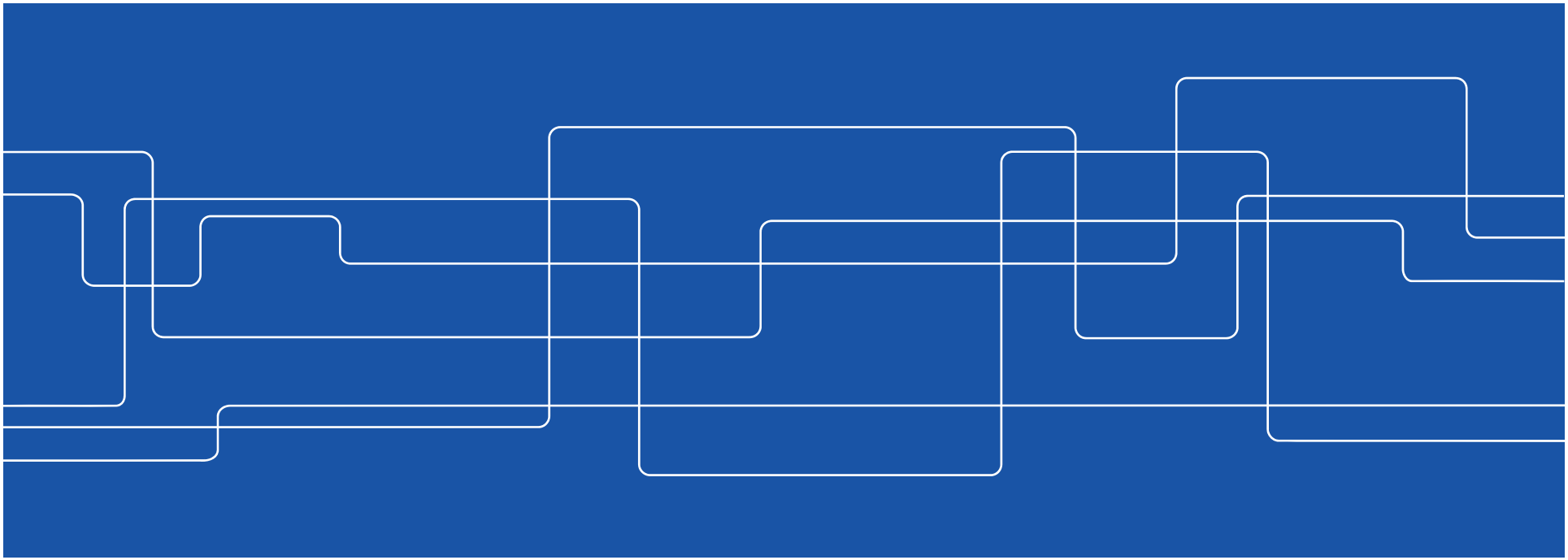
II2202: Presenting your data

prof. Gerald Q. Maguire Jr. http://people.kth.se/~maguire/

School of Information and Communication Technology (ICT), KTH Royal Institute of Technology

II2202 Fall 2015, Period 1 and Periods 1&2 2015.08.21



Presenting your data

1. Data visualization for **yourself** – to understand your data, to “listen to your data”, to exploit your data, …
2. **Presenting it to your audience** – so that you can explain what you have found in the data, so that you can facilitate their understanding the data, so that they can make use of your data to build upon it, …

Presenting information with images

"A picture is worth a thousand words."

-- Popular saying

Pictures, graphs, flow charts, UML, state machines, … can convey an enormous amount of information if used well.

Consider "a wink" at a party

Why use graphical presentations?

- Very compact – you can present at lot of data in a small space – in contrast to a table
- To bring out difference and make comparisons
- To help abstract a general (abstract) “picture” (conception) from the data
- Many people are good at seeing **patterns** in visual scenes
- To achieve clarity and objectivity
- To support your text (i.e., to help you tell your story)

A graph is a encoding, when you look at it you need to visually decode it

“When a graph is made, quantitative and categorical information is encoded by a display method. Then the information is visually decoded. This visual perception is a vital link. No matter how clever the choice of the information, and no matter how technologically impressive the encoding, a visualization fails if the decoding fails. Some display methods lead to efficient, accurate decoding, and others lead to inefficient, inaccurate decoding. It is only through scientific study of visual perception that informed judgments can be made about display methods. The display methods of Elements rest on a foundation of scientific enquiry.”

From the preface of William S. Cleveland’s
“The Elements of Graphing Data” [Cleveland 1989]

Edward Tufte's books

Examples of how to present information well and even beautifully:

- Beautiful Evidence [Tufte 2006]
- The Visual Display of Quantitative Information [Tufte 2001]
- Visual Explanations: Images and Quantities, Evidence and Narrative [Tufte 1997]
- Envisioning Information [Tufte 2008]

http://www.edwardtufte.com/tufte/index

Measuring a FASP file transfer

Inspired by National Center for Biotechnology Information's 'Aspera Transfer Guide' [NCBI 2014]

Downloaded and installed Aspera Connect software from: http://downloads.asperasoft.com/connect2/

Transferred a 1Gbyte test file – while collecting data using:

tcpdump –I eth0 –w /tmp/xxxxxxx

Start ascp to transfer 1G from test server

```
maguire@ccsser2:~/.aspera/connect/bin> ls
ascp  asperaconnect  asperaconnect.bin  asperacrypt  asunprotect  plugins
maguire@ccsser2:~/.aspera/connect/bin> env ASPERA_SCP_PASS=demoaspera
./ascp -L- -T -l100m aspera@demo.asperasoft.com:aspera-test-dir-large/1GB /tmp/
LOG Aspera Connect version 3.6.0.106805

LOG Alternate log directory: "-"
LOG Configuration: using v2 configuration file
"/home/maguire/.aspera/connect/etc/aspera.conf", user -
LOG Initializing FASP version 3.5.4.103990, license max rate=(unlimited), account
no.=1, license no.=1 product=6
LOG Configured symlink actions: create=1, follow=1, follow_wide=0, skip=0
LOG [asssh] remote host-key fingerprint f34dfcda4110604e4ecf53e6e18c6559a38cbb43
LOG [asssh] authentication succeeded, proceeding.
LOG changing session job size from 0 to 2 to match server configuration
```

-L-	log to standard output
-T	disable encryption
-l100m	maximum bandwidth of request – in this case 100 Mbps

FASP session starts

LOG FASP Session Start uuid=a9063e44-f785-4bca-8e71-3eaa20a64b32 op=recv status=started source=aspera-test-dir-large/1GB (1) dest=/tmp source_prefix=- local=130.237.209.248:42132 peer=198.23.89.123:33001 tcp_port=22 os="Linux 3.7.10-1.45-desktop #1 SMP PREEMPT" ver=3.5.4.103990 lic=6:1:1 peeros="Linux 2.6.32-504.3.3.el6.x86_64 #1 SMP W" peerver=3.5.4.100392 peerlic=10:1:22001 proto_sess=20002 proto_udp=20000 proto_bwmeas=20000 proto_data=20008

LOG FASP Session Params uuid=a9063e44-f785-4bca-8e71-3eaa20a64b32 userid=0 user="aspera" targetrate=100000000 minrate=0 rate_policy=fair cipher=none resume=0 create=0 ovr=1 times=0 precalc=yes mf=0 mf_path=- mf_suffix=.aspera-inprogress partial_file_suffix=- files_encrypt=no files_decrypt=no file_csum=none dgram_sz=0 prepostcmd=- tcp_mode=no rtt_auto=yes cookie="-" vl_proto_ver=1 peer_vl_proto_ver=1 vl_local=0 vlink_remote=0 vl_sess_id=3840 srcbase=- rd_sz=0 wr_sz=0 cluster_num_nodes=1 cluster_node_id=0 range=0-0 keepalive=no test_login=no proxy_ip=- net_rc_alg=alg_delay exclude_older/newer_than=0/0

