How to compile a Linux Kernel for your 100Gbps router

Or, how to keep doing Software Routing until you really have to start using an ASIC, with 100% open source software.

CLUG - 13 Jun 2023, Woodstock Brewery

Joe Botha 🖖

Past Co-founder:
Frogfoot - when it was an ISP
Teraco - DC
Octotel - FNO

Currently:
Atomic Access - Fibre ISP in Cape of Good Hope
www.atomic.ac
twitter.com/swimgeek

“ASICs, magic and pro-wrestling are closely guarded secrets”

Avoid ASICs with NDAs and buggy / closed SDKs

Run 100% Debian

My quick 20 year history with Linux Routing

2000 - Frogfoot's 1st router: x86 PC
2006 - Amobia x86 & wireless
2018 - Atomic software routing
2022 - Atomic ASIC routing

Software vs Hardware routing?

Read: ’The world in which IPv6 was a good design’

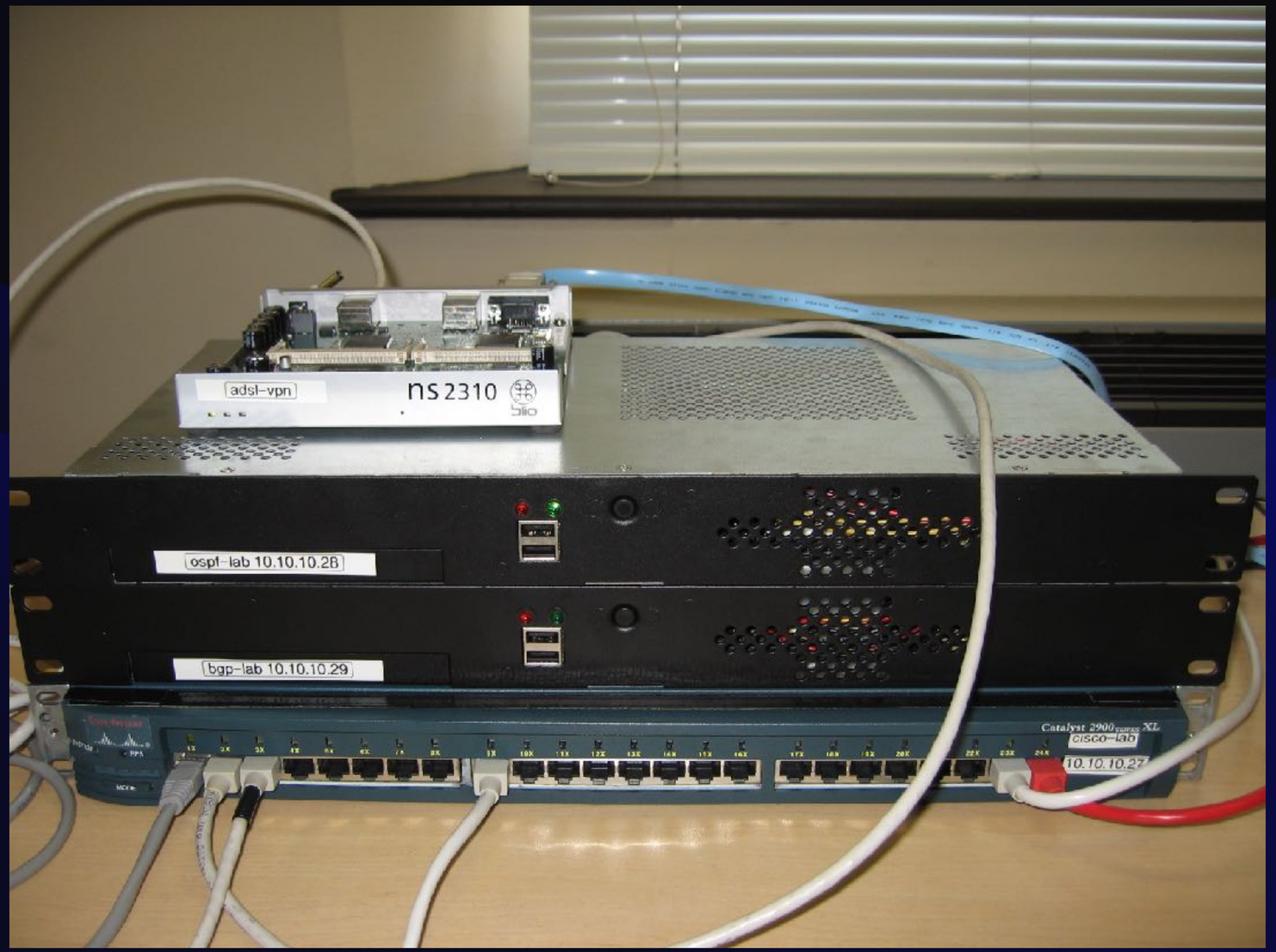
IETF vs IEEE & Routing vs Switching

tldr; Route packets with a CPU, until you can’t.

