# Hybrid network traffic engineering system (HNTES) Project 1

Zhenzhen Yan, Zhengyang Liu, Chris Tracy, Malathi Veeraraghavan
University of Virginia and ESnet
Jan 12-13, 2012

Acknowledgment: Thanks to the US DOE ASCR program office for UVA grants DE-SC002350 and DE-SC0007341 and ESnet grant DE-AC02-05CH11231

# Outline

- Problem statement (What?)
- Solution approach for designing HNTES (How?)
  - Formulate questions
  - Test hypotheses through
    - Analyses of ESnet NetFlow data
    - Analyses of GridFTP logs
- Why? (work from Oct. 2010-Jan. 2012)
- HNTES Project 1 planned work: Jan.-Aug. 2012



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

# Problem statement

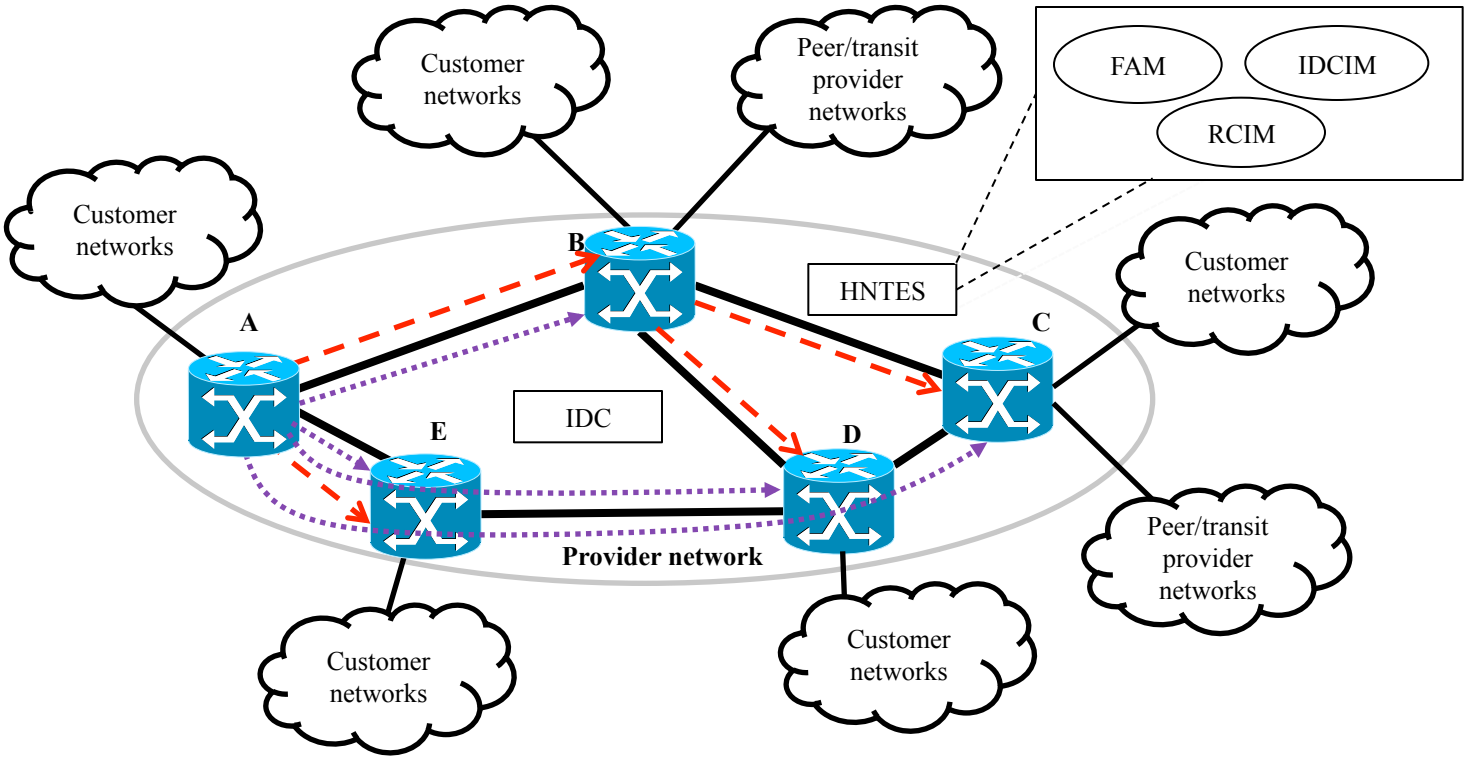
- Hybrid network is one that supports both IP-routed and circuit services on:
  - Separate networks as in ESnet4, or
  - An integrated network
- A hybrid network traffic engineering system (HNTES) is one that moves data flows between these two services as needed
  - engineers the traffic to use the service type appropriate to the traffic type

