LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

DRBDとPacemakerによるLinux-HA環境

株式会社サードウェア　岩崎のぼる

HighAvailability

LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

Profile

名前	岩崎のぼる／橘べるちぇ
所属	株式会社サードウェア
Twitter	http://twitter.com/bellche
活動	（会社） ・サーバ構築とか簡単なスクリプトとか（会社） （個人） ・書籍・雑誌の執筆 （日経Linux、SoftwareDesign、Perl中毒、その他） ・Linux-HA Japan Project（半分仕事） ・日本Unboundユーザ会 主にLinux関連の業界を右往左往。
その他	・猫好き ・最近Pythonを勉強中 ・iPhone/iPadを持ってもMacが無くて寂しい ・Androidの開発環境があっても端末なくて寂 ・自宅に自分の部屋が無くて寂しい

SoftwareDesign「DRBDで始める今どきクラスタリング」

↓

@IT「DRBD＋iSCSI夢の共演（前/後編）」→

HighAvailability

アウトライン

- DRBDの概要
- Linux-HAとDRBDを組み合わせ
- DRBDで起こりがちなトラブル
- 事例紹介
- Pacemaker概要
- HeartbeatからPacemakerへ
- Linux-HA Japan Project（コミュニティ）紹介

LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

HighAvailability

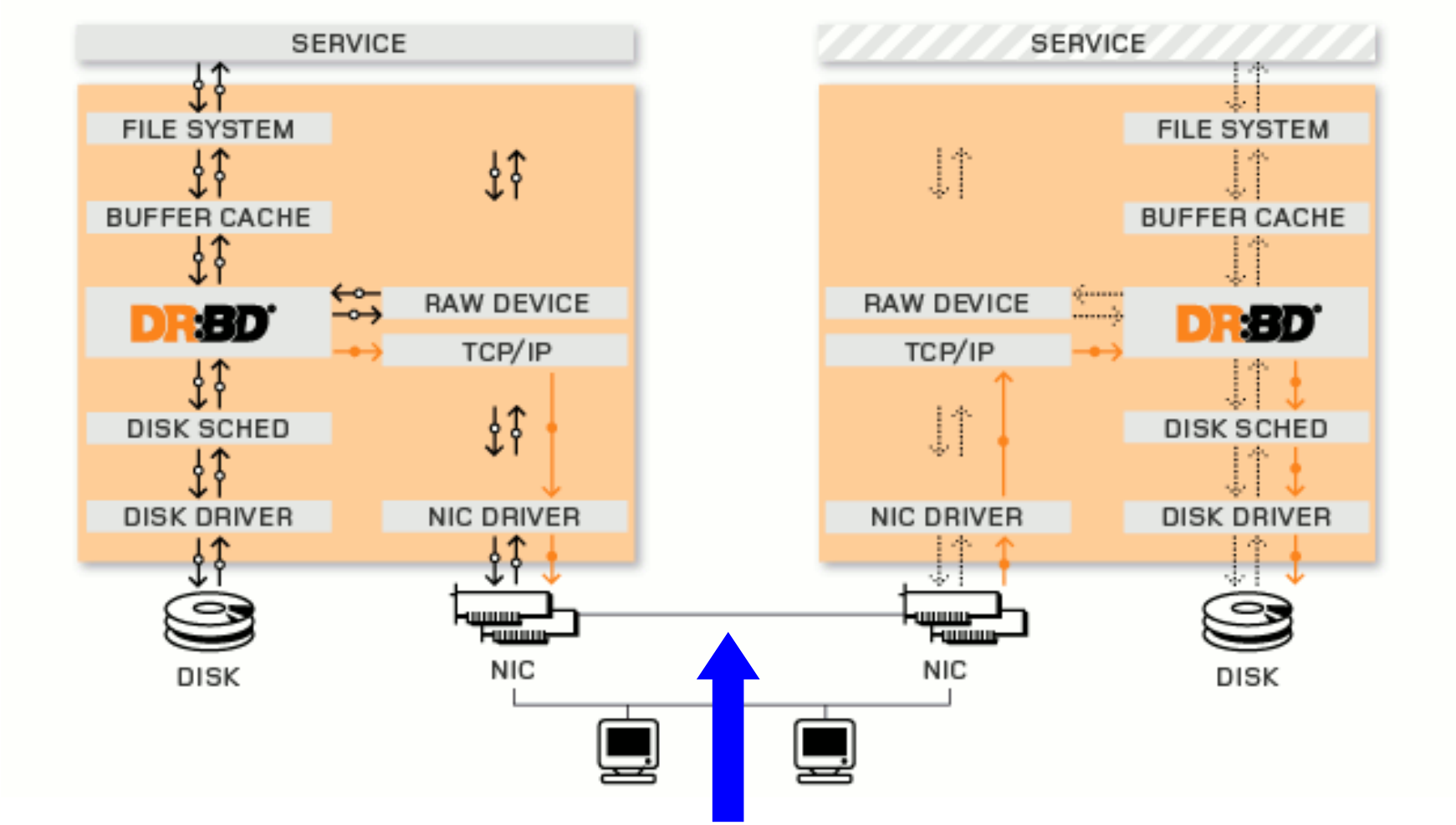
DRBDとは

Duplicated Replicated Block Device

Duplicated = 複製された
Replicated = 重ねられた
Block Device = ブロックデバイス
（なんて言えばいいのかw）

ブロックデバイスが複製されているという感じです。

ただの複製ではありません



TCP/IPネットワーク

いろいろ細かい説明は飛ばします。

- ハードウェア構成とか
- 設定方法とか
- 初期同期とか
- 基本的な管理コマンドとか

ブロックデバイスが複製できるということは？

- 通常のHDD（/dev/sda1 とか）はもちろん
- SSD デバイスが複製できる（早い）
- RAMディスクなんかも複製できる（超早い）

ブロックデバイスで認識されてば
結構いろんなものがレプリケーションされます。

ネットワークを使っているのに
なんとリアルタイムでレプリケーションされる！

TCP/IPネットワークを使うということは？

- 専用のレプリケーションI/Fがいらない
- TCP/IPなので速度はともかくVPNで遠隔地可能

夢が広がる

2台分のサーバの面倒をみないといけなくなるため、メンテナンスする
仕事の範囲も広がる…。

重要なところ

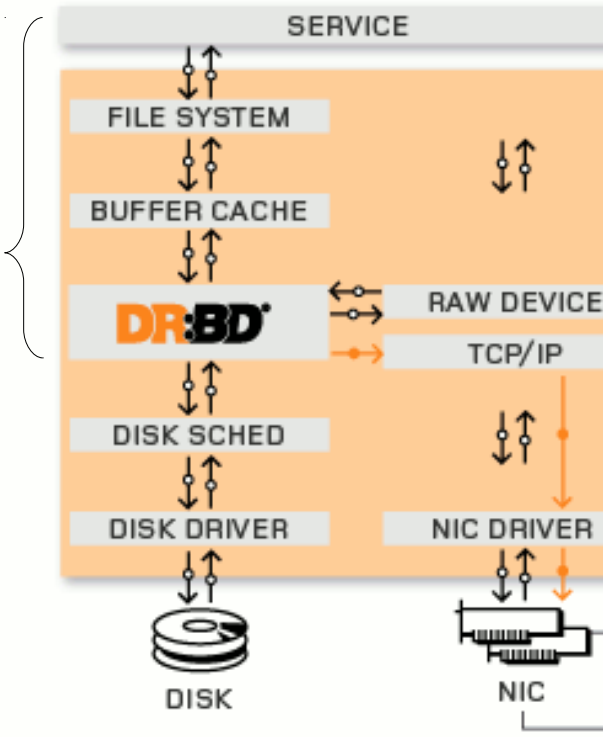
DRBDはFilesystemの下で動いている。

Filesystemを選ばない！

- ext3/ext4
- XFS
- FAT32
- NTFS

：
：

まさになんでも来い



弱点もある

- 単純にサーバが2台以上必要
- やっぱり同期するためディスクI/Oは遅くなる
- ネットワークの遅延にモロ影響される
- 壊れたデータもレプリケーションしてしまう
- NFSとかSambaのディレクトリは複製できない
  （DRBDでミラーされた領域を共有してください。）

