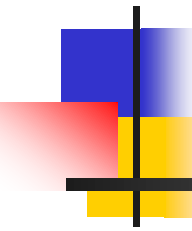
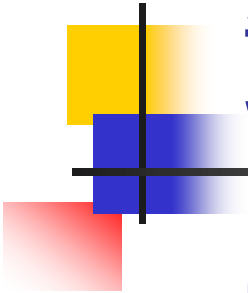


Pacemakerによる高可用化NFS

筑波大学情報学群情報科学類
三戸 健一 (@mittyorz)
mitty [at] coins.tsukuba.ac.jp

高可用NFSとは

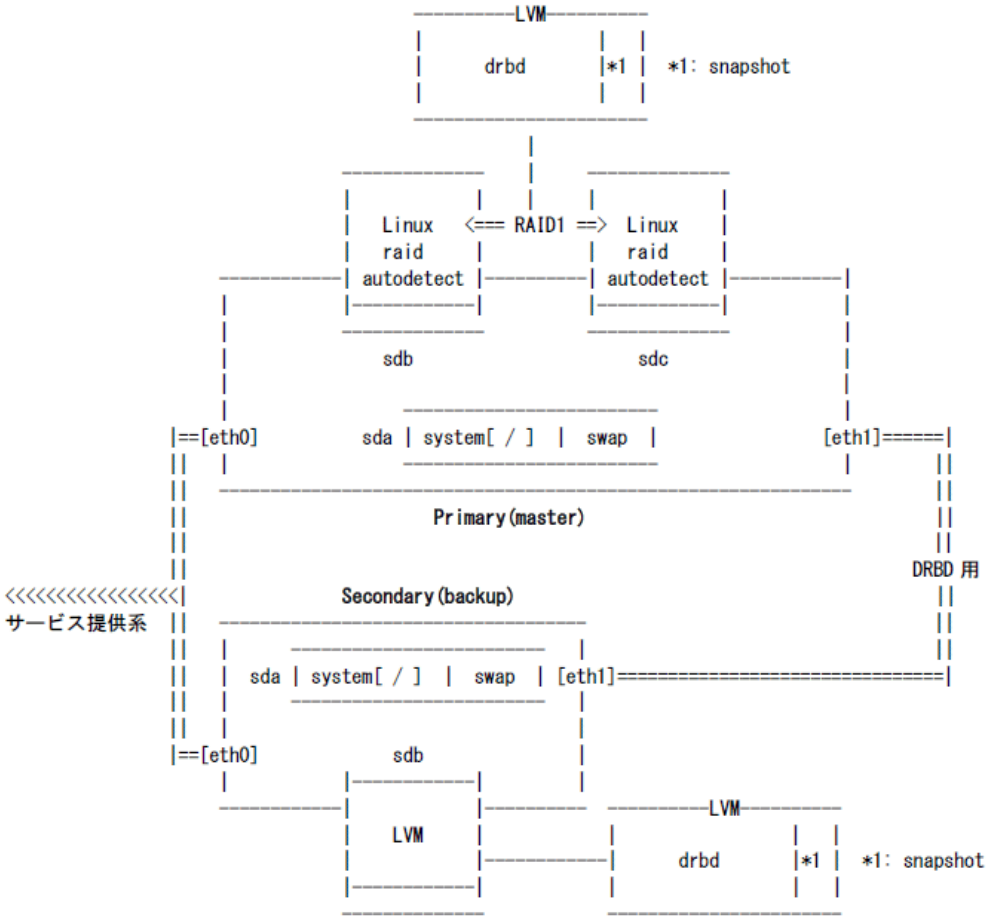
- NFSとは、TCP/IPネットワークを介してリモートのストレージを利用するクライアント・サーバモデルの分散ファイルシステム
- クライアント・サーバモデルなので、多数のクライアントに対し、本来はサーバは一つ
- NFSサーバを冗長化し、高可用化する
  - 本件では本番系・バックアップ系の2ノード構成
  - Pacemaker, DRBD, LVMの使用



