Linux Conference Australia 2022 – Virtual

NorNet – A Linux- and Open-Source-Software-based International Platform for Networking Research

Thomas Dreibholz (托马斯博士)
Simula Metropolitan Centre for Digital Engineering
dreibh@simula.no

January 15, 2022
Contents

- Motivation
- Multi-Homing and Multi-Path Transport
- The NorNet Testbed
- The Software: VMs, Containers and Multi-Homed Networking
- Conclusion
Overview:
Motivation

• Motivation
• Multi-Homing and Multi-Path Transport
• The NorNet Testbed
• The Software: VMs, Containers and Multi-Homed Networking
• Conclusion
„Classic“ Internet Communication

- Example: World-Wide Web

- Client ↔ Server Communication
  - 1 network interface per device → 1 IPv4 address
  - Communication with Transmission Control Protocol (TCP)
IPv6
- Devices are frequently IPv4/IPv6 dual stack
- Usually multiple addresses per interface

Mobility $\rightarrow$ address change

Devices with multiple interfaces
- Router
- Smartphone (LTE/UMTS, WLAN, Bluetooth?)
- Laptop (Ethernet, WLAN, LTE/UMTS?)
Multi-Homing and Multi-Path Transport

• Multi-Homing
  - Multiple interfaces (addresses)
  - **Redundancy** → Communication even when some paths fail

• Multi-Path Transport
  - Also utilise paths **simultaneously** → better throughput
  - **MPTCP:** Multi-Path TCP
  - **CMT-SCTP:** Concurrent Multi-Path Transfer for SCTP

**SCTP:** Stream Control Transmission Protocol
**TCP:** Transmission Control Protocol

Hot topic in research and standardisation!
Overview:
Multi-Homing and Multi-Path Transport

- Motivation
- Multi-Homing and Multi-Path Transport
- The NorNet Testbed
- The Software: VMs, Containers and Multi-Homed Networking
- Conclusion
Multi-Path Transport with MPTCP and CMT-SCTP

- Subflow ↔ path
- Fairness
  - Paths may overlap (fully or partially)
- Scheduling
  - Different path characteristics
    - Bandwidth
    - Latency and jitter
    - Packet loss

How can I use multi-path transport?
Multi-Path TCP (MPTCP)

- RFC 6824
- Features: multi-homing + multi-path transport
- Backwards-compatible to TCP
  - Communicate with old TCP implementations
  - Works (mostly) even over non-MPTCP middleboxes (e.g. NAT/PAT)
- Linux:
  - UC Louvain → https://www.multipath-tcp.org
- FreeBSD:
  - Swinburne → http://caia.swin.edu.au/newtcp/mptcp/

How can I use MPTCP under Linux?
Routing Tables and Routing Rules for Multi-Path Transport

• Example:
  - Device eth0 → ISP 1
  - Device eth1 → ISP 2

• Problem:
  - First default route (with lowest metric) via ISP 1
  - All traffic uses ISP 1 😞

• Solution:
  - **Routing rules**
  - Separate routing tables for each ISP
  - “Selector” for actually used table, per source address
A Linux Routing Rule Example

- **Configure eth0:**
  - `ip addr add 10.1.1.42/24 dev eth0`
  - `ip route add default via 10.1.1.1 dev eth0`
  - `ip route add 10.1.1.0/24 scope link dev eth0 table 1`
  - `ip route add default via 10.1.1.1 dev eth0 table 1`

- **Configure eth1:**
  - `ip addr add 172.16.1.42/24 dev eth1`
  - `ip route add 172.16.1.0/24 scope link dev eth1 table 2`
  - `ip route add default via 172.16.1.1 dev eth1 table 2`

- We have 2 new routing tables now! Set up routing rules based on source address:
  - `ip rule add from 10.1.1.42 table 1`
  - `ip rule add from 172.16.1.42 table 2`

- Table numbers difficult to remember? Set name mapping in `/etc/iproute2/rt_tables`!
The Resulting Configuration

- The routing rules: `ip rule show`
  
  0:   from all lookup local
  32764: from 172.16.1.42 lookup 2
  32765: from 10.1.1.42 lookup 1
  32766: from all lookup main
  32767: from all lookup default

- The name mappings: `cat /etc/iproute2/rt_tables`
  
<table>
<thead>
<tr>
<th>Table</th>
<th>IP Configuration</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>from all lookup local</td>
<td></td>
</tr>
<tr>
<td>255</td>
<td>local</td>
<td></td>
</tr>
<tr>
<td>254</td>
<td>main</td>
<td></td>
</tr>
<tr>
<td>253</td>
<td>default</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>unspec</td>
<td></td>
</tr>
</tbody>
</table>

