Using VirtIO for high speed container IPC with the Yocto Project and LXC

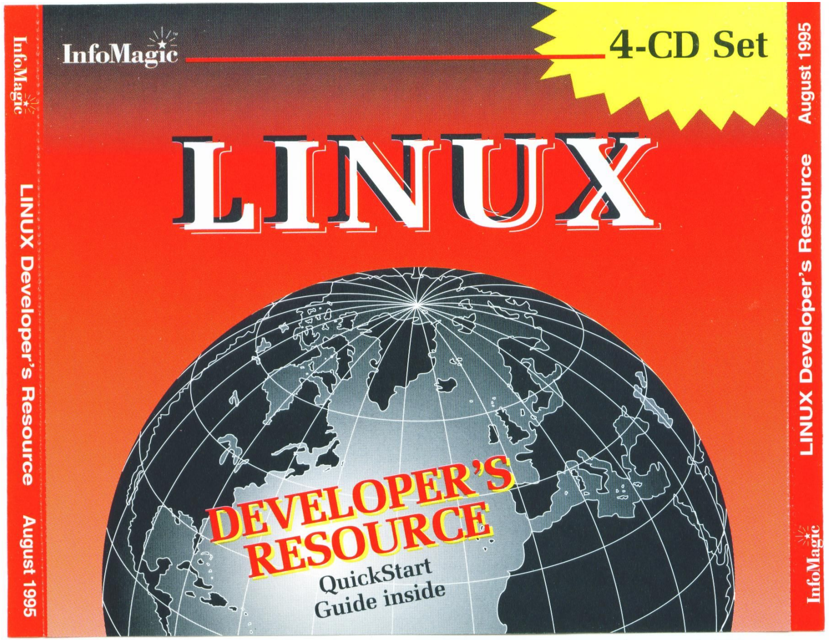
Eilís “pidge” Ní Fhlannagáin

Contents

- whoami
- Virtio in Virtual Machines
- Virtio with Containers?
- Yocto Setup
- Demo
- Conclusions

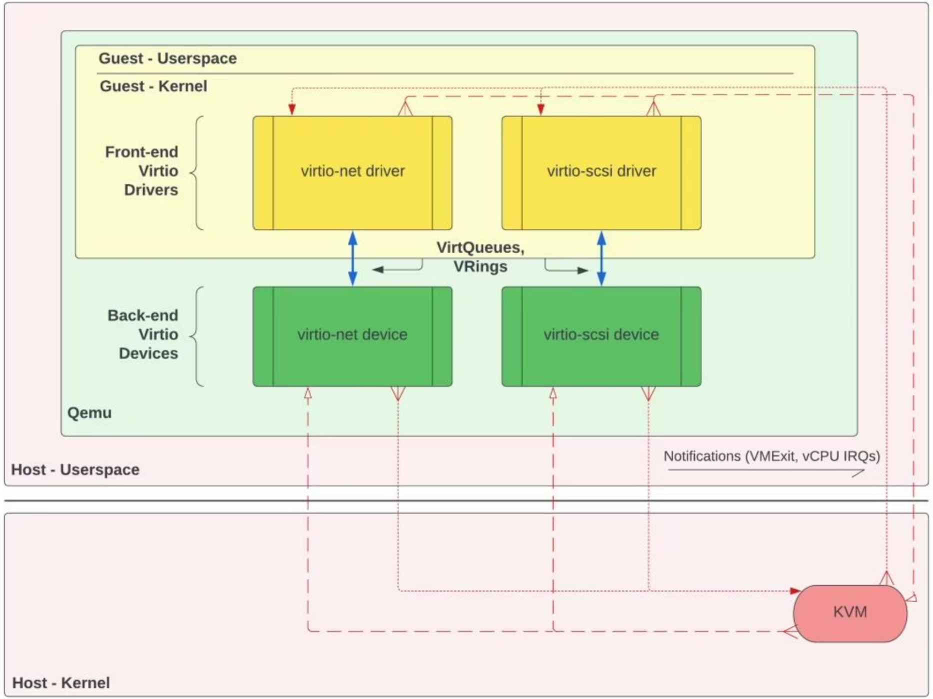
whoami

- Eilis (Eye Leash)
  - aka pidge
- Long time Linux User/Developer
  - Slackware pre-elf/pre-kmods
- Long time Yocto Project Contributor
  - Build statistics bbclass, licensing, SPDX
  - yocto-autobuilder
  - Oryx Linux/Network Grade Linux
  - Creator of weird YP demos
    - meta-zephyr midi glove
    - YP powered vielle á roue
- Baylibre
  - Mostly Yocto Project Things