https://apenwarr.ca/log/20170810

2000 - Frogfoot's 1st routers: x86 PC 🐸

2006 - x86, but embedded / rackmount



Open Networking History

2013 - Cumulus & ONIE 😎

2016 - Mellanox Spectrum1 & SwitchDev

2019 - DENT and Marvell driver for Switchdev

2018 - Atomic software routing 🤨

- Xeon D & Intel i40e
- Debian 9
- FRR v3

vs commercial options / Arista 7280

2018 - Atomic software routing

2021 - Atomic recent routing 🧐

- Xeon D & Intel i40e & SR-IOV
- Debian 10 & Proxmox
- FRR v7

Limits: softIRQ/core
~8Gbps with 8 cores
Intel NIC drivers are not great
Port density is a problem

Software routing can’t scale, interrupts.

What now?

Searching for Open Networking

2018-2021 - found nothing really nice 😭

- DPDK & VPP, OVS, Vyos etc
- Broadcom, IPinfusion, OCNOS etc
- ONL and lots more…

> found Mellanox NICs & Switches

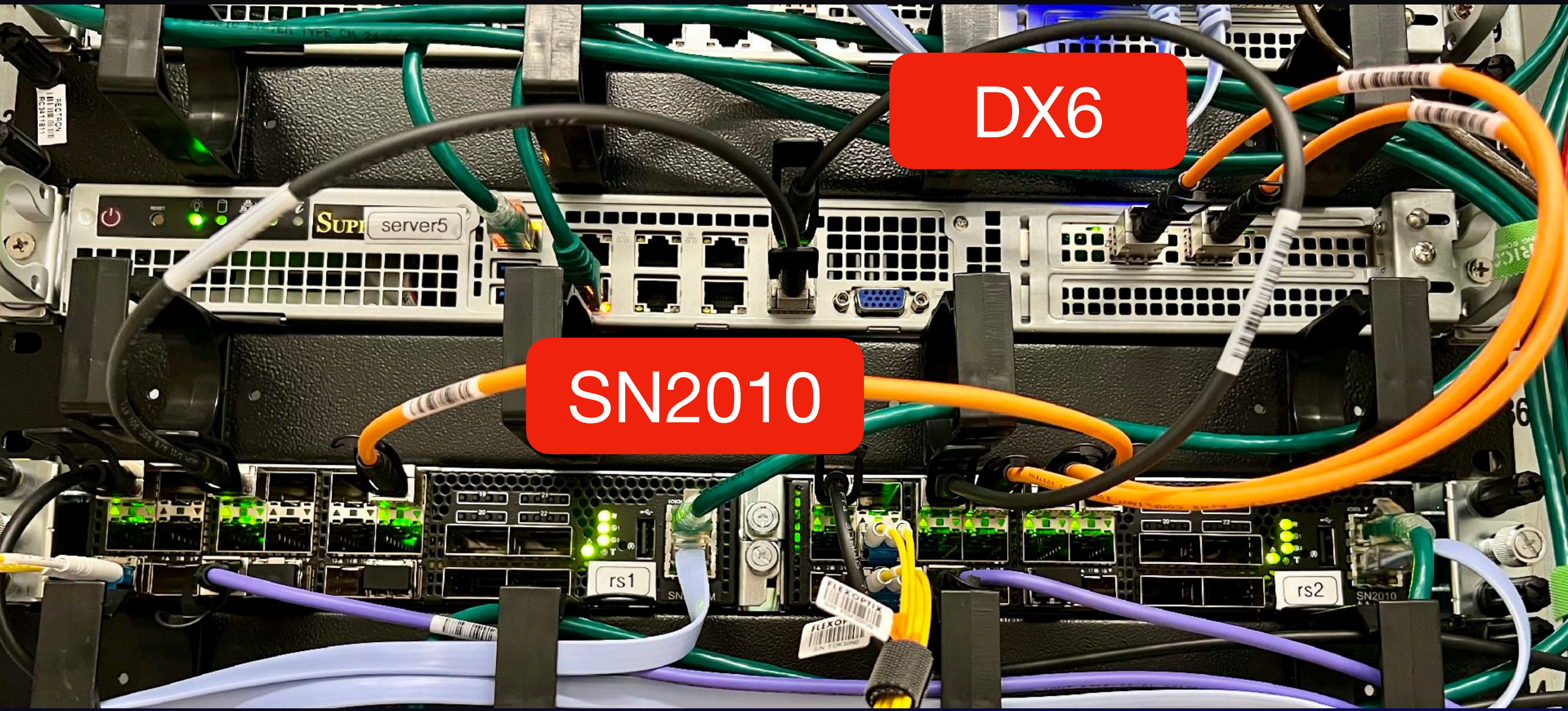
Could you build a 100Gbps router with Debian?

