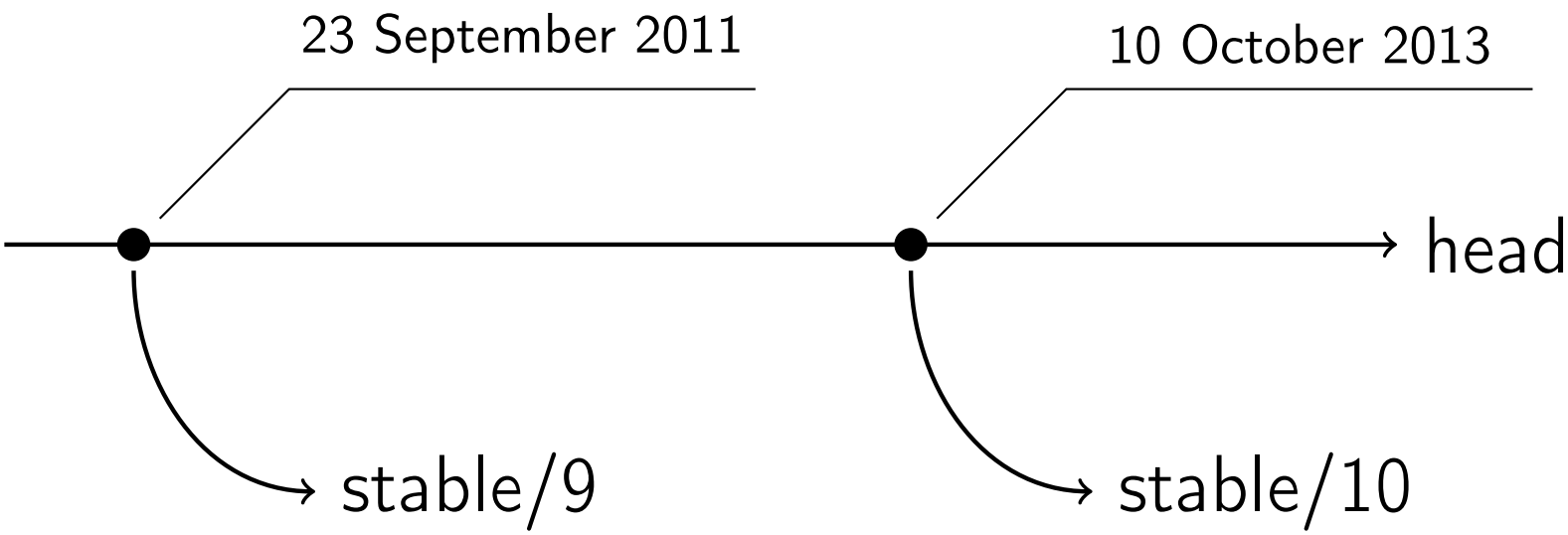
What's new in FreeBSD 10?

