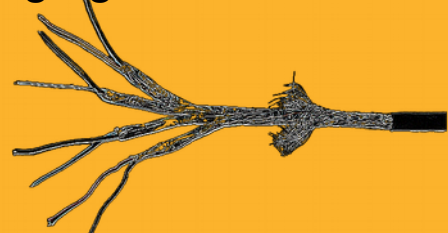
The long and Winding road to 10Gbit(+) in the home

& the current state of Network Connectivity options generally

- Joel Pauling | joel@aenertia.net | @aenertia – on the tweeters/IRC/g+/github/$etc

- LCA 2018 - 24-01-2017



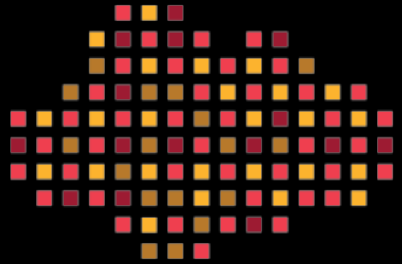
https://github.com/aenertia/lca2018-talk/tree/talk

Who am I

https://github.com/aenertia

@aenertia

https://plus.google.com/u/0/+JoelWir%C4%81muPauling

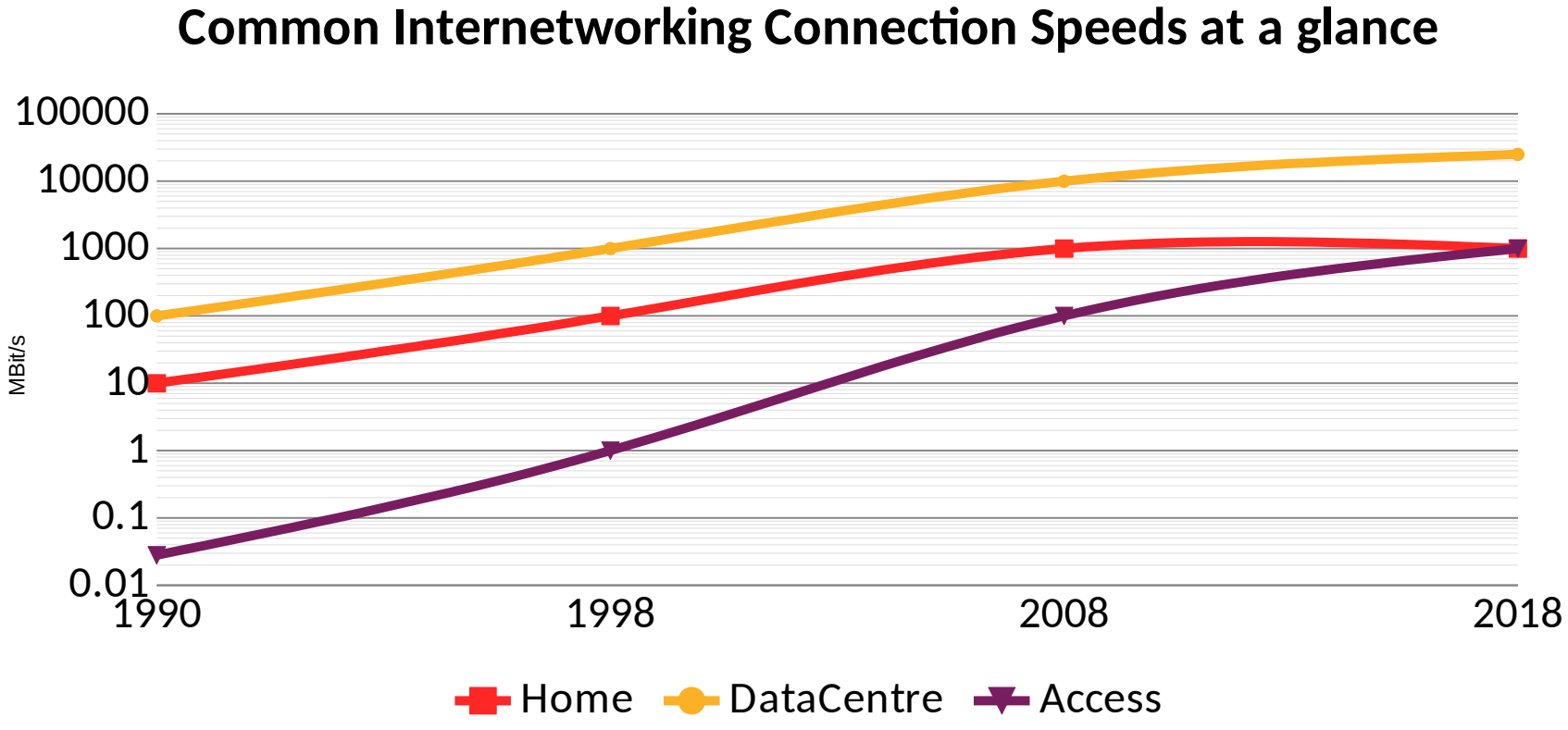


nuagenetworks

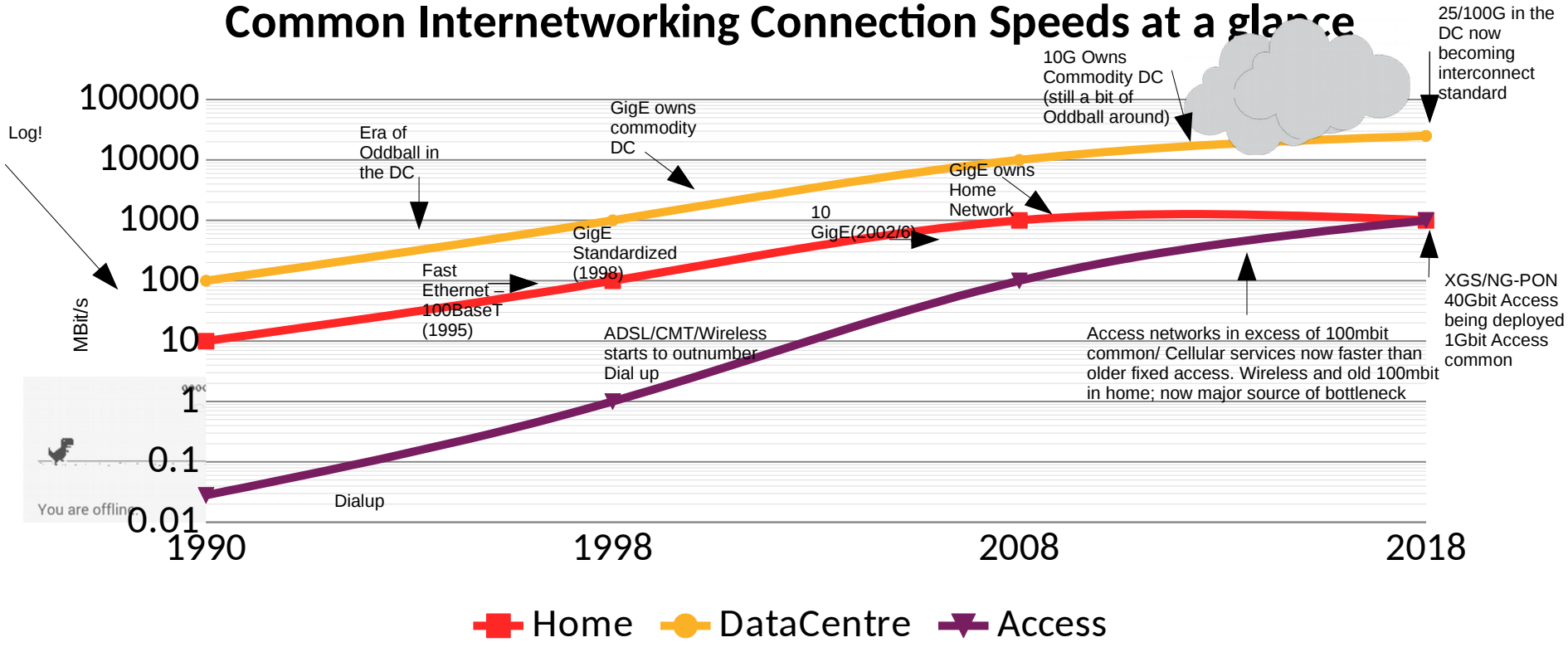
From Nokia

NOKIA

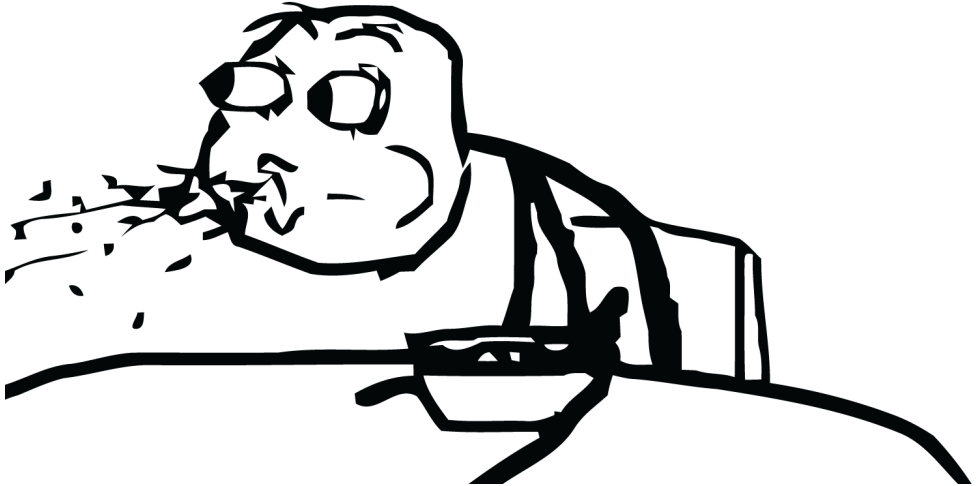
The Problem


Common Internetworking Connection Speeds at a glance
MBit/s
100000
10000
1000
100
10
1
0.1
0.01
1990
1998
2008
2018
Home
DataCentre
Access

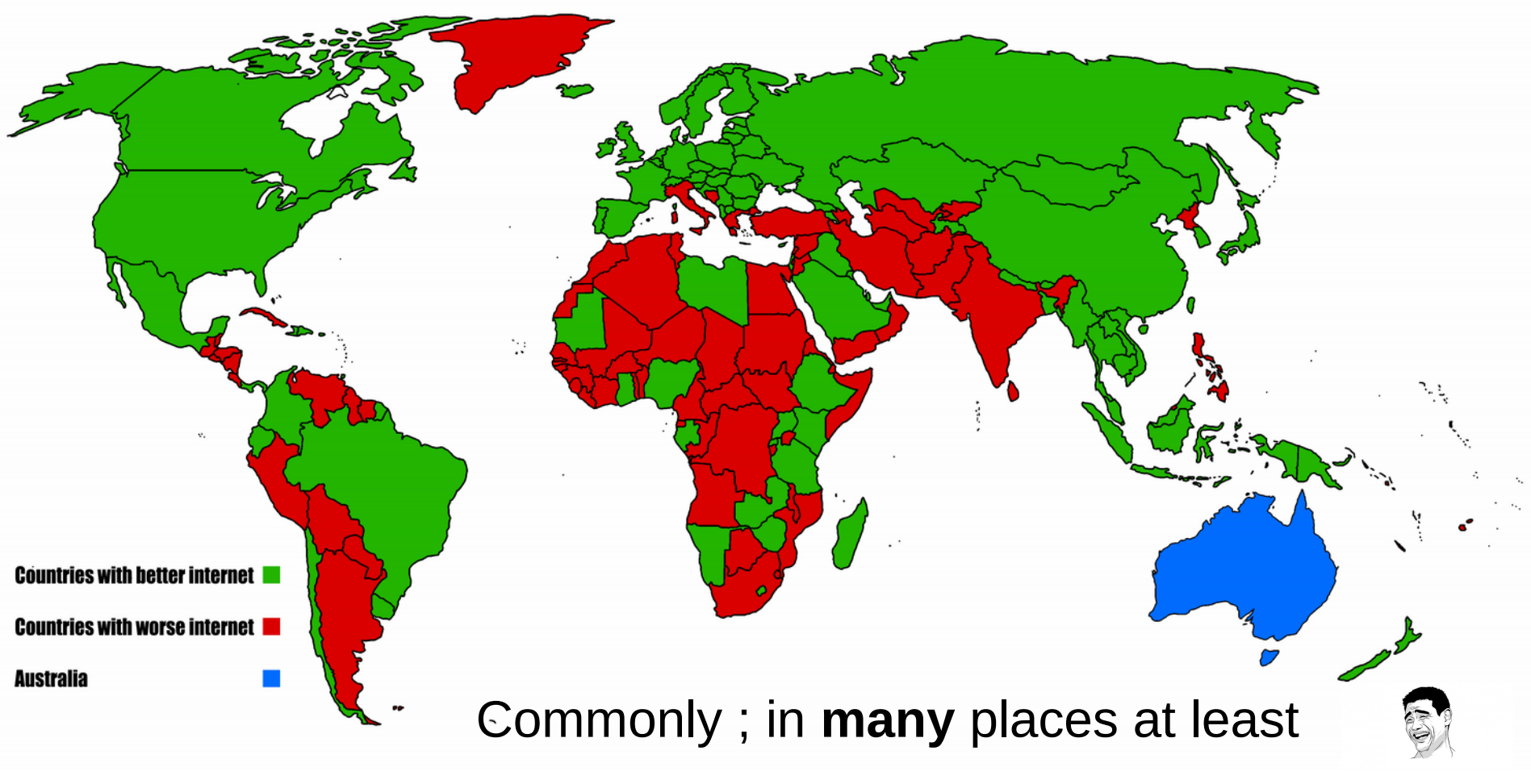
The Problem – Some more context



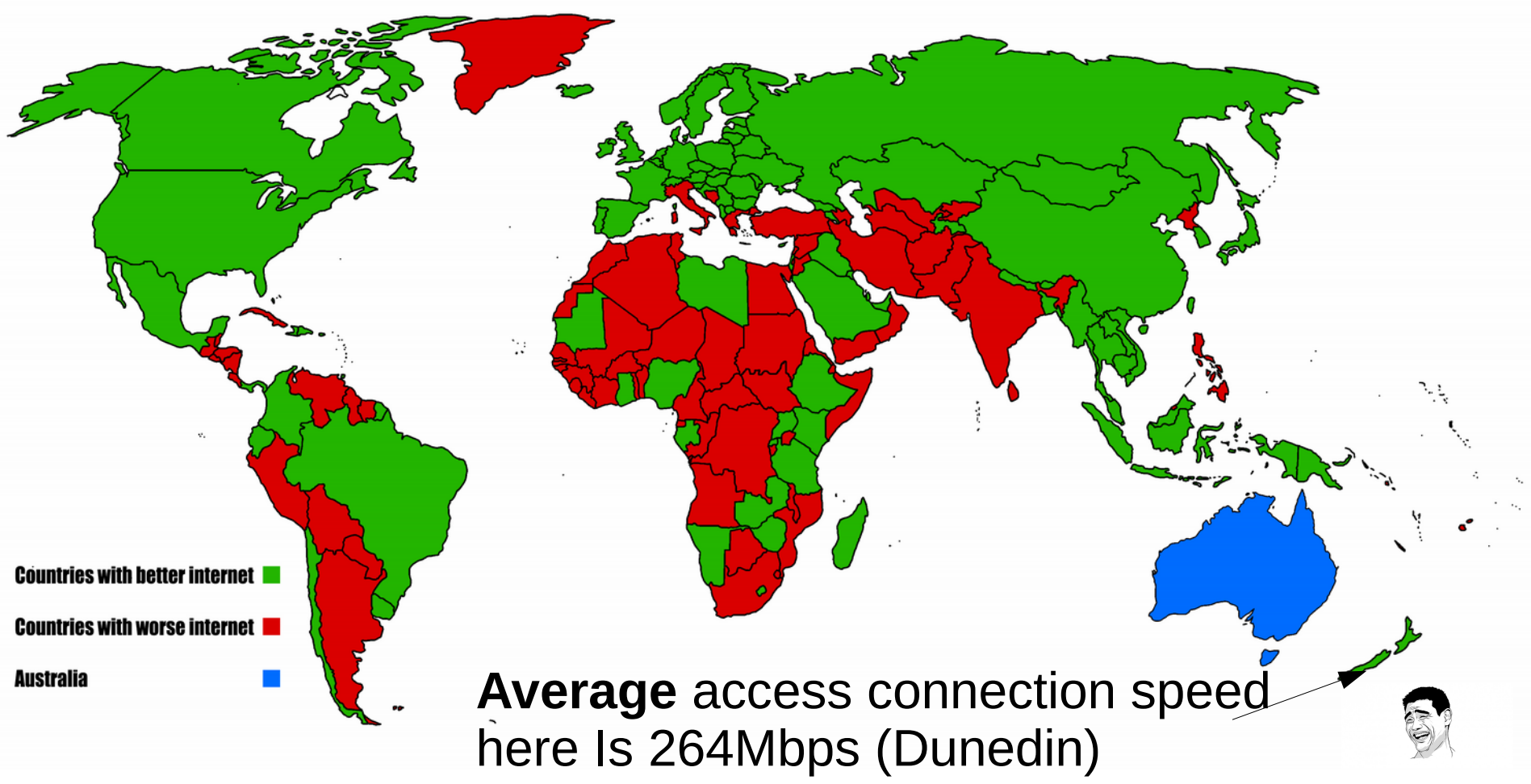
Access Network hand-off is now becoming commonly better than what Home Networks can Handle



