Hybrid network traffic engineering system (HNTES)

Zhenzhen Yan, Zhengyang Liu, Chris Tracy, Malathi Veeraraghavan
University of Virginia and ESnet
Jan 12-13, 2012
mvee@virginia.edu, ctracy@es.net

Project web site: http://www.ece.virginia.edu/mv/research/DOE09/index.html

Thanks to the US DOE ASCR program office and NSF for UVA grants DE-SC002350, DE-SC0007341, OCI-1127340 and ESnet grant DE-AC02-05CH11231

Problem statement

- A hybrid network supports both IP-routed and circuit services on:
  - Separate networks as in ESnet4, or
  - An integrated network as in ESnet5
- A hybrid network traffic engineering system (HNTES) is one that moves science data flows to circuits
- Problem statement: Design HNTES

Two reasons for using circuits

1. Offer scientists rate-guaranteed connectivity
2. Isolate science flows from general-purpose flows

Circuit scope \ Reason	Rate-guaranteed connections	Science flow isolation
End-to-end (inter-domain)	✔	✖
Per provider (intra-domain)	✖	✔

Request to sites:

- Any information on trouble tickets created by science flows would be appreciated

What type of flows should be isolated?

- Dimensions
  - size (bytes): elephant and mice
  - rate: cheetah and snail
  - duration: tortoise and dragonfly
  - burstiness: porcupine and stingray

Kun-chan Lan and John Heidemann, A measurement study of correlations of Internet flow characteristics. *ACM Comput. Netw.* 50, 1 (January 2006), 46-62.

alpha flows

- number of bytes in any T-sec interval $\geq$ H bytes
  - if H = 1 GB and T = 60 sec
    - throughput exceeds 133 Mbps
- alpha flows responsible for burstiness
- alpha flows are caused by transfers of large files over fast links
  - Let's look at GridFTP usage statistics

S. Sarvotham, R. Riedi, and R. Baraniuk, "Connection-level analysis and modeling of nework traffic," in ACM SIGCOMM Internet Measurement Workshop 2001, November 2001, pp. 99–104.

GridFTP log analysis

- Two goals:
  - Determine durations of high-throughput GridFTP transfers
    - to use dynamic circuits, since current IDC circuit setup delay is ~1 min, need transfer durations to be say 10 mins
  - Characterize variance in throughput
    - identify causes

GridFTP data analysis findings

- GridFTP transfers from NERSC dtn servers that > 100 MB in one month (Sept. 2010)
- Total number of transfers: 124236
- GridFTP usage statistics

TABLE I: Summary of all NERSC transfers larger than 100 MB; the three columns are independent, e.g., the transfer with the largest size is not the same transfer as the one with the longest duration or the one with the highest throughput

	Size (Bytes)	Duration (s)	Throughput (bps)
Min	1.000e+08	0.2488	1.266e+06
1st Quartile	1.049e+08	1.9229	1.713e+08
Median	1.049e+08	2.4919	3.480e+08
Mean	2.531e+08	35.4022	3.557e+08
3rd Quartile	1.261e+08	8.8897	4.445e+08
Max	9.679e+10	9952.2382	4.315e+09

Thanks to Brent Draney, Jing Tie and Ian Foster for the GridFTP data

Top quartile highest-throughput transfers NERSC (100MB dataset)

	Min	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Throughput (Mb/s)	444.5	483.0	596.3	698.8	791.9	4315

- Total number: 31059 transfers
- 50% of this set had duration < 1.51 sec
- 75% had duration < 1.8 sec
- 95% had duration < 3.36 sec
- 99.3% had duration < 1 min
- 169 (0.0054%) transfers had duration > 2 mins
- Only 1 transfer had duration > 10 mins
- Need to look for multi-transfer sessions

