Building high-performing infrastructures for research Workshop

NETWORK INFRASTRUCTURE AND ARCHITECTURES

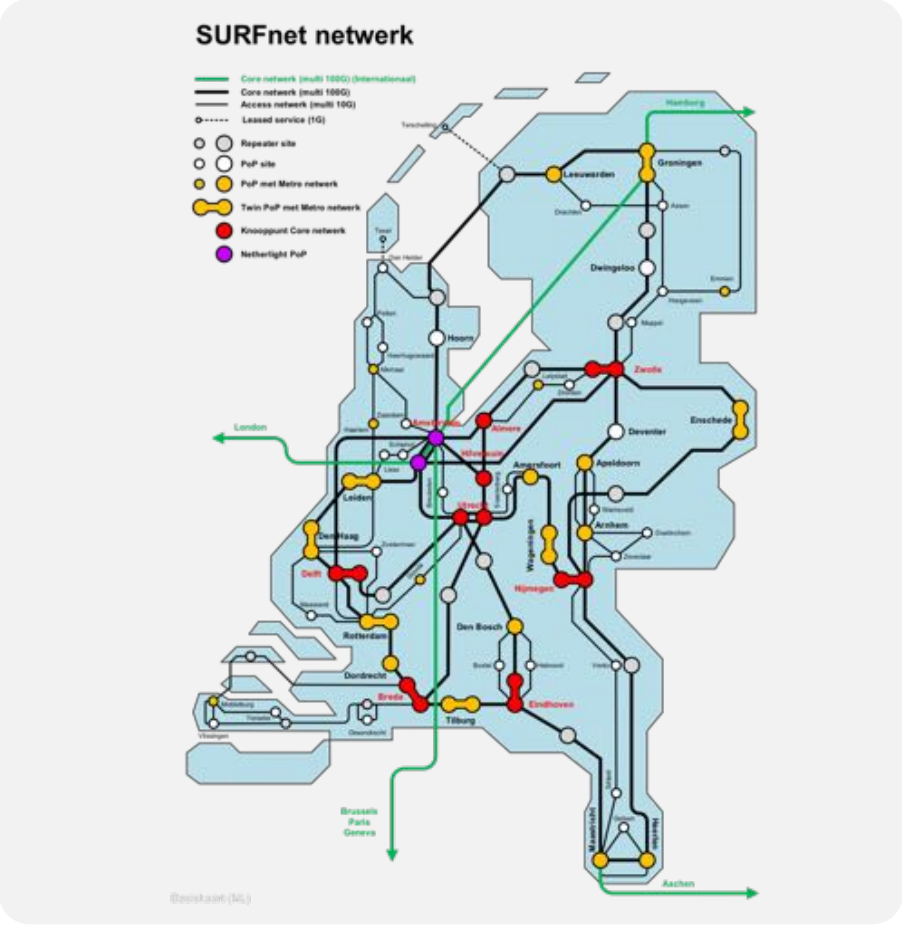
Pieter de Boer
SURFnet Network Engineer
ASTRON workshop
24 September 2018

Agenda

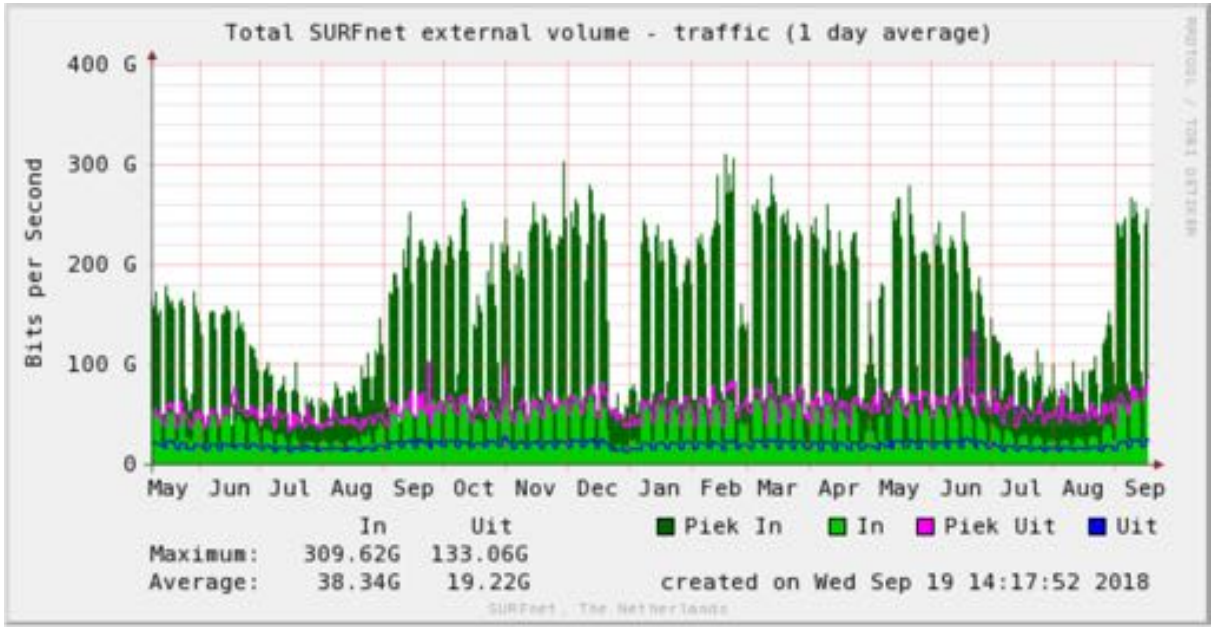
- Motivation
  - Overview of SURFnet & traffic statistics
- Research data workflows
- Applications & TCP
  - 80% of SURFnet traffic is TCP
  - Some may have heard of QUIC (UDP based) – TCP vs QUIC header stack
    - QUIC is still less than 3% of internet traffic – main application HTTP
  - TCP behavior and congestion
- What parts of the network architecture are important to examine?
  - Firewalls
  - General purpose networking
  - Building out the architecture for research requirements
- Privacy while in transit
- Recommendations