* Slide from 2019

2023 - Atomic routing, sw & hw 😊

- Border: Xeon D & Mellanox DX6 NICs
- Peering/BNG: SN2010 Mellanox NVIDIA

- Debian 12
- FRR v8.5

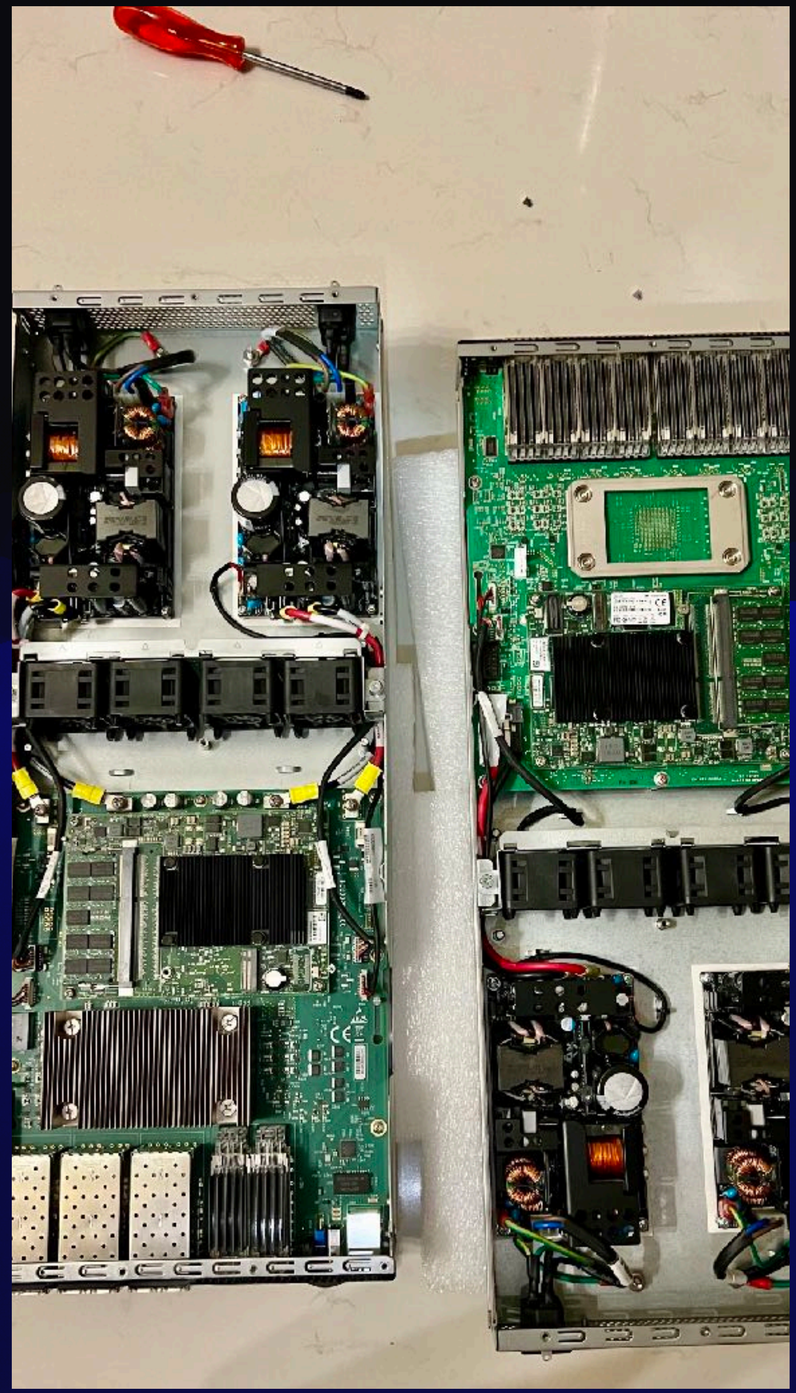

DX6
SN2010
server5
rs1
rs2
SN2010

What you need to compile:

- Linux kernel 6.1 with SwitchDev
- Various hw sensors
- hw-mgmt
- ethtool
- iproute2
- ifupdown2
- resmon
- psample

Pro tip: use equivs


SN2010



```
     /\
    /  \
   / /\ \ //\tomic //\ccess
  /_/\__/   www.atomic.ac

Last login: Mon Jun 19 19:57:00 2023 from 100.100.100.1
root@cpt-ter-rs1 ~ # cat /etc/debian_version
12.0
root@cpt-ter-rs1 ~ # uname -a
Linux cpt-ter-rs1 6.1.32-atomic #1 SMP PREEMPT_DYNAMIC Mon Jun  5 16:12:08 SAST 2023 x86_6
4 GNU/Linux
root@cpt-ter-rs1 ~ # ethtool -i swp1
driver: mlxsw_spectrum
version: 1.0
firmware-version: 13.2010.3146
expansion-rom-version:
bus-info: 0000:01:00.0
supports-statistics: yes
supports-test: no
supports-eeprom-access: no
supports-register-dump: no
supports-priv-flags: no
root@cpt-ter-rs1 ~ # _
```

Why? 😎

- purist ❤️ debian
- most open, no lame NDAs
- small, low power
- pretty good port density
- peering traffic at ASIC speeds