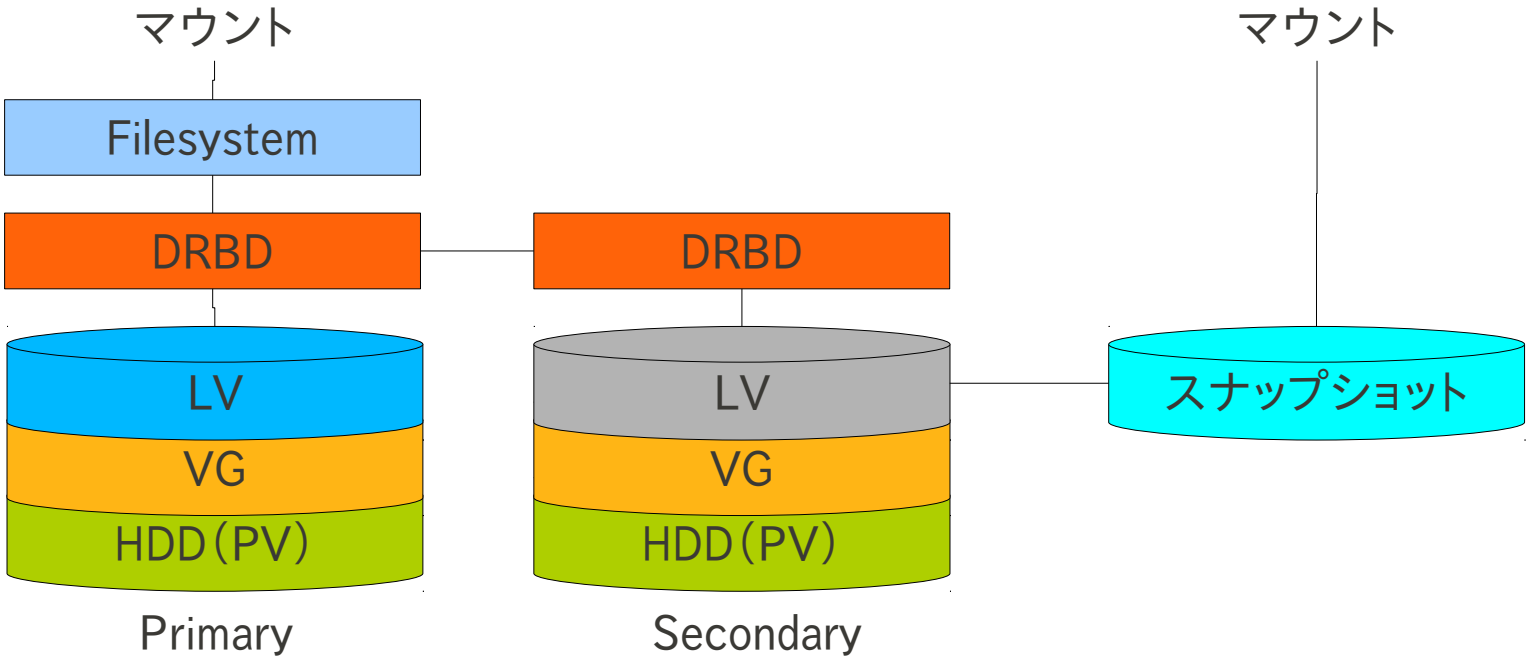
ここでちょっとしたテクニックをご紹介

DRBDとLVMの組み合わせ

- LVをミラーリング
- PVをミラーリング

LVをDRBDでミラーリングすると？

- 設定は簡単。普通にLVをミラーすればいい。

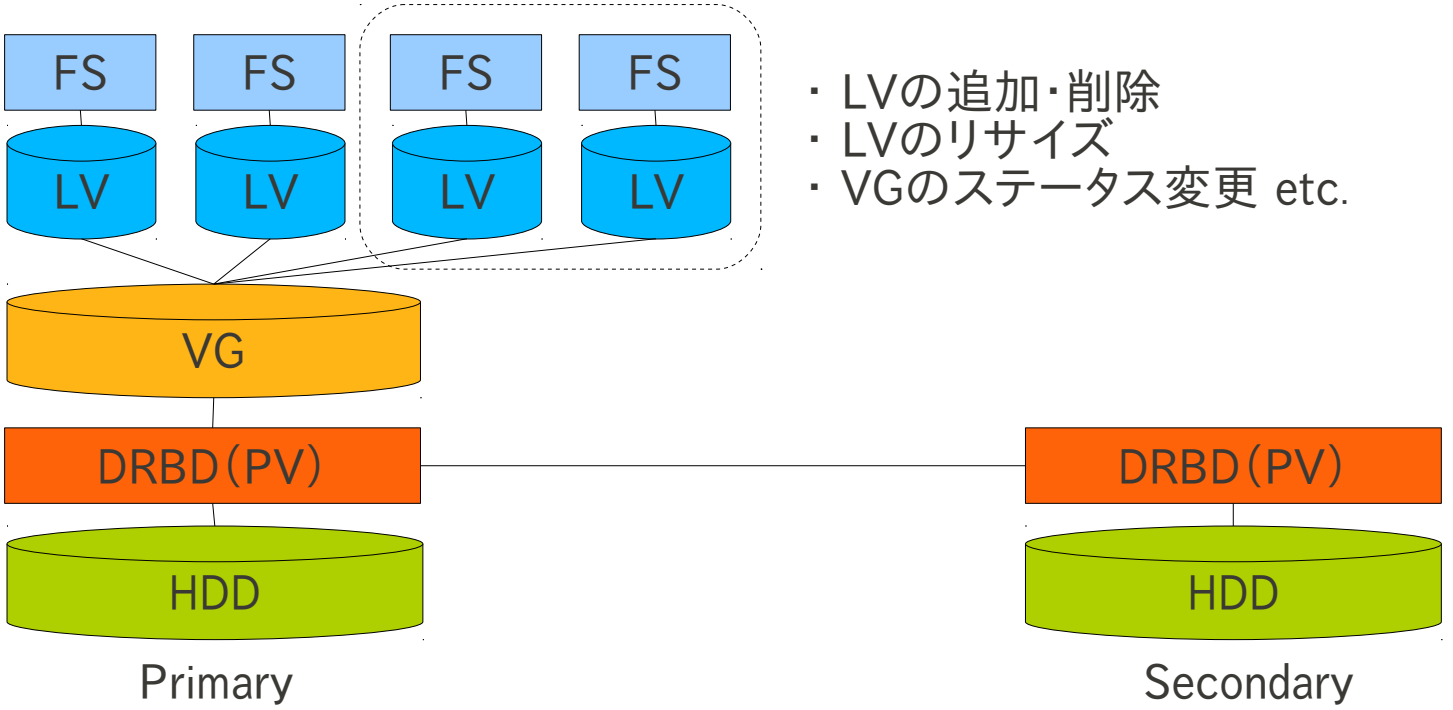


セカンダリ側でもスナップショットを取ればマウントできちゃう！

・運用側に負荷をかけずに世代バックアップを取ることができる。
・セカンダリのI/O負荷があがるなら、一旦DRBD同士の接続を切っちゃえばいい☆
・多分これDRBDでしか実現できない技な気がする。（他知らない）