SURFnet Introduction

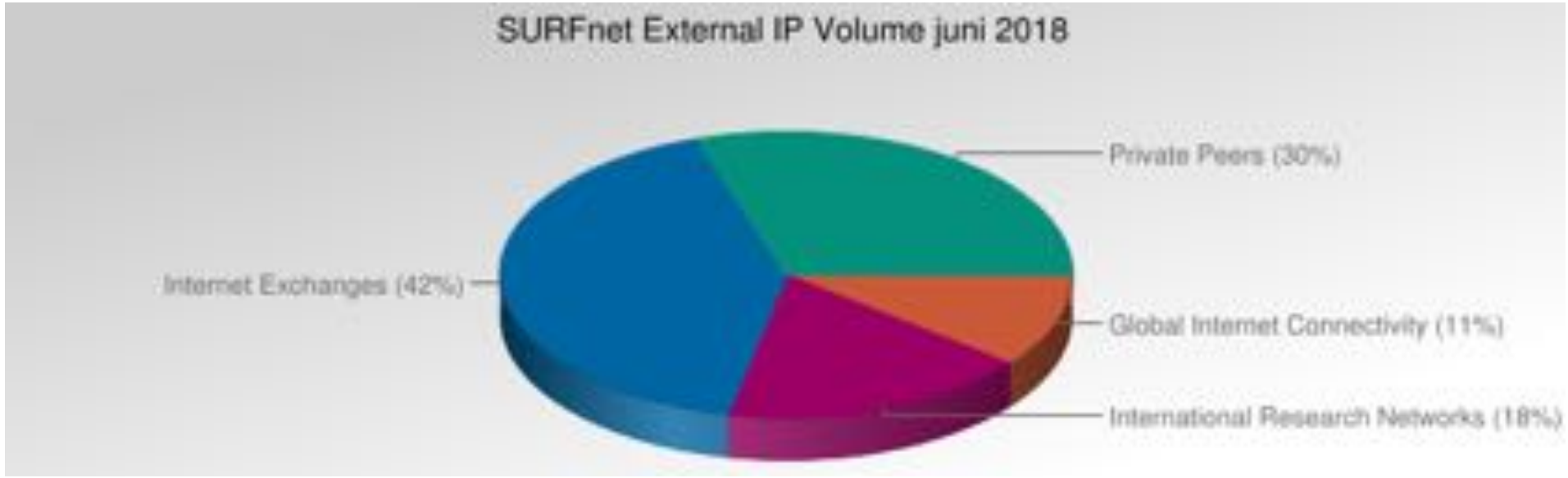
- National NREN of the Netherlands
- 12.000km of dark fiber
- Serving all Universities/HBO + research and MBO
- 1.5 million end users
- In the middle of migration to SURFnet8
- We peer at 100G with GÉANT and other providers
- Cross border fibers going to
  - Geneva (CERN),
  - Hamburg, Brussels/Paris/Geneva, Aachen
- Advanced North Atlantic (ANA) project partner
  - New link to Singapore



Data & traffic growth in SURFnet



Traffic categories



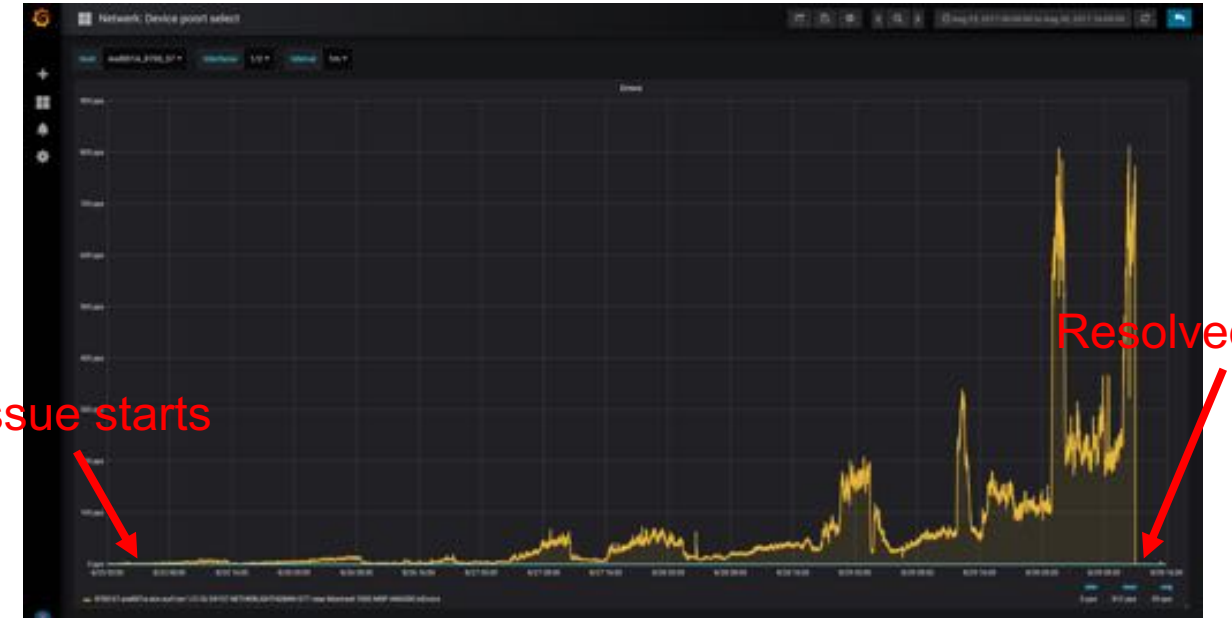
Data Characteristics

- Example researcher workflow:
  - Bulk data transfers
  - Streaming data
  - LOSF…
- Applications that researchers use are commonly TCP-based
- ~80% SURFnet traffic is TCP

