http://people.allbsd.org/~hrs/FreeBSD/sato-FBSDW20141226-1.pdf

第3回FreeBSDワークショップ

(18:30から)

佐藤　広生 <hrs@FreeBSD.org>

東京工業大学/ FreeBSD Project

2014/12/26

開催背景

- **日本国内の＊BSD活動は2000年以降、縮小の一途です**
  - 少なくともユーザ数は大幅に減った
  - 海外では明るい話題がそれなりにあるのに...

- 盛り上げたいのはやまやまですが、何をするのが良いですか？

本ワークショップの進行

- 18:30〜19:30　自己紹介＋話題にしたいトピックの提示
- 19:30〜20:00　提示トピック
- 20:00〜20:15　休憩
- 20:15〜21:30　FreeBSD 10系の話

意見は自由に発言ください！

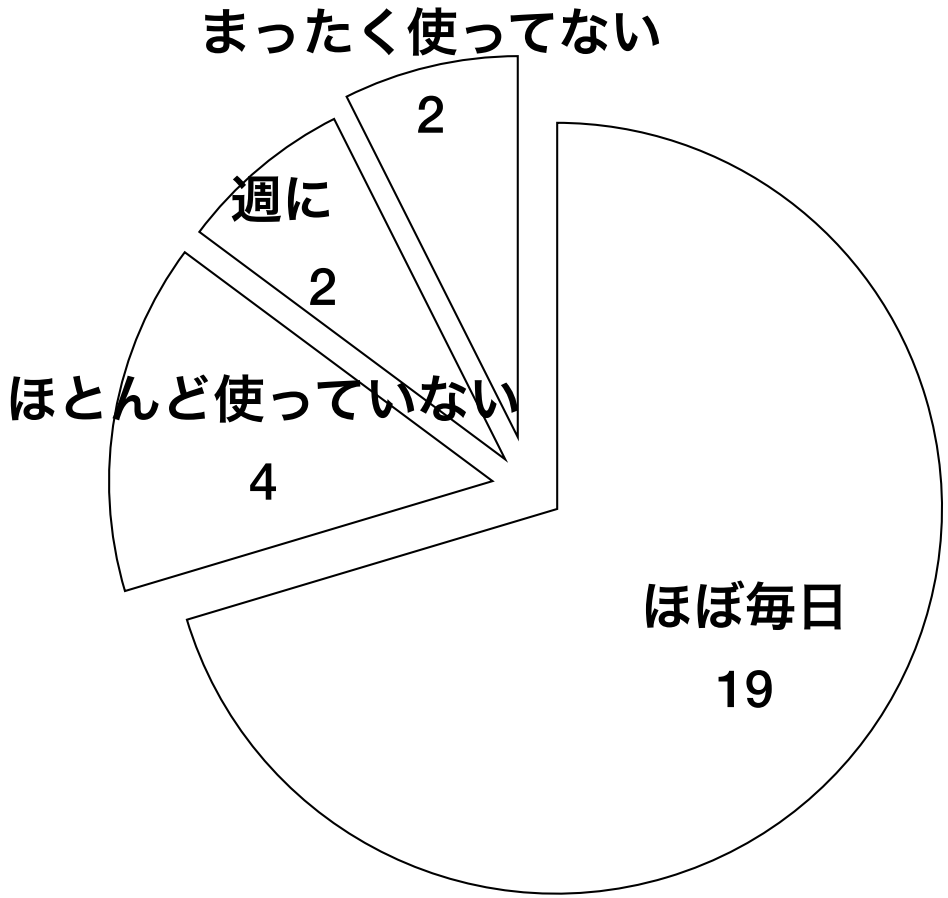
オーガナイザの自己紹介

- 名前：佐藤 広生
  - FreeBSD コアチームメンバ、リリースエンジニア(2006-)
  - FreeBSD Foundation 理事(2008-)
  - その他の＊BSD／オープンソース関連の活動いろいろ
  - 東京工業大学助教(2009-)

自己紹介タイム

- 名前（所属）
- 開発者 or 利用者
- 興味がある／話題にしたい内容

をどうぞ



今回の出席者内訳：新規16名、再参加者11名

本ワークショップの進行

- 18:30〜19:30　自己紹介＋話題にしたいトピックの提示
- 19:30〜20:00　提示トピック
- 20:00〜20:15　休憩
- 20:15〜21:30　FreeBSD 10系の話

意見は自由に発言ください！

FreeBSD10 系の変更点

- **カーネル**
  - スケーラビリティに関する改良多数

    例：unmapped I/O, イベントタイマ、direct dispatch GEOM、pf
- **ユーザランド**
  - GCCを捨ててClang + libc++ へ
  - pkg_tools(8)を捨ててpkg(8)へ
  - GPL排除：patch(1), sort(1), yacc(1), lex(1)
  - BIND削除
  - 再帰リゾルバとしてunboundを一時的に導入
  - iconv導入(Citrus iconv)
  - bhyve (BSD hyper visor)
  - Capsicum sandboxの適用

Unmmaped I/O

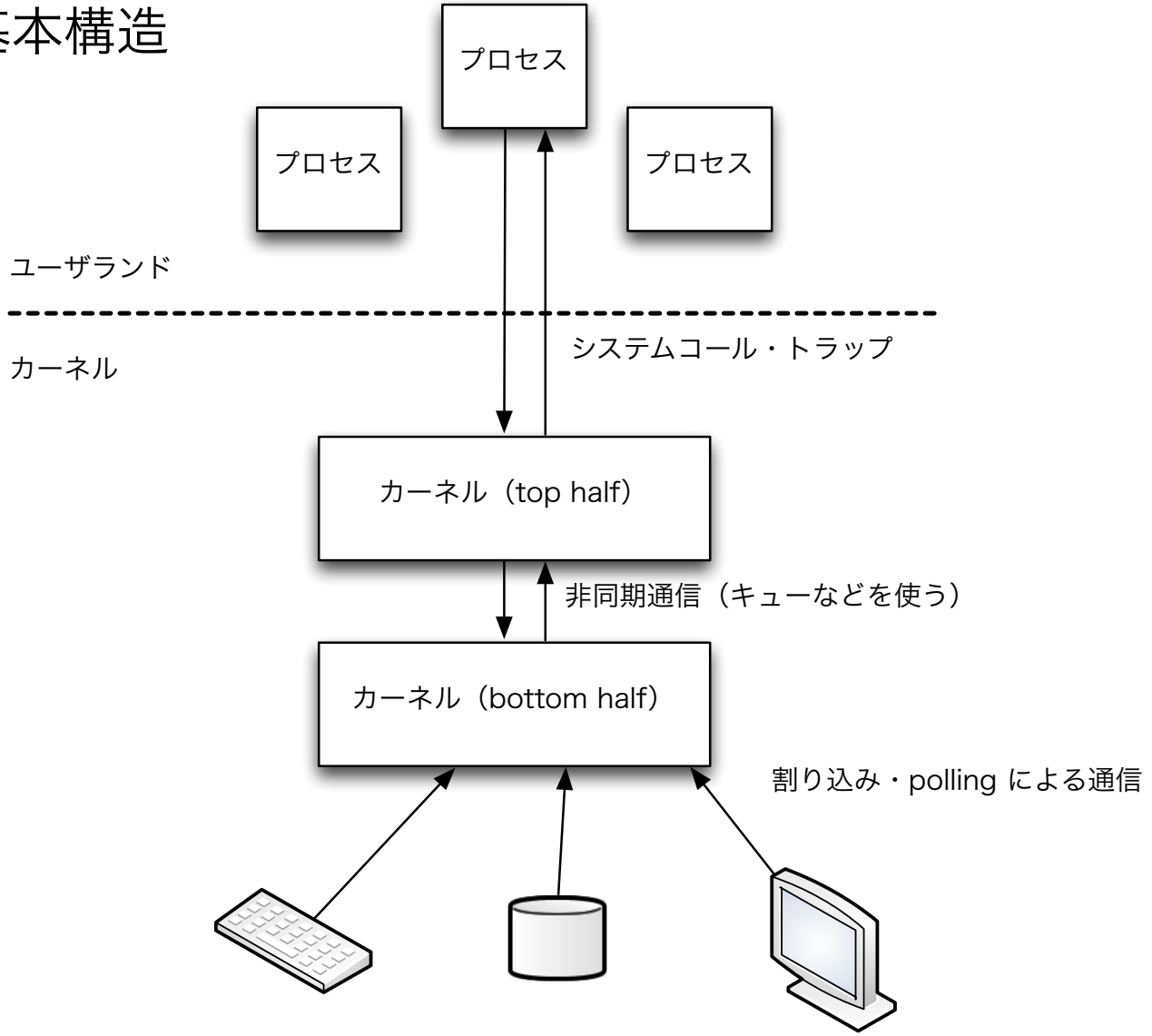
- マルチプロセッサ環境におけるストレージI/Oの性能低下対策

ざっくりしたまとめ

- FreeBSDのVMは unified buffer cacheに なっている
- ストレージからの読み出しは、常にバッファページマップを伴う
- バッファページマップは、TLBのフラッシュが必要
- SMP環境では、このマップの度にIPIが発生して並列処理性能が低くなる