Gleb Smirnoff
glebius@FreeBSD.org

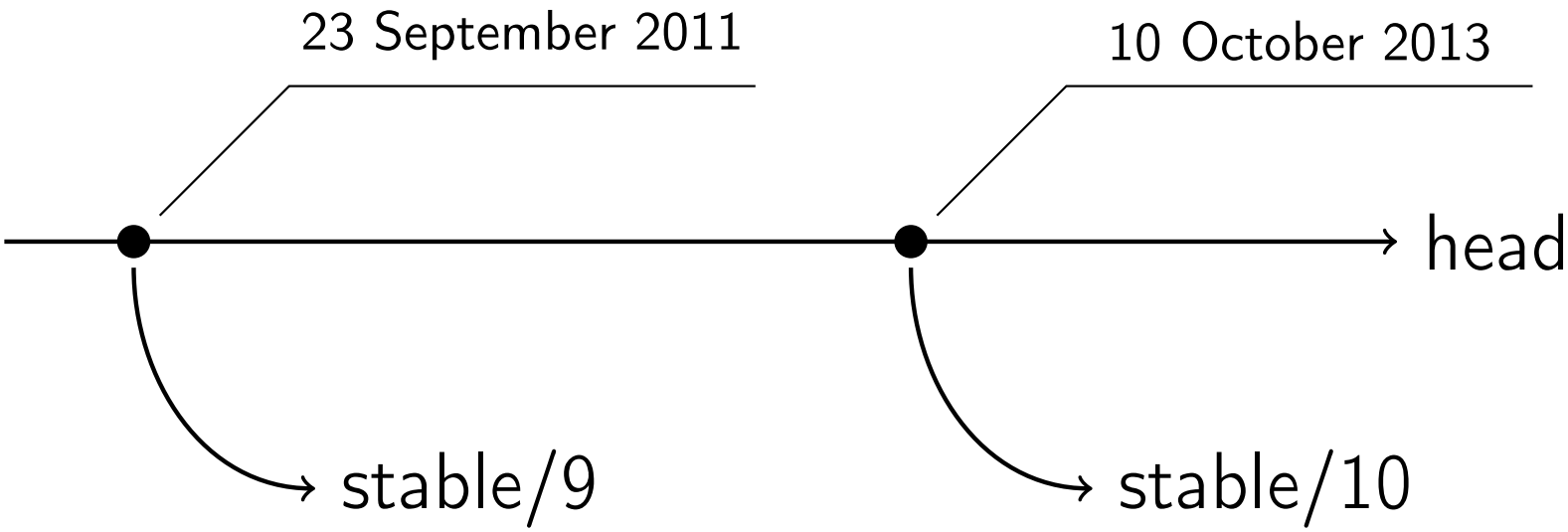
ruBSD 2013
Yandex
Moscow

December 14, 2013

Two years of development



Two years of development



- 10.0-RC1 available now
- 10.0-RELEASE planned for 2 January 2014

Outline

1. Userland changes
   - Packaging system
   - Toolchain
   - Developers tools
   - DNS tools
   - Other userland updates
2. Kernel: virtualization
   - bhyve
   - guest improvements
3. Kernel: ARM port

Outline

4. Kernel: security
   - capsicum(4) update
   - /dev/random improvements
5. Kernel: general improvements
   - callout(9) new generation
   - unmapped I/O
   - memory management
   - atomic close-on-exec
6. Kernel: I/O and storage
   - improvements
   - GEOM
   - 3rd party filesystems

Outline

7. Kernel: networking
   - changes
   - carp
   - packet filters
8. Conslusion
   - looking forward to FreeBSD 11

New generation packaging system

pkg(1)

- Replaces pkg_tools in FreeBSD 10.0
- Updates packages from remote repository
- Is developed as a library + command line frontend

New generation packaging system

pkg(1)

- Replaces pkg_tools in FreeBSD 10.0
- Updates packages from remote repository
- Is developed as a library + command line frontend

Don't miss section at 11:40 by Vsevolod Stakhov!

Compiler change

LLVM/Clang 3.3 is default compiler
(amd64, arm and i386)

Compiler change

LLVM/Clang 3.3 is default compiler
(amd64, arm and i386)

Why?

- BSD licensed (gcc > 4.2.1 is GPLv3)
- Fully C++11 compliant. Includes LLVM libc++.
- Always cross compiler.

Compiler change

LLVM/Clang 3.3 is default compiler
(amd64, arm and i386)

Why?

- BSD licensed (gcc > 4.2.1 is GPLv3)
- Fully C++11 compliant. Includes LLVM libc++.
- Always cross compiler.

We still support gcc 4.2+ to build tier 2 arches.

Toolchain

- Moving towards external toolchain.

Toolchain

- Moving towards external toolchain.
- Portable make(1) imported from NetBSD

Toolchain

- Moving towards external toolchain.
- Portable make(1) imported from NetBSD
- Tools updated:
  - patch(1): ~~GNU~~ BSD licensed fork of original Larry Wall
  - sort(1): ~~GNU~~ own implementation
  - byacc for yacc(1)
  - flex for lex(1)

Developers tools

- ~~CVS~~ -> subversion (lite)
- ATF/kyua from NetBSD
- Work in progress: ~~gdb~~ -> lldb

DNS tools

- Recursive resolver & tools
  - ~~*BIND*~~ -> *unbound*
  - ~~dig(1)~~ -> drill(1)
  - new host(1) implementation
  - ~~nslookup~~

DNS tools

- Recursive resolver & tools
  - ~~*BIND*~~ -> *unbound*
  - ~~dig(1)~~ -> drill(1)
  - new host(1) implementation
  - ~~nslookup~~
- LDNS library
  - Feature rich API, providing control over recursion, DNSSEC, TSIG, etc.
  - Utilized by OpenSSH, drill(1)

