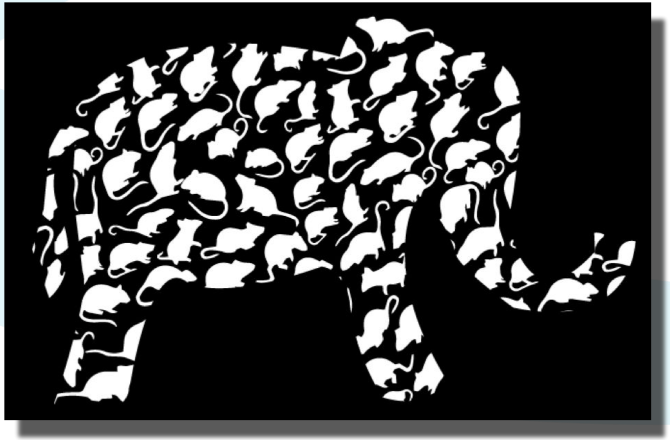
ESnet
ENERGY SCIENCES NETWORK

Should we engineer for Elephant flows ?

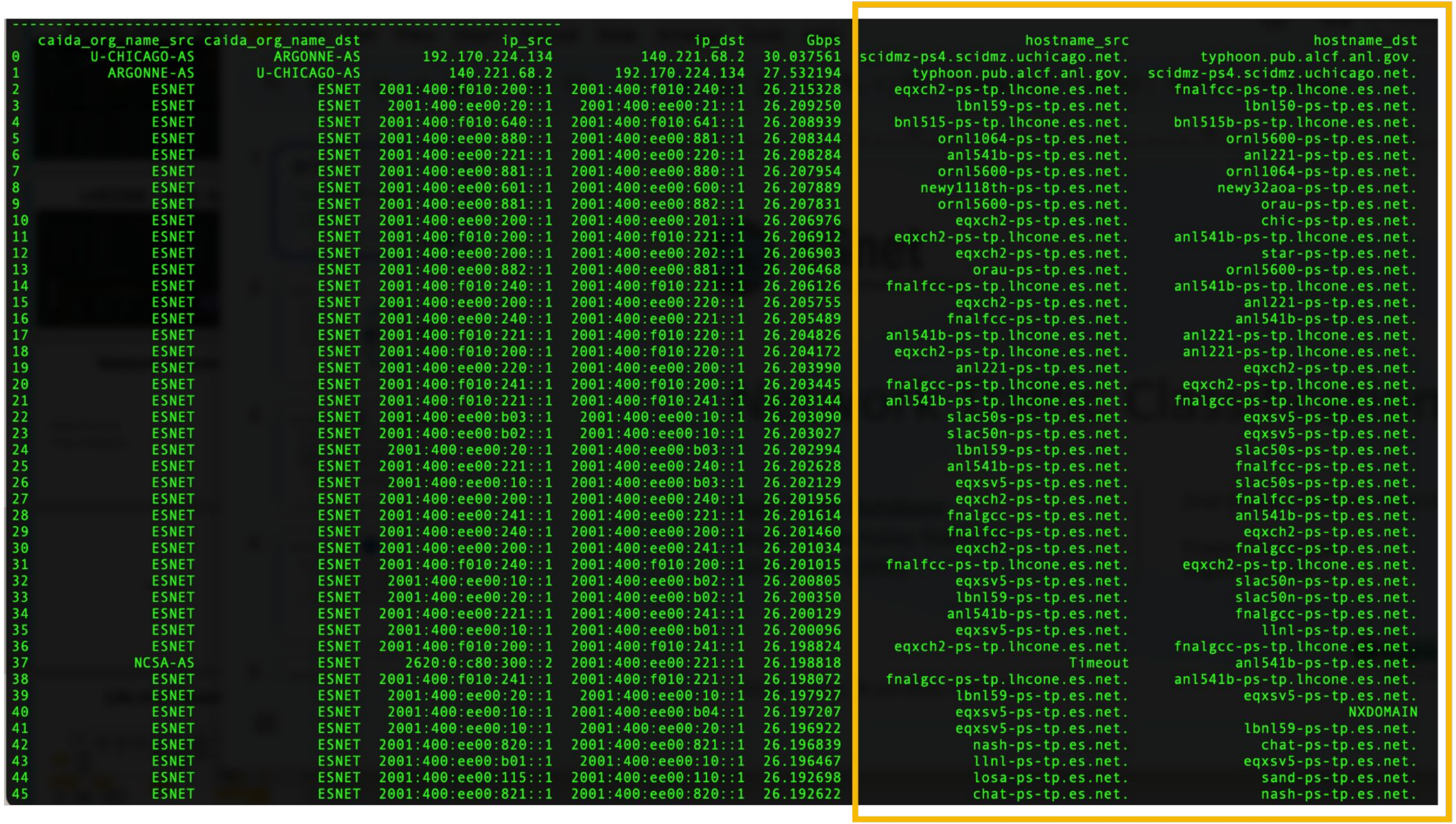
Yatish Kumar
ESnet



U.S. DEPARTMENT OF ENERGY
Office of Science

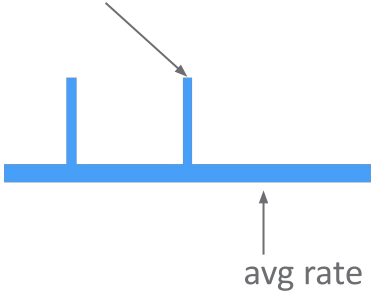
BERKELEY LAB

We searched our High Touch Database .....