Unmmaped I/O

- BSDの基本構造



Unmmaped I/O

- BSDのメモリ管理

直接メモリにアクセスするのではなく、
アドレス変換を行ってアクセスする

CPUから見えるアドレス
（例：32bitなら4GB）

搭載メモリのアドレス
（例：2GBあれば2GB）

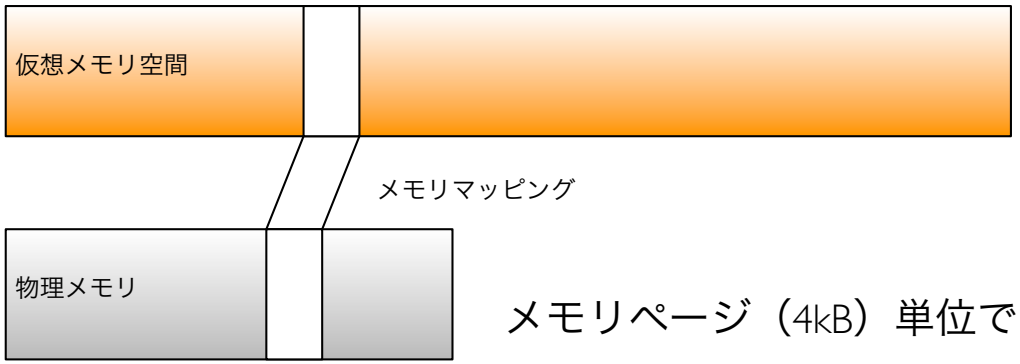
Unmmaped I/O

- BSDのメモリ管理

直接メモリにアクセスするのではなく、
アドレス変換を行ってアクセスする

CPUから見えるアドレス
（例：32bitなら4GB）

搭載メモリのアドレス
（例：2GBあれば2GB）

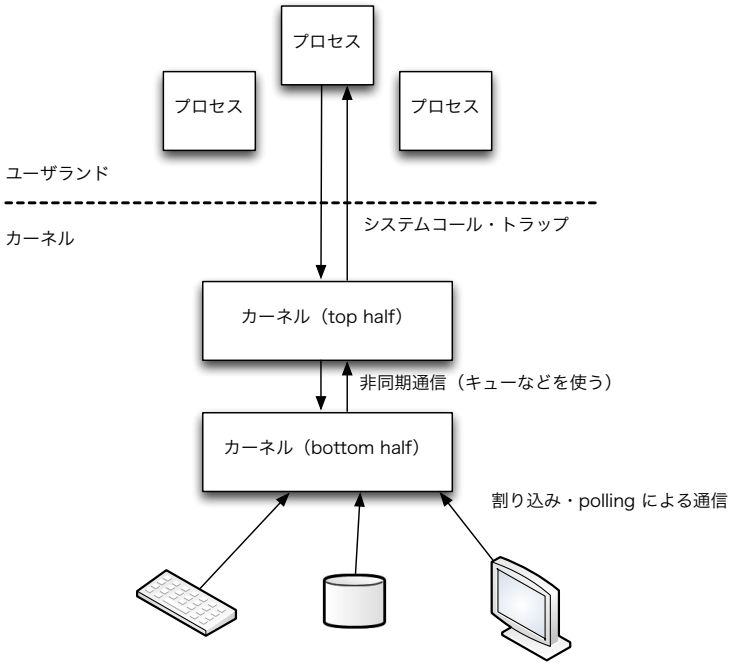


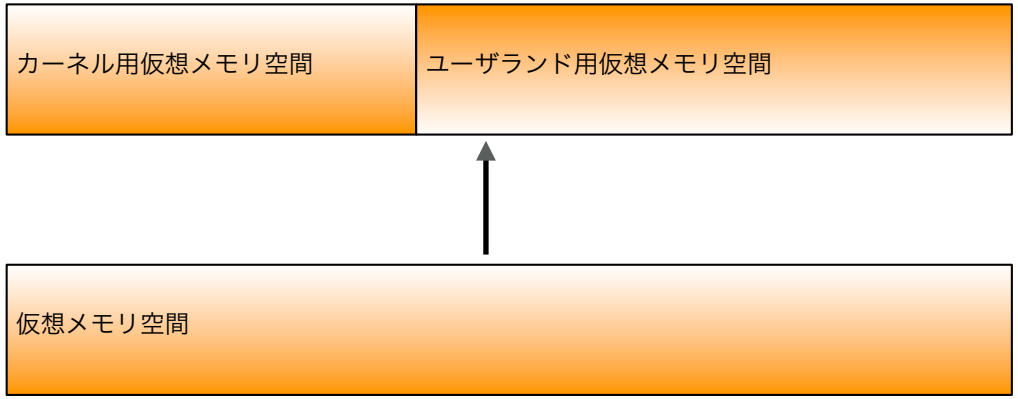
メモリページ（4kB）単位で
対応関係をつくることができる

Unmmaped I/O

- BSDのメモリ管理

仮想メモリ空間は、
カーネル用とユーザランド用に分けてある

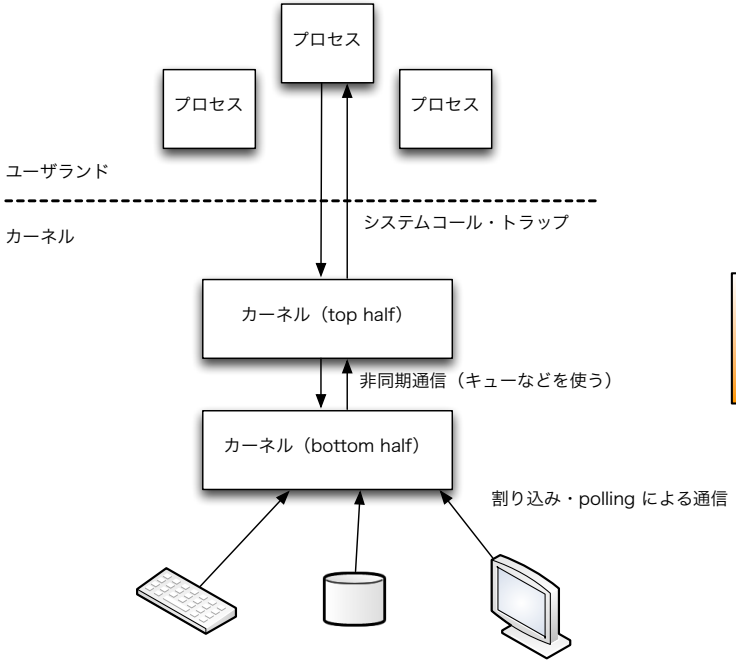


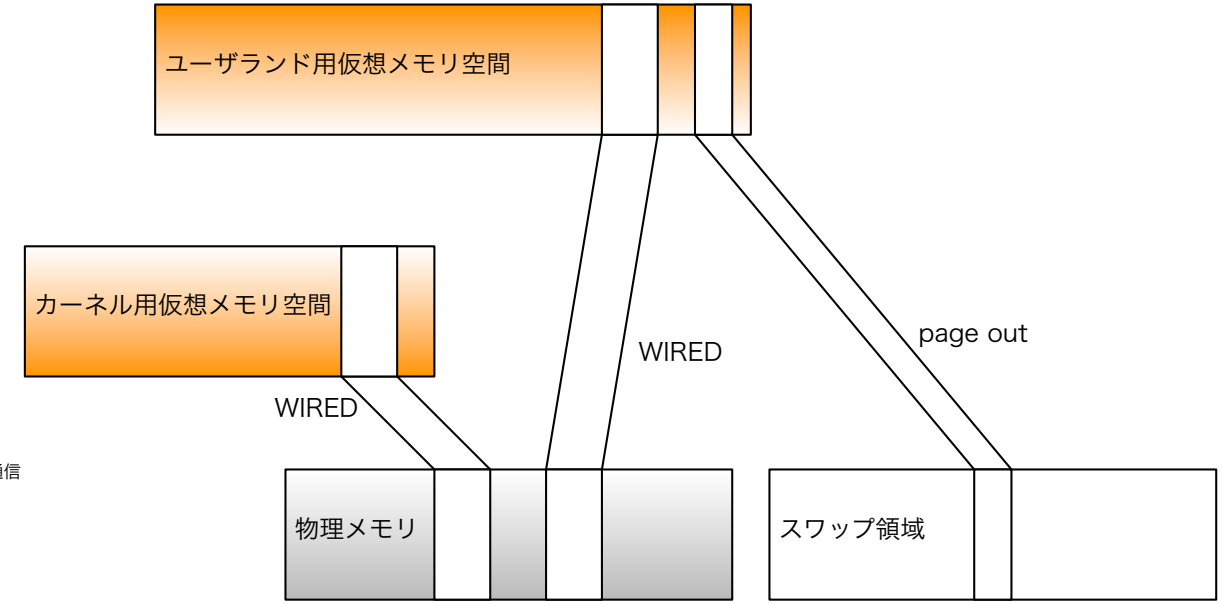


Unmmaped I/O

- BSDのメモリ管理

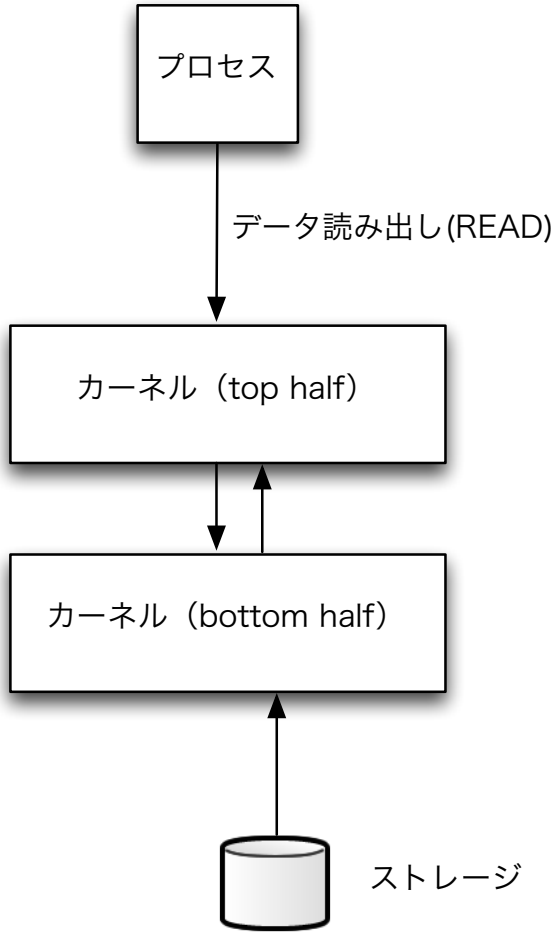
・必要な分だけ物理メモリにマップして使う
・ユーザランド用のメモリは足りなければスワップも使う





