http://groups.google.com/group/sysadminstudy, 2010.3.17, 이종준

CHAPTER 03
무중단 인프라를 향한 새로운 연구

3.1 DNS서버의 다중화

DNS서버 다중화의 중요성

DNS서버의 이상을 검출하려면 Timeout 시간을 기다리는 것 외에는 방법 없음

- 좀처럼 발생하지 않지만, 일단 발생하면 원인이 판명되기까지 시간이 걸림
- /etc/resolv.conf에 여러 DNS서버를 지정하는 방법 → 가장 쉬움
- "질의가 Timeout된 경우, 다음 네임서버에 질의해본다" 라는 동작의 문제점
- DNS서버가 다운되면 Timeout을 (기본값 5초)대기한 후 다음서버로 질의
  → 1시간에 1000통의 메일을 전송해야하는 메일서버의 경우,
  최소한 3초에 1통씩 메일을 전송할 수 있어야한다
  → DNS서버 1대가 장애시, DNS질의 1회에 5초의 Timeout 발생
  → 1통을 메일을 송신하는데 10초가 소요되어 360통 정도만 처리 가능

Active/Backup 구성

- keepalived의 VRRP를 이용한 구성
- Active 서버가 VIP를 가진다.
- /etc/resolv.conf에 VIP만을 설정하고, 2대의 DNS서버중에 하나가 VIP를 가지도록 설정
- dig 명령을 통해 DNS서버의 상태를 확인하는 스크립트 필요

DNS서버의 부하분산

- Active/Active 구성
- 로드밸런서가 VIP를 가진다.
- 동일 서브넷 상에서는 DSR로 구성한다.
  → 각 DNS서버에서 VIP를 향하는 패킷을 처리할 수 있도록 loopback 장치에
  VIP를 할당하거나 iptables를 이용해 리다이렉트하는 등의 처리가 필요

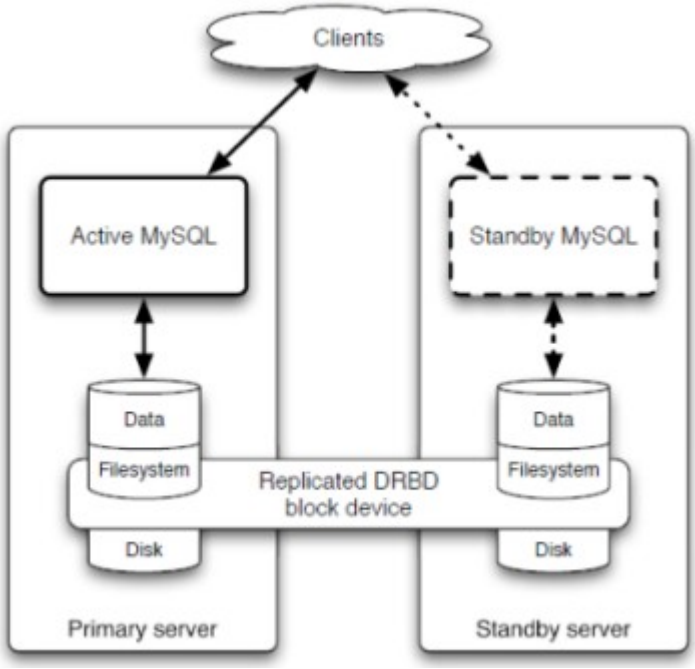
궁금한 사항

- Master/Slave 구성의 DNS서버 구축과는 다른관점의 이중화인가?
  → Master 를 이중화 하는게 더 좋은 것인지?