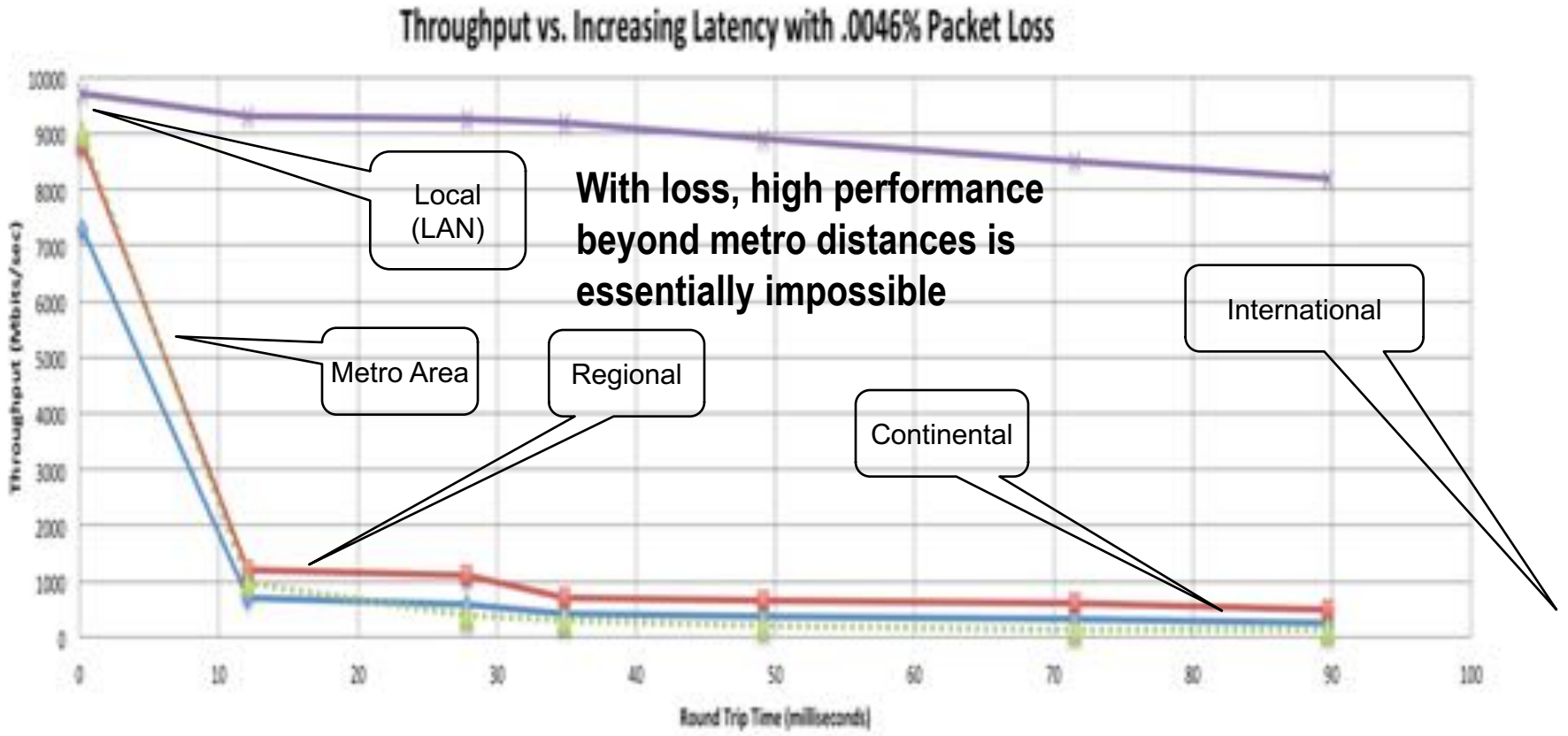
What does this mean?

- Doesn’t cope well with packet loss
- With packet loss assumes congestion and thus scales back

TCP Performance vs Interface errors

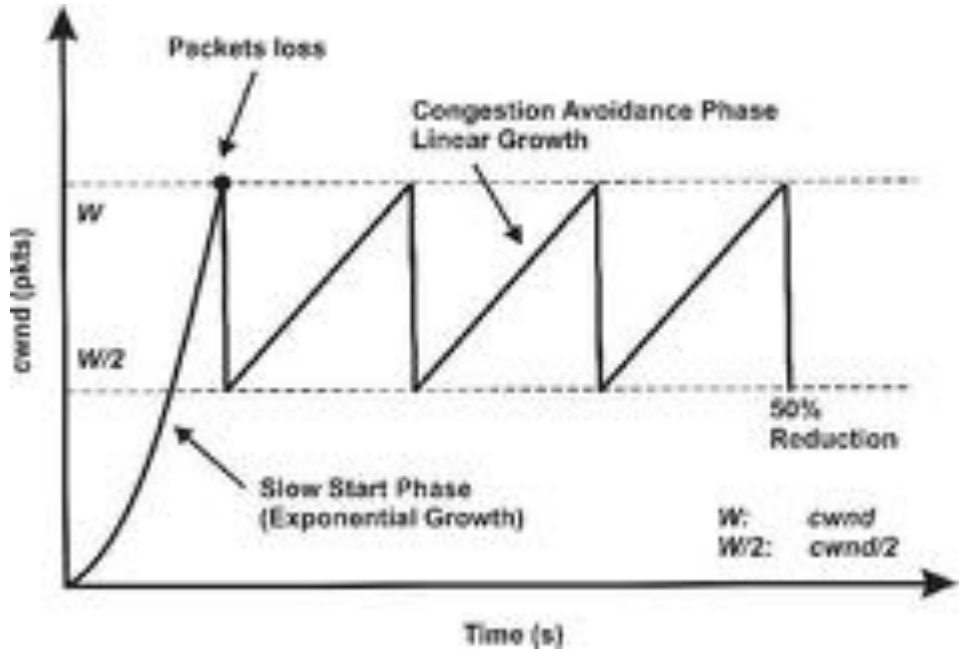


International Data Transfers with TCP for Example Radio Astronomy Data



© 2014, Energy Sciences Network

TCP congestion control



TCP Congestion Control (6)

With fast recovery, we get the classic sawtooth (TCP Reno)

- Retransmit lost packet after 3 duplicate ACKs
- New packet for each dup. ACK until loss is repaired

CNSE by Tanenbaum & Wetherall, © Pearson Education-Prentice Hall and D. Wetherall, 2011

TCP tuning

A lot of TCP tuning is going on

- Somewhat asocial, gives you advantage to other using less/untuned
- Risk of earlier collapse with aggressive ramp up (trying to get a bigger piece of the pie)
- What is the influence of your tuning on others?

