Virtual kernel update

My (re)V_iew

Bjoern A. Zeeb

bz@FreeBSD.org

The FreeBSD Project

EuroBSDCon 2009, DevSummit

Overview

No time for that.
Need to tell you about 10 years in 10 minutes.

Where the story began

1999:

- April:

  ```
  Revision 46155
  Added Wed Apr 28 11:38:52 1999 UTC (10y, 4m ago) by phk

  This Implements the mumbled about "Jail" feature.
  ...
  If somebody wants to take it from here and develop it into
  more of a "virtual machine" they should be most welcome!
  ...
  ```

- May: VMware Workstation 1.0
- September 17th: FreeBSD 3.3-Release.

2000:

- March: FreeBSD 4.0

Jails 3-5-6-7 years later

- 2002
  Marko Zec *BSD network stack virtualization*
  BSDCon Europe, based on FreeBSD 4.7.
- 2002/2003
  pjd posts multi-IPv4 jail patches.
- 2005
  raw socket support, global sysctl, submitted by csjp.
- 2006
  Initial try on multi-IPv4/v6 based on pjd multi-IPv4 and cognet single-IPv6 patches.
  Jail resource limits by Chris Jones (GSoC ’06).
- 2007
  Jail friendly file systems (ZFS) by pjd.

9-10 years we missed

- 2008
  Updated resource limits patch by Christopher Thunes.
  Multi-IPv4/v6/no-IP patches by bz.
  Jail Wiki page.
  November - multi-IPv4/v6/no-IP patches in HEAD.
- 2009
  February - multi-IPv4/v6/no-IP patches in 7-STABLE
  and the later 7.2-RELEASE.
  Hierachical jails, new, flexible syntax and new syscalls from jamie.
  Hierachical resource limits by trasz (GSoC '09).

But there was more happening the last years. Let's see.

The trail

Whatever was between 2002 and this is beyond my knowledge apart from that I had a bookmark to the vimage work since.

- EuroBSDCon 2006 - Whiteboard session.
- Dec 2006 - FF newsletter: Network Stack Virtualization Project. Sponsorship from NLNet. Protoype for 7-CURRENT in early 2007.
- EuroBSDCon 2007 - Danish country side.
- BSDCan 2008 - The famous schedule #1.
- Cambridge DevSummit 2008 - Let the games begin.
  The famous schedule #2 ”4 and a half steps”.
- EuroBSDCon 2008 - An update.
- BSDCan 2009 - We are done - but the very last . . . .
- The last 4 months.

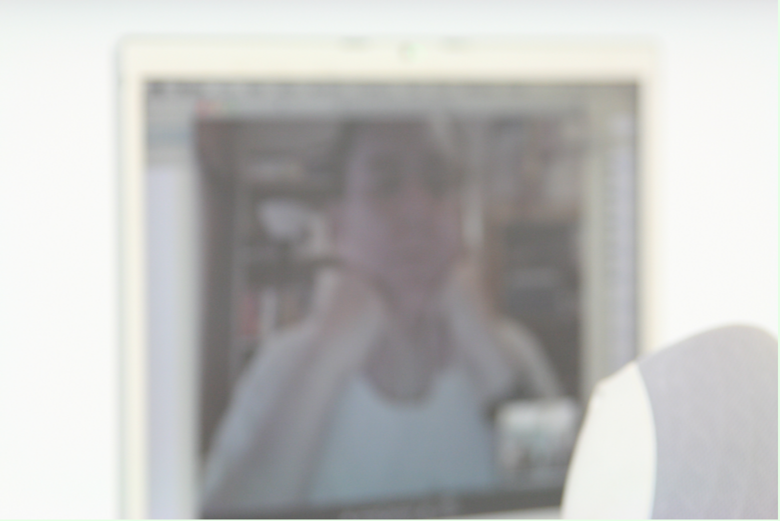
Last year

- Cambridge started at BSDcan 2008
  two days after the famous schedule #1.
- Many thanks to Robert - I still feel guilty.
- What were we doing?

Punting and fun

(Kris Kennaway, FreeBSD DevSummit, Cambridge, UK, August 2008, http://people.freebsd.org/ kris/Cambridge/DSC03042.JPG)

