Universiteit van Amsterdam | Hogeschool van Amsterdam
System and Network Engineering

Trans-oceanic performance engineering and monitoring:
reverse engineering other people's networks, policies or assumptions about your traffic

JP Velders, UvA SNE & UvA/HvA ICTS
GLIF 2017, Sydney
September 27th 2017

Agenda

- Data
- Leaders of the Pack(et)
- From A(msterdam) to (o)Z
- Other People's Networks
- Other People's Policies
- Other People's Assumptions
- Other People's Hosts/DTN's
- Observations
- Discussion !
- Utilized Equipment @ UvA

Data

- Quality: Who are the **current** data pushers ?
  - Power Users: HPx researchers, HEP, etc
  - inFamous huge data sources: CERN, SKA, etc
- Quantity: Who **will** be the biggest data consumers ?
  - Novice Users: Life Sciences, Climate Research, etc
  - Lots of ”small” data sources: make really big ones!
  - They don’t and shouldn’t need to know about (high performance) networking, computing or storage !
- How do we disseminate our research and expertise on high-speed/long-distance data transfer methods into cool touch-screen drag-and-drop interfaces ?

Leaders of the Pack(et)

We're the Alpha of the pack:
- We help, run and support (Packet based) networking
- Plan capacity by offering (virtual) circuit services

Data consumers:
- 40-100Gbps experiments focus on L2 Ethernet circuits
- End-to-end IP is what most users want and need !

Commodity vs R&E:
- R&E Networks still have a slight edge on L2.5/L3
- Feedback ~100Gbps Ethernet into ~100Gbps IP needed

From A(msterdam) to (o)Z

- 100G DTN within AARnet
- 40G DTN within UvAnet
- Round and round she goes, how she routes, God only knows:

Other People's Networks

- Great that we can use them, collaboration !
- Not so great if you hit limits and can't explain them. ☹
  - A limit of 10Gbps US East-West on NORDUnet
  - A 5x10Gps LAG between GÉANT and SURFnet
  - A 10Gbps peering fabric at PacWave ?
- OK, so how to identify these hurdles ?
  - Looking glasses (routes, interfaces, peerings)
  - Traceroutes to other places (bi-di)
  - Playing with BGP announcements
  - Connectivity listings on various websites
- Reverse Engineering almost like attack reconnaissance

Other People's Policies

- Some R&E networks provide IP transit, let's use that !
- Some R&E networks have oceanic 100GE's, use them !
- But, what about their BGP/Routing policies and peerings ?
- How do other parties treat your traffic/routes ?
- How do other parties do traffic engineering ?

Other People's Assumptions

- GÉANT NOC:
  - Why would you need >10Gbps single-flow !?
  - A 5x10Gbps LAG is sufficient for all your needs !
- LAG's aren't the solution:
  - <10Gbps single flows impacted other traffic
  - Back in 1Gbps days per-packet-hashing worked
  - FlexE ?
- GridFTP works !
  - HEP folks are thinking about streaming to hosts
  - Doesn't work end-to-end for novice users
  - Multi-stream, so requires store-and-then-forward

Other People's Hosts/DTN's

- Which OS + Kernel is the other DTN running ?
- What kernel TCP/IP-stack parameters been tuned ?
- Setting A works best with trans-atlantic TCP single-stream
- Setting B works best with trans-atlantic TCP multi-stream
- Setting C works best with trans-pacific TCP single-stream
- Setting D works best with trans-pacific TCP multi-stream
- Oh, setting A does not work with the DTN right next to it ☹
- Current network measurement tools don't scale endlessly:
    - iPerf TCP Window Size limits
    - iPerf >40Gbps single-stream is "interesting"

Observations

- ~27 Gbps single stream TCP @ ~320ms (Z->A)
- Adjusted routing policies in various places
- Kernel TCP/IP tuning is complex due to other factors
- NIC offloading can improve or ***reduce*** performance
- TCP Window Sizes >1GB sometimes work, sometimes…
- Intel Architectures require significant work vs Power8
- Whitebox switches have all kinds of unknown behavior

Take-aways:

- IBM Power8 is easy to use: no CPU I/O affinity issues
- IP was able to support single flow >10Gbps AU-USA-EU

Discussion !

- When you're sharing paths:
  - What do you monitor and is it meaningful ?
  - Impact on other traffic and vice versa ?
  - Can SD<*whatever*> help with traffic engineering ?
  - Is one monitoring dashboard referral site possible ?
- R&E networks and Exchange points need to collaborate:
  - Testpoints within their networks and at GOLE's
  - Design **end-to-end** routing policies
  - Where does L2/L2.5 make sense ***and where not*** ?
  - BGP Communities for 10G vs 100G links/routes ?

Utilized Equipment @ UvA

- IBM S821LC:
  - dual Power8 8 core @ 2.1GHz base
  - 2 threads per core (max 8)
  - 64GB RAM
- CentOS7:
  - PPC64LE (Little Endian just like x64)
  - kernel 3.10.0-514.6.1 (newer kernels: different behavior ☹)
  - RMEM/WMEM MAX @ ~2GB, CUBIC, etc
- Chelsio T6:
  - Chelsio drivers v3.2.0.0 with High Capacity settings
  - Tuned for NIC and OFLD profiles
- Juniper MX960:
  - 100GE up to SURFnet with 100GE to GÉANT
  - 40GE down to IBM box