Virtio for Virtual Machines
(overly simplified for embedded developers)

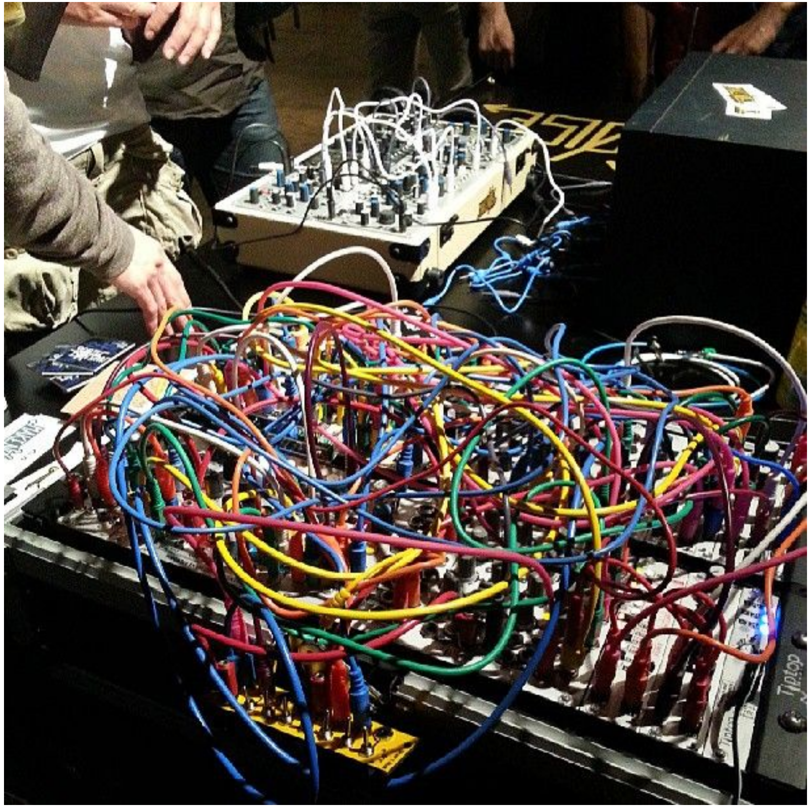
- Set of standards to provide virtualized interfaces to VMs
  - Virtual Devices in the hypervisor
  - Virtual Drivers in the guest
  - Data transport via virtqueue/vring
- virtqueues/vrings coordinate in guest memory
- VM process handles all the I/O however
- Data plane all within VM process
- All this is great BUT….



https://blogs.oracle.com/linux/post/introduction-to-virtio

Virtio for Virtual Machines

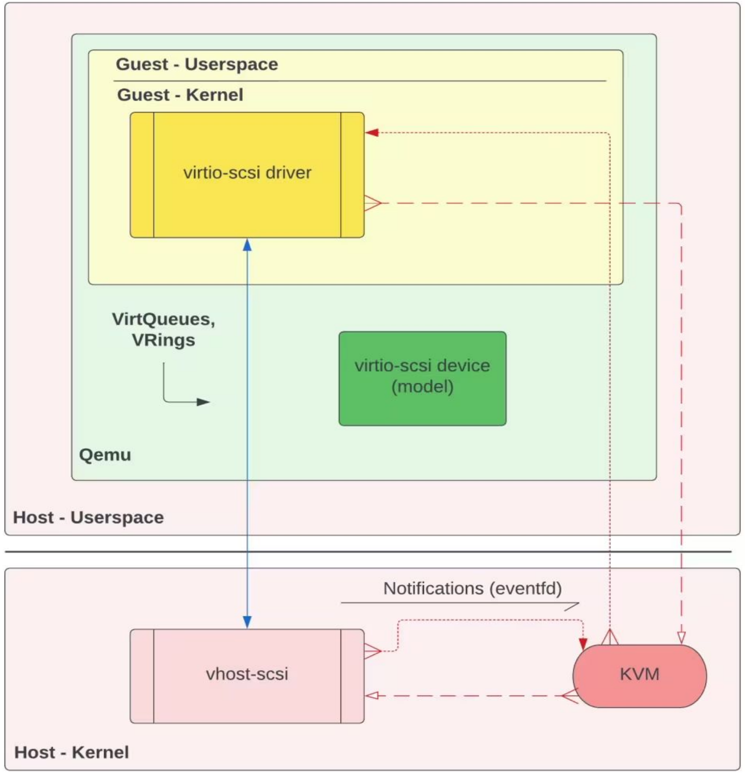
- Context switching/syscalls can be expensive
- Pushing a lot of data down the pipe
- If only there was an **Exceptional** way taking the **Data Path** and **Accelerating** it.
    - Exceptional Data Path Acceleration
    - Kernel Bypass
-