動機

- 情報科学類生の有志で運用している、open-coinsのシステムの耐障害性を上げたい
  - 主にメーリングリストなどを提供中
  - http://www.open.coins.tsukuba.ac.jp/
- どの程度のシステムが構築出来るか、実地で試したい
  - ストレージをNFSで集中させ、ストレージ自体の耐障害性を上げる

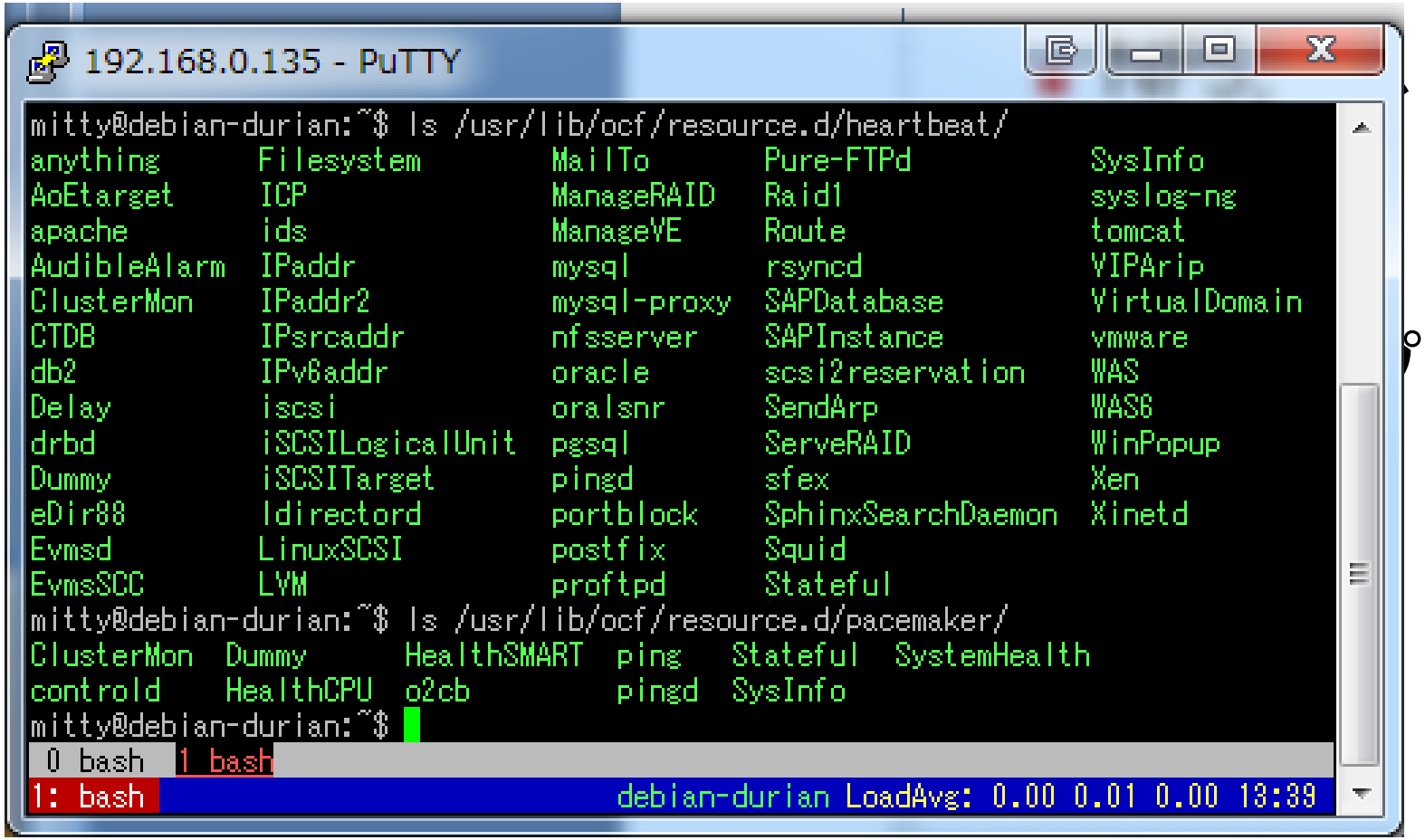
ストレージ構成(RAID+LVM+DRBD)