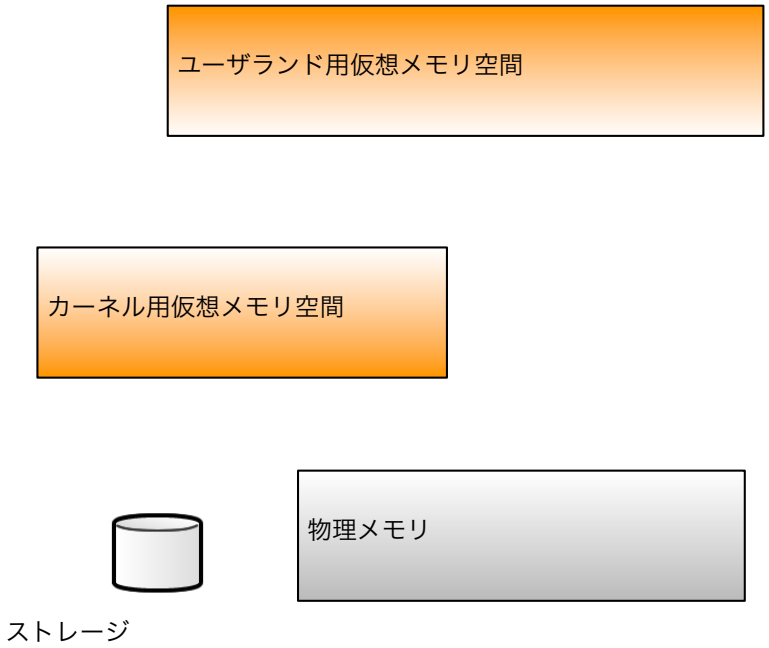
Unmmaped I/O

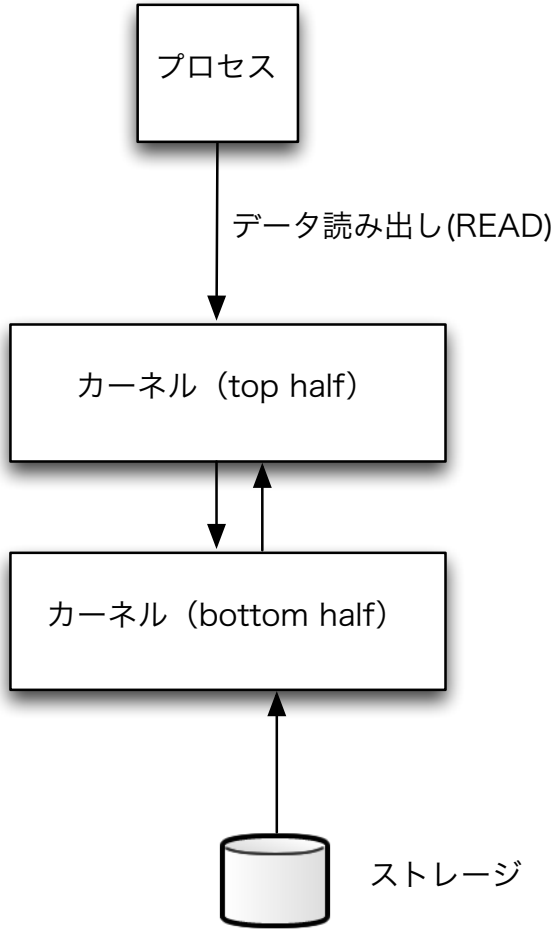
- プロセスがストレージにアクセスする時のメモリ管理



Unmmaped I/O

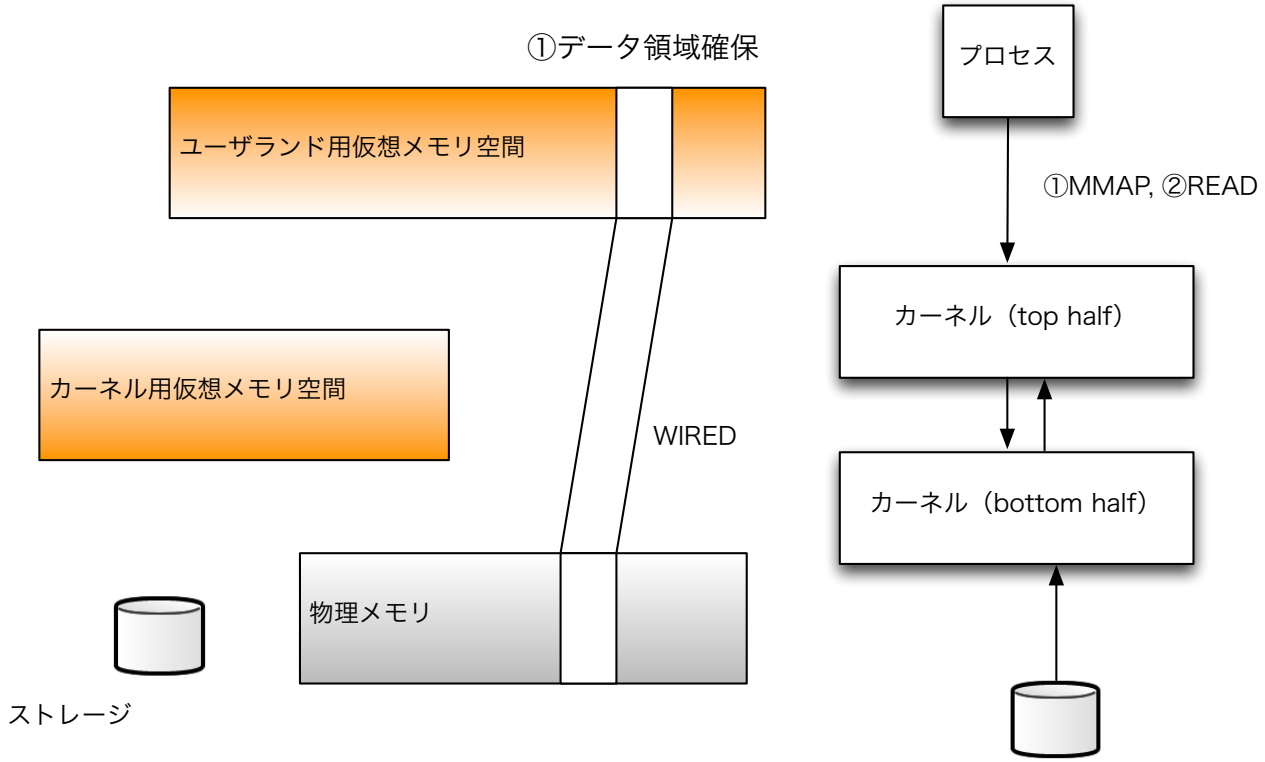
- プロセスがストレージにアクセスする時のメモリ管理





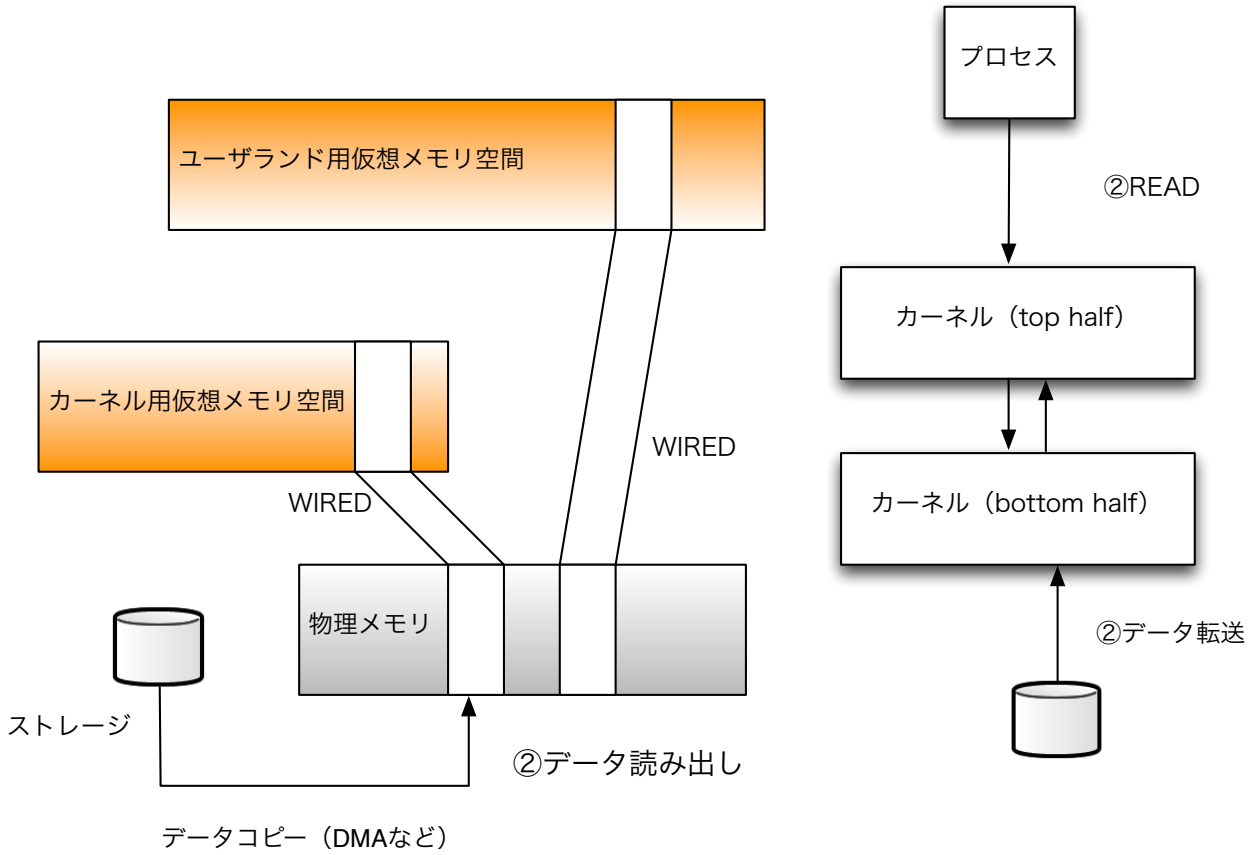
Unmmaped I/O

- プロセスがストレージにアクセスする時のメモリ管理



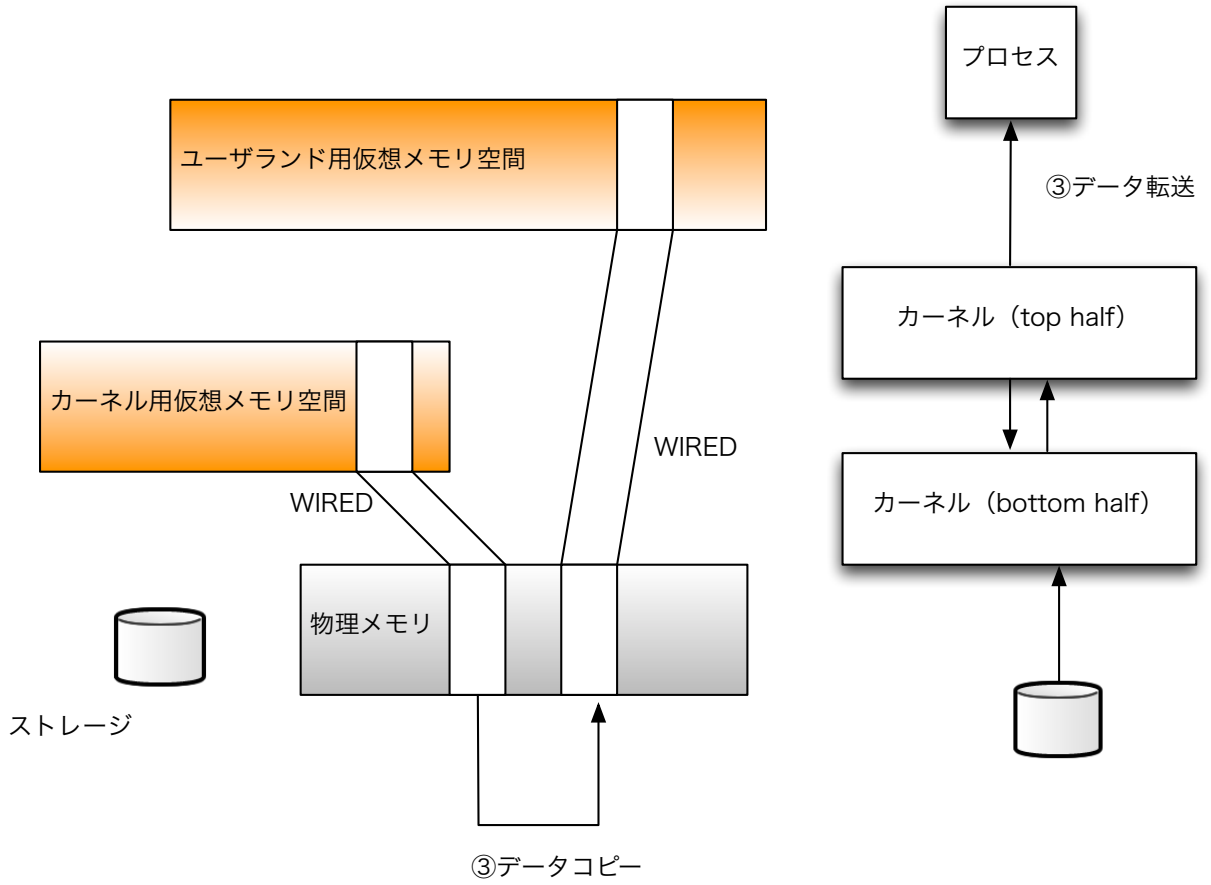
Unmmaped I/O

- プロセスがストレージにアクセスする時のメモリ管理



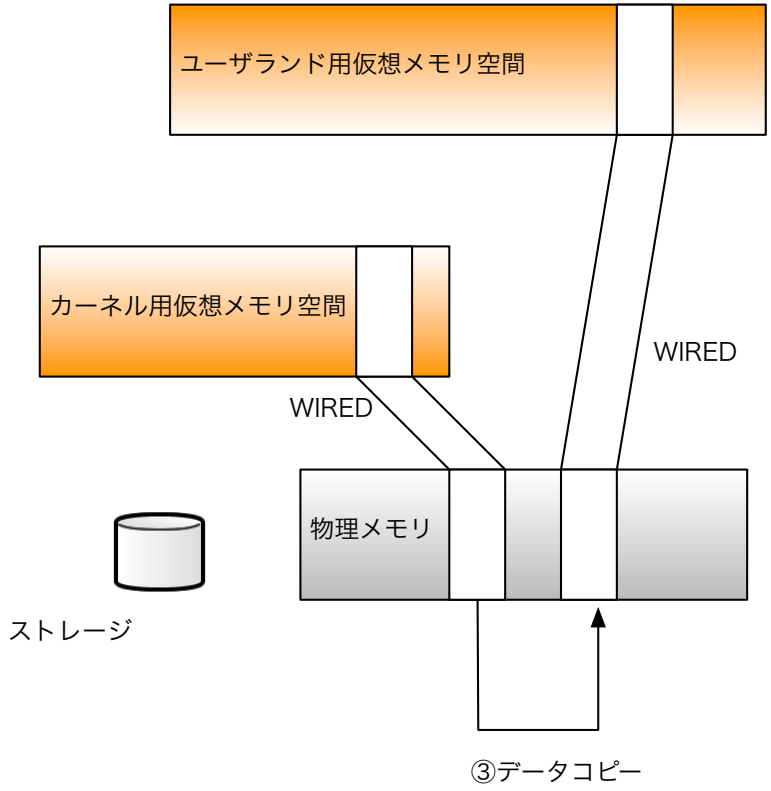