If you had told me this 15 years ago I likely would have laughed



Commonly ; in **many** places at least

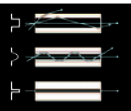


Average access connection speed here Is 264Mbps (Dunedin)

There are a lot of moving pieces to the Puzzle

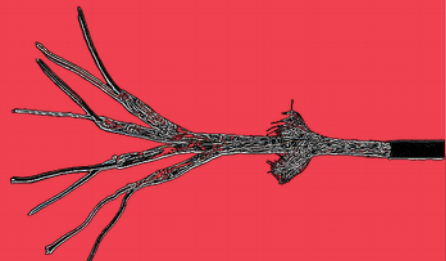
Cracks In the Path

- Wiring and Cable Connector Standards
- Architectures on the PC as well as SoC (System on a Chip) in Consumer Gateway kit being fit for purpose and more importantly having **fully usable** Open Architectures
- Latency and PPS (Packet Per Second) processing in the Mainline Kernel (is being worked on – this isn't really a problem for home networks… yet)
- Wireless and consumer perceptions and expectations have muddied the situation
- Consumer gear STILL using 100Mbit interfaces in brand new product I.e SmartTV's
- Power Consumption and Legacy compatibility
- Cost



In-home wiring standards

A non-exhaustive search of stuff i’ve found lying around in my own and others houses through the last 20 years

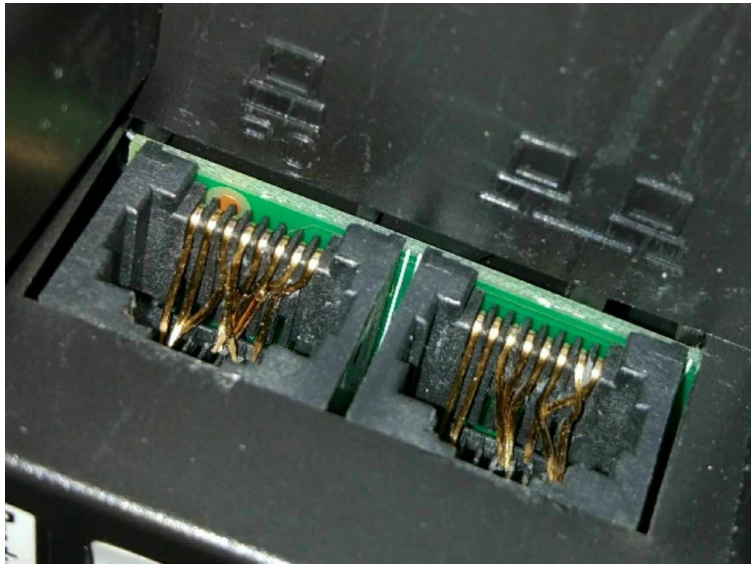


In home Structured Cabling Standard is just gone for review in New Zealand and Australia – Open submissions requested in Sept – unsure where this process is at the moment (but the document is closed paid for subscription, boo!)

AS/NZS 14763.3:2017

AS/NZS 3080

Strongly suggest anyone involved with Open Governance, or and or has an interest– try and figure out why this stuff is 300$ a copy and not open. And perhaps approach local representatives to look into it please.



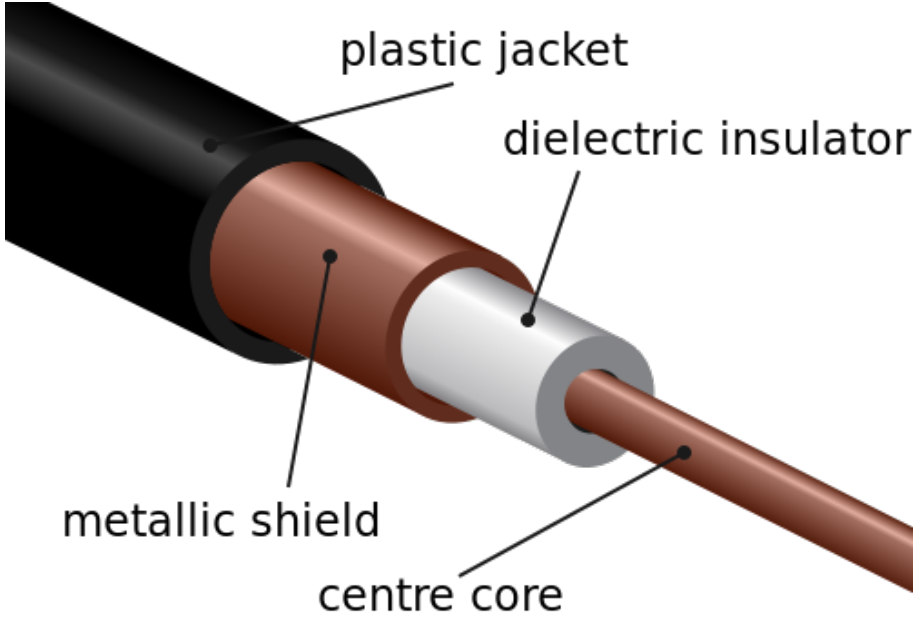
If your organisation is in a position to support this body of work could you, please send a letter or email of support to office@vti.net.au in order for your support to be included with the submission to Standards Australia.

The submission included the recommendation for the adoption of the following documents:

1. ISO/IEC 11801-1 Information technology -- Generic cabling for customer premises -- Part 1: General requirements
2. ISO/IEC 11801-2 Information technology -- Generic cabling for customer premises -- Part 2: Office premises
3. ISO/IEC 11801-3 Information technology -- Generic cabling for customer premises -- Part 3: Industrial premises
4. ISO/IEC 11801-4 Information technology -- Generic cabling for customer premises -- Part 4: Single-tenant homes
5. ISO/IEC 11801-5 Information technology -- Generic cabling for customer premises -- Part 5: Data centres
6. ISO/IEC 11801-6 Information technology -- Generic cabling for customer premises -- Part 6: Distributed building services
7. ISO/IEC 30129 Information technology -- Telecommunications bonding networks for buildings and other structures
8. ISO/IEC TS 29125 Information technology -- Telecommunications cabling requirements for remote powering of terminal equipment
9. ISO/IEC TR 11801-9902 Information technology -- Generic cabling for customer premises -- Part 9902: Specifications for End-to-end link configurations

Cables and Connectors - Coaxial

- Been around a Long time
- Was first widely used as Home cabling type for fast networks (BNC 2.9-10Mbit Eth/IPX) late 1980-1990's
- Bus architecture (signals share common medium/carrier)
- Widely used today still for Access Networks (Docsis/Cable Modems) in Hybrid Fibre/Coax networks
- Pretty damn good electrical scaling characteristics for a medium thats been around this long (10Gbit Docsis3.1 ratified 2017)
- Rarely seen today in home networks beyond the CPE interconnect. But it's still everywhere in homes (AV/etc)
- BUT ….

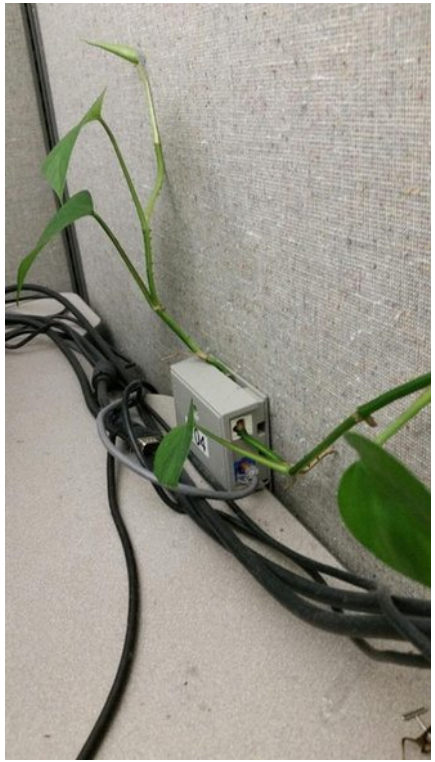


Cables and Connectors - Twinax

- TwinAx is the current undisputed owner of in-rack Network to Switch connection Technologies in the Datacentre
- It’s basically two (or more) of Coaxial cables in a single Jacket
- Short runs up to 8 Meters
- Happily supports 100Gbit Ethernet. 400Gbit CR4 is about
- Used to be common for some older oddball network Kit – seen resurgence during switch to 10Gbit mid 2000’s due to cost
- I’ve yet to see it in homes – short runs, low bend radius/can be pretty bulky – is coupled to SFP+/QSFP+ form factor Electro-Electrical generally (I.e the ‘optics’ are sold pre-attached to cable) – known as DAC (Direct Attach)
- Commonly used as ‘squid’ breakout (100G – 4*25G)

Cables and Connectors (s)UTP

- UTP is the Undisputed owner of Home and Commercial building cabling
- Cat5e is currently the standard most widely found
- Cat6a is needed for 10Gbit (although 802.3bz – allows Cat6 to work potentially at 2.5/5Gbit if tolerances are inadequate for 10Gbit)
- Cat7 passed over for 40Gbit …
- Cheap(ish)
- Familiar – tooling needed for 8P8C (rj45) crimping is common in toolboxes of geeks worldwide.
- Relatively durable depending on variant low bend radius/long runs



Cables and Connectors (s)UTP – Cat6a for 10Gbit?

- Cat6a for 10Gbit applications hasn't taken off
- Personal experiences are tolerances for 10G Base-T make Cat6A only really viable for greenfields and short in-rack runs; and then WHY not use it over TwinAx – which needs less power and has better tolerances whilst being cheaper?
- It's ratified for 100M runs at 10G … i've yet to see a stable 100M 10Gbit run in real world DC conditions (multiple patch ports/use of older non tiered 8P8C connectors)
- Too many little 'gotchas' needed to ensure it works
- In DC SFP+ Standard is out of Spec for what 10G BaseT Silicon needs to power it(changing now with move to 28nm silicon). So SFP+ Opto-electrical converters rare or out of spec. Whereas for 1Gbit it was easy to swap in an electrical 'optic' you can't easily do that with 10Gbit
- The cables are kinda bulky compared to cat5e
- No path beyond 10Gbit currently…
- In home Switch/Gateways?

Cables and Connectors (s)UTP – Cat6a for 10Gbit?

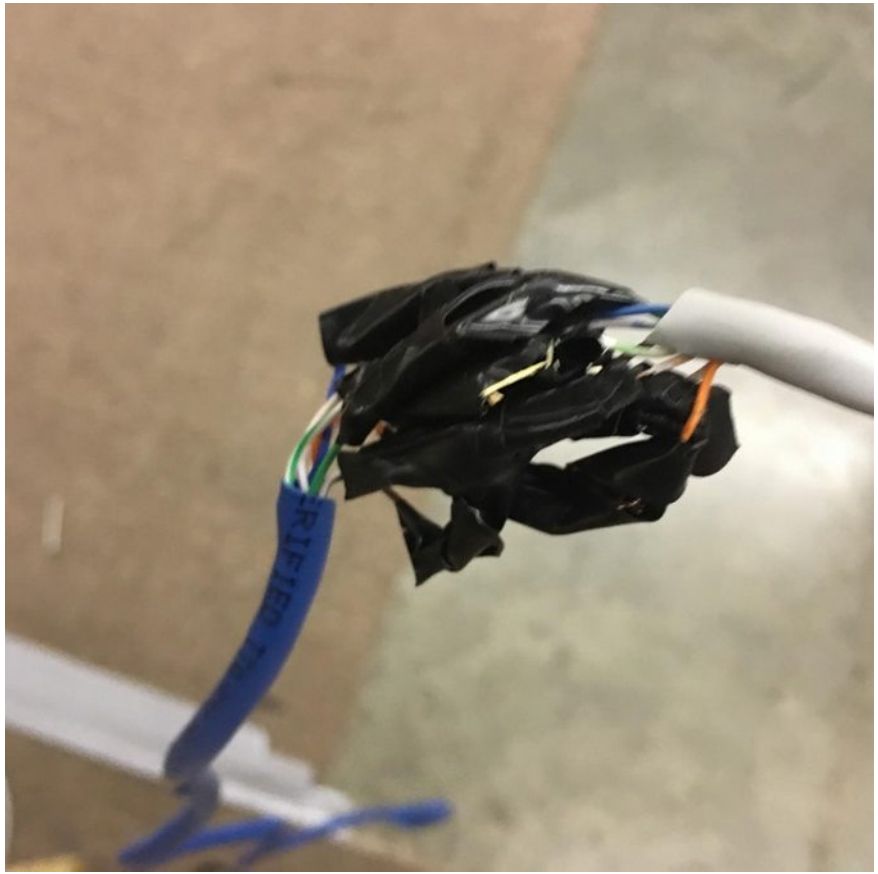
- Cat6a for 10Gbit applications hasn't taken off
- Personal experiences are tolerances for 10G Base-T make Cat6A only really viable for greenfields and short in-rack runs; and then WHY not use it over TwinAx – which needs less power and has better tolerances whilst being cheaper?
- It's ratified for 100M runs at 10G … i've yet to see a stable 100M 10Gbit run in real world DC conditions (multiple patch ports/use of older non tiered 8P8C connectors)
- Too many little 'gotchas' needed to ensure it works
- In DC SFP+ Standard is out of Spec for what 10G BaseT Silicon needs to power it(changing now with move to 28nm silicon). So SFP+ Opto-electrical converters rare or out of spec. Whereas for 1Gbit it was easy to swap in an electrical 'optic' you can't easily do that with 10Gbit
- The cables are kinda bulky compared to cat5e
- No path beyond 10Gbit currently…
- In home Switch/Gateways?

This is actually not quite in Spec. Why? Bonus what is different about this to older 8P8C ?

(s)UTP – the Future?

