Flow Data Sharing

Building a data set for network research

LHCONE - Prague Meeting

April 2023

Yatish Kumar, Eli Dart, Ed Balas, Chin Guok

ESnet

Our Goal

Provide a standard data set to networking researchers based on High Touch and Netflow data **seen by ESnet on LHCONE**

Flow Records are of the following form:

start_time_ns	end_time_ns	dir	ip_src	ip_dst	l4_src_port	l4_dst_port	prot	ipv	packets	bytes
2023-04-15 20:47:41.616506225	2023-04-15 20:47:42.016214841	out	129.59.59.77	128.227.78.71	1094	33588	TCP	4	882	7889280
2023-04-15 20:47:41.987611159	2023-04-15 20:47:42.017260428	in	128.227.78.43	131.225.189.210	39435	9618	TCP	4	2	149
2023-04-15 20:45:32.018488664	2023-04-15 20:47:42.017833250	in	192.188.182.212	134.158.208.104	38410	1095	TCP	4	800	42904
2023-04-15 20:47:41.209844131	2023-04-15 20:47:42.019981550	in	128.227.78.192	128.211.143.87	52750	1101	TCP	4	343	14384
2023-04-15 20:47:42.020991600	2023-04-15 20:47:42.020991600	in	2620:0:210:1::2d	2605:9a00:10:200a:f816:3eff:feaf:ac66	39922	2881	TCP	6	1	72
2023-04-15 20:47:41.943912389	2023-04-15 20:47:42.023970099	out	128.142.248.159	129.59.197.8	9618	42139	TCP	4	2	104
2023-04-15 20:47:35.798252250	2023-04-15 20:47:42.024050231	out	131.225.188.190	129.59.197.42	24054	37682	TCP	4	4779	42980947
2023-04-15 20:47:41.488099159	2023-04-15 20:47:42.024491413	in	128.227.78.16	128.142.248.159	38285	9618	TCP	4	155	231283
2023-04-15 20:47:41.611047163	2023-04-15 20:47:42.024491433	in	128.227.78.71	129.59.59.77	33588	1094	TCP	4	257	11000
2023-04-15 20:47:41.611062238	2023-04-15 20:47:42.024508159	out	128.227.78.71	129.59.59.77	33588	1094	TCP	4	257	11000

With HT they are 1:1 accurate.

So they represent all packets for all flows. Large or small.

Netflow is similar, but a random subset based on netflow sampling

ESnet

What does this allow:

1. Capacity and traffic pattern analysis for every flow as seen by ESnet on LHCONE, between two entities.
2. Frequency, Peak Rate and Duration of flows
3. Network congestion analysis
4. Correlation between CRIC and what we see
5. Anomalies.
    a. Eg. connection attempts with no syn-acknowledgement
    b. Round Trip Times
    c. Asymmetric routing

These topics are all interesting to Networking Researchers, but presently only ESnet engineers can access this data. We would like to share it with the community.

ESnet

What we need

1. Before sharing such information, we need agreement from a governing body that represents this traffic. (WLCG for LHCONE ?)
2. We need to address any concerns regarding Individual IP profiles
3. We need to gather any other concerns from sites or peered networks

- Depending on need for anonymization, information can be obscured
- More anonymization → less information available for research
- What is the right balance?