Throughput variance

TABLE II: Summary of all 32 GB NERSC transfers

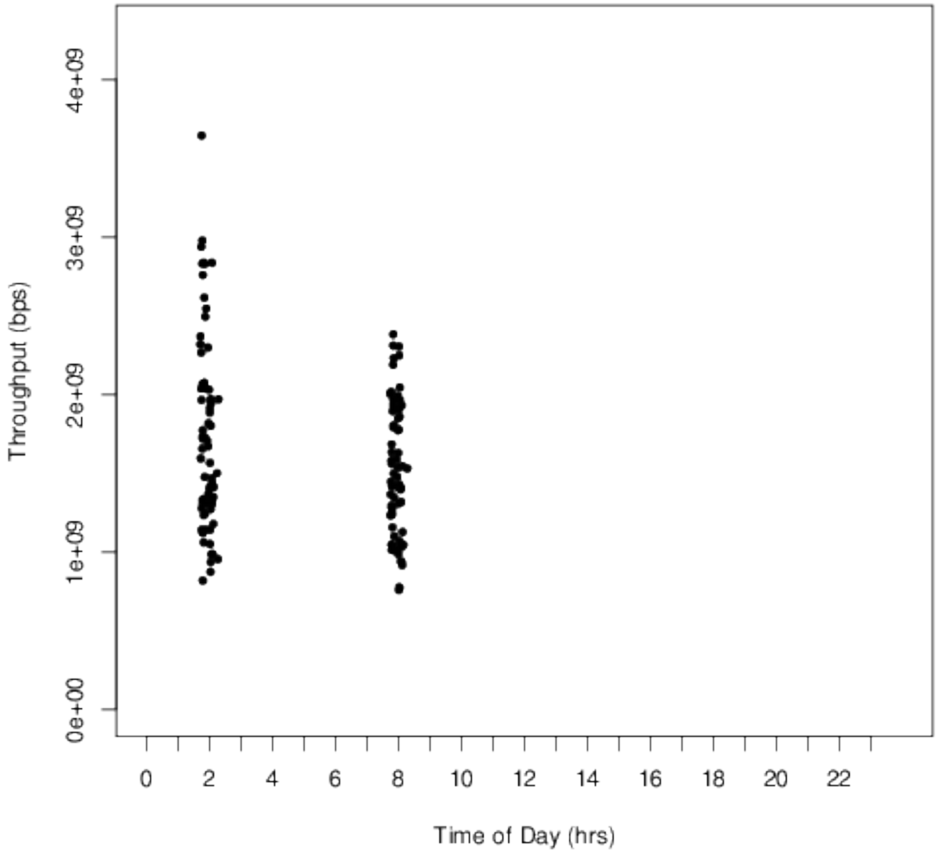
	Duration (s)	Throughput (bps)
Min	75.4	7.579e+08
1st Qu.	141.20	1.251e+09
Median	183.40	1.499e+09
Mean	186.60	1.625e+09
3rd Qu.	219.70	1.947e+09
Max	362.70	3.644e+09

- There were 145 file transfers of size 32 GB to same client
  - Same round-trip time (RTT), bottleneck link rate and packet loss rate
- IQR (Inter-quartile range) measure of variance is 695 Mbps
- Need to find an explanation for this variance

Potenial causes of throughput variance

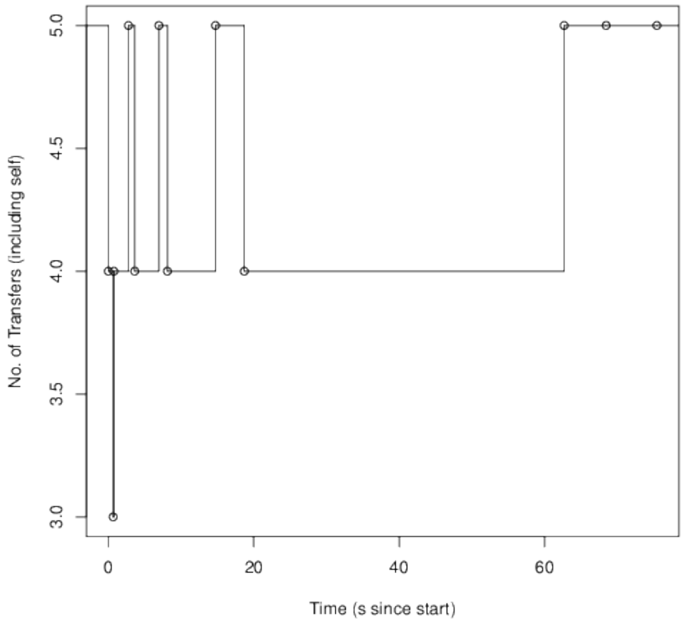
- Path characteristics:
  - RTT, bottleneck link rate, packet loss rate
  - Usage stats do not record remote IP address
  - Can extract from NetFlow data for alpha flows
- Number of stripes
- Number of parallel TCP streams
- Time-of-day dependence
- Concurrent GridFTP transfers
- Network link utilization (SNMP data)
- CPU usage, I/O usage on servers at the two ends

Time-of-day dependence (NERSC 32 GB: same path)



- Two sets of transfers: 2 AM and 8 AM
- Higher throughput levels on some 2 AM transfers
- But variance even among same time-of-day flows