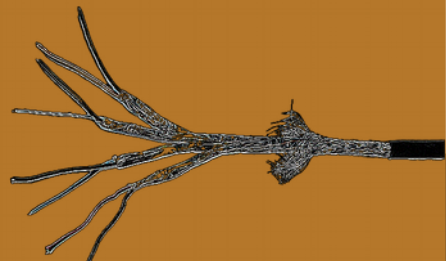
- No current path beyond 10Gbit
- 10Gbit needs 6A anyway
- Kinda weird we are going backwards with speeds on this format
- Cat8 is crazy; not for the home. Only s(UTP) currently ratified for anything higher than 10Gbit
- Maybe it’s time to ditch it?
- Fibre is actually cheaper…

(s)UTP – the Future?

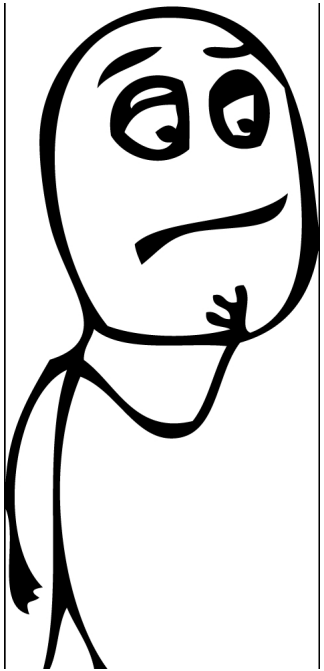


In Home Hardware

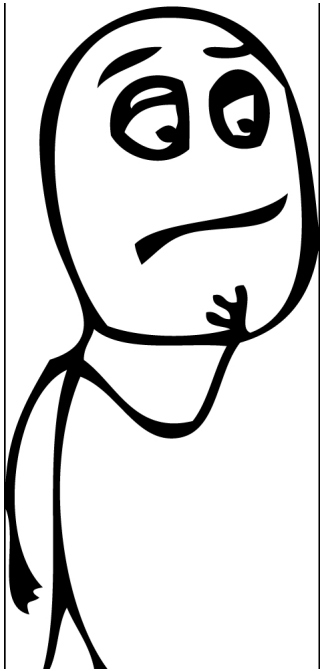
Or – what can I actually do today?



So I want to build a 10Gbit+ Capable Home Network – Off the Shelf switch and gateway options?



So I want to build a 10Gbit+ Capable Home Network – Off the Shelf switch and gateway options?

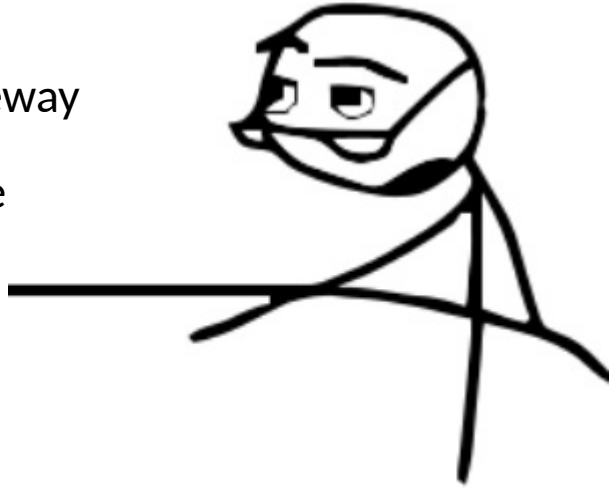


Wait… I’m **sure** i’ve seen some pro-sumer 10G gear starting to be advertised

Yeah, nah.

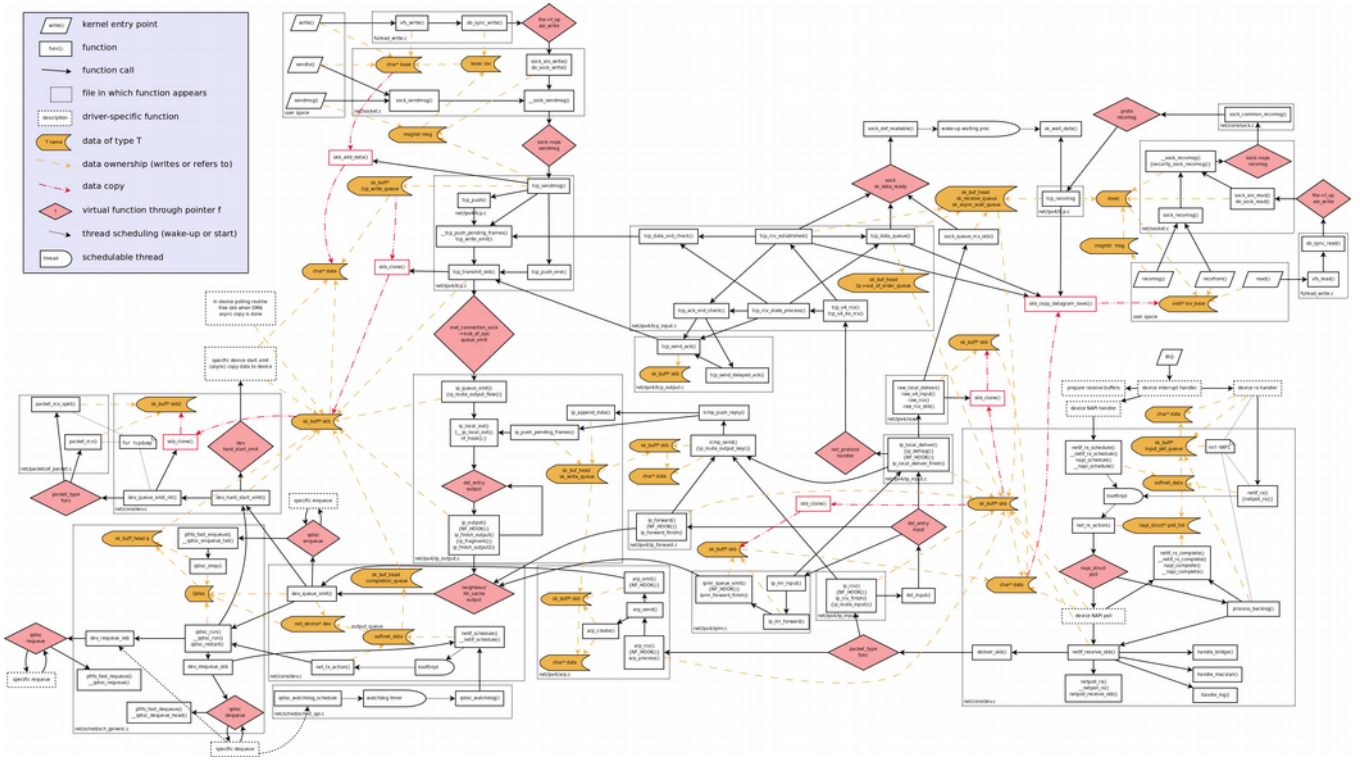
Even doing 1Gbit **slow path** Layer3 processing at 1Gig line rate on the Majority of Top end Expensive home gateway gear is often impossible. Yes you can do it on Desktop class x86 relatively easily (the worst case for **line rate Gigabit is ~1.5MPPS**), but the MIPS or ARM CPU in your gateway devices generally are not up to par and can't get close to this on a good day*

So whilst 10G BaseT NIC's are now becoming available at a reasonable price point – switches and more-so switch/gateway combo's which can do **14MPPS slowpath** style operations required for Linerate 10G Ethernet* are non-existent in the Consumer space

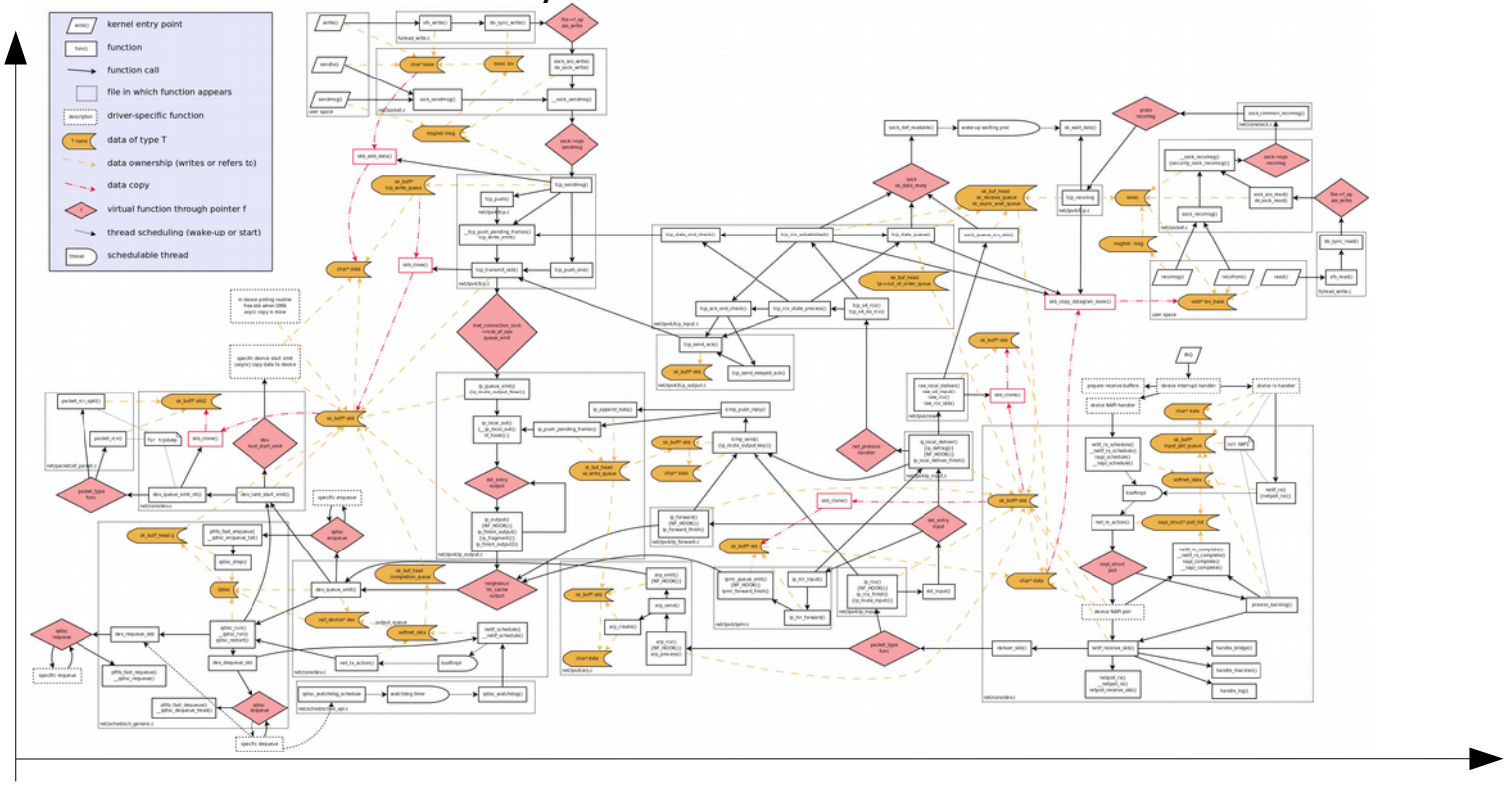


* On fully open firmware/Without offload tricks

Slow what? - Linux Network Stack Flow chart...