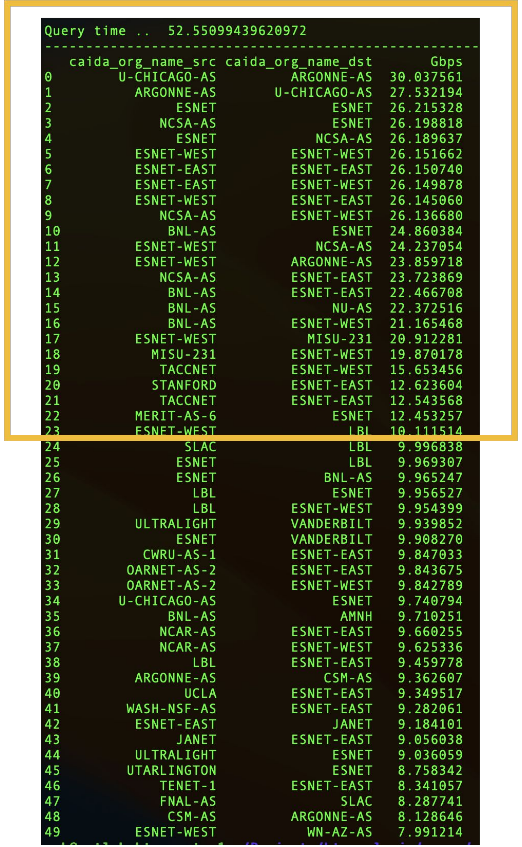
```
--------------------------------------------------------------------------------
   caida_org_name_src caida_org_name_dst               ip_src               ip_dst      Gbps                     hostname_src                     hostname_dst
0        U-CHICAGO-AS         ARGONNE-AS      192.170.224.134         140.221.68.2 30.037561  scidmz-ps4.scidmz.uchicago.net.        typhoon.pub.alcf.anl.gov.
1          ARGONNE-AS       U-CHICAGO-AS         140.221.68.2      192.170.224.134 27.532194        typhoon.pub.alcf.anl.gov.  scidmz-ps4.scidmz.uchicago.net.
2               ESNET              ESNET 2001:400:f010:200::1 2001:400:f010:240::1 26.215328      eqxch2-ps-tp.lhcone.es.net.     fnalfcc-ps-tp.lhcone.es.net.
3               ESNET              ESNET  2001:400:ee00:20::1  2001:400:ee00:21::1 26.209250             lbnl59-ps-tp.es.net.             lbnl50-ps-tp.es.net.
4               ESNET              ESNET 2001:400:f010:640::1 2001:400:f010:641::1 26.208939      bnl515-ps-tp.lhcone.es.net.     bnl515b-ps-tp.lhcone.es.net.
5               ESNET              ESNET 2001:400:ee00:880::1 2001:400:ee00:881::1 26.208344           ornl1064-ps-tp.es.net.           ornl5600-ps-tp.es.net.
6               ESNET              ESNET 2001:400:ee00:221::1 2001:400:ee00:220::1 26.208284            anl541b-ps-tp.es.net.             anl221-ps-tp.es.net.
7               ESNET              ESNET 2001:400:ee00:881::1 2001:400:ee00:880::1 26.207954           ornl5600-ps-tp.es.net.           ornl1064-ps-tp.es.net.
8               ESNET              ESNET 2001:400:ee00:601::1 2001:400:ee00:600::1 26.207889         newy1118th-ps-tp.es.net.          newy32aoa-ps-tp.es.net.
9               ESNET              ESNET 2001:400:ee00:881::1 2001:400:ee00:882::1 26.207831           ornl5600-ps-tp.es.net.               orau-ps-tp.es.net.
10              ESNET              ESNET 2001:400:ee00:200::1 2001:400:ee00:201::1 26.206976             eqxch2-ps-tp.es.net.               chic-ps-tp.es.net.
11              ESNET              ESNET 2001:400:f010:200::1 2001:400:f010:221::1 26.206912      eqxch2-ps-tp.lhcone.es.net.     anl541b-ps-tp.lhcone.es.net.
12              ESNET              ESNET 2001:400:ee00:200::1 2001:400:ee00:202::1 26.206903             eqxch2-ps-tp.es.net.               star-ps-tp.es.net.
13              ESNET              ESNET 2001:400:ee00:882::1 2001:400:ee00:881::1 26.206468               orau-ps-tp.es.net.           ornl5600-ps-tp.es.net.
14              ESNET              ESNET 2001:400:f010:240::1 2001:400:f010:221::1 26.206126     fnalfcc-ps-tp.lhcone.es.net.     anl541b-ps-tp.lhcone.es.net.
15              ESNET              ESNET 2001:400:ee00:200::1 2001:400:ee00:220::1 26.205755             eqxch2-ps-tp.es.net.             anl221-ps-tp.es.net.
16              ESNET              ESNET 2001:400:ee00:240::1 2001:400:ee00:221::1 26.205489            fnalfcc-ps-tp.es.net.            anl541b-ps-tp.es.net.
17              ESNET              ESNET 2001:400:f010:221::1 2001:400:f010:220::1 26.204826     anl541b-ps-tp.lhcone.es.net.      anl221-ps-tp.lhcone.es.net.
18              ESNET              ESNET 2001:400:f010:200::1 2001:400:f010:220::1 26.204172      eqxch2-ps-tp.lhcone.es.net.      anl221-ps-tp.lhcone.es.net.
19              ESNET              ESNET 2001:400:ee00:220::1 2001:400:ee00:200::1 26.203990             anl221-ps-tp.es.net.             eqxch2-ps-tp.es.net.
20              ESNET              ESNET 2001:400:f010:241::1 2001:400:f010:200::1 26.203445     fnalgcc-ps-tp.lhcone.es.net.      eqxch2-ps-tp.lhcone.es.net.
21              ESNET              ESNET 2001:400:f010:221::1 2001:400:f010:241::1 26.203144     anl541b-ps-tp.lhcone.es.net.     fnalgcc-ps-tp.lhcone.es.net.
