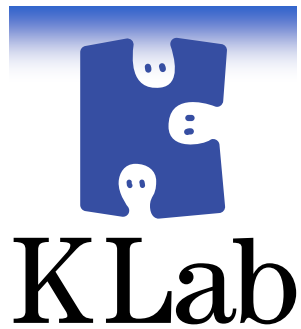
DSASのあそこ

～ストレージサーバ編～

第2回 KLab 勉強会

http://dsas.blog.klab.org



2007年6月22日

KLab 株式会社
Kラボラトリー
ひろせ まさあき

KLab

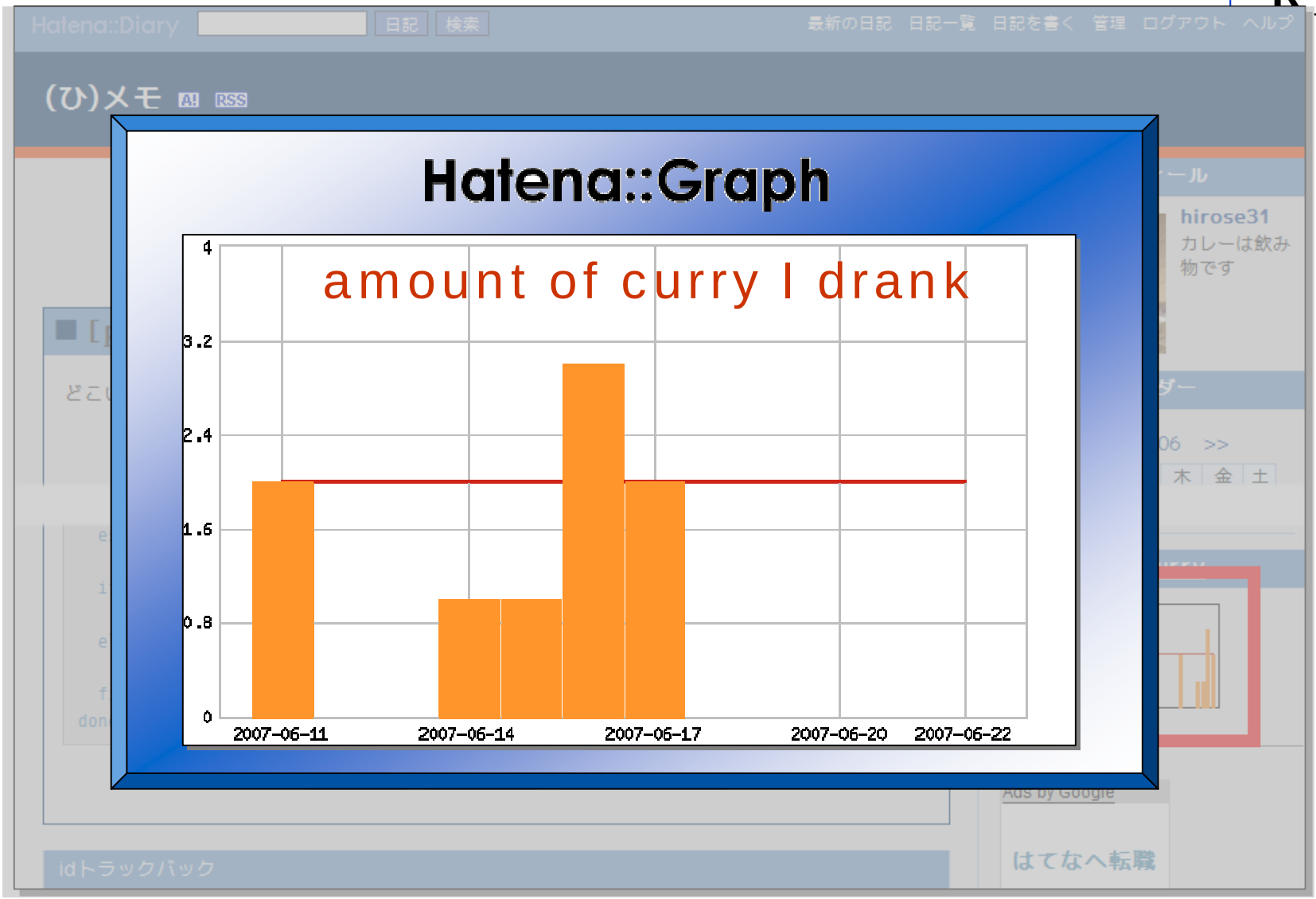
自己紹介

自己紹介

- 名前：ひろせ まさあき
- はてなID：id:hirose31

- KLab株式会社
  - Kラボラトリーに所属

http://d.hatena.ne.jp/hirose31/



アジェンダ

- ストレージサーバの作り方
- 要件の整理
  - データの保全
  - 可用性の確保
- 実装方法の検討
  - とまらないストレージサーバ
  - ストレージへのアクセス方法
- 構築
- 落とし穴
  - DRBD
  - NFS

アジェンダ

KLab

- 要件の整理
  - データの保全
  - 可用性の確保
- 実装方法の検討
  - とまらないストレージサーバ
  - ストレージへのアクセス方法
- 構築
- 落とし穴
  - DRBD
  - NFS

こんなストレージサーバが欲しい（1）

KLab

- こんなファイルを保存することを考えます

- サイズ
  - そこそこでかい
  - 数MB～数十MB
- 数
  - とてもたくさん
- 具体的には
  - 画像、音、動画、着メロデータなど

こんなストレージサーバが欲しい（2）

KLab

- ストレージサービスの要件
  - データがなくならない
  - サービスがとまらない

データがなくならない

KLab

- ディスク故障からデータを守る方法を考える

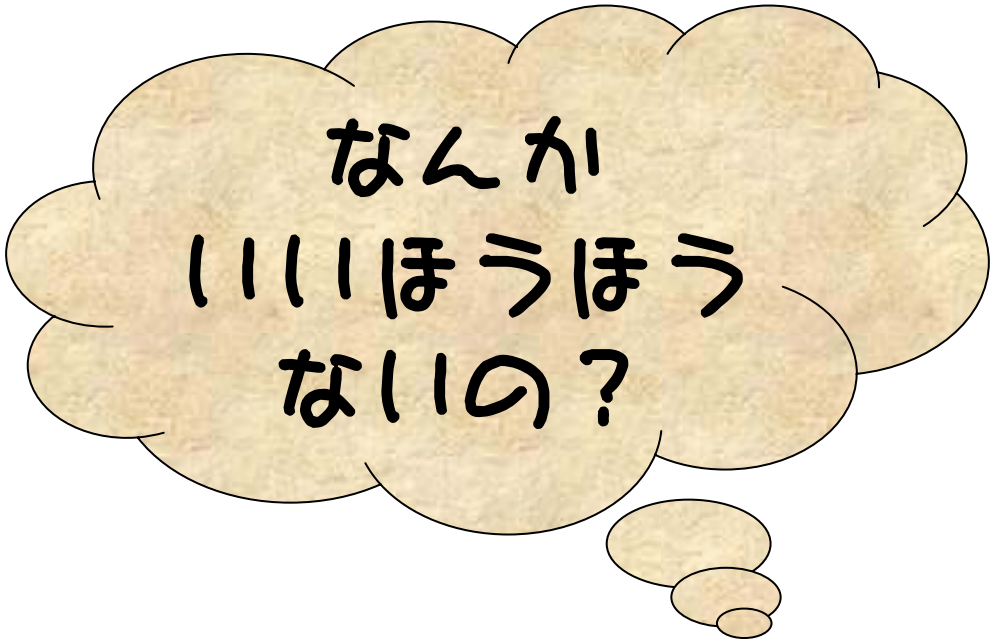
- RAIDはどうか？
  - ☺ 有益だが、万能ではない
  - 💣ディスクが同時に2台壊れるとデータ破損 (RAID1,5)
- バックアップはどうか？
  - フルバックアップ
    - 💣領域が大きいと時間がかかる
  - 差分バックアップ
    - 💣ファイル数が多いと差分判定に時間がかかる
  - 💣バックアップ中は負荷がかかる
  - 💣定期バックアップは時間が経つと本データと乖離する

サービスがとまらない

KLab

- RAIDはどうか？
  - ☺ ディスク故障
  - 💣RAIDコントローラが壊れたら？
  - 💣RAIDと関係ない部分が壊れたら？
    - 電源
    - メモリ
- バックアップはどうか？
  - 💣リストアに時間がかかる
    - ➔ 復旧に時間がかかる
  - 💣リカバリできないかもしれない
    - ➔ 障害時データとバックアップデータの乖離

なんかいいのないの？



DRBD

アジェンダ

KLab

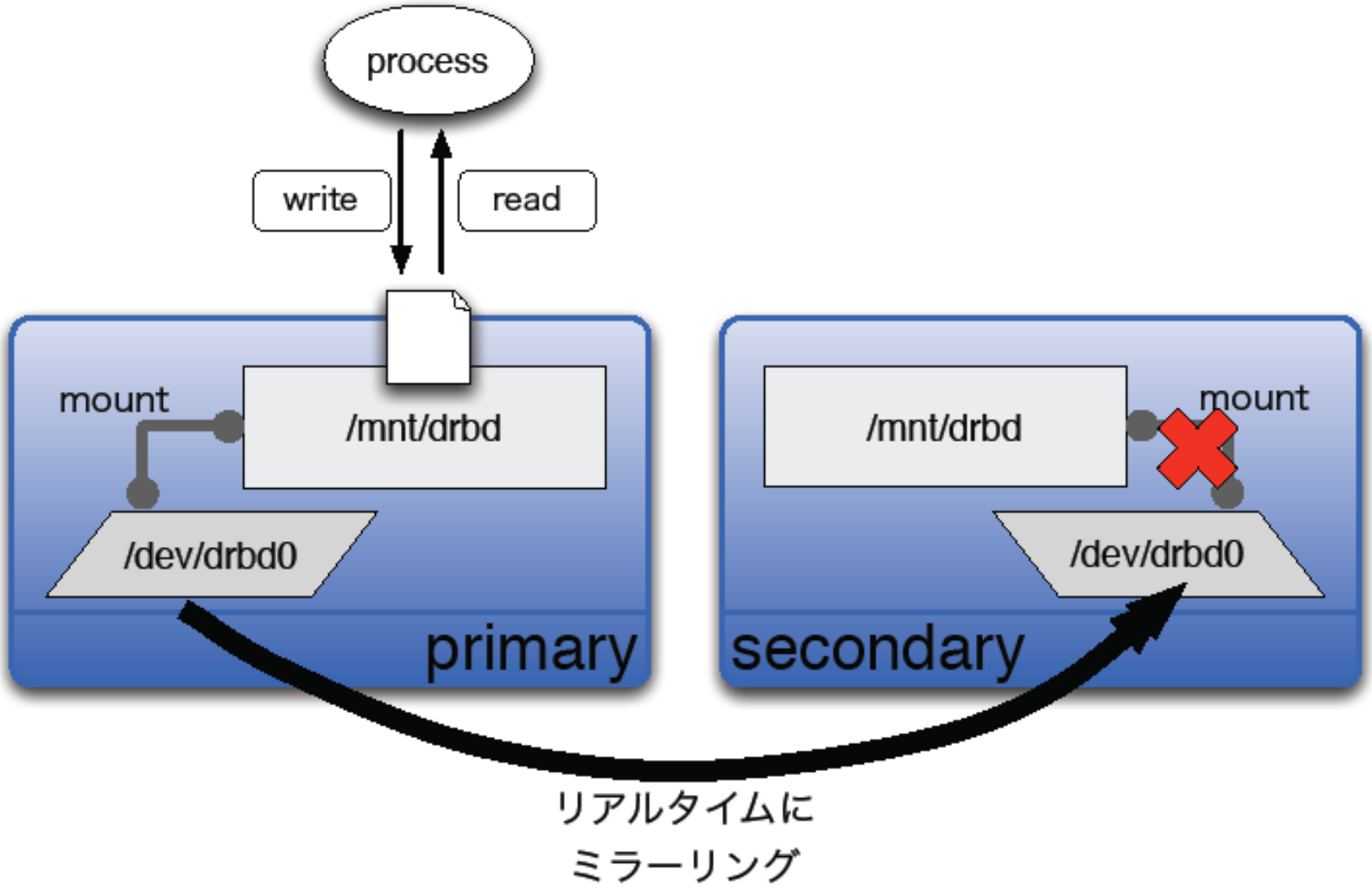
- 要件の整理
  - データの保全
  - 可用性の確保
- **実装方法の検討**
  - **とまらないストレージ**
  - ストレージへのアクセス方法
- 構築
- 落とし穴
  - DRBD
  - NFS

