

Paul Schenkeveld

"The bhyve hypervisor"

OPENFEST

November 2, 2013



BSDEurope.eu



The European BSD gateway



About the speaker



Paul Schenkeveld

> 30 years UNIX experience

Independent consultant since 1991

FreeBSD since 1.0



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway

Paul Schenkeveld - 2 November 2013

Copyright (c) 2011 - 2013 BSDEurope.eu - printed: Sun Nov 3 15:51:15 CET 2013

Notes:

Hypervisors

Robert P. Goldberg, 1973:

"Architectural Principles for Virtual Computer Systems"

- Type 1 native/bare metal hypervisor (VMware, Virtualbox)
- Type 2 hosted hypervisor (Xen, KVM)

Gerald J. Popek and Robert P. Goldberg, 1974:

"Formal Requirements for Virtualizable Third Generation Architectures"

Properties of a Hypervisor:

<http://en.wikipedia.org/wiki/Hypervisor>

[http://en.wikipedia.org/wiki/\](http://en.wikipedia.org/wiki/Popek_and_Goldberg_virtualization_requirements)

[Popek_and_Goldberg_virtualization_requirements](http://en.wikipedia.org/wiki/Popek_and_Goldberg_virtualization_requirements)



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

Properties

- Equivalence / Fidelity

A program running under the VMM should exhibit a behavior essentially identical to that demonstrated when running on an equivalent machine directly.

- Resource control / Safety

The VMM must be in complete control of the virtualized resources.

- Efficiency / Performance

A statistically dominant fraction of machine instructions must be executed without VMM intervention.



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

Landscape

We already have many hypervisors:

- Proprietary and dominant: VMware
- Linux: KVM/QEMU Hypervisor / LXC Containers
- SmartOS: KVM Hypervisor / Zones
- FreeBSD: Xen EC2 / jail(8)
- Honorable mention: NetBSD Xen
- Virtualbox
- Parallels (MacOS)
- Hyper-V



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

Do we need another Hypervisor?

MeetBSD California 2010

- 100% Virtualization session attendance
- Follow-up session the next day

Conclusion:

- We need a BSD hypervisor!



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

bhyve - The "BSD HyperVisor"

Written by:

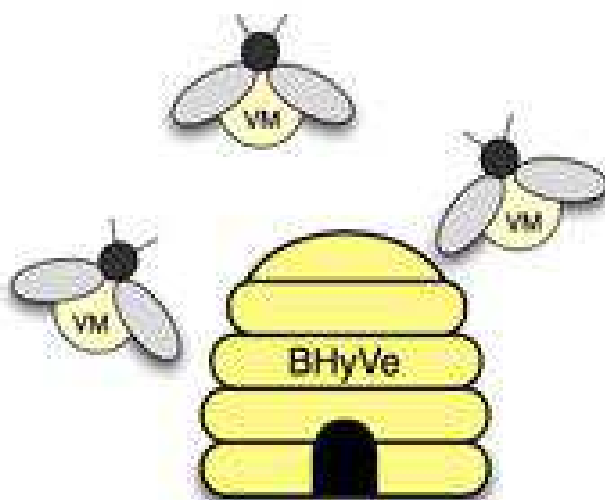
- Neel Natu
- Peter Grehan

First announcement:

- BSDcan 2011 Devsummit

Slides and audio are online:

- <http://wiki.freebsd.org/201105DevSummit?action=AttachFile&do=view&target=BHyVe.pdf>
- <http://www.bsdcan.org/2011/audio/BHyVeNativeBSDHypervisor.mp3>



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

bhyve - The "BSD HyperVisor"

- Requires Intel Extended Page Tables (EPT)
- Guests are booted from disk images
- Easiest to test on 9.0 and 10-BETA (Not 8.x or 9.1)
- 8.3 guests feasible, 9.0 and newer supported
- **Support for OpenBSD and Linux guests!**
- Works with VMware Fusion VT-x pass-through (on Mac)
- Imported to HEAD (10-CURRENT) January 2013
- **Will be in upcoming 10.0 RELEASE**



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

Hardware-Assisted Virtualization

”It’s all built on VT-x exits. I/O exits are used to build the PCI emulation since I/O instructions are used for PCI config space. VT-x sets up state to enter/exit "non-root" mode, aka VM mode.