Unmmaped I/O

- プロセスがストレージにアクセスする時のメモリ管理



Unmmaped I/O

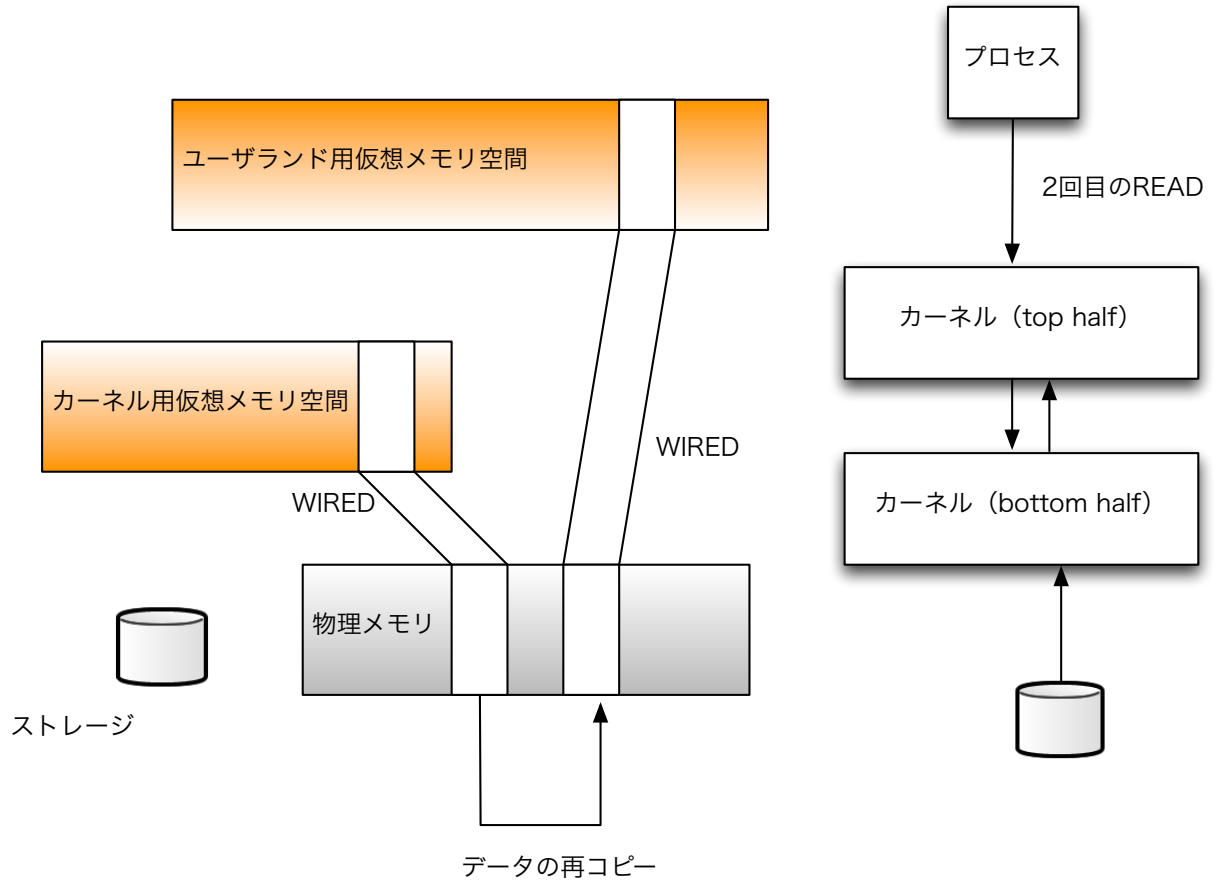
- プロセスがストレージにアクセスする時のメモリ管理



疑問点：なぜカーネルのメモリ空間に一度コピーしているのか？

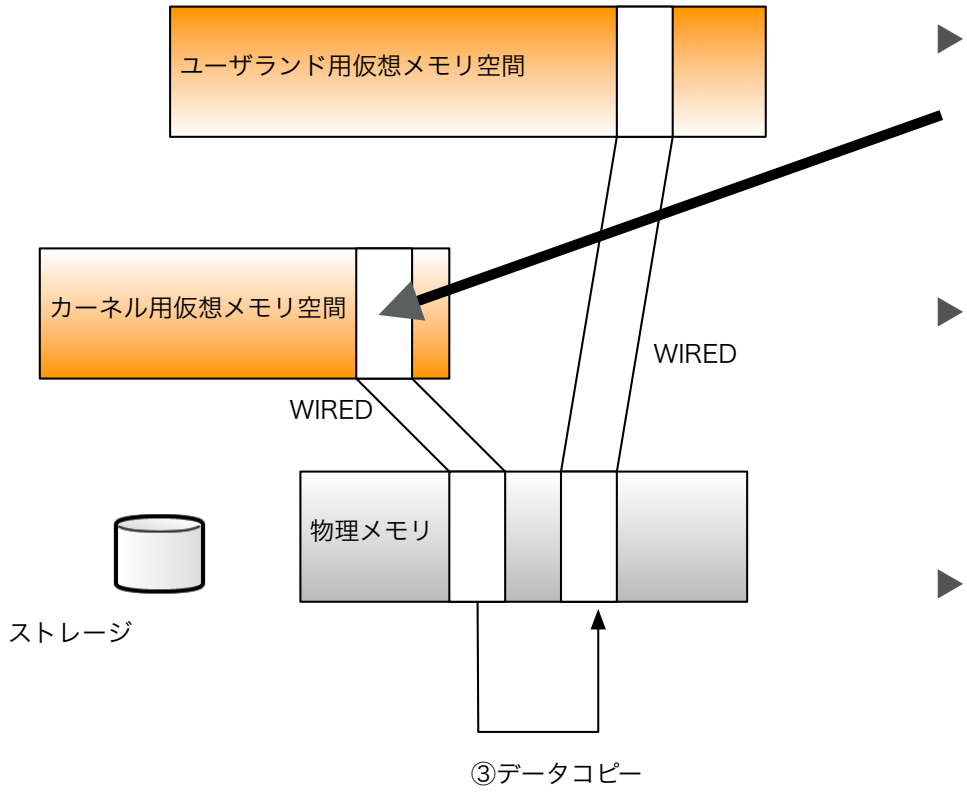
Unmmaped I/O

- プロセスがストレージにアクセスする時のメモリ管理



Unmmaped I/O

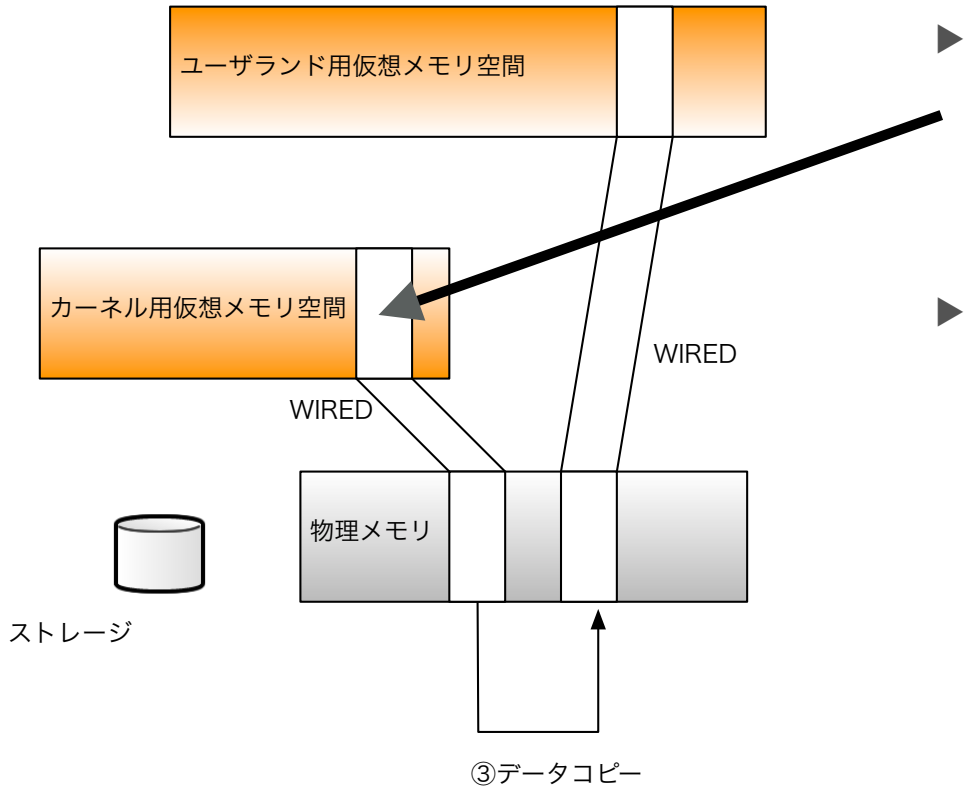
- プロセスがストレージにアクセスする時のメモリ管理



- **バッファマップ:**
  カーネル空間に確保されるページ
- 空きメモリに余裕があるうちは、
  データは破棄されない
- どのストレージのデータなのかを
  覚えている
- ディスクキャッシュとして機能

Unmmaped I/O

- プロセスがストレージにアクセスする時のメモリ管理



- **バッファマップ**：
  カーネル空間に確保されるページ

- **問題点**
  ・確保するにはマッピングを変えないといけない

  ・TLBというMMUのキャッシュを破棄しなければならない

  ・全CPUが一瞬停止する

Unmmaped I/O

- **結局何が変わったのか**
  - バッファマップを使わないモードを追加した
    ・カーネルのバッファ領域をあらかじめ確保
    ・カーネルでデータを変更しない転送に、その領域を使う
    （例えばUFSへのアクセス等）
    ・キャッシュ機能は変わらない
    （個別のマップではなく、まとまったバッファになっただけ）
  - 内容を変更するアクセスは、従来どおりマップする
  - I/O request を出す側でもどっちを使うか選択できる