22              ESNET              ESNET 2001:400:ee00:b03::1  2001:400:ee00:10::1 26.203090            slac50s-ps-tp.es.net.             eqxsv5-ps-tp.es.net.
23              ESNET              ESNET 2001:400:ee00:b02::1  2001:400:ee00:10::1 26.203027            slac50n-ps-tp.es.net.             eqxsv5-ps-tp.es.net.
24              ESNET              ESNET  2001:400:ee00:20::1 2001:400:ee00:b03::1 26.202994             lbnl59-ps-tp.es.net.            slac50s-ps-tp.es.net.
25              ESNET              ESNET 2001:400:ee00:221::1 2001:400:ee00:240::1 26.202628            anl541b-ps-tp.es.net.            fnalfcc-ps-tp.es.net.
26              ESNET              ESNET  2001:400:ee00:10::1 2001:400:ee00:b03::1 26.202129             eqxsv5-ps-tp.es.net.            slac50s-ps-tp.es.net.
27              ESNET              ESNET 2001:400:ee00:200::1 2001:400:ee00:240::1 26.201956             eqxch2-ps-tp.es.net.            fnalfcc-ps-tp.es.net.
28              ESNET              ESNET 2001:400:ee00:241::1 2001:400:ee00:221::1 26.201614            fnalgcc-ps-tp.es.net.            anl541b-ps-tp.es.net.
29              ESNET              ESNET 2001:400:ee00:240::1 2001:400:ee00:200::1 26.201460            fnalfcc-ps-tp.es.net.             eqxch2-ps-tp.es.net.
30              ESNET              ESNET 2001:400:ee00:200::1 2001:400:ee00:241::1 26.201034             eqxch2-ps-tp.es.net.            fnalgcc-ps-tp.es.net.
31              ESNET              ESNET 2001:400:f010:240::1 2001:400:f010:200::1 26.201015     fnalfcc-ps-tp.lhcone.es.net.      eqxch2-ps-tp.lhcone.es.net.
32              ESNET              ESNET  2001:400:ee00:10::1 2001:400:ee00:b02::1 26.200805             eqxsv5-ps-tp.es.net.            slac50n-ps-tp.es.net.
33              ESNET              ESNET  2001:400:ee00:20::1 2001:400:ee00:b02::1 26.200350             lbnl59-ps-tp.es.net.            slac50n-ps-tp.es.net.
34              ESNET              ESNET 2001:400:ee00:221::1 2001:400:ee00:241::1 26.200129            anl541b-ps-tp.es.net.            fnalgcc-ps-tp.es.net.
35              ESNET              ESNET  2001:400:ee00:10::1 2001:400:ee00:b01::1 26.200096             eqxsv5-ps-tp.es.net.               llnl-ps-tp.es.net.
36              ESNET              ESNET 2001:400:f010:200::1 2001:400:f010:241::1 26.198824      eqxch2-ps-tp.lhcone.es.net.     fnalgcc-ps-tp.lhcone.es.net.
37            NCSA-AS              ESNET    2620:0:c80:300::2 2001:400:ee00:221::1 26.198818                          Timeout            anl541b-ps-tp.es.net.
38              ESNET              ESNET 2001:400:f010:241::1 2001:400:f010:221::1 26.198072     fnalgcc-ps-tp.lhcone.es.net.     anl541b-ps-tp.lhcone.es.net.
39              ESNET              ESNET  2001:400:ee00:20::1  2001:400:ee00:10::1 26.197927             lbnl59-ps-tp.es.net.             eqxsv5-ps-tp.es.net.
40              ESNET              ESNET  2001:400:ee00:10::1 2001:400:ee00:b04::1 26.197207             eqxsv5-ps-tp.es.net.                         NXDOMAIN
41              ESNET              ESNET  2001:400:ee00:10::1  2001:400:ee00:20::1 26.196922             eqxsv5-ps-tp.es.net.             lbnl59-ps-tp.es.net.
42              ESNET              ESNET 2001:400:ee00:820::1 2001:400:ee00:821::1 26.196839               nash-ps-tp.es.net.               chat-ps-tp.es.net.
43              ESNET              ESNET 2001:400:ee00:b01::1  2001:400:ee00:10::1 26.196467               llnl-ps-tp.es.net.             eqxsv5-ps-tp.es.net.
44              ESNET              ESNET 2001:400:ee00:115::1 2001:400:ee00:110::1 26.192698               losa-ps-tp.es.net.               sand-ps-tp.es.net.
45              ESNET              ESNET 2001:400:ee00:821::1 2001:400:ee00:820::1 26.192622               chat-ps-tp.es.net.               nash-ps-tp.es.net.
```

```
select (*)
where
  Peak Rate > 10Gbps
for at least 10
seconds
Order by Rate
```