Dep. on concurrent transfers: Predicted throughput



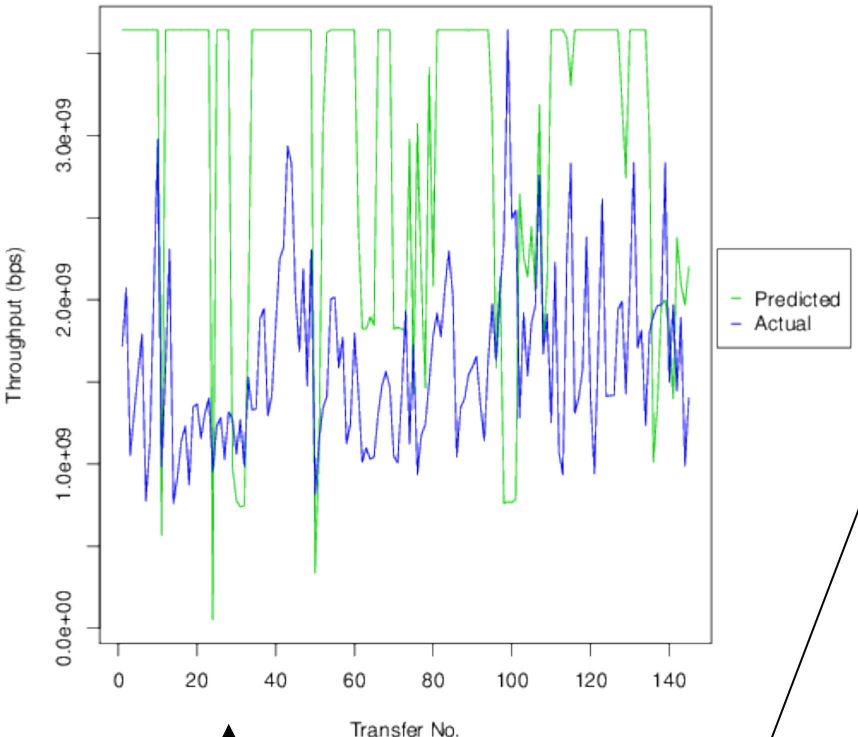
$$\tilde{T}_i = T_{max} \sum_{j=1}^{j_{max}} \frac{1}{n_{ij}} \times \frac{d_{ij}}{D_i}$$

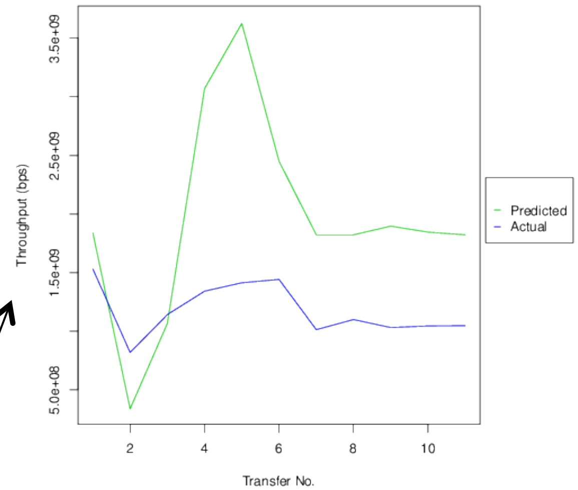
- Find number of concurrent transfers from GridFTP logs for i^{th} 32 GB GridFTP transfer: NERSC end only
- Determine predicted throughput
- d_{ij}: duration of j^{th} interval of i^{th} transfer
- n_{ij}: number of concurrent transfers in j^{th} interval of ith transfer

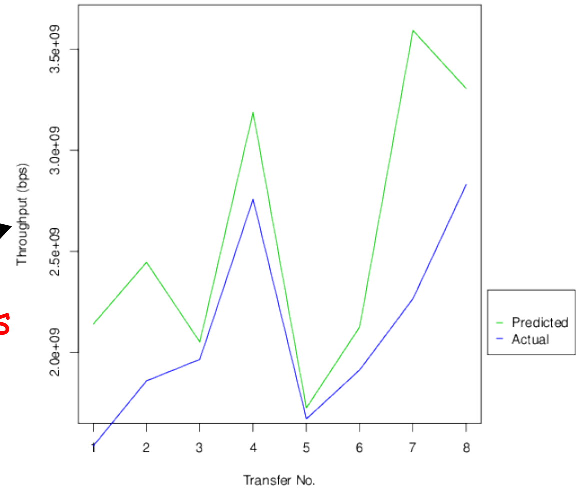
ESnet Energy Sciences Network

University of Virginia 1819

Dependence on concurrent transfers (NERSC 32 GB transfers)







Correlation seen for some transfers
But overall correlation low (0.03)
expl: Other apps besides GridFTP

Correlation with SNMP data

Correlation between GridFTP bytes and total SNMP reported bytes

	if1	if2	if3	if4	if5
1st Qu.	0.677	0.604	0.719	0.750	0.749
2nd Qu.	0.419	0.147	0.138	0.327	0.294
3rd Qu.	0.538	0.592	0.543	0.415	0.371
4th Qu.	0.782	0.872	0.797	0.789	0.790
All	0.902	0.922	0.919	0.918	0.918

Correlation between GridFTP bytes and other flow bytes

	if1	if2	if3	if4	if5
1st Qu.	0.254	0.188	0.429	0.505	0.486
2nd Qu.	0.269	-0.067	-0.110	0.089	0.071
3rd Qu.	0.059	0.157	0.110	0.015	-0.039
4th Qu.	0.196	0.328	0.239	0.287	0.276
All	0.351	0.365	0.443	0.524	0.527