# Two reasons for using circuits

1. Offer scientists rate-guaranteed connectivity
   - necessary for low-latency/low-jitter applications such as remote instrument control
   - provides low-variance throughput for file transfers
2. Isolate science flows from general-purpose flows

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

# Role of HNTES (what is HNTES?)




IP router/ MPLS LSR

IP-routed paths

MPLS LSPs

IDC: Inter-Domain Controller

HNTES: Hybrid Network Traffic Engineering System

FAM: Flow Analysis Module

IDCIM: IDC Interface Module

RCIM: Router Control Interface Module

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

# Three tasks executed by HNTES

1. Heavy-hitter flow identification
   - Offline flow analysis
   - Online flow analysis
   - End-host assisted

online:
upon flow arrival

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

# Outline

- Problem statement (What?)
- ➢ Solution approach for designing HNTES (How?)
  - Formulate questions
  - Test hypotheses through
    - Analyses of ESnet NetFlow data
    - Analyses of GridFTP logs
- Why? (work from Oct. 2010-Jan. 2012)
- HNTES Project 1 planned work: Jan.-Aug. 2012



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

# Questions for HNTES design

- Is a Flow monitoring module(FMM) that can capture all packets necessary, or is NetFlow data sufficient (given 1-in-1000 sampling)?
- Should circuit setup and PBR config. be online or offline?
- If offline, should PBRs be set for raw IP flow identifiers or prefix flow identifiers?
- But do IP addresses of nodes that create alpha flows stay unchanged? /24 or /32?
- Should prefix flow IDs added to PBR table be aged out (parameter A days)?