ストレージ構成(詳細)

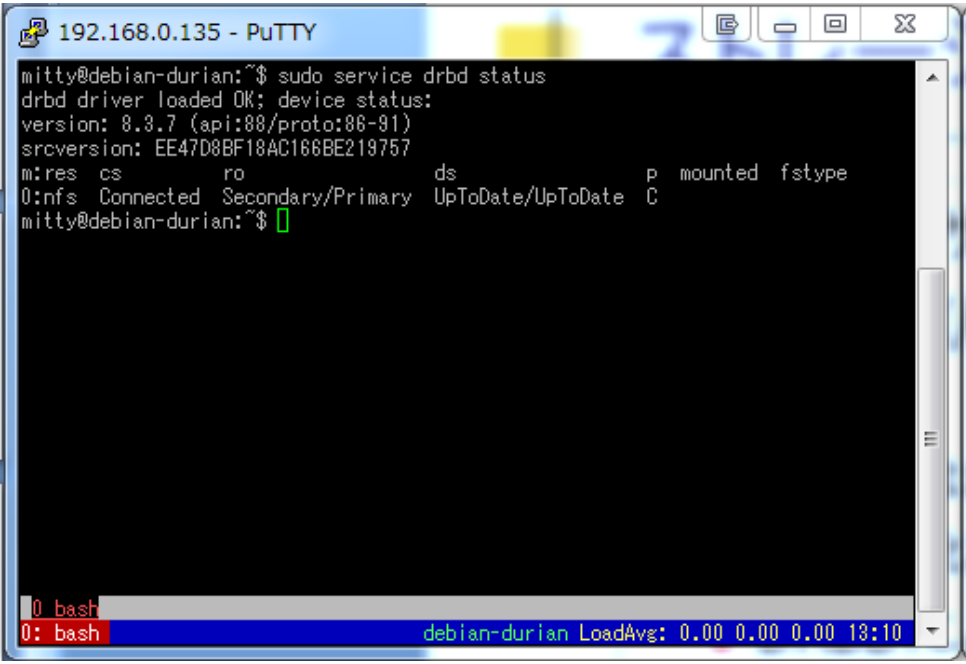
- マスターサーバ
  - RAID1 (/dev/sdb + /dev/sdc)
  - LVM
    - これはオンラインスナップショットを取るため
    - 冗長化とは別に、データの世代バックアップを取る
  - DRBDによるマスターサーバ・スレーブサーバ間のミラーリング
    - リアルタイムミラーのため、誤操作(rm –rf hoge)等によるデータ喪失へは、世代バックアップなどによる対応が別途必要

