An Analysis of Bulk Data Movement Patterns in Large-scale Scientific Collaborations

W. Wu, P. DeMar, A. Bobyshev (Fermilab)

CHEP 2010

Academia Sinica

Taipei, Taiwan

Office of Science

U.S. DEPARTMENT OF ENERGY

Fermilab

Topics

- Background
- Issues
- Fermilab Flow Data Collection & Analysis System
- Bulk Data Movement Pattern Analysis
  - Data analysis methodology
  - Bulk Data Movement Patterns

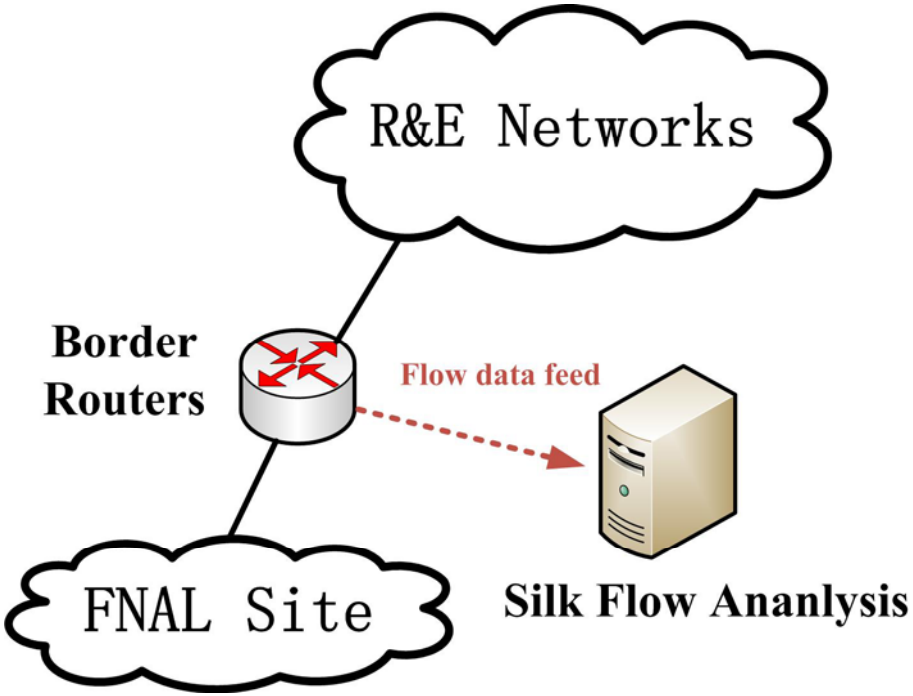
Background

- Large-scale research efforts such as LHC experiments, are built upon large, globally distributed collaborations.
  - Depend on predictable and efficient bulk data movement between collaboration sites.
- Data centers comprising hundreds to thousands of nodes are utilized as massively scaled, highly distributed cluster computing platforms.
- Existing transport protocols such as TCP do not work well in long fat pipes (LFPs).
- Parallel data transmission tools such as GridFTP have been widely applied to bulk data movements.

Issues

- What is the current status of bulk data movement in support of Large-scale Scientific Collaborations?
- What are the bulk data movement patterns in Large-scale Scientific Collaborations?

Fermilab Flow Data Collection & Analysis System



Flow-based analysis produces data that are more fine-grained than those provided by SNMP but still not as detailed and high-volume as required for packet trace analysis.

Fermilab Flow Data Collection and Analysis System

- Cisco NetFlow & CMU's SiLK toolset used for flow analysis
- NetFlow collection is complete; not sampled
  - Flow records exported to SiLK traffic analysis system.
- SiLK analysis suite is a collection of command-line tools for processing Flow records.

Bulk Data Movement Pattern Analysis

Data Analysis Methodology

- Flow records analyzed from 11/11/2009 to 12/23/2009.
  - Flow record database size = 60GBytes
    - 2,679,598,772 flow records.
  - Represented 23,764,073GByte of moved data between Fermilab & other sites
    - 2.221×10^{12} packets…
- Only TCP bulk data transfers analyzed
- Data transfers analyzed between Fermilab & /24 subnets
  - TOP 100 /24 sites that transfer to FNAL.
  - TOP 100 /24 sites that transfer from FNAL.

Bulk Data Movement Patterns & Status

TOP 100 Sites

- Analyzed data transfers between Fermilab and /24 subnets.
  - In the IN direction, the TOP 100 Sites transfer 99.04% traffic.
  - In the OUT direction, the TOP 100 sites transfer 95.69% traffic

INPUT: 525473576 Records for 1292588 Bins and 1342120011089719 Total Bytes
OUTPUT: Top 20 Bins by Bytes

dIP	Bytes	%Bytes	cumul_%
193.48.99.0	104925686638110	7.817906	7.817906
129.93.239.0	75583209636217	5.631628	13.449535
117.103.110.0	54956332106193	4.094741	17.544275
130.246.179.0	48627993633168	3.623222	21.167498
202.122.33.0	46725107353838	3.481440	24.648938
169.228.131.0	44228254392375	3.295402	27.944340
10.31.62.0	38363157396724	2.858400	30.802740
193.190.247.0	37788352590212	2.815572	33.618312
128.211.143.0	36018885849158	2.683731	36.302043
129.59.197.0	35477111134534	2.643364	38.945406
130.246.178.0	34008222083051	2.533918	41.479324
128.142.208.0	31483103104034	2.345774	43.825098
193.205.76.0	30965938511286	2.307241	46.132339
18.12.8.0	27271793859623	2.031994	48.164333
141.108.36.0	27170640242278	2.024457	50.188790
192.101.161.0	23196817209626	1.728371	51.917161
144.16.111.0	20022547614028	1.491860	53.409021
134.61.24.0	19812780178410	1.476230	54.885251
128.142.178.0	19345695468381	1.441428	56.326679
192.84.86.0	19152516130382	1.427035	57.753713

TOP20 Sites that transfer from FNAL

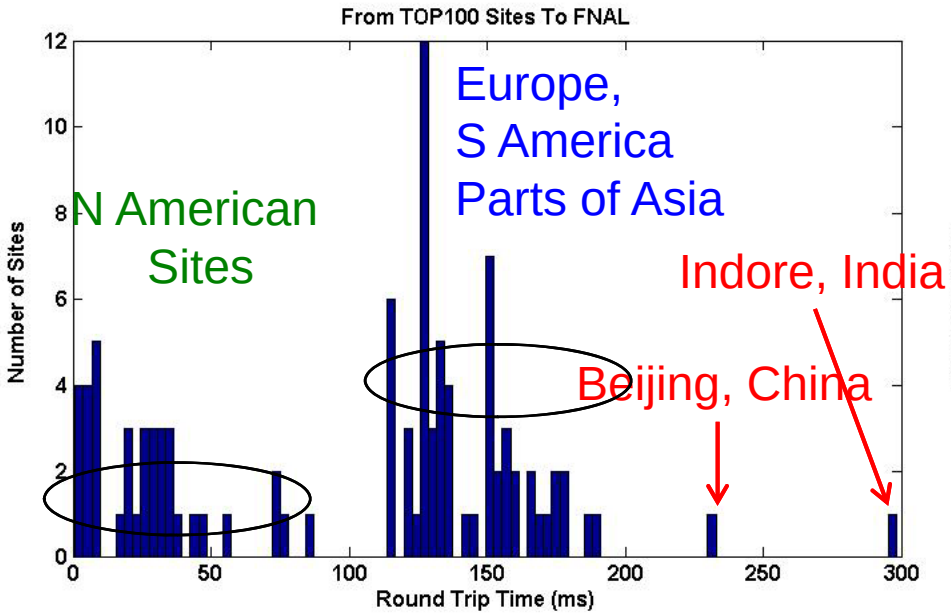
INPUT: 739219728 Records for 1257685 Bins and 971912972320223 Total Bytes
OUTPUT: Top 20 Bins by Bytes

