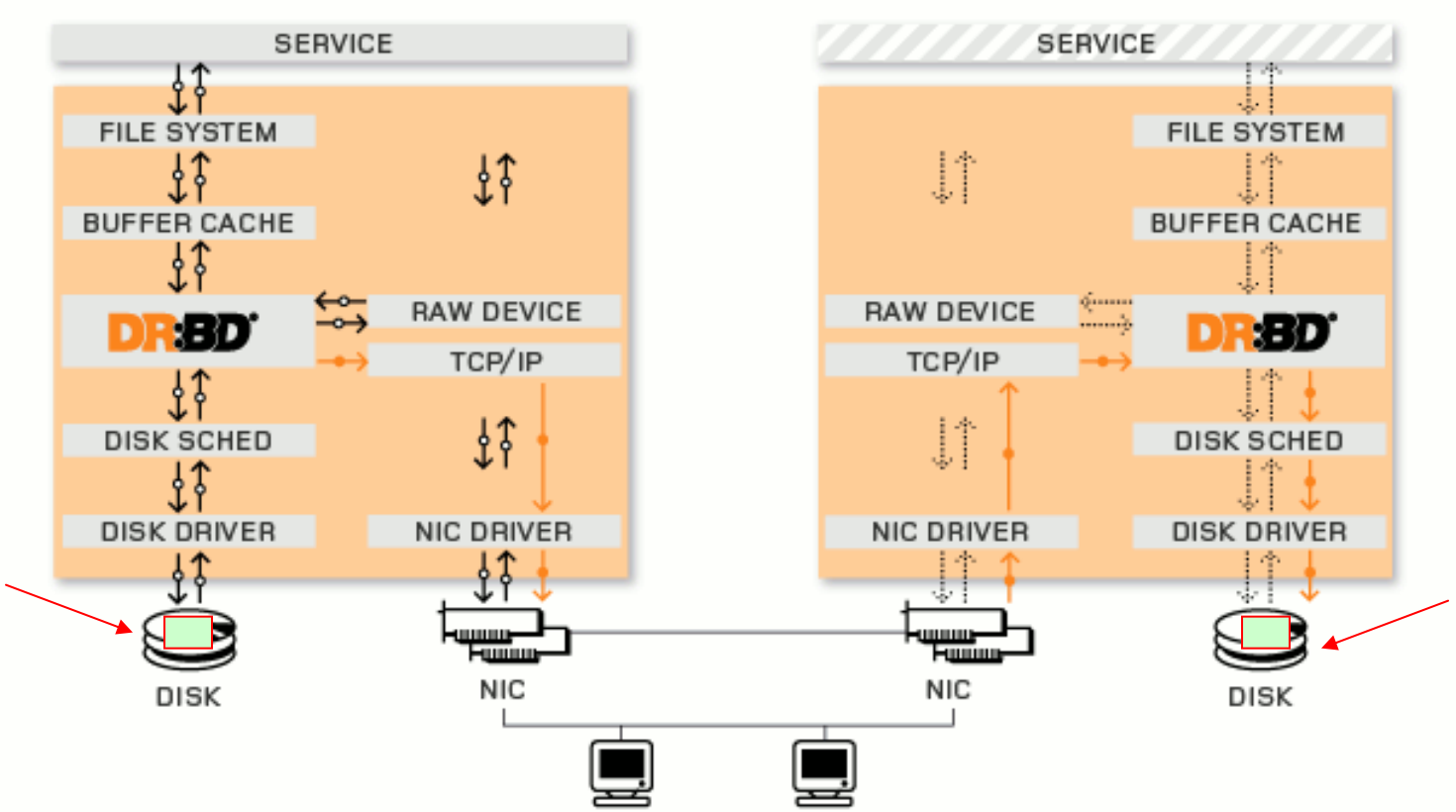
文件级别



- DRBD
  - Distributed Replicated Block Device
  - A software-based, shared-nothing, replicated storage solution mirroring the content of block devices (hard disks, partitions, logical volumes etc.) between servers
  - DRBD's core functionality is implemented by way of a Linux kernel module
- DRBD mirrors data
  - In real time
  - Transparently
  - Synchronously or asynchronously