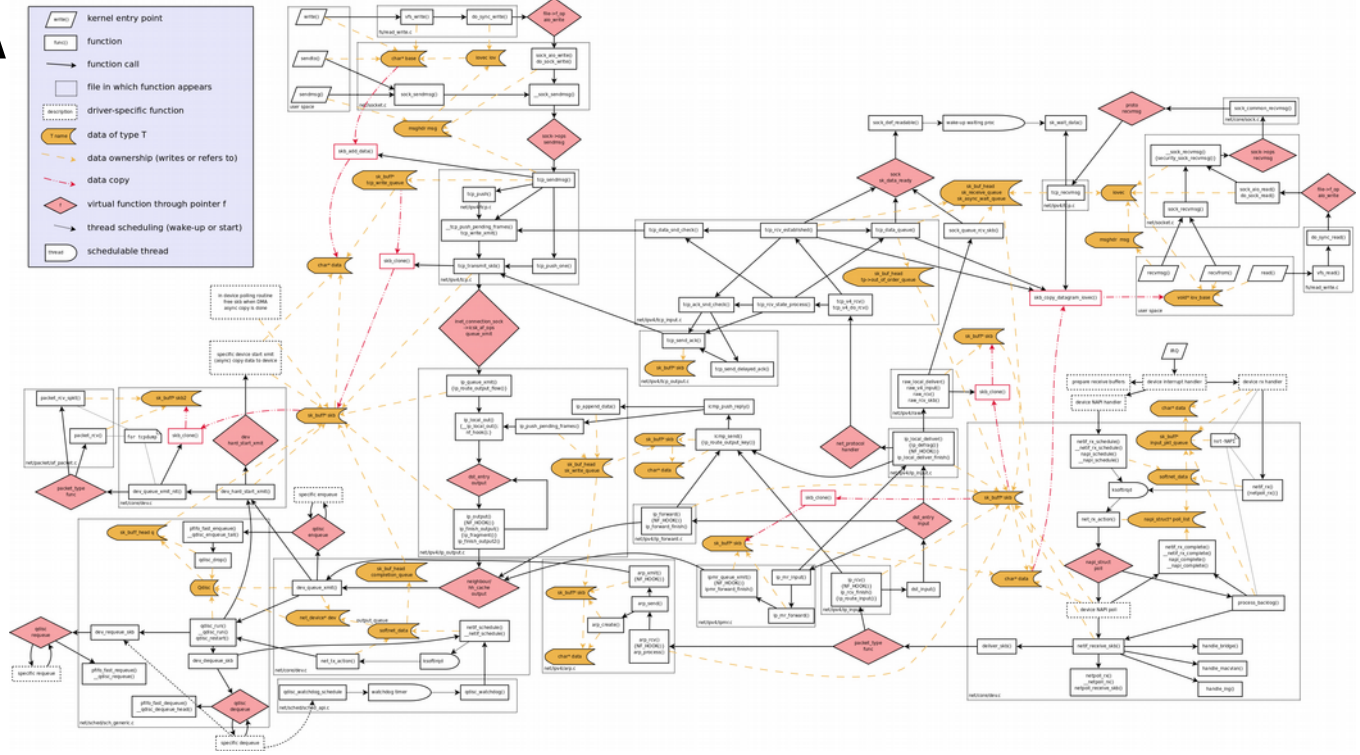


Roughly if we put these Axis on that beauty



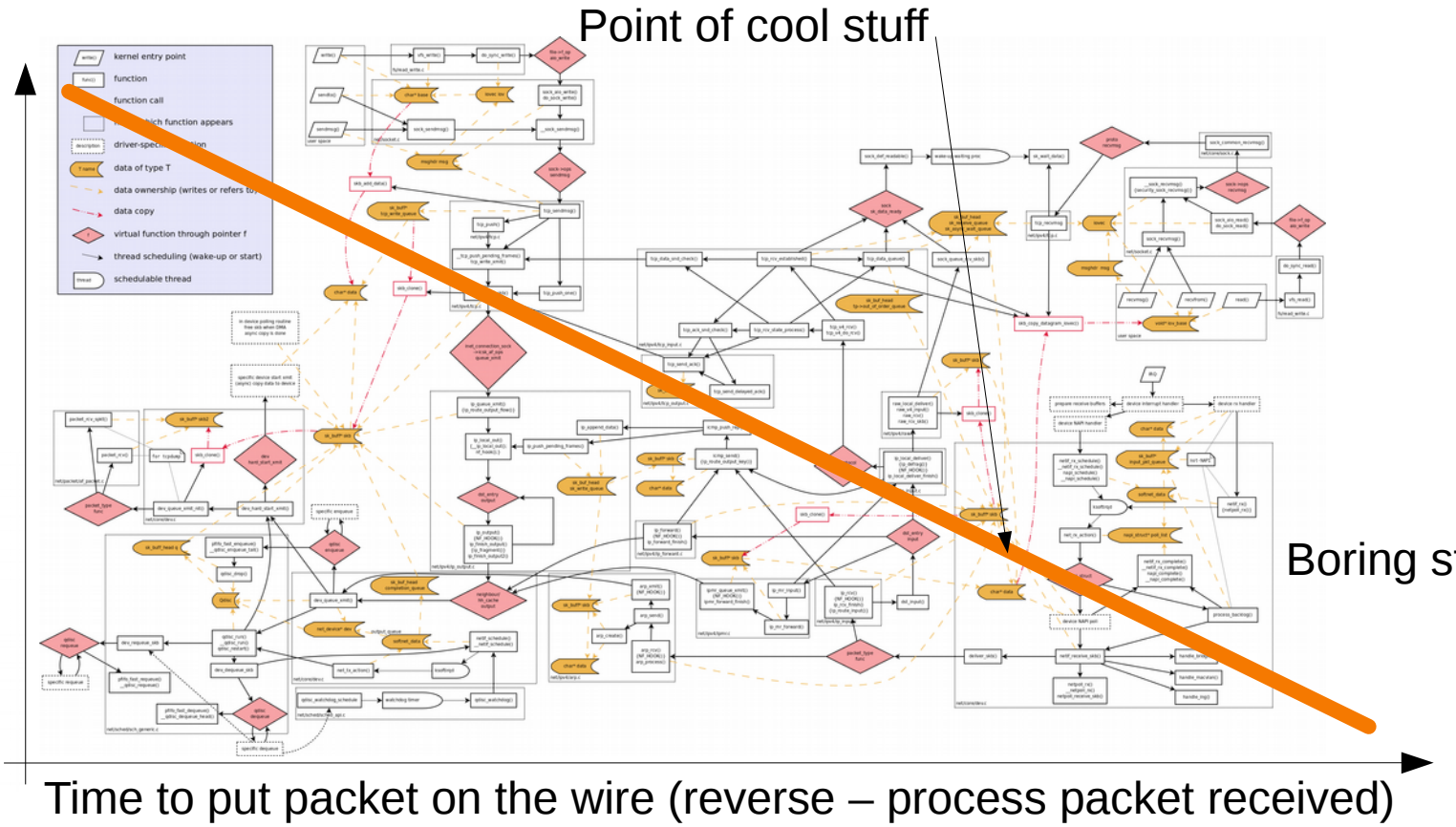
Roughly if we put these Axis on that beauty add some labels

More interaction with the CPU



Time to put packet on the wire (reverse – process packet received)

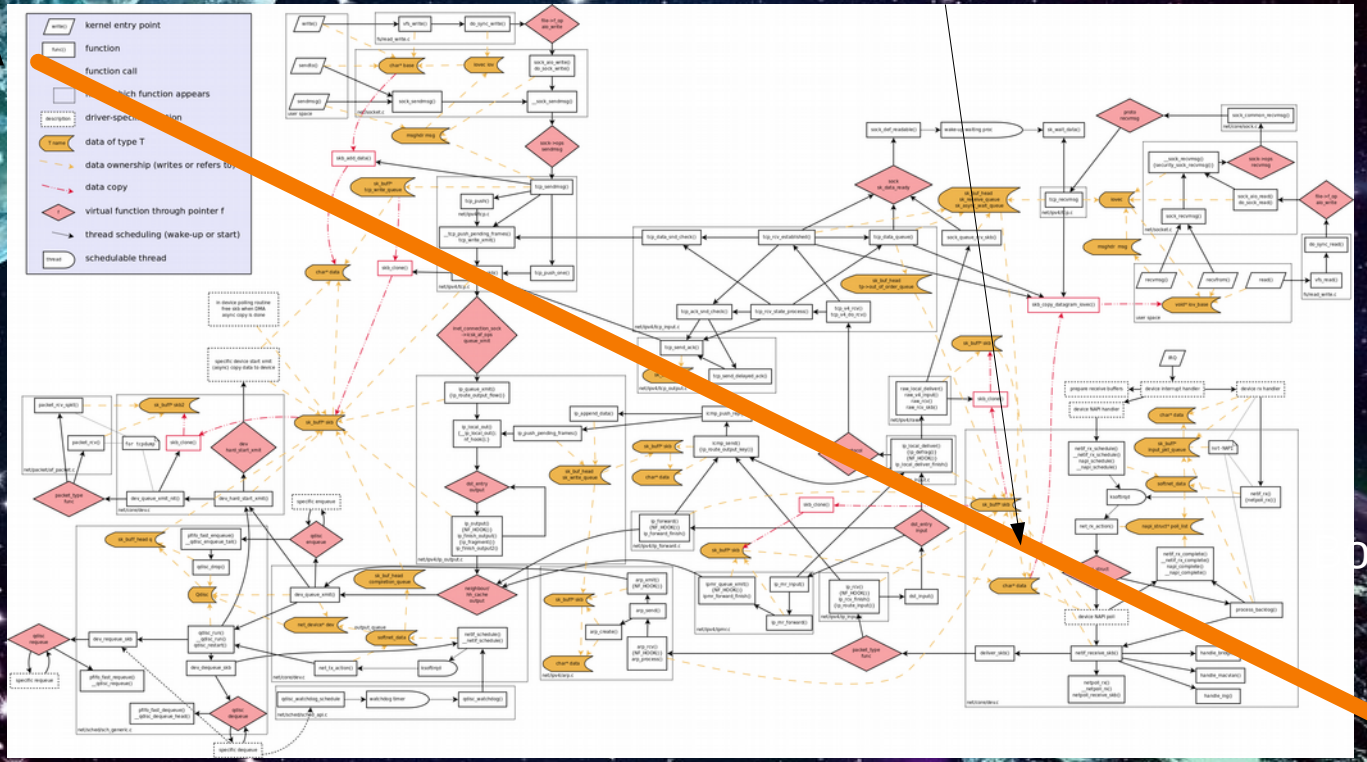
We get this –



More interaction with the CPU

User Space

Correlation of cool stuff



More interaction with the CPU

oring stuff

Time to put packet on the wire (reverse – process packet received)

What is ‘cool stuff’ - I am hip, you can’t tell me what to do!

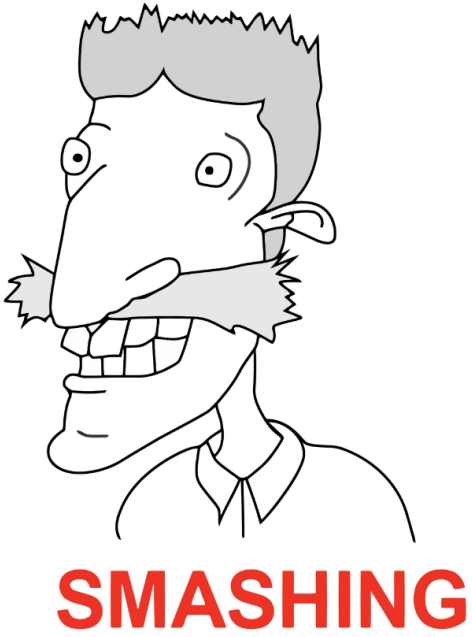
Ok maybe let’s look at what is boring stuff AKA operations that are mostly ‘fastpath’

- Bridging/Switching Packets for which we already know the Destination
- Reading from the NIC Hardware (NAPI Poll) – chunking read/writes (LRO/GRO)
- Offload mechanisms hooks

XDP and eBPF are allowing more cool stuff happen in Kernel paths that were traditionally considered ‘slowpath’ and/or needing to cross into user space – at ‘fastpath’ speeds

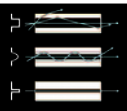
Recommend watching the NetDev presentations from November

https://www.youtube.com/channel/UCribHdOMgiD5R3OUDgx2qTg/videos



Being CPU Bound shouldn’t be a problem in fact, it’s kind of the goal. Networks should be matched as close as possible to compute power. Sans trickery – generalized kernel path should be good enough for general purpose computing tasks matched to PHY rates.

If the Architecture inside my PC is already Fast and Packetised – why can’t I just spread that around my general vicinity?



But – if the Datacentre is now hitting 25-100Gigabit right now? And It's the same basic commodity parts for the Motherboard/CPU that is in my home stuff?... right?

In the Datacentre – there are a heap of Vendor Specific work arounds commonly employed to to enable larger amounts of $coolstuff to happen at higher PPS rates than the default kernel stack normally would. Conversely achieving line-rate or close to line rate - for what would normally be slow-path/user-space limited by the CPU Clock and Bus Architecture.

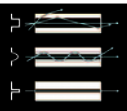
This turns out to be about ~30Gbps for Linux slowpath on recent high clocked(3.2ghz) intel x86 *

Regardless of Vendor/NIC Model - Traditional approach has (almost) universally been achieved by effectively making more stuff appear as 'boring stuff' to the Linux Network Stack with approach specific shortcuts – often via some sort of kernel bypass

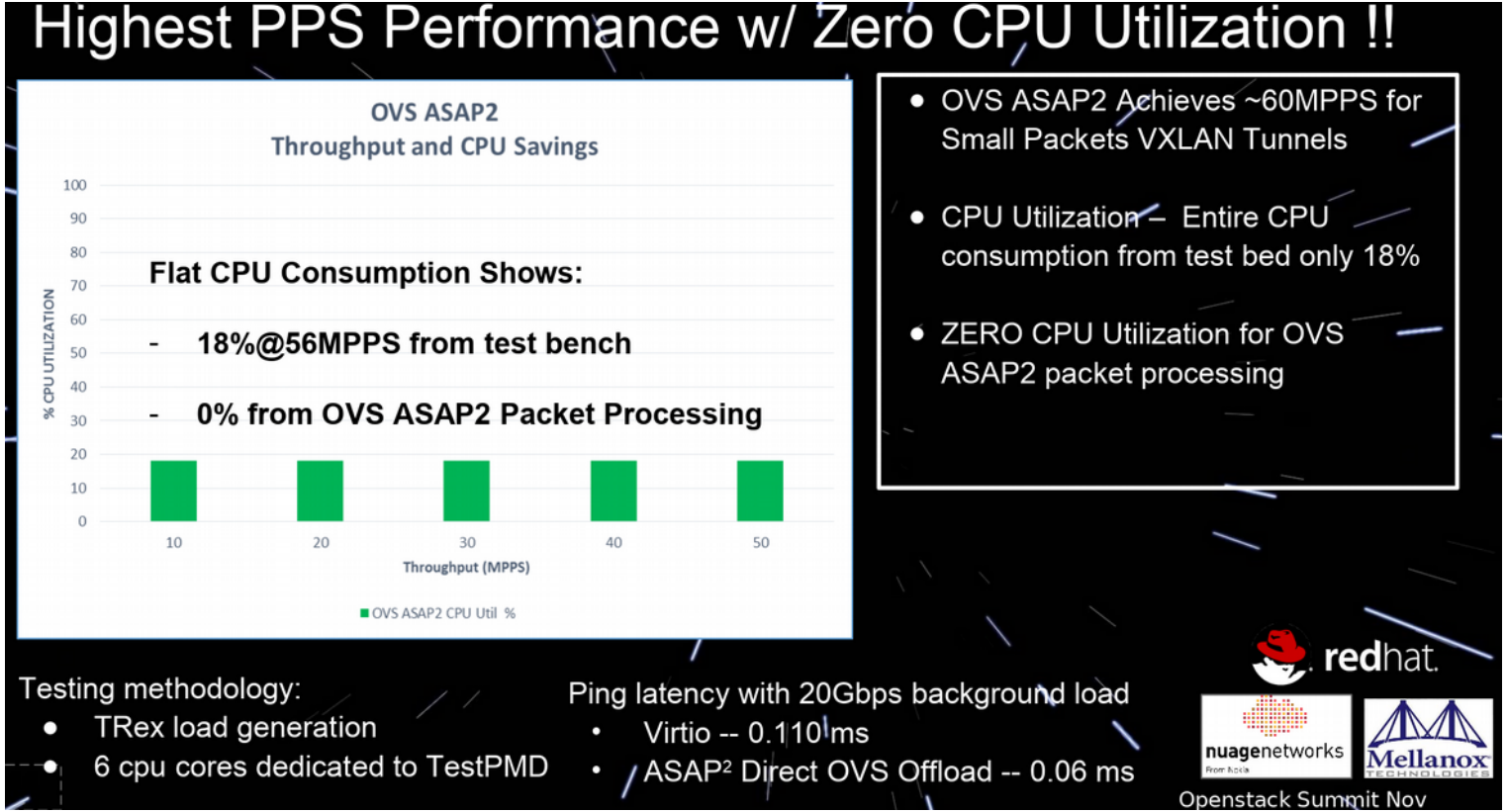
* Openvswitch virtio ports

Example – Mellanox eSwitch on x5 Series NIC with enhanced userland Openvswitch (Nuage VRS) with ASAP2 extensions

- Only really useful on x5 series – (older x4 NIC's don't support Layer3 rules)
- Nuage VRS is pretty much mainline-standard OVS with an skb_buf hook for first instance of a flow (this by itself allows pretty big gains – irregardless of NIC logic)
- Openflow rules are sourced from centralized controller to program local ovs with acl's and mac information on first occurrence of a flow
- Eswitch takes the local ovs rule-set programmed by userspace and pushes it directly onto NIC ASIC
- Anything modeled on NIC 'eswitch' can be offloaded from the Kernel path after that completely

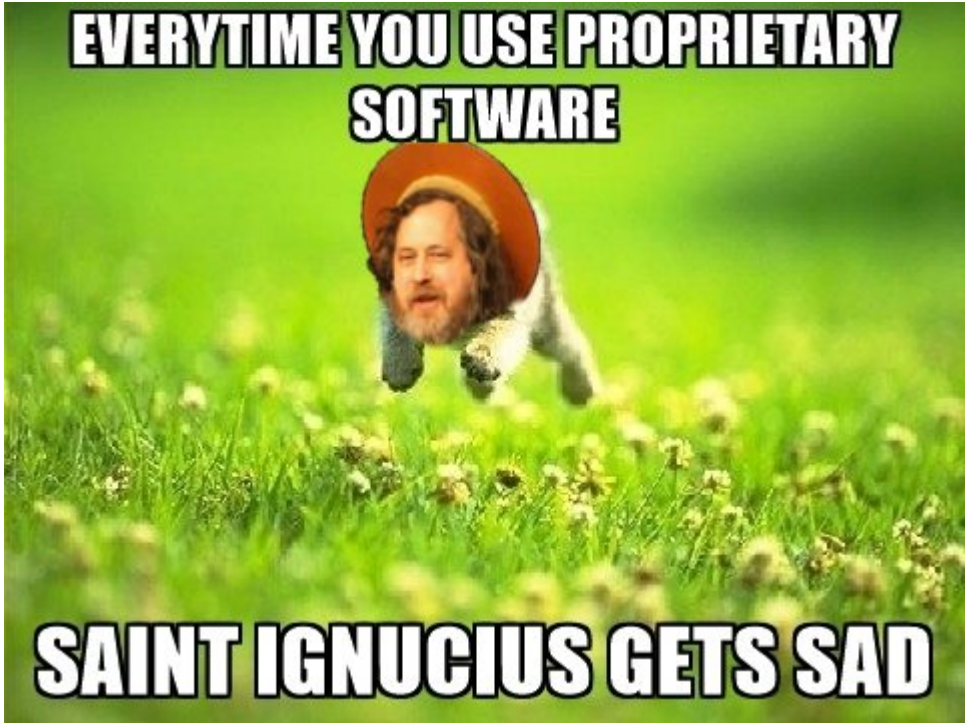


This approach (and others) allows for some fairly impressive results. And for BIG VNF workloads (Mobile Packet Core) spanning multiple Hyper-visors etc… it's needed