DNS tools

- Recursive resolver & tools
  - ~~*BIND*~~ -> *unbound*
  - ~~dig(1)~~ -> drill(1)
  - new host(1) implementation
  - ~~nslookup~~
- LDNS library
  - Feature rich API, providing control over recursion, DNSSEC, TSIG, etc.
  - Utilized by OpenSSH, drill(1)
- Plan for FreeBSD 11: caching, validating, secure resolver library with standard API

Other userland updates

- freebsd-version(1) tool introduced
- libyaml added to base

Other userland updates

- freebsd-version(1) tool introduced
- libyaml added to base
- Citrus iconv(3) in libc
- newest *jemalloc* 3.4.1 in libc

Other userland updates

- freebsd-version(1) tool introduced
- libyaml added to base
- Citrus iconv(3) in libc
- newest *jemalloc* 3.4.1 in libc
- *nvi* editor supports wide character locales

Other userland updates

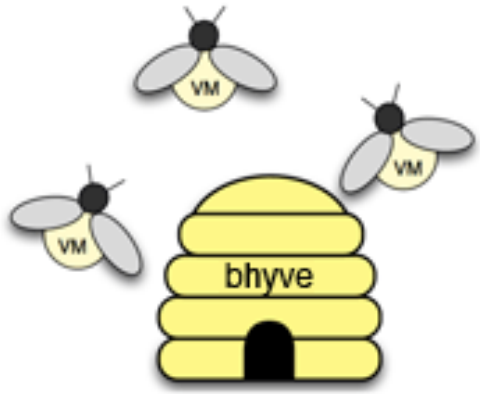
- freebsd-version(1) tool introduced
- libyaml added to base
- Citrus iconv(3) in libc
- newest *jemalloc* 3.4.1 in libc
- *nvi* editor supports wide character locales
- wpa_supplicant/hostapd updated to 2.0
- OpenSSH updated to 6.4
- OpenSSL updated to 1.0.1e

Installer

- *bsdinstall* features ZFS root installation
- Removed old installer *sysinstall* and auxiliary tools *libdisk*, *libftpio*, *sade*

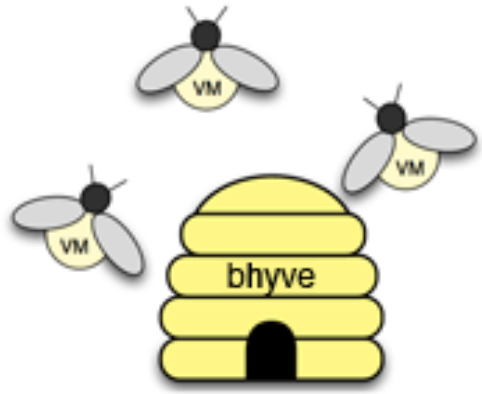
bhyve(4) hypervisor

BSD hyper visor
(pronounced as “bee hive”)



bhyve(4) hypervisor

BSD hyper visor
(pronounced as “bee hive”)

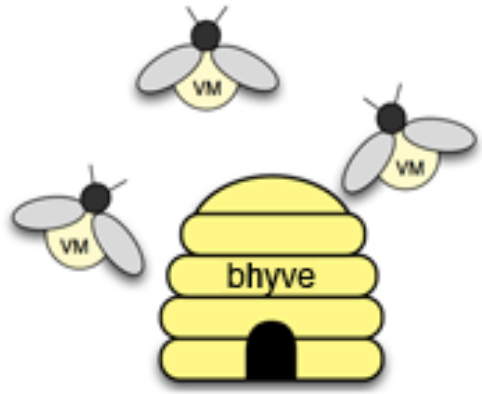


Requirements:

- host is amd64: Intel CPU with VT-x feature or AMD CPU with AMD-V feature
- no BIOS provided

bhyve(4) hypervisor

BSD hyper visor
(pronounced as “bee hive”)