DRBDとは？

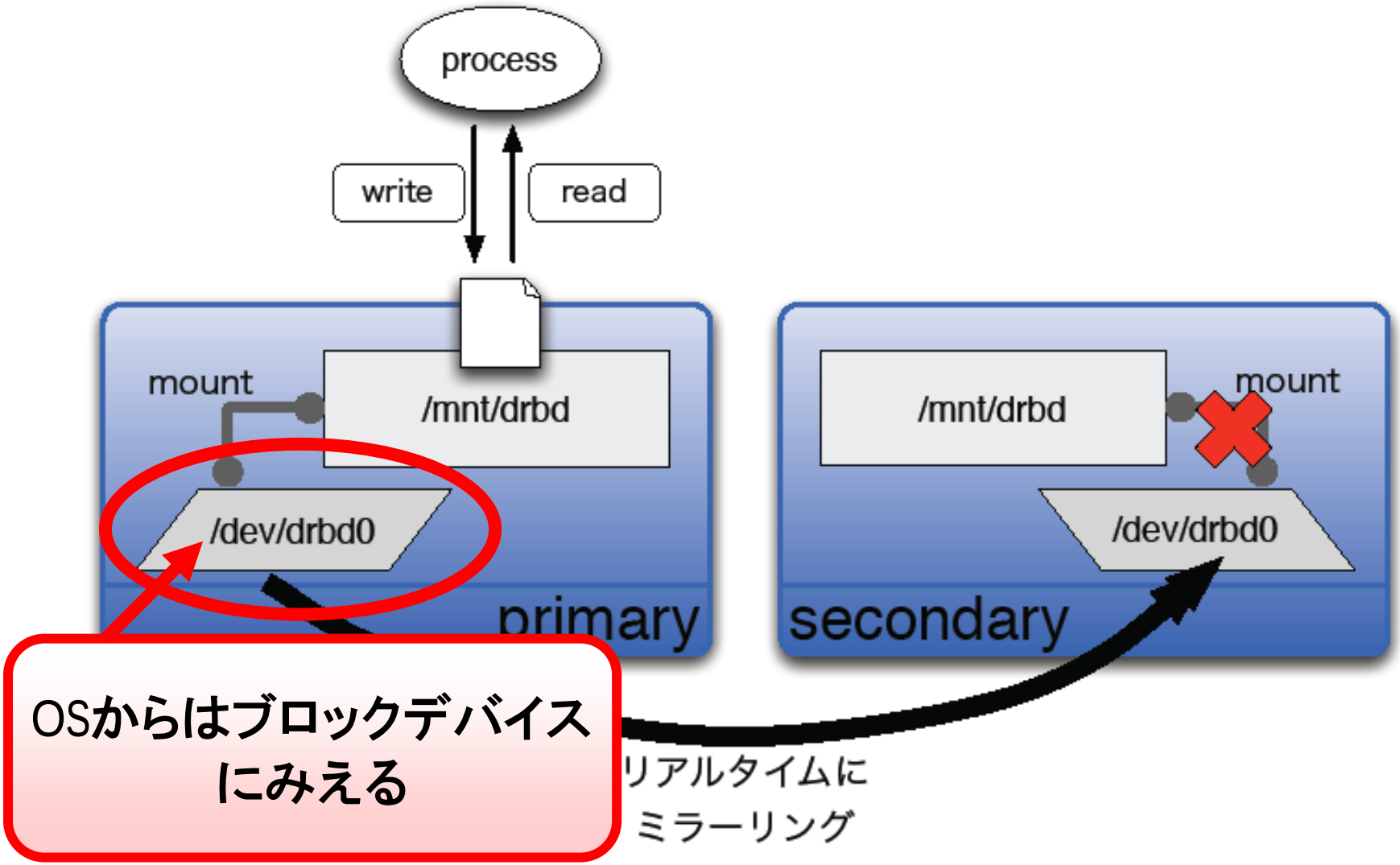
KLab

- DRBD - Distributed Replicated Block Device
  - http://www.drbd.org
  - kernel moduel ＋ drbdadm コマンド
- しくみ
  - DRBD
    - サーバ対サーバの　ネットワーク越しの　ミラーリング
  - RAID 1
    - ディスク対ディスクの ローカルバス越しの ミラーリング