LOG Measured pMTU: 1492 Bytes, start_brtt: 174 ms

LOG datagram size 1492B, block size 1452B, path MTU 1492B

Intermediate output

```
1GB                                   0% 2904             --:--     LOG FASP Transfer Start uuid=a9063e44-f785-4bca-8e71-3eaa20a64b32 op=recv
status=started file="/tmp/1GB" size=1048576000 start_byte=0 rate=100.00Mbps loss=0.00 rexreqs=0 overhead=0 mtime="2014-04-10 19:49"
LOG Receiver bl t/o/r/d/ts=2223/2223/0/0/1970 rex_rtt l/h/s/o=0/0/174/8 ooo_rtt l/h/s/o=0/0/174/8 rate_rtt b/l/h/s/r/f=174/175/177/176/0/2 ctl bm/bs=0/0
rex n/s/q/v/a/r=0/0/0/0/0/0 bl l/d/o/r/a/x/dl/df/dm/ds=0/0/0/0/0/0/0/0/0/0 disk l/h/b=0/1/0 vlink lq/lo/rq/ro=0/0/0/0 rate
t/m/c/n/vl/vr/r=100000000/0/59680000/59680000/100000000/100000000/100000000 prog t/f/e=3227796/3227796/1000221 rcvD=0
LOG Receiver DS Qs ds/n/rq/ao/ap/rd/ru/no/po/pc/do=1/0/0/0/0/0/1/0/0/0/0 Rs i/o=1/1 mgmt backlog i/s/n =
1GB                                  22%  221MB 97.3Mb/s    01:10 ETALOG Receiver bl t/o/r/d/ts=168010/168010/0/0/167517 rex_rtt l/h/s/o=0/0/174/8
ooo_rtt l/h/s/o=0/0/174/8 rate_rtt b/l/h/s/r/f=174/175/182/175/0/1 ctl bm/bs=0/0 rex n/s/q/v/a/r=0/0/0/0/0/0 bl l/d/o/r/a/x/dl/df/dm/ds=0/0/0/0/0/0/0/0/0/0 disk
l/h/b=0/1/0 vlink lq/lo/rq/ro=0/0/0/0 rate t/m/c/n/vl/vr/r=100000000/0/100000000/100000000/100000000/100000000/100000000 prog
t/f/e=243950520/243950520/21060992 rcvD=0
LOG Receiver DS Qs ds/n/rq/ao/ap/rd/ru/no/po/pc/do=1/0/0/0/0/0/1/0/0/0/0 Rs i/o=1/1 mgmt backlog i/s/n =
1GB                                  45%  453MB 97.3Mb/s    00:48 ETALOG Receiver bl t/o/r/d/ts=336110/336110/0/0/335617 rex_rtt l/h/s/o=0/0/174/8
ooo_rtt l/h/s/o=0/0/174/8 rate_rtt b/l/h/s/r/f=174/174/182/176/0/1 ctl bm/bs=0/0 rex n/s/q/v/a/r=0/0/0/0/0/0 bl l/d/o/r/a/x/dl/df/dm/ds=0/0/0/0/0/0/0/0/0/0 disk
l/h/b=0/1/0 vlink lq/lo/rq/ro=0/0/0/0 rate t/m/c/n/vl/vr/r=100000000/0/100000000/100000000/100000000/100000000/100000000 prog
t/f/e=488031720/488031720/41122269 rcvD=0
LOG Receiver DS Qs ds/n/rq/ao/ap/rd/ru/no/po/pc/do=1/0/0/0/0/0/1/0/0/0/0 Rs i/o=1/1 mgmt backlog i/s/n =
1GB                                  68%  686MB 97.3Mb/s    00:27 ETALOG Receiver bl t/o/r/d/ts=504210/504210/0/0/503717 rex_rtt l/h/s/o=0/0/174/8
ooo_rtt l/h/s/o=0/0/174/8 rate_rtt b/l/h/s/r/f=174/174/182/176/0/1 ctl bm/bs=0/0 rex n/s/q/v/a/r=0/0/0/0/0/0 bl l/d/o/r/a/x/dl/df/dm/ds=0/0/0/0/0/0/0/0/0/0 disk
l/h/b=0/1/0 vlink lq/lo/rq/ro=0/0/0/0 rate t/m/c/n/vl/vr/r=100000000/0/100000000/100000000/100000000/100000000/100000000 prog
t/f/e=732112920/732112920/61183179 rcvD=0
LOG Receiver DS Qs ds/n/rq/ao/ap/rd/ru/no/po/pc/do=1/0/0/0/0/0/1/0/0/0/0 Rs i/o=1/1 mgmt backlog i/s/n =
1GB                                  91%  919MB 97.3Mb/s    00:07 ETALOG Receiver bl t/o/r/d/ts=672310/672310/0/0/671817 rex_rtt l/h/s/o=0/0/174/8
ooo_rtt l/h/s/o=0/0/174/8 rate_rtt b/l/h/s/r/f=174/174/182/175/0/1 ctl bm/bs=0/0 rex n/s/q/v/a/r=0/0/0/0/0/0 bl l/d/o/r/a/x/dl/df/dm/ds=0/0/0/0/0/0/0/0/0/0 disk
l/h/b=0/1/0 vlink lq/lo/rq/ro=0/0/0/0 rate t/m/c/n/vl/vr/r=100000000/0/100000000/100000000/100000000/100000000/100000000 prog
t/f/e=976194120/976194120/81244539 rcvD=0
LOG Receiver DS Qs ds/n/rq/ao/ap/rd/ru/no/po/pc/do=1/0/0/0/0/0/1/0/0/0/0 Rs i/o=1/1 mgmt backlog i/s/n =
1GB                                 100% 1000MB 97.3Mb/s    01:26
```

FASP transfer stops

LOG FASP Transfer Stop uuid=a9063e44-f785-4bca-8e71-3eaa20a64b32 op=recv status=success file="/tmp/1GB" size=1048576000 start_byte=0 rate=96.36Mbps elapsed=87.05s loss=0.00 rexreqs=0 overhead=0 mtime="2014-04-10 19:49"

LOG Receiver bl t/o/r/d/ts=722162/722160/0/2/722160 rex_rtt l/h/s/o=0/0/174/8 ooo_rtt l/h/s/o=0/0/174/8 rate_rtt b/l/h/s/r/f=174/175/178/175/0/1 ctl bm/bs=0/0 rex n/s/q/v/a/r=0/0/0/0/0/0 bl l/d/o/r/a/x/dl/df/dm/ds=0/0/0/0/0/2/0/0/0/0 disk l/h/b=0/1/0 vlink lq/lo/rq/ro=0/0/0/0 rate t/m/c/n/vl/vr/r=100000000/0/100000000/100000000/100000000/100000000/100000000 prog t/f/e=1048576000/1048576000/87494969 rcvD=1

LOG Receiver DS Qs ds/n/rq/ao/ap/rd/ru/no/po/pc/do=0/0/0/0/0/0/0/0/0/0/0 Rs i/o=1/1 mgmt backlog i/s/n =

Completed: 1024000K bytes transferred in 87 seconds

(95875K bits/sec), in 1 file.

LOG FASP Session Stop uuid=a9063e44-f785-4bca-8e71-3eaa20a64b32 op=recv status=success source=aspera-test-dir-large/1GB (1) dest=/tmp source_prefix=- local=130.237.209.248:42132 peer=198.23.89.123:33001 tcp_port=22 os="Linux 3.7.10-1.45-desktop #1 SMP PREEMPT" ver=3.5.4.103990 lic=6:1:1 peeros="Linux 2.6.32-504.3.3.el6.x86_64 #1 SMP W" peerver=3.5.4.100392 peerlic=10:1:22001 proto_sess=20002 proto_udp=20000 proto_bwmeas=20000 proto_data=20008

LOG FASP Session Params uuid=a9063e44-f785-4bca-8e71-3eaa20a64b32 userid=0 user="aspera" targetrate=100000000 minrate=0 rate_policy=fair cipher=none resume=0 create=0 ovr=1 times=0 precalc=yes mf=0 mf_path=- mf_suffix=.aspera-inprogress partial_file_suffix= files_encrypt=no files_decrypt=no file_csum=none dgram_sz=0 prepostcmd=- tcp_mode=no rtt_auto=yes cookie="-" vl_proto_ver=1 peer_vl_proto_ver=1 vl_local=0 vlink_remote=0 vl_sess_id=3840 srcbase=- rd_sz=0 wr_sz=0 cluster_num_nodes=1 cluster_node_id=0 range=0-0 keepalive=no test_login=no proxy_ip=- net_rc_alg=alg_delay exclude_older/newer_than=0/0