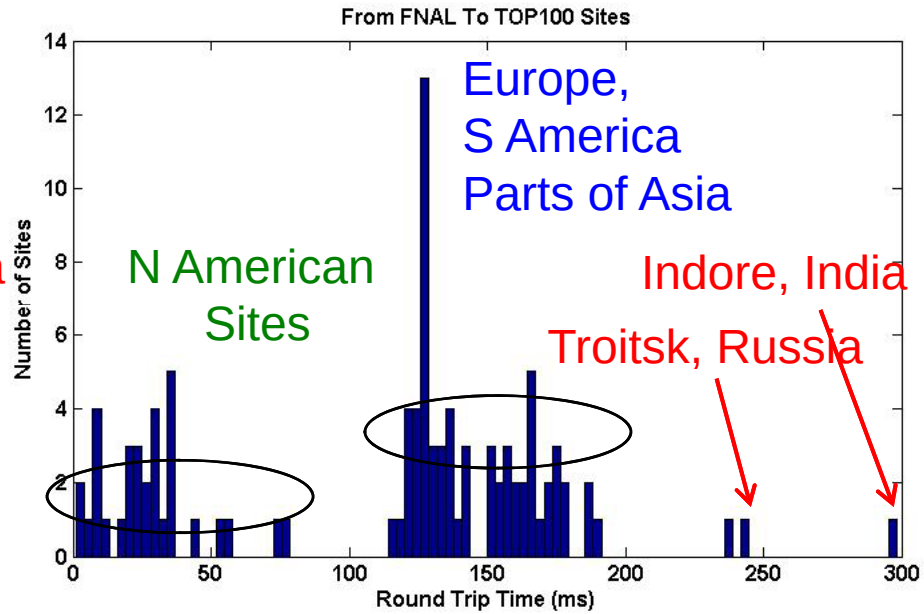
sIP|	Bytes|	%Bytes|	cumul_%|
128.142.178.0|	181261448306775|	18.649967|	18.649967|
128.142.208.0|	161148186915150|	16.580516|	35.230483|
128.142.218.0|	54066767661262|	5.562923|	40.793406|
128.142.215.0|	43732497666926|	4.499631|	45.293037|
152.54.1.0|	40493661525725|	4.166388|	49.459424|
128.142.220.0|	38698917020295|	3.981727|	53.441151|
128.211.143.0|	32047915439628|	3.297406|	56.738557|
131.154.129.0|	27023885838518|	2.780484|	59.519041|
128.135.70.0|	25718235171048|	2.646146|	62.165187|
129.93.227.0|	25659406689105|	2.640093|	64.805280|
192.84.86.0|	21846698205156|	2.247804|	67.053084|
128.227.221.0|	17649352242955|	1.815940|	68.869023|
140.221.82.0|	15946876669688|	1.640772|	70.509796|
18.12.8.0|	15922867380709|	1.638302|	72.148097|
141.142.31.0|	15306444330596|	1.574878|	73.722975|
192.108.47.0|	14289856548832|	1.470281|	75.193257|
144.92.181.0|	13591462491691|	1.398424|	76.591681|
10.31.62.0|	13106429140824|	1.348519|	77.940199|
128.142.219.0|	13079789714222|	1.345778|	79.285977|
169.228.131.0|	12077140266718|	1.242615|	80.528593|

TOP20 Sites that transfer to FNAL

TOP 100 Sites (Cont)

- TOP 100 sites distributed around the world.
- We measured the Round Trip Time (RTT) between FNAL and these TOP 100 sites both in IN and OUT directions.





A few sites have circuitous paths to/from FNAL.
Many factors: peering, traffic engineering, or cable path constraints.

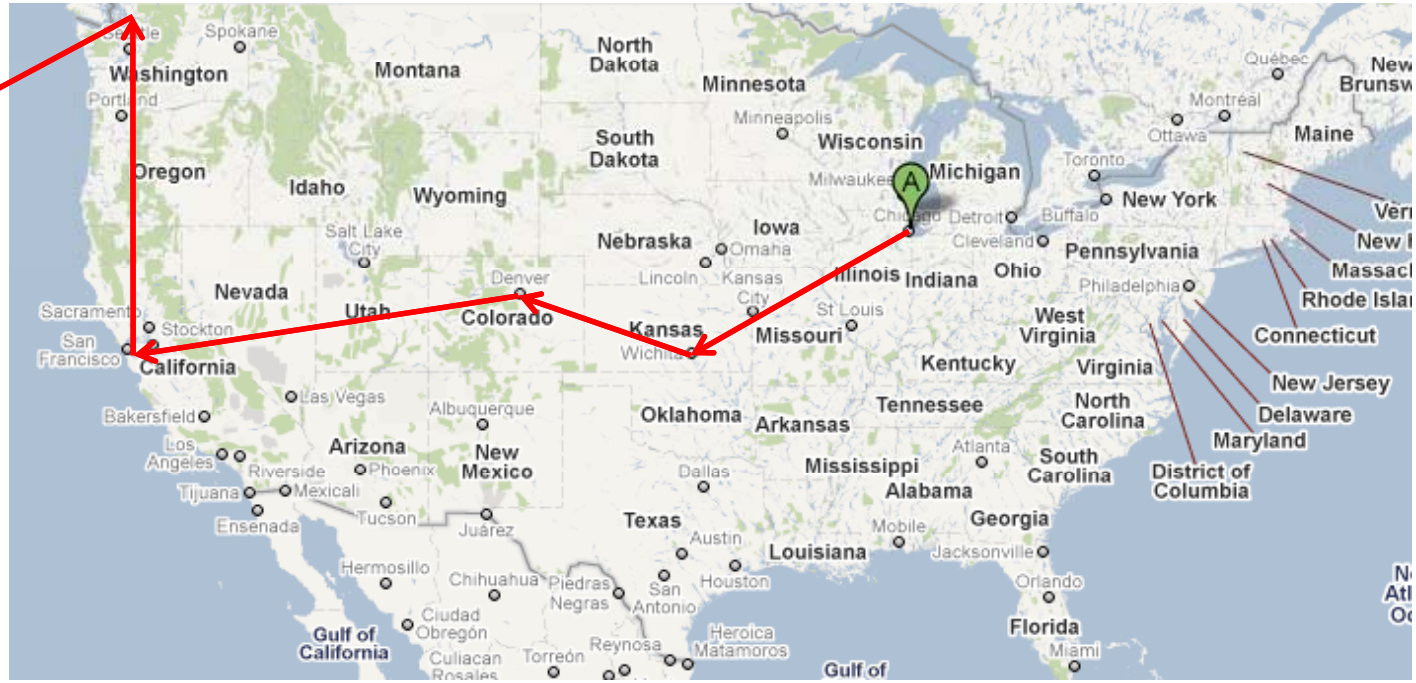
```
traceroute to 144.16.111.0 (144.16.111.0), 64 hops max, 40 byte packets
 1  r-s-edge-1-vlan-173-2.fnal.gov (131.225.173.2)  3.915 ms  1.375 ms  1.393 ms
 2  vlan375.r-s-bdr.fnal.gov (131.225.15.202)  1.031 ms  1.346 ms  1.004 ms
 3  fnal-mr2.fnal.gov (198.49.208.229)  0.963 ms  1.297 ms  0.928 ms
 4  fnalmr3-ip-fnalmr2.es.net (134.55.41.41)  0.998 ms  1.182 ms  1.489 ms
 5  chiccr1-ip-fnalmr3.es.net (134.55.219.121)  2.653 ms  2.767 ms  3.239 ms
 6  kanscr1-ip-chiccr1.es.net (134.55.221.57)  13.783 ms  13.226 ms  13.469 ms
 7  denvcr2-ip-kanscr1.es.net (134.55.209.45)  26.581 ms  27.121 ms  26.833 ms
 8  sunncr1-denvcr2.es.net (134.55.220.50)  55.199 ms  53.522 ms  53.430 ms
 9  transpac-1-is-jmb-780.lsanca.pacificwave.net (207.231.246.136)  62.130 ms  94.630 ms  61.829 ms
10  tokyo-losa-tp2.transpac2.net (192.203.116.146)  177.212 ms  177.060 ms  177.086 ms
11  tyo-gate1.jp.apan.net (203.181.248.249)  177.312 ms  178.064 ms  177.349 ms
12  sg-so-03-v4.bb.tein3.net (202.179.249.49)  257.951 ms  257.618 ms  259.626 ms
13  mb-so-01-v4.bb.tein3.net (202.179.249.82)  319.240 ms  318.853 ms  319.461 ms
14  in-pr-v4.bb.tein3.net (202.179.249.90)  329.601 ms  318.982 ms  318.906 ms
15  202.141.153.25 (202.141.153.25)  319.072 ms  319.058 ms  319.188 ms
16  tifr-mum.backbone.ernet.in (202.141.144.106)  302.666 ms  304.148 ms *
```