Don’t do this over aggressive without realizing the consequences

What about UDP?

- Interesting discussions around UDP
  - Google's QUIC and IETF version of QUIC
  - >8% of traffic (depending on the point of analysis)
  - Interesting APNIC blog: https://blog.apnic.net/2018/05/15/how-much-of-the-internet-is-using-quic/
  - More interesting research done by: *Jan Rüth, PhD student at the Chair for Communication and Distributed Systems at RWTH Aachen University in Germany.*
- Other commercial UDP based applications are available
  - I.e., Aspera (IBM)
- Other open source UDP tools…
  - UDT

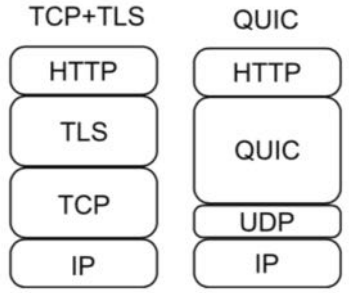


Figure 1. Comparison of TLS and QUIC

https://www.ietfjournal.org/quic-performance-and-security-at-the-transport-layer/

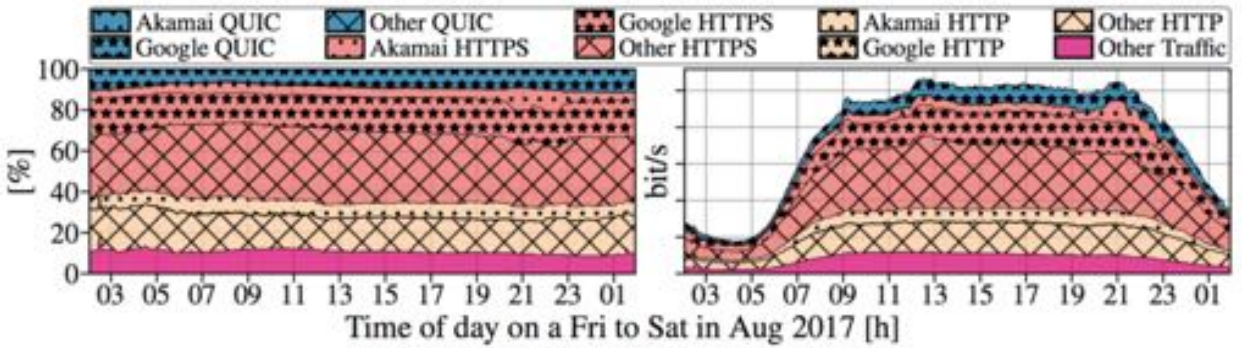


Figure 4: Relative QUIC shares (left) and absolute traffic (right) in the mobile network of a major European Tier-1 network. QUIC shares (blue) in contrast to HTTP (yellow) and HTTPS (red). Note: the ISP requested the actual traffic volume not be disclosed. From: https://blog.apnic.net/2018/05/15/how-much-of-the-internet-is-using-quic/

TCP is here to stay

Introduction new protocols and acceptance

- **IPv6**
- **Ipsec**
  - **Firewalls only allow TCP/UDP/ICMP if your unlucky**

Therefore

- TCP is here, far easier to support it than change all applications➔ people have tried
- Avoid congestion
- Build highly optimized section in your network for research data and work within research requirements
- Create a place to receive data, then transfer to end-users (automated or double-copy). I.e.,
  - Science DMZ: http://fasterdata.es.net/science-dmz/
  - Research Data Zone: https://www.surf.nl/innovatieprojecten/verbindende-infrastructuren/research-data-zone.html

Debugging network transfer is a pain

- Especially dedicated circuits (EPL/Ethernet pipe/SURFnet Lightpath)
- Routed gives you somewhat more view (traceroute)
- Assume route a-symmetry
- Juniper routers/switches have create firewall filter options to just count packets
  - Select source/destination set and just count
  - Can even apply layer3 filters on a switch port
- Checking every link for errors is such a pain ➔ have monitoring setup
- Checking mac tables for dedicated links is annoying

Interface errors / Monitor capacity

Fast data transfers not a given

- Network issues
  - QoS
  - Firewalls
  - Dirty fibers
  - Misconfiguration
  - Issues (fiber cuts, hardware faults, router OS faults)
- Hosts issues
  - CPU IO bandwidth limitation
  - OS tuning
  - Too many protocol layers

Quality of Service (QoS)

- **QoS, just don’t do it**
  - Sub-rate policers expect non-bursty traffic
    - 1G policer on a 10G interface, what does it
    - Burst will allow you briefly more
  - Research data is highly bursty (more than regular traffic)

Dedicated Firewalls (a.k.a. state full firewalls)

- Deep packet inspection, keep state, stuff that can't be done in ASICS
- Will not necessarily perform at the line speeds on the box
  - Firewall's with 100G interfaces exist
    …doesn't mean you'll get 100G
- Look at what happens when a DDOS hits the firewall
  - It collapses under the load

If you want some basic firewalling stay with firewall filters on your router, no state but wire-speed

Stateless filtering on a router

- Routers can do very useful firewalling, totally stateless
- Things like:
  - L3: Source/destination
  - L4: protocol + ports
  - State of TCP

- In Cisco land ➔ Access-lists
- In Juniper land ➔ Firewall filters

- It is Basic but all in can be done wire speed in hardware
- Great way to add more security in addition to host based firewall.