LOG FASP Session Statistics [Receiver] id=a9063e44-f785-4bca-8e71-3eaa20a64b32 delay=176ms rex_delay=8ms ooo_delay=8ms solicited_rex=0.00% rcvd_rex=0.00% rcvd_dups=0.00% ave_xmit_rate 98.63Mbps effective=100.00% effective_rate=98.63Mbps (detail: good_blks 722160 bl_total 722162 bl_orig 722160 bl_rex 0 dup_blks 0 dup_last_blks 0 drop_blks_xnf 2) (sndr ctl: sent 112 rcvd 112 lost 0 lost 0.00%) (rcvr ctl: sent 879 rcvd 877 lost 2 lost 0.23%) (rex ctl: sent 0 rcvd 0 lost 0 lost 0.00%) (progress: tx_bytes 1048576000 file_bytes 1048576000 tx_time 87494969) rex_xmit_blks 0 xmit_total 722162 rex_xmit_pct 0.00%

Completed: 1024000K bytes transferred in 87 seconds (95875K bits/sec), in 1 file

delay=176ms rex_delay=8ms
ave_xmit_rate 98.63Mbps
(sndr ctl: sent 112 rcvd 112 lost 0 lost 0.00%)
(rcvr ctl: sent 879 rcvd 877 lost 2 lost 0.23%)

Final transfer statistics

```
LOG ======= File Transfer statistics =======
LOG ------- Source statistics -------
LOG Source argument scans attempted           :         1
LOG - Source argument scans completed         :         1
LOG Source path scans attempted               :         1
LOG - Source path scans failed                :         0
LOG - Source path scans skipped since irregular :       0
LOG - Source path scans excluded              :         0
LOG - Source directory scans completed        :         0
LOG - Source file scans completed             :         1
LOG Source directory creates attempted        :         0
LOG - Source directory creates failed         :         0
LOG - Source directory created or existed     :         0
LOG Source file transfers attempted           :         1
LOG - Source file transfers failed            :         0
LOG - Source file transfers passed            :         1
LOG - Source file transfers skipped           :         0
LOG Source bytes transferred                  : 1048576000
LOG ======= end File Transfer statistics =======
```

Wireshark: UDP conversion

client server

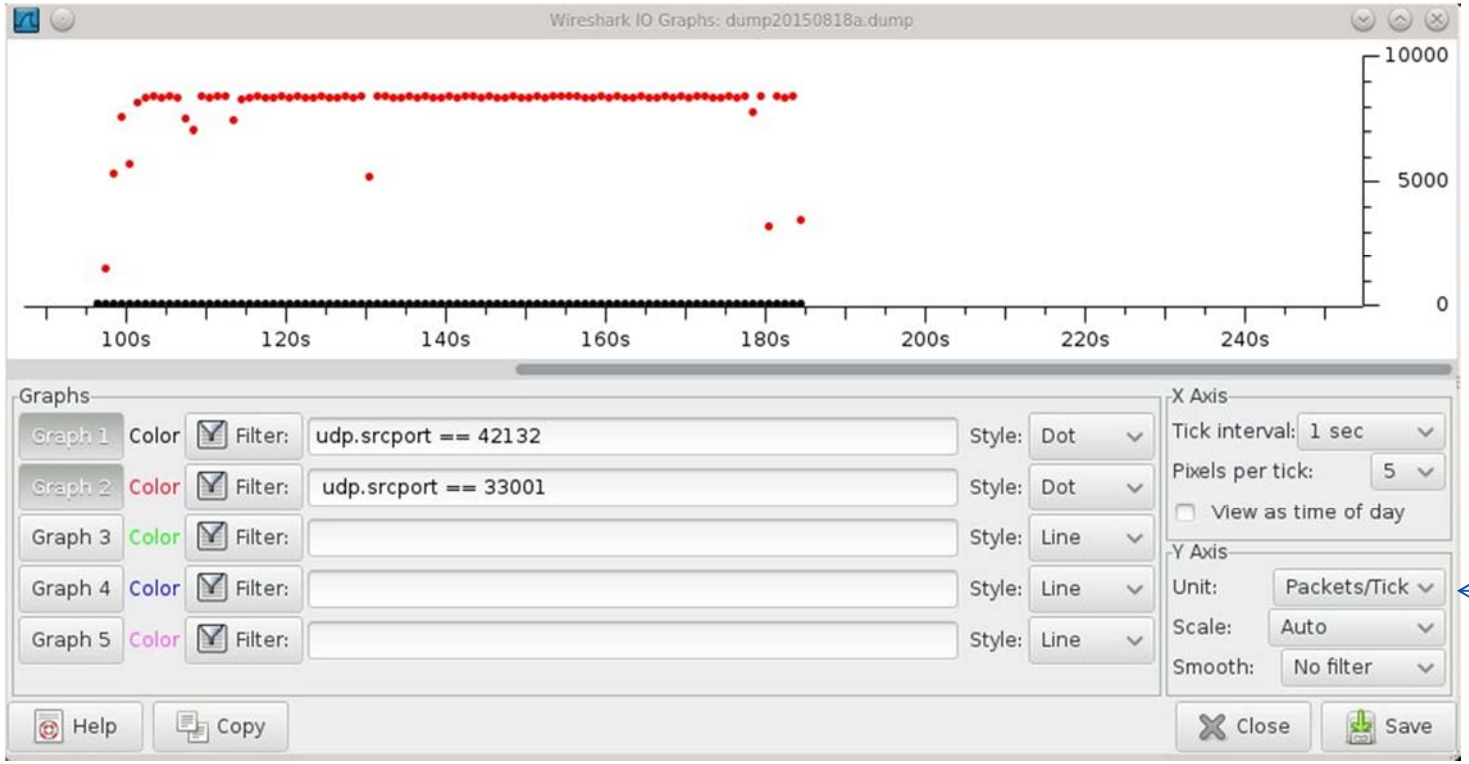
A: 130.237.209.248:42132 ↔ B: 198.23.89.123:33001

Packets: 703,728 Bytes: 1,058,302,232

	Packets	Bytes	bps
A→B	961	88518	8,058.85
A←B	702,767	1,058,213,714	96,341,791.88

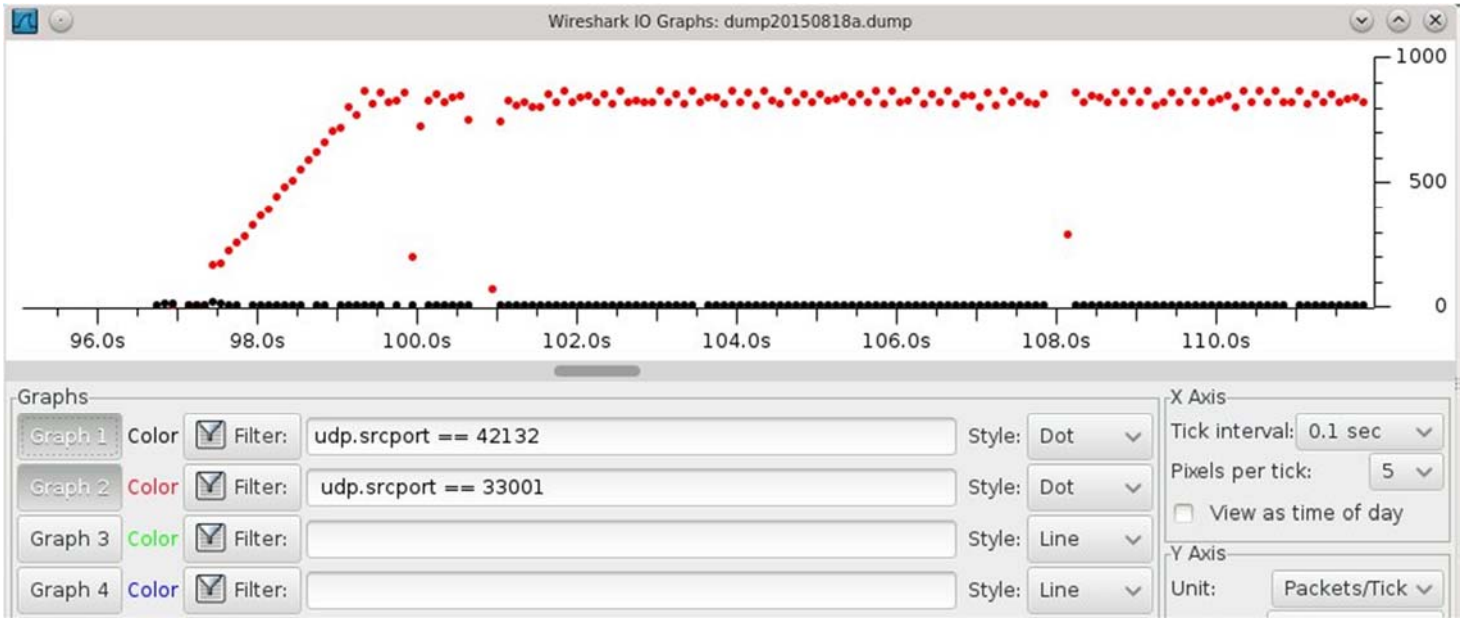
File size = 1,048,576,000 bytes

33001 is the source UDP port of server
42132 is the source UDP port of the client