3.2 스토리지 서버의 다중화

스토리지 서버의 장애대책 → 스토리지 서버의 동기화 문제
- 기존 스크립트를 이용한 동기화는 신뢰성면에서 불안 요소 → 정합성 점검 문제

DRBD(Distributed Replicated Block Device)
- 네트워크 상에서 디스크 RAID1을 구현
- 미러링의 대상을 블록 디바이스로 지정
- 기본적으론 Active/Backup 을 지원하나, OCFS2, GFS에 한해 Active/Active 지원
- 커널 2.6.33 에 들어감, CentOS/RHEL의 경우 extra 저장소에서 drbd83 패키지로 지원



이미지 출처 : http://www.cs.helsinki.fi/u/jplindst/ps/af-HA-MySQL.pdf

Active/Backup 서버의 drbd.conf	cat /proc/drbd

global { usage-count yes; }
common { syncer { rate 10M; } }
resource drbd0 {
  protocol C;
  startup {
  degr-wfc-timeout 120;
  }
  net {
  cram-hmac-alg sha1;
  shared-secret "security";
  }
  on test1 {
  device /dev/drbd0;
  disk /dev/sdb1;
  address 192.168.0.11:7789;
  meta-disk internal;
  }
  on test2 {
  device /dev/drbd0;
  disk /dev/sdb1;
  address 192.168.0.12:7789;
  meta-disk internal;
  }
}
	• cs — connection state
	• st — node state (local/remote)
	• ld — local data consistency
	• ds — data consistency
	• ns — network send
	• nr — network receive
	• dw — disk write
	• dr — disk read
	• pe — pending (waiting for ack)
	• ua — unack'd (still need to send ack)
	• al — access log write count