DRBDのしくみ

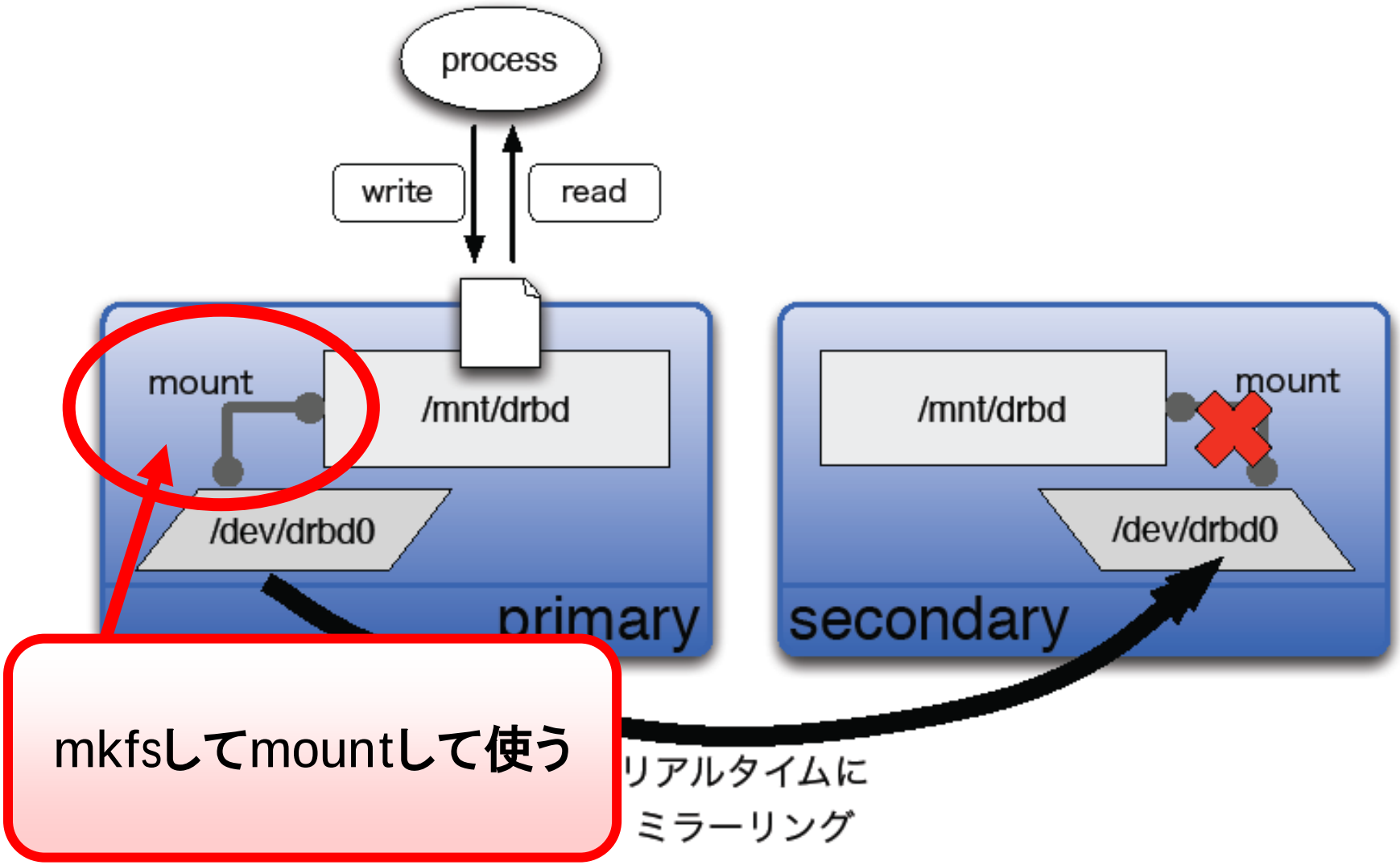


DRBDの特徴



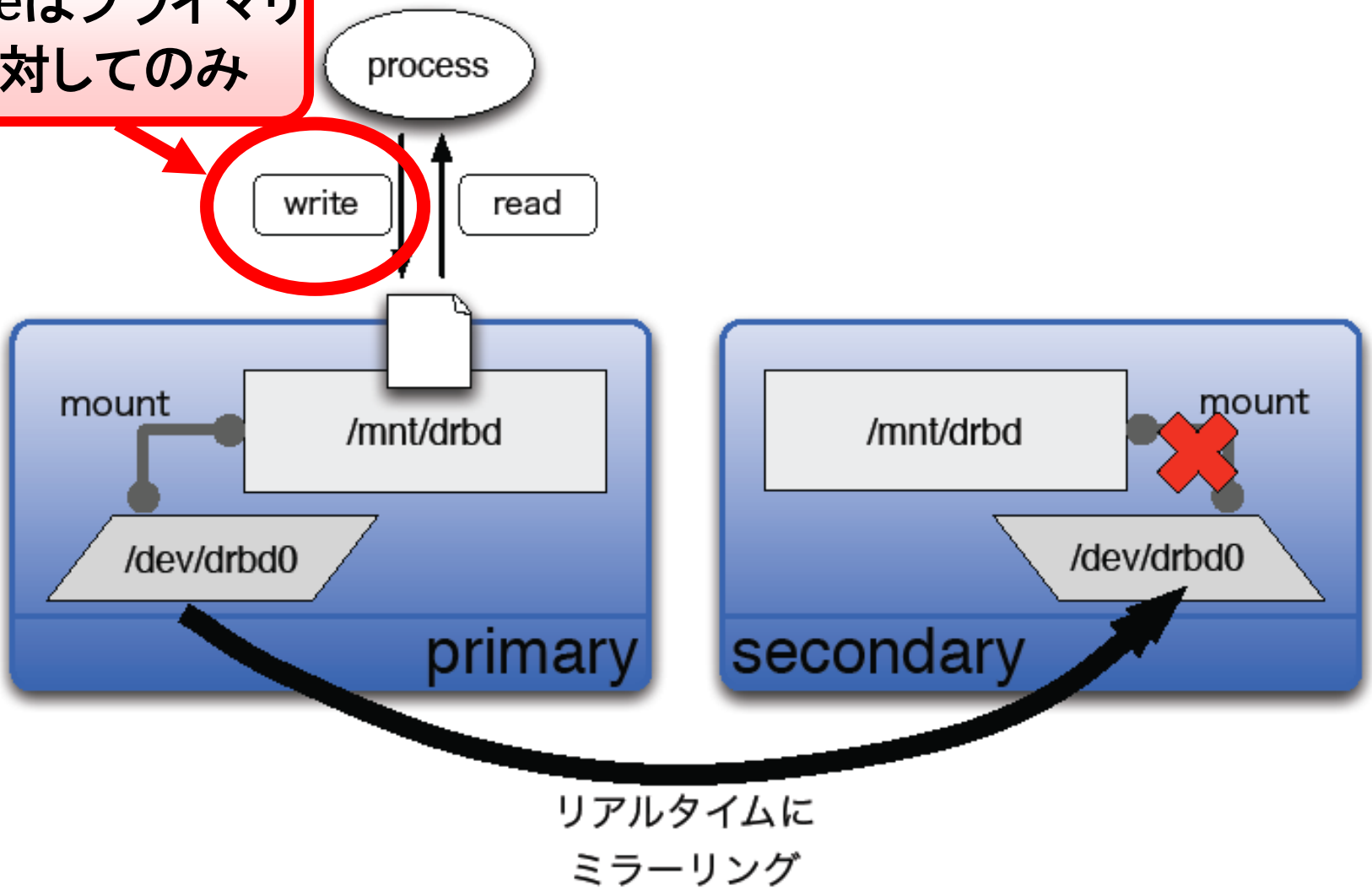
DRBDの特徴

KLab

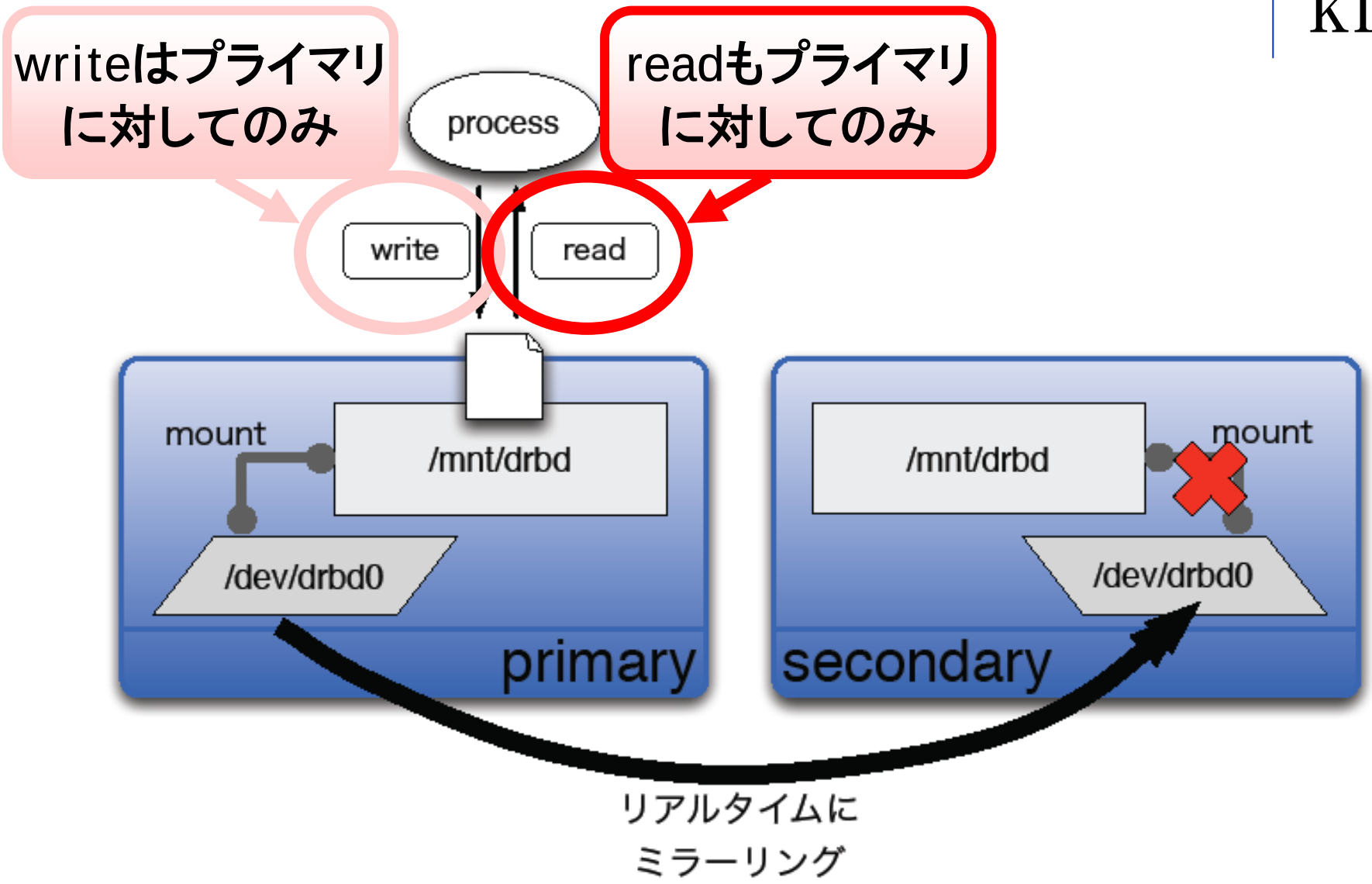


DRBDの特徴

writeはプライマリに対してのみ



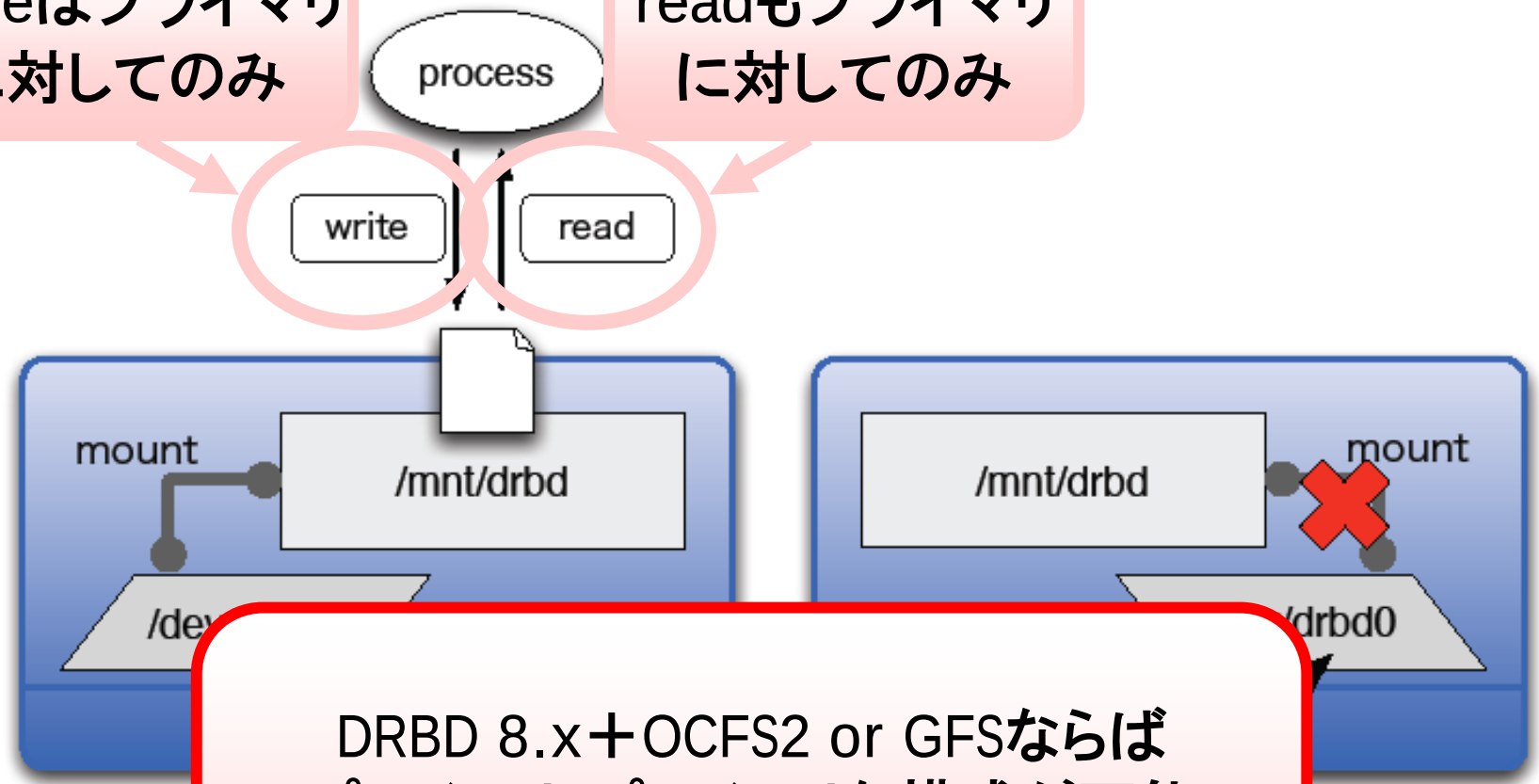
DRBDの特徴



DRBDの特徴

writeはプライマリに対してのみ

readもプライマリに対してのみ



DRBD 8.x＋OCFS2 or GFSならば
プライマリ/プライマリな構成が可能
（らしい）

解決したい問題の再確認

KLab

- データがなくならない
- サービスがとまらない

DRBDで解決できる問題（1）

KLab

- データがなくならない
  - リアルタイムに別サーバにミラーリング

DRBDで解決できる問題（2）

KLab

- サービスがとまらない
  - プライマリが故障した場合はセカンダリを昇格すればおk（フェイルオーバ）

➔ダウンタイムを短くできる

フェイルオーバ(Ｆ/Ｏ)はどう実装するか？

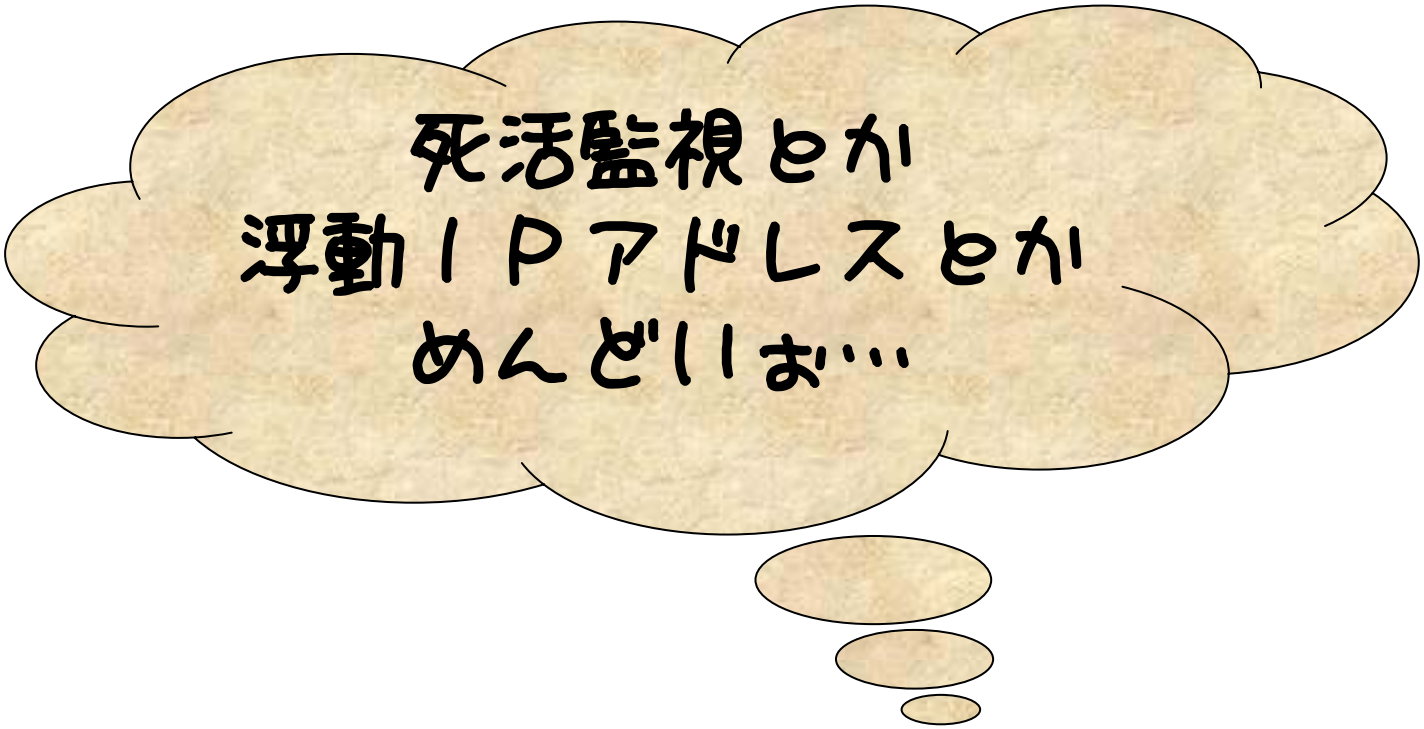
KLab