Chicago IL, USA

Kansas, USA
Denver, USA
Sunnyvale, USA
Seattle, USA
Tokyo, JP
Singapore, SG

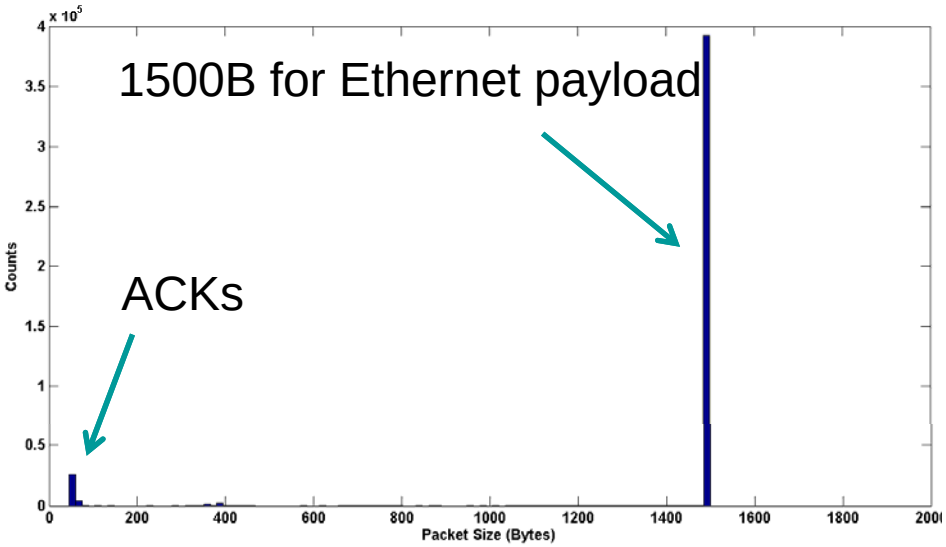
Traceroute from FNAL to Indore, India

To Tokyo

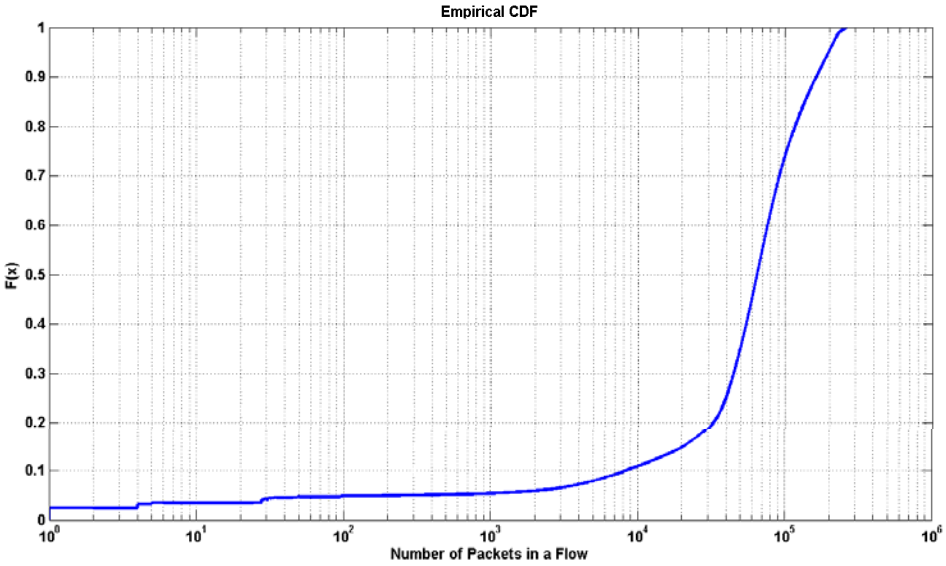


Single Flow Throughputs

- We calculated statistics for single flow throughputs between FNAL and TOP100 Sites in both IN and OUT directions.
  - Each flow record includes data such as the number of packets and bytes in the flow and the timestamps of the first and last packet.
- A few concerns
  - Should exclude the flow records for pure ACKs of the reverse path
  - A bulk data movement usually involves frequent administrative message exchanges between sites. These flow records usually contain a small number of packets with short durations. These flow records are excluded from throughput calculation.