- Table #1: `ip route show table 1`
  
  default via 10.1.1.1 dev eth0
  10.1.1.0/24 dev eth0 scope link

- Table #2: `ip route show table 2`
  
  default via 172.16.1.1 dev eth1
  172.16.1.0/24 dev eth1 scope link

- Table “main” (254): `ip route show table main`
  
  default via 10.1.1.1 dev eth0
  172.16.1.0/24 dev eth1 proto kernel \ scope link src 172.16.1.42
  10.1.1.0/24 dev eth0 proto kernel \ scope link src 10.1.1.42
What about Routing Rules with IPv6?

- It works with IPv6 as well, of course!
  
  - `ip -6 addr add 3ffe:cafe:affe:1234::2a/64 dev eth0`
  
  - `ip -6 route add default via 3ffe:cafe:affe:1234::1 dev eth0`
  
  - `ip -6 route add 3ffe:cafe:affe:1234::/64 scope link dev eth0 table 1`
  
  - `ip -6 route add default via 3ffe:cafe:affe:1234::1 dev eth0 table 1`
  
  - `ip -6 addr add 3ffe:dead:beef:ffff::2a/64 dev eth1`
  
  - `ip -6 route add 3ffe:dead:beef:ffff::/64 scope link dev eth1 table 2`
  
  - `ip -6 route add default via 3ffe:dead:beef:ffff::1 dev eth1 table 2`
  
  - `ip -6 rule add from 3ffe:cafe:affe:1234::2a table 1`
  
  - `ip -6 rule add from 3ffe:dead:beef:ffff::2a table 2`

- Depending on source address, a packet leaves via network 1 or network 2
  
  - MPTCP (and CMT-SCTP) will make this choice, depending on subflow

- Connections can even have IPv4- and IPv6 subflows simultaneously!
Finally: Testing MPTCP

- First, boot a MPTCP-enabled kernel
  - See https://multipath-tcp.org for sources
  - MPTCP enabled by default → all TCP connections are MPTCP-capable!
- Configure (and check) the routing
  - Connect to two (or more) ISPs, if possible
  - IPv4 + IPv6 may also give you partially independent paths
  - Just 1 IP address? → multiple paths to a multi-homed remote side!
- Test:
  - NetPerfMeter → https://www.uni-due.de/~be0001/netperfMeter/
  - Wireshark/T-Shark → https://www.wireshark.org
  - ...
Overview:
The NorNet Testbed

- Motivation
- Multi-Homing and Multi-Path Transport
- The NorNet Testbed
- The Software: VMs, Containers and Multi-Homed Networking
- Conclusion
Testing Multi-Path Transport (1)
The First Step – A Lab Setup

- Surprisingly big effort:
  - Strange effects of cheap network components: „It's only cheap on the paper!“
  - Debugging of SCTP in FreeBSD

- But valuable:
  - Prior simulations were useful!
  - Bugfixes for the FreeBSD community
  - Open Source software „NetPerfMeter“
  - Learning effects and new ideas!

Internet protocols → testbed in the Internet!
Testing Multi-Path Transport (2)
Real Internet: 3 Cities and 2 Continents

- 3 connected lab setups
  - Establishment of an international cooperation
  - Essen, Burgsteinfurt (FH Münster), Haikou 海口 (Hainan University)
- Very interesting scenario:
  - CMT-SCTP and MPTCP evaluation
  - Very different path characteristics
    → Ideas for further experiments

Many new ideas!

Now really big: NorNet testbed!
Testing Multi-Path Transport (3)
The NorNet Testbed

- **NorNet Core**
  - Cable, up to 4 providers, IPv4+IPv6 (fibre, “consumer-grade” DSL, etc.)
  - Hosts for virtual machines
  - 21 locations (11 in Norway, 10 abroad)

[ simula.research laboratory ]

- **NorNet Edge**
  - Embedded systems (3 generations) running customised Debian Linux
  - Up to 4× 2G/3G/4G (+ 1× CDMA, 1× Ethernet)
  - Hundreds of locations (in Norway)

https://www.nntb.no
## NorNet Core Site Deployment Status (October 2016)