- DRBD自体には自動F/Oの機構がない
  - いくつかの管理コマンドを実行する必要がある

  ➔死活監視＋管理コマンド実行が必要

- F/O時にIPアドレスが変わらないようにしたい
  - F/Oをクライアントに意識させない

  ➔浮動する仮想IPアドレスをプライマリに付与すればいい



keepalivedでできるよ

KLab

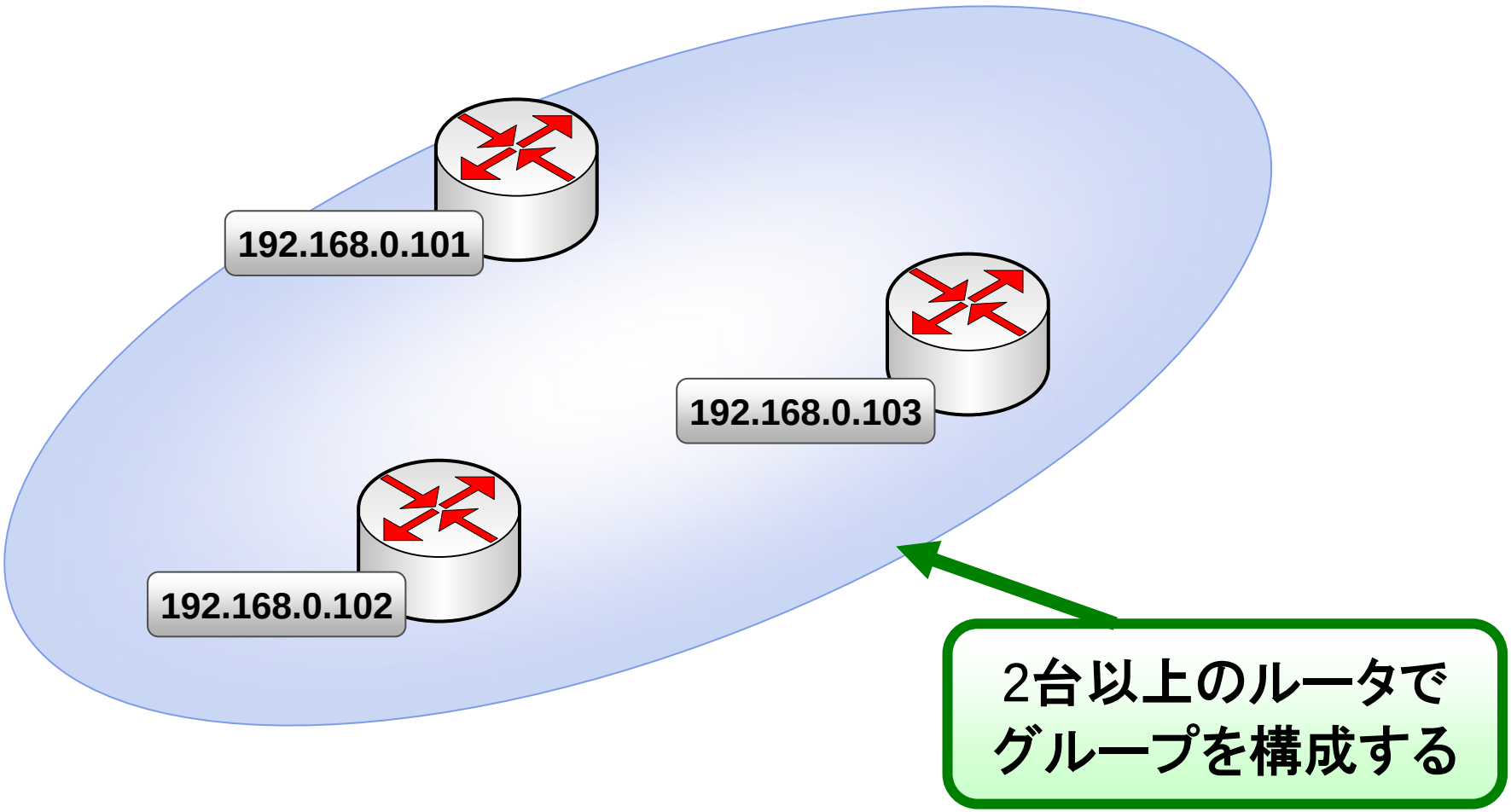
- keepalivedの2つの機能
  - 1) IPVSによるロードバランスとリアルサーバの死活監視
  - 2) VRRPによるアクティブ / バックアップ構成の冗長化 ←今日はこっち

VRRPって？

KLab

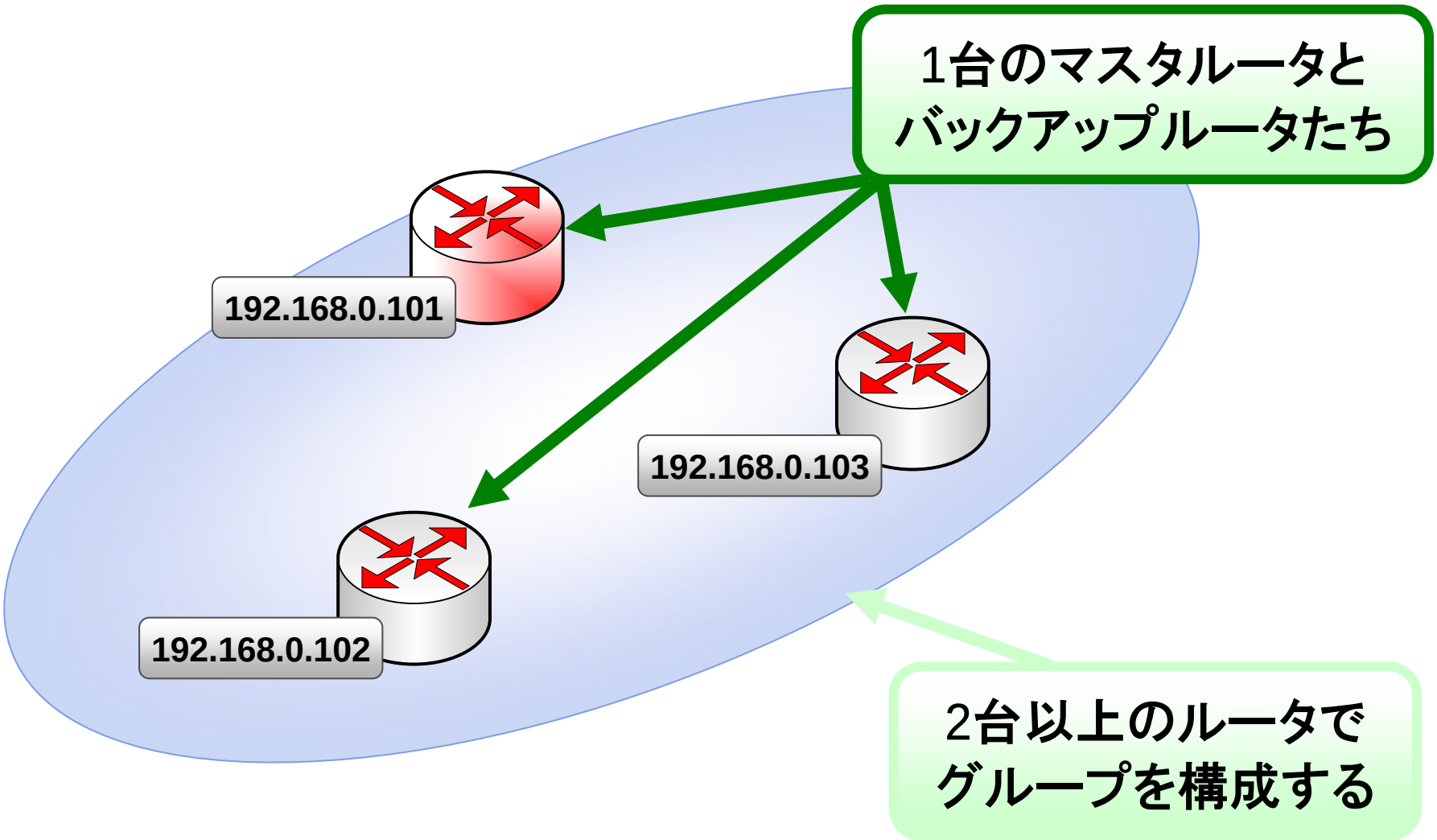
- VRRP - Virtual Router Redundancy Protocol
  - RFC 3768
- VRRPの概要
  - 2台以上のルータでグループを構成する
  - 1台のマスタルータと1台以上のバックアップルータ
  - マスタルータは、浮動する仮想ルータアドレスを保持する
  - マスタルータが停止すると、バックアップルータのいずれかがマスタに昇格する
- VRRPの死活監視
  - マスタルータが定期的にマルチキャストする
    - VRRP Advertisement Message