ESnet

Proposed Minimum Data Set

start_time_ns	end_time_ns	dir	ip_src	ip_dst	l4_src_port	l4_dst_port	prot	ipv	packets	bytes
2023-04-15 20:47:41.616506225	2023-04-15 20:47:42.016214841	out	129.59.59.77	128.227.78.71	1094	33588	TCP	4	882	7889280
2023-04-15 20:47:41.987611159	2023-04-15 20:47:42.017260428	in	128.227.78.43	131.225.189.210	39435	9618	TCP	4	2	149
2023-04-15 20:45:32.018488664	2023-04-15 20:47:42.017833250	in	192.188.182.212	134.158.208.104	38410	1095	TCP	4	800	42904
2023-04-15 20:47:41.209844131	2023-04-15 20:47:42.019981550	in	128.227.78.192	128.211.143.87	52750	1101	TCP	4	343	14384
2023-04-15 20:47:42.020991600	2023-04-15 20:47:42.020991600	in	2620:0:210:1::2d	2605:9a00:10:200a:f816:3eff:feaf:ac66	39922	2881	TCP	6	1	72
2023-04-15 20:47:41.943912389	2023-04-15 20:47:42.023970099	out	128.142.248.159	129.59.197.8	9618	42139	TCP	4	2	104
2023-04-15 20:47:35.798252250	2023-04-15 20:47:42.024050231	out	131.225.188.190	129.59.197.42	24054	37682	TCP	4	4779	42980947
2023-04-15 20:47:41.488099159	2023-04-15 20:47:42.024491413	in	128.227.78.16	128.142.248.159	38285	9618	TCP	4	155	231283
2023-04-15 20:47:41.611047163	2023-04-15 20:47:42.024491433	in	128.227.78.71	129.59.59.77	33588	1094	TCP	4	257	11000
2023-04-15 20:47:41.611062238	2023-04-15 20:47:42.024508159	out	128.227.78.71	129.59.59.77	33588	1094	TCP	4	257	11000

start_time_ns	end_time_ns	dir	ip_src	ip_dst	l4_src_port	l4_dst_port	prot	ipv	packets	bytes
2023-04-15 01:00:55.000190312	2023-04-15 01:00:55.000190312	in	17735692664596054083	9052400364753418059	50478	873	TCP	4	1	44
2023-04-15 01:00:55.001898626	2023-04-15 01:00:55.001898626	in	14620496478008448314	3007831929922001348	55504	8015	TCP	4	1	44
2023-04-15 01:00:55.002905119	2023-04-15 01:00:55.002905119	in	15038903950705291078	13685088522320632395	54113	8080	TCP	4	1	44
2023-04-15 01:00:55.003023507	2023-04-15 01:00:55.003023507	in	9240684160268551578	18441205285899465754	50789	8081	TCP	4	1	44
2023-04-15 01:00:55.003429979	2023-04-15 01:00:55.003429979	in	9382868531200459780	7554311246434819712	53027	1521	TCP	4	1	44
2023-04-15 01:00:55.004042297	2023-04-15 01:00:55.004042297	in	9078314055143389604	11369579337328394915	50998	82	TCP	4	1	44
2023-04-15 01:00:55.004159586	2023-04-15 01:00:55.004159586	in	3026927348603649639	6281812981236989392	50998	82	TCP	4	1	44
2023-04-15 01:00:55.005132147	2023-04-15 01:00:55.005132147	in	3586909256860327984	8799974332569612893	55504	8015	TCP	4	1	44
2023-04-15 01:00:55.005196976	2023-04-15 01:00:55.005196976	in	5998037206129486313	5343166444818954477	51051	8531	TCP	4	1	44
2023-04-15 01:00:55.005648778	2023-04-15 01:00:55.005648778	in	16547538881454434279	5829288755111577961	50182	8159	TCP	4	1	44

Replace all IP addresses with a 64 bit random number.

- Preserves per flow resolution, but hides the identity of a source

Or use CryptoPan (https://en.wikipedia.org/wiki/Crypto-PAn). This is prefix preserving.

This limits traffic analysis for :