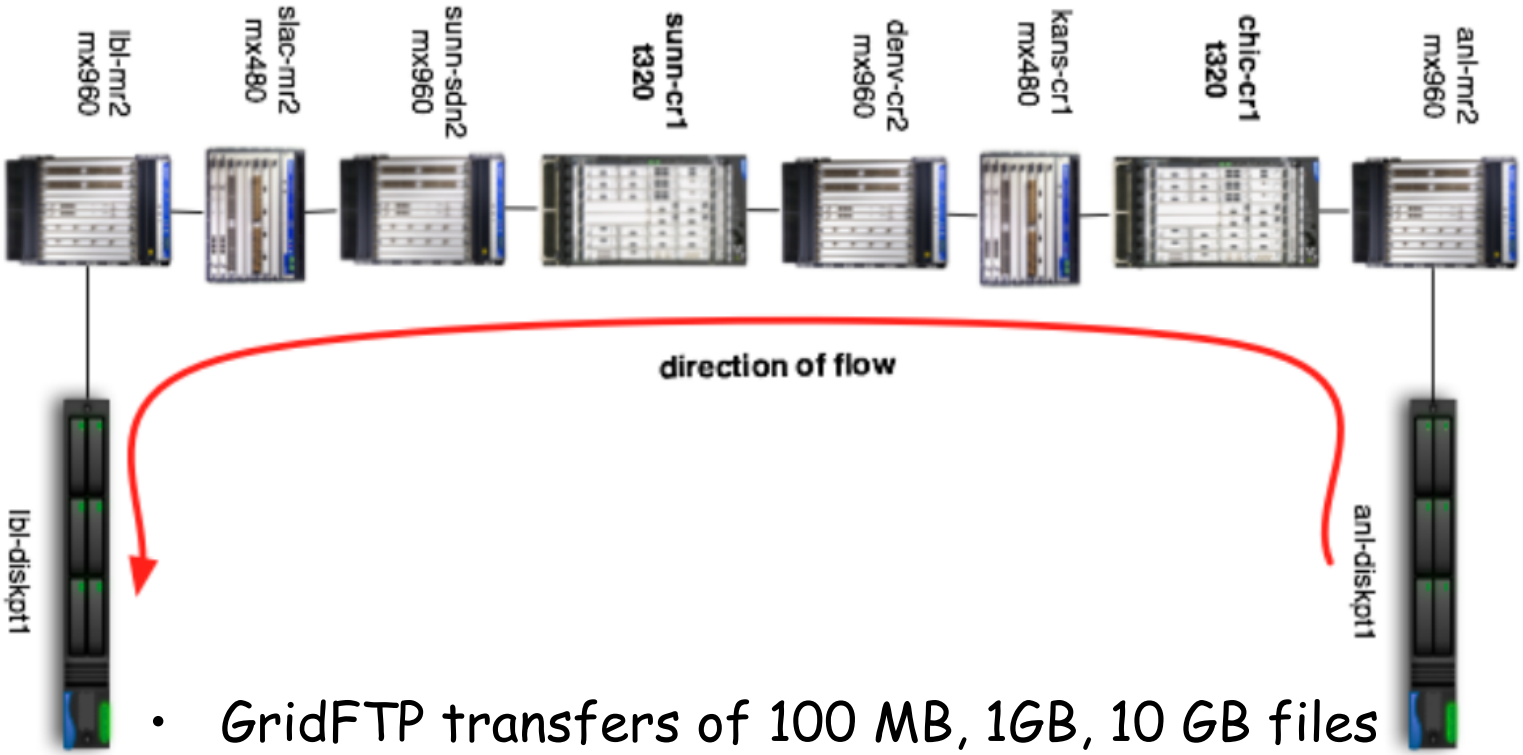
# Flow identification

- Flow monitoring module (FMM) or NetFlow?
  - FMM: challenging at high rates
  - ➡ NetFlow: 1/1000 sampling

# Validation of size estimation from NetFlow data

- Hypothesis
  - Flow size from concatenated Netflow records for one flow can be multiplied by 1000 (since the ESnet Netflow sampling rate is 1 in 1000 packets) to estimate actual flow size

# Experimental setup




- GridFTP transfers of 100 MB, 1GB, 10 GB files
- sunn-cr1 and chic-cr1 Netflow data used

Chris Tracy set up this experiment

# Flow size estimation experiments

- Workflow inner loop (executed 30 times):
  - obtain initial value of firewall counters at sunn-cr1 and chic-cr1 routers
  - start GridFTP transfer of a file of known size
  - from GridFTP logs, determine data connection TCP port numbers
  - read firewall counters at the end of the transfer
  - wait 300 seconds for Netflow data to be exported
- Repeat experiment 400 times for 100MB, 1 GB and 10 GB file sizes

Chris Tracy ran the experiments

# Create log files

- Filter out GridFTP flows from Netflow data
- For each transfer, find packet counts and byte counts from all the flow records and add
- Multiply by 1000 (1-in-1000 sampling rate)
- Output the byte and packet counts from the firewall counters
- Size-accuracy ratio = Size computed from Netflow data divided by size computed from firewall counters

Chris Tracy wrote scripts to create these log files and sent UVA these files for analysis

# Size-accuracy ratio

| | Netflow records obtained from Sunnyvale ESnet router | | Netflow records obtained from Chicago ESent router | |
|---|---|---|---|---|
| | Mean | Standard deviation | Mean | Standard deviation |
| 100 MB | 0.949 | 0.2780 | 1.0812 | 0.3073 |
| 1 GB | 0.996 | 0.1708 | 1.032 | 0.1653 |
| 10 GB | 0.990 | 0.0368 | 0.999 | 0.0252 |

- Sample mean shows a size-accuracy ratio close to 1
- Standard deviation is smaller for larger files.
- Dependence on traffic load
- Sample size = 50

# Answer to Question 1

- Is a Flow monitoring module(FMM) that can capture all packets necessary, or is NetFlow data sufficient (given 1-in-1000 sampling)?
  - GridFTP flows were both elephants (large size) and alpha (high rate) flows
  - Experiment conclusion: NetFlow data is sufficient
  - No FMM in HNTES 2.0