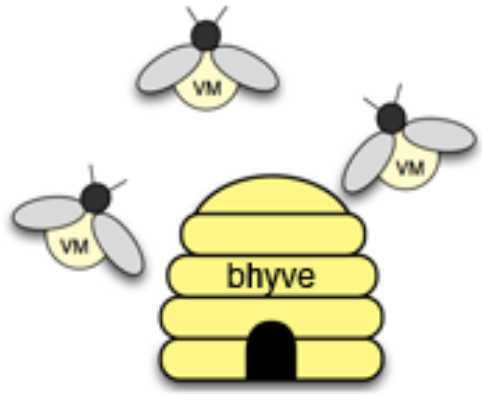


Results in:

- 12k lines of code in kernel
- 14k lines of code in userland

bhyve(4) hypervisor

BSD hyper visor
(pronounced as “bee hive”)



Guest OSes supported:

- FreeBSD, OpenBSD
- GNU/Linux

Guest improvements

- Xen and Xen HVM in GENERIC kernel
- Microsoft Hyper-V drivers added
- VMware VMXNET3 driver added

ARM port

ARM soon to become Tier 1 platform

- compiled with clang
- superpages support
- EABI by default

capsicum(4) update

Capsicum - hybrid capability + UNIX access control model. Introduced in FreeBSD 9.0.

capsicum(4) update

Capsicum integrates further into FreeBSD:

- notions of “capability” and “file descriptor” merge
- new APIs: ~~cap_new(2)~~ cap_rights_limit(2)

capsicum(4) update

Capsicum integrates further into FreeBSD:

- notions of "capability" and "file descriptor" merge
- new APIs: ~~cap_new(2)~~ cap_rights_limit(2)
- capsicum(4) in GENERIC by default
- sandboxed applications: tcpdump(1), dhclient(8), rwhod(8), kdump(8), hastd(8), auditdistd(8), ctld(8), iscsid(8)

capsicum(4) update

Future integration in 10.1-RELEASE:

- casperd(8) daemon
- libcapsicum(3) library
- sandboxing a lot of applications

better random

Problem: hardware assisted randomness (RDRAND and Padlock) no longer trusted.

Solution: run them through Yarrow.

better random

Problem: hardware assisted randomness (RDRAND and Padlock) no longer trusted.

Solution: run them through Yarrow.

Problem: not enough entropy on early boot.

Solution: we can get some from device attach time.

better random

Problem: hardware assisted randomness (RDRAND and Padlock) no longer trusted.

Solution: run them through Yarrow.

Problem: not enough entropy on early boot.

Solution: we can get some from device attach time.

Problem: not enough entropy on first boot.

Let bsdinstall save an entropy cookie for future boot.

better random

Problem: hardware assisted randomness (RDRAND and Padlock) no longer trusted.

Solution: run them through Yarrow.

Problem: not enough entropy on early boot.

Solution: we can get some from device attach time.

Problem: not enough entropy on first boot.

Let bsdinstall save an entropy cookie for future boot.

FreeBSD 11.0 plan: substitute Yarrow with Fortuna.

callout(9) improvements

callout(9) - kernel subsystem to schedule delayed events.

callout(9) improvements

callout(9) - kernel subsystem to schedule delayed events.
New improvements:

- tickless
- event coalescing
- direct execution

unmapped I/O

Problem: kernel doing I/O on behalf of userland process maps the I/O region into kernel address space.

unmapped I/O

Problem: kernel doing I/O on behalf of userland process maps the I/O region into kernel address space. Change of virtual memory map requires notification of other CPUs.

unmapped I/O

Problem: kernel doing I/O on behalf of userland process maps the I/O region into kernel address space. Change of virtual memory map requires notification of other CPUs.

Solution: unmapped I/O. Required modification of file system layer, GEOM classes, disk drivers.

unmapped I/O

Problem: kernel doing I/O on behalf of userland process maps the I/O region into kernel address space. Change of virtual memory map requires notification of other CPUs.

Solution: unmapped I/O. Required modification of file system layer, GEOM classes, disk drivers.