PVをDRBDでミラーリングすると？

- 設定は少しおまじないが必要。(lvm.conf)



・LVの追加・削除
・LVのリサイズ
・VGのステータス変更 etc.

LVMの利便性をそのままにDRBDで全部ミラーリング可能！

・LV作る度にDRBDのリソース追加とかしなくていい
・16TBを越えた大容量に対応できるようになる。(VG使用)←DRBDは一応16TBまでサポート

PVをDRBDでミラーリングすると？

- /etc/lvm/lvm.conf のおまじないって？

```
# By default we accept every block device:
#filter = [ "a/.*/" ]

# DRBD on LVM Configuration
filter = [ "r|/dev/sd.*|" ]
```

①デフォルトだと/dev/drbd*デバイスをPVとして認識してくれない。

②物理デバイス（sda3とかsdb1とか）を認識しないように指定する。

これをやってからlvm2-monitorを再起動すると/dev/drbd*がPVとして認識される。

注) これをやらなくても「pvcreate /dev/drbd0」が通りますが、pvdisplayで確認すると物理デバイスをPVとして認識している状態となるので、DRBDでミラーリングされません。

LVをミラーするパターンのメリット・デメリット

- メリット
  - セカンダリ側でスナップショットをとってマウントできる。
  - 構築が簡単

- デメリット
  - LVのリサイズができなくなる
  - LVを追加する場合、DRBDリソースも追加設定する必要がある。
  - LVMの機能は結構制限される。

PVをミラーするパターンのメリット・デメリット

- メリット
  - LVのサイズ変更・追加等LVMの機能全開にできる
  - DRBDデバイス（PV）を増やしてVGに登録できる。

- デメリット
  - セカンダリ側でスナップショットがマウントできない
  - LVMで障害が発生するとシングルポイントになる
  - 設定と運用に少し手間がかかる。(vgchangeとか)

（直前追加）3ノードについて

- DRBDは最大4台まで同期できます。
- スタックノードというのを使ってDRBDの上にDRBDをかぶせるイメージで理解してください。
- 3ノード目をWAN越えする要求がけっこうあります。
- WANになると回線が遅いのがネックになります。
- 回線が遅いのをごまかすために
  DRBD Proxy（有料ソリューション）があります。

図はあとで追加してアップ…するかも

LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

キーワード

Disaster Recovery

（ディザスタリカバリ）

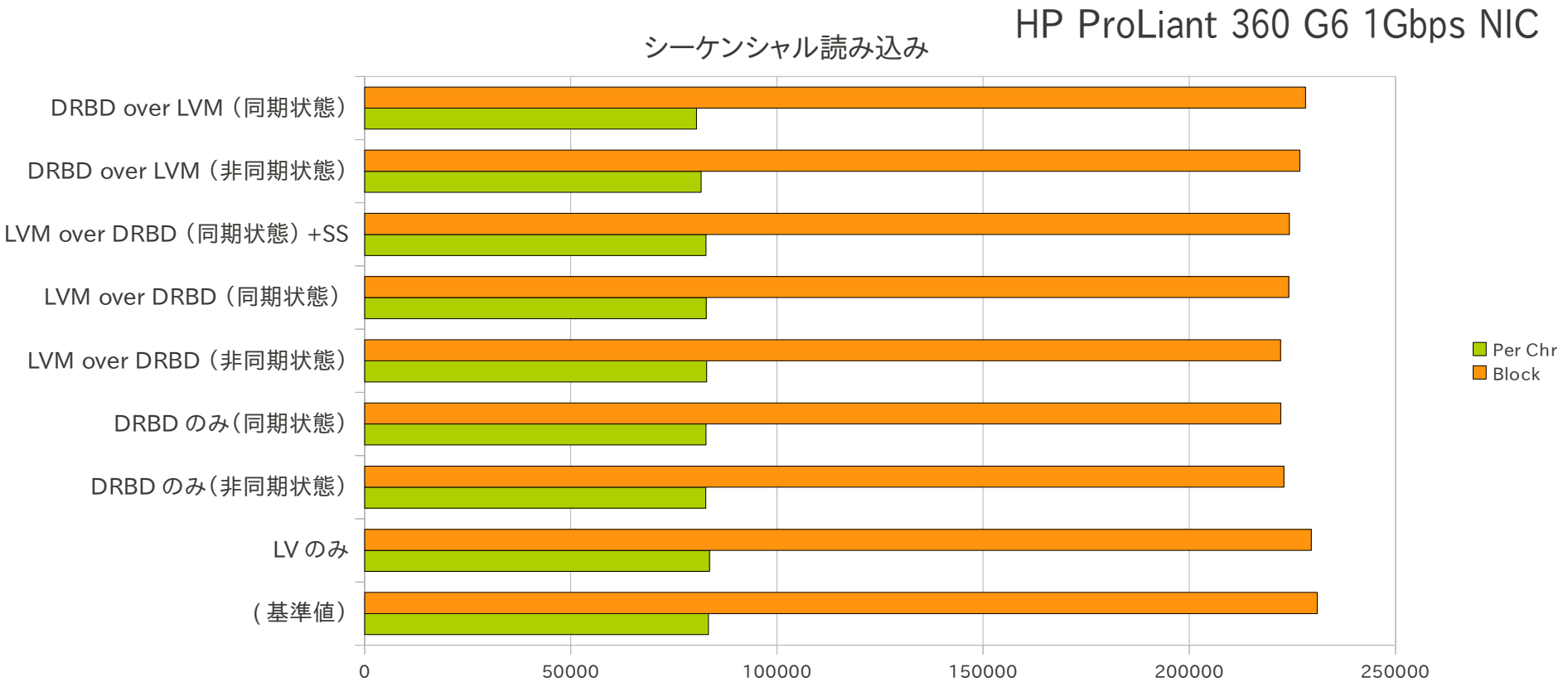
遠隔地にバックアップを置くことで、災害時にデータセンタ事態に被害が発生してもデータを消失しないための体制やしくみを言います。
FTPはVPNを使って定期的なバックアップを遠隔地に保存する手法が現在一般的なのではないでしょうか。