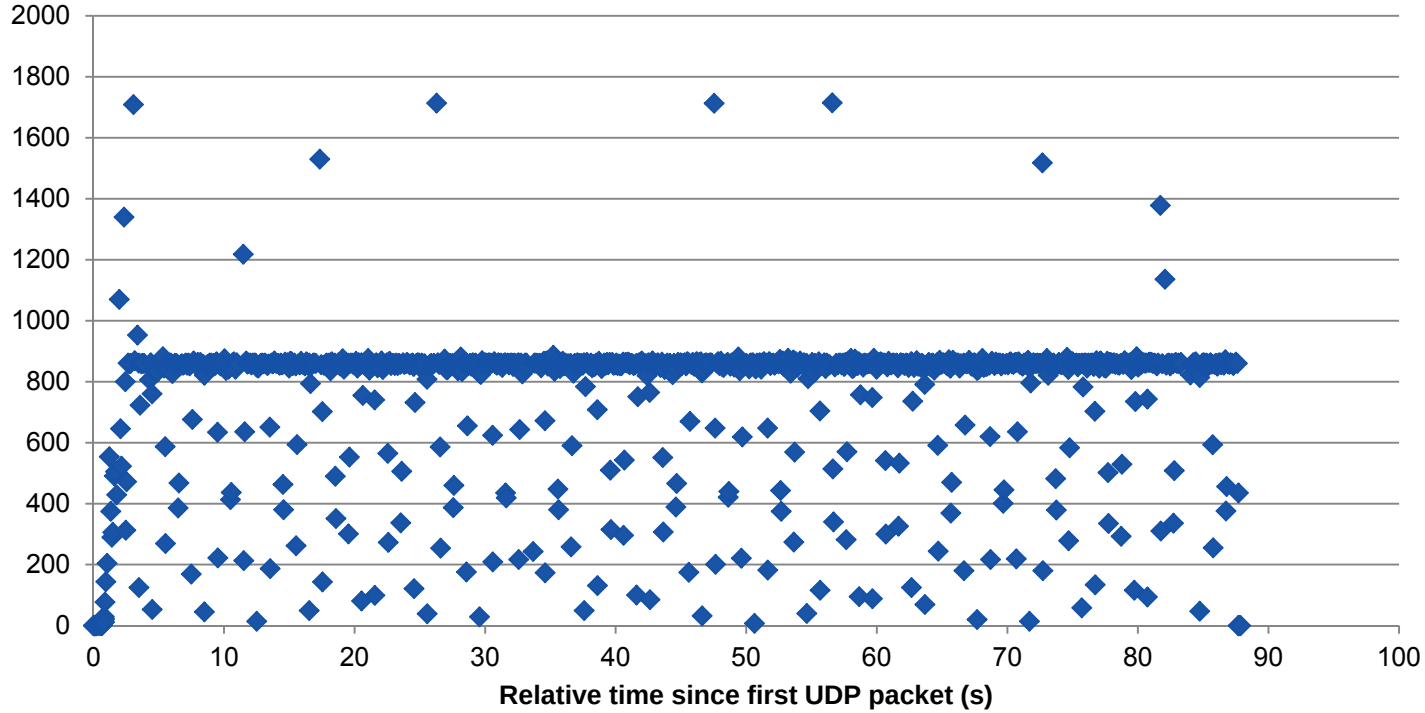
← Packet rate

33001 is the source UDP port of server
42132 is the source UDP port of the client