DRBD's position within the Linux I/O stack



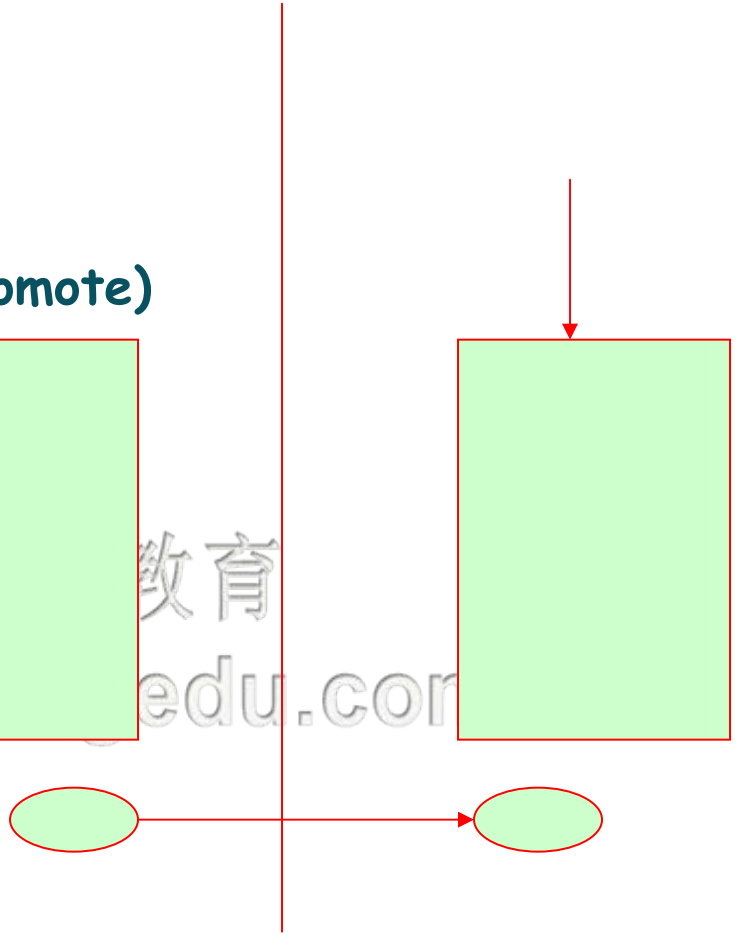

SERVICE
FILE SYSTEM
BUFFER CACHE
DRBD
RAW DEVICE
TCP/IP
DISK SCHED
DISK DRIVER
NIC DRIVER
DISK
NIC
SERVICE
FILE SYSTEM
BUFFER CACHE
RAW DEVICE
TCP/IP
DRBD
DISK SCHED
NIC DRIVER
DISK DRIVER
NIC
DISK

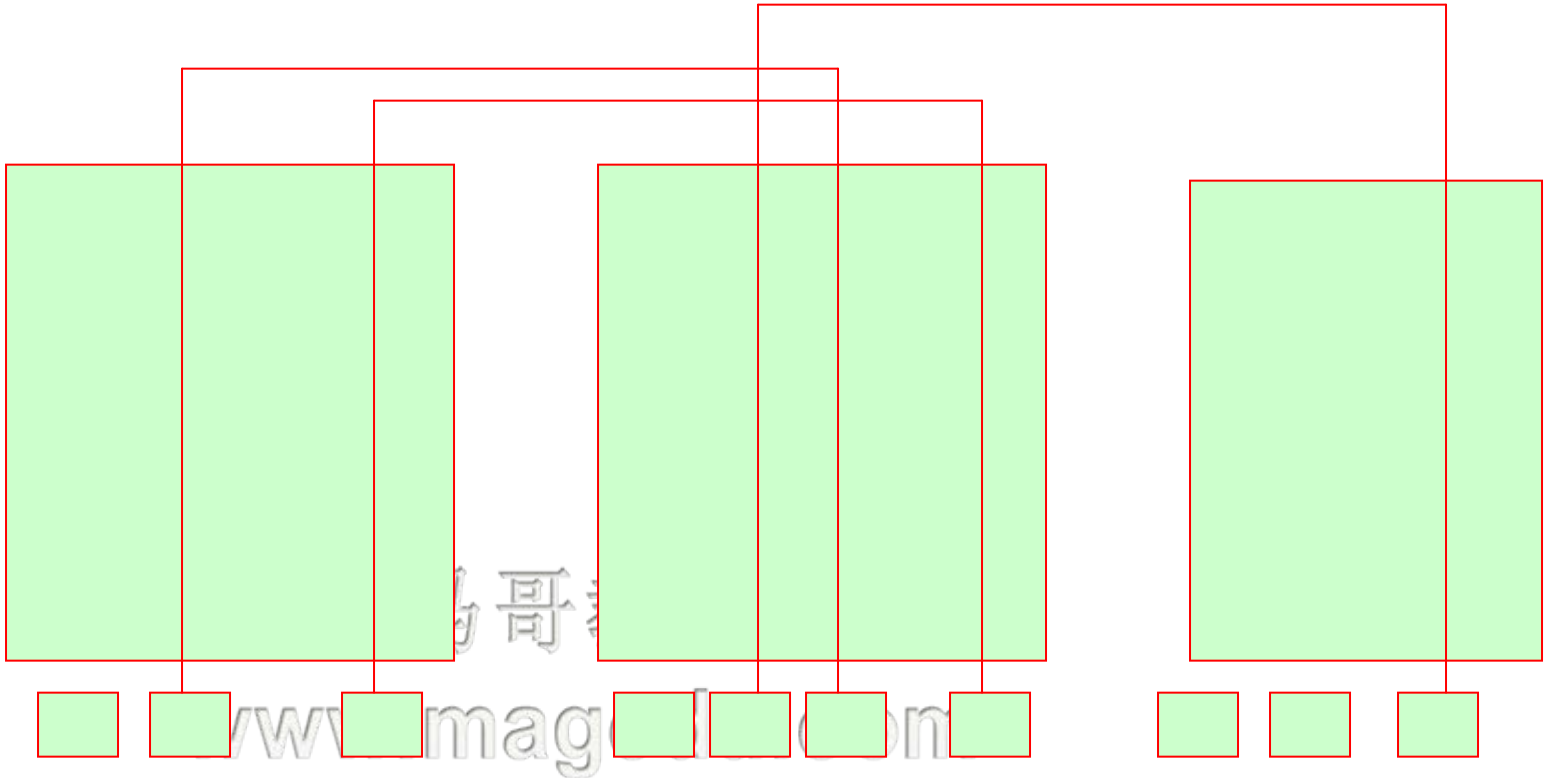
- master/master
  - cluster filesystem: GFS2, OCFS2
  - HA: ra
  - dlm
- master/slave