DRBDを使うとリアルタイムなDR環境が実現します

HighAvailability

気になるパフォーマンス

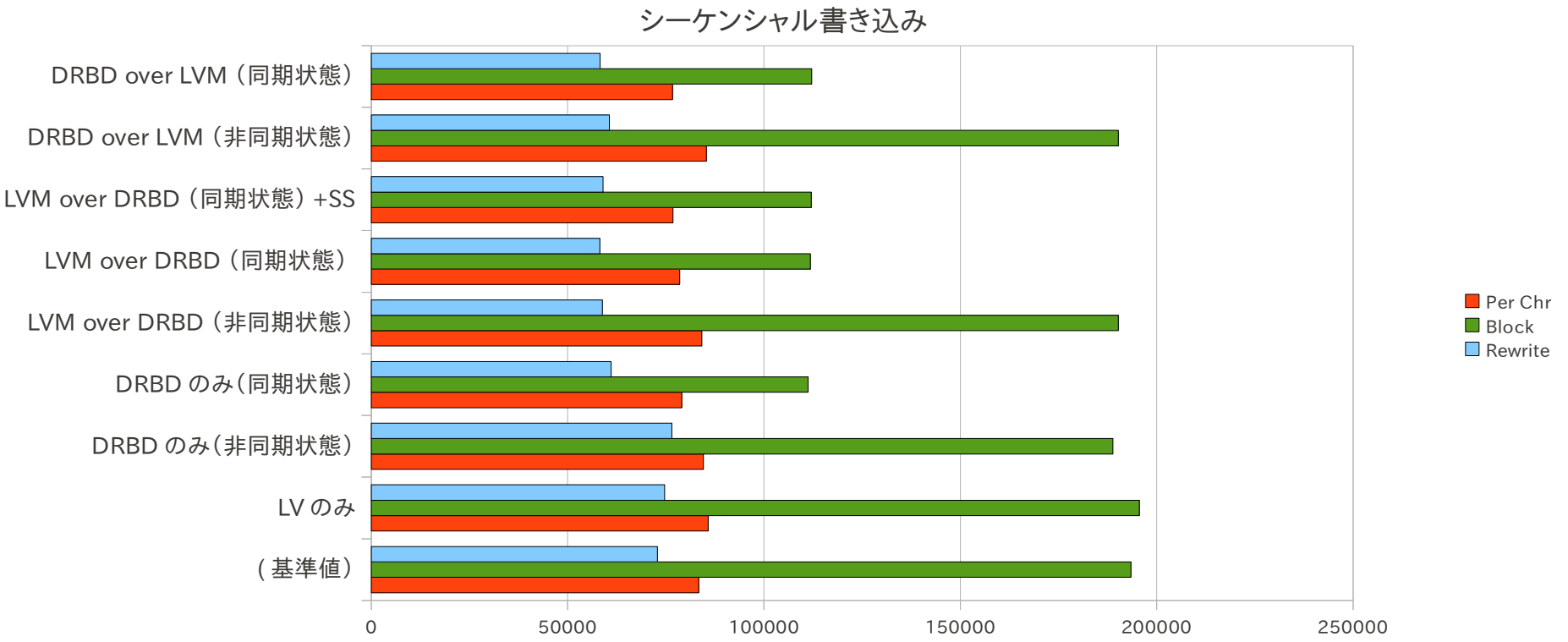
ベンチマークを取ってみました



読み込みはローカルからしか読まないので、同期非同期に関わらず一定値を出す。

ベンチマークを取ってみました

HP ProLiant 360 G6 1Gbps NIC



ChrとRewriteはもともと遅いのであまり差は出ないがBlockでは結構差が出る。

ベンチマークを取ってみました

HP ProLiant 360 G6 1Gbps NIC

すみません。このデータは「普通には」出せないことになっていますw

ファイルの作成や削除を繰り返すテストでは同期のオーバーヘッドが多くなりかなりの差が出る。

結　論

DRBDは向き不向きがあるのでDRBDを使用するアプリケーションの特性を考えて使いましょう。

当然と言えば当然なんですが…

LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

HighAvailability

LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

Pacemakerとは

HAクラスタ環境を

実現

HighAvailability

>> キーワード

これと ↓ これ ↓

HighAvailability

（ハイ アベイラビリティ）

略して **HA**

HighAvailabilityとは

High = 高い

Availability = 有用性

要するに「結構使える」ということ？

よくわかりませんw

HighAvailabilityとは2

HighAvailability

高可用性

高可用性

出典: フリー百科事典『ウィキペディア(Wikipedia)』

高可用性(こうかようせい、英: *High Availability*, **ハイ・アベイラビリティ**、*HA*)は可用性が高いことを示すIT用語。システムなどにおいて、サービス提供が出来なくなる事態の発生頻度が少ないことを指す。また、そのようなシステムを**HA構成**などと呼称する。システム的には冗長化構成を組んでいたり、バックアップ手段の確立、災害対策システムを講じることでHAを実現させる。IT業界においてはほぼクラスタリングあるいはクラスタサーバと同義で使用される用語。

LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

どのようにサービスの可用性(使える度)を高めるか

予備を用意する

同じサービスを提供できるサーバをもう一大用意して寝かせておき、
いざとなったら交換してしまえば、同じサービスが提供できる。

「Active/Standby型のフェイルオーバーHAクラスター環境」

という長い名前で呼んで説得力を無駄に稼いだりします。

HighAvailability

HA環境の基本形

提供するサービスが停止したときに、予備のサーバを
自動で起動してサービスを継続する

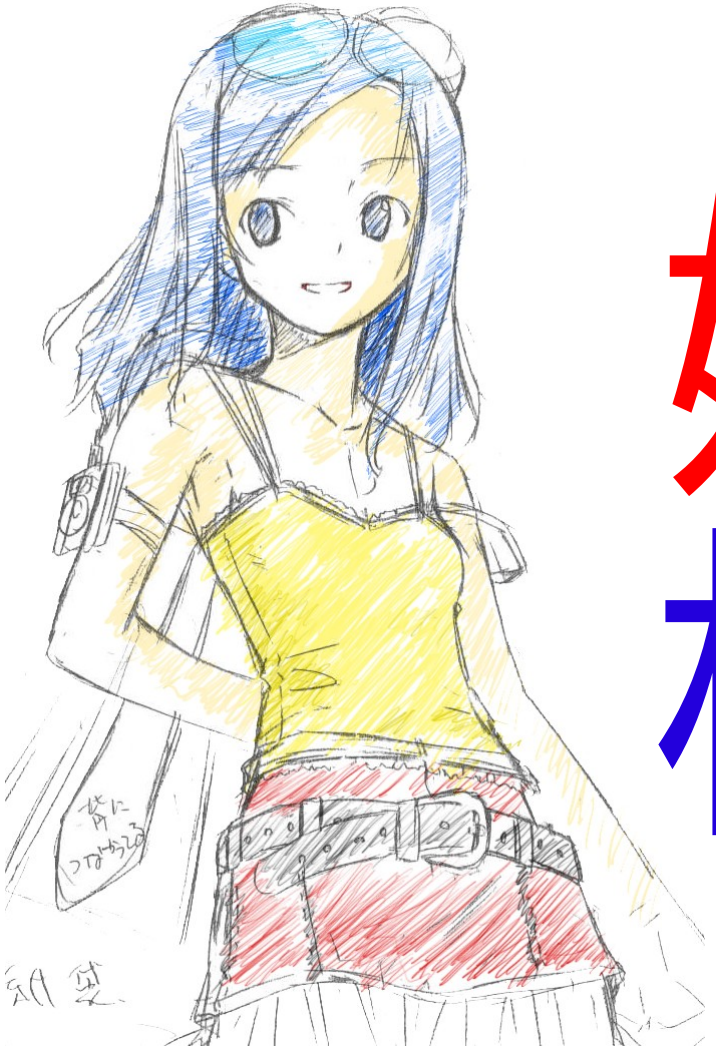
2台構成

つまり…

LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

前置き終わり



姉妹構成

HighAvailability

LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

姉妹だから

使える

（可用性が高い）

HighAvailability

LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

姉妹

＋

＝

双子の姉妹

HighAvailability

良い！

（可用性が高い）

LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

愛着が沸く

（最重要事項）

妹（姉）「ごめん、ちょっとなんか今日調子悪くて…」← 障害発生

妹（妹）「わかった、今日は私ががんばるよ！」← フェイルオーバー

俺　「しょうがないな、調子が悪いところを俺が診てやるよ」← メンテナンス

↓

妹（姉）「調子良くなったよ。心配かけてごめんね」← メンテナンス終了

妹（妹）「よかった、でもお姉ちゃんはまだゆっくりしててっ」← HA環境復旧

HighAvailability

設定

名前「高良 かな」
・CV：田中理恵
・身長165cm、体重49kg、胸はAカップ。　スポーティ。
・性格　：明るく元気。 結構気が強い姉御肌。
・髪の毛：黒で若干グレーが入るくらい。 セミロング。
　普段はアップにしていることが多い。
　髪留めに、ペースメーカーのロゴ。（姉妹おそろい）
・スポーツウーマン、結構頭もいい。
・年齢：20歳
・職業：某工学部情報工学科の2年生。
　周囲は男性が8割以上だけど、平気。
そこそこもてるけど、あまり恋愛に興味はない。
　お父さんはIT系の会社で仕事をしている。
　将来は、自分もIT系の会社に入るんだろうなぁと考えている。
　2年生だけどそろそろ就活も視野に入ってきている。
・スポーツウーマン：高校時代、陸上部。種目は走り高跳び。
　インターハイに出たことがある。
　（でも優勝とか言うレベルではない）
　大学に入ってからは、もっぱらバイトの日々。
・家族構成：長女。 高校2年生の妹がいる。 4人家族。
　両親は健在。 お母さんは専業主婦。
・生い立ちなどの設定　江戸っ子。 普通のサラリーマン家庭で育つ。
　パソコンも使うけど、妹ほどではない。
　どちらかと言うと身体を動かしていた方が好き。
・格好　：スーツ。
　ボーイッシュな、さっぱりした格好が中心。
　色は寒色系。
・口癖はある？　：ちょっと姉御肌な声
　「てやんでぃ」
・特技　：運動
　字は汚い。
・趣味　：ネットサーフィン。
　身体を動かす。
・妹が胸が大きいことが、ちょっとだけコンプレックス。

これイラスト制作用のラフ画
なので、公式キャラクターは
もうちょっと変わる予定です

色も適当です

設定

・名前「高良 かよ」
・CV：丹下桜
・身長160cm、体重49kg、胸はEカップ。
・性格　：明るいオタク。
・髪の毛：黒で若干茶色が入るくらい。 ロング。
　普段はポニーテールで、大きなリボンを付けている。
　髪留めに、ペースメーカーのロゴ。（姉妹おそろい）
・年齢：17歳
・職業：ミッション系女子高の2年生。 成績は中の上。
・オタク：ネットとアニメが大好き。
　ファンタジー系のラノベも好き。
　コスプレも少しだけするけど、人前でやるのはちょっと恥ずかしい。
・家族構成：次女。 お姉さんはかなちゃん。 4人家族。
　両親は健在。 お母さんは専業主婦。
　お父さんはIT系の会社で仕事をしている。
・生い立ちなどの設定　江戸っ子。 普通のサラリーマン家庭で育つ。
　小4からパソコンを触っている
　パソコンとかネットには詳しいがITはそれほどでもない。
・格好　：かわいい系の格好。
　メイド服（コスプレ）
　学校の制服はセーラー服（東京女学館参照）
・口癖はある？　：「萌え～」
　舌っ足らずで、少し噛む。
・特技　：パソコン
　ブラインドタッチが出来る。
　携帯を打つ早さは誰にも負けない。
・趣味　：ブログ
アフリエイトで小遣い稼ぎ
コスプレを少々。
　身体を動かすことは苦手。
　ピンク系の色が好き。 服もそんな感じが多い。
・部屋は姉妹別。 となりどうし。
・お姉ちゃんがなんでもできることがちょっとだけコンプレックス。