Result: 30% of system CPU time saved in I/O bound tasks.

memory management changes

- Kernel memory maps:
  - vmem(9) generic allocator from NetBSD
  - kernel memory map allocation backed by vmem(9)

memory management changes

- Kernel memory maps:
  - vmem(9) generic allocator from NetBSD
  - kernel memory map allocation backed by vmem(9)
- Mach VM
  - radix tree instead of splay tree for vm_pages in vm_object

memory management changes

- Kernel memory maps:
  - vmem(9) generic allocator from NetBSD
  - kernel memory map allocation backed by vmem(9)
- Mach VM
  - radix tree instead of splay tree for vm_pages in vm_object
- UMA
  - performance/efficiency improvements
  - per-CPU zones
  - log warning when a zone hits limit

atomic close-on-exec

- Prevents descriptor leak in presence of threads or signals
- Suggested for future POSIX

storage changes

- NAND flash support
  - NAND controller/chip/bus APIs
  - NAND disk GEOM class
  - NAND file system

storage changes

- NAND flash support
  - NAND controller/chip/bus APIs
  - NAND disk GEOM class
  - NAND file system
- Resizing
  - general support of “resize” notion in GEOM
  - resizing of GEOM mirror (in 10.1-RELEASE)
  - growfs(1) works on mounted filesystems

storage changes

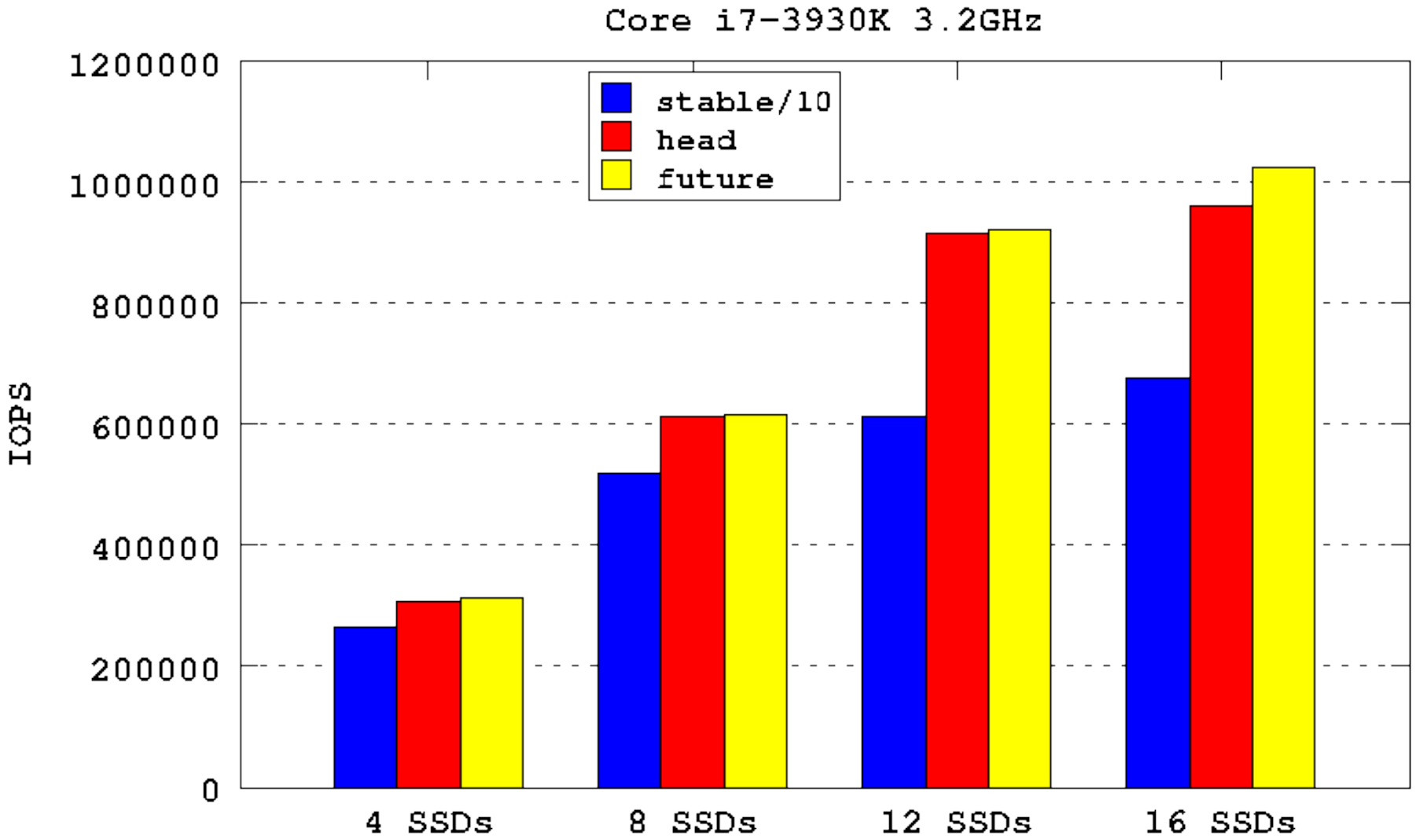
- NAND flash support
  - NAND controller/chip/bus APIs
  - NAND disk GEOM class
  - NAND file system
- Resizing
  - general support of “resize” notion in GEOM
  - resizing of GEOM mirror (in 10.1-RELEASE)
  - growfs(1) works on mounted filesystems
- legacy ATA layer removed