Video conferencing with Julian and Jamie

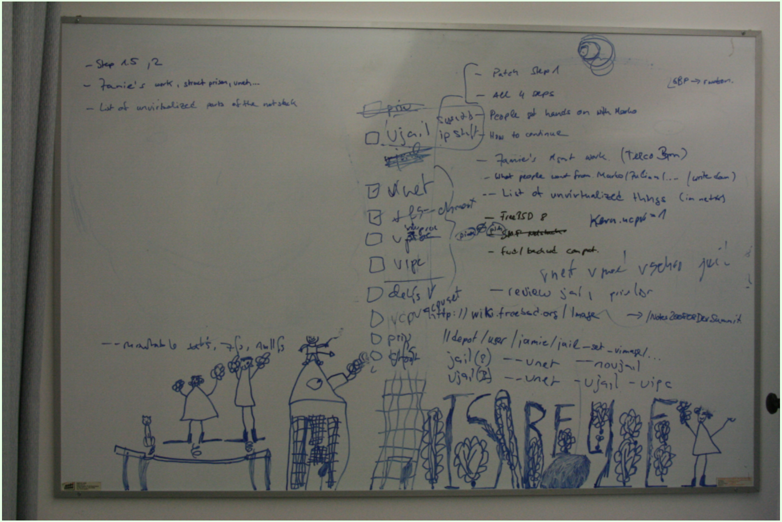


(Bjoern Zeeb, FreeBSD DevSummit, Cambridge, UK, August 2008)

A full moon night

(Bjoern Zeeb, FreeBSD DevSummit, Cambridge, UK, August 2008)

Future Dreams^WDesigns



- Step 1.5 ,2
- Jamie's work, struct prison, vnet...
- List of unvirtualized parts of the netstack

~~□ priv~~
□ Ujail
☑ vnet
☑ + fs -- chmount
□ v ~~pgc~~ proc
□ vipc
□ devfs ✓
□ vcpu acpuset
○ priv
○ chroot

switch
ip shift

- Patch Step 1
- All 4 steps
- People get hands on with Marko
- How to continue
- Jamie's Mgmt work (Telco 3pm)
- What people want from Marko/Julian/... (write down)
-- List of unvirtualized things (in netstack)
- FreeBSD 8
- ~~SMP netstack~~
- fwd/backwd compat.

piano pith

kern.ncpus=1

BGP → routers.

vnet vproc! vsched jail!

— review jail priv lsr

http://wiki.freebsd.org/Image → /Notes200808DevSummit

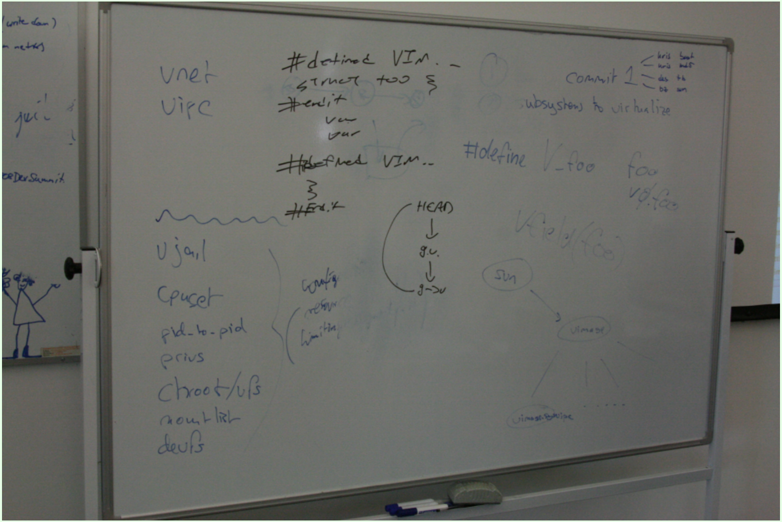
//depot/user/jamie/jail-set-vimage/...

jail(8) --vnet --novjail
vjail(8) --vnet -vjail -vipc

-- mountable stats, zfs, nullfs

(Bjoern Zeeb, FreeBSD DevSummit, Cambridge, UK, August 2008)

Future Dreams^WDesigns cont.ed



(Bjoern Zeeb, FreeBSD DevSummit, Cambridge, UK, August 2008)

13 months back to the day

The V_commit.

```
Date: Sun Aug 17 23:27:27 2008 UTC
Log Message:

Commit step 1 of the vimage project, (network stack)
virtualization work done by Marko Zec (zec@).
[...]
V_Commit_Message_Reviewed_By: more people than the patch
```

Summary: among other presentations and the first commit
- we had a sponsored dinner.
- we came up with the famous schedule #2.
- we talked about jail branding and integration.
- we talked about future virtualization.

Before BSDCan 2009

- Step-by-step merging of VImage changes to HEAD.
- Status update at the DevSummit of EuroBSDCon 2008 by Marko.
- Large structs with collections of (formerly) global variables.
- #include poisoning problems also for modules and ABI problem with struct vnet_net and MRT.
- The global variables under #ifdef.
- 3 different kernel options:
  GENERIC: classic globals,
  VIMAGE_GLOBALS: container structs but no indirection,
  VIMAGE: container structs; multiple network stacks possible.
- Wiki pages with basic information.
- Independently from spring 2009 on:
  merging of hierachical jails from Jamie.