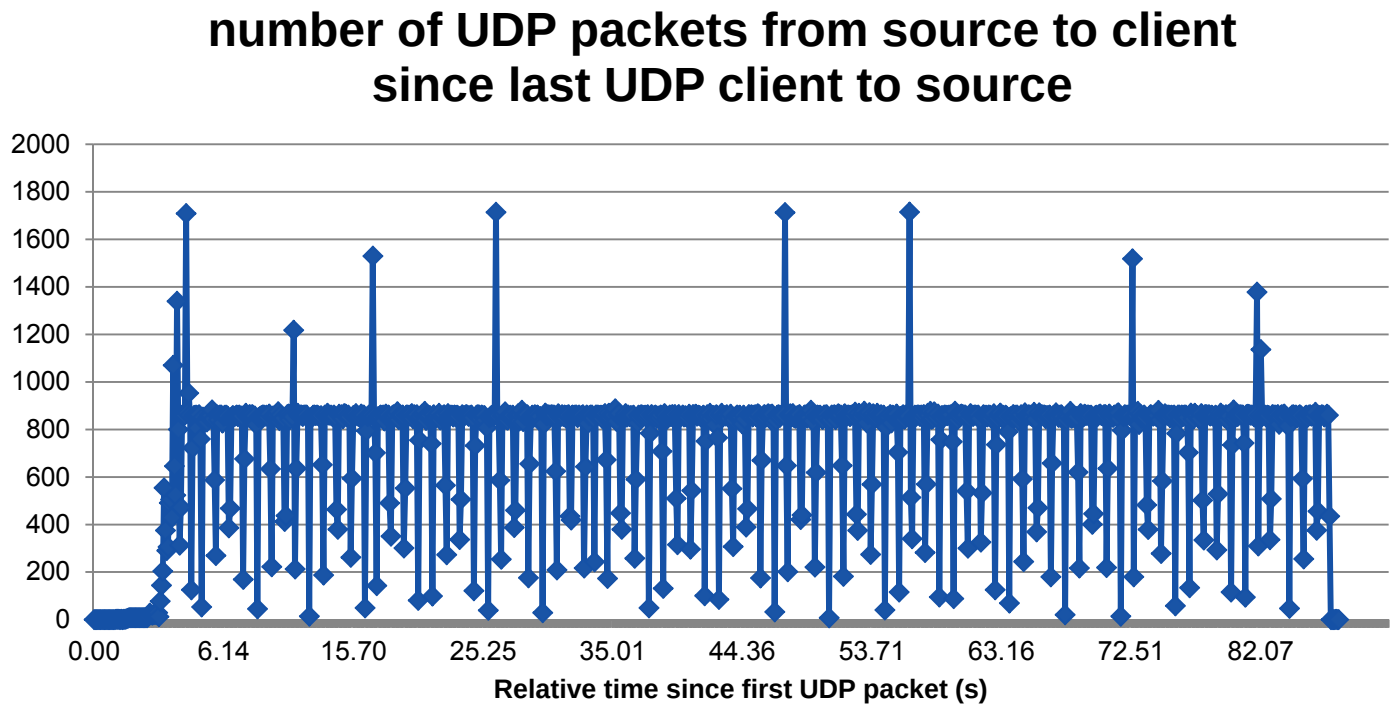


The first ~15 seconds of the transfer

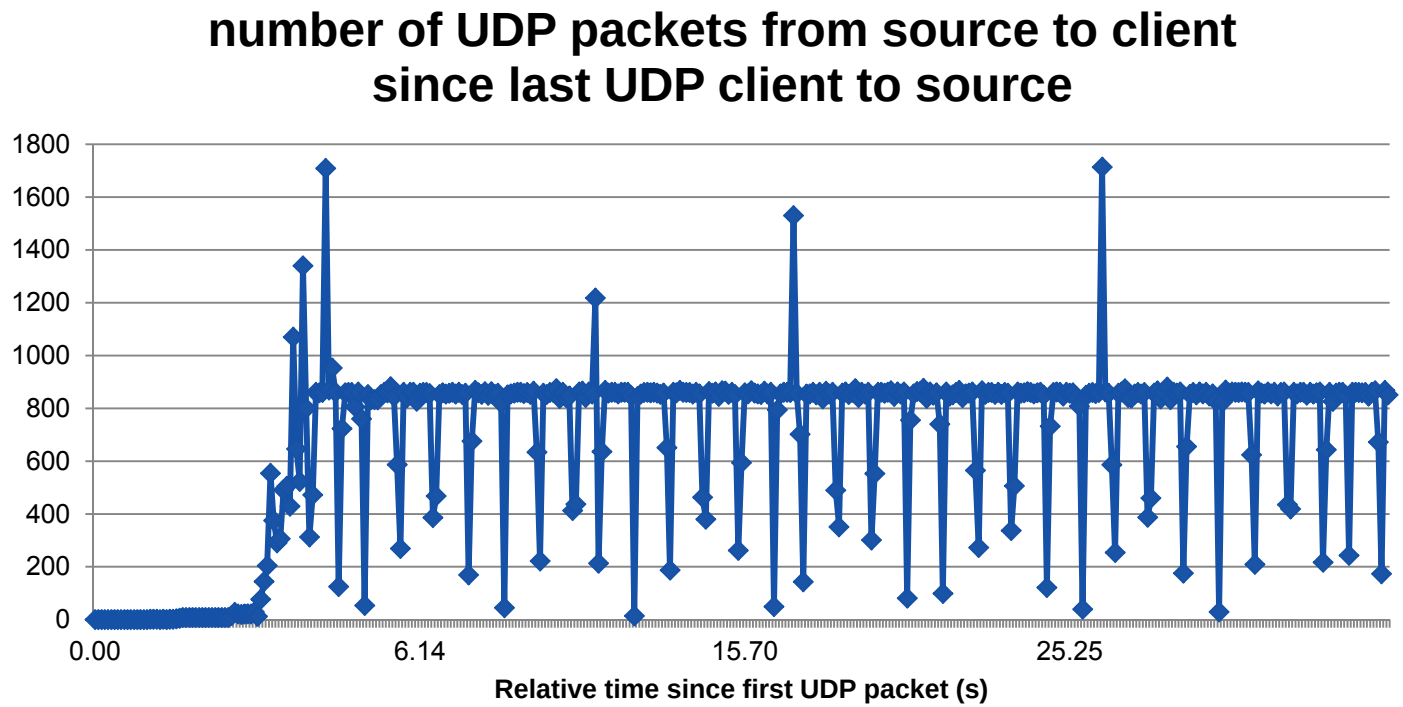
Number of UDP packets from source ⇒ client since last UDP packet client ⇒ source



Connecting the points with lines to see their order

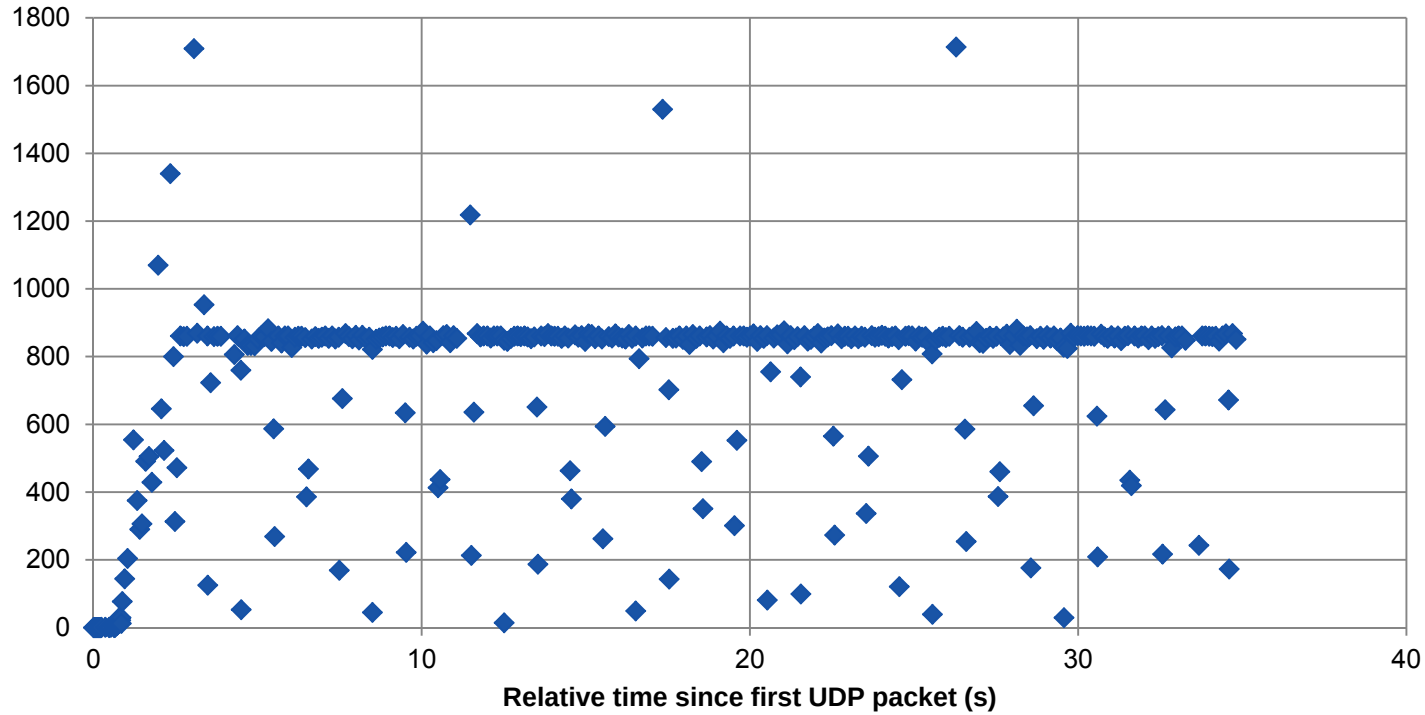


Zooming in more we can see very periodic behavior



Showing only the first 330 bursts

Is the burst length periodicity a result of sever using different periodic processes that are out of phase wrt each others?



Showing only the first 330 bursts

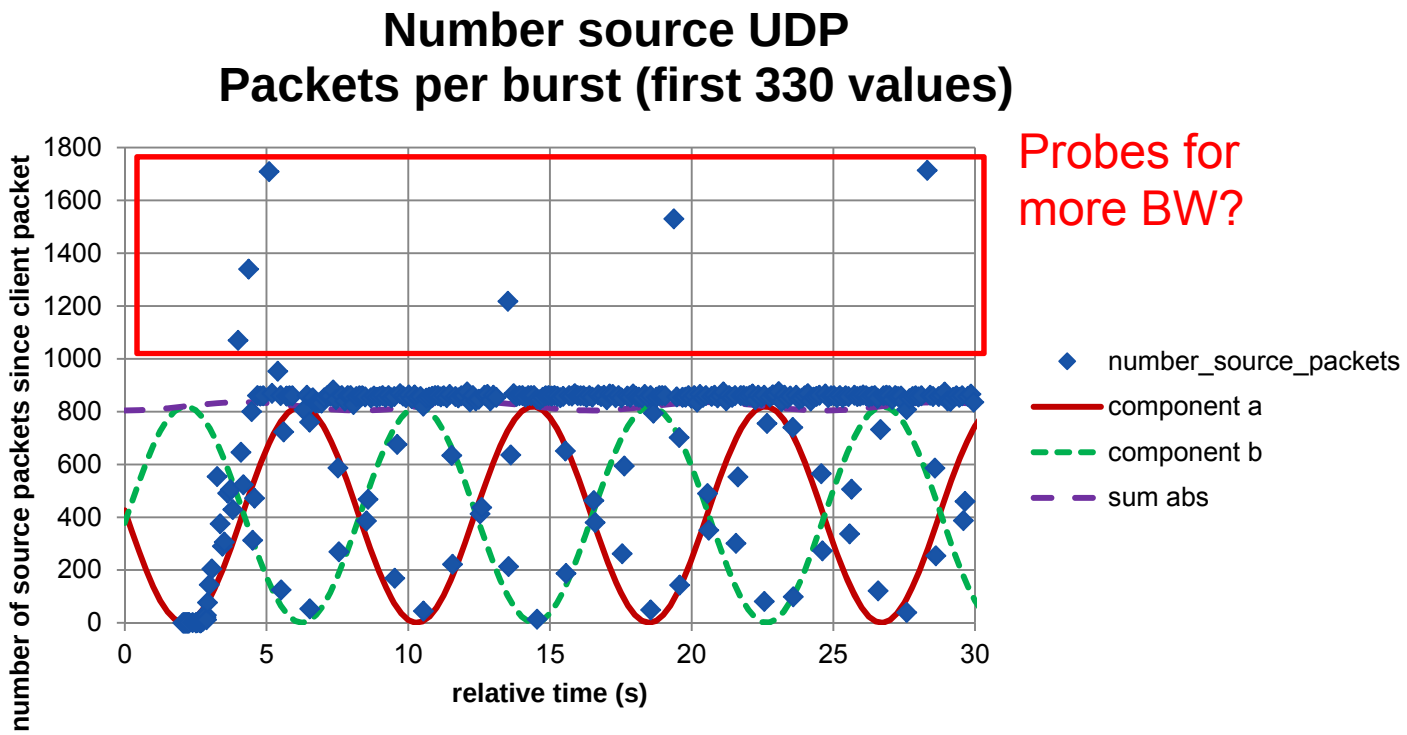
Two sinograms – out of phase with each other – shaping the short burst lengths?