VRRPって？



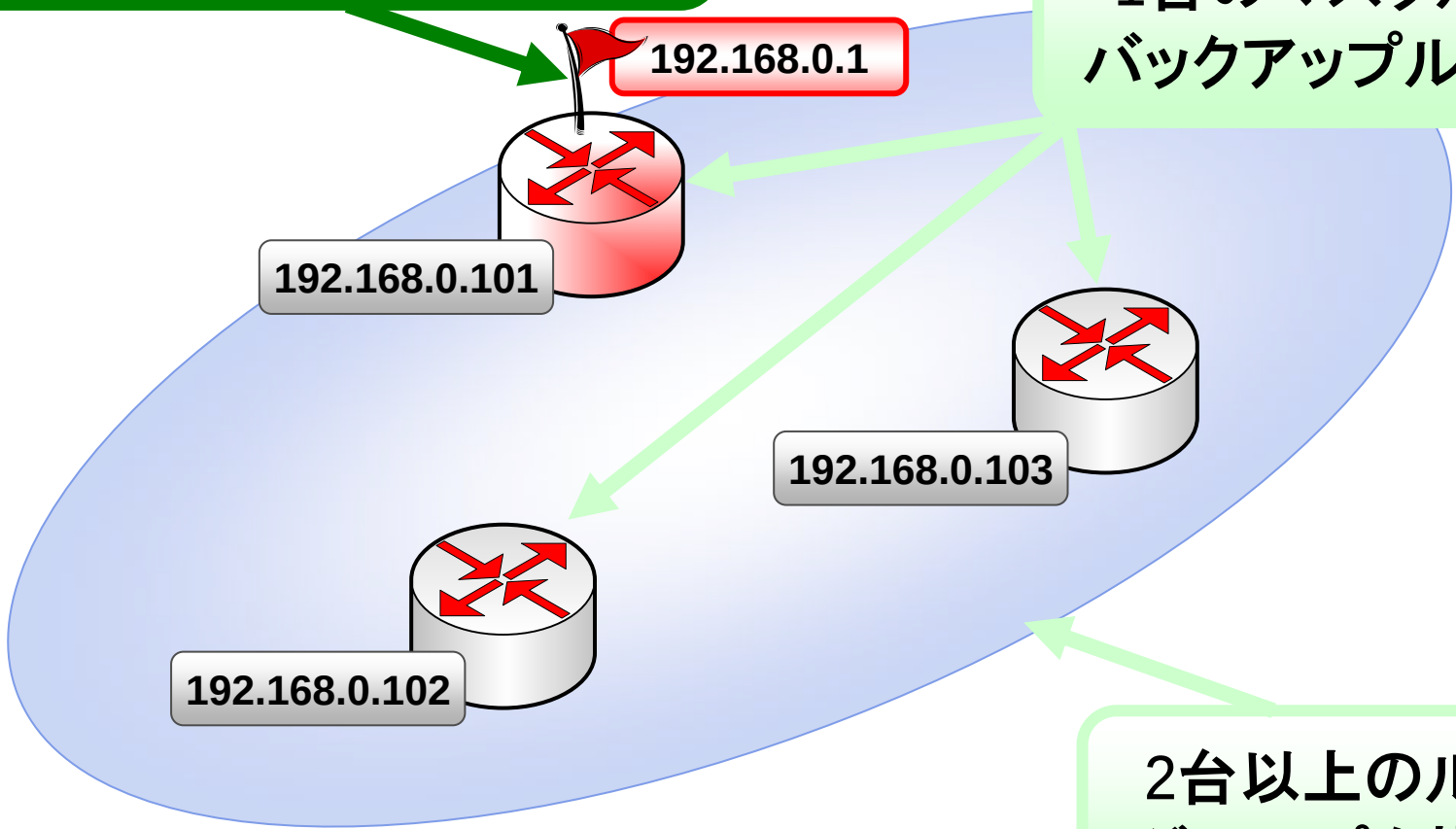
VRRPって？

KLab



VRRPって？

マスタが、浮動する仮想
ルータアドレスを保持する

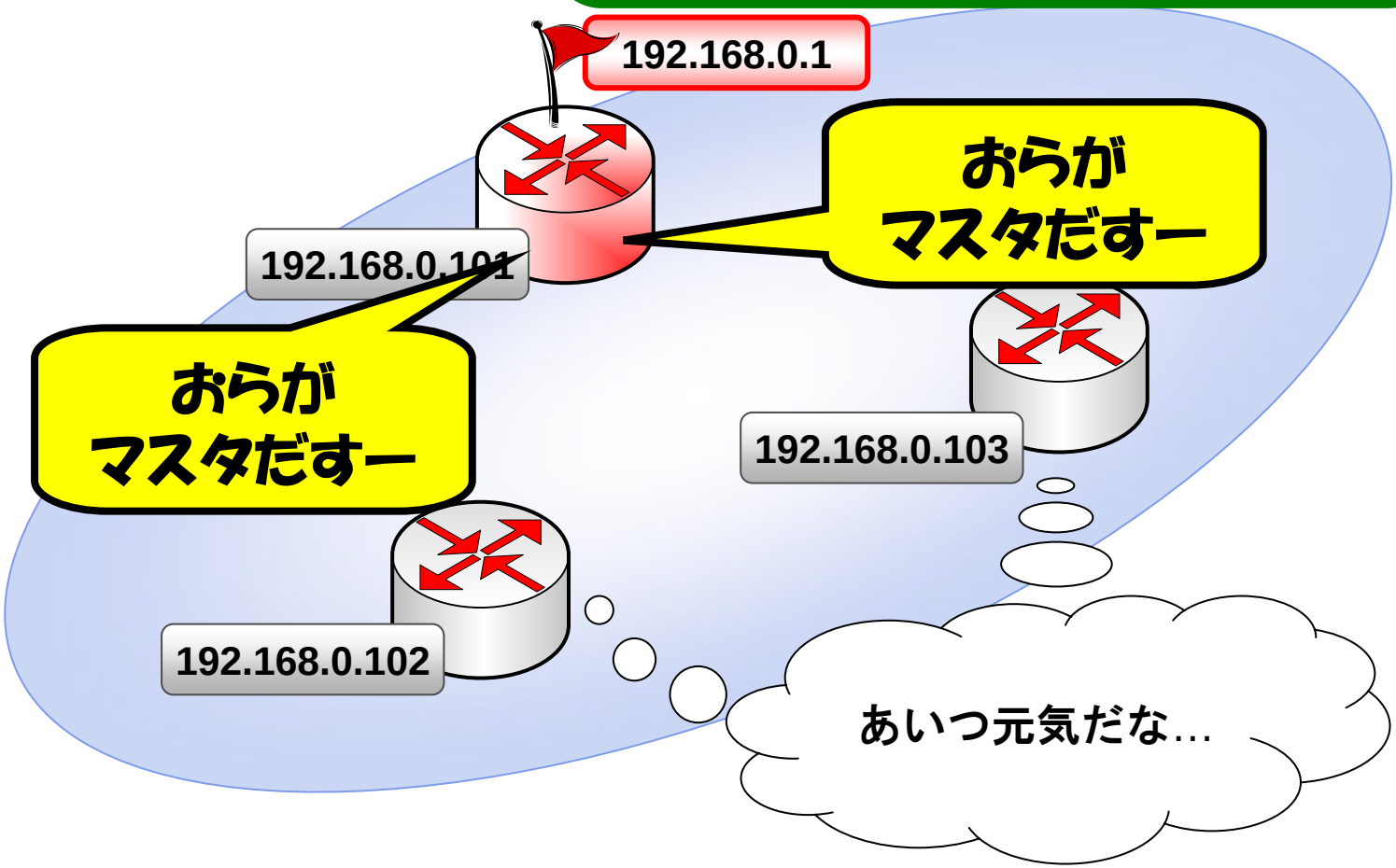


1台のマスタルータと
バックアップルータたち

2台以上のルータで
グループを構成する

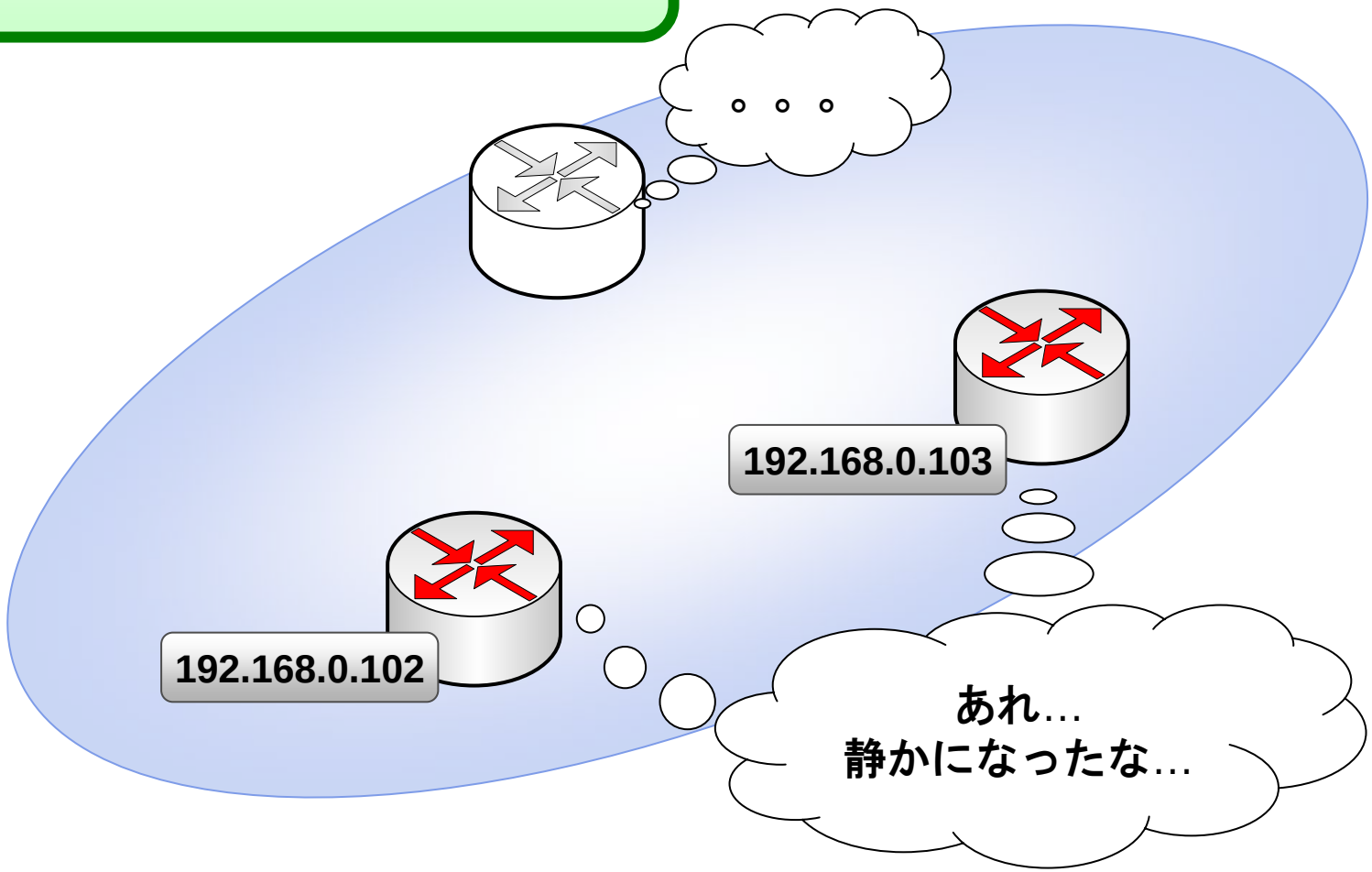
VRRPって？

マスタが定期的に
マルチキャストする
VRRP Advertisement Message



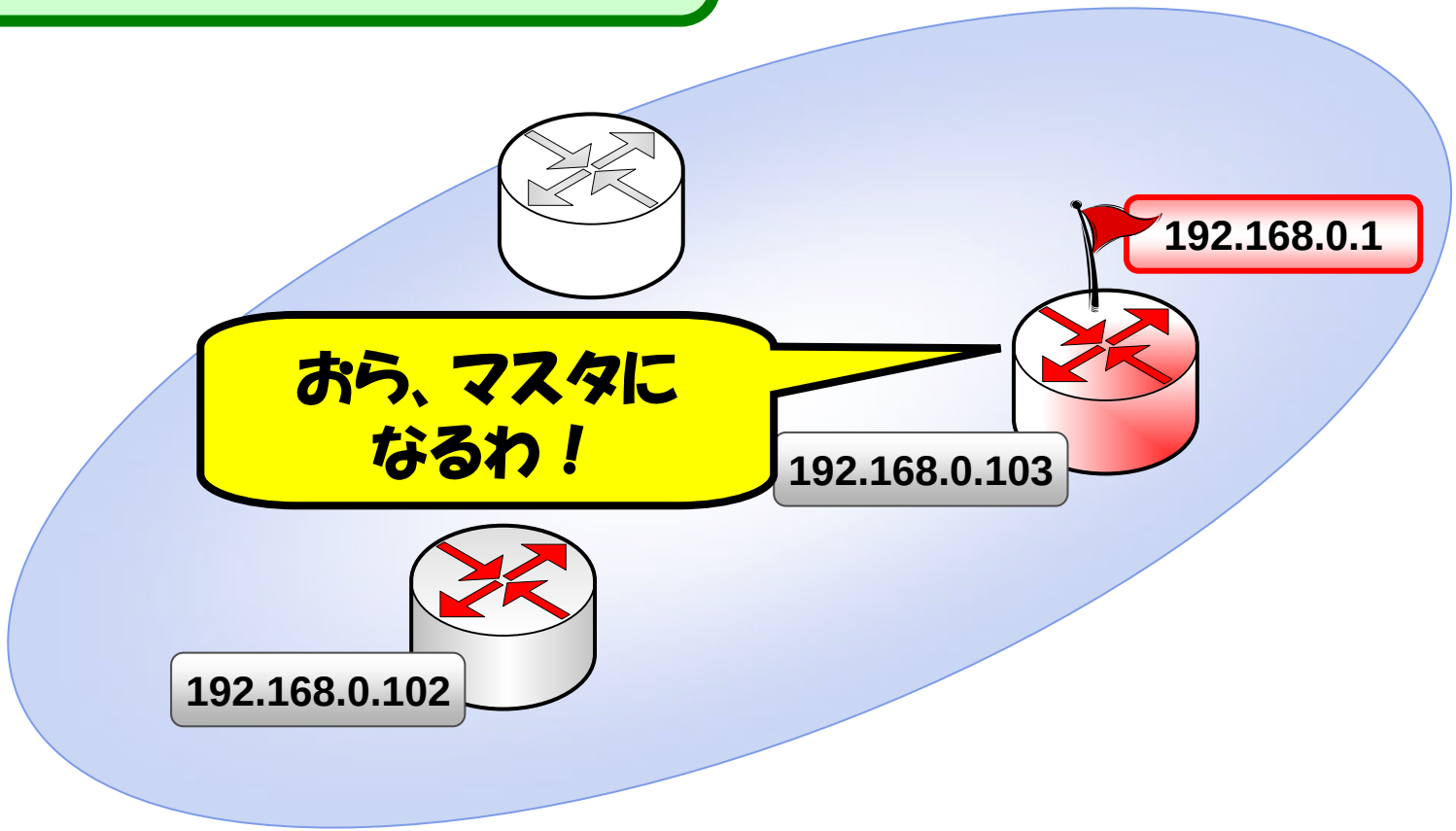
VRRPって？

マスタがダウンすると…