At the time of BSDCan 2009 - after 1 year

- Basically done apart from the userland interface to create virtual network stacks.
- Plans for before 8.0-Release:
  - Add temporary classic vimage API to test things.
  - Have Jamie merge the jail parts for vnets.
  - Remove the classic API again before 8.0 and have a vimage(8) compat for early adopters.
  - Test, virtualize more missing parts, talk to re@ to ship a VIAMGE kernel, ...
  - Probably more I cannot remember.

After BSDCan 2009

- Jails manage virtual network stacks. Done and merged by Jamie.
- After 6 months I got hold of Peter. He explains the linker set idea.
- Robert takes dpcpu and linker sets and develops the infrastructure to get rid of the huge container structs.
- That comitted we have only one place left where variables have to be defined. New macros. Only GENERIC and VIMAGE.
- Replacing more “vnet” infrastructure like using VNET_SYSINITS (jhb, rwatson) instead of a vimage internal dependency framework. libkvm support for vnets for netstat (rwatson, bz).
- Virtualization of new code that came in unvirtualized.
- Fix things that did not work properly like ipfw, flowtables (Julian, Marko, . . .).

Today

```
int answertoallquestions = 42;
```

becomes:

```
  #include <net/vnet.h>

  VNET_DEFINE(int, answertoallquestions) = 42;
  #define V_answertoallquestions  VNET(answertoallquestions)
```

and as this is not file local static in the header file you change:

```
extern int answertoallquestions;
```

becomes:

```
VNET_DECLARE(int, answertoallquestions);
```

More of today’s way of V_irtualizing

Change all uses of

```
answertoallquestions
```

to

```
V_answertoallquestions
```

and that variable is virtualized!

Pitfalls:

- Pre-initialization of lists or uma zones, . . . need an initializer function.
- Be carefull with locks (keep them global where possible).
- Timers, callouts, ...

More of today’s way of V_irtualizing

Initializer functions?

- SYSINIT / SYSUNINIT:
  runs once for the “base”
- VNET_SYSINIT / VNET_SYSUNINIT:
  runs once for the “base” and for every “vnet”.
- Use those to (de-)initialize things from functions.

What about SYSCTLs?

- SYSCTL_VNET_<type> exist:

```
SYSCTL_VNET_INT(_hhg, OID_AUTO, anser_to_all_questions,
  CTLFLAG_RW, &VNET_NAME(answertoallquestions), 0, "42");
```

More of today’s way of V_irtualizing

What else?

- CRED_TO_VNET / TD_TO_VNET / P_TO_VNET
- IS_DEFAULT_VNET
- CURVNET_SET / CURVNET_SET_QUIET / CURVNET_RESTORE
- VNET_ASSERT
- Julian’s “primer” in p4.
- Wiki pages updated any millenium now.

How to test it out?

Compile a kernel with `options VIMAGE` (and remove SCTP).

- jail -c (create), jail -r (remove), jls (-s) list jails. See man pages.
- option “vnet” to jail -c gives you a network stack (not in man page).

Create 3 jails with a network stack:

```
jail -c vnet host.hostname=lefty.example.net path=/ persist
jail -c vnet host.hostname=middy.example.net path=/ persist
jail -c vnet host.hostname=righty.example.net path=/ persist
jexec 2 sysctl net.inet.ip.forwarding\=1
jexec 2 sysctl net.inet6.ip6.forwarding\=1
```

And networking?

```
ifconfig epair20 create
ifconfig epair20a vnet 1
jexec 1 ifconfig lo0 127.0.0.1/8
jexec 1 ifconfig epair20a inet  192.0.2.1/30 up
ifconfig epair20b vnet 2
jexec 2 ifconfig lo0 127.0.0.1/8
jexec 2 ifconfig epair20b inet  192.0.2.2/30 up
jexec 2 ping -c 3 192.0.2.1
```

or use netgraph nodes. You can route, bridge etc. between vnets.

Why not XYZ?

- Still lightwight jails.
- Very low memory footprint (180k atm. for a jail+netstack)
- Being able to run services from Megabytes of storage rather than 100 MBs or GBs like a full OS installation inside other VMs.
- 1 Kernel, 1 Scheduler, no heavy switching between host and guest VMs, ...
- . . . .

Near future

- Post-lunch session in FW09.
- More hands - it's not hard.
- vnet allocator / vstorage (Robert Watson)
- Jails and priviledge sets.
- Test more things like IPv6, MC, IPsec, . . . , not just v4.
- Fix bugs, virtualize missing parts, http://wiki.freebsd.org/Image/TODO .
- New management interface.
- Update more Dokumentation.

Far future

- Be able to fully support and ship it with 8.2.
- More subsystems.
- jailinit(8).
- Whatever ideas you can come up with.

Lessons learnt so far?

- Developers are no full-time employees so RL interfers with schedules.
- Things take longer than you plan for.
- Too many developers say “not my playground”.
- Not convinced if “break the tree for a week or two” would have been better?
- It takes FreeBSD too long to adopt and merge larger 3rd party work in.
- ....