What is .ps-tp ?
Because it generates our largest elephant flows.

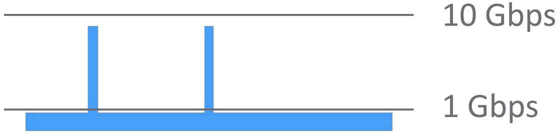
Zooming out a bit

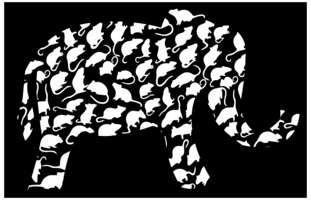


```
Query time ..  52.55099439620972
------------------------------------------------------
   caida_org_name_src caida_org_name_dst       Gbps
0        U-CHICAGO-AS         ARGONNE-AS  30.037561
1          ARGONNE-AS       U-CHICAGO-AS  27.532194
2               ESNET              ESNET  26.215328
3             NCSA-AS              ESNET  26.198818
4               ESNET            NCSA-AS  26.189637
5          ESNET-WEST         ESNET-WEST  26.151662
6          ESNET-EAST         ESNET-EAST  26.150740
7          ESNET-EAST         ESNET-WEST  26.149878
8          ESNET-WEST         ESNET-EAST  26.145060
9             NCSA-AS         ESNET-WEST  26.136680
10             BNL-AS              ESNET  24.860384
11         ESNET-WEST            NCSA-AS  24.237054
12         ESNET-WEST         ARGONNE-AS  23.859718
13            NCSA-AS         ESNET-EAST  23.723869
14             BNL-AS         ESNET-EAST  22.466708
15             BNL-AS              NU-AS  22.372516
16             BNL-AS         ESNET-WEST  21.165468
17         ESNET-WEST           MISU-231  20.912281
18           MISU-231         ESNET-WEST  19.870178
19            TACCNET         ESNET-WEST  15.653456
20           STANFORD         ESNET-EAST  12.623604
21            TACCNET         ESNET-EAST  12.543568
22         MERIT-AS-6              ESNET  12.453257
23         ESNET-WEST                LBL  10.111514
24               SLAC                LBL   9.996838
25              ESNET                LBL   9.969307
26              ESNET             BNL-AS   9.965247
27                LBL              ESNET   9.956527
28                LBL         ESNET-WEST   9.954399
29         ULTRALIGHT         VANDERBILT   9.939852
30              ESNET         VANDERBILT   9.908270
31          CWRU-AS-1         ESNET-EAST   9.847033
32        OARNET-AS-2         ESNET-EAST   9.843675
33        OARNET-AS-2         ESNET-WEST   9.842789
34       U-CHICAGO-AS              ESNET   9.740794