VRRPって？

バックアップルータのどれかがマスタに昇格する



VRRPはルータじゃないと使えないの？

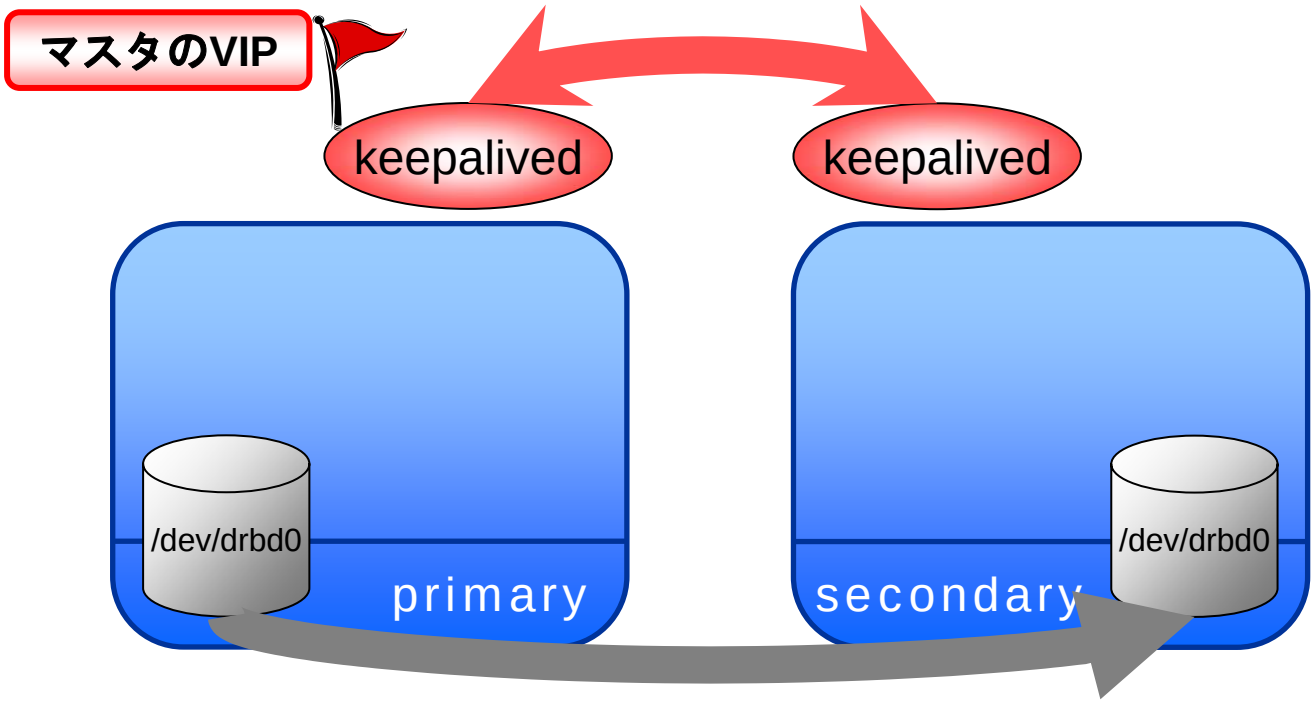
KLab

- ルータじゃなくても使えます！！
  - 複数のノードでグループを構成し、
  - その中の1つのノードがマスタになり
  - なにかしらのサービスを提供する

- VRRPを使う利点
  - マスタに付与される仮想IPアドレスの管理を任せられる
  - ネットワーク的な疎通監視を任せられる
- ＋keepalivedの利点
  - keepalivedの場合は、VRRPの状態変化をフックして任意のプログラムを実行できる
    ➔ ここでDRBDのF/Oの処理を実行すればいい