GEOM: work in progress

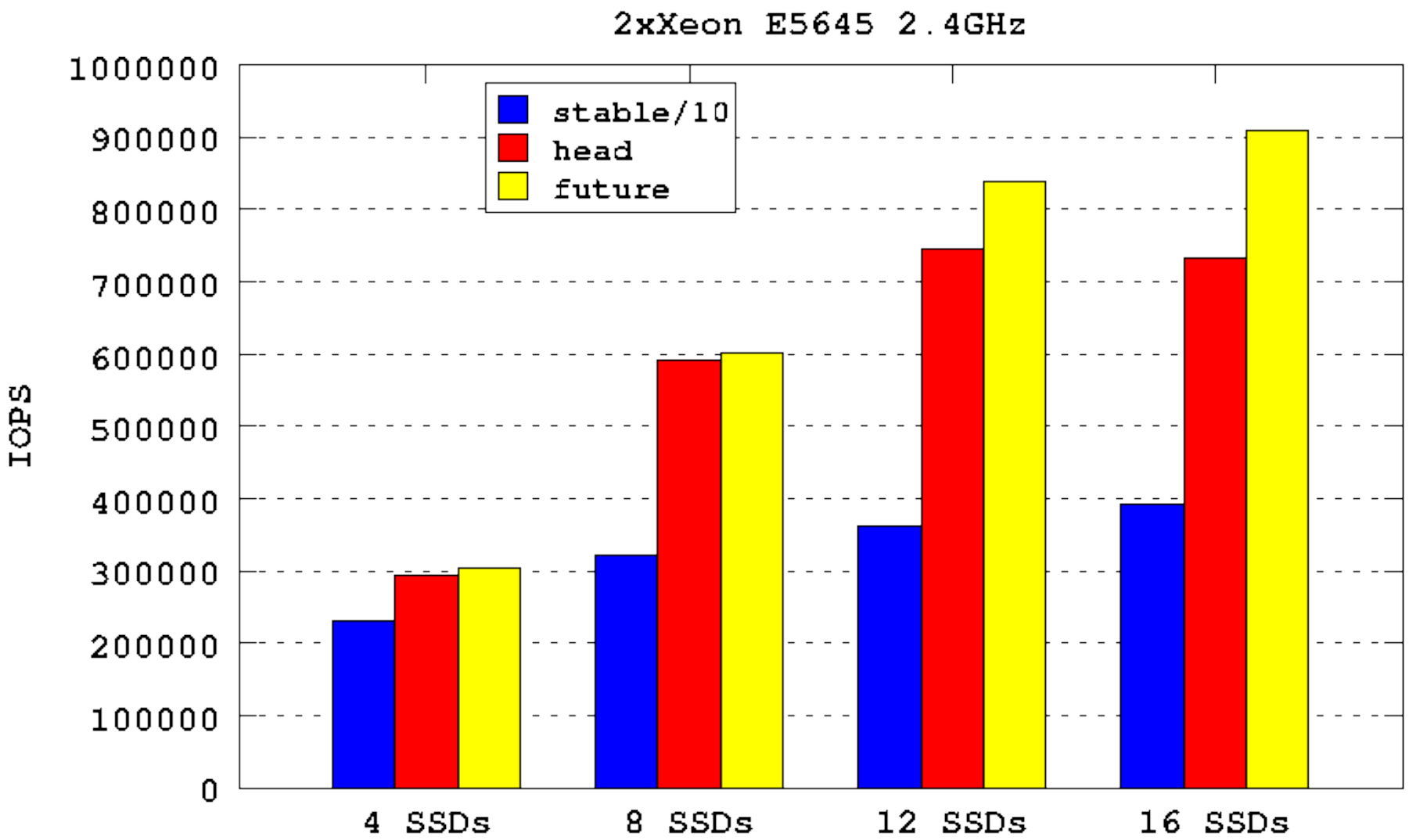
Targeted for 10.1-RELEASE:

- direct dispatch in GEOM instead of two threads
- fine grained locking of CAM layer

GEOM: work in progress



GEOM: work in progress



FUSE

- FUSE moved to base from ports to improve stability

FUSE

- FUSE moved to base from ports to improve stability
- Giant-locked and GPL-contaminated filesystems removed from kernel: hpfs, ext2fs, ntfs, reiserfs, coda, xfs, nwfs, portalfs.

networking changes

- newest Infiniband OFED stack
- native iSCSI Target and Initiator
- etherswitch(4): embedded Ethernet switch driver

networking changes

- ~~ZERO_COPY_SOCKETS~~
- sendfile(2) on shared memory fd

networking changes

- network byte order throughout the stack
- counter(9): raceless and cheap statistic per-CPU counters
- IP/TCP/UDP dtrace(1) providers

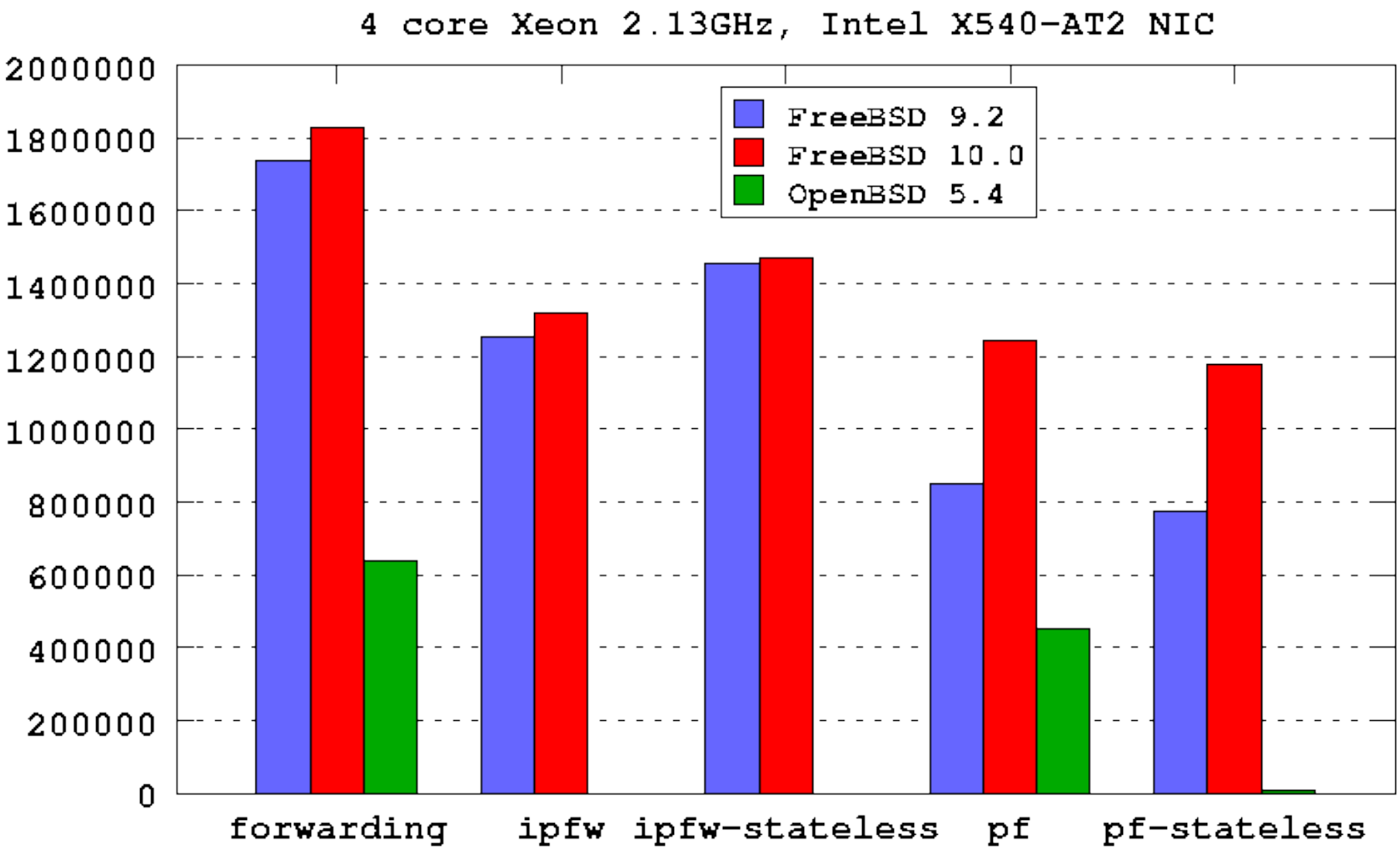
new carp(4)