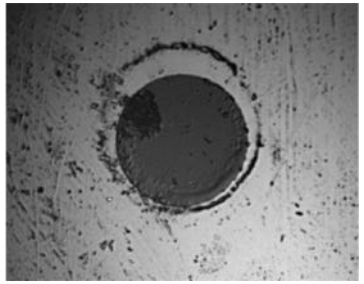
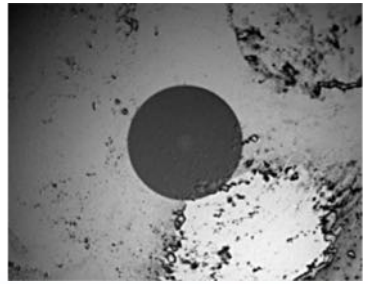
More on security in Michael's talk

Default Deny

- Only allow what you really need
- Block everything else
- Yes it is a pain at first
- It will get you to a better place
- And if you get hacked you're less fucked

Network issues – Dirt

- Dirty fibers
  - GE cleaning was optional
  - 10GE fiber cleaning is strongly advised
  - 100GE fiber cleaning necessity
  - Actually just always clean them, get a fiber clearer
  - Optic interface cleaning with a fiber cleaning pen
  - SUNET has a great blog https://www.sunet.se/blogg/long-read-cleanliness-is-a-virtue/



Network issues - other

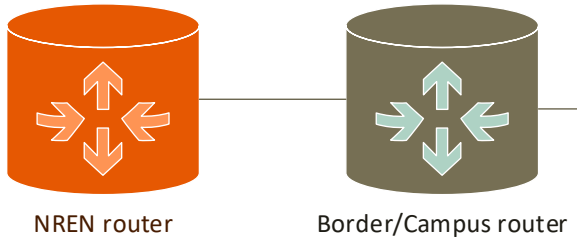
- Fiber cuts and other issues, do happen
- Misconfiguration
- Human errors
- Hardware/software issues

Hosts issues (briefly)

- Intel CPU's severely limited in IO
  - AMD Zen and IBM Power 8/9 much better
  - see Tristan Suerink's work ➔ https://indico.cern.ch/event/676324/contributions/2967991/attachments/1651172/2640923/Hepix-2018-Madison.pdf
- Receiving is harder then sending
  - Jumbo MTU's help
  - Capable NIC's with TCP offloading
- OS tuning
  - In the past 64byte TCP window
  - Great info on ESnet fasterdata website https://fasterdata.es.net/host-tuning/