scale	410	410
v_offset	410	410
offset	4.05	8.1
frequency	0.121951	0.121951

1/8.2

median=857

963 bursts in ~90 s

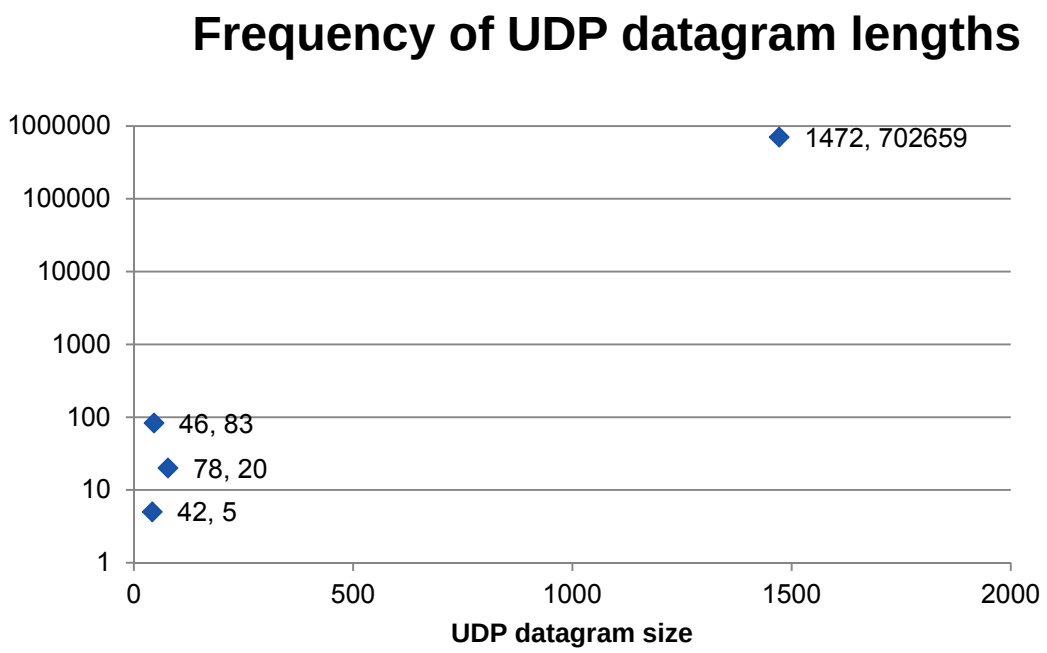


Probes for more BW?

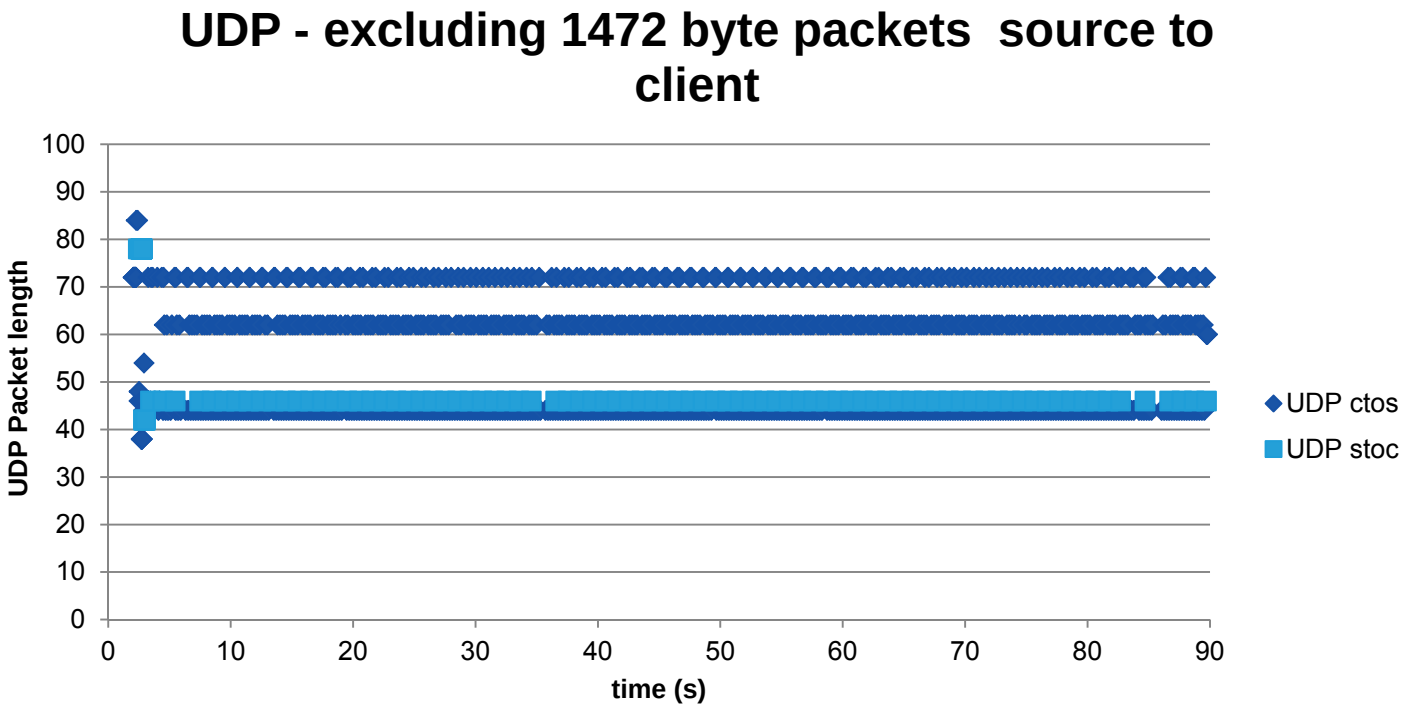
Shorter bursts to detect congestion or decrease in BW?