Unmmaped I/O

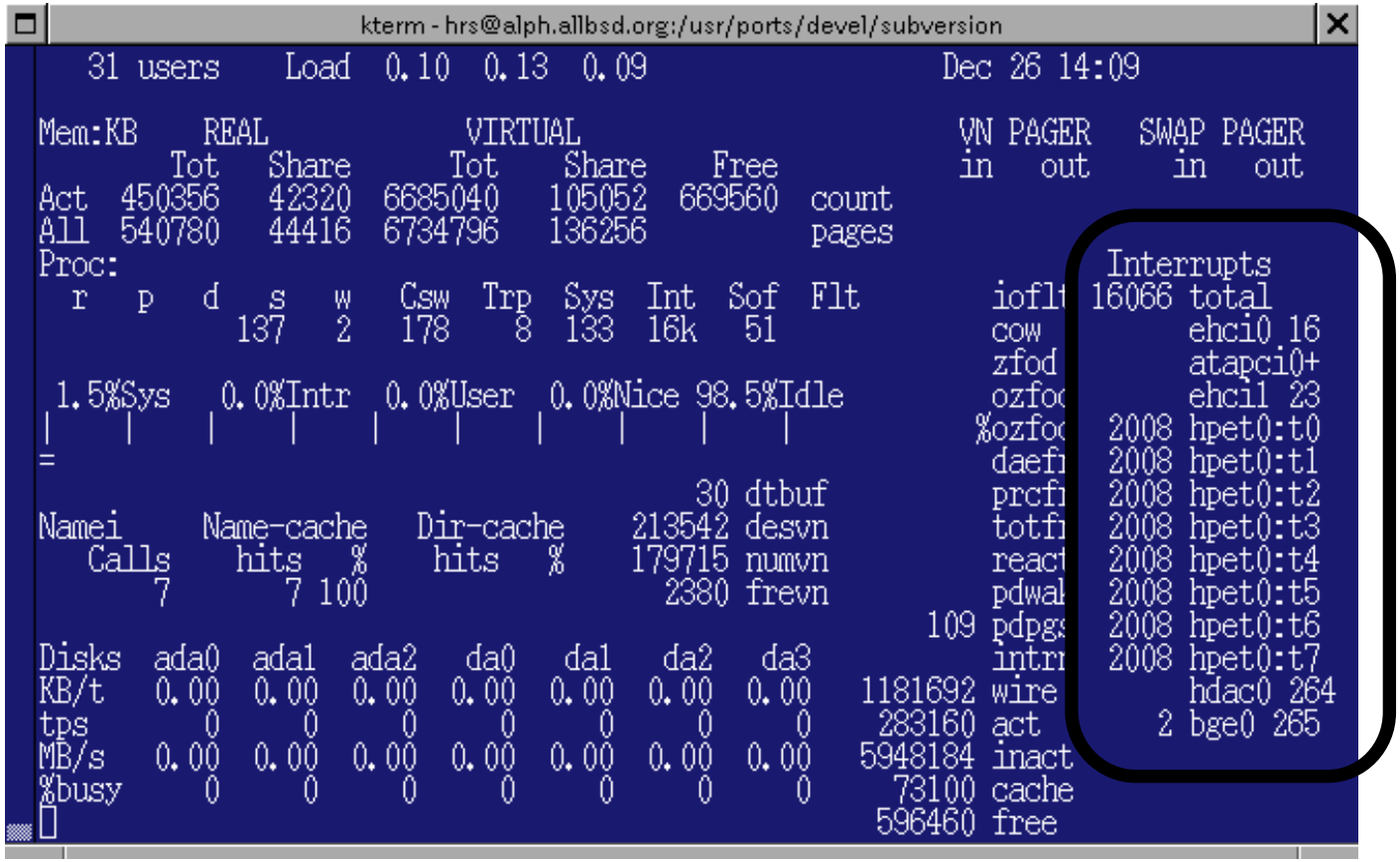
- **結局何が変わったのか**
  - 得られる成果：CPU負荷の減少
    - 1 CPU では特に変わらず
    - SMPで system の %CPU が減るはず
      （30%くらい減ったというマイクロベンチマークがある）
  - UFSが高速になったと紹介している記事とかありますが
    嘘八百です。

eventtimer

- 問題：CPU tick のタイマ割り込みの負荷が大きい
- 解決策：タイマ割り込みをone-shot modeに変えた。
  （＋タイマソース管理のフレームワークを刷新した）
- これもsystemの%CPUがだいぶ下がります。
  （9の途中あたりからは入ってます）