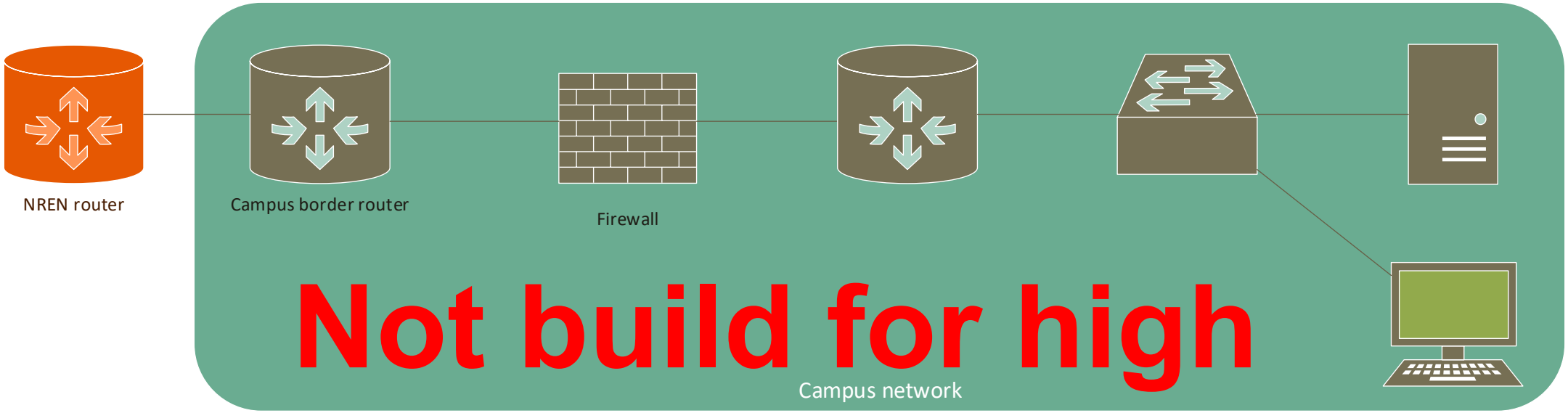
More on this in Ron's talk later today

From an NREN perspective



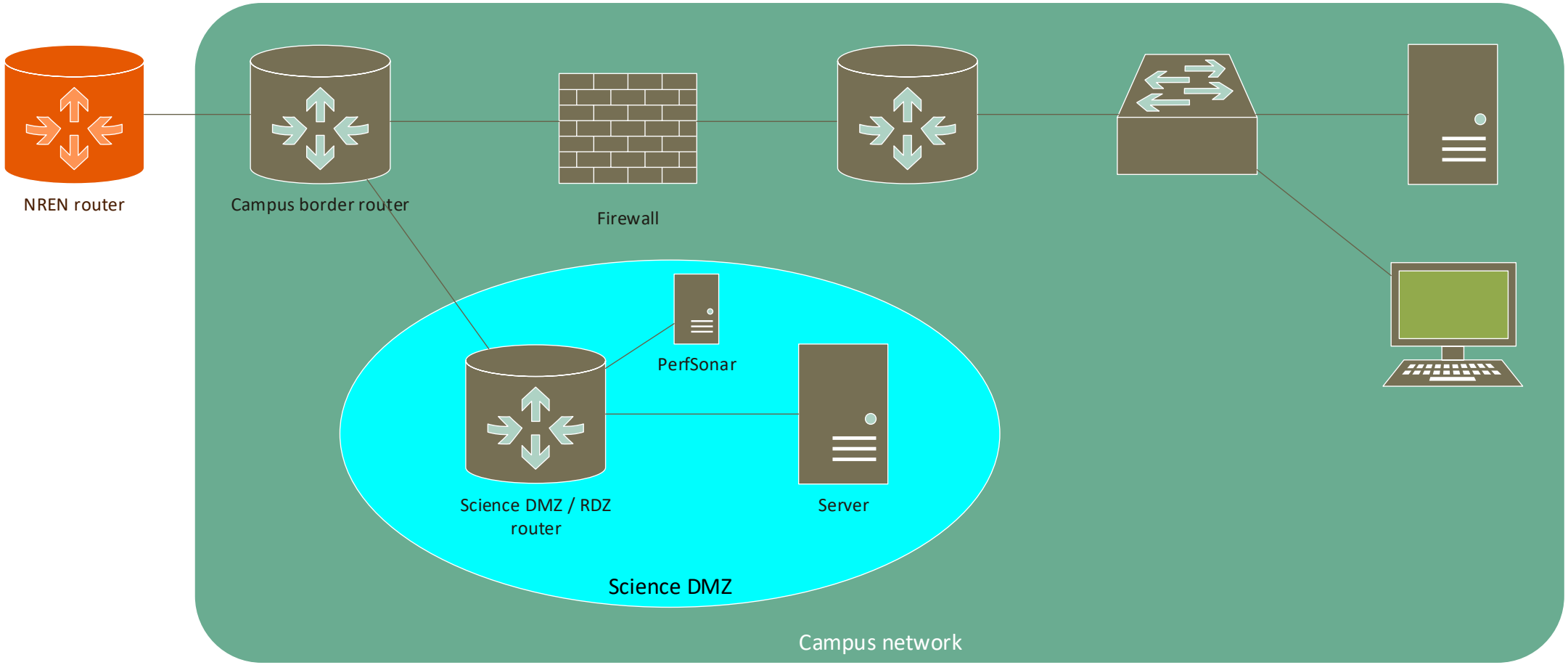
- Path to researcher is NOT limited by provided bandwidth
  - This link can be 10G, 100G, or multiples of…
- Both of these systems are on the customer site/on premise POP
- After this demarcation point, usually the Campus ICT is managing the network

Typical campus network (very simple)



Not build for high performance

Network architecture solution…

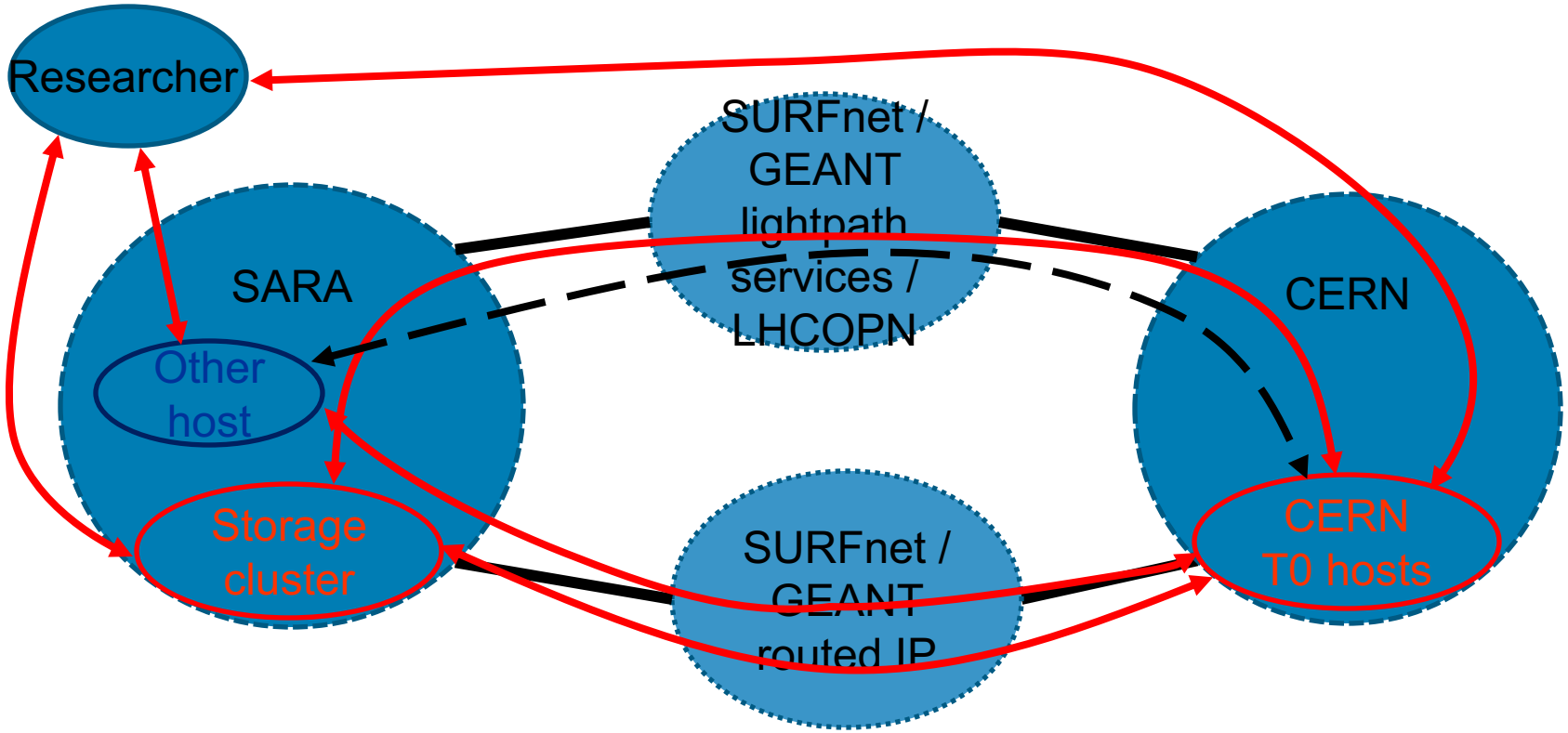


Network challenges from an connecting institute perspective (my previous SURFsara life)