これイラスト制作用のラフ画
なので、公式キャラクターは
もうちょっと変わる予定です

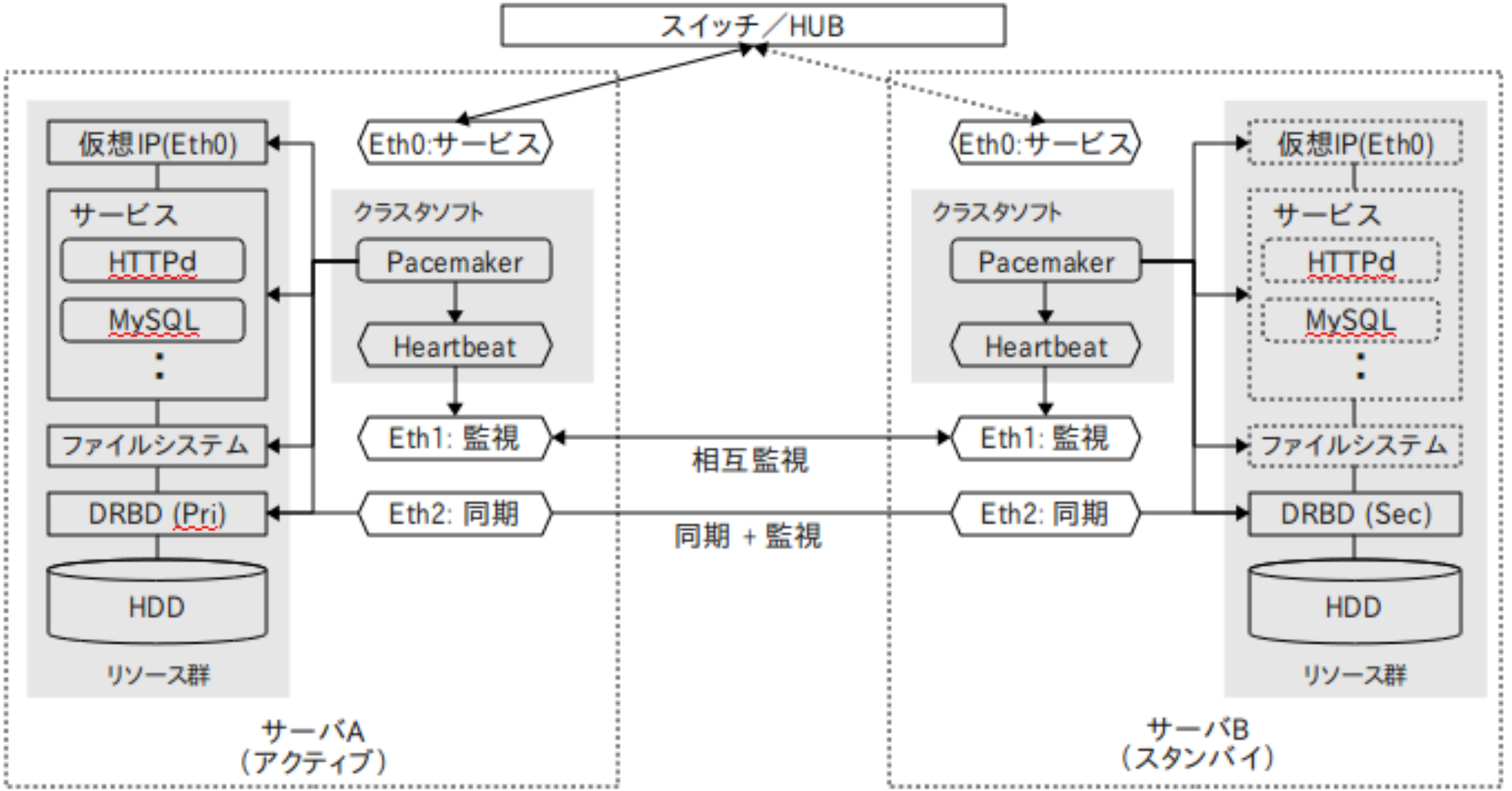
色も適当です

設定終了

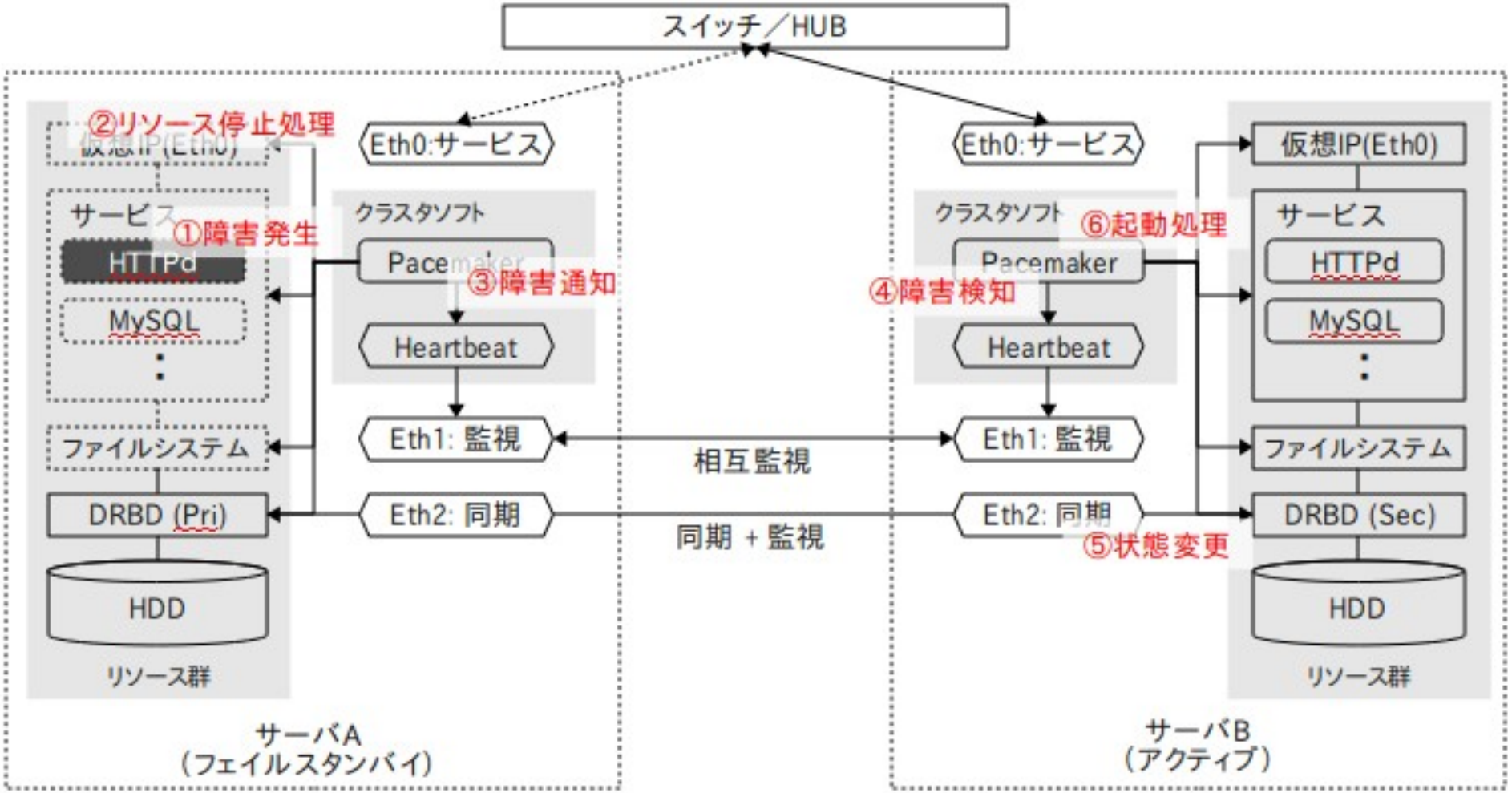
Pacemakerの動き

- リソースはCRM(クラスターリソースマネージャ)で管理されます。
- CRMコマンドという管理コマンドで設定します
- CIB.xmlというファイルで設定値と現在の状態を一元管理しています
- Heartbeat V2にもCRMが実装されていますが、分離してPacemakerとなりました。

HA環境の例


スイッチ／HUB
仮想IP(Eth0)
Eth0:サービス
サービス
HTTPd
MySQL
クラスタソフト
Pacemaker
Heartbeat
ファイルシステム
Eth1: 監視
DRBD (Pri)
Eth2: 同期
HDD
リソース群
サーバA
（アクティブ）
相互監視
同期 + 監視
Eth0:サービス
クラスタソフト
Pacemaker
Heartbeat
Eth1: 監視
Eth2: 同期
仮想IP(Eth0)
サービス
HTTPd
MySQL
ファイルシステム
DRBD (Sec)
HDD
リソース群
サーバB
（スタンバイ）

HA環境の例


スイッチ／HUB
②リソース停止処理
Eth0:サービス
Eth0:サービス
仮想IP(Eth0)
サービス
①障害発生
クラスタソフト
クラスタソフト
⑥起動処理
サービス
HTTPd
Pacemaker
Pacemaker
③障害通知
④障害検知
HTTPd
MySQL
MySQL
Heartbeat
Heartbeat
ファイルシステム
Eth1: 監視
Eth1: 監視
相互監視
ファイルシステム
DRBD (Pri)
Eth2: 同期
Eth2: 同期
同期 + 監視
⑤状態変更
DRBD (Sec)
HDD
HDD
リソース群
リソース群
サーバA
（フェイルスタンバイ）
サーバB
（アクティブ）

LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

キーワード

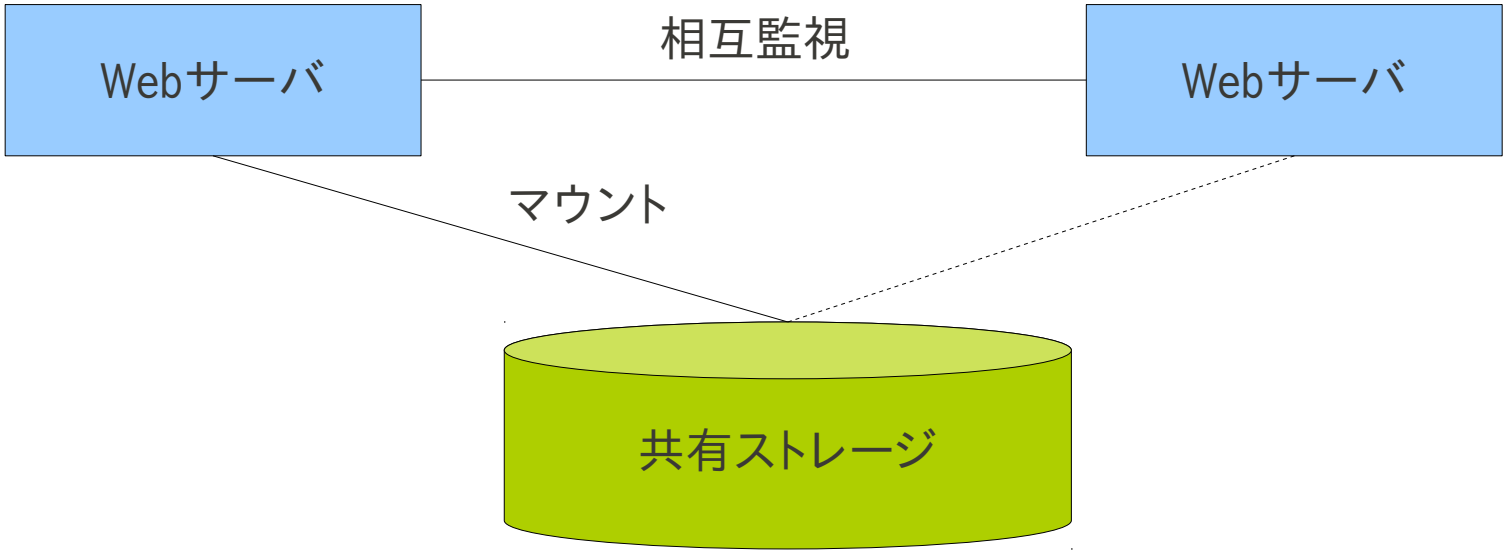
SHARED Nothing Cluster

(シェアドナッシングクラスター)

HighAvailability

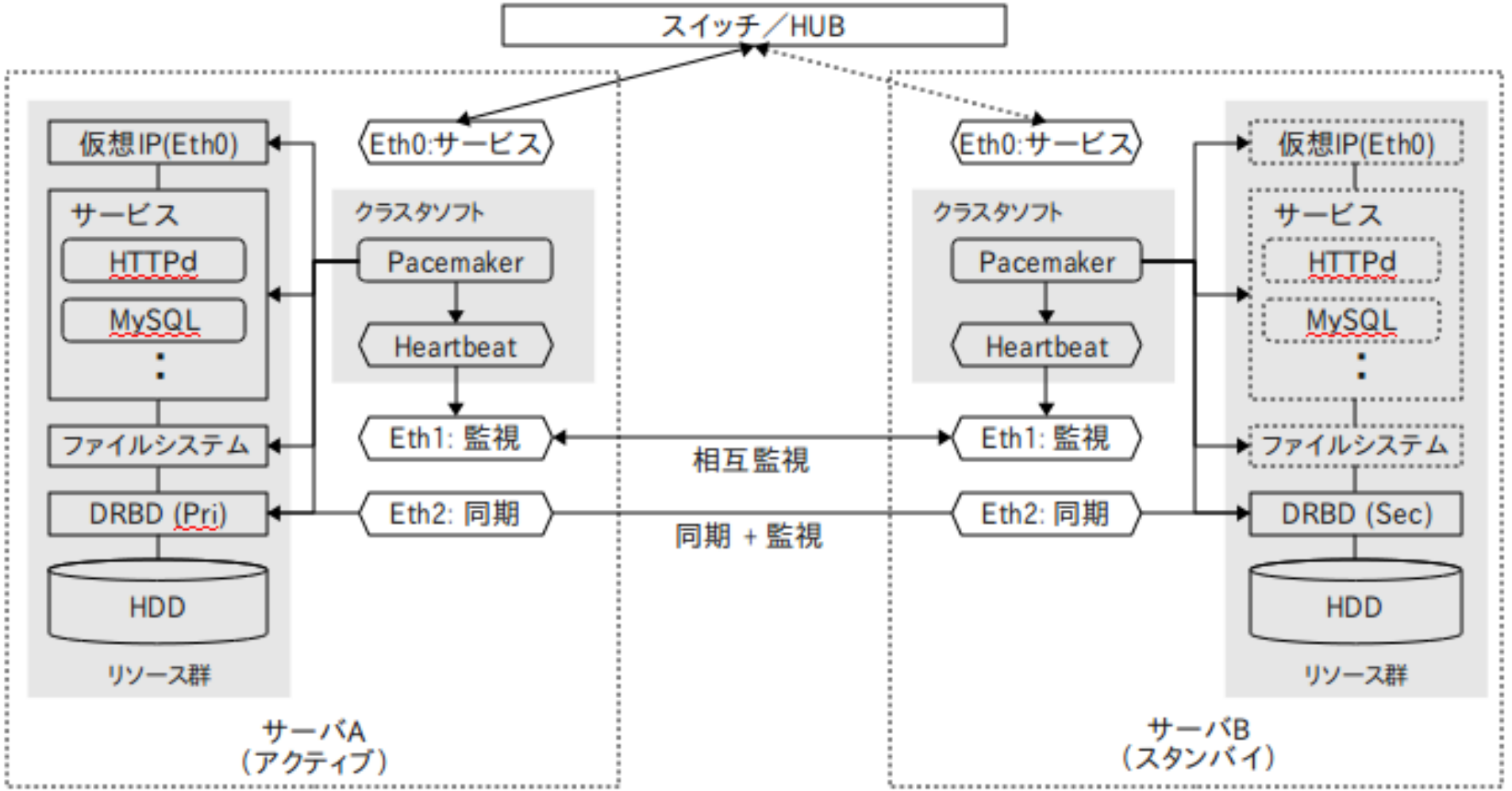
DRBDを使わないクラスター環境

サービスを提供するサーバは、2台構成のHAクラスタを組んでおり、データは共有ストレージを使用している。



・共有ストレージダウンしてしまうとサービス継続は不可能となる。
・共有ストレージのデータが壊れてしまうとどうしようもない。
（RAIDを組んだりして対応している）