# Questions for HNTES design

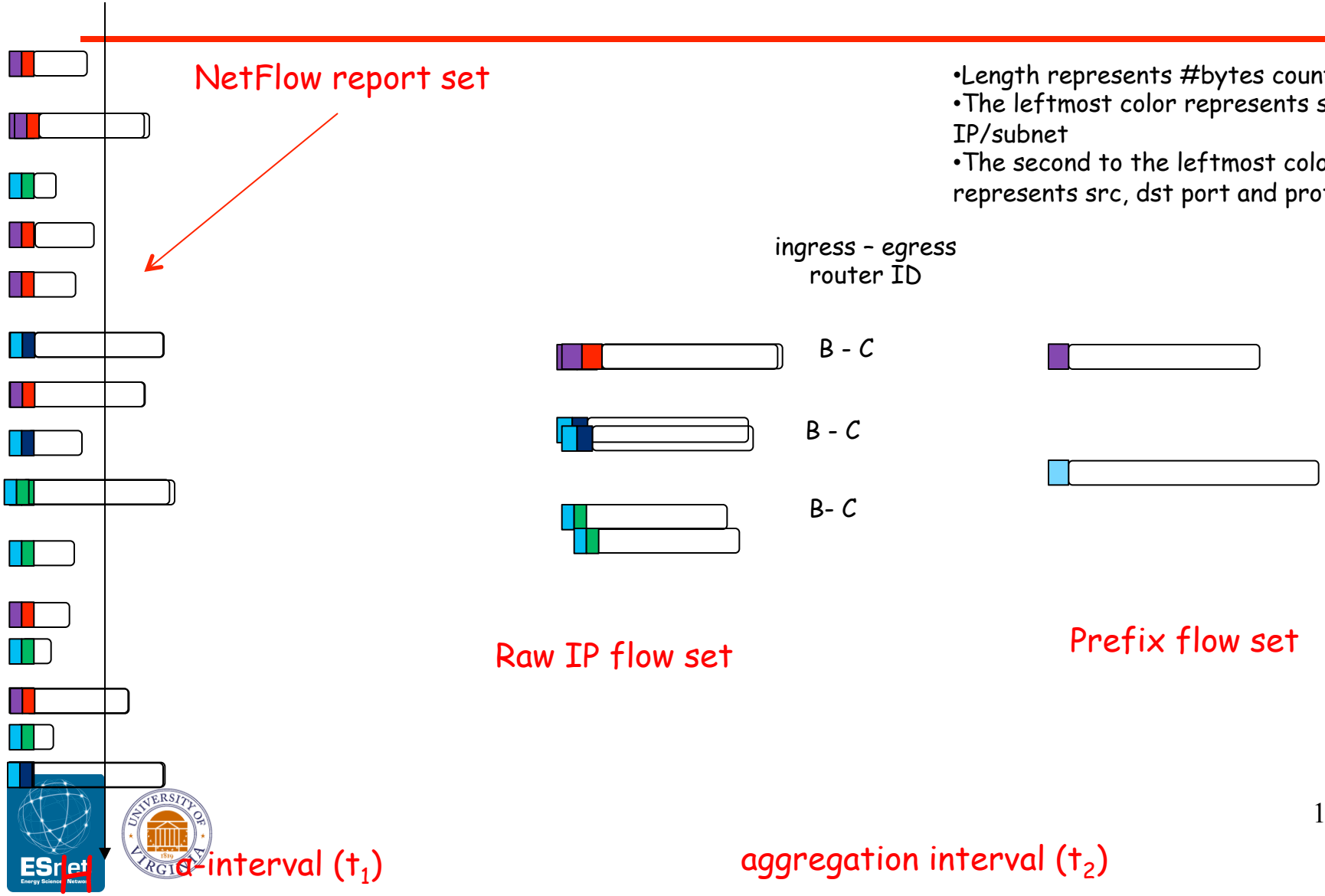
- Is a Flow monitoring module(FMM) that can capture all packets necessary, or is NetFlow data sufficient (given 1-in-1000 sampling)?
- ➢ Should circuit setup and PBR config. be online or offline?
- If offline, should PBRs be set for raw IP flow identifiers or prefix flow identifiers?
- But do IP addresses of nodes that create alpha flows stay unchanged? /24 or /32?
- Should prefix flow IDs added to PBR table be aged out (parameter A days)?