Spaghetti. Tokyo Festival of Modular 2013 - Kazuhisa Otsubo

Vhost-user in Virtual Machines

- Hardware is fast (sometimes), software (can be) slow
- Offload data plane to vhost-driver on host
- Virtio is still there but is just dealing with control plane
- Embedded devs see this with QEMU



https://blogs.oracle.com/linux/post/introduction-to-virtio

Untangling virtio/vhost* for embedded developers

- virtio
    - Standard for interfaces for virtual devices
    - Between VM and Hypervisor/host
- vhost
    - Guest Kernel bypass
- vhost-user
    - Guest to userspace vhost-user backend drivers
    - Bypass all the kernels
    - DPDK etc
- vhost-device
    - Same as vhost-user but kind of cooler
    - GPU passthrough/i2c passthrough etc
    - Mostly focusing on VMs
    - Jake Howard's GPU and LXC passthrough
        - https://theorangeone.net/posts/lxc-nvidia-gpu-passthrough/
    - Stratos vmm devices

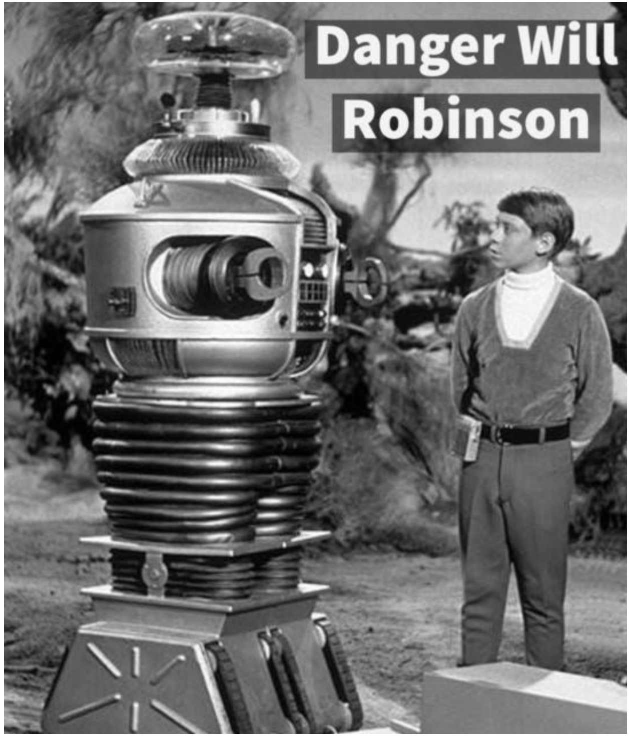
Wait?
This is a talk about containers and embedded, right?

“Can you implement virtio for my container?”

- Why? Isn’t virtio for VMs?
  - Sounds like a cheese submarine
  - But… wait…. why?
- Virtual Machines are obviously not containers
  - No Guest Kernel.
  - Data goes same places data always went
- Remember what virtio and vhost user can do?
  - Hardware abstraction layer
  - Kernel bypass
    - Speed up data path
  - Interesting for things that need Data NOW

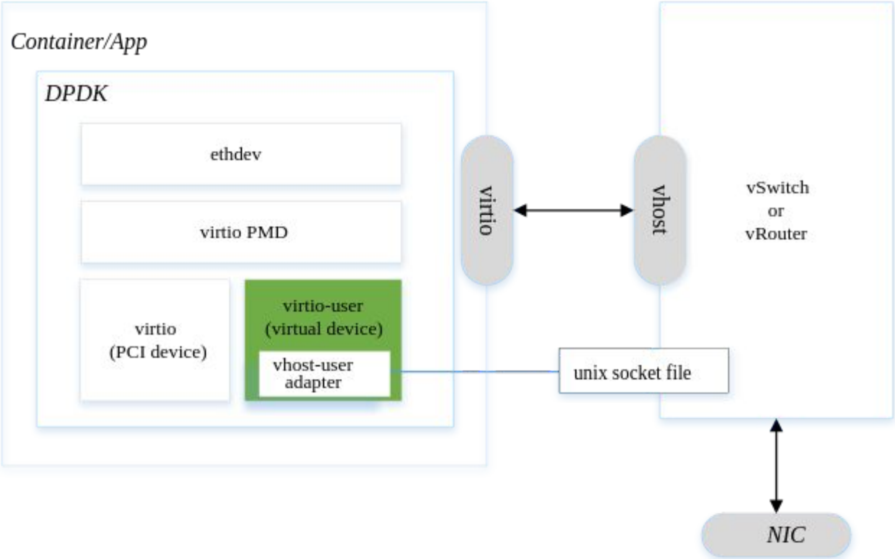
Why vhost in containers for embedded?