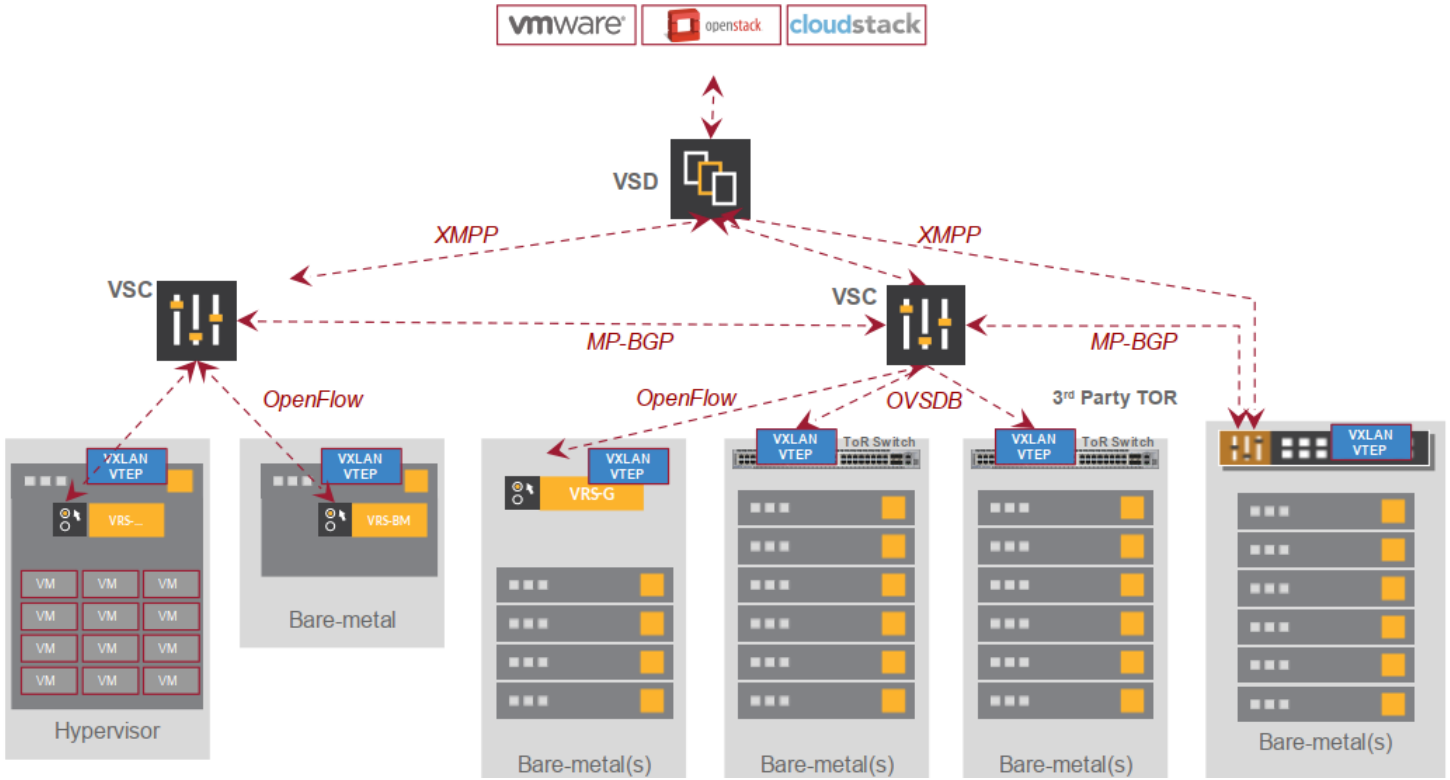


https://www.openstack.org/videos/sydney-2017/warp-speed-openvswitch-turbo-charge-vnfs-to-100gbps-in-nextgen-sdnnfv-datacenter

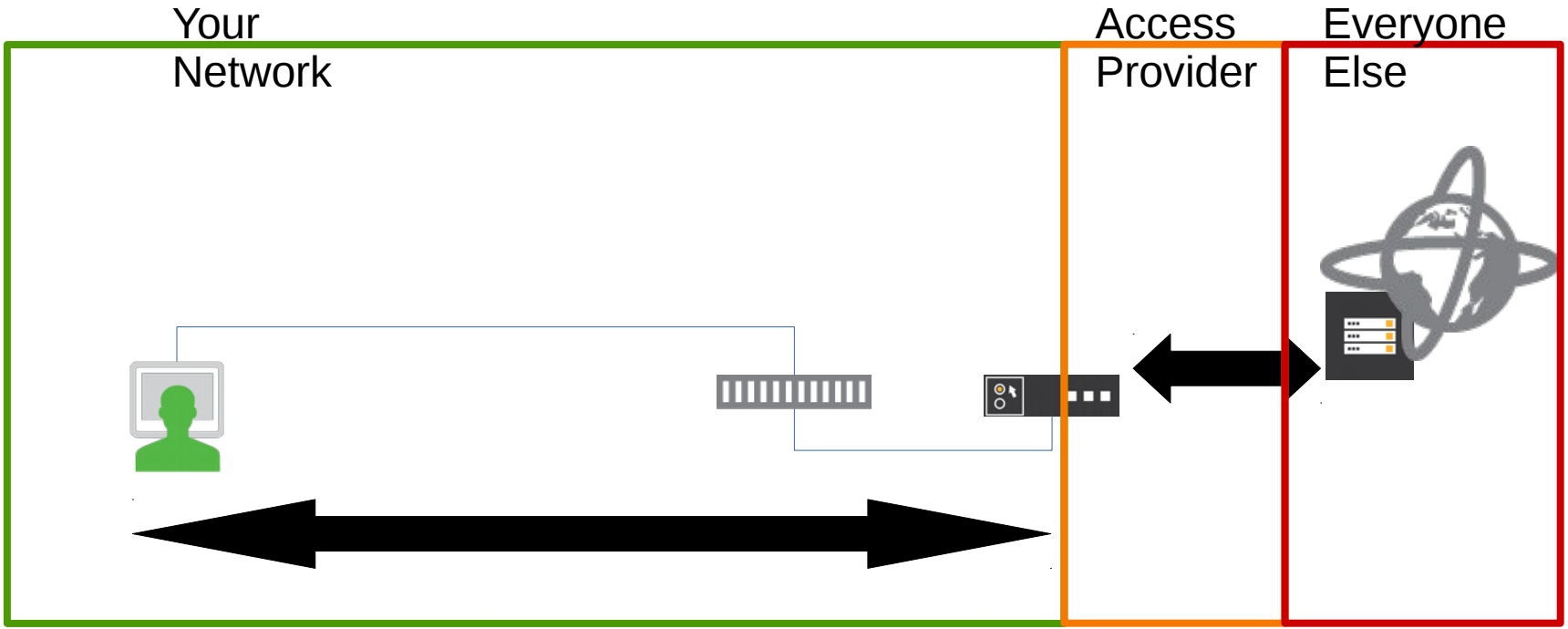
BUT this approach is horribly proprietary AND vendor specific in several places – but at least it uses an Open API for programming… your stack of deceit



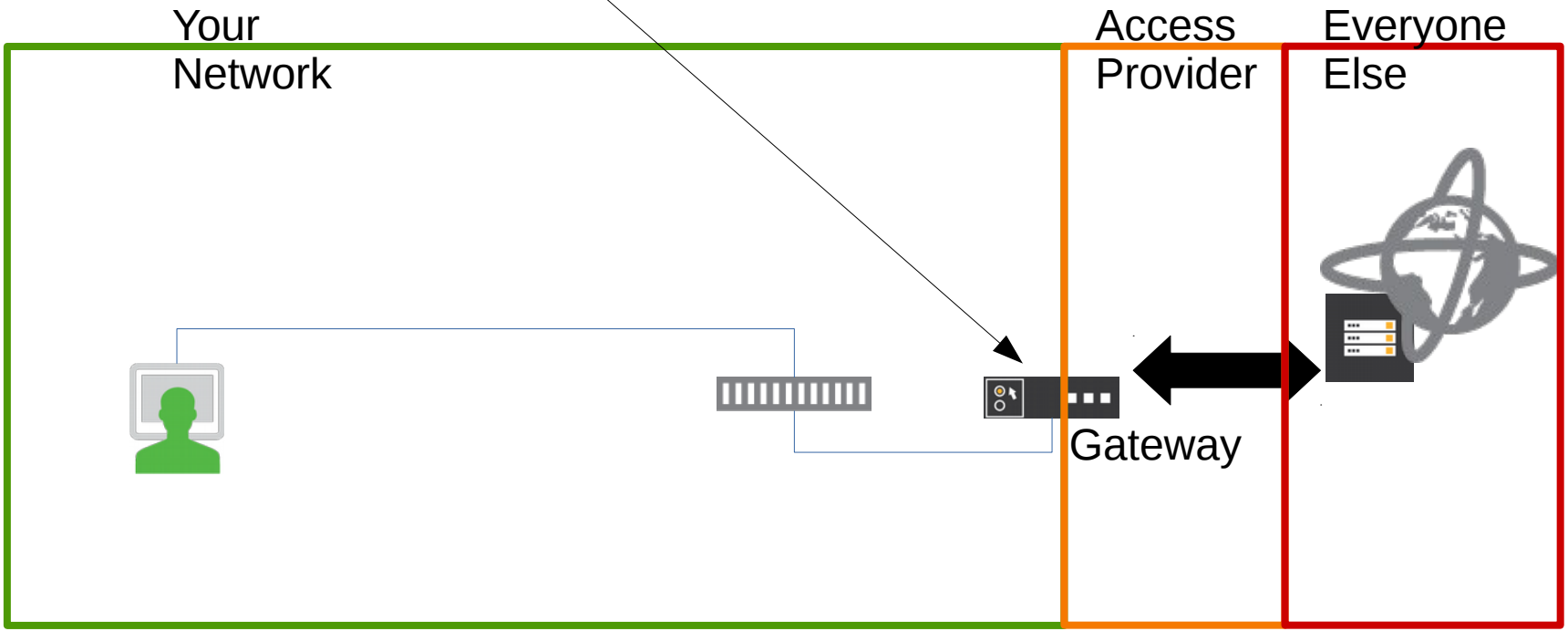
But it’s unlikely your home network looks like this



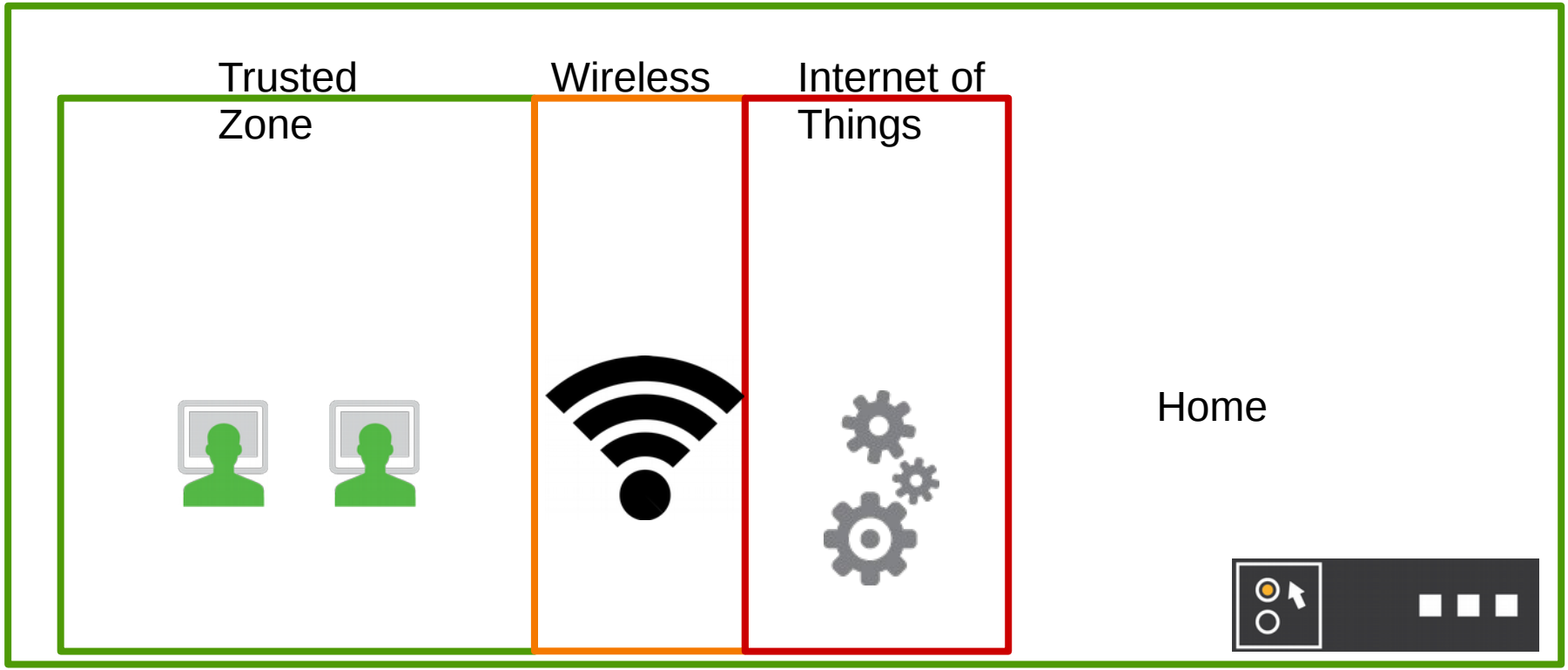
your network probably looks something more like this:



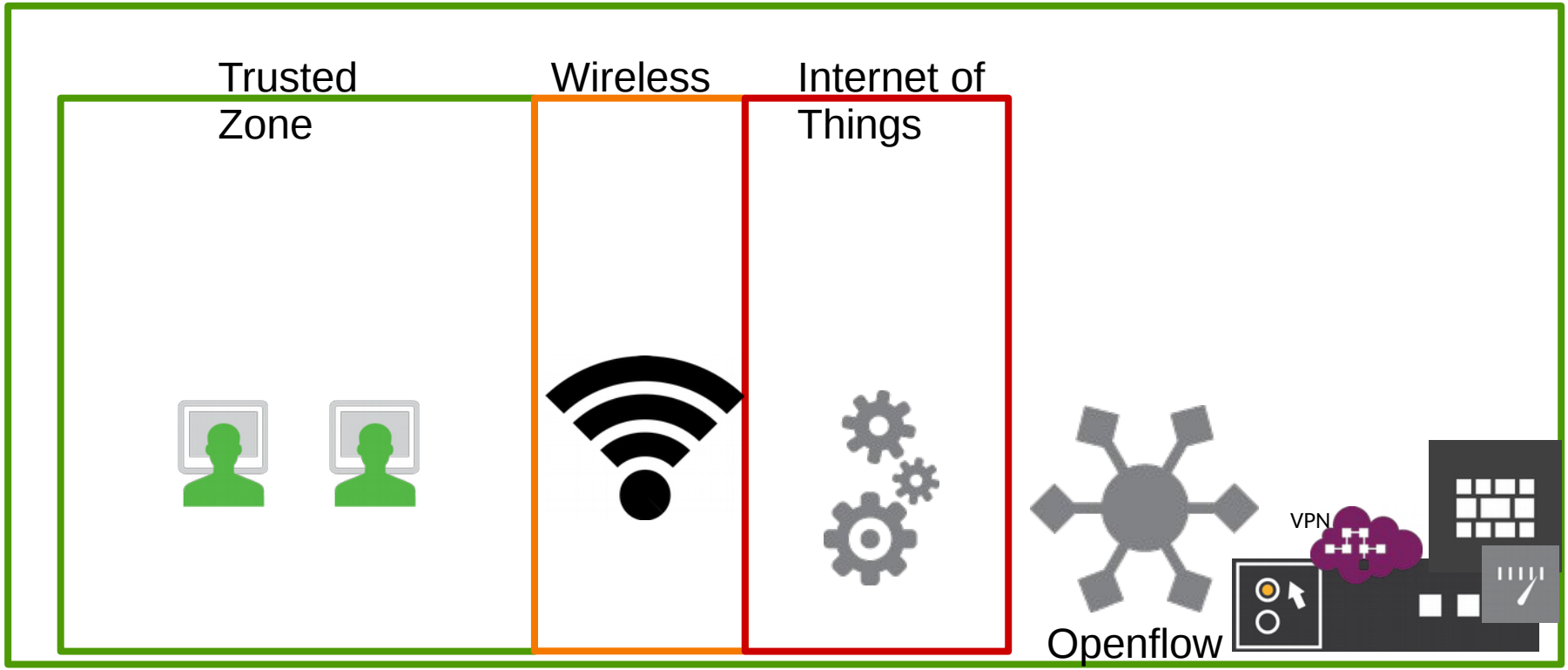
but this box is doing cool stuff pretty much 24/7 and probably runs linux (openwrt etc)



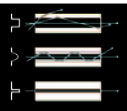
So that's a bottleneck right now. Also if your doing this sort of thing; which you probably should be. Then even for your home network that box is doing quite a bit of 'cool stuff'



And if you want a proper F/LOSS stack all the way down to the switch port management & VPN, & QoS & IDS & Firewall AKA **REALLY** COOL STUFF – that Box is getting a workout and is pretty likely slowpath bound a majority of the time

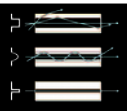


But basically most things you would normally do with a Home Gateway outbound your home (and depending on use case sometimes inside it) network is likely going to be some level of $coolstuff – that ends up being CPU bound somewhere along the path in and out



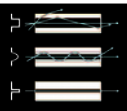
Kernel Development activities working on this problem ?

- XDP – eXpress Data Path - https://www.iovisor.org/technology/xdp (4.12 >=)
- In kernel Layer7 Proxies I.e kTLS stack (religious war ensues)
- eBPF – These are now mainline as of 4.15 net-next – and allow for some really cool monitoring stuff currently (and not just Network Stack!) - http://www.brendangregg.com/ebpf.html
- af_packet extensions and improvements (4.0.9 >= really made huge improvements to PPS/Lookup times)
- Vendor/Hardware Specific patch-sets – DPDK, Qualcomm Fastpath, Mellanox eSwitch/ASAP2, Nuage VRS (Openvswitch), Windriver accelerated Openvswitch (Nuage AVRS).
  - Generally not mainline, often proprietary or needs some external widget to work. Or sits in Userspace and takes away from normal Kernel stack workings.



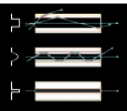
Team Blue (Intel/x86)

- x86 is Currently your best bet for a fully open Router/Gateway platform
- Widely available – Last gen (Broadwell) is sub 100$ for fanless SoC's with Dual NIC that can do 5Gbit Duplex Slowpath – perfect for 1G capable Homegateways, better than any ARM or MIPS consumer or pro-sumer gear i've tested
- There are a few Enthusiast Motherboards that have 10GBase-T, nothing with SFP+ unfortunately they all are designed around Desktop/Server chips… so not exactly low power or cheap
- Denverton range (Atom SoC) replacing Rangely – with SFP+ and 10Gbase Options – yet to see consumer devices with Denverton – but you can order it ; price-point still +1000$
- Thunderbolt3 as option for 10G or SFP+ breakout using Adaptor boxes – but not on Denverton??!! NuC's are an option



Open(ish)ARM things

- PI and Similar based SBC boards – all are pretty terrible
- Some of the newer A53 stuff can do close to 1Gbit Slow Path – might be ok as a replacement for your existing MIPs gateway
- Most lack good bus lane configuration (some exceptions, Firefly Rockchip 3399) so even looking to expansion options are not great
- There are some up and coming…

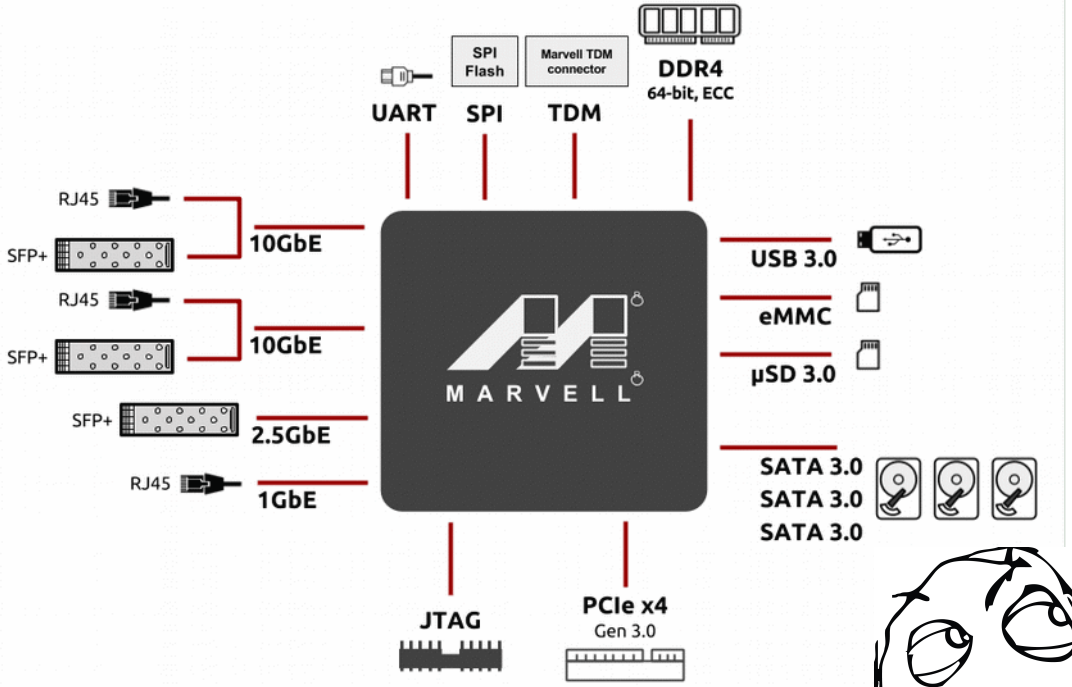


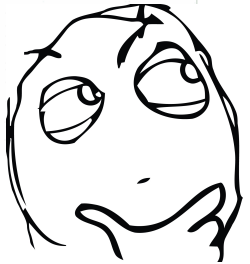
10Gbit+ Capable Gateway Hardware – Open System on a Chip/Router Platforms

Marvell MACCHIATObin

The Marvell MACCHIATObin is a first-of-its-kind Cost-Effective and High-Performance networking community development board targeting OpenDataPlane (ODP), OpenFastPath (OFP) and ARM network functions virtualization (NFV) ecosystem communities.

With a software offering that include a fully open source ODP implementation, U-Boot 2015.x, mainline U-Boot, UEFI EDK2, Linux LTS kernel 4.4.x, mainline Linux, Yocto 2.1 and netmap , the Marvell MACCHIATObin is an optimal platform that community developers and Independent Software Vendors (ISVs) can use for development around ODP and OFP and for delivering ARM based VNFs.





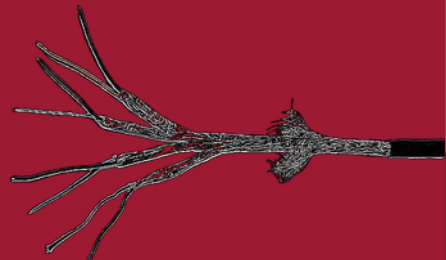
Micchiactobin – looks like the first SoC platform which meets all our requirements for a home 10Gbit+ gateway – the good

- Mainline Kernel Support
- Good Bus Lane Design
- Provides 2 * SFP+ 10Gbit & 10Gbit Base-T RJ45 Sockets (although shared with SFP+)
- Has a PCI-E x4 Slot – so could be extended with MIMO Wireless
- CPU - ‘should’ be fast enough to do slow path L3 processing
- Isn’t x86
- Fan-less, and ‘should’ be low power

Micchiactobin – looks like the first SoC platform which meets all our requirements for a home 10Gbit+ gateway – the not so good

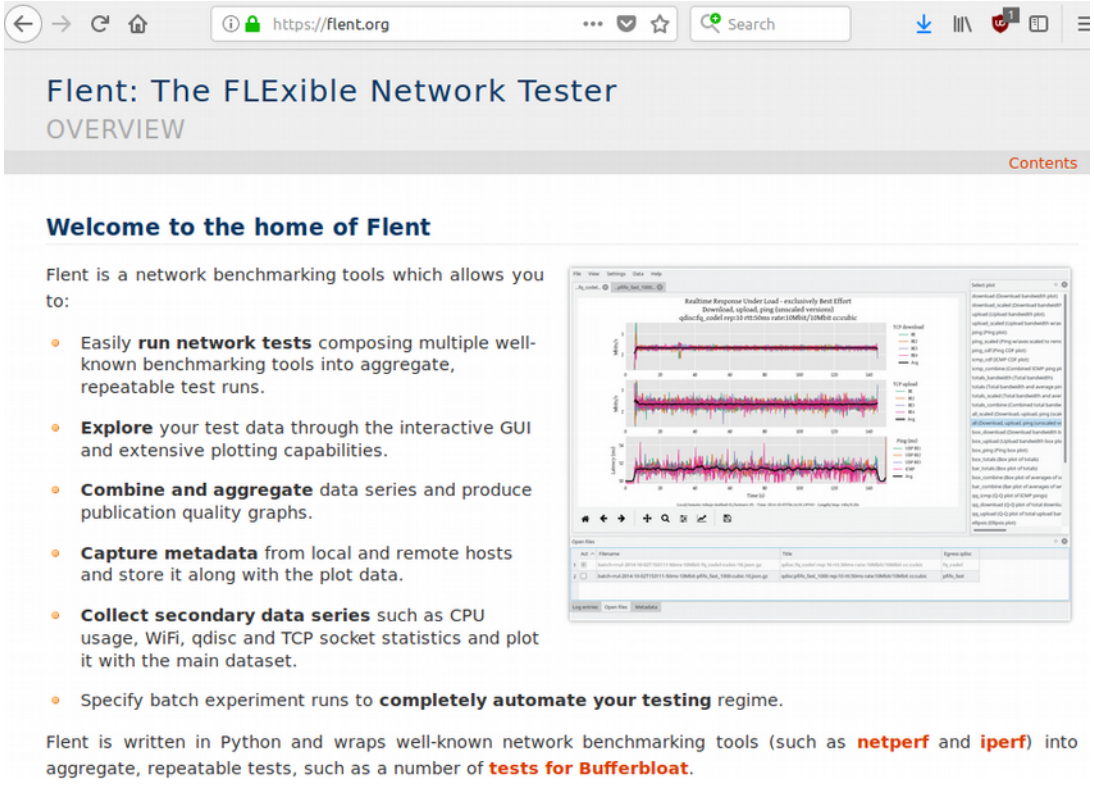
- Single PCIx4 Slot
- Requires DIY Housing (although there is the Google 8K variant which comes with case)
- Is an uncertified PoC (proof of concept) board
- I don’t have one/I don’t know anyone who has one
- Is on the steeper end of Consumer price range (400-800$ NZD depending on config)
- Isn’t x86 (Meaning Heavy VM/VNF workloads might not have arm version) – move to ‘cloud native’ containers however minimizes this problem somewhat.
- No thunderbolt3 – this is a rather large omission for connectivity to say NAS etc and/or SFP expansion
- No Wireless
- Probably relies on at least some of the Vendor tricks (Open FastPath to achieve line rate – but at least it’s Opensource)

I was told there would be Thunderbolt3

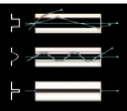


Yes! - But first let me introduce you to FLENT

http://flent.org

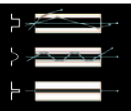


Provides a python wrapper around standard tools; such as netperf, iperf, tc, irtt (built specifically for the bufferbloat project). And matplotlib to do lovely graphing



Provides a python wrapper around standard tools; such as netperf, iperf, tc, irtt (built specifically for the bufferbloat project). And matplotlib to do lovely graphing

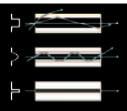
Extensible, repeatable test platform. It ain’t perfect but It in combination with RRUL* Suite is my ‘baseline’ testsuite.



https://www.bufferbloat.net/projects/codel/wiki/RRUL_test_suite/

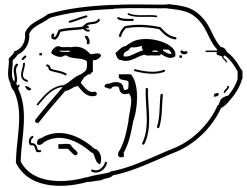
Thunderbolt - Overview

- Royalty and License free as of this year
- Provides standardized access to PCIE 3.0 X4 bus via an external connector
- Uses Serialized channels to allow slow and fast rate devices to co-exist on bus
- Low latency – as in almost as good as installing a card in a PCIE slot
- 40Gbit throughput available today
- You get 10Gbit network essentially for free ; 6 Devices daisy chained. Doesn’t need DMA access/BAR PCI extension in this mode (secure)
- Uses USB-C style connectors and provides for USB3.0 fallback modes (that don’t affect other Channel operation) + USB-C Power Delivery
- Available as Add-in cards, and Intel has said they will include Titan Ridge capability in all next generation Chips
- Not limited to Intel Chipsets! Although there are very few products to date that are non intel due to previously stringent certification process.
- SFP+ and QSFP+ thunderbolt NIC adapters available – take your laptop into the DC and plug into 10/40G Switches!
- Long run optical cables are available (but rare in market right now)



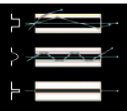
But what about USB3 /USB-C I hear you cry!