참조 : http://www.drbd.org/users-guide/

DRBD의 성능
참조 : http://blog.trivadis.com/blogs/gregorysteulet/default.aspx

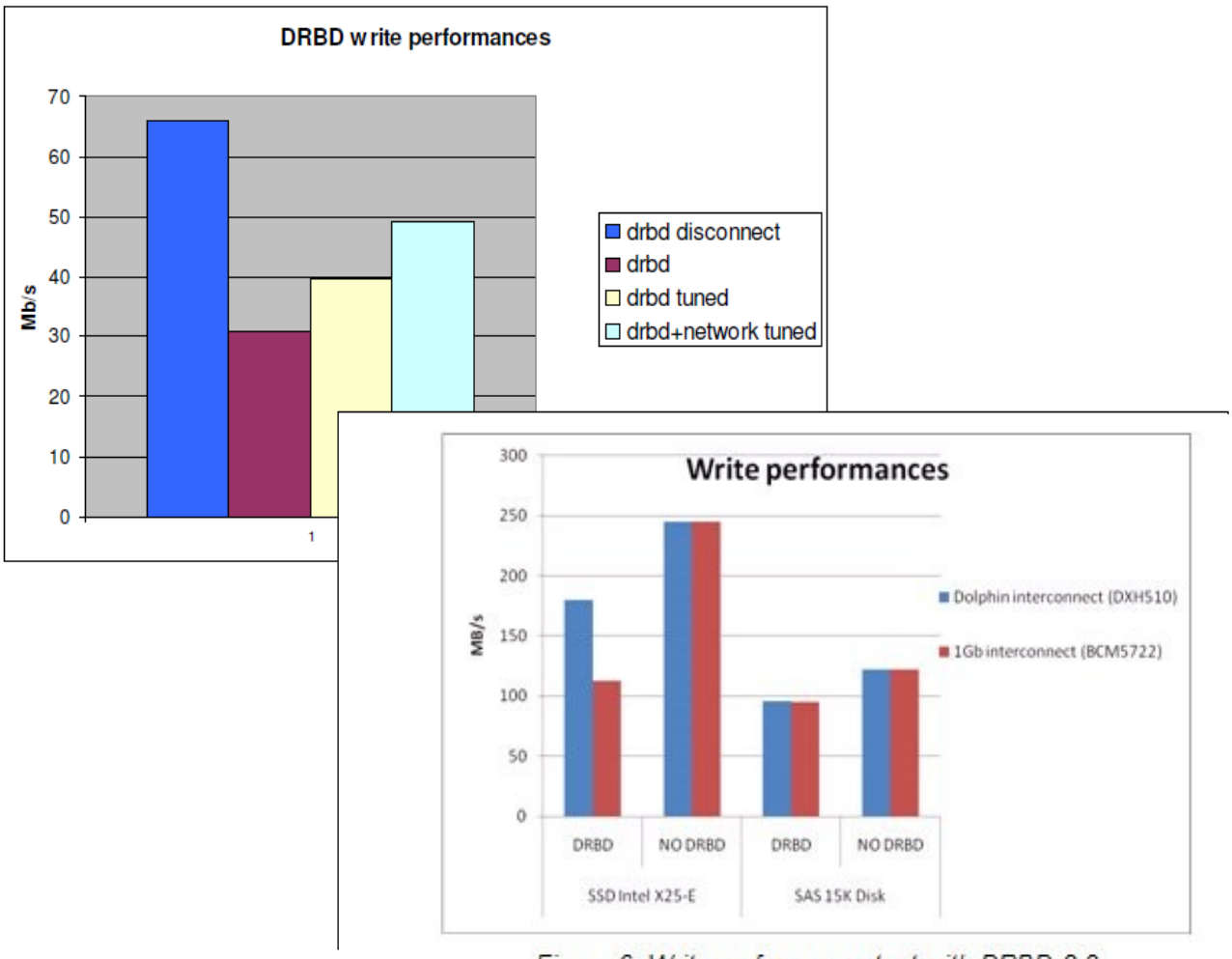


Figure 6: Write performance test with DRBD 8.3

Write Performance with iozone, 4GB file size

- xfs file system without drbd, single thread: 28,9 MB/s
- with drbd (connected): 17,4 MB/s → 60 %
- unconnected: 24,2 MB/s → 84 %
- 4 threads: 15,0 MB/s
- with drbd (connected), but protocol A: 21,4MB/s → 74 %
- unconnected: 24,2 MB/s → 84 %

출처 : https://www.racf.bnl.gov/Facility/hepix/talks/041020am/miers.pdf
참조 : http://j2k.naver.com/j2k_frame.php/korean/d.hatena.ne.jp/shakemid/20080210/1202665383

DRBD의 프로토콜

- A : 로컬 디스크에 쓰기가 끝나고 TCP 버퍼에 데이터를 송신한 시점에서 쓰기작업 완료(비동기)
- B : 로컬 디스크에 쓰기가 끝나고 원격 호스트로 데이터가 도달한 시점에서 쓰기작업 완료(중간)
- C : 원격 호스트의 디스크에도 쓰기가 완료된 시점에서 쓰기작업 완료(동기)

궁금한 사항 DRBD의 split brain 시나리오?

- The split brain is the fact of having two primary resources while the network between these resources failed.
- 복구 관련 : http://j2k.naver.com/j2k_frame.php/korean/d.hatena.ne.jp/hogem/searchdiary?word=*[drbd] 와 http://blog.lovediary.co.kr/entry/DRBD-split-brain-%EC%B2%98%EB%A6%AC%ED%95%98%EA%B8%B0

DRBD의 장애조치

keepalived 를 이용한 DRBD의 장애조치

- DRBD는 Active 서버에 문제가 발생하더라고 자동으로 Backup서버가 Active서버로 되지 않는다.
- Keepalived 기능에 Active서버에 장애가 발생했을때 자동으로 실행될 작업을 추가하여 장애조치/장애복구

Active 서버의 keepalived.conf	Backup 서버의 keepalived.conf

vrrp_instance DRBD {
state MASTER
interface eth0
garp_master_delay 5
virtual_router_id 200
priority 101
nopreempt
advert_int 1
authentication {
auth_type PASS
auth_pass Add-Your-Password-Here
}
virtual_ipaddress {
192.168.0.15
}
notify_master "/usr/local/sbin/drbd_master.sh"
notify_backup "/usr/local/sbin/drbd_backup.sh"
notify_fault "/usr/local/sbin/drbd_backup.sh"
}
	vrrp_instance DRBD {
	state BACKUP
	interface eth0
	garp_master_delay 5
	virtual_router_id 200
	priority 102
	nopreempt
	advert_int 1
	authentication {
	auth_type PASS
	auth_pass Add-Your-Password-Here
	}
	virtual_ipaddress {
	192.168.0.15
	}
	notify_master "/usr/local/sbin/drbd_master.sh"
	notify_backup "/usr/local/sbin/drbd_backup.sh"
	notify_fault "/usr/local/sbin/drbd_backup.sh"
	}

drbd_master.sh

```
#!/bin/sh
drbdadm primary all
mount /dev/drbd0 /mnt/drbd0
```

drbd_backup.sh

```
#!/bin/sh
umount /mnt/drbd0
drbdadm secondary all
```

!! notify_master(Active상태), notify_backup(Backup상태), notify_falut(link 장애)에 정의된 스크립트가 실행되는지 반드시 확인하자.
→ keepalived 데몬이 종료되었을때 대응 필요(daemontools, mon 등 이용)