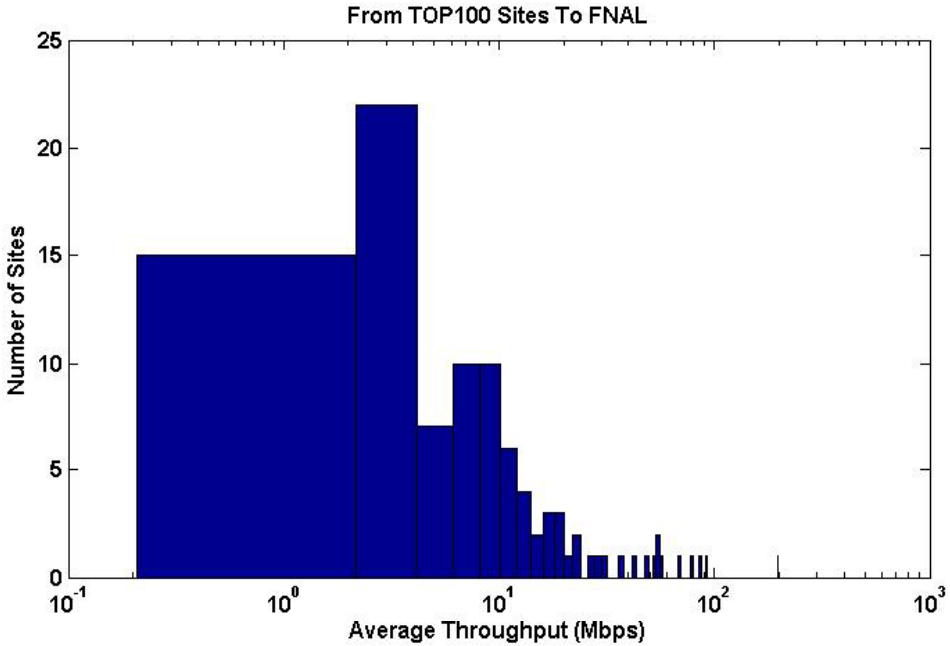


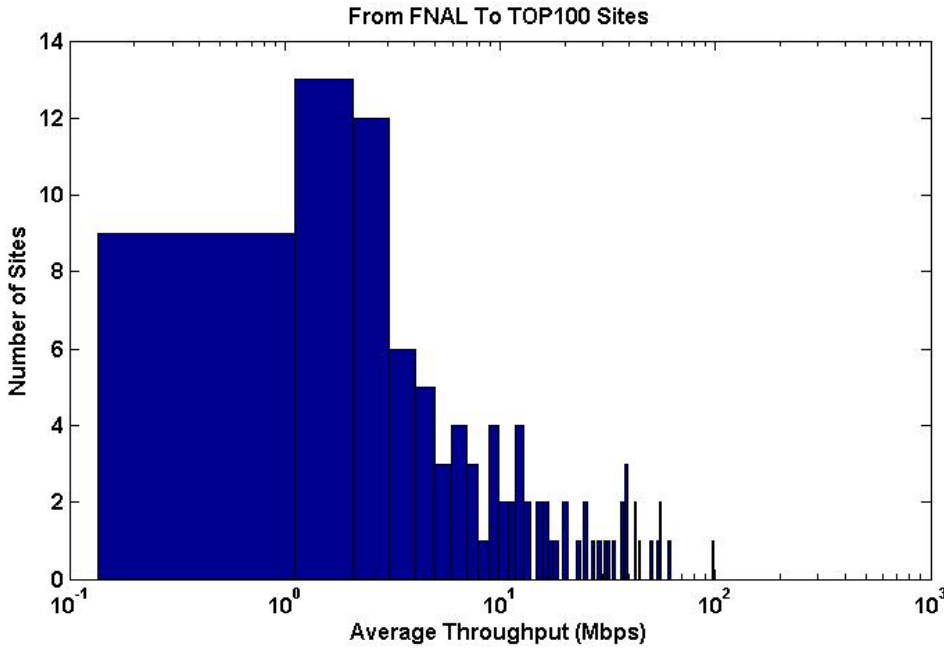
Histogram for Packet Size



Histogram for Flow Size

Single Flow Throughputs (cont)





Most average throughputs are less than 10Mbps!
But 1Gbps NICs are widely deployed!

The slowest throughput is 0.135 Mbps,

Houston, TX!!!

The slowest throughput is 0.207 Mbps,

Indore, India

Single Flow Throughputs (cont)

- In the IN direction
  - 2 sites' average throughputs are less than 1Mbps
    - 63 sites' average throughputs are less than 10Mbps
  - Only 1 site' average throughput is greater than 100Mbps
- In the OUT direction
  - 7 sites' average throughput are less than 1Mbps
  - 60 sites' average throughput are less than 10Mbps
  - No site's average throughput is greater than 100Mbps

(Max) TCP Throughput <= ~0.7 * MSS / (rtt * sqrt(packet_loss))

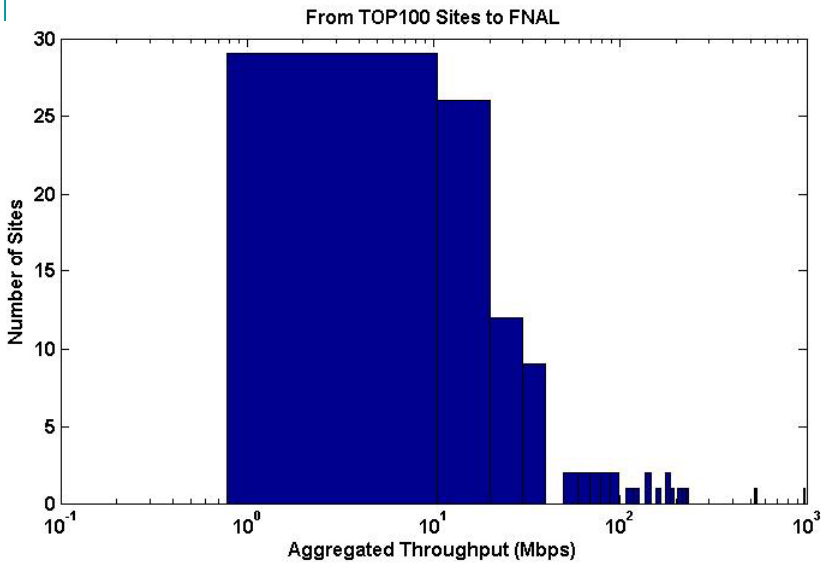
We calculate the correlation of "average throughput vs. RTT"

- In the IN direction, it is -0.426329
- In the OUT direction, it is -0.37256

Aggregated Throughputs

- Parallel data transmission tools such as GridFTP have been widely applied to bulk data movements
- In both IN and OUT directions, for each of the TOP100 sites, we bin traffic at 10 minute intervals
  - We calculate the aggregated throughputs
  - We collect the flow statistics
  - We collect the statistics of # of different IPs involved (hosts)