- Had an opportunity to introduce a proper Network Model for USB3.2 Gen 1 Spec – didn't
- High Latency comparatively (but better than USB2)– requires routing packets to host as first device in chain always.
- Throughput seems quite Dependant on USB Hub/Switch implementation
- Doesn't provide direct Bus/DMA access – only as good as your intermediary
- Short run copper only <=1M for Superspeed+ (10Gbit+)
- No optical cables
- Network model requires Kernel/Userspace gadget support – cross platform support isn't going to be a thing unless someone writes it.
- Have yet to see SFP+ / NIC adapters



Right now – Thunderbolt3 offers the best path to 10G+ networking in the home

The alternative is re-purposing Desktop Motherboards with Server NIC cards and a bunch of SFP+ (or try your luck with 10G Base T)




SHOW ME

- Setup

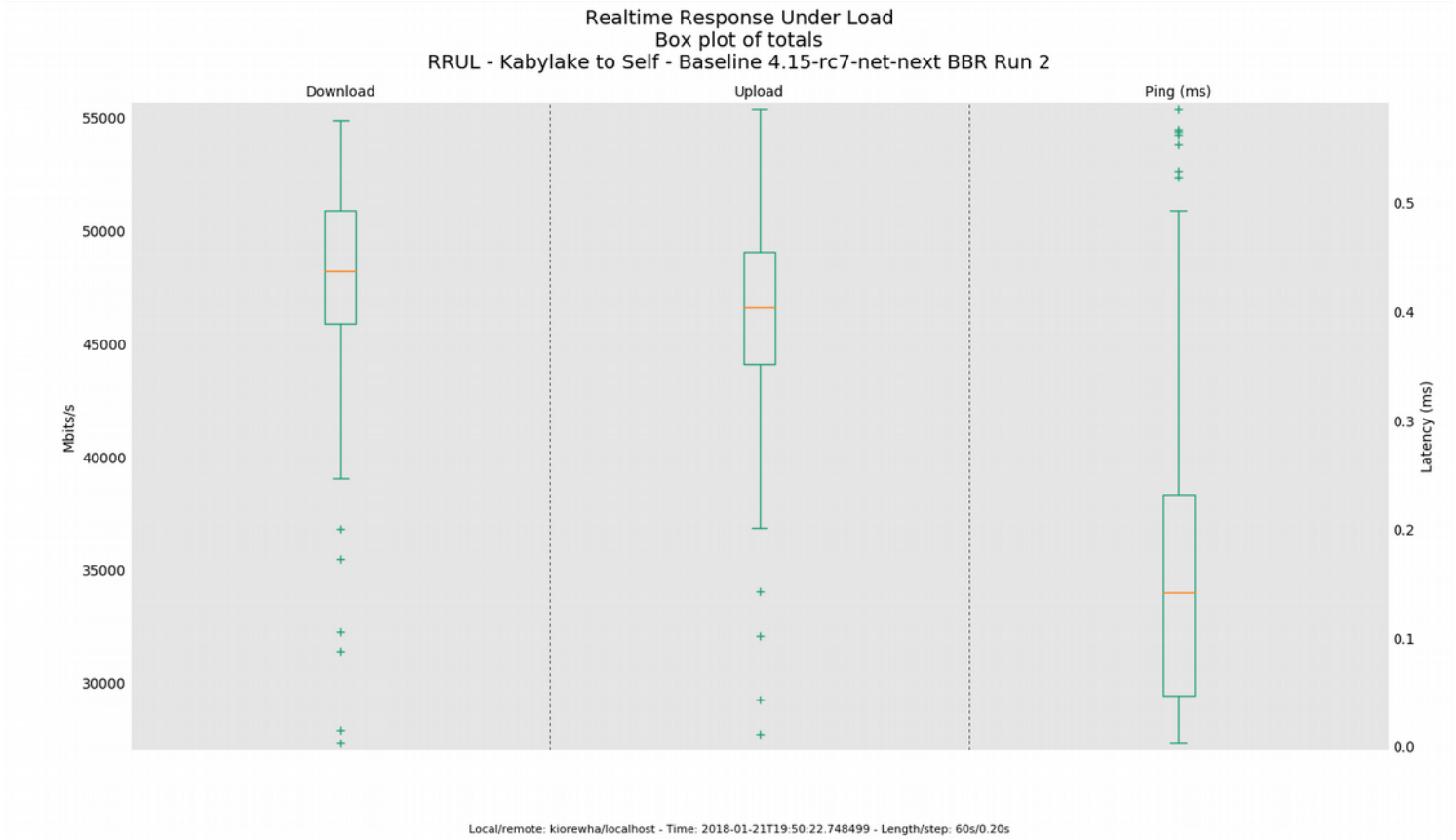
4.15-rc7 + net-next (includes KPTI patches)

Hurarongo – i7700k (kabylake)
Kiorewha – Xeon E3-1505M (skylake)

Both using Alpine Ridge TB3 Controllers
* Also tested to a Lenovo with Win10
* With Switch/Hub/Daisy Chain
* With older out of tree patches
* With low latency (the so called ‘zen’ patches)

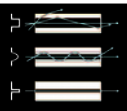
```
aenertia@kiorewha:~/go/bin$ ./irtt sleep
Testing sleep accuracy...

Sleep Duration         Mean Error        % Error
          1ns               243ns        24395.1
         10ns               234ns         2341.2
        100ns            13.371µs        13371.7
          1µs            53.065µs         5306.6
         10µs            53.177µs          531.8
        100µs            53.901µs           53.9
          1ms            69.326µs            6.9
         10ms            80.508µs            0.8
        100ms            86.692µs            0.1
        200ms            66.221µs            0.0
        500ms            82.322µs            0.0
 aenertia@hurarongo:~$ ./irtt sleep
Testing sleep accuracy...

Sleep Duration         Mean Error        % Error
          1ns               399ns        39991.9
         10ns               395ns         3956.5
        100ns               304ns          304.2
          1µs            55.184µs         5518.5
         10µs             55.21µs          552.1
        100µs            56.183µs           56.2
          1ms           160.201µs           16.0
         10ms           223.237µs            2.2
        100ms           215.309µs            0.2
        200ms           213.357µs            0.1
        500ms           182.279µs            0.0
```

Baseline RRUL loopback test on Kabylake (i7700K)



Linux Thunderbolt components

- kernel thunderbolt handler (enumeration and Memory mapping address space/PCI resources)
- Networking (IpoTB) module thunderbolt-net – compatible with Win/Mac
  - Appears as thunderbolt# device – manipulate as standard NIC
    - Currently doesn't fake Media rate, limited ethtool ; could be improved
    - Seems to be tuned for Duplex 10G ; although …
- User-space programs for authenticating device chain and listing topology (user cli tbtadm, udev hooks for acl tbtacl)
- Firmware update/capability handler (although no official userspace for this as yet) – hooking into mainline fwupd – initial out of tree util released on Monday (https://github.com/01org/thunderbolt-software-user-space/tree/fwupdate)
- Three security levels – none, user, secure
  - SL1 – user is the default recommended (requires device approval)
  - SL2 – requires additional cryptographic hash to be supplied during approval
  - Security levels DO NOT APPLY to the network component

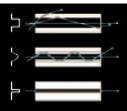


tbtadm

```
root@hurarongo:~# tbtadm
Usage: devices | peers | topology | approve [--once] <route-string> | approve-all [--once] | acl | add <route-string> |
 remove <uuid>|<route-string> | remove-all
```

```
root@hurarongo:~# tbtadm topology
Controller 0
    ├─ Details:
    │   ├─ Name: HP ZBook Studio G3, HP Inc.
    │   └─ Security level: SL1 (user)
    │
    └─ HP Thunderbolt 3 Dock, HP Inc.
        └─ Details:
            ├─ Route-string: 0-3
            ├─ Authorized: No
            ├─ In ACL: No
            └─ UUID: d3010000-0000-8f08-a224-34ca4ed02217
```

```
aenertia@hurarongo:~$ sudo tbtadm approve-all
[sudo] password for aenertia:
Sorry, try again.
[sudo] password for aenertia:
Found domain "/sys/bus/thunderbolt/devices/domain0"
Found child "/sys/bus/thunderbolt/devices/domain0/0-0/0-3"
Authorizing "/sys/bus/thunderbolt/devices/domain0/0-0/0-3"
Added to ACL
Authorized
```



Native devices appear as PCI resources after approval

```
aenertia@hurarongo:~$ lspci
00:00.0 Host bridge: Intel Corporation Skylake Host Bridge/DRAM Registers (rev 07)
00:01.0 PCI bridge: Intel Corporation Skylake PCIe Controller (x16) (rev 07)
00:02.0 VGA compatible controller: Intel Corporation HD Graphics P530 (rev 06)
00:04.0 Signal processing controller: Intel Corporation Skylake Processor Thermal Subsystem (rev 07)
00:14.0 USB controller: Intel Corporation Sunrise Point-H USB 3.0 xHCI Controller (rev 31)
00:14.2 Signal processing controller: Intel Corporation Sunrise Point-H Thermal subsystem (rev 31)
00:15.0 Signal processing controller: Intel Corporation Sunrise Point-H Serial IO I2C Controller #0 (re
00:16.0 Communication controller: Intel Corporation Sunrise Point-H CSME HECI #1 (rev 31)
00:17.0 SATA controller: Intel Corporation Sunrise Point-H SATA controller [AHCI mode] (rev 31)
00:1c.0 PCI bridge: Intel Corporation Sunrise Point-H PCI Express Root Port #1 (rev f1)
00:1c.1 PCI bridge: Intel Corporation Sunrise Point-H PCI Express Root Port #2 (rev f1)
00:1c.4 PCI bridge: Intel Corporation Sunrise Point-H PCI Express Root Port #5 (rev f1)
00:1d.0 PCI bridge: Intel Corporation Sunrise Point-H PCI Express Root Port #9 (rev f1)
00:1f.0 ISA bridge: Intel Corporation Sunrise Point-H LPC Controller (rev 31)
00:1f.2 Memory controller: Intel Corporation Sunrise Point-H PMC (rev 31)
00:1f.3 Audio device: Intel Corporation Sunrise Point-H HD Audio (rev 31)
00:1f.4 SMBus: Intel Corporation Sunrise Point-H SMBus (rev 31)
00:1f.6 Ethernet controller: Intel Corporation Ethernet Connection (2) I219-LM (rev 31)
01:00.0 VGA compatible controller: NVIDIA Corporation GM107GLM [Quadro M1000M] (rev a2)
02:00.0 Network controller: Intel Corporation Wireless 8260 (rev 3a)
03:00.0 Unassigned class [ff00]: Realtek Semiconductor Co., Ltd. RTS525A PCI Express Card Reader (rev 0
04:00.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
05:00.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
05:01.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
05:02.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
05:04.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
06:00.0 System peripheral: Intel Corporation DSL6540 Thunderbolt 3 NHI [Alpine Ridge 4C 2015]
6f:00.0 Non-Volatile memory controller: Samsung Electronics Co Ltd NVMe SSD Controller SM951/PM951 (rev
```

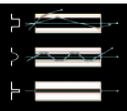
Before

```
aenertia@hurarongo:~$ lspci
00:00.0 Host bridge: Intel Corporation Skylake Host Bridge/DRAM Registers (rev 07)
00:01.0 PCI bridge: Intel Corporation Skylake PCIe Controller (x16) (rev 07)
00:02.0 VGA compatible controller: Intel Corporation HD Graphics P530 (rev 06)
00:04.0 Signal processing controller: Intel Corporation Skylake Processor Thermal Subsystem (rev 07)
00:14.0 USB controller: Intel Corporation Sunrise Point-H USB 3.0 xHCI Controller (rev 31)
00:14.2 Signal processing controller: Intel Corporation Sunrise Point-H Thermal subsystem (rev 31)
00:15.0 Signal processing controller: Intel Corporation Sunrise Point-H Serial IO I2C Controller #0 (rev 31)
00:16.0 Communication controller: Intel Corporation Sunrise Point-H CSME HECI #1 (rev 31)
00:17.0 SATA controller: Intel Corporation Sunrise Point-H SATA controller [AHCI mode] (rev 31)
00:1c.0 PCI bridge: Intel Corporation Sunrise Point-H PCI Express Root Port #1 (rev f1)
00:1c.1 PCI bridge: Intel Corporation Sunrise Point-H PCI Express Root Port #2 (rev f1)
00:1c.4 PCI bridge: Intel Corporation Sunrise Point-H PCI Express Root Port #5 (rev f1)
00:1d.0 PCI bridge: Intel Corporation Sunrise Point-H PCI Express Root Port #9 (rev f1)
00:1f.0 ISA bridge: Intel Corporation Sunrise Point-H LPC Controller (rev 31)
00:1f.2 Memory controller: Intel Corporation Sunrise Point-H PMC (rev 31)
00:1f.3 Audio device: Intel Corporation Sunrise Point-H HD Audio (rev 31)
00:1f.4 SMBus: Intel Corporation Sunrise Point-H SMBus (rev 31)
00:1f.6 Ethernet controller: Intel Corporation Ethernet Connection (2) I219-LM (rev 31)
01:00.0 VGA compatible controller: NVIDIA Corporation GM107GLM [Quadro M1000M] (rev a2)
02:00.0 Network controller: Intel Corporation Wireless 8260 (rev 3a)
03:00.0 Unassigned class [ff00]: Realtek Semiconductor Co., Ltd. RTS525A PCI Express Card Reader (rev 01)
04:00.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
05:00.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
05:01.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
05:02.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
05:04.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
06:00.0 System peripheral: Intel Corporation DSL6540 Thunderbolt 3 NHI [Alpine Ridge 4C 2015]
3b:00.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
3c:00.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
3c:01.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
3c:02.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
3c:03.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
3c:04.0 PCI bridge: Intel Corporation DSL6540 Thunderbolt 3 Bridge [Alpine Ridge 4C 2015]
3d:00.0 USB controller: ASMedia Technology Inc. ASM1042A USB 3.0 Host Controller
3e:00.0 Ethernet controller: Broadcom Limited NetXtreme BCM57762 Gigabit Ethernet PCIe (rev 01)
6f:00.0 Non-Volatile memory controller: Samsung Electronics Co Ltd NVMe SSD Controller SM951/PM951 (rev 01)
```