NFS서버를 장애극복할 때 주의점

- NFS서버의 접속정보를 저장하는 /var/lib/nfs 를 동기화(DRBD 마운트영역에 링크)
  → 배포판에 따라서는 exportfs명령으로 접속정보를 초기화 하는 경우도 있음
- NFS서버를 다중화 하기 위해 만들어진 고유 기능인 nfsd 를 이용
  → mount -t nfsd nfsd /proc/fs/nfsd 를 한 상태로 실행된 nfs서버는 /var/lib/nfs 디렉토리를 이용하지 않음

백업의 필요성
DRBD로 미러링하고 있더라도 누군가의 실수로 삭제한 파일을 복구 할 수는 없다.
→ 최악의 사태에 대비해서 항상 백업을 한다.

DRBD에 대한 구글링과 네이버 일본웹 검색 결과

DRBD 설치 방법을 상세히 알고 싶다면 다음 링크를 참고한다.
http://blog.lovediary.co.kr/tag/DRBD
https://wiki.ubuntu.com/Testing/Cases/UbuntuServer-drbd
http://translate.google.co.kr/translate?hl=ko&sl=th&tl=en&u=http%3A%2F%2Fspalinux.com%2F2010%2F02%2Finstall_and_configure_drbd

DRBD의 구축 실적

- 1999년 부터 개발이 계속되고 있음, 개발사인 LINBIT사에 의하면 약 20만 세트 이상임
- 일본의 경우 LINBIT 파트너사가 있음(기술지원, 교육 등)
  (http://j2k.naver.com/j2k_frame.php/korean/www.3ware.co.jp/DRBD.html)
- 64Bit 구조에서 최대 16TB의 디스크 영역을 동기화 가능
- 검증된 어플리케이션 : LAMP stack, MySQL, PostgreSQL, Oracle, DB2, NFS, iSCSI, Samba, Vmware, Xen, Kvm, 클러스터매니져(Heartbeat, Pacemaker, RHCS, Lifekeeper)

참조 : http://www.linbit.com/en/products-services/drbd/drbd-for-high-availability/

왜 DRBD가 필요한가?
참조 : (http://j2k.naver.com/j2k_frame.php/korean/www.conversion.co.jp/drbd_faq.html)
공유 스토리지가 RAID구성이면 DRBD가 필요없는 것이 아닌가?
→ SPOF 발생할 가능성이 있기 때문에 다중화하는 것을 권장
→ RAID구성의 경우 HDD 고장에는 대응하지만, RAID콘트롤러나, 전원등의 장애는 대응 못함
(물론 콘트롤러, 전원 이중화가 되는 스토리지가 있으나 고가임)

스토리지를 정기 백업하고 있으면 DRBD는 필요 없는 것이 아닌가?
→ 정기 백업은 다음 번의 백업까지 데이터의 갱신이 반영되지 않는다.

DRBD를 리눅스 이외의 OS에서 사용이 가능한가?
→ 리눅스 커널 모듈이므로 사용은 불가능하다, 단, iSCSI 타켓 머신에 DRBD를 도입해 내부에서 DRBD를 설정하여 이용가능
참조 : http://www.howtoforge.com/installing-and-configuring-openfiler-with-drbd-and-heartbeat 와 https://help.ubuntu.com/community/HighlyAvailableiSCSITarget

상대적으로 느린 Write 속도를 어떻게 개선할것인가?
참조 : http://lists.linbit.com/pipermail/drbd-user/2008-July/009893.html

Troughput on LV: 239.9 MiByte/sec DRBD standalone: 239.8 MiByte/sec DRBD connected: 163.8 MiByte/sec (protocol C, over 2x 1GbE bonding balance-rr, e1000 ethernet card)