- can do cool things with tc rules

…not really cheaper

Switchd vs Switchdev

Old
ASIC managed by closed SDK
Linux is just driving the fans and LEDs
Userspace SDK app is running the front panel ports and ASIC
ASIC forwarding plane is not part of the OS

Better (switchd)
Create network devices in Linux
Make the SDK app listen for netlink messages
FRR, 'iproute2', 'ifupdown2' start working
Linux becomes the API / management interface

Open (switchdev framework + spectrum driver)
'ethtool' fully works
devlink to talk to ASIC via PCI bus
ASIC driver is now in the standard kernel
ASIC API is open

Specs 🤓 SN2010

- 180k IPv4, 30k IPv6, 8k MAC (256k)
- 10 / 25 / 40 / 100 Gbps
- 57 Watt power
- 8G memory, 256G storage (DIY)
- 1.3 Bpps
- 16MB buffers
- 300ns latency

What's missing? 😕

- Deep buffers
- 2M routes (DFZ)

Links

- Watch: Building a better NOS: https://www.youtube.com/watch?v=CfgjbHivdQ8

- Switchdev + Spectrum: https://github.com/Mellanox/mlxsw/wiki

- Switchdev + Marvell: https://github.com/Marvell-switching/switchdev-prestera

Questions? 😵‍💫

- Cumulus vs Debian
- Sectrum1 vs Spectrum2
- Buffers - all 10G ports
- Ops / upgrades