- Hardware Abstraction for Containers
  - I no longer care (as much) what i2c/net/gpio chip exists on host
  - Virtualized access to hardware
  - Device tree, vendor kernels, I don’t need to care as much
  - Con: Not as much frontend support
    - Write your own guest drivers!
- Exceptional Data Path Acceleration
  - Most everything in userspace
  - Poses “some” security concerns
    - Physical access
    - Work towards mitigation



What is DPDK (Data Plane Development Kit)

- Linux Foundation Project
- Accelerate packet processing workloads
- Good documentation/support
- Good example of the technology
  - FD.io VPP
  - OpenVSwitch



Why LXC?

- Pros
  - Automotive Grade Linux uses it
  - Lightweight
- Cons
  - Not as widely used as other container runtimes
  - LXD/LXC or LXC?
    - Please don’t do this
    - Snaps
- Other better choices
  - OCI/runc
  - Docker

DON'T MAKE ME TAP THE SIGN.
STOP CIRCUMVENTING OS PACKAGE MANAGERS!

Virtio/Vhost/container support in the Yocto Project?

- meta-virtualization
    - Lxc support
    - Linaro rust vmm backend vhost-device drivers
        - Excited about these
    - Kernel config fragments
- meta-dpdk
    - DPDK
    - Split out by swold from meta-intel
        - Older YP releases might still need meta-intel
- Tying this all together with meta-lxc-dpdk

meta-lxc-dpdk

- Multiconfig container builds
    - Prepopulate image with container config and roofs
        - Based on work done by Paul Barker, Scott Murray, myself
        - meta-agl containers use this
    - Based off of what is in AGL (Scott Murray)
    - Might try hand at making this generic enough for upstream
- Kernel Config
    - Missing CONFIG_HUGETLBFS
        - And friends
    - Some of this work should be upstreamed to meta-virtualization
- Templates
    - MIA

meta-lxc-dpdk

- Issue with DPDK
  - COMPATIBLE_MACHINE:pn-dpdk = "none"
  - COMPATIBLE_MACHINE:pn-dpdk-module = "none"
  - Currently fixing in image but this is wrong!
- TODO:
  - Some DISTRO/image based configs shoved into local.conf
  - Cleanup and Release

Virtio Container Demo

- Quick layer walkthrough
  - Show how we do prepopulation of containers
- Bring up QEMU with LXC Container
  - Setup hugetblfs, etc
- Run testpmd on host using virtio/vhost, socket, hugetables
- Run testpmd on guest using virtio/vhost, socket, hugetables
- No Container to Container IPC (sorrrry)
  - Exercise for Reader

Layer Walkthrough/Demo

Conclusions

- Is it faster?
  - Yes, but….
  - “Root privilege is a must. DPDK resolves physical addresses of hugepages which seems not necessary, and some discussions are going on to remove this restriction.”
- Is it worth it?
  - Maybe, but….
    - Very specific use cases
    - Is it needed for your temp sensor?
      - Is your temp sensor for a rocket ship?
  - Lots of data that needs speed?
    - Yes
  - There might be better ways to do all of this

Contact/Links/etc

Email:
pidge@pidge.org
pidge@baylibre.com

meta-lxc-dpdk: https://git.yoctoproject.org/poky-contrib/log/?h=pidge/meta-lxc-dpdk

Thanks to:

- Dr. Luca Abeni: https://retis.santannapisa.it/luca/ (LXC/DPDK)
- Scott Murray/Paul Barker for mc::containers work
- Redhat's series on virtio/vhost etc https://www.redhat.com/en/blog/virtio-devices-and-drivers-overview-headjack-and-phone
- Oracle series of virtio https://blogs.oracle.com/linux/post/introduction-to-virtio
- Jake Howard for Nvidia GPU passthrough in LXC
    - https://theorangeone.net/posts/lxc-nvidia-gpu-passthrough/

Questions?