SMOLBSD

Making NetBSD a fast(er) booting microvm

Emile 'iMil' Heitor - BSDCan 2024

$ whoami

- Emile *'iMil'* Heitor, from Valencia, Spain 🇪🇸
- Freelance 🌴
- Flying phobia ✈️
- Using NetBSD since 1998 🟧
- NetBSD Commiter since 2009 ✏️
- Initial author of the `pkgin` package manager 📦

HOW IT ALL BEGAN

- Passionated by small systems, NetBSD Live Key (2006)
- Back in 2016, sailor, container-like for NetBSD
- Intrigued by Firecracker
- Wanted to boot NetBSD with `qemu -kernel`
- https://imil.net/blog/posts/2020/fakecracker-netbsd-as-a-function-based-microvm/
- Trim up NetBSD binary kernel: mksmolnb
- Posts about FreeBSD booting from Firecracker

PVH?

- PVH Introduced by Xen in 4.4, PVHv2 in Xen 4.10
  - Starts kernel from a different entry point
  - informations are passed by the hypervisor
- Linux PVH boot from Qemu in 2019
- Firecracker PVH boot introduced in FreeBSD 14

PVH NOTES

- https://xenbits.xen.org/docs/unstable/misc/pvh.html
- New, simplified entry point for the kernel
- NetBSD has Xen PVH support since version 10
- `Error loading uncompressed kernel without PVH ELF Note`
- Getting to know `locore.S`

```
#define ELFNOTE(name, type, desctype, descdata...) \
    .pushsection .note.name, "a", @note ;
/* [...] */
ELFNOTE(Xen, XEN_ELFNOTE_PHYS32_ENTRY,   .long,  RELOC(start_xen32))
```

©Colin

GREAT SUCCESS!

After a couple of seconds...

```
KVM internal error. Suberror: 1
emulation failure
EAX=0000ffc8 EBX=00000000 ECX=00000000 EDX=00090000
ESI=00000000 EDI=00000000 EBP=00000000 ESP=0000ffb8
EIP=0000001e EFL=00010082 [--S----] CPL=0 II=0 A20=1 SMM=0 HLT=0
ES =0000 00000000 0000ffff 00009300
CS =a171 000a1710 0000ffff 00009b00
SS =0000 00000000 0000ffff 00009300
DS =0200 00002000 0000ffff 00009300
FS =0000 00000000 0000ffff 00009300
GS =0000 00000000 0000ffff 00009300
LDT=0000 00000000 0000ffff 00008200
TR =0000 00000000 0000ffff 00008b00
GDT=     00000000 0000ffff
IDT=     00000000 0000ffff
```

REMOTE GDB

- Where are we failing? Are we in the new entry point?
- `qemu -S -s` 🔌 `gdb target remote`

```
(gdb) p &start_xen32
$1 = (<text variable, no debug info> *) 0xffffffff8020b440 <start_xen32>
(gdb) b *0x20b440
```

AN INNOCENT COMMENT

in `locore.S`

```
/*
 * save addr of the hvm_start_info structure. This is also the end
 * of the symbol table
 */
movl  %ebx, RELOC(hvm_start_paddr)
movl  %ebx, %eax
addl  $KERNBASE_LO,%eax
```

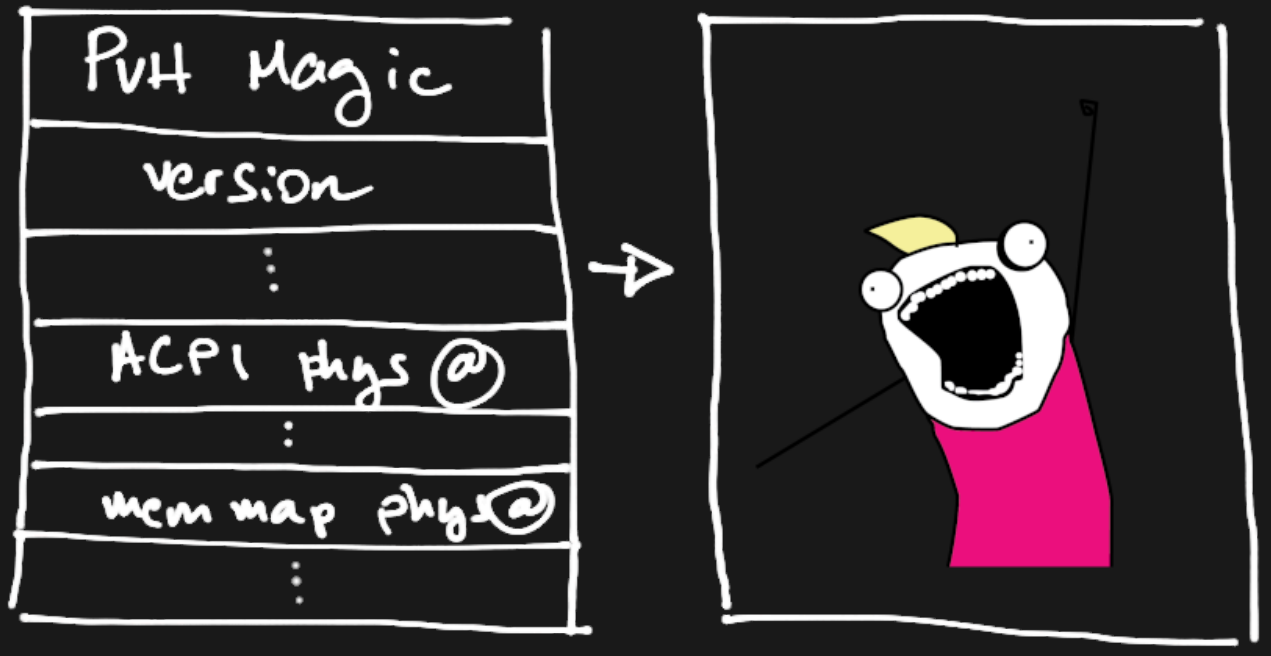
/* Now, zero out the BOOTSTRAP TABLES */

COPY ALL THE THINGS!