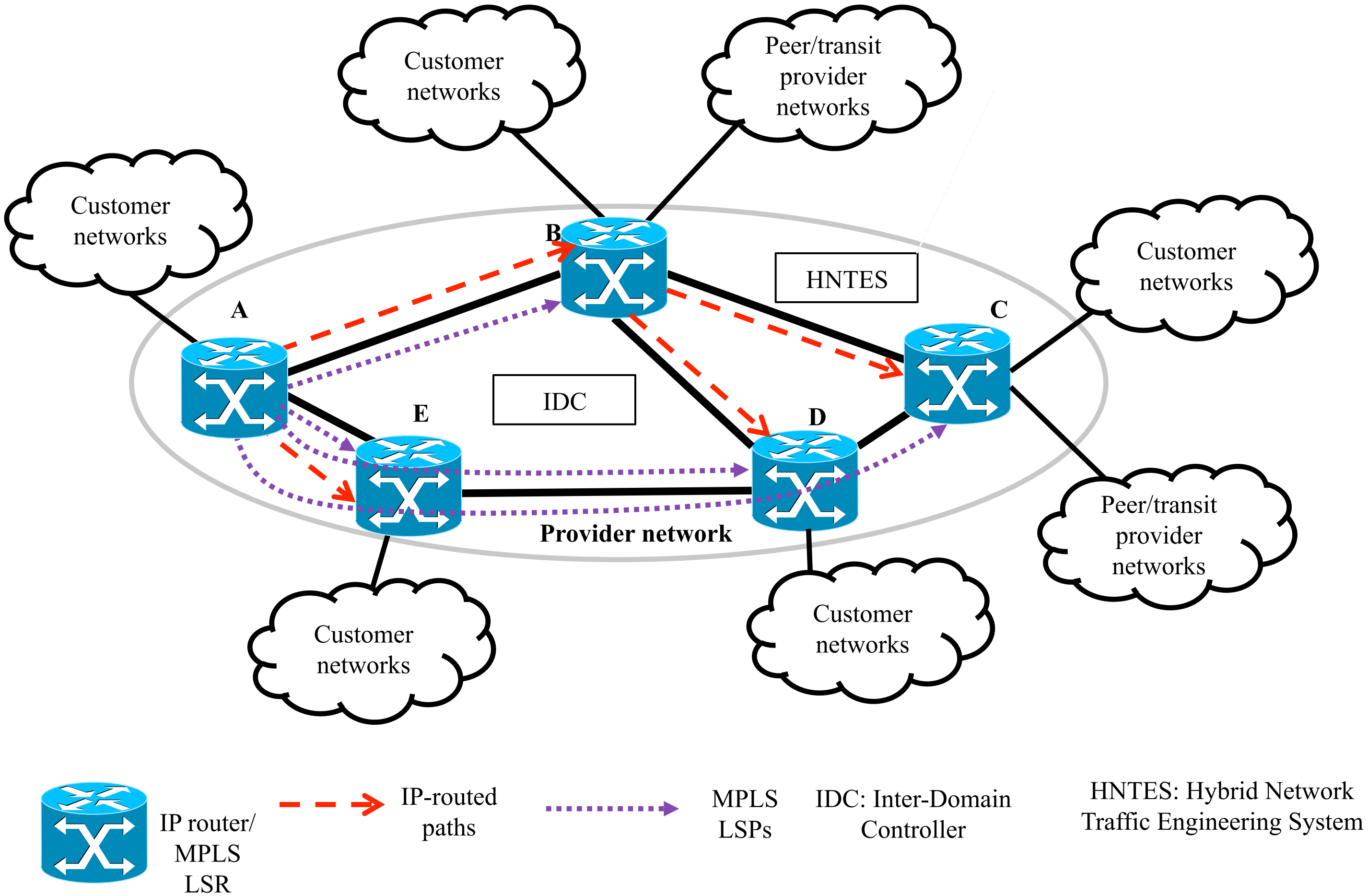
- SNMP raw byte counts: 30 sec polling
- Assume GridFTP bytes uniformly distributed over duration
- Ordered GridFTP transfers by throughput
- Conclusion: GridFTP bytes dominate and are not affected by other transfers – consistent with alpha behavior

Thanks to Jon Dugan for the SNMP data

Request from sites

- Permission to view GridFTP usage statistics
- Performance monitoring of DTN servers
  - File system usage
  - CPU usage
- MRTG data from site internal links
- Trouble ticket information

Back to HNTES: Role
Usage within domains for science flow isolation



- Ingress routers would be configured by HNTES to move science flows to MPLS LSPs

Three tasks executed by HNTES

1. alpha flow identification
   - Offline flow analysis
   - Online flow analysis
   - End-host assisted

2. Circuit Provisioning
   - Rate-unlimited MPLS LSPs initiated offline
   - Rate-unlimited MPLS LSPs initiated online
   - Rate-specified MPLS LSPs initiated online

3. Policy Based Route (PBR) configuration at ingress/egress routers
   - Set offline
   - Set online

online:
upon flow arrival

offline: periodic process (e.g., every hour or every day)

Questions for HNTES design

- Online or offline?
- PBRs: 5-tuple identifiers or just src/dst addresses?
- /24 or /32?
- How should PBR table entries be aged out?