# Offline 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
- Details on why alpha flows is explained in next talk

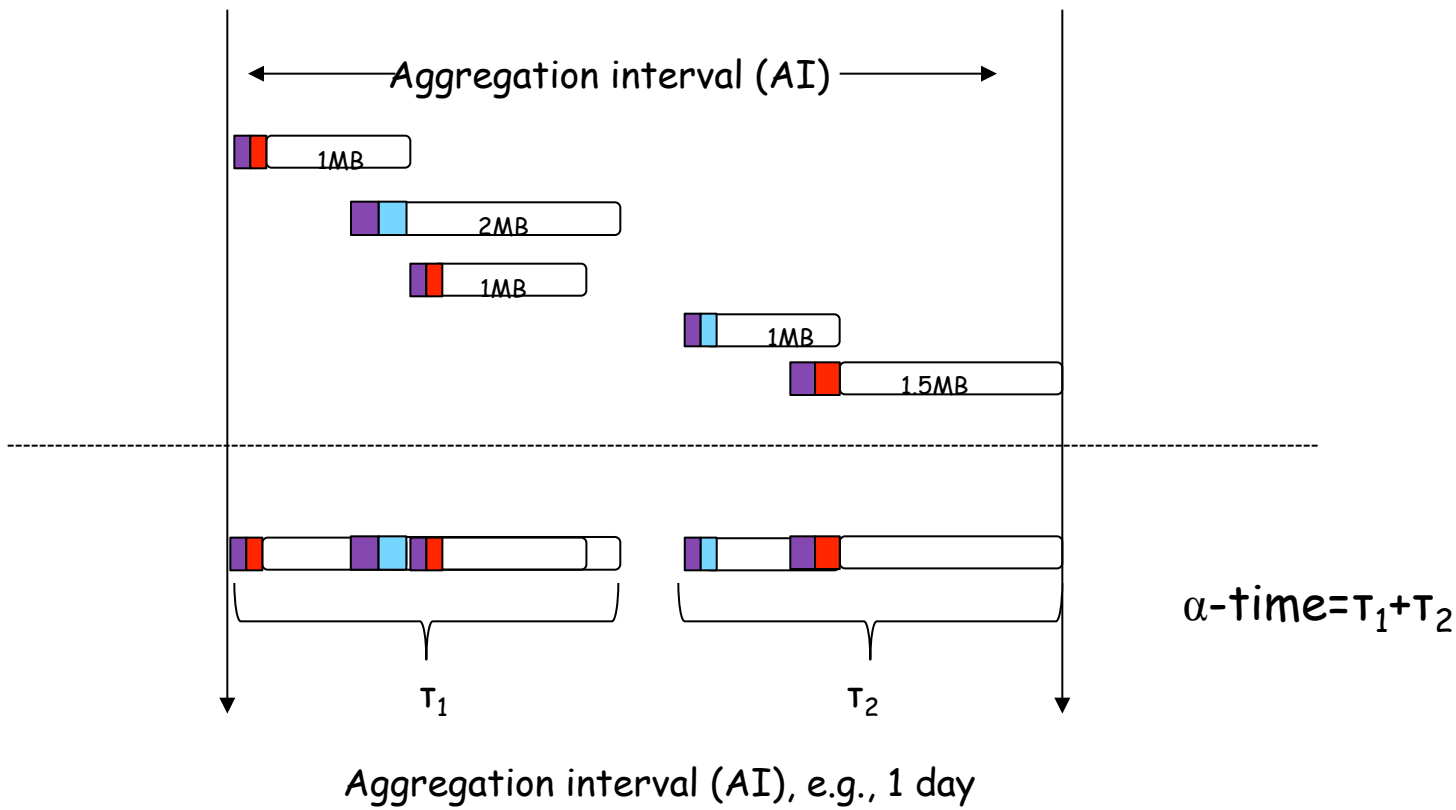
S. Sarvotham, R. Riedi, and R. Baraniuk, "Connection-level analysis and modeling of nework traffic," in ACM SIGCOMM Internet Measurement Workshop 2001, November 2001, pp. 99–104.

# Flow aggregation from NetFlow



# Terminology