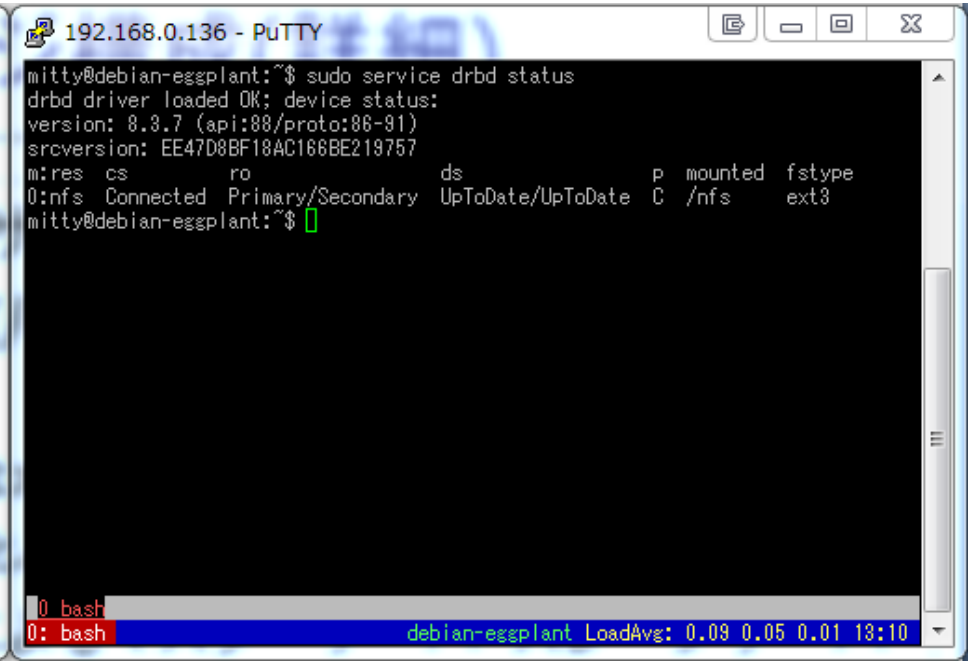
Pacemakerによる冗長化



Pacemakerによる制御

- Pingによるネットワーク監視
  - サービス系NICが外部ネットワークへ到達出来るか





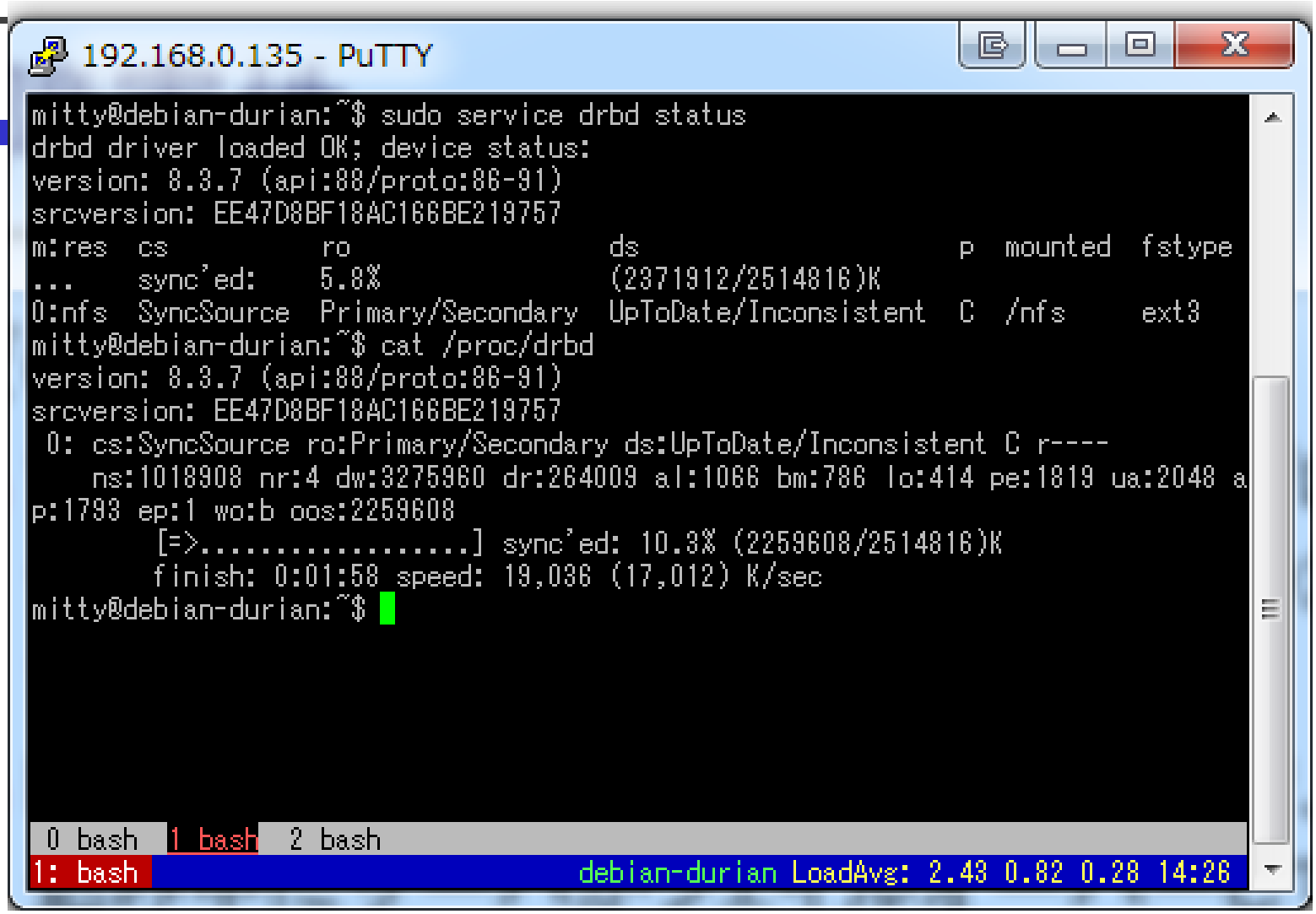
DRBDとpacemakerの構築

- 必要なプログラムはパッケージから導入出来る
  - Ubuntu 10.04 / Debian 6.0では確認
- DRBDの設定は比較的簡単
  - ミラーリング対象とするブロックデバイス
  - 提供するブロックデバイス名
  - 対向ノードのIPアドレス
  - 不整合が起きたときの挙動　　etc...