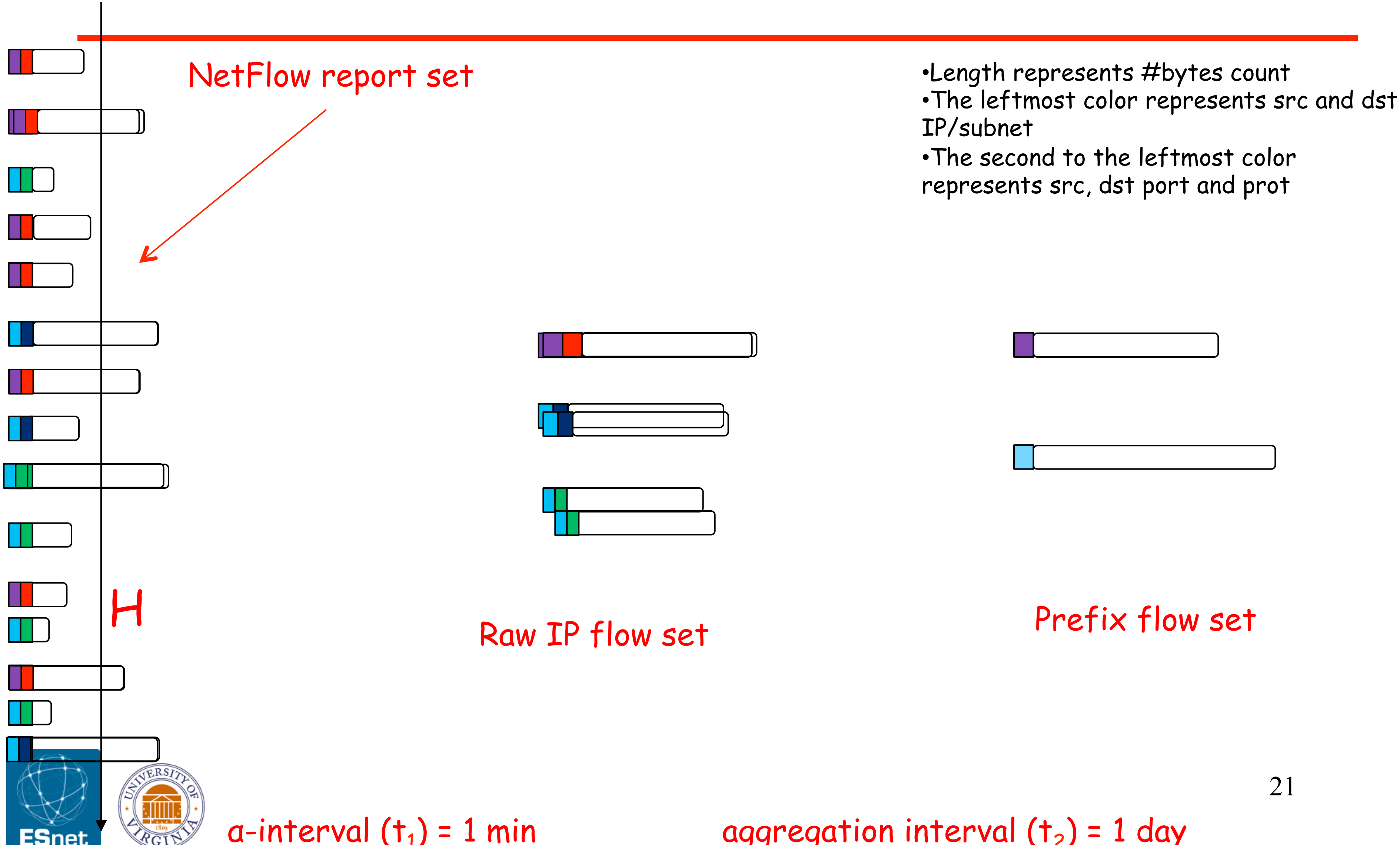
NetFlow data analysis

- NetFlow data over 7 months (May-Nov 2011) collected at ESnet site PE router
- Three steps
  - UVA wrote R analysis and anonymization programs
  - ESnet executed on NetFlow data
  - Joint analysis of results