<table>
<thead>
<tr>
<th>No.</th>
<th>Site</th>
<th>ISP 1</th>
<th>ISP 2</th>
<th>ISP 3</th>
<th>ISP 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Simula Research Laboratory</td>
<td>Uninett</td>
<td>Kvantel</td>
<td>Telenor</td>
<td>PowerTech</td>
</tr>
<tr>
<td>2</td>
<td>Universitetet i Oslo</td>
<td>Uninett</td>
<td>Broadnet</td>
<td>PowerTech</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Høgskolen i Gjøvik</td>
<td>Uninett</td>
<td>PowerTech</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Universitetet i Tromsø</td>
<td>Uninett</td>
<td>Telenor</td>
<td>PowerTech</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Universitetet i Stavanger</td>
<td>Uninett</td>
<td>Altibox</td>
<td>PowerTech</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Universitetet i Bergen</td>
<td>Uninett</td>
<td>BKK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Universitetet i Agder</td>
<td>Uninett</td>
<td>PowerTech</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Universitetet på Svalbard</td>
<td>Uninett</td>
<td>Telenor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Universitetet i Trondheim</td>
<td>Uninett</td>
<td>PowerTech</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Høgskolen i Narvik</td>
<td>Uninett</td>
<td>Broadnet</td>
<td>PowerTech</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Høgskolen i Oslo og Akershus</td>
<td>Uninett</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Karlstads Universitet</td>
<td>SUNET</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Universität Kaiserslautern</td>
<td>DFN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Universität Duisburg-Essen</td>
<td>DFN</td>
<td>Versatel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Hainan University 海南大学</td>
<td>CERNET</td>
<td>China Unicom</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>The University of Kansas</td>
<td>KanREN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Korea University 高麗大學校</td>
<td>KREONET</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>National ICT Australia (NICTA)</td>
<td>AARNet</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>HAW Hamburg</td>
<td>DFN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Technische Universität Darmstadt</td>
<td>DFN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Haikou Cg. of Econ. 海口经济学院</td>
<td>China Telecom</td>
<td>CERNET</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**
- IPv4 and IPv6
- IPv4 only (ISP without IPv6 support)
- IPv4 only (site’s network without IPv6 support)
- ISP negotiation in progress

[https://www.nntb.no/pub/nornet-configuration/NorNetCore-Sites.html](https://www.nntb.no/pub/nornet-configuration/NorNetCore-Sites.html)
Routing Visualisation

HiPerConTracer Observed Routes and Autonomous Systems, from July 1 to July 31, 2016

Work in Progress!
Remote Systems

Our servers may be really remote!

The "road" to Longyearbyen på Svalbard, 78.2°N
Overview:
The Software: VMs, Containers and Multi-Homed Networking

- Motivation
- Multi-Homing and Multi-Path Transport
- The NorNet Testbed
- The Software: VMs, Containers and Multi-Homed Networking
- Conclusion
Virtualisation

“Anything that can go wrong, will go wrong.”
[Murphy’s law]

- Experimentation software is experimental
- How to avoid software issues making a remote machine unusable?
- Idea: virtualisation
  - Lightweight, stable software setup: **Ubuntu Server** 18.04 LTS → 20.04/22.04 LTS
  - KVM (Kernel-based Virtual Machine)
  - Other software runs in VMs:
    - Tunnelbox (router) VM
    - Research Node VMs
  - In case of problem: manual/automatic restart or reinstall of VM
Physical Machine Setup

- Ubuntu Server LTS – due to long-term support
- Customisation:
  - File system: ReiserFS 3
    - Ext4 resilience is really awful – manual fsck at remote machines 😞
    - BTRFS has nice features, but awful performance for hosting VMs
    - => ReiserFS!
      - very reliable (it never killed the data of a machine)
      - good performance, also for hosting VMs
    - Newer installations: BTRFS (after removal of ReiserFS from installer)
  - Virtualisation: now KVM
    - Originally: VirtualBox (custom package with Open Source VNC, instead of Oracle’s closed source blob)
Tunnelbox – The Router (1)

- **Tunnelbox**
  - Router at each site
  - Handles all network communication, over multiple ISPs
    - 1 public IPv4 address (+ 1 public IPv6 address) per site and ISP
    - Tunnels among the sites (GRE-over-IPv4, IPv6-over-IPv6)
    - Own systematic addressing scheme
      - IPv6: 2001:700:4100:<Provider><Site>::<Node>/64
  - Direct communication between sites over tunnels
  - Communication between sites and Internet over Simula’s site
- Security + avoiding legal issues (DE: “Mitstörerhaftung” ...)
Tunnelbox – The Router (2)