192.168.0.135 - PuTTY

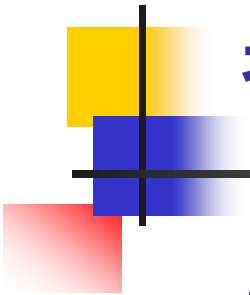
```
mitty@debian-durian:~$ sudo crm configure show
node debian-durian
node debian-eggplant
primitive nfs_common lsb:nfs-common \
        op monitor interval="5s"
primitive nfs_drbd ocf:linbit:drbd \
        params drbd_resource="nfs" \
        op start interval="0s" timeout="240s" on-fail="restart" \
        op monitor interval="5s" role="Master" timeout="60s" on-fail="restart" \
        op monitor interval="10s" role="Slave" timeout="60s" on-fail="restart" \
        op stop interval="0s" timeout="100s" on-fail="block"
primitive nfs_fs ocf:heartbeat:Filesystem \
        params device="/dev/drbd/by-res/nfs" directory="/nfs" fstype="ext3" \
        op start interval="0s" timeout="120s" on-fail="restart" \
        op monitor interval="10s" timeout="60s" on-fail="restart" \
        op stop interval="0s" timeout="60s" on-fail="block"
primitive nfs_ip ocf:heartbeat:IPaddr2 \
        params ip="192.168.0.130" nic="eth0" cidr_netmask="24" \
        op start interval="0s" timeout="90s" on-fail="restart" \
        op monitor interval="10s" timeout="60s" on-fail="restart" \
        op stop interval="0s" timeout="100s" on-fail="block"
primitive nfs_ping ocf:pacemaker:pingd \
        params name="nfs_ping_to_keepalive" host_list="192.168.0.254 192.168.0.250" multiplier="100" dampen="0" \
        meta migration-threshold="10" \
        op start interval="0s" timeout="90s" on-fail="restart" \
        op monitor interval="10s" timeout="60s" on-fail="restart" \
        op stop interval="0s" timeout="100s" on-fail="block"
primitive nfs_server lsb:nfs-kernel-server \
        op monitor interval="5s"
group group_nfs nfs_ip nfs_fs nfs_common nfs_server
ms ms_nfs_drbd nfs_drbd \
        meta master-max="1" master-node-max="1" clone-max="2" clone-node-max="1" notify="true"
clone clone_nfs_ping nfs_ping \
        meta clone-max="2" clone-node-max="1"
location loc_group_nfs group_nfs \
        rule $id="loc_group_nfs-rule" 200: #uname eq debian-eggplant \
        rule $id="loc_group_nfs-rule-0" 100: #uname eq debian-durian \
        rule $id="loc_group_nfs-rule-1" -inf: defined nfs_ping_to_keepalive and nfs_ping_to_keepalive lt 100
location loc_ms_nfs_drbd ms_nfs_drbd \
        rule $id="loc_ms_nfs_drbd-rule" 200: #uname eq debian-eggplant \
        rule $id="loc_ms_nfs_drbd-rule-0" 100: #uname eq debian-durian \
        rule $id="loc_ms_nfs_drbd-rule-1" $role="master" -inf: defined nfs_ping_to_keepalive and nfs_ping_to_keepalive lt 100
colocation col_nfs_on_drbd inf: group_nfs ms_nfs_drbd:Master
colocation col_nfs_on_pingd inf: group_nfs clone_nfs_ping
order order_drbd_before_nfs inf: ms_nfs_drbd:promote group_nfs:start
property $id="cib-bootstrap-options" \
        dc-version="1.0.9-74392a28b7f31d7ddc86689598bd23114f58978b" \
        cluster-infrastructure="openais" \
        expected-quorum-votes="2" \
        stonith-enabled="false" \
        no-quorum-policy="ignore" \
        startup-fencing="false"
rsc_defaults $id="rsc-options" \
        resource-stickiness="INFINITY" \
        migration-threshold="1"
mitty@debian-durian:~$
 0 bash  1 bash
0: bash                                                    debian-durian LoadAvg: 0.04 0.01 0.00 13:51
```

耐障害性


192.168.0.135 - PuTTY
mitty@debian-durian:~$ sudo service drbd status
drbd driver loaded OK; device status:
version: 8.3.7 (api:88/proto:86-91)
srcversion: EE47D8BF18AC166BE219757
m:res cs ro ds p mounted fstype
... sync'ed: 5.8% (2371912/2514816)K
0:nfs SyncSource Primary/Secondary UpToDate/Inconsistent C /nfs ext3
mitty@debian-durian:~$ cat /proc/drbd
version: 8.3.7 (api:88/proto:86-91)
srcversion: EE47D8BF18AC166BE219757
0: cs:SyncSource ro:Primary/Secondary ds:UpToDate/Inconsistent C r----
ns:1018908 nr:4 dw:3275960 dr:264009 al:1066 bm:786 lo:414 pe:1819 ua:2048 a
p:1793 ep:1 wo:b oos:2259608
[=>.....] sync'ed: 10.3% (2259608/2514816)K
finish: 0:01:58 speed: 19,036 (17,012) K/sec
mitty@debian-durian:~$
0 bash 1 bash 2 bash
1: bash debian-durian LoadAvg: 2.43 0.82 0.28 14:26

考えられるその他の用途

- Apache等と組み合わせて、webサービスの冗長化
- iSCSIによるストレージエリア
  - DRBDはmaster/slaveではなくmaster/master も可能なので、上に分散ファイルシステムを用いることも出来る
- KVMやXenなどと組み合わせて仮想化インフラとすることも出来る模様



参考資料

- 参考にしたサイトのまとめ
  - http://lab.mitty.jp/trac/lab/wiki/TipAndDoc/HA
- LT資料の置き場
  - http://lab.mitty.jp/svn/lab/trunk/TipAndDoc/HA/