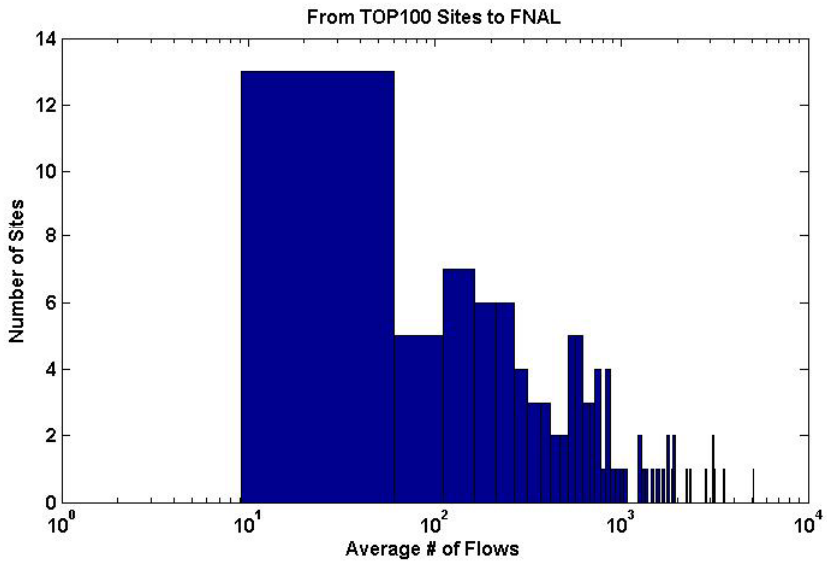
From TOP100 Sites to FNAL



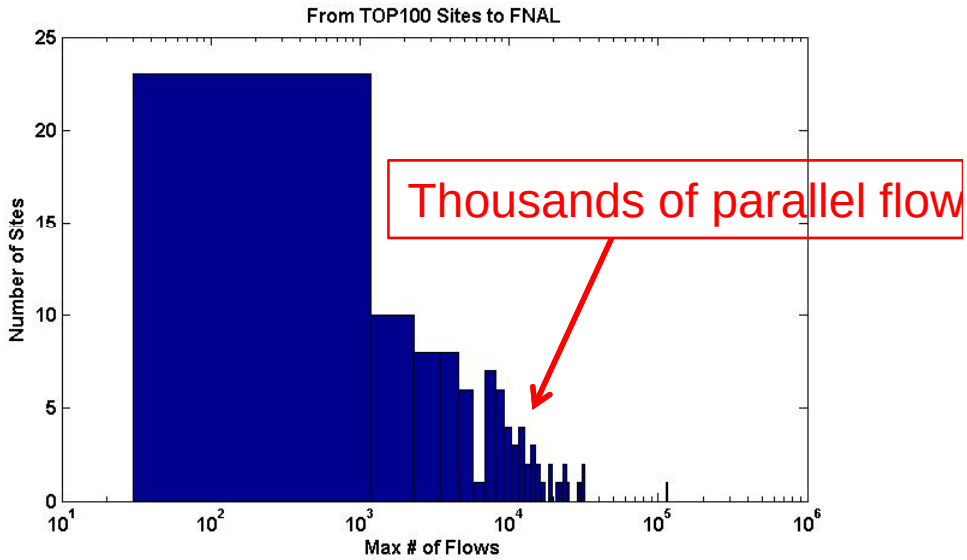
Histogram of Aggregated Throughputs

In general, the aggregated throughputs are higher

We see the effect of parallel data transmission

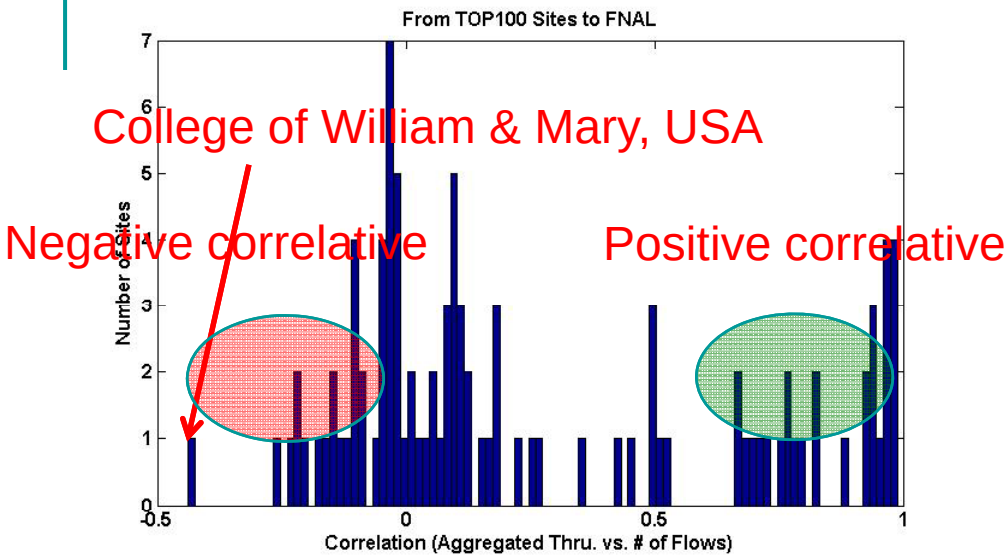


Histogram of Average # of Flows

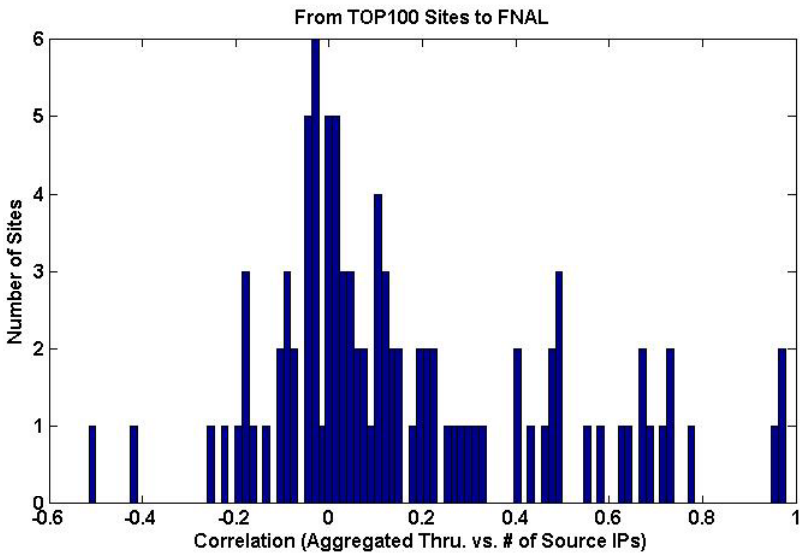


Histogram of Max # of Flows

From TOP100 Sites to FNAL



His. of Correlation (Aggregated Thru vs. # of Flows)