CARP isn't pseudo-interface any more. Redundant address is configured directly on a real interface.

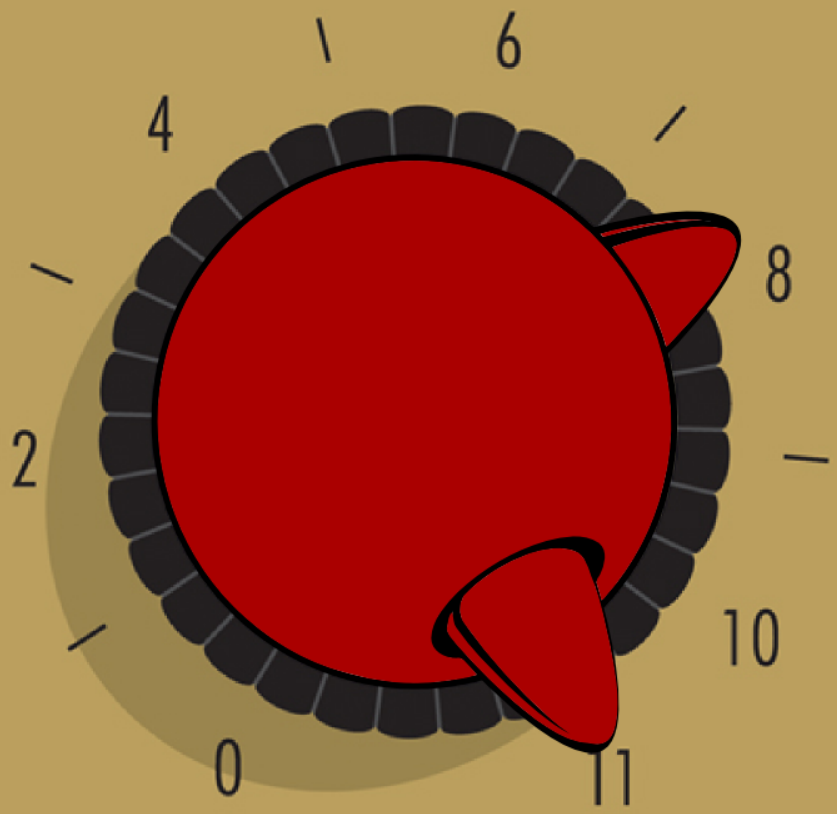
```
% ifconfig igb0 10.0.0.112/27 vhid 112
% ifconfig igb0
igb0: flags=8943<UP,BROADCAST,RUNNING,PROMISC,SIMPLEX,MULTICAST> metric 0 mtu 1500
        ether 00:25:90:03:0e:fa
        inet 10.0.0.112 netmask 0xffffffe0 broadcast 10.0.0.127 vhid 112
        media: Ethernet autoselect (1000baseT <full-duplex>)
        status: active
        carp: BACKUP vhid 112 advbase 1 advskew 0
```

packet filters

- pf(4): fork off OpenBSD, bringing in multithreading
- ipfilter(4): update to 5.1.2 (BSD license pledged)
- ipfw(4): no significant changes


4 core Xeon 2.13GHz, Intel X540-AT2 NIC
2000000
1800000
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
FreeBSD 9.2
FreeBSD 10.0
OpenBSD 5.4
forwarding
ipfw
ipfw-stateless
pf
pf-stateless

FreeBSD



Turn it up all the way to 11

FreeBSD

Questions?

Turn it up all the way to 11