Org to Org

Study of paths or latencies between orgs

ESnet

Slightly Better Data Set

start_time_ns	end_time_ns	dir	ip_src	ip_dst	asn_src	asn_dst	l4_src_port	l4_dst_port	prot	ipv	packets	bytes
2023-04-15 16:15:01.333194920	2023-04-15 16:15:02.315679700	out	17073150574832218571	14469603570784687024	ESNET-EAST	ESNET-EAST	179	23975	TCP	4	2	123
2023-04-15 16:15:16.342846663	2023-04-15 16:15:16.342846663	in	14469603570784687024	17073150574832218571	ESNET-EAST	ESNET-EAST	23975	179	TCP	4	1	71
2023-04-15 16:15:31.362744040	2023-04-15 16:15:31.362744040	in	14469603570784687024	17073150574832218571	ESNET-EAST	ESNET-EAST	23975	179	TCP	4	1	71
2023-04-15 16:15:32.915626726	2023-04-15 16:15:32.915626726	out	17073150574832218571	14469603570784687024	ESNET-EAST	ESNET-EAST	179	23975	TCP	4	1	71
2023-04-15 10:41:31.989233218	2023-04-15 10:41:34.415497925	in	17970276053418248896	3277551430657147755	NERDCNET	ASGARR	58972	1094	TCP	4	175	7676
2023-04-15 10:39:14.571348554	2023-04-15 10:41:34.419258527	out	13540611744320629726	2032478640527493105	ASGARR	NERDCNET	1094	49592	TCP	4	2767	24876652
2023-04-15 10:41:33.989511643	2023-04-15 10:41:34.424056518	in	10596795559813743378	154049852835516020	VANDERBILT	CERN	1094	55546	TCP	6	14	7312
2023-04-15 10:41:34.040437335	2023-04-15 10:41:34.424296412	in	4843683896311383801	12840386780741658199	NERDCNET	FNAL-AS	1094	44582	TCP	6	250	370730
2023-04-15 10:41:33.147360570	2023-04-15 10:41:34.425015024	in	10173424899132320331	3282474128008594986	ORNL-MSRNET	IN2P3	56858	1095	TCP	4	10	712
2023-04-15 10:41:14.432745355	2023-04-15 10:41:34.426219226	out	15690536830131726380	17290534916254316874	ASGARR	NERDCNET	1094	34992	TCP	4	1671	15031646

Include the ASN names

Or .. we can include site names from CRIC

We can also filter out ANY traffic not registered in CRIC

ESnet

Further Enhancements

end_time_ns	dir	ip_src	ip_dst	tcp_f_syn_ack_only	tcp_f_syn_only	tcp_f_cwr	tcp_f_ece	tcp_f_urg	tcp_f_ack	tcp_f_psh	tcp_f_rst	tcp_f_syn	tcp_f_fin	packets	bytes	pkt_size_hist
2023-04-15 20:56:31.225751077	in	19464260758327567O4	13439401633455902264	false	false	false	false	false	true	true	false	false	false	2	149	[1,1,0,0,0,0,0,0]
2023-04-15 20:56:31.225922713	out	4584612987610159525	774417625577374267	false	false	false	false	false	true	true	false	false	false	2	227	[0,1,1,0,0,0,0,0]
2023-04-15 20:56:31.226058684	out	166111255126598993O0	18102868303733436120	false	false	false	false	false	true	true	false	false	false	8621	77229122	[0,0,17,8,0,15,3,8578]
2023-04-15 20:56:31.227981961	out	14784553695144767B5	9864286103658761992	false	false	false	false	false	true	false	false	false	false	4	160	[4,0,0,0,0,0,0,0]
2023-04-15 20:56:31.229563771	out	59621691431416949B7	11500399955916603686	false	false	false	false	false	true	true	false	false	false	1	121	[0,1,0,0,0,0,0,0]
2023-04-15 20:56:31.229946580	out	35843051246248411O0	2539667937134499323	false	false	false	false	false	true	true	false	false	false	118	79257	[0,0,67,6,4,41,0,0]
2023-04-15 20:56:31.230919389	in	58202906321811347S3	17244459572524761147	false	true	false	false	false	true	true	false	true	true	15	5880	[7,1,1,3,0,3,0,0]
2023-04-15 20:56:31.231520047	in	774417625577374267	4584612987610159525	false	false	false	false	false	true	true	false	false	false	2	270	[1,0,1,0,0,0,0,0]
2023-04-15 20:56:31.231533454	out	12189054256328962750	3588980711754583130	false	false	false	false	false	true	true	false	false	false	1	83	[0,1,0,0,0,0,0,0]
2023-04-15 20:56:31.232430958	in	6199317915428408193	10858316836202061411	false	false	false	false	false	true	false	false	false	false	1	72	[0,1,0,0,0,0,0,0]

We can include roll ups of all tcp flags set during a flow interval.
Histograms of packet sizes per flow interval.

Useful for large scale studies of TCP behaviour

Experiment meta-data, from Fireflies can also be incorporated into the annotation

ESnet

Reproducible Research

1. Release a 1 Terabyte data set that can be used for comparing work between different research papers, even if it is done decades apart.
2. Release a monthly data set, that has a limited lifetime (storage and use bounded), which can be used for current trend analysis.
3. Encourage researchers to build long term summaries from 2.