User space administration tools

- drbdadm
  - The high-level administration tool of the DRBD program suite
  - It obtains all DRBD configuration parameters from the configuration file /etc/drbd.conf(/etc/drbd.d/)
- drbdsetup
  - The program that allows users to configure the DRBD module that has been loaded into the running kernel
  - It is the low-level tool within the DRBD program suite
- drbdmeta
  - The program which allows users to create, dump, restore, and modify DRBD's meta data structures

- secondary/secondary
  - promote：提升
    - primary/secondary
  - demote
    - secondary/secondary
    - secondary/primary(promote)





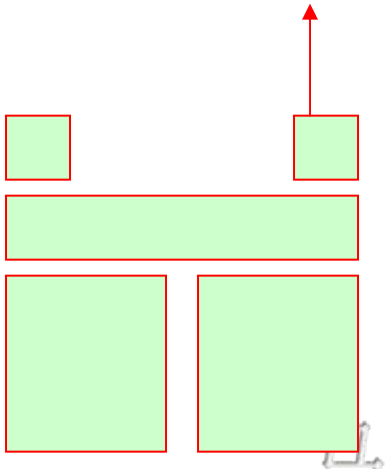
Resources

- the collective term that refers to all aspects of a particular replicated storage device
- Include
  - Resource name
    - This can be any arbitrary, US-ASCII name not containing whitespace by which the resource is referred to
  - DRBD device
    - This is the virtual block device managed by DRBD
    - It has a device major number of 147, and its minor numbers are numbered from 0 onwards, as is customary
    - The associated block device is always named /dev/drbd*n*, where *n* is the device minor number
  - Disk configuration
    - This entails the local copy of the data, and meta data for DRBD's internal use
  - Network configuration
    - This entails all aspects of DRBD's communication with the peer node

Resource roles

- In DRBD, every resource has a role, which may be Primary or Secondary
  - Primary
    - A DRBD device in the primary role can be used unrestrictedly for read and write operations
    - It may be used for creating and mounting file systems, raw or direct I/O to the block device, etc
  - Secondary
    - A DRBD device in the secondary role receives all updates from the peer node's device, but otherwise disallows access completely
    - It can not be used by applications, neither for read nor write access

❖ promote, demote



DRDB Modes

- Single-primary mode
  - Any resource is, at any given time, in the primary role on only one cluster member
  - Only one cluster node manipulates the data at any moment
  - This mode can be used with any conventional file system (ext3, ext4, XFS etc.)
  - Deploying DRBD in single-primary mode is the canonical approach for high availability (fail-over capable) clusters

DRDB Modes

- Dual-primary mode
  - Any resource is, at any given time, in the primary role on both cluster nodes
  - Since concurrent access to the data is thus possible, this mode requires the use of a shared cluster file system that utilizes a distributed lock manager
    - Examples include GFS2 and OCFS2
  - Deploying DRBD in dual-primary mode is the preferred approach for load-balancing clusters which require concurrent data access from two nodes
  - Disabled by default, and must be enabled explicitly in DRBD's configuration file
  - Available in DRBD 8.0 and later