Source to client UDP 4 different datagram sizes

Bin	Frequency
42	5
46	83
78	20
1472	702659

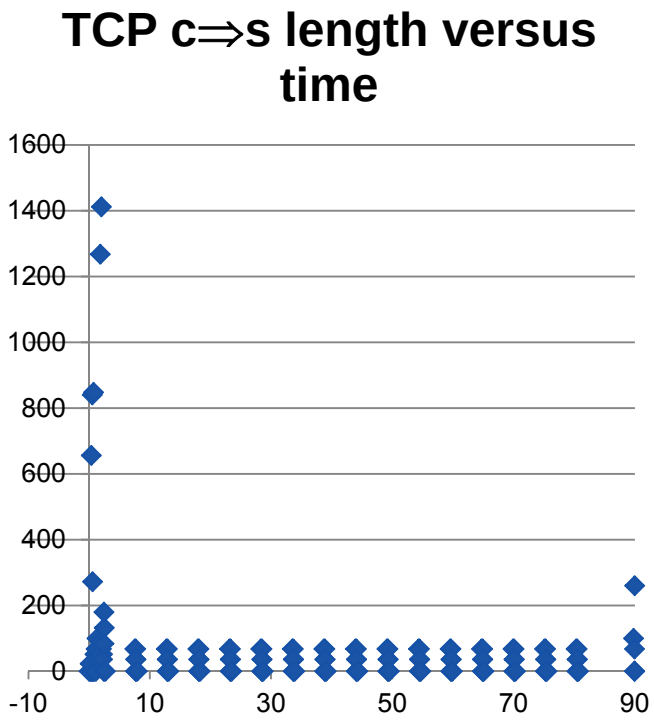


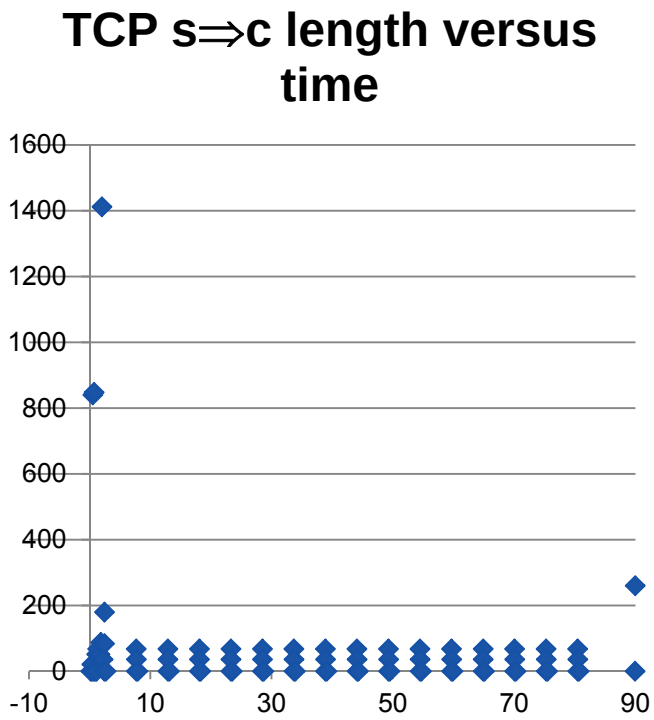
When are these shorter UDP datagrams sent?



Even the source is regularly sending short datagrams!