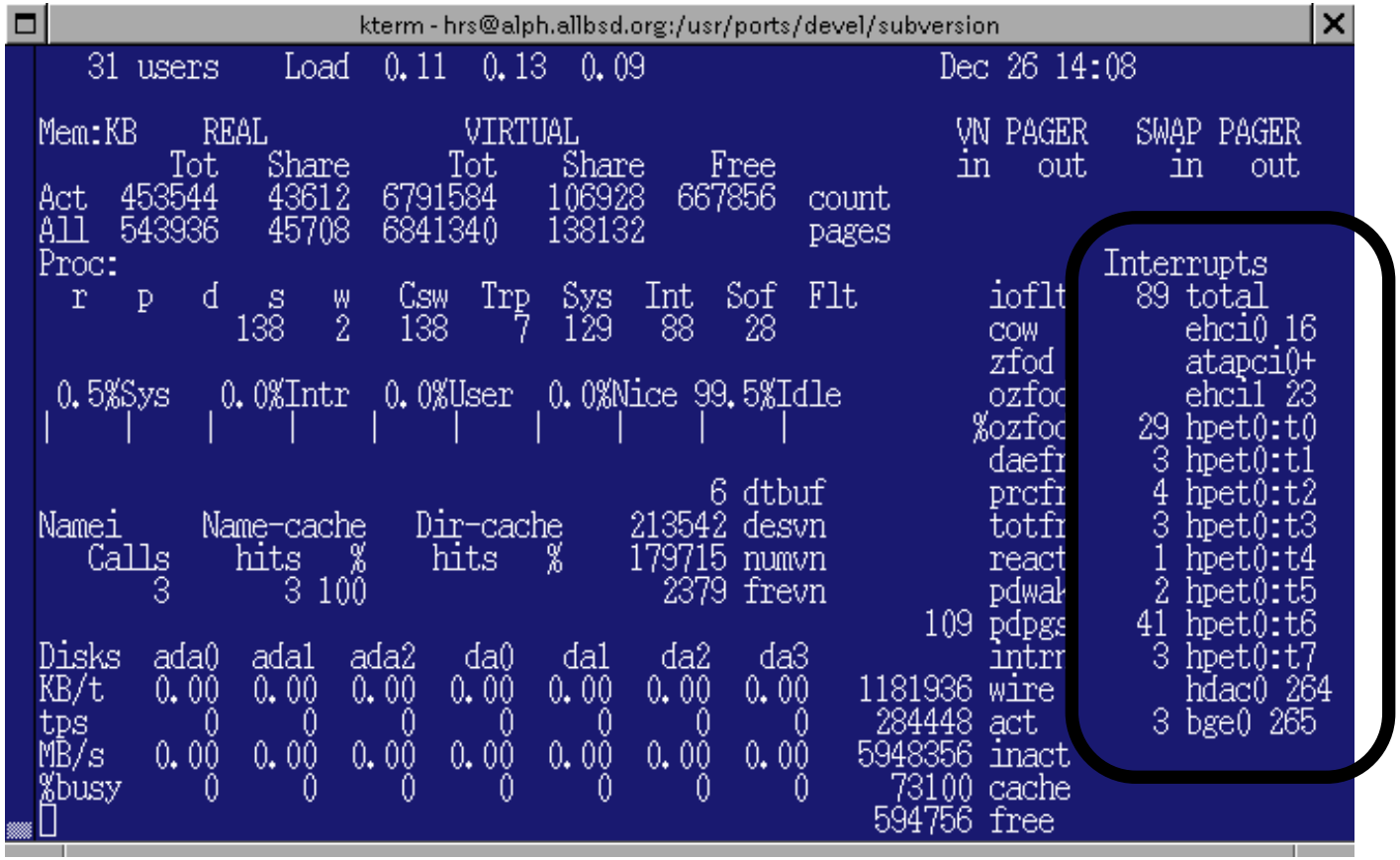
eventtimer

- % systat -vm 1



eventtimer

- % systat -vm 1



- kern.eventtimer.periodic を 1 にすると戻る

direct dispatch GEOM

- **GEOMとは：**

  ・伝統的なUNIXにおけるstruct bufを階層化したもの

  ・カーネルのディスクI/Oとデバイスドライバの間に
  スタッカブルなI/O処理用の層を挟むことができる

- **問題点**：

  ・GEOMのI/O処理は、基本的にキューで逐次処理

  ・並列処理が効率良くできない（IOPSに影響）

direct dispatch GEOM

- **対策：**

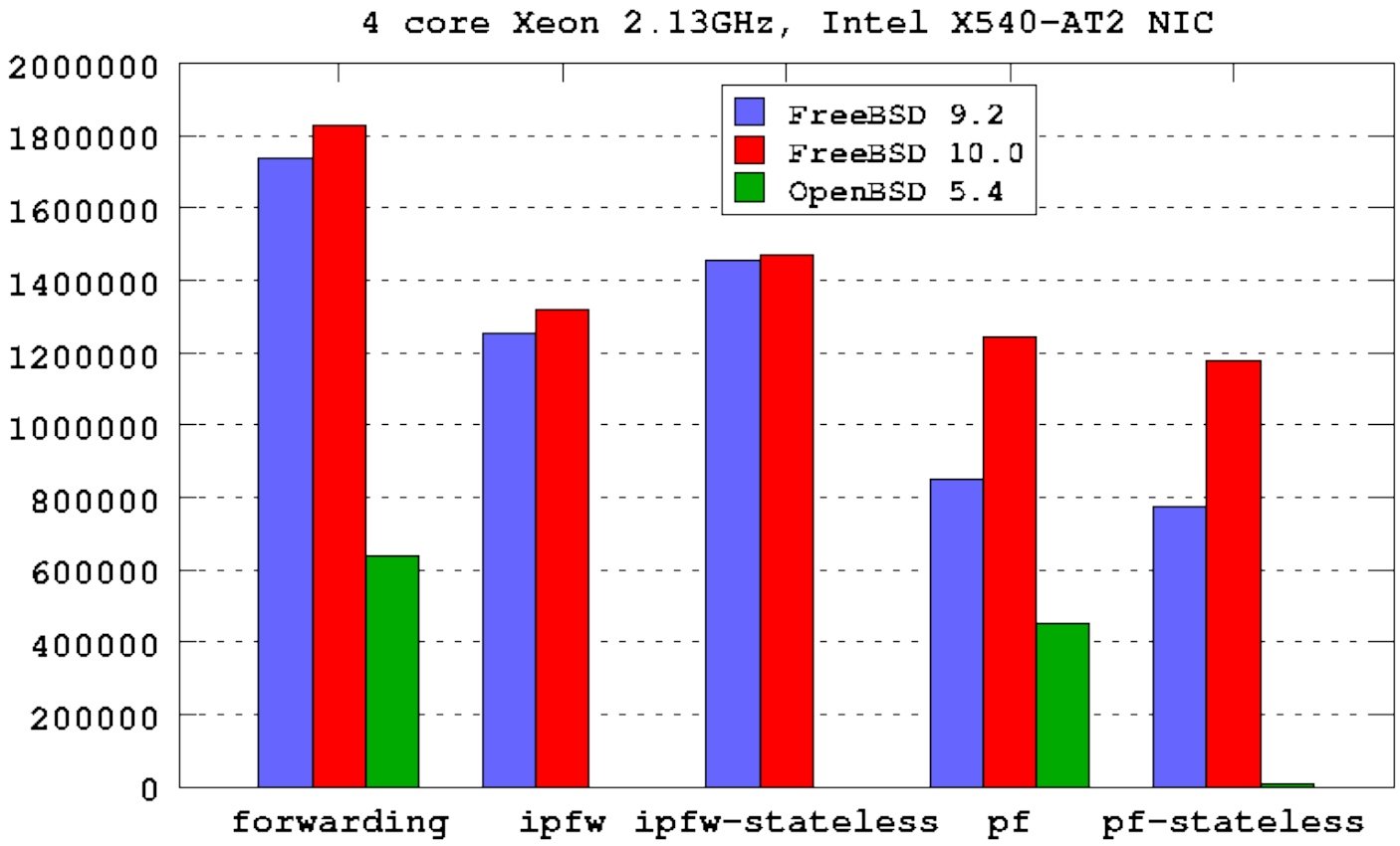
  ・一部のGEOMクラスだけ、逐次ではなく並列に
  処理するように変更

  ・例えばDISKクラス（ディスクそのもの）への要求は
  複数並列に送りつけても良いはず

  ・9系は 200k IOPSの水準は出せない

pf scalability

- **SMPスケーラビリティに関する改良**



はまりやすいところ

- BIND 消えた
- CARP の変更
- pkg_tools なくなった

BIND消えた

- dig がなくなって drill に （LDNSライブラリベース）

- di

kterm - hrs@alph.allbsd.org:/usr/ports/devel/subversion

```
hrs@alph % dig www.freebsd.org

; <<>> DiG 9.9.6-P1 <<>> www.freebsd.org
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 41779
;; flags: qr rd ra; QUERY: 1, ANSWER: 2, AUTHORITY: 13, ADDITIONAL: 1

;; OPT PSEUDOSECTION:
; EDNS: version: 0, flags:; udp: 4096
;; QUESTION SECTION:
;www.freebsd.org.               IN      A

;; ANSWER SECTION:
www.freebsd.org.        577     IN      CNAME   wfe0.ysv.freebsd.org.
wfe0.ysv.freebsd.org.   577     IN      A       8.8.178.110

;; AUTHORITY SECTION:
.                       290093  IN      NS      b.root-servers.net.
.                       290093  IN      NS      i.root-servers.net.
.                       290093  IN      NS      e.root-servers.net.
.                       290093  IN      NS      m.root-servers.net.
.                       290093  IN      NS      a.root-servers.net.
.                       290093  IN      NS      h.root-servers.net.
.                       290093  IN      NS      d.root-servers.net.
.                       290093  IN      NS      k.root-servers.net.
.                       290093  IN      NS      f.root-servers.net.
.                       290093  IN      NS      l.root-servers.net.
.                       290093  IN      NS      g.root-servers.net.
.                       290093  IN      NS      j.root-servers.net.
.                       290093  IN      NS      c.root-servers.net.

;; Query time: 1 msec
;; SERVER: 127.0.0.1#53(127.0.0.1)
;; WHEN: Fri Dec 26 14:40:44 JST 2014
;; MSG SIZE  rcvd: 294

hrs@alph %
```

DNSゾーン

- C

```
hrs@alph % drill www.freebsd.org
;; ->>HEADER<<- opcode: QUERY, rcode: NOERROR, id: 3990
;; flags: qr rd ra ; QUERY: 1, ANSWER: 2, AUTHORITY: 13, ADDITIONAL: 0
;; QUESTION SECTION:
;; www.freebsd.org.      IN      A

;; ANSWER SECTION:
www.freebsd.org.         545     IN      CNAME   wfe0.ysv.freebsd.org.
wfe0.ysv.freebsd.org.    545     IN      A       8.8.178.110

;; AUTHORITY SECTION:
.        290061  IN      NS      a.root-servers.net.
.        290061  IN      NS      j.root-servers.net.
.        290061  IN      NS      h.root-servers.net.
.        290061  IN      NS      m.root-servers.net.
.        290061  IN      NS      g.root-servers.net.
.        290061  IN      NS      b.root-servers.net.
.        290061  IN      NS      k.root-servers.net.
.        290061  IN      NS      l.root-servers.net.
.        290061  IN      NS      d.root-servers.net.
.        290061  IN      NS      e.root-servers.net.
.        290061  IN      NS      i.root-servers.net.
.        290061  IN      NS      c.root-servers.net.
.        290061  IN      NS      f.root-servers.net.

;; ADDITIONAL SECTION:

;; Query time: 5 msec
;; SERVER: 192.168.0.90
;; WHEN: Fri Dec 26 14:41:16 2014
;; MSG SIZE  rcvd: 283
hrs@alph %
```

BIND消えた

- /etc/namedb, /var/named が消えた
- dns/bind910 を入れましょう
- /usr/local/etc/namedb を使ってください

- **FreeBSD 9系／BIND 9.8系までを使っていた人は...**
  - dns/bind910 には chroot 機能がありません。
  - BIND 9.10 系へ移行すると：
    slave zone ファイルフォーマットが変わります。
    IPv6でのlistenがデフォルトで入ります

BIND消えた

- **FreeBSD 9系／BIND 9.9系までを使っていた人は...**
  - dns/bind910をインストールした後、/etc/namedb もしくは /var/named/etc/namedb を /usr/local/etc/namedbと置き換えましょう
  - 「ln -s /usr/local/etc/namedb /etc」として symlinkを張りましょう
  - 今までの設定を変えたくなければ、named.conf に次の設定を入れましょう

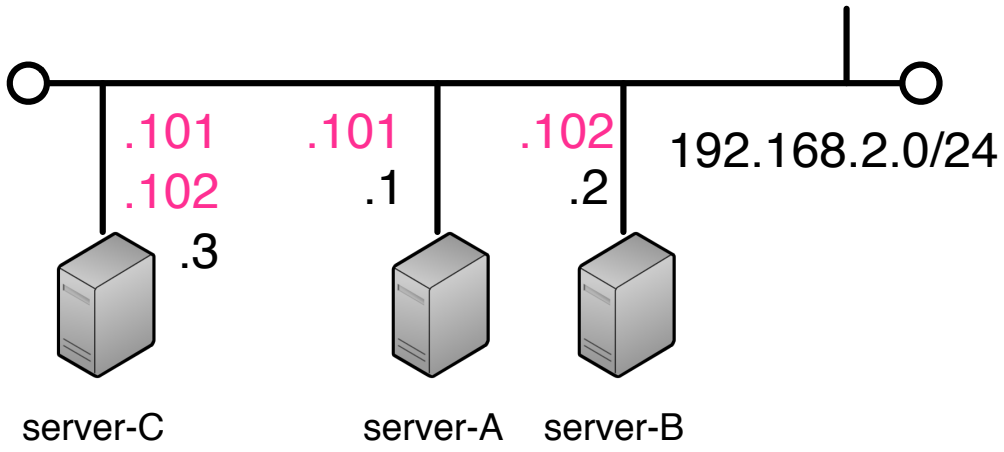
namedb/named.conf

```
options {
        listen-on-v6            { none; };
        masterfile-format       text;
};
```

CARP が変わった

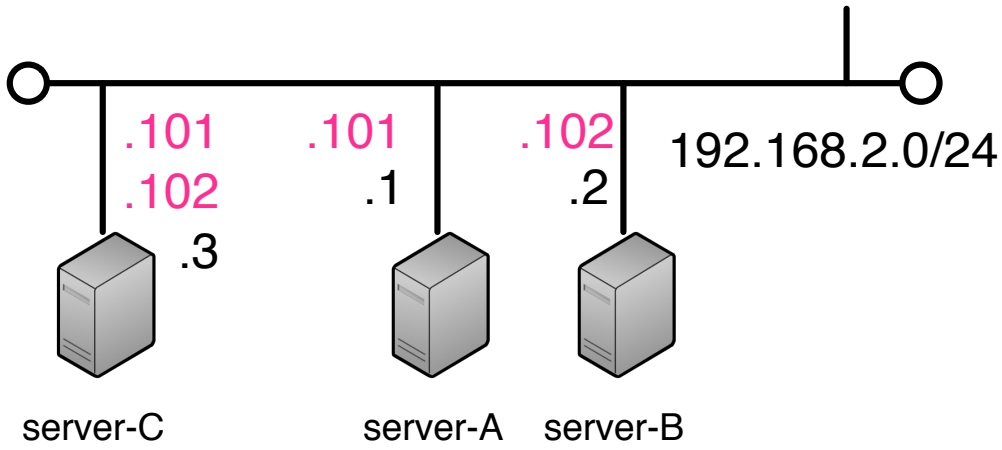
- **CARP ってなんだ**
  同一のIPアドレスを複数のマシンに付けて、常に
  そのうち1台だけが通信できるようにする仕組み。