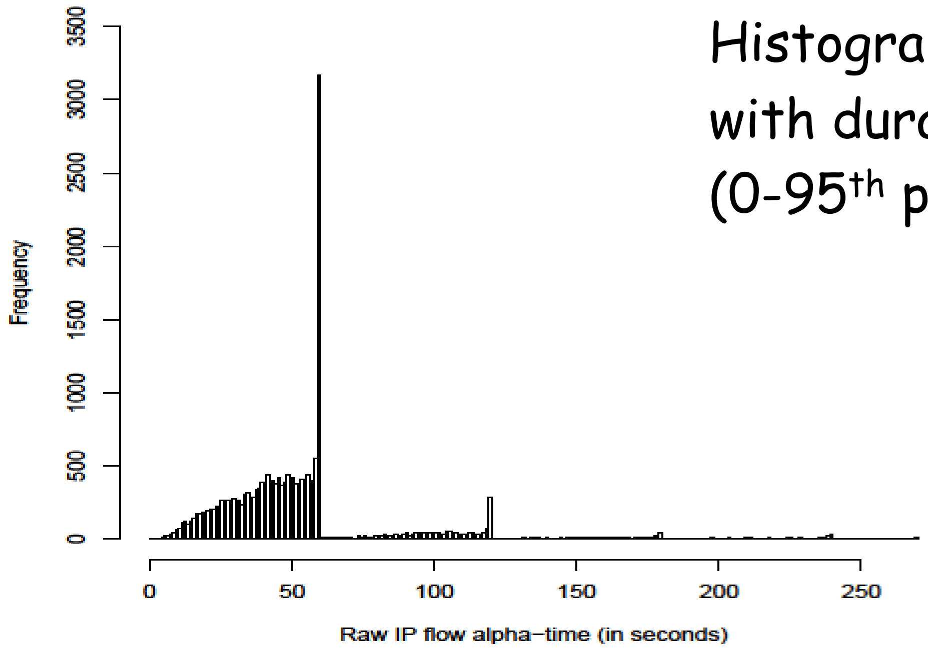
Flow identification algorithm

- alpha flows: high rate flows
  - NetFlow reports: subset where bytes sent in 1 minute > H bytes (1 GB)
  - Raw IP flows: 5 tuple based aggregation of reports on a daily basis
  - Prefix flows: /32 and /24 src/dst IP
  - Super-prefix flows: (ingress, egress) router based aggregation of prefix flows
- 7-month data set
  - 22041 raw IP flows, 125 (/24) prefix flows, and 1548 (/32) prefix flows

Flow aggregation from NetFlow



Online vs. offline



Histogram of α-flows with duration < 4.5mins (0-95th percentile)

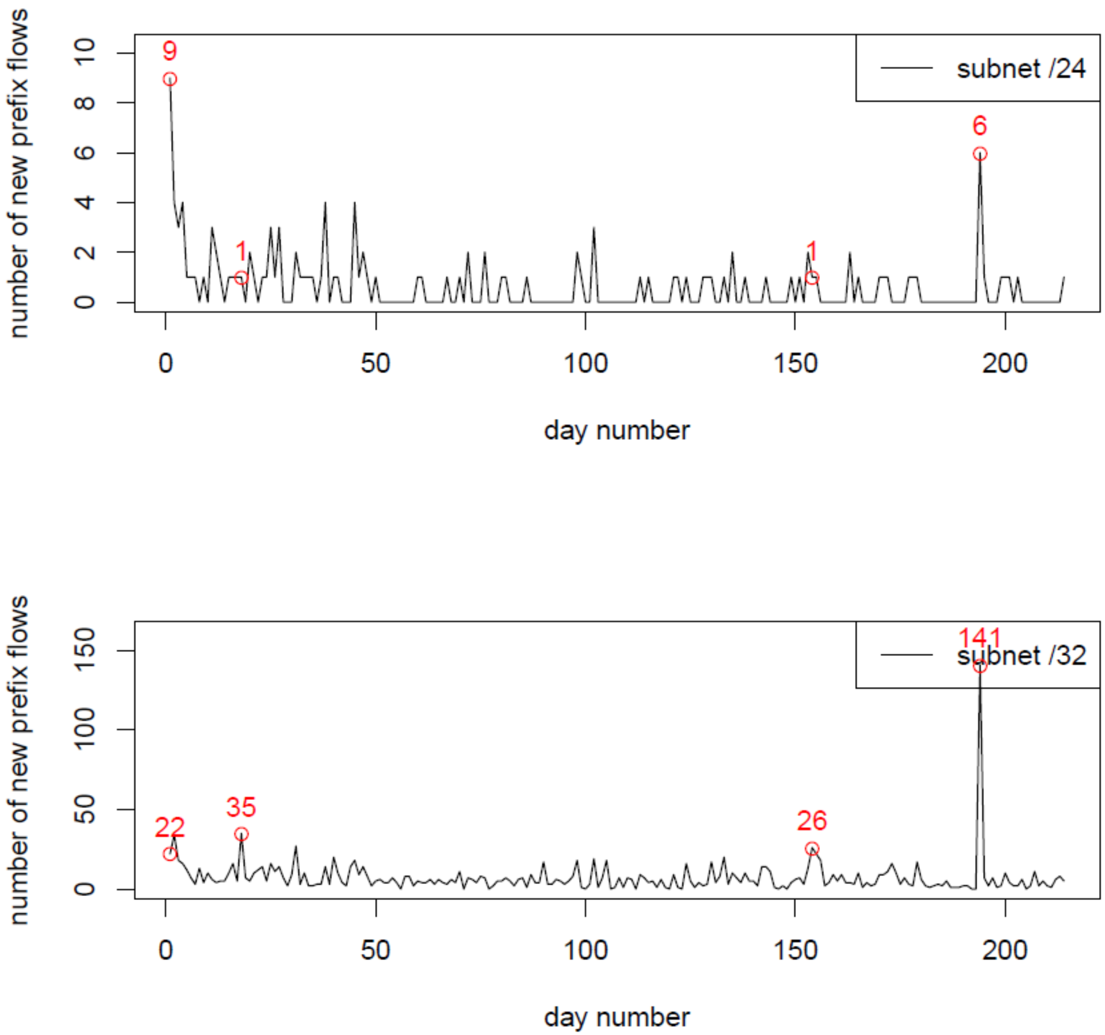
- 89.84% α-flows are less than 2 min, virtual circuit setup delay is 1 min
- 0.99% of the flows are longer than 10 minutes, but same ID for long and short flows (how then to predict)