그외 Dolphin Express (http://www.mysqlconf.com/mysql2009/public/schedule/detail/6760) 와 Infiniband, 10 GbE Network

세미나 동영상 자료(이메일 주소만 입력하면 wmv 포멧 다운로드 가능)
http://www.linbit.com/en/training/on-demand-webinars/

- DRBD Performance Tuning
- High-Performance Replication with DRBD for Dolphin Express
- DRBD 8.3

What is DRBD?
Distributed: spans 2 cluster nodes
Replicated: all write I/O synchronously replicated to second node
Block **D**evice: abstraction layer for block-backed I/O in the Linux kernel

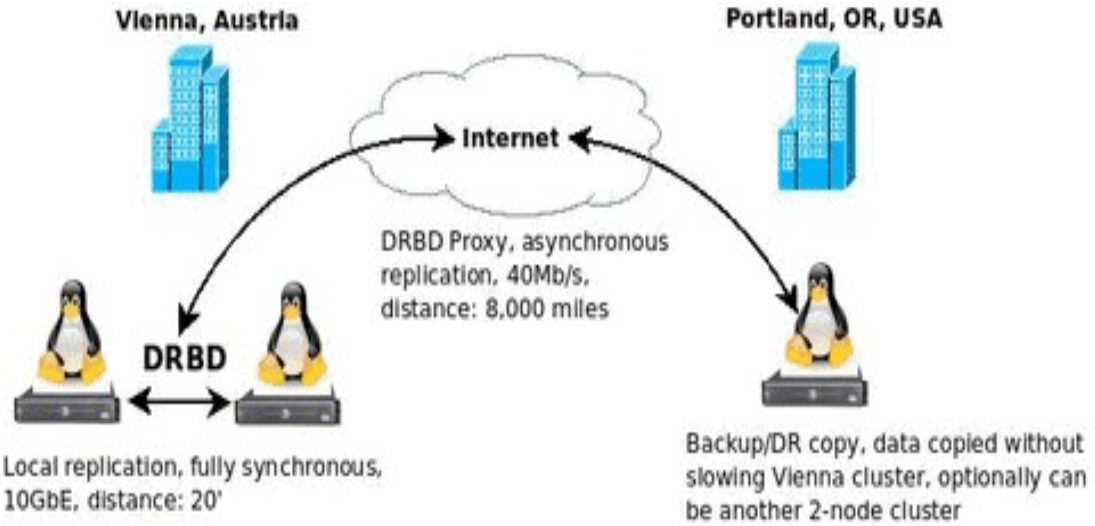
MySQL과 DRBD : http://mysql.com/drbd
http://downloads.mysql.com/docs/mysql-ha-drbd-en.a4.pdf

MySQL의 고가용성 증대
이미지 출처 : http://programm.froscon.org/2009/attachments/109_MySQL-HA-FrOSCon-2009-08-22.pdf

	Requirements	MySQL Replication	MySQL Replication & Heartbeat	MySQL, Heartbeat & DRBD	MySQL Cluster
Availability	Automated IP Failover	No	Yes	Yes	No
	Automated DB Failover	No	No	Yes	Yes
	Typical Failover time	Varies	Varies	< 30s	< 3s
	Auto resync of data	No	No	Yes	Yes
	Geographic redundancy	Yes	Yes	MySQL Replication	MySQL Replication
Scalability	Built-in load balancing	MySQL Replication	MySQL Replication	MySQL Replication	Yes
	Read-intensive	Yes	Yes	MySQL Replication	Yes
	Write-intensive	No	No	Possible	Yes
	#Nodes/Cluster	Master/Slave(s)	Master/Slave(s)	Active/Passive	255

참조 : MTV Japan의 사례 : http://www.mysqlconf.com/mysql2008/public/schedule/detail/3287

WAN구간에서의 DRBD
Closed로 제공되는 DRBD Proxy를 이용하면 WAN구간에서도 가능하다. 재해복구 환경을 저렴하게 구축이 가능해질것으로 보인다.
이미지 출처 : http://www.linbit.com/en/products-services/drbd-proxy/why-drbd-proxy/



참조 : https://www.os3.nl/_media/2008-2009/students/attilla_de_groot/virt_migration.pdf