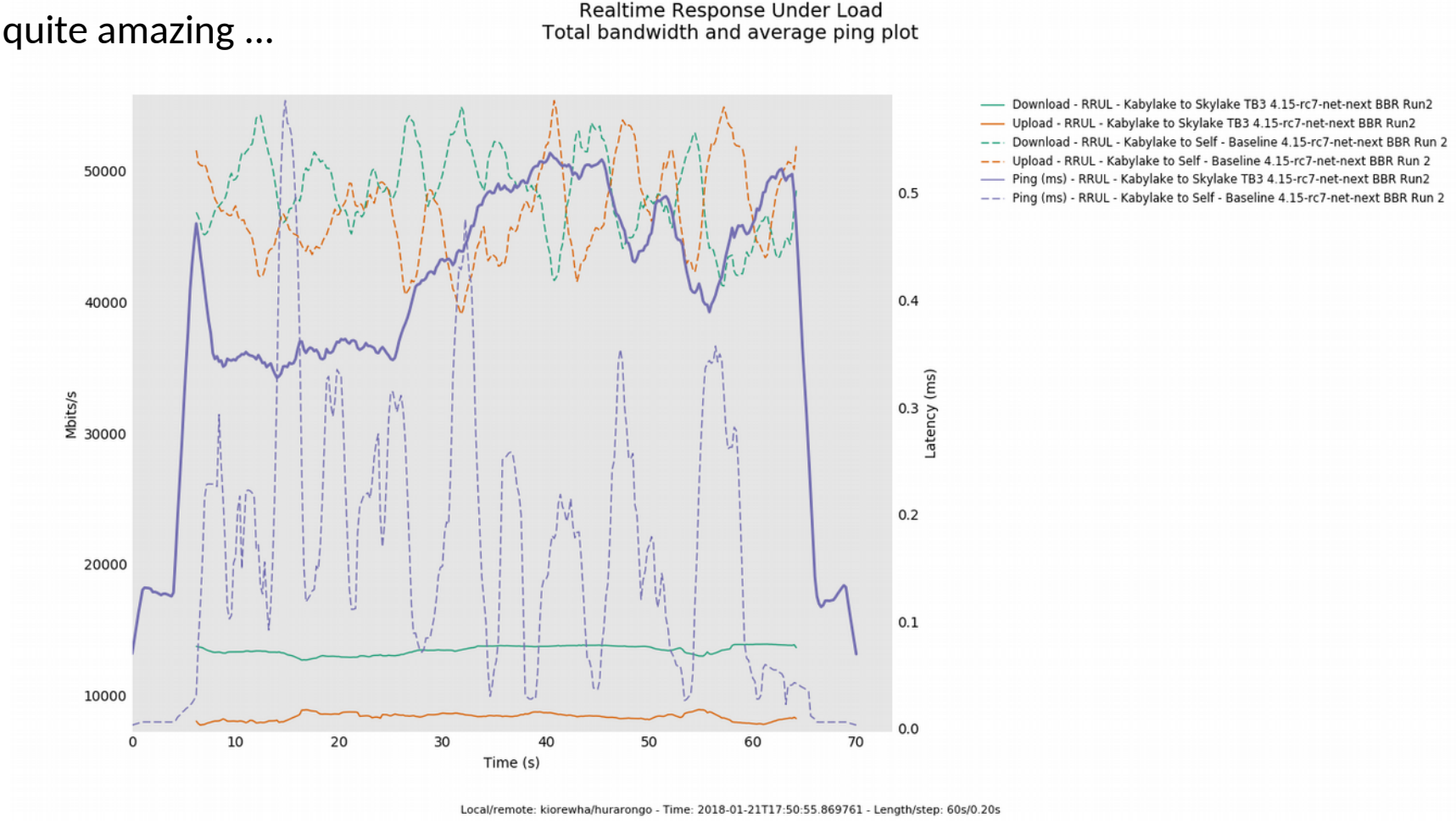
After

My Experiences from a year of use

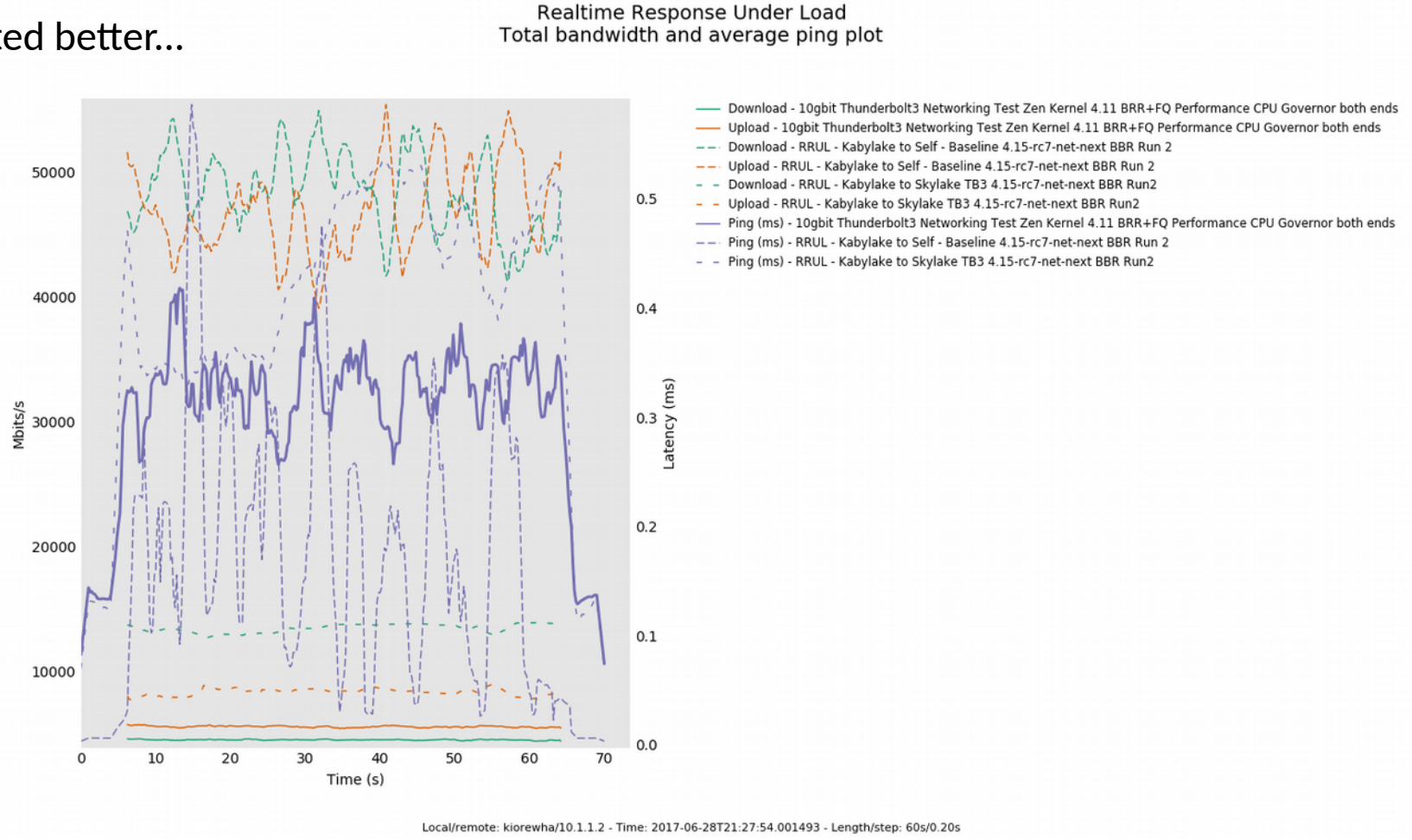
- Generally ‘just works’
- Some issues with PCI BAR memory allocation mapping with some BIOS – seems to have gone away sometime around 4.14-net-next
- Have tested most things, including stuff I don’t really use (eDP output)
- Sometimes network module get’s confused after suspend resume (likely other weirdness I am inducing, I am not a kind master to my devices)
- REALLY wish there was a switch/hub that was just 6 ports of TB3
- REALLY wish the Mythical Optical cables would make an appearance
- TCP Congestion control makes a HUGE difference at these speeds/latency – BBR is freaking awesome
- Latency is amazing ; I mean it’s as good as the expensive fancy NIC’s I use in my day job…



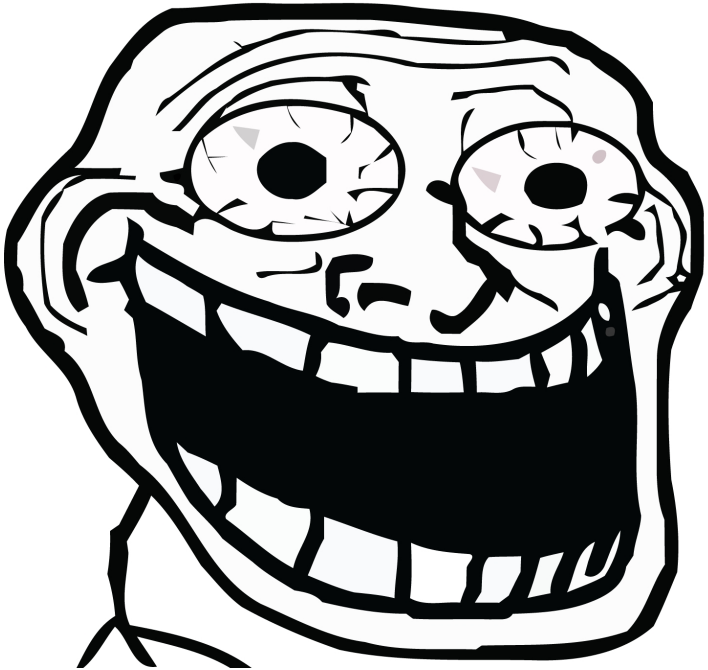
I mean quite amazing ...



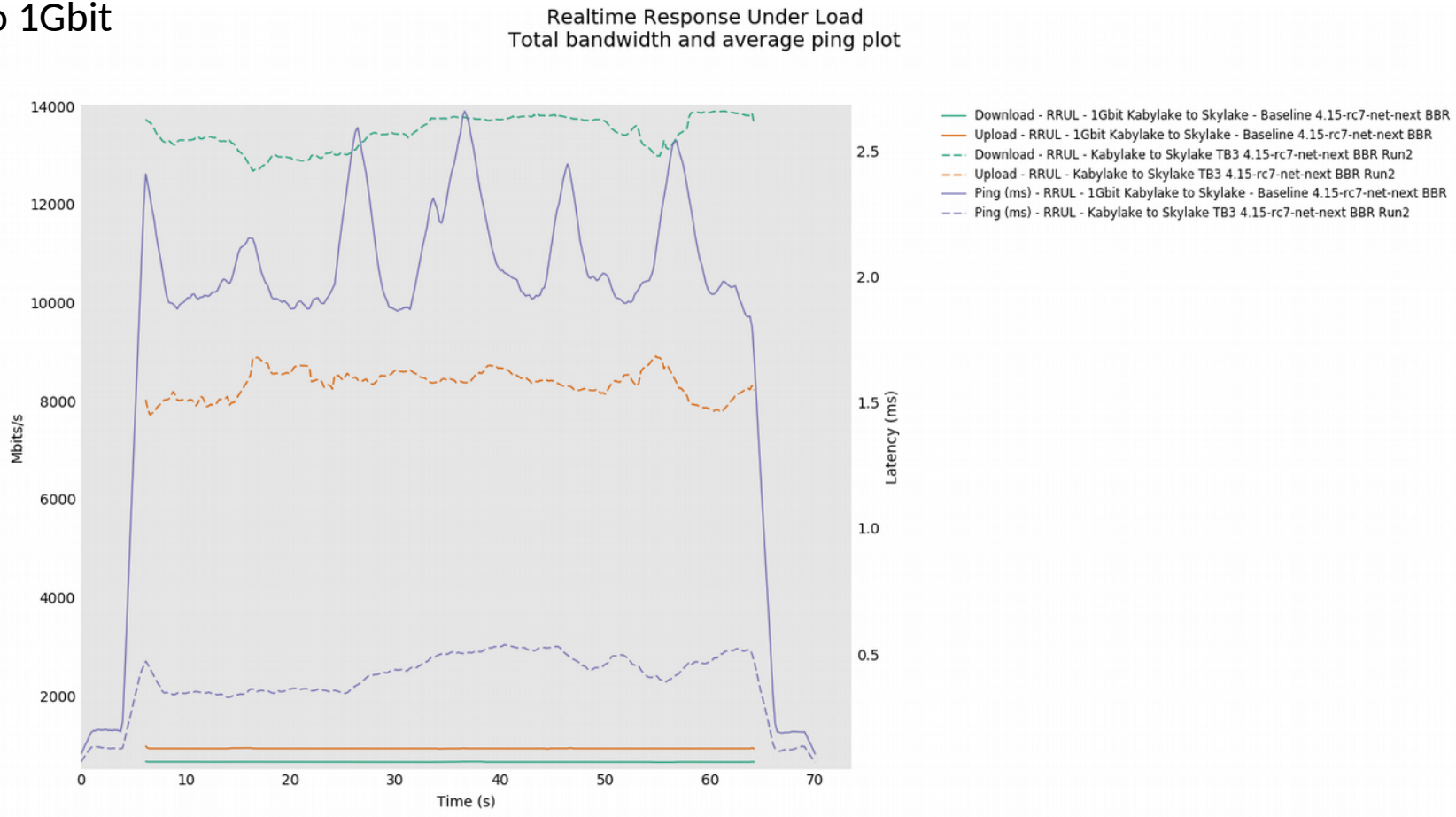
But I wanted better...



Within 50-100µs of Native CPU latency



Compare to 1Gbit

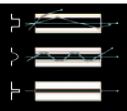


Yeah alright – real world things?

```
aenertia@kiorewha:~/Downloads$ mkdir /tmp/ramdisk
aenertia@kiorewha:~/Downloads$ sudo chmod 777 /tmp/ramdisk
[sudo] password for aenertia:
aenertia@kiorewha:~/Downloads$ sudo mount -t tmpfs -o size=4096m ramdisk /tmp/ramdisk
aenertia@kiorewha:~/Downloads$ mount |tail -n 1
ramdisk on /tmp/ramdisk type tmpfs (rw,relatime,size=4194304
aenertia@kiorewha:/tmp/ramdisk$ dd if=/dev/zero of=test bs=1G count=4
4+0 records in
4+0 records out
4294967296 bytes (4.3 GB, 4.0 GiB) copied, 1.09898 s, 3.9 GB/s

- roughly 30-32 Gigabit/Second - PCI Bus Rate... For Local Memory Copy operations ; about as good as we can get.

aenertia@kiorewha:~/Downloads$ rsync -vPh We\ Love\ Katamari\ \(USA\).iso /tmp/ramdisk/
We Love Katamari (USA).iso
          3.55G 100%  651.89MB/s    0:00:05 (xfr#1, to-chk=0/1)

sent 3.55G bytes  received 35 bytes  646.06M bytes/sec
total size is 3.55G  speedup is 1.00
aenertia@hurarongo:~/Downloads$ rsync -vPh We\ Love\ Katamari\ \(USA\).iso /tmp/ramdisk/
We Love Katamari (USA).iso
          3.55G 100%  411.70MB/s    0:00:08 (xfr#1, to-chk=0/1)

sent 3.55G bytes  received 35 bytes  418.04M bytes/sec
total size is 3.55G  speedup is 1.00
--
aenertia@kiorewha:~/Downloads$ rsync -vPh /tmp/ramdisk/We\ Love\ Katamari\ \(USA\).iso hurarongo:/tmp/ramdisk/
We Love Katamari (USA).iso
          3.55G 100%  232.43MB/s    0:00:14 (xfr#1, to-chk=0/1)

sent 3.55G bytes  received 35 bytes  229.25M bytes/sec

Roughly: 1.85 Gigabit/Second


---
Hrm rsync/ssh is hitting some sort of lower bound - no time to investigate further, to teh net cats;
```

Discovery - userspace tooling is woeful for fast networks – whole other talk…

```
aenertia@hurarongo:/tmp/ramdisk$ nc -l 9999 > Katamari.iso

total size is 3.55G  speedup is 1.00aenertia@kiorewha:/tmp/ramdisk$ pv We\ Love\ Katamari\ \(USA\).iso |nc hurarongo 9999
3.31GiB 0:00:02 [1.13GiB/s] [=================================================================================================================
==================================>] 100%


UDP? (don't try this at home folks)



aenertia@kiorewha:/tmp/ramdisk$ pv We\ Love\ Katamari\ \(USA\).iso |nc -u hurarongo 9999
3.31GiB 0:00:02 [1.47GiB/s] [=================================================================================================================
==================================>] 100%
^C
aenertia@kiorewha:/tmp/ramdisk$ md5sum We\ Love\ Katamari\ \(USA\).iso
94e909dfd5e1d573bc68e59585b200f6  We Love Katamari (USA).iso
aenertia@hurarongo:/tmp/ramdisk$ nc -u -l 9999 > Katamari.iso
^C
aenertia@hurarongo:/tmp/ramdisk$ md5sum Katamari.iso
57639f3f13580906b4941c3f3a80a66d  Katamari.iso

---

Roughly 11 Gigabits/Second
Again

aenertia@kiorewha:/tmp/ramdisk$ pv -i 0  -ptebar We\ Love\ Katamari\ \(USA\).iso |nc -u hurarongo 9999
3.31GiB 0:00:02 [1.41GiB/s] [1.41GiB/s] [=================================================================================
==================================>] 100%
```



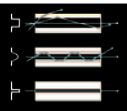
More Information

My TB Enabled ZEN Kernel Patchset (4.11 – pre mainline merge)
https://github.com/aenertia/zen-kernel

TB Userspace:
https://github.com/01org/thunderbolt-software-user-space

Mailing list:

thunderbolt-software@lists.01.org
Register at: https://lists.01.org/mailman/listinfo/thunderbolt-software
Archives at: https://lists.01.org/pipermail/thunderbolt-software/



THANK YOU

https://github.com/aenertia/lca2018-talk/tree/talk