Where they are expected

EBX
0x21c0

__kern_end
0xa00000



S/XEN/GEN/

- Adding a new *VM* type: `VM_GUEST_GENPVH`
- Get console, physical memory and *ACPI* addresses from `start_info` structure
- Don't use *Xen*'s hypercalls


iMil
@iMilnb
IT BOOTS
FUCK!!!
qemu/PVH ON NETBSD BABY!!!!
12:25 PM · Dec 6, 2023 · 1,531 Views

VIRTIO/MMIO

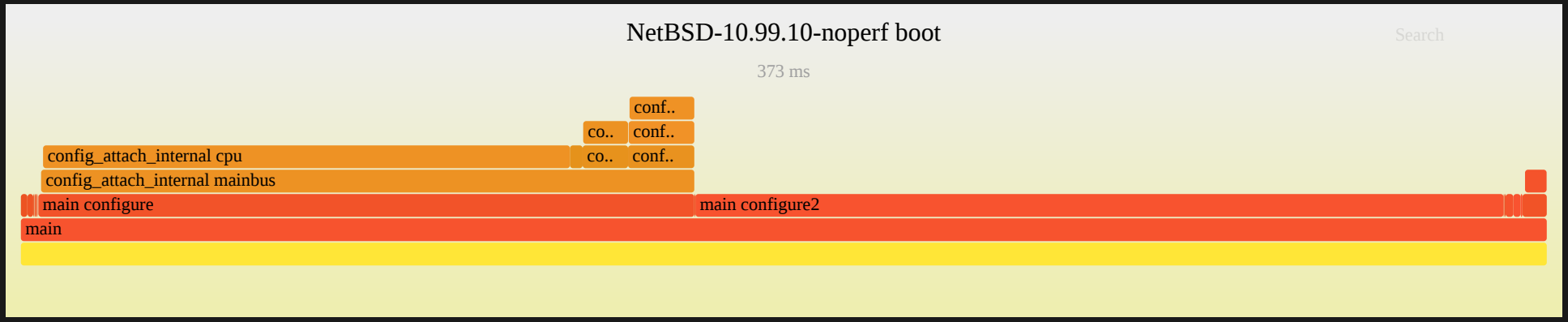
- *Firecracker* doesn't expose a PCI bus
- Both *Firecracker* and *Qemu microvm* support *MMIO*
- Device address and `irq` passed via kernel parameters
  - Ex: `virtio_mmio.device=512@0xfeb00e00:12`
- New virtual bus inspired from OpenBSD's `pv(4)`
- Ported Colin's `mmio_cmdline` driver

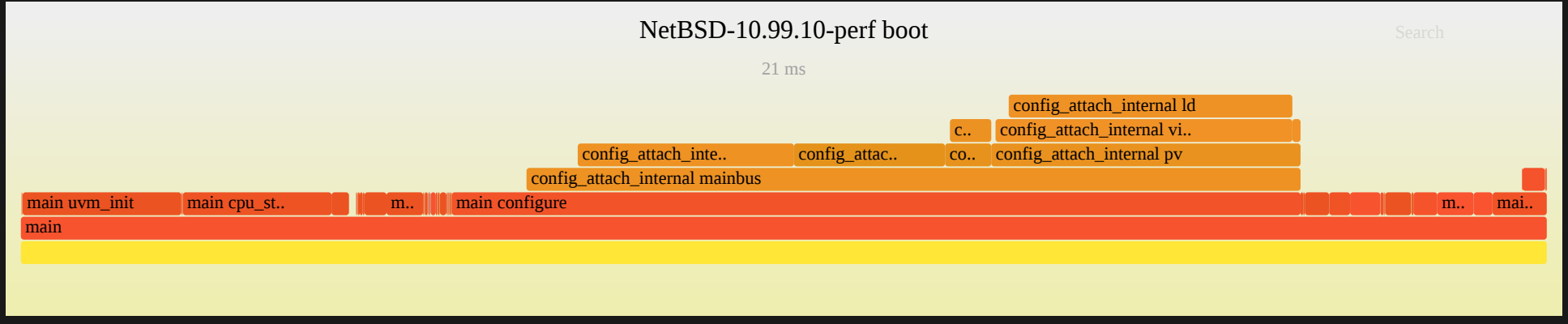
FASTER© --CRYO

- Get CPU frequency from `cpuid` or `rdmsr` instead of calibration loop
- Various *TSC* methods implemented
  - *VMWare* `cpuid 0x40000010`
  - *AMD MSR* (OpenBSD)
  - *Intel* read freq from brand (FreeBSD) ©Colin
- Import `pvclock(4)` from OpenBSD
  - Use it instead of `LAPIC` when possible
- Kill `DELAY()`
  - `lapic_calibrate_timer()` and `com_attach_subr()`

COLIN, COLIN EVERYWHERE

tslog(4)





DEMO!

QUESTIONS?