Raw IP flow vs. prefix flow

- Port numbers are ephemeral for most high-speed file transfer applications, such as GridFTP
  - Answer to Q: Use prefix flow IDs
- Hypothesis:
  - Computing systems that run the high-speed file transfer applications don't change their IP addresses and/or subnet IDs often
  - Flows with previously unseen prefix flow identifiers will appear but such occurrences will be relatively rare

Number of new prefix flows daily

- When new collaborations start or new data transfer nodes are brought online, new prefix flows will occur



Effectiveness of offline design

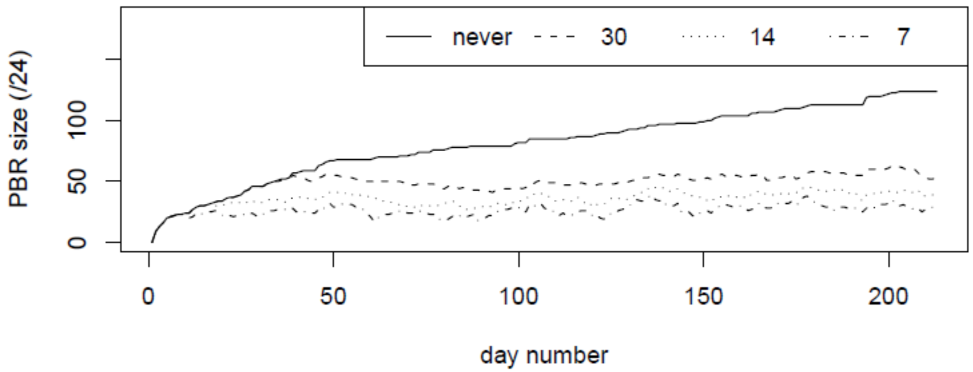
TABLE II: Number of days during which the percentages of α-bytes, and number of raw IP flows, that are not redirected to MPLS LSPs exceed different thresholds. The total number of days is 214.

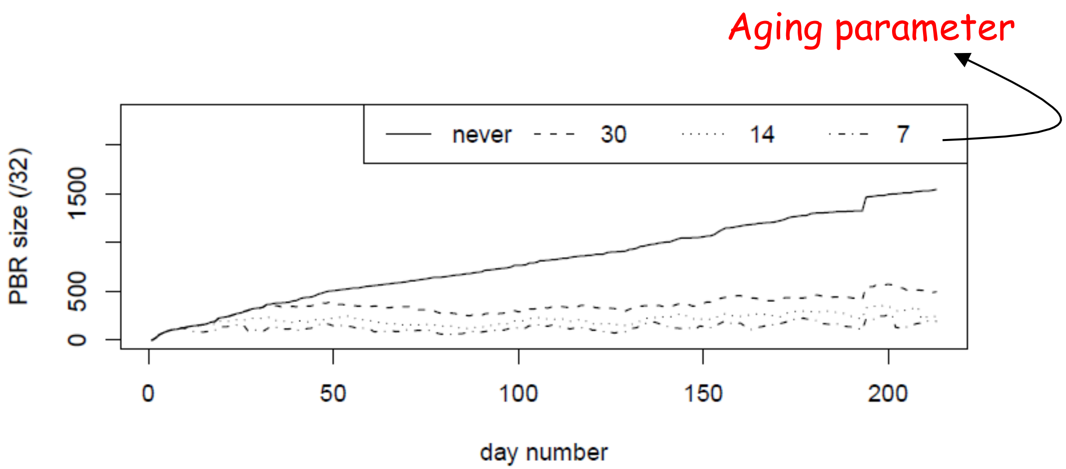
Aging parameter	Three measures	≥ 100%		≥ 75%		≥ 50%		≥ 25%	
		/24	/32	/24	/32	/24	/32	/24	/32
7 days	α-bytes	2	15	20	66	37	107	62	145
	Number of raw IP flows	2	15	17	76	32	117	64	161
14 days	α-bytes	2	8	12	47	27	86	49	125
	Number of raw IP flows	2	8	10	51	21	91	41	146
30 days	α-bytes	1	5	8	34	19	62	37	105
	Number of raw IP flows	1	5	6	35	14	67	32	126
∞ days	α-bytes	1	4	8	22	12	45	22	82
	Number of raw IP flows	1	4	6	23	10	48	16	101