- 複数のサーバを用意してフェイルオーバさせる
  （Active-Standby構成）

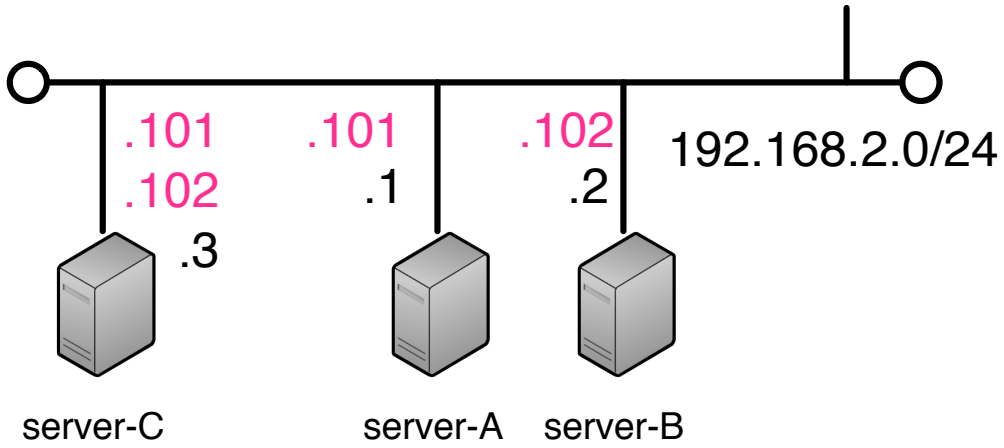


CARP が変わった

- **変更点**
  - ・CARP インタフェースはなくなりました。
  - ・機能はふつうのインタフェースへ統合



CARP が変わった



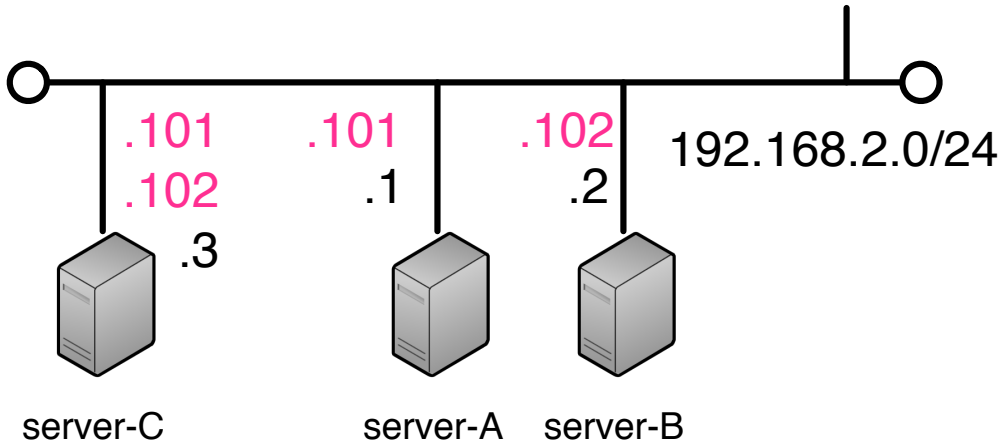
10系の書き方

```
hostname="server-A"
ifconfig_em0="inet 192.168.2.1/24"
ifconfig_em0_alias0="vhid 1 pass testtest alias 192.168.2.101/32"
```

```
hostname="server-B"
ifconfig_em0="inet 192.168.2.2/24"
ifconfig_em0_alias0="vhid 2 pass passpass alias 192.168.2.102/32"
```

```
hostname="server-C"
ifconfig_em0="inet 192.168.2.3/24"
ifconfig_em0_alias0="vhid 1 advskew 100 pass testtest alias 192.168.2.101/32"
ifconfig_em0_alias1="vhid 2 advskew 100 pass passpass alias 192.168.2.102/32"
```

CARP が変わった



9系の書き方

```
hostname="server-A"
cloned_interfaces="carp0"
ifconfig_em0="inet 192.168.2.1/24"
ifconfig_carp0="vhid 1 pass testtest alias 192.168.2.101/24"

hostname="server-B"
cloned_interfaces="carp0"
ifconfig_em0="inet 192.168.2.2/24"
ifconfig_carp0="vhid 2 pass passpass alias 192.168.2.102/24"


hostname="server-C"
cloned_interfaces="carp0 carp1"
ifconfig_em0="inet 192.168.2.3/24"
ifconfig_carp0="vhid 1 advskew 100 pass testtest alias 192.168.2.101/24"
ifconfig_carp1="vhid 2 advskew 100 pass passpass alias 192.168.2.102/24"
```

pkg_tool(8)がなくなった

- packageは pkg(8)に全面移行

  - 「`pkg_foo`」が「`pkg foo`」になったととりあえず思おう

  - パッケージの扱いはあまり変わっていない
    → ports を使って構築したものが package
    → portupgrade や portmaster はこれまでどおり
    → 新しい機能をつかおうとしなければ、
    今までの作業フローは大きく変わらない
  - packageだけで良ければ、pkg upgrade で
    バイナリアップグレードできる
  - ports と pre-compiled package を組み合わせる場合に
    工夫が必要

SAたくさん

SA-13:14.openssh	19 November 2013	OpenSSH AES-GCM memory corruption vulnerability
SA-14:01.bsnmpd	14 January 2014	bsnmpd remote denial of service vulnerability
SA-14:02.ntpd	**14 January 2014**	**ntpd distributed reflection Denial of Service vulnerability**
SA-14:03.openssl	**14 January 2014**	**OpenSSL multiple vulnerabilities**
SA-14:04.bind	14 January 2014	BIND remote denial of service vulnerability
SA-14:05.nfsserver	8 April 2014	Deadlock in the NFS server
SA-14:06.openssl	8 April 2014	OpenSSL multiple vulnerabilities
SA-14:07.devfs	30 April 2014	Fix devfs rules not applied by default for jails
SA-14:08.tcp	30 April 2014	Fix TCP reassembly vulnerability
SA-14:09.openssl	**30 April 2014**	**Fix OpenSSL use-after-free vulnerability**
SA-14:10.openssl	**15 May 2014**	**Fix OpenSSL NULL pointer deference vulnerability**
SA-14:11.sendmail	3 June 2014	Fix sendmail improper close-on-exec flag handling
SA-14:13.pam	3 June 2014	Fix incorrect error handling in PAM policy parser
SA-14:14.openssl	**5 June 2014**	**Multiple vulnerabilities**
SA-14:15.iconv	24 June 2014	NULL pointer dereference and out-of-bounds array access
SA-14:16.file	24 June 2014	Multiple vulnerabilities
SA-14:17.kmem	8 July 2014	Kernel memory disclosure in control messages and SCTP notifications
SA-14:18.openssl	**9 September 2014**	**Multiple vulnerabilities**
SA-14:19.tcp	16 September 2014	Denial of Service in TCP packet processing.
SA-14:20.rtsold	**21 October 2014**	**Remote buffer overflow vulnerability.**
SA-14:21.routed	**21 October 2014**	**Remote denial of service vulnerability.**
SA-14:22.namei	21 October 2014	Memory leak in sandboxed namei lookup.
SA-14:23.openssl	**21 October 2014**	**Multiple vulerabilities.**
SA-14:25.setlogin	04 November 2014	Kernel stack disclosure.
SA-14:26.ftp	**04 November 2014**	**Remote code execution.**
SA-14:27.stdio	10 December 2014	Buffer overflow in stdio
SA-14:28.file	10 December 2014	Multiple vulnerabilities in file(1) and libmagic(3)
SA-14:29.bind	10 December 2014	BIND remote denial of service vulnerability
SA-14:30.unbound	17 December 2014	unbound remote denial of service vulnerability
SA-14:31.ntp	**23 December 2014**	**Multiple vulnerabilities in NTP suite**

何か他にはありますか？

- なんか変だと思ったら、とりあえず声をあげましょう

AsiaBSDCon


AsiaBSDCon2015
A Technical Conference for Users and Developers on BSD-based Systems
Redistribution and use in sou nd bi forms, with or hout modification, are
permitted provided that the met
1. Redistributions of source t notice, this list of
conditions and the follow
2. Redistributions in binary copyright notice, this list of
conditions and the fol entation and/or other
materials provide it
THIS SOFTWA Y TH CONTRIBUTORS "AS IS" AND

告知

- AsiaBSDCon 2015（JR飯田橋駅付近）
  2015/3/12-15。ベンダーサミットもやります
- FreeBSD勉強会（有楽町線・麹町駅付近）
  不定期（おおよそ月一回）
- FreeBSDワークショップ（JR飯田橋駅付近）
  月一回のしゃべる会。