シェアドナッシングクラスター



DRBDとPacemakerを組み合わせると

簡単低コストで
シェアドナッシングの
HAクラスター環境の
構築が可能となります。

よく遭遇する障害

LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

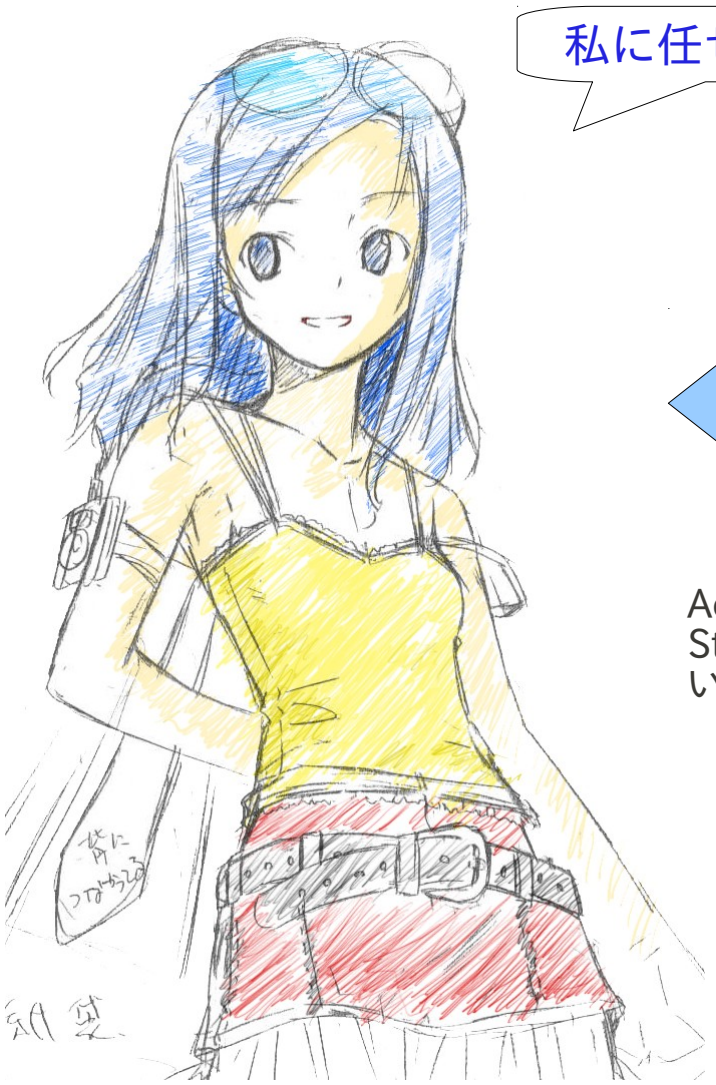
キーワード

SPLIT BRAIN
（スプリットブレイン）

頭が2つに割れてしまうという意味？

HighAvailability

正常な時は…



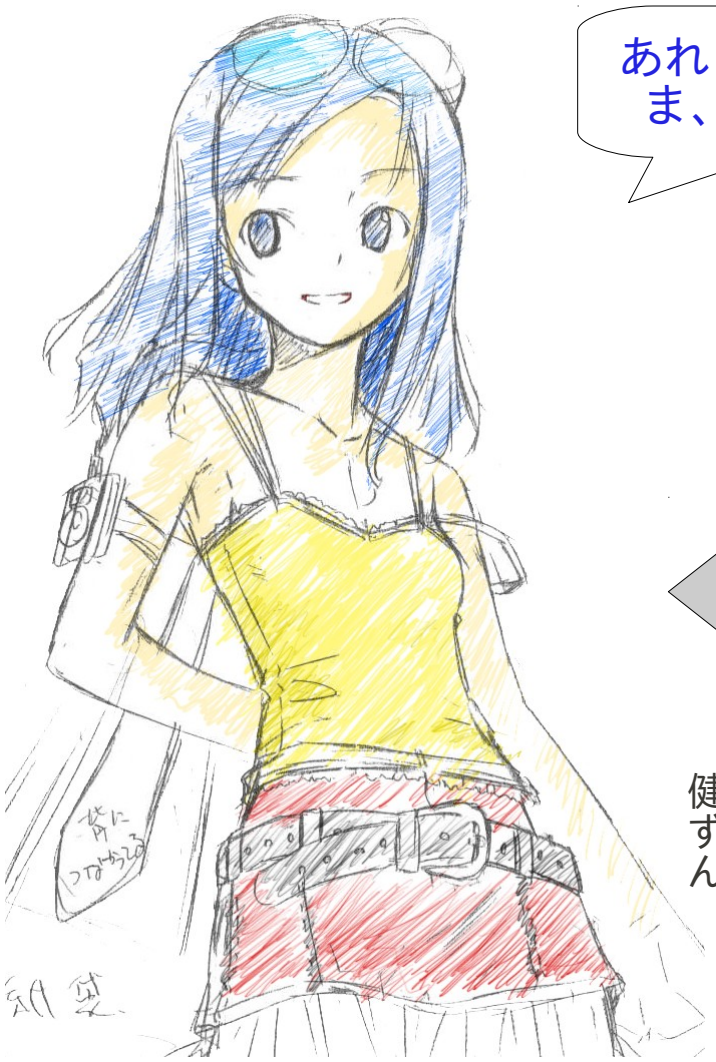
Active

Standby

意思の疎通ができてる

Active側でサービスが起し、
Standby側ではサービスが上がらない状態で待機される。

スプリットブレイン状態だと…



Active

Active

意思の疎通ができない

健気な妹たちは、意思の疎通ができず、お互いがActiveになろうとしてがんばってしまいます。

スプリットブレインを起こしてしまうと…

- IP重複が起きる可能性がある
- 同時に2つのノードから共有ストレージをマウントしてしまい、ファイルシステムが壊れる
- その他いろいろ

重大な障害へ発展する可能性高い

Pacemaker＋DRBD環境でのスプリットブレイン

- IP重複は起きてしまう場合がある
- それぞれがストレージを持っているため、同時マウントの障害は発生しない。
- ただデータの同期が停止する（安全装置）

重大なデータ障害へ発展する可能性が比較的低い。（落ちれば全部重大）

同期されなかった差分は自動または人間の目でデータを確認しながら抽出し、再度同期を開始させることで元の状態に戻る。
データが破損されるリスクも共有ストレージに加えて低くなります。

まとめ

- DRBDはファイルシステムを問わない
- LVMと組み合わせて楽しくも辛くもなれる
- リアルタイムなディザスタリカバリ環境の実現
- Pacemakerと組み合わせると世界がぐっと広がる
- 単一障害点のHA化にもってこい（シェアドナッシングクラスター環境）
- パフォーマンス低下に注意。

姉妹構成で愛情溢れる保守が可能となる。

事例紹介

- シチズン時計（DRBD over LVM 3ノード)
- 岐阜女子大学（LVM over DRBD 3ノード）
- 米国911センター（業務ごと冗長化4ノード？）

コミュニティ紹介

LINUX-HA JAPAN

HIGH-AVAILABILITY CLUSTERING ON LINUX

Linux-HA Japan 活動内容

- MLの運営
- 技術情報の公開
- OSC等のイベントに参加
- 飲んだり／食べたり
- つぶやいたり
- いろいろ作ってみたり？

ご静聴ありがとうございました。