35             BNL-AS               AMNH   9.710251
36            NCAR-AS         ESNET-EAST   9.660255
37            NCAR-AS         ESNET-WEST   9.625336
38                LBL         ESNET-EAST   9.459778
39         ARGONNE-AS             CSM-AS   9.362607
40               UCLA         ESNET-EAST   9.349517
41        WASH-NSF-AS         ESNET-EAST   9.282061
42         ESNET-EAST              JANET   9.184101
43              JANET         ESNET-EAST   9.056038
44         ULTRALIGHT              ESNET   9.036059
45        UTARLINGTON              ESNET   8.758342
46            TENET-1         ESNET-EAST   8.341057
47            FNAL-AS               SLAC   8.287741
48             CSM-AS         ARGONNE-AS   8.128646
49         ESNET-WEST           WN-AZ-AS   7.991214
```

These are largely, if not entirely PerfSonar

- We have found the enemy, and it is us !

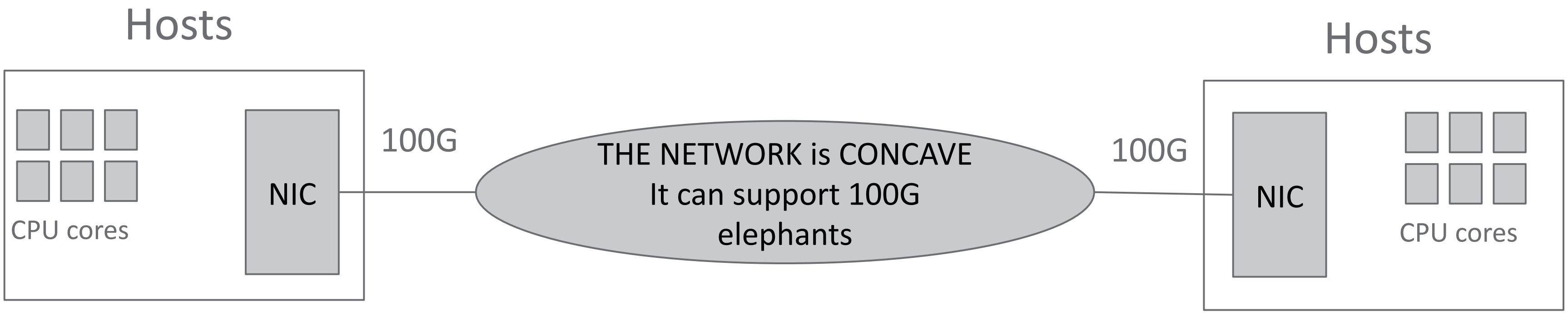




- Everything else has a “peak” of 10Gbps and 99% have an average < 1 Gbps

Globus , FTS etc.. move data at 100Gbps, but as multiple parallel transfers at this scale.

WHY ?



A single core, peaks out at 2Mpps. IF you use zero copy / DPDK.

Much less if you open a socket like most people do.

The receiver has the same limitations.

No one on the host side wants to make a superhuman effort to use 'JUST ONE CORE'. Far easier to just use 3 or 4 cores and get on with life.

Should we engineer our network for elephant flows ?

1. Networks can already correctly forward elephants on any 100G/400G link. So the question is moot !

2. But when we have 100G campus connections into a 400G WAN. How important is it to worry about 4x100G vs. 1x400G ?
   - The MICE don't care. So if 4x100G is cheaper / more redundant / doesn't require a new router chassis. Pick the better option. Or at least check your netflow and ask your users to show you a mythical pachyderm.

3. If you have a 10G network, then you will see 5G flows, and by all means engineer for elephants.

4. Next talk: "6 Gauge speaker wire, and Tube amplifiers" how do we engineer for that really deep bass.