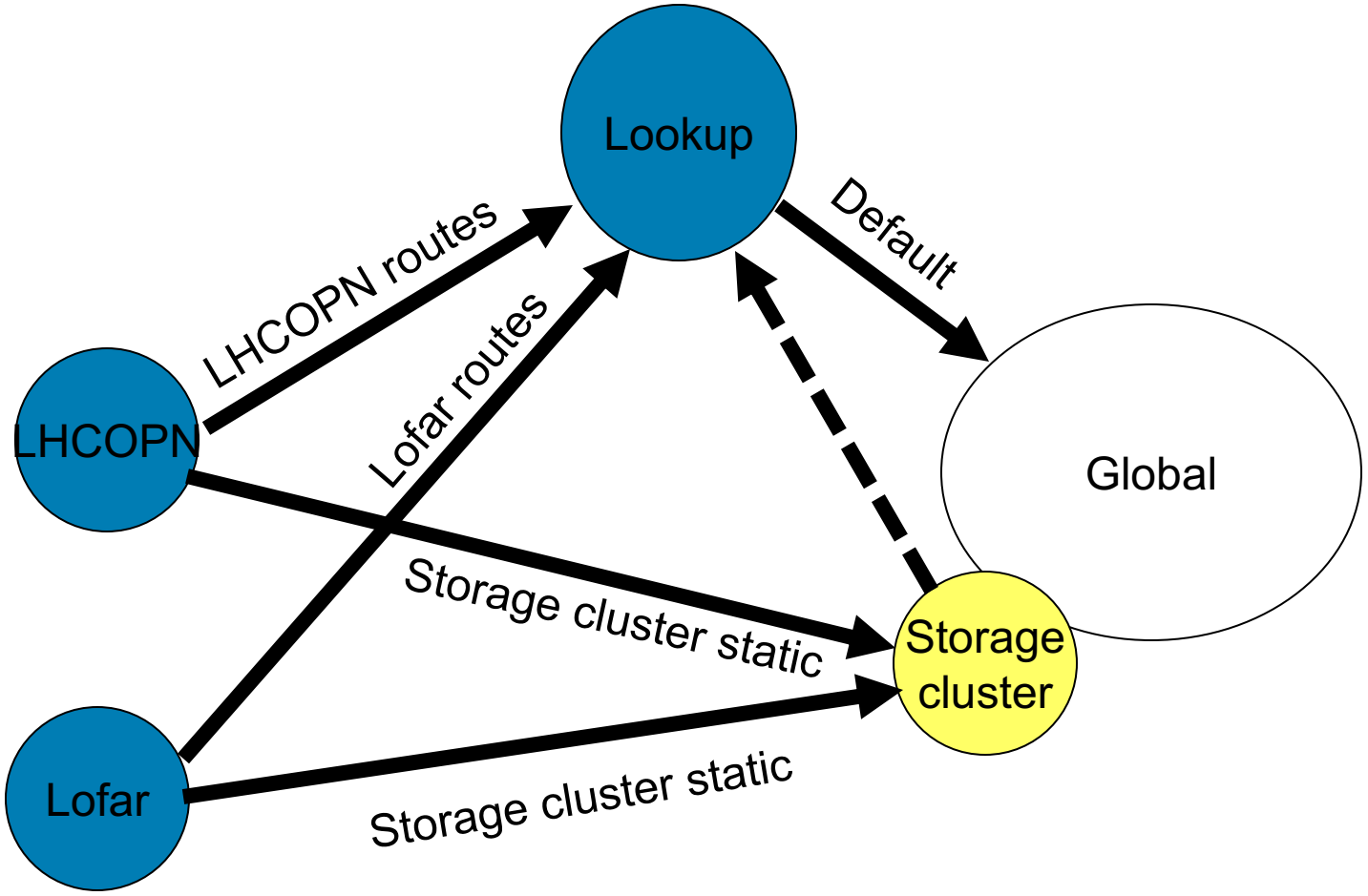
- Multiple paths to the outside world is a challenge.
  - Dedicated links for projects are nice but have a challenge
  - And often a use policy
  - Days with a default towards outside world where it was so easy and then dedicated links came along (OPN's / EPL's / Lightpaths)