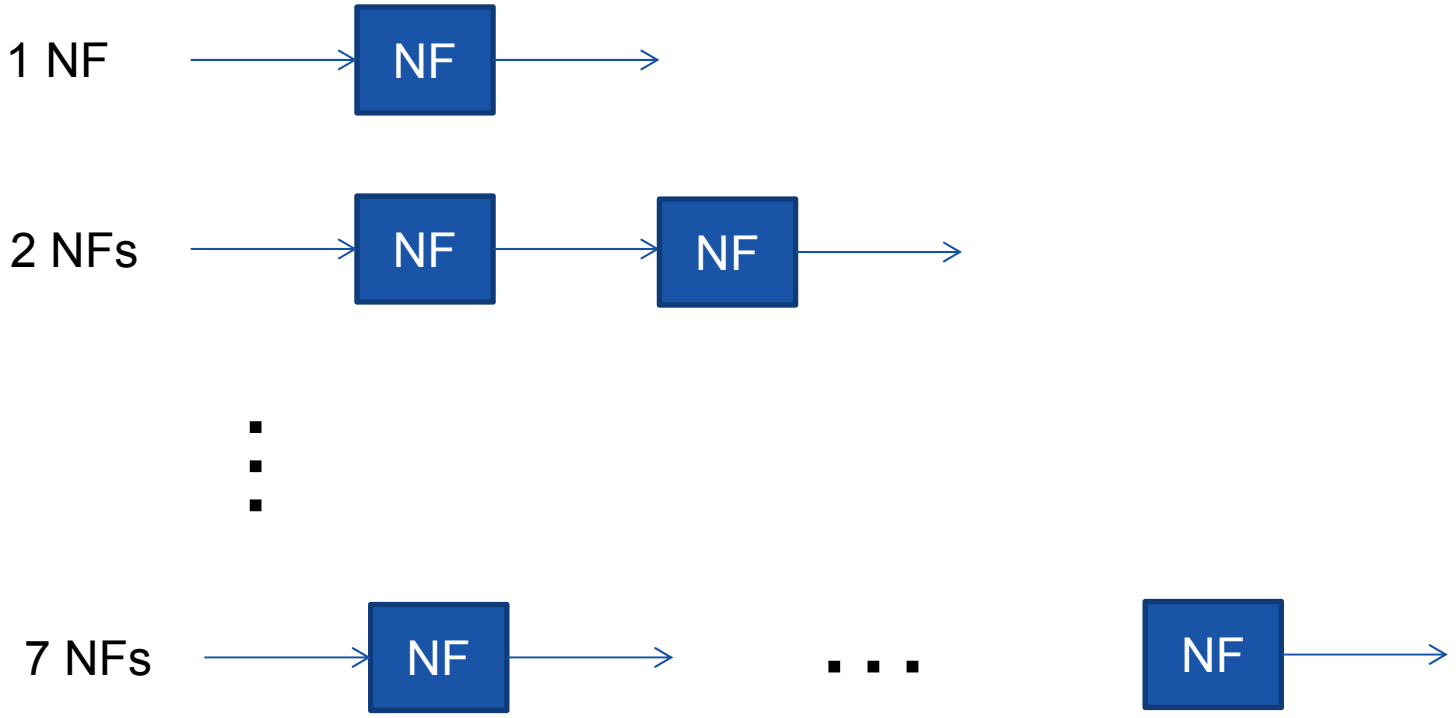
There is also TCP traffic





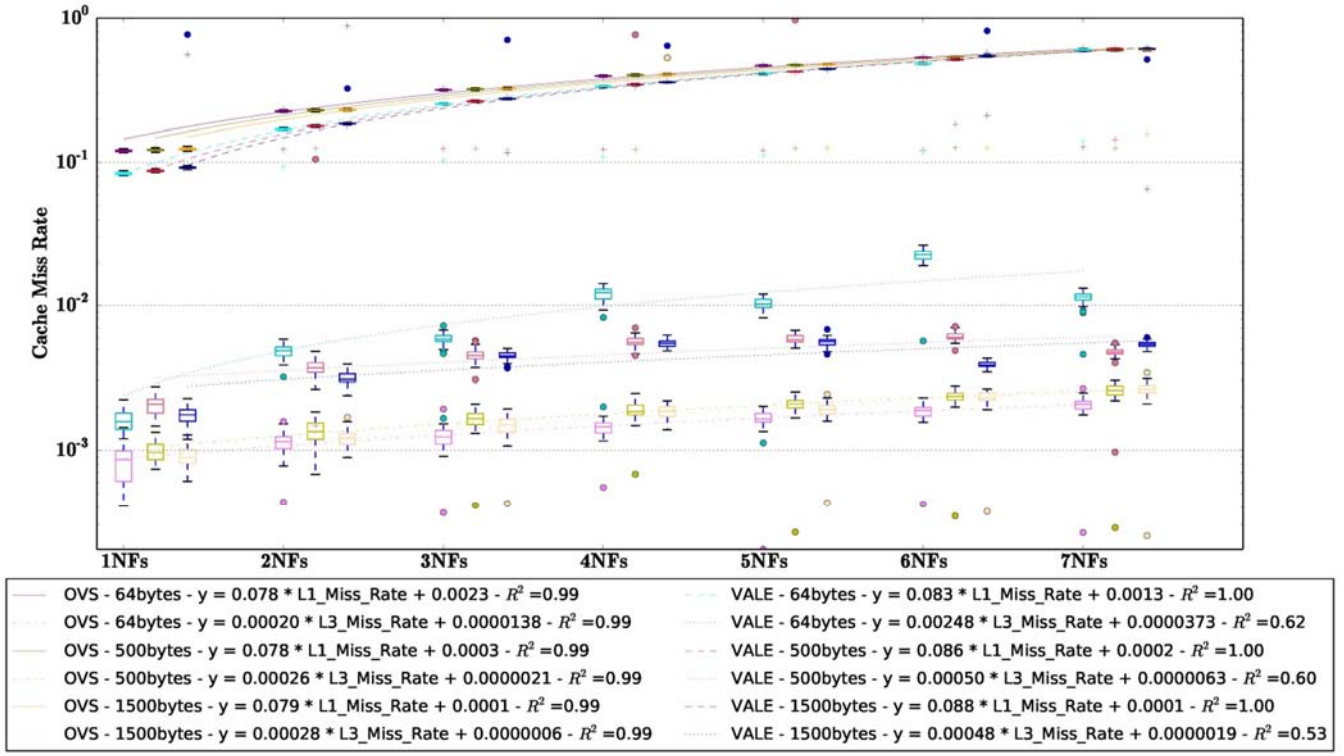
~5.196 s between bursts of TCP packets

Another experiment: Sending packets through chains of network functions



Importance of visualizing outliers

Cache miss rates as a function of the length of a chain of network functions

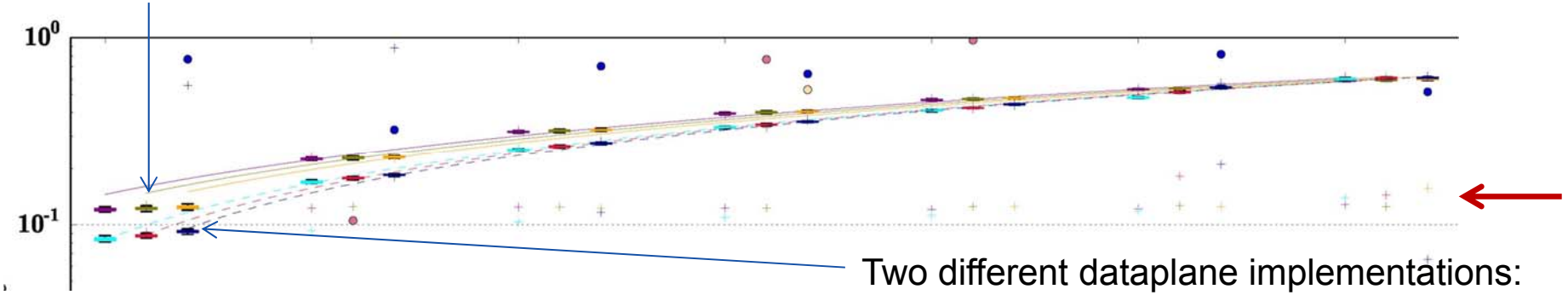


Data provided by Georgios Katsikas, doctoral student at Network Systems Lab (NSL), CoS

Importance of visualizing outliers

Zooming in on L1 cache miss rates as a function of the length of a chain

Three different lengths of packets (64, 500, 1500)



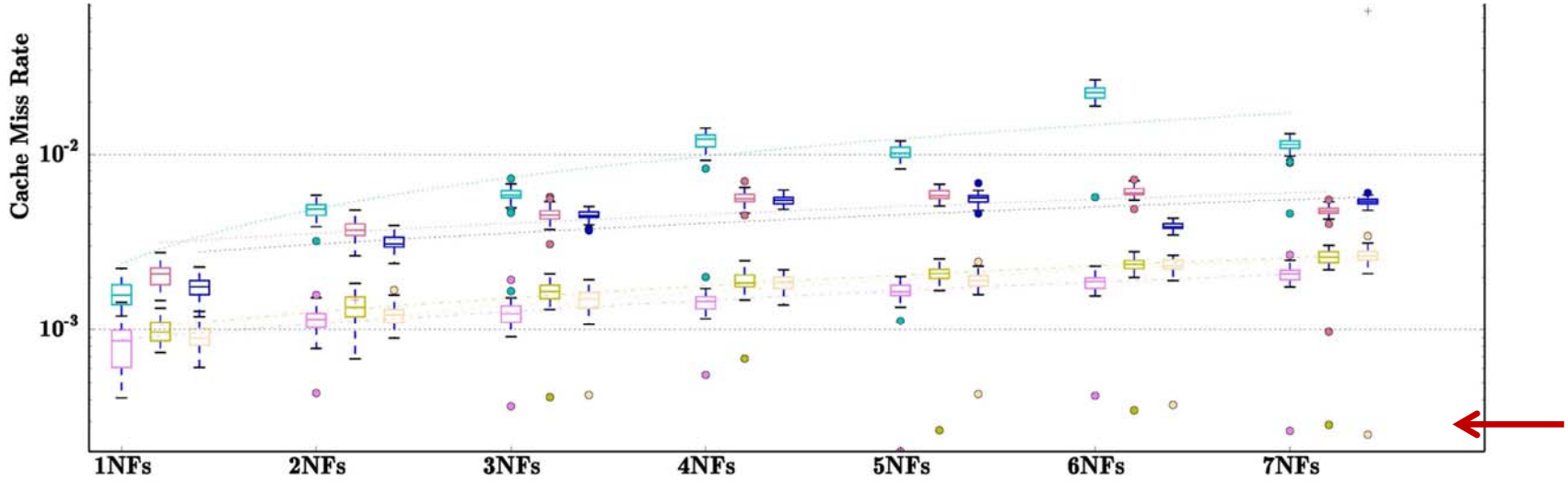
Two different dataplane implementations:
Open vSwitch (OVS) and netmap/VALE

OVS - 64bytes - y = 0.078 * L1_Miss_Rate + 0.0023 - R^2 =0.99	VALE - 64bytes - y = 0.083 * L1_Miss_Rate + 0.0013 - R^2 =1.00
OVS - 64bytes - y = 0.00020 * L3_Miss_Rate + 0.0000138 - R^2 =0.99	VALE - 64bytes - y = 0.00248 * L3_Miss_Rate + 0.0000373 - R^2 =0.62
OVS - 500bytes - y = 0.078 * L1_Miss_Rate + 0.0003 - R^2 =0.99	VALE - 500bytes - y = 0.086 * L1_Miss_Rate + 0.0002 - R^2 =1.00
OVS - 500bytes - y = 0.00026 * L3_Miss_Rate + 0.0000021 - R^2 =0.99	VALE - 500bytes - y = 0.00050 * L3_Miss_Rate + 0.0000063 - R^2 =0.60
OVS - 1500bytes - y = 0.079 * L1_Miss_Rate + 0.0001 - R^2 =0.99	VALE - 1500bytes - y = 0.088 * L1_Miss_Rate + 0.0001 - R^2 =1.00
OVS - 1500bytes - y = 0.00028 * L3_Miss_Rate + 0.0000006 - R^2 =0.99	VALE - 1500bytes - y = 0.00048 * L3_Miss_Rate + 0.0000019 - R^2 =0.53

Data provided by Georgios Katsikas, doctoral student at Network Systems Lab (NSL), CoS

Importance of outliers

Zooming in on the lowest L3 miss rates – shows there are some that experience an order of magnitude lower cache miss rates. Why?



Data provided by Georgios Katsikas, doctoral student at Network Systems Lab (NSL), CoS

Lots of sources for more information

There are lots of different ways of presenting data graphically, see for example:

- Ray Lyons, 'Best Practices in Graphical Data Presentation', [Lyons 2010] http://libraryassessment.org/bm~doc/workshop_lyons_ray.pdf
- Dona M. Wong, *The Wall Street journal guide to information graphics: the dos and don'ts of presenting data, facts, and figures* [Wong 2010] (Edward R. Tufte was her thesis advisor)

References

[Cleveland 1989] William S. Cleveland, *The elements of graphing data*, 10.[print.] ed. Monterey, Cal: Wadsworth, 1989, ISBN: 978-0-534-03729-1.

[Cleveland 1993] William S. Cleveland, *Visualizing data*. Murray Hill, N.J. : [Summit, N.J: At&T Bell Laboratories ; Published by Hobart Press, 1993, ISBN: 978-0-9634884-0-4.

[Lyons 2010] Ray Lyons, 'Best Practices in Graphical Data Presentation', Baltimore, MD, USA, 25-Oct-2010 [Online]. Available: http://libraryassessment.org/bm~doc/workshop_lyons_ray.pdf. [Accessed: 18-Aug-2015]

[Tufte 2008] Edward Rolf Tufte, *Envisioning information*, 12. printing. Cheshire, Conn: Graphics Press, 2008, ISBN: 978-0-9613921-1-6.

[Tufte 1997] Edward R. Tufte, *Visual explanations: images and quantities, evidence and narrative*. Cheshire, Conn: Graphics Press, 1997, ISBN: 978-0-9613921-2-3.

[Tufte 2006] Edward R. Tufte, *Beautiful evidence*. Cheshire, Conn: Graphics Press, 2006, ISBN: 978-0-9613921-7-8.

[Tufte 2001] Edward R. Tufte, *The visual display of quantitative information*, 2nd ed. Cheshire, Conn: Graphics Press, 2001, ISBN: 978-0-9613921-4-7.

[Wong 2010] Dona M. Wong, *The Wall Street journal guide to information graphics: the dos and don'ts of presenting data, facts, and figures*, 1st ed. New York: W.W. Norton & Co, 2010, ISBN: 978-0-393-07295-2.

¿Questions?