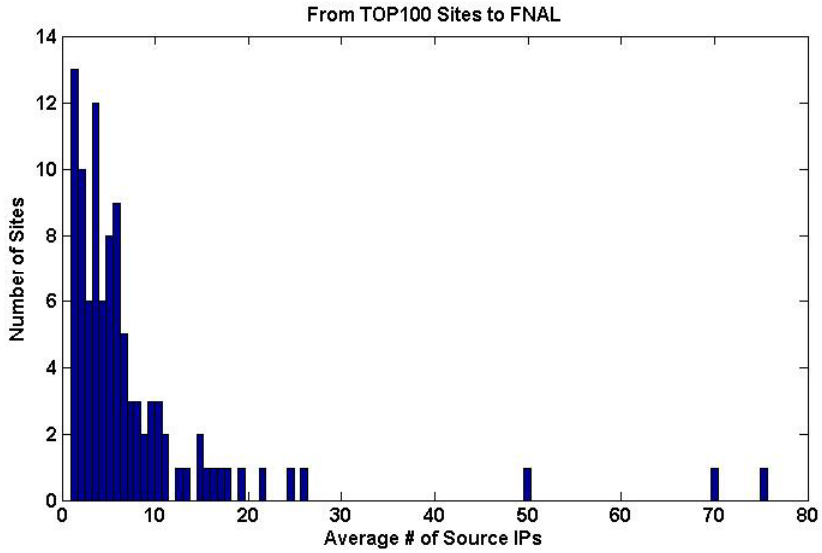
His. of Correlation (Aggregated Thru vs. # of Source IPs)
His. Of Corr. (Aggr. Thru vs. # of Dest. Ips) is similiar

- We calculate the correlation between aggregated throughputs vs. number of flow
- In general, more parallel data transmission (# of flows) generate higher aggregated throughputs
- But for some sites, more parallel data transmission generate less aggregated throughputs
  - More parallel data transmission causes network congestion
  - parallel data transmission make disk I/O less efficient.

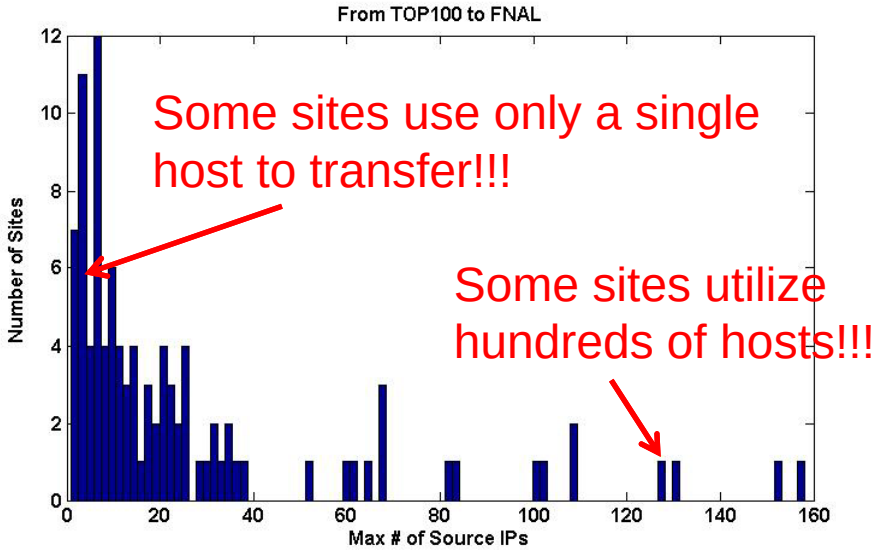
From TOP100 Sites to FNAL

- Totally, there are 35 sites that the “correlation between aggregated throughputs vs. number of flow” is negative.
  - The worst case is from “College of William & Mary, USA” to FNAL, the correlation is only -0.439
- There are 31 sites that the “correlation between aggregated throughputs vs. number of flow” is greater than 0.5, which implies that increasing the number of flows can effectively enhance the throughputs.
  - The best case is from “University of Glasgow, UK” to FNAL.