- 94.4% of the days, at least 50% of the alpha bytes would have been redirected.
- For 89.7% of the days, 75% of the alpha bytes would have redirected (aging parameter = never; prefix identifier is /24)

Effect of aging parameter on PBR table size

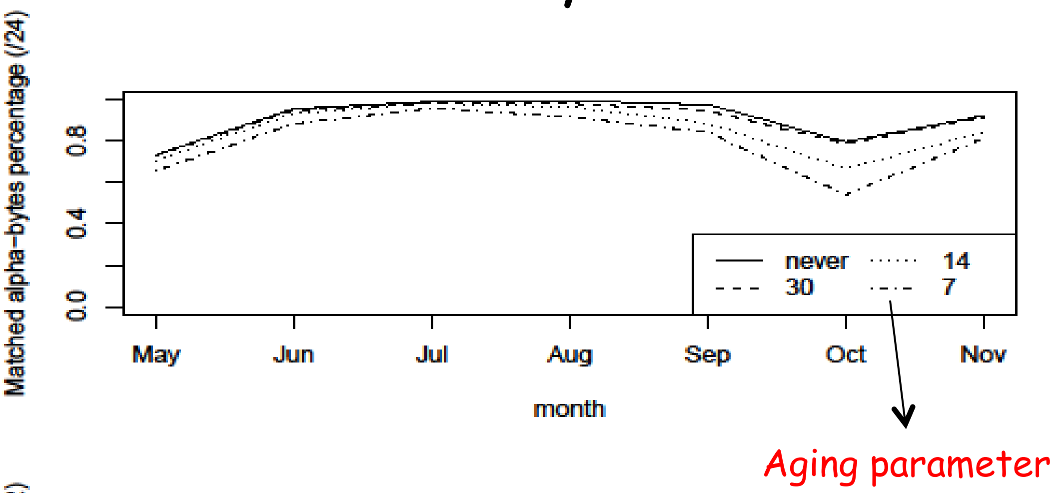
- For operational reasons, and forwarding latency, this table should be kept small

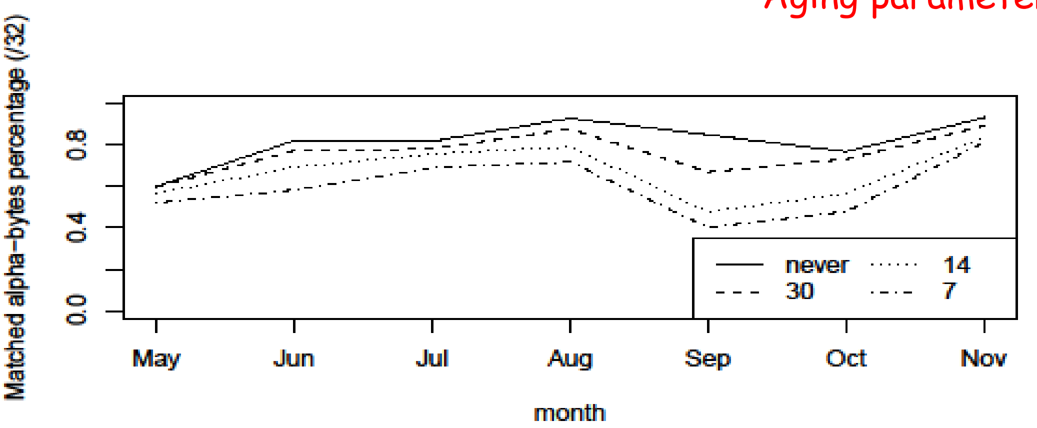




Matched a-bytes percentage

Monthly:





All 7 month:

Aging parameter	/24	/32
7	82%	67%
14	87%	73%
30	91%	82%
never	92%	86%

92% of the alpha bytes received over the 7-month period would have been redirected (aging parameter = never; prefix identifier is /24)

Key points for HNTES 2.0 design

- From current analysis:
  - Offline design appears to be feasible
    - IP addresses of sources that generate alpha flows relatively stable
    - Most alpha bytes would have been redirected in the analyzed data set
  - /24 seems better option than /32
- Aging parameter:
  - 30 days: tradeoff PBR size with effectiveness

Future NetFlow data analyses

- other routers' NetFlow data
- redirected beta flow bytes experience competition with alpha flows (/24)
- utilization of MPLS LSPs
- multiple simultaneous alpha flows on same LSPs
- match with known data doors

Discussion

- To determine cause of throughput variance
  - Feedback?
  - Need your support to obtain data
- Would trouble ticket log mining be useful to help answer "why isolate science flows"?
- Automatic flow identification and redirection appears feasible
  - How do you feel about this?