- IP addresses are difficult to remember
  - DNS setup (bind9) with private TLD “.nornet”
  - Convenience 1: easy naming scheme:
    - fjellrev.telenor.unis.nornet: node “fjellrev” with ISP “Telenor” at “UNIS”
    - borbeck.ude.nornet: node “borbeck” with primary ISP at “UDE”
    - østhorn.kvantel.simula.nornet → xn--sthorn-9xa.kvantel.simula.nornet
    - 三亚.cnunicom.hu.nornet → xn--ehqrn.cnunicom.hu.nornet
  - Convenience 2: SSHFP (SSH key fingerprints) and LOC (geolocation) RRs
- Squid HTTP proxy
  - Caching HTTP accesses (mainly: package updates)
  - If necessary: forward every request to Simula (to avoid legal issues)
Research Node

• “Usual” research node:
  – A VM (usually KVM), managed by PlanetLab Central (PLC)-based software
    • 2.5 GiB RAM, 1-2 cores
    • Fedora Core 25, Linux kernel v4.4 with MPTCP v0.94 + API patch
  – User gets a “sliver” of the research node → LXC container
    • “Own” Fedora installation, with development tools, T-Shark, ...
    • “Own” IP addresses (IPv4 + IPv6, for each ISP of the site)
    • Root permission (only within the LXC container, with limitations)
    • But: A slice is basically a BTRFS clone of a template
      – File duplicates only necessary upon changes
      – Very lightweight setup per user
  – Custom VMs as research nodes are also possible (currently requires manual setup)
The Different Entities: Server, Node, Slice and Sliver

Server (physical)

Node (virtual)
• Sliver hu_multipath
• Sliver srl_tutorial
• Sliver ntnu_test
• Sliver due_rserpool
• ...

Node (virtual)
KVM VM

Slice:
• User list
• Node list

Sliver = an instance of a slice on a node

ssh <Slice>@<Node>

Forwarding to sliver!

KVM VM

LXC containers
A Look into a Sliver

- SSH login to the sliver
- Here: 2 ISPs
  - 2x IPv4
  - 2x IPv6
- Kernel with MPTCP
- For security:
  - SSH key fingerprints
  - Superuser ("su" or "sudo")
  - dnf install ...
  - tcpdump -i eth0 ...
  - tshark -i eth0 ...
Overview:

Conclusion

- Motivation
- Multi-Homing and Multi-Path Transport
- The NorNet Testbed
- The Software: VMs, Containers and Multi-Homed Networking
- Conclusion
Conclusion and Future Work

• Multi-homed devices increasingly widespread → multi-path transfer
• Realistic, large-scale Internet testbed infrastructure is available: NorNet
• NorNet Core is an open testbed!
  - Interested in using NorNet? Just ask!

NorNet

• Future work: extend NorNet Core's scope beyond multi-path transport topic:
  - Network Function Virtualisation (NFV) and Software-Defined Networking (SDN)
  - Cloud Computing and applications

See https://www.nntb.no for more information!
Open Source Software

- NorNet Project: https://www.nntb.no
  - NorNet Core Site: https://www.nntb.no/pub/nornet-configuration/NorNetCore-Sites.html
  - NorNet Software: https://www.nntb.no/software/
    - Management: https://github.com/simula/nornet-control
    - Research Node: https://benlomond.nntb.no/releases/
- HiPerConTracer: https://github.com/dreibh/hipercontracer
- NetPerfMeter: https://www.uni-due.de/~be0001/netperfmeter/
- RSPLIB: https://www.uni-due.de/~be0001/rserpool/
- Linux Multi-Path TCP: https://multipath-tcp.org/
- Multi-Path TCP Page: https://www.uni-due.de/~be0001/mptcp/
- SCTP Project Page: http://www.iem.uni-due.de/~dreibh/sctp/
- Simula Research Laboratory: https://www.simula.no
Literature (1)


- Dreibholz, T.: "NorNet at Hainan University in 2021: Getting Started with NorNet Core – A Remote Tutorial", Tutorial at Hainan University, College of Information Science and Technology (CIST), Haikou, Hainan/People's Republic of China, January 2021.


- Dreibholz, T.: "NorNet at the University of Sydney: From Simulations to Real-World Internet Measurements for Multi-Path Transport Research", Invited Talk at University of Sydney, Sydney, New South Wales/Australia, January 2019.

Literature (2)


Also see https://www.nntb.no/publications/!
Thank you for your attention!
Any questions?

Thomas Dreibholz
dreibh@simula.no
https://www.simula.no/people/dreibh