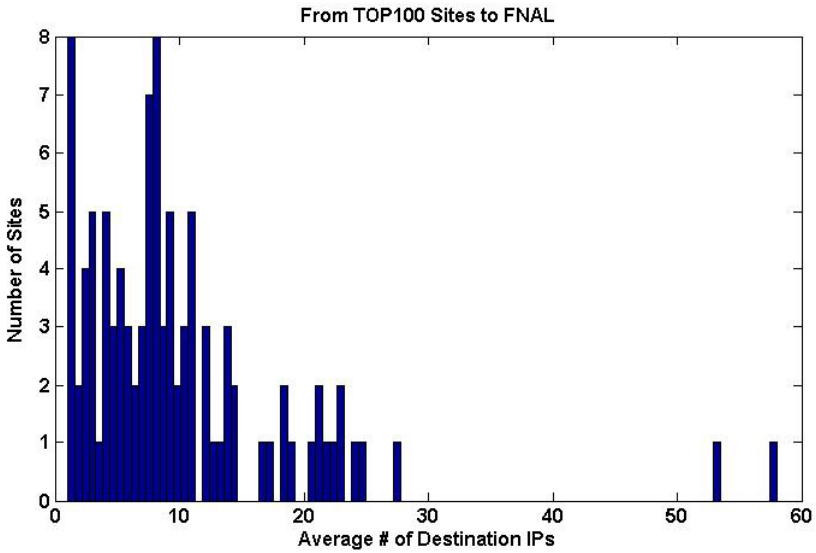
From TOP100 Sites to FNAL



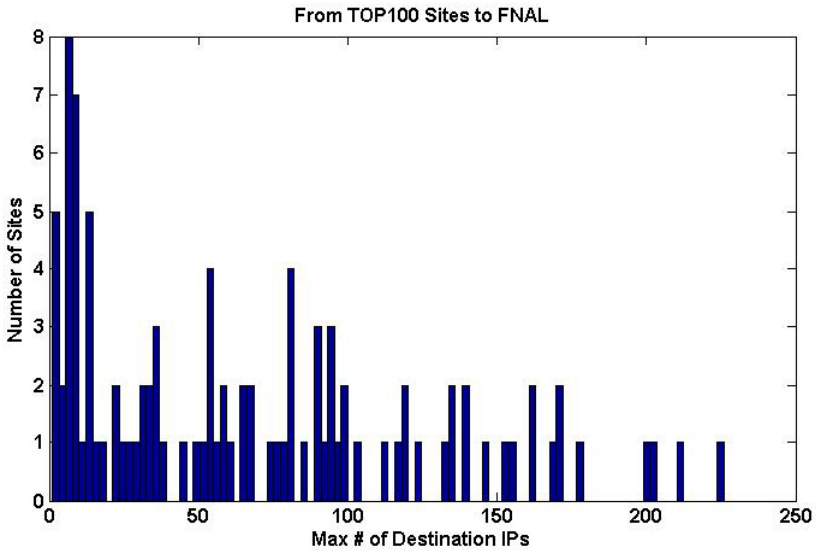
Histogram of Average # of Source IPs



Histogram of Max # of Source IPs

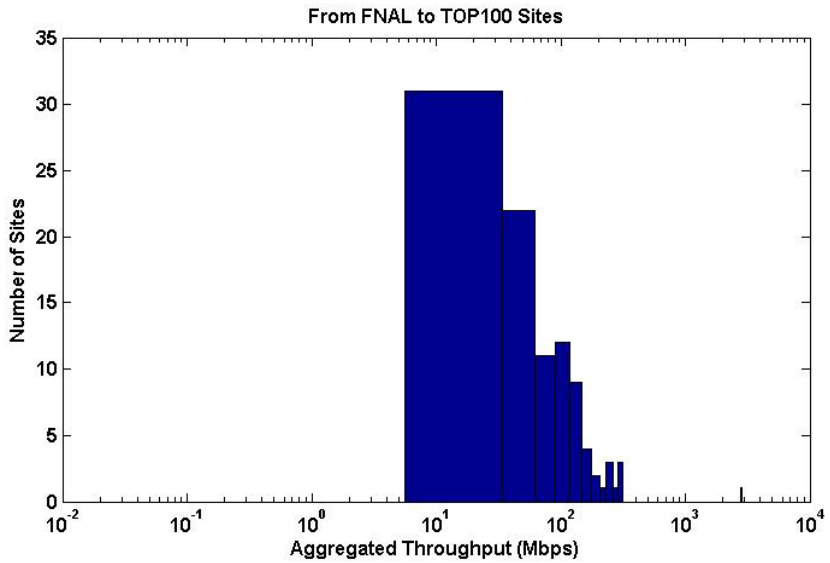


Histogram of Average # of Destination IPs



Histogram of Max # of Destination IPs

From FNAL to TOP100 Sites

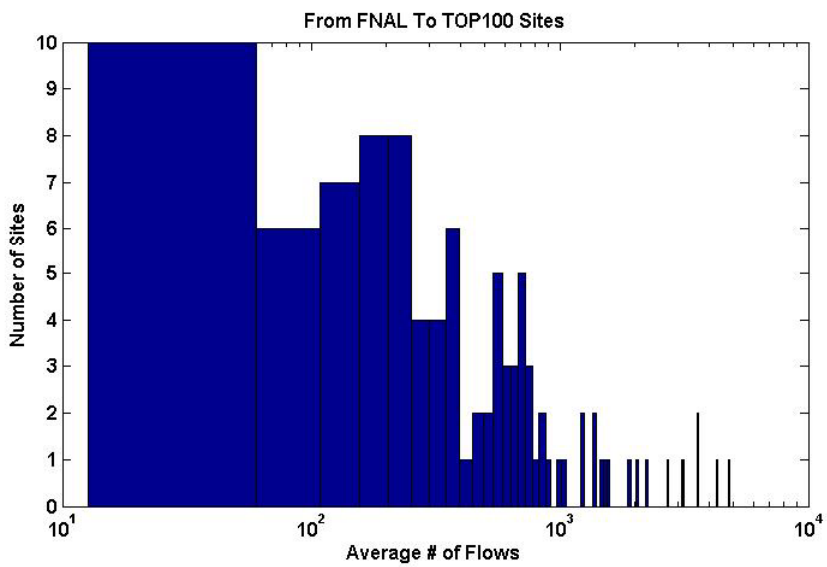


Histogram of Aggregated Throughputs

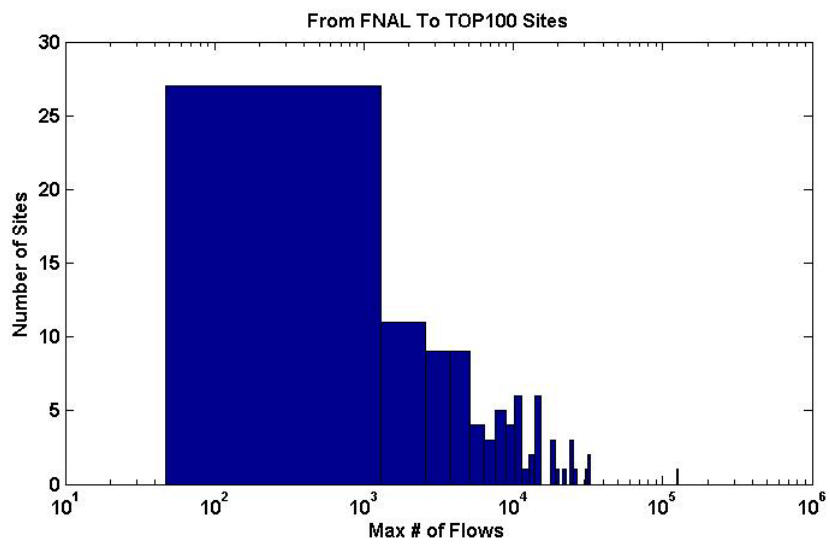
In general, the aggregated throughputs are highe

We see the effect of parallel data transmission

The transmission from FNAL to TOP100 Sites is better than the other way around.