Replication modes

- Protocol A
  - Asynchronous replication protocol
  - Local write operations on the primary node are considered completed as soon as the local disk write has occurred, and the replication packet has been placed in the local TCP send buffer
  - In the event of forced fail-over, data loss may occur
  - The data on the standby node is consistent after fail-over, however, the most recent updates performed prior to the crash could be lost

Replication modes

- Protocol B
  - Memory synchronous (semi-synchronous) replication protocol
  - Local write operations on the primary node are considered completed as soon as the local disk write has occurred, and the replication packet has reached the peer node
  - Normally, no writes are lost in case of forced fail-over
  - However, in the event of simultaneous power failure on both nodes and concurrent, irreversible destruction of the primary's data store, the most recent writes completed on the primary may be lost

Replication modes

- Protocol C
  - Synchronous replication protocol
  - Local write operations on the primary node are considered completed only after both the local and the remote disk write have been confirmed
  - Loss of a single node is guaranteed not to lead to any data loss
  - Data loss is, of course, inevitable even with this replication protocol if both nodes (or their storage subsystems) are irreversibly destroyed at the same time

Efficient synchronization

- Synchronization is necessary if the replication link has been interrupted for any reason
  - failure of the primary node
  - failure of the secondary node
  - or interruption of the replication link
- (Re-)synchronization is distinct from device replication

Split brain

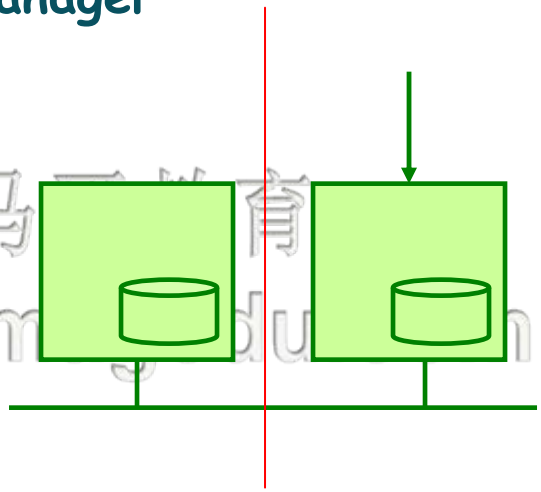
- Split brain is a situation where, due to temporary failure of all network links between cluster nodes, and possibly due to intervention by a cluster management software or human error, both nodes switched to the primary role while disconnected

Split brain notification and automatic recovery

- DRBD allows for automatic operator notification (by email or other means) when it detects split brain
- RBD has several resolution algorithms available for resolving the split brain automaticly
  - Discarding modifications made on the "younger" primary
  - Discarding modifications made on the "older" primary
  - Discarding modifications on the primary with fewer changes
  - Graceful recovery from split brain if one host has had no intermediate changes

- HA
  - Messaging Layer
  - CRM
    - heartbeat v2 + haresources/crm
    - heartbeat v3 + pacemaker
    - corosync + pacemaker
    - cman(openais) + rgmanager

- mysql + drbd:
  - primary/secondary



马哥教育

drbd + pacemaker

主讲：马永亮(马哥)
QQ群:169777636
客服QQ：2813150558，1661815153
http://www.magedu.com
http://mageedu.blog.51cto.com

Field	Description
clone-max	*How many copies of the resource to start. Defaults to the number of nodes in the cluster.*
clone-node-max	*How many copies of the resource can be started on a single node; default 1.*
notify	*When stopping or starting a copy of the clone, tell all the other copies beforehand and when the action was successful. Allowed values: false,* **true**
globally-unique	*Does each copy of the clone perform a different function? Allowed values: false,* **true**
ordered	*Should the copies be started in series (instead of in parallel). Allowed values: false,* **true**
interleave	*Changes the behavior of ordering constraints (between clones/masters) so that instances can start/stop as soon as their peer instance has (rather than waiting for every instance of the other clone has). Allowed values: false,* **true**

Field	Description
master-max	*How many copies of the resource can be promoted to master status. Defaults to 1.*
master-node-max	*How many copies of the resource can be promoted to master status on a single node. Defaults to 1.*

关于马哥教育

- 博客：http://mageedu.blog.51cto.com
- 主页：http://www.magedu.com
- QQ：1661815153，113228115
- QQ群：203585050，279599283

马哥教育

Thank You!