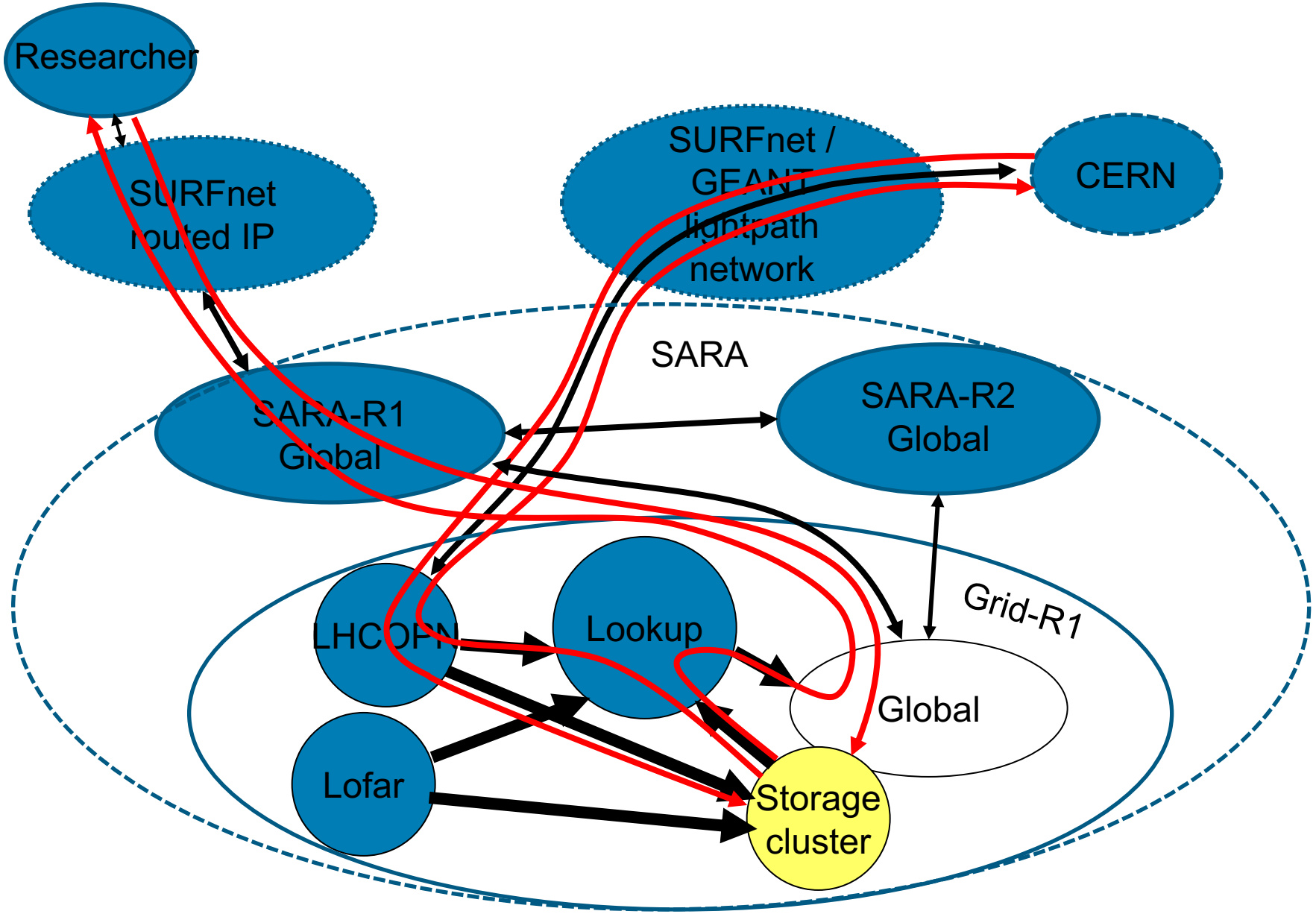
LHCOPN challenges

- Back in April 2007 (Munich) and January (Cambridge) description of problem/setup
- Struggling to keep traffic flows separate



VRFs to the rescue




Researcher
SURFnet
routed IP
SURFnet /
GEANT
lightpath
network
CERN
SARA
SARA-R1
Global
SARA-R2
Global
Grid-R1
LHCOPN
Lookup
Global
Lofar
Storage
cluster

In short

- Every OPN in its own VRF
- Every OPN has a static to SURFsara's storage cluster
- Routes from VRF exported into a look up VRF, so our storage cluster sees them all, takes the most specific and defaults to the global
- Since no routes between OPN VRFs are exchanged there is absolutely no risk of traffic leaking between OPNs
- In addition the global doesn't now anything of the OPN's
- Same construction used for the perfSONAR boxes, except they only see the LHCOPN and have the default option
- And it's scalable!!!

Data security while in transit

Your data while being transported are just Ethernet frames

- On SURFnet DWDM backbone:
  - Colored WDM wave
  - Possibly 10G signals muxed into ONT4 container

Fiber tapping is a serious risk

- Only need 1% of optical light, which is below variations caused by temperature/pressure/other
- Don’t need to interrupt link to setup
- (nearly) impossible to detect

Consider your data to be at risk while in transit ➔ protect it (encryption)

Data security while in transit - solutions

- Network
  - DCI or optical encryption (partial solution)
  - MACsec on links (even more partial solution) – link based

- Host
  - Encrypt your data transfer on your DTN node (protect while in transfer) - better
  - Encrypt your data before that (smaller security concern if someone hack your DTN, they have the encrypted data and no keys, good luck) - best

Recommendations for network architecture

- Campus side..
  - Make sure you are able to get to the end user with enough capacity
  - Be prepared for multiple routing tables
    - Interface membership determines first lookup
    - Rest is what you put in
  - Make sure you've hardware that can do multiple route lookups
  - Avoid QoS and Firewalls for bulk data transfers
  - Take fiber hygiene seriously

- NREN side
  - Know what your users are trying to do and what you can offer
  - Monitor your links
  - Get monitoring boxes (Perfsonar)

- – what can they do? (i.e., monitoring, knowledge sharing, consult on optimised architectures for a campus...)

Questions? Comments?

Pieter de Boer

E-mail: Pieter.deboer@surfnet.nl

Linkedin: https://www.linkedin.com/in/pgcdeboer

CC BY

SURF NET

Data transfers takes time

- Single flow data transfers
- Link typically maxes out at 97%
  - You need space for pre-amble and interframe gap
  - Only achievable with a lot of transfers over a link

Data set size	Speed					
10PB	Faster	1,333.33 Tbps	266.67 Tbps	66.67 Tbps	22.22 Tbps	925.83Gbps
1PB		133.33 Tbps	26.67 Tbps	6.67 Tbps	2.22 Tbps	92.58 Gbps
100TB		13.33 Tbps	2.67 Tbps	666.67 Gbps	222.22 Gbps	9.25 Gbps
10TB		1.33 Tbps	266.67 Gbps	66.67 Gbps	22.22 Gbps	925.92 Mbps
1TB		133.33 Gbps	26.67 Gbps	6.67 Gbps	2.22 Gbps	92.59 Mbps
100GB	100Gbit	13.33 Gbps	2.67 Gbps	666.67 Mbps	222.22 Mbps	9.26Mbps
10GB	10Gbit	1.33 Gbps	266.67 Mbps	66.67 Mbps	22.22 Mbps	0.93 Mbps
1GB	1GBit	133.33 Mbps	26.67 Mbps	6.67 Mbps	2.22 Mbps	0.09 Mbps
100MB	0.1Gbit or 100Mbit	13.33 Mbps	2.67 Mbps	0.67 Mbps	0.22 Mbps	0.01 Mbps
Time to transfer		1 Minute	5 Minutes	20 Minutes	1 Hour	1 Day