EPT-violation exits are used for memory-mapped I/O. Guest physical memory is set up in EPT tables. Guest access to anything else causes an EPT-violation exit. Then instruction emulation is used to determine what is written and where reads should go.”

– Peter Grehan



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

Hardware-Assisted Virtualization

- VT-x: Virtualization Extensions:
Interception of privileged instructions
- VT-d: Virtualization for Directed I/O:
IOMMU Virtualization / PCI Pass-Through
- EPT: Extended Page Tables: MMU Virtualization
Previously handled in software
- AMD-V, AMD-Vi and RVI/Nested Page Tables Planned

http://en.wikipedia.org/wiki/X86_virtualization

http://en.wikipedia.org/wiki/Extended_Page_Table



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

Does my hardware support bhyve?

Look at the dmesg(8) output:

```
Features=0xbfebfbff<FPU,VME,DE,PSE,TSC,MSR,PAE,  
MCE,CX8,APIC,SEP,MTRR,PGE,MCA,CMOV,PAT,PSE36,  
CLFLUSH,DTS,ACPI,MMX,FXSR,SSE,SSE2,SS,HTT,TM,PBE>  
Features2=0x17bae3ff<SSE3,PCLMULQDQ,DTES64,MON,  
DS_CPL,VMX,SMX,EST,TM2,SSSE3,CX16,xTPR,PDCM,  
PCID,SSE4.1,SSE4.2,x2APIC,POPCNT,TSCDLT,AESNI,  
XSAVE,AVX>
```

- POPCNT Accompanies EPT



BSDDeurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

Hardware-Assisted Virtualization

Most, if not all "Nehalem", "Sandy Bridge" and "Ivy Bridge" Core, Xeon, Pentium and Celeron Processors:

- EPT: Intel Core i3, i5, i7 Processors
- EPT: Most same-generation Xeon Processors
- EPT: Some Pentium Mobile and Celeron Processors (!)
- VT-d: Many Core i5, i7 and Xeon Processors
- Not on Atom processors

Caveat: It may be disabled in BIOS or blocked entirely



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

Hardware-Assisted Virtualization

ark.intel.com is your friend:

Intel® Pentium® 2020M Processor (2M Cache, 2.40 GHz)

Advanced Technologies	
Intel® Turbo Boost Technology	No
Intel® vPro Technology	No
Intel® Hyper-Threading Technology	No
Intel® Virtualization Technology (VT-x)	Yes
Intel® Virtualization Technology for Directed I/O (VT-d)	No
Intel® VT-x with Extended Page Tables (EPT)	Yes



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

as(1) Implications

- The GNU binutils assembler with the EPT instructions is GPLv3-licensed and will not be included in BASE
- Revision 238123: Add support for the `xsave`, `xrstor`, `xsaveopt`, `xgetbv`, and `xsetbv` instructions. I reimplemented this from scratch based on the Intel manuals and the existing support for handling the `fxsave` and `fxrstor` instructions. This will let us use these instructions natively with GCC rather than hardcoding the opcodes in hex. -- jhb@
- Revision 238167: Add support for the `invept` and `invvpid` instructions. Beyond simply adding appropriate table entries, the assembler had to be adjusted as these are the first non-SSE instructions to use a 3-byte opcode (and a mandatory prefix to boot). -- jhb@

Openfest 2013 - bhyve



BSDEurope.eu



The European BSD gateway



Notes:

The Versioning Moving Target

- *bhyve* began life in FreeBSD 8.1
- Is manageable in FreeBSD 9.0
- FreeBSD 9.1 suffers from AVX floating point changes
- FreeBSD 10 projects/bhyve is^W was the official home
- *bhyve* is included in FreeBSD-10 since January 2013
- **Will be in 10.0 RELEASE**

My recommendation:

- Use a FreeBSD 10.0BETA2 or later
- Just works out-of-the-box



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

bhyve Host Components

- `/usr/sbin/bhyve`
the user-space sequencer and I/O emulation
- `/usr/sbin/bhyveload`
the user-space FreeBSD loader
- `/usr/sbin/bhyvectl`
a utility to dump hypervisor register state
- `/usr/lib/libvmmapi.{a,so*}`
the front-end to the `vmm.ko` chardev interface
- `/boot/kernel/vmm.ko`
kernel module for VT-x, VT-d and hypervisor control
- `/boot/kernel/if_tap.ko`
not part of `*bhyve*` but needed for networking



BSDDeurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

bhyve Guest Kernel Components

Your guest system needs the following kernel modules to interact with the hypervisor:

- device virtio # Generic VirtIO bus (required)
- device virtio_pci # VirtIO PCI device
- device vtnet # VirtIO Ethernet device
- device virtio_blk # VirtIO Block device
- device virtio_scsi # VirtIO SCSI device
- device virtio_balloon # VirtIO Memory Balloon device

These are included in GENERIC since `*bhyve*` import but can also be loaded as modules.



BSDeurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

Host Preparation

- Add these lines to `/boot/loader.conf`:

```
# Suppress noise when running on -CURRENT
# (or build non-debugging kernel):
debug.witness.watch="0"

# Load modules (loadable with kldload too):
vmm_load="YES"
if_tap_load="YES"
bridgestp_load="YES"
if_bridge_load="YES"
```

- Or load `vmm`, `if_tap`, `bridgestp` and `if_bridge` using `kldload(8)`



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

DIY bhyve

- Create the tap0 interface `ifconfig tap0 create`
- Download `vmrun.sh` and `release.iso` from <http://people.freebsd.org/~neel/bhyve>

Copy both into the same directory in which you have write permission.

Please be aware that my `release.iso` is most likely stale and you can get the latest `release.iso` here:

```
ftp://ftp.freebsd.org/pub/FreeBSD/\n    snapshots/amd64/amd64/ISO-IMAGES/10.0/
```



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

DIY bhyve

- As root (or using `sudo`) execute the following commands:
`./vmrun.sh vm1`
- Select default console type `vt100` or whatever is appropriate for you. Install on disk device `vtbd0` (appears as a 8GB disk device). At the end select `yes` when the "Manual Configuration" box appears.

Type in the following:

```
cat >> /etc/ttys << EOF
console "/usr/libexec/getty std.9600" vt100 on secure
EOF
```

And then reboot

- Enjoy your virtual machine



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

DIY bhyve

And then you see:

```
Launching virtual machine "vml" ...
Consoles: userboot
```

```
FreeBSD/amd64 User boot, Revision 1.1
(paul@bhyve.psconsult.nl, Sun Jan 27 17:53:58 CET 2013)
Loading /boot/defaults/loader.conf
/boot/kernel/kernel text=0xd00880 data=0x15ed20+0x2bbc90 syms=[0x8+0x14c2c8+0x8+0x1a8066]
/boot/kernel/virtio.ko size 0x5a50 at 0x1810000
/boot/kernel/virtio_pci.ko size 0x6d68 at 0x1816000
/boot/kernel/virtio_blk.ko size 0x6a30 at 0x181d000
/boot/kernel/if_vtnet.ko size 0xd8e8 at 0x1824000
\
Hit [Enter] to boot immediately, or any other key for command prompt.
Booting [/boot/kernel/kernel]...
GDB: no debug ports present
KDB: debugger backends: ddb
KDB: current backend: ddb
Copyright (c) 1992-2012 The FreeBSD Project.
Copyright (c) 1979, 1980, 1983, 1986, 1988, 1989, 1991, 1992, 1993, 1994
The Regents of the University of California. All rights reserved.
FreeBSD is a registered trademark of The FreeBSD Foundation.
FreeBSD 10.0-CURRENT #23 r243595:243640: Tue Nov 27 17:11:16 PST 2012
neelnatu@neelnatu4:/usr/obj/usr/freebsd/projects/bhyve/sys/GENERIC amd64
WARNING: WITNESS option enabled, expect reduced performance.
CPU: Intel(R) Xeon(R) CPU E5620 @ 2.40GHz (2399.90-MHz K8-class CPU)
Origin = "GenuineIntel" Id = 0x206c2 Family = 0x6 Model = 0x2c Stepping = 2
Features=0x8fa3ab7f<FPU,VME,DE,PSE,TSC,MSR,PAE,CX8,APIC,SEP,PGE,CMOV,PAT,PSE36,DTS,MMX,FXSR,SSE,SSE2,SS,PBE>
Features2=0x80bee255<SSE3,DTES64,DS_CPL,SMX,SSSE3,CX16,xTPR,PDCM,PCID,DCA,SSE4.1,SSE4.2,x2APIC,POPCNT,HV>
AMD Features=0x2c100800<SYSCALL,NX,Page1GB,RDTSCP,LM>
AMD Features2=0x1<LAHF>
TSC: P-state invariant
real memory = 2147483648 (2048 MB)
avail memory = 2042015744 (1947 MB)
```