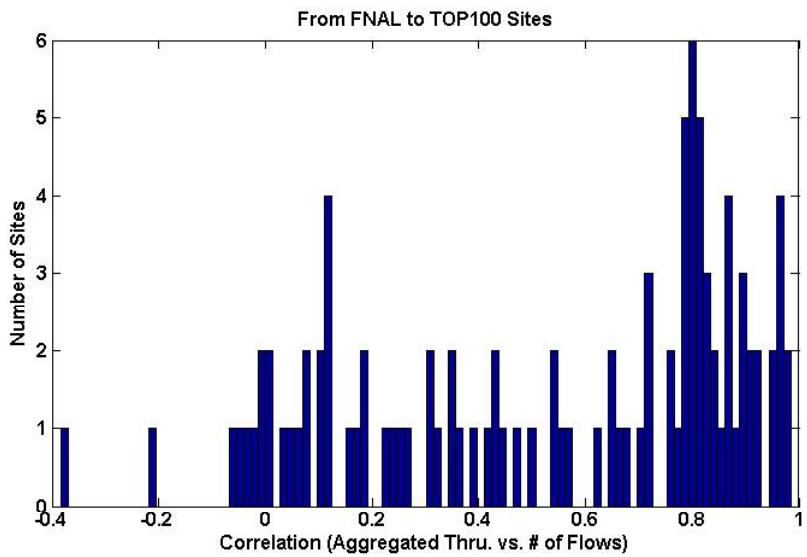


Histogram of Average # of Flows

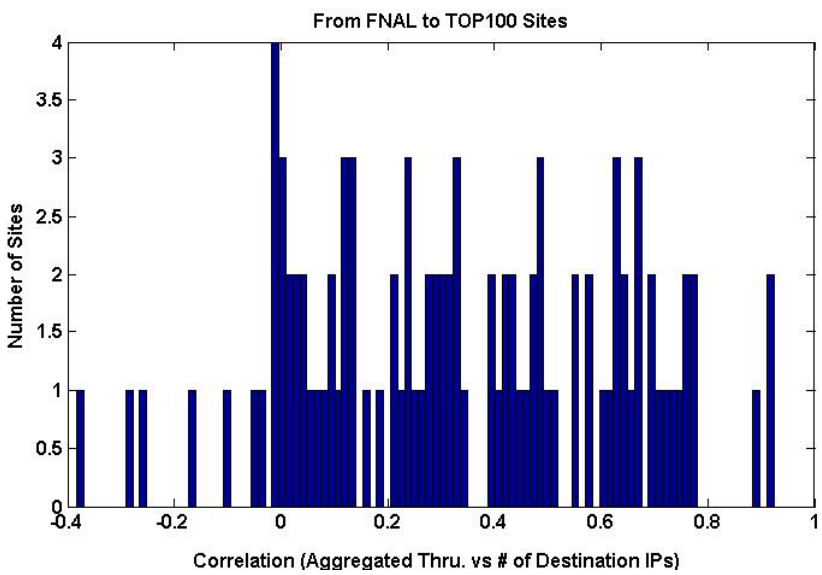


Histogram of Max # of Flows

From FNAL to TOP100 Sites



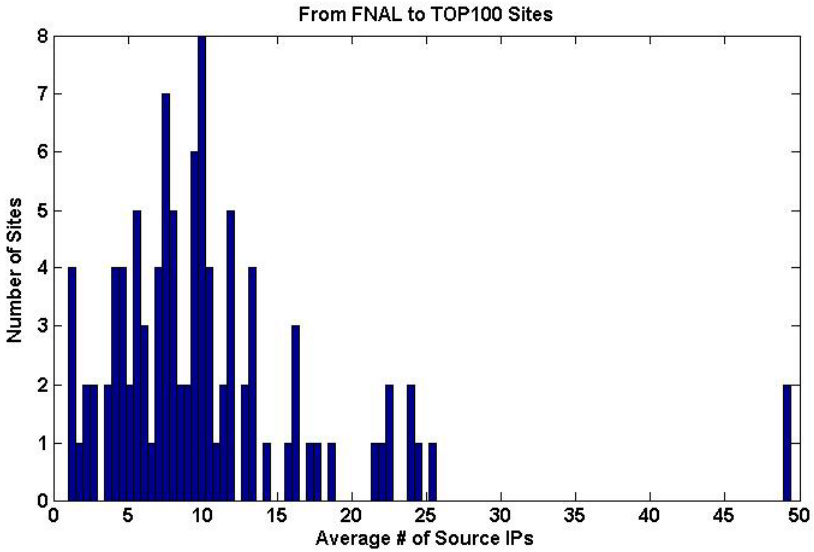
His. of Correlation (Aggregated Thru vs. # of Flows)



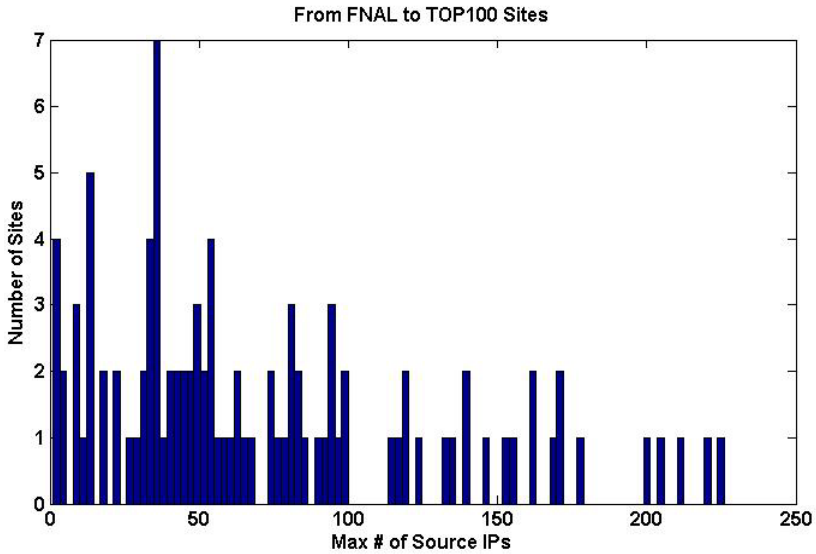
His. of Correlation (Aggregated Thru vs. # of Dest. IPs)
His. Of Corr. (Aggr. Thru vs. # of Src. Ips) is similiar

- We calculate the correlation between aggregated throughputs vs. number of flow
- In general, more parallel data transmission (# of flows) generate higher aggregated throughputs
- But for some sites, more parallel data transmission generate less aggregated throughputs
  - More parallel data transmission causes network congestion
  - parallel data transmission make disk I/O less efficient.

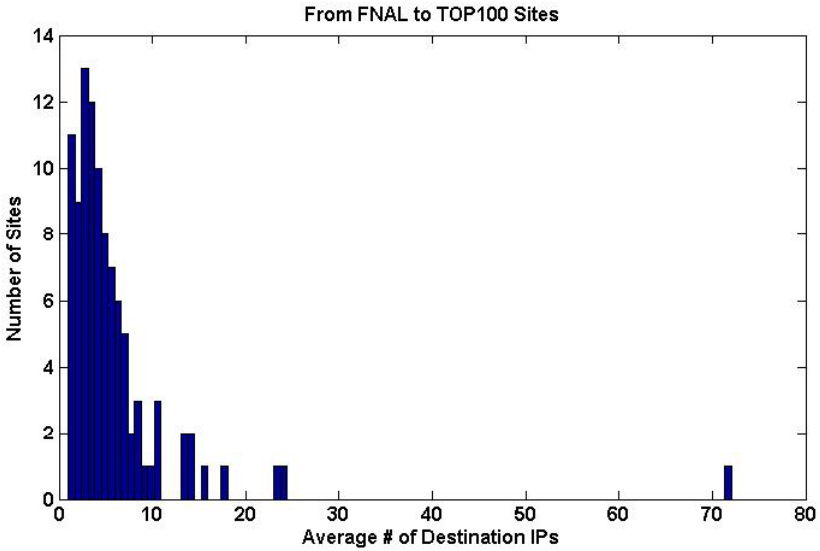
From FNAL to TOP100 Sites



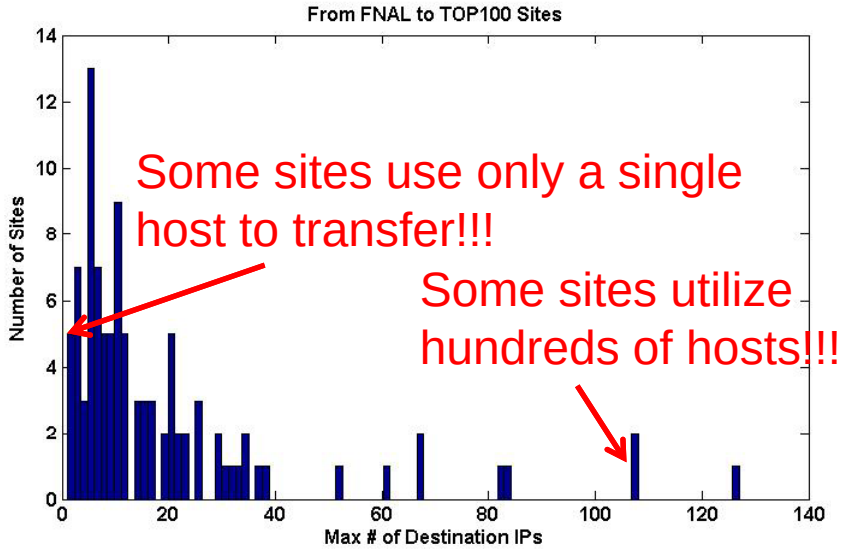
Histogram of Average # of Source IPs



Histogram of Max # of Source IPs



Histogram of Average # of Destination IPs



Histogram of Max # of Destination IPs