ここまでのまとめ

VRRPによるフェイルオーバ
と死活監視



DRBDによる
リアルタイムミラー

アジェンダ

KLab

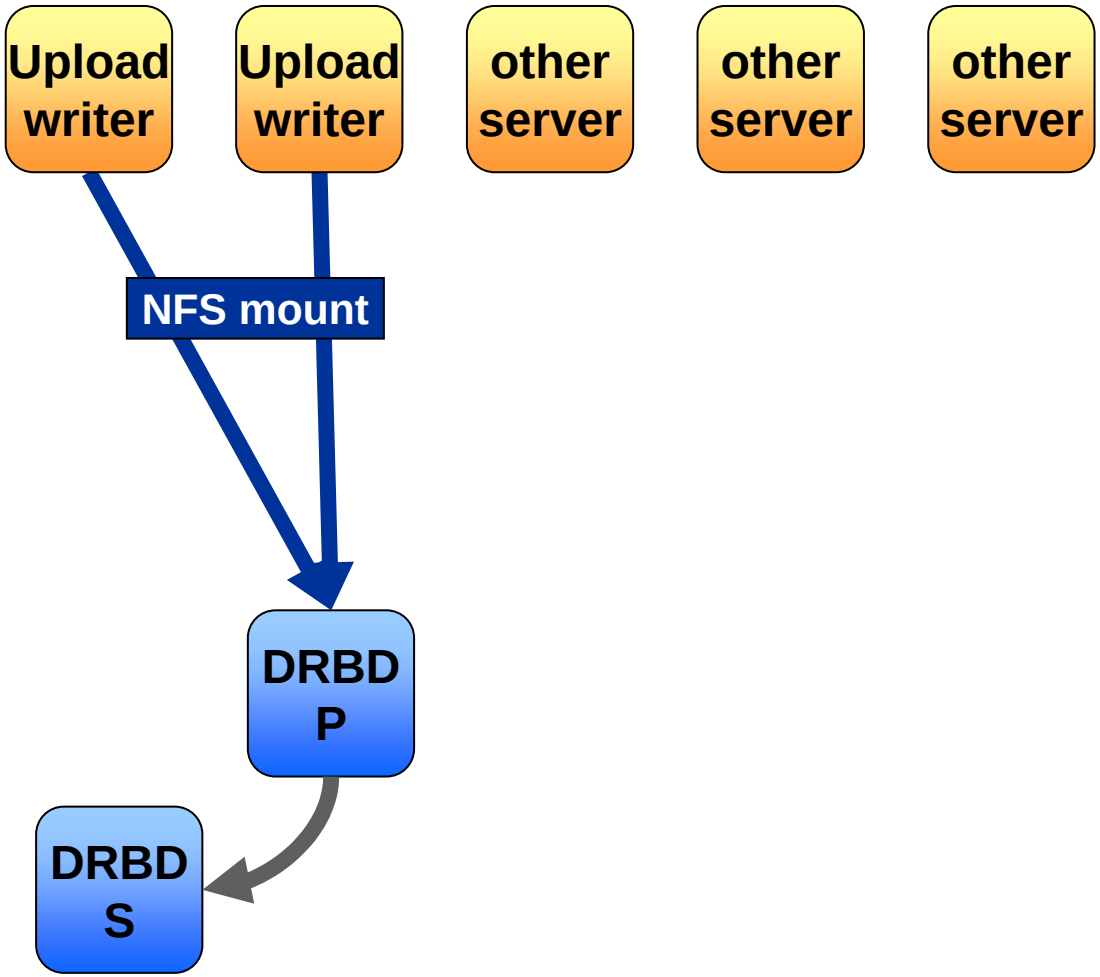
- 要件の整理
  - データの保全
  - 可用性の確保
- **実装方法の検討**
  - とまらないストレージ
  - **ストレージへのアクセス方法**
- 構築
- 落とし穴
  - DRBD
  - NFS

ストレージサーバへのアクセス – write

KLab

- NFSv3
  - 少数のアップロードサーバ(NFSクライアント)がストレージサーバ(NFSサーバ)に書き込む

ストレージサーバへのアクセス – write

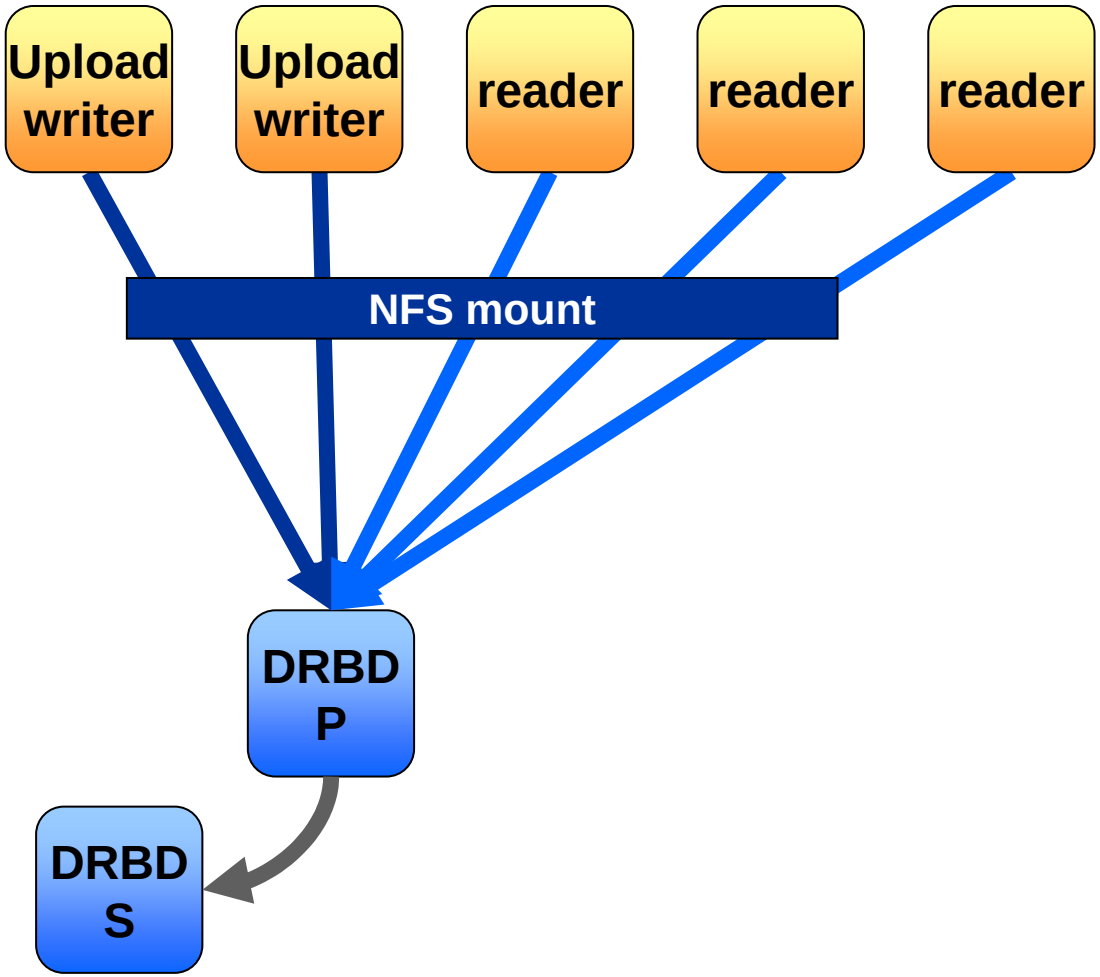


ストレージサーバへのアクセス – read (1)

KLab

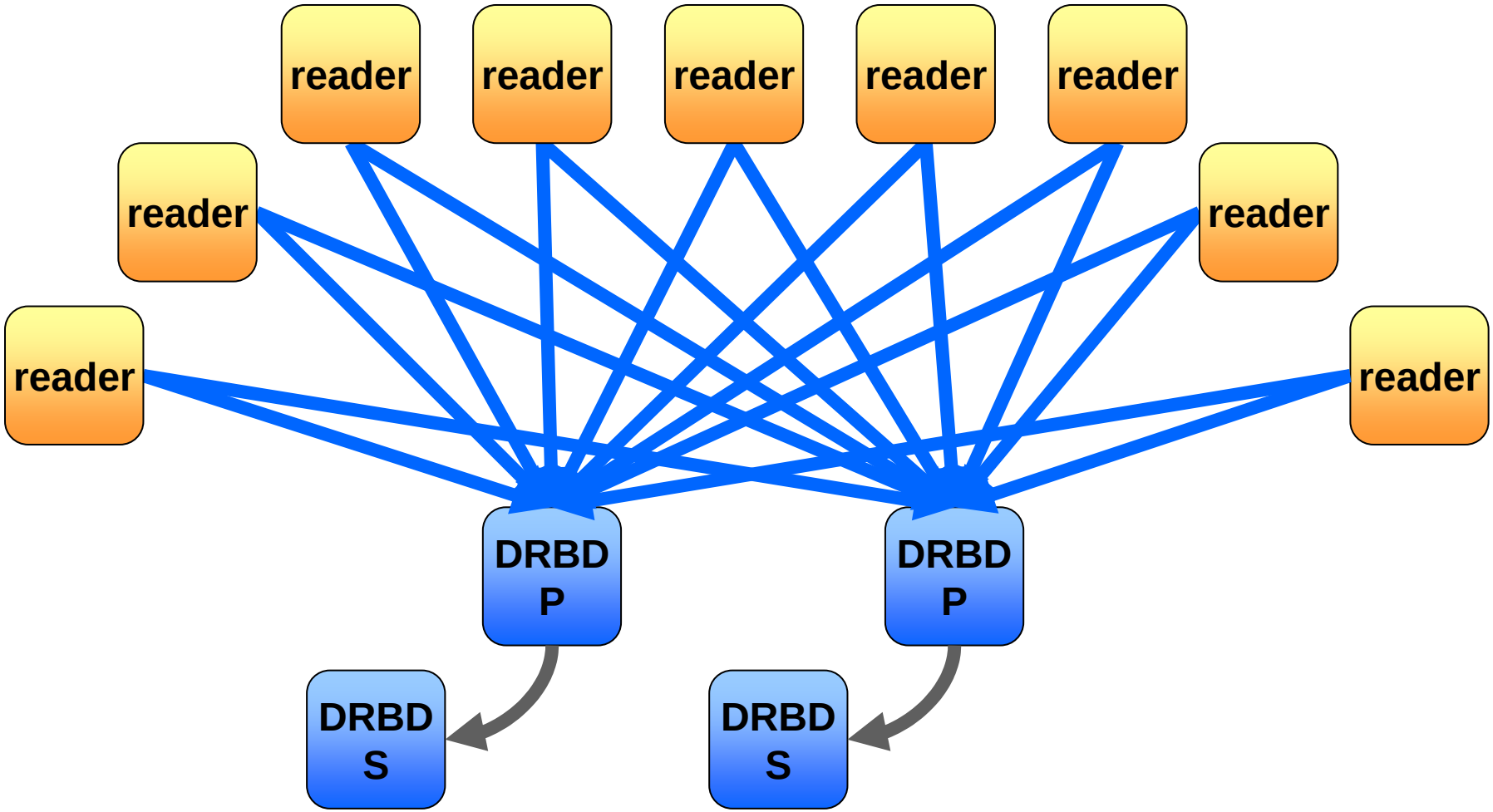
- NFSv3
  - でもいいけど、NFSはサーバ間の結合が強いのでできれば避けたいのが本音
  - Webサービスの場合、readするHTTPサーバがたくさんいるのでなおさらイヤ

ストレージサーバへのアクセス – read (1)

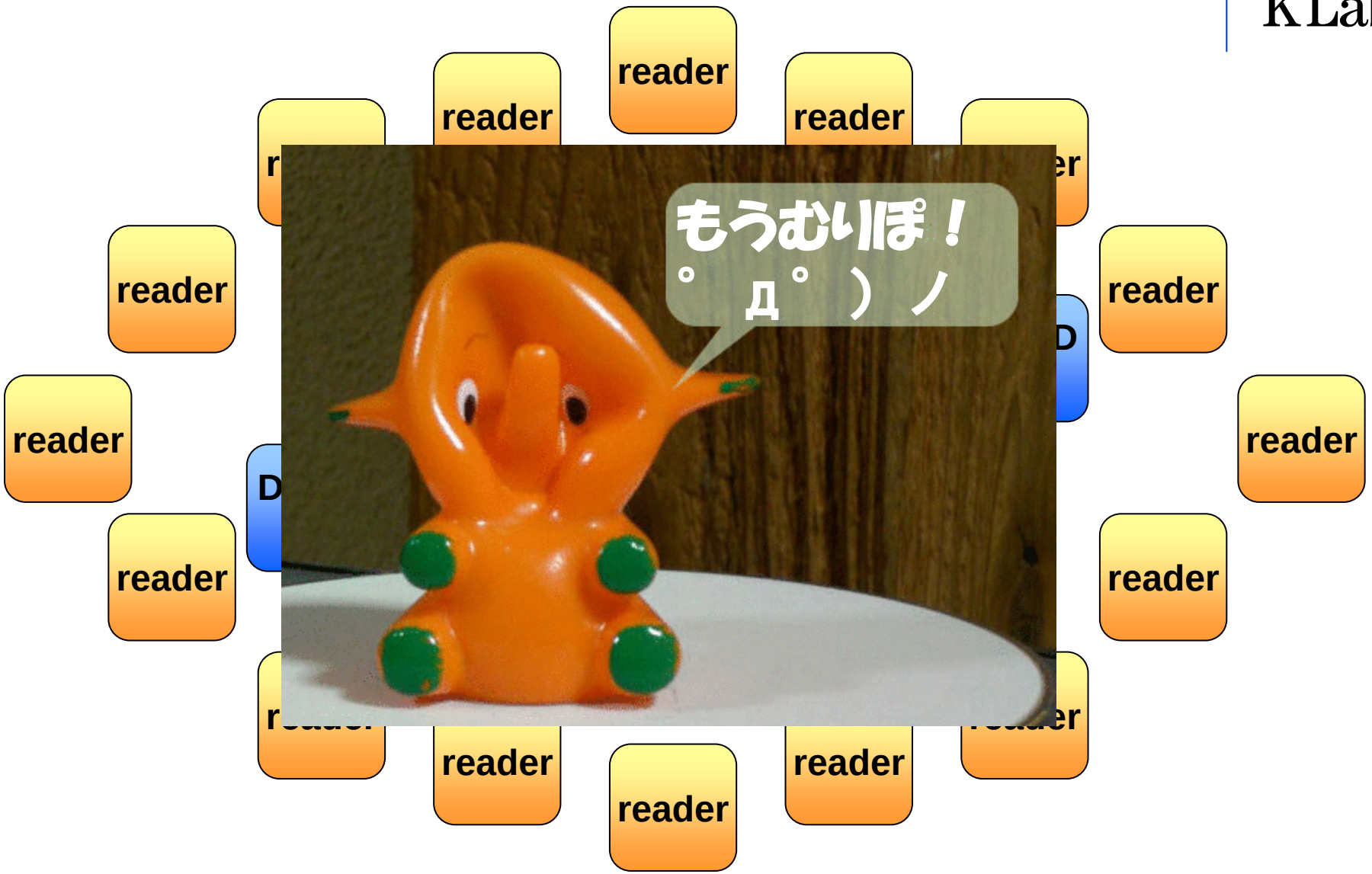


ストレージサーバへのアクセス – read (1)

KLab



ストレージサーバへのアクセス – read (1)



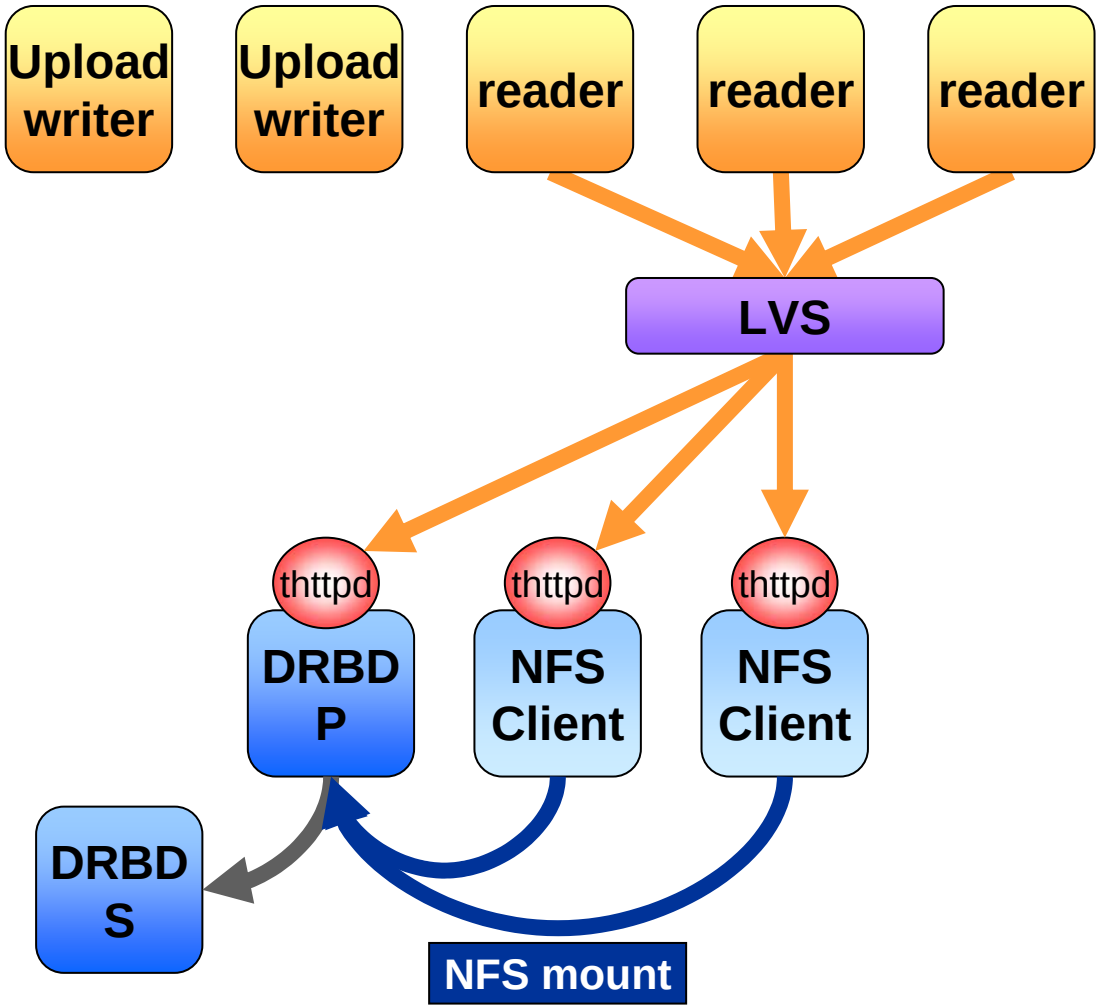
ストレージサーバへのアクセス – read (2)

KLab

- HTTP
  - readならHTTPでもいいんでないかと
  - **内部のロードバランサ経由でアクセス**
    - ☺**分散**
    - ☺**高可用性**

ストレージサーバへのアクセス – read (2)

KLab



NFSのフェイルオーバ時の注意点

KLab

- /var/lib/nfs
  - 接続中のNFSクライアントの情報を保持している場所
- F/O時にこの情報が失われてしまう
  - 昇格したNFSサーバは見知らぬNFSクライアントからのアクセスにみえるため拒否してしまう！
- 解決方法
  - (1) DRBDで/var/lib/nfsもミラーリングしちゃう
  - (2) mount –t nfsd nfsd /proc/fs/nfs ←オススメ
    - kernel 2.6以降

アジェンダ

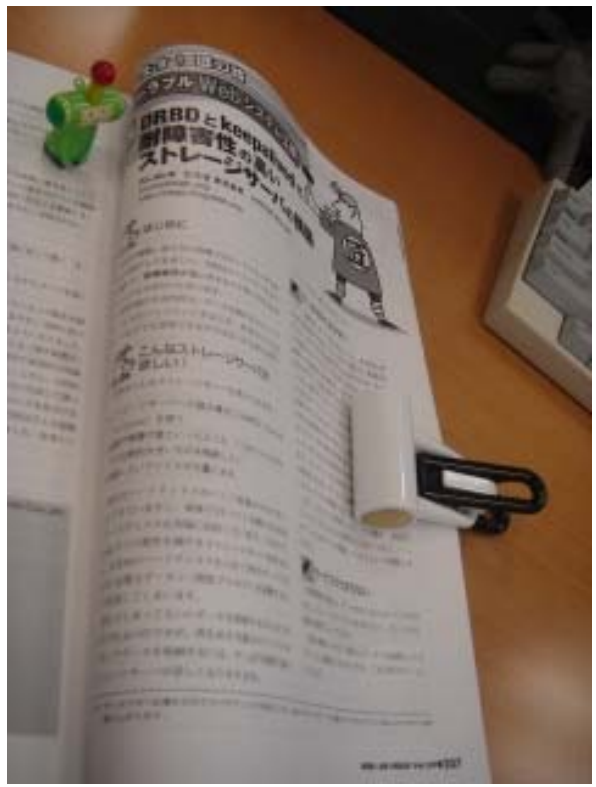
KLab

- 要件の整理
  - データの保全
  - 可用性の確保
- 実装方法の検討
  - とまらないストレージ
  - ストレージへのアクセス方法
- **構築**
- 落とし穴
  - DRBD
  - NFS

で、実際の構築方法の説明は。。。

KLab

- 6/2x発売のWEB+DB PRESS Vol.39を見てください＞＜



それだけではなんなので。。。

KLab

- 過去にハマった落とし穴情報をいくつか…

アジェンダ

KLab

- 要件の整理
  - データの保全
  - 可用性の確保
- 実装方法の検討
  - とまらないストレージ
  - ストレージへのアクセス方法
- 構築
- 落とし穴
  - DRBD
  - NFS

落とし穴（1）
– DRBDとXFS

KLab

- ちょっと前（kernel 2.6.11か2.6.12ぐらい）のお話
- DRBDをXFSで使うと、ある日突然サーバが何もいわずに落ちる問題が発生
  - 再現性あり：bonnie++で負荷をかけ続けると3日目ぐらいで落ちる
- kernelのスタックサイズを4KBから8KBに変更すればOK
  - Kernel hacking → Use 4Kb for kernel stacks instead of 8Kbのチェックを外す
- 最近のバージョンでは直ってる情報もアリ

落とし穴（2）
– drbd.confのon-io-error

KLab

- on-io-error ：ディスクでIOエラーが発生した場合の挙動の指定
  - pass_on
    - 上位レイヤ（ファイルシステム）にスロー
  - panic
    - 当該ノードでkernel panicを起こし停止させる
  - detach ←こいつがクセモノ
    - 物理デバイスを切り離す
    - と同時に相方サーバをkernel panicさせる

    ➔プライマリもセカンダリもダウン＞＜

落とし穴（3）
– メモリ系のデバイスとDRBD

KLab

- tmpfs
  - ✖ デバイスに見えないのでDRBDで使えない
- tmpfs＋loop back device
  - ✖ フリーズする
- ramdisk
  - 😐 使えるが、1つのramdiskの最大サイズは512MB
- ramdisk＋LVM
  - ✖ pvcreateできなかった
- ramdisk＋md(raid0)
  - ✖ 領域を超えるサイズのファイルを作るとOSが落ちる

✪ 現在の最新バージョンでは異なる可能性があります

KLab

今日はここまで

KLab

ご清聴、
ありがとうございました～
＞＜