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

DIY bhyve

To stop your running guest:

- Type `reboot` or `shutdown -r` inside the guest
- Press `ESC` at the loader and type `quit`



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

Legitimate Concerns

"x86 virtualization is about basically placing another nearly full kernel, full of new bugs, on top of a nasty x86 architecture which barely has correct page protection. Then running your operating system on the other side of this brand new pile of sh*t.

You are absolutely deluded, if not stupid, if you think that a worldwide collection of software engineers who can't write operating systems or applications without security holes, can then turn around and suddenly write virtualization layers without security holes."

– Theo de Raadt



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Paul Schenkeveld - 2 November 2013

Copyright (c) 2011 - 2013 BSDEurope.eu - printed: Sun Nov 3 15:51:15 CET 2013

Notes:

Legitimate Concerns

"But *bhyve* code (kernel + userland libraries and utilities) adds only 250KB of source to the existing FreeBSD code base."

– Me

The *bhyve* heavy lifting is done by:

```
/usr/src/sys/amd64/vmm/intel/vmx.c  
(1,300 LOC)
```

– Michael Dexter



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

The Future

What's coming:

- ACPI Tables
- AHCI Device Emulation
- VirtIO MSIx Support

On the horizon:

- Takuya Asada's BIOS Emulation Work
Allows to run other operating systems as guest.
- **AMD Support - In progress**



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

The Future

On the horizon:

- Better Integration with Host Scheduler
- Memory Over-Commit
- Suspend and Resume
- Generalization of CPUID Features for Guest Portability
- Sparse Images (QCOW, VDI, VMDK, ZVOL?)
- Non-tap VirtIO Back-End



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

More info:

The main places for *bhyve* information are:

- <http://bhyve.org/>
- <https://wiki.freebsd.org/BHyVe>
- <http://callfortesting.org/bhyve/>
Here you can also find a package to install *bhyve* on FreeBSD-9 easily
- The FreeBSD virtualisation mailing list, subscribe at <http://lists.freebsd.org/mailman/listinfo/freebsd-virtualization>

Manual pages are not complete (yet)

- Any volunteers to help writing docs?



BSDeurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

Acknowledgements

Thank you:

- * Neel and Peter for this beautiful piece of code!
- * Michael Dexter for his endless testing, publications, presentations and scripts
- * The Openfest organisation for inviting this talk
- * Ken Thompson, Dennis Ritchie et. al. for giving us UNIX and C.
- * The CSRG team at Berkeley University for BSD.
- * The FreeBSD project for FreeBSD.
- * You, for not making me look like a fool talking to an empty room.
- * All others who I should thank for making this talk possible but who are not yet mentioned above.



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

Questions?

You should have asked: "How can I help?"



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway



Notes:

Final words

Thank you and "Happy hacking"!



BSDEurope.eu

Openfest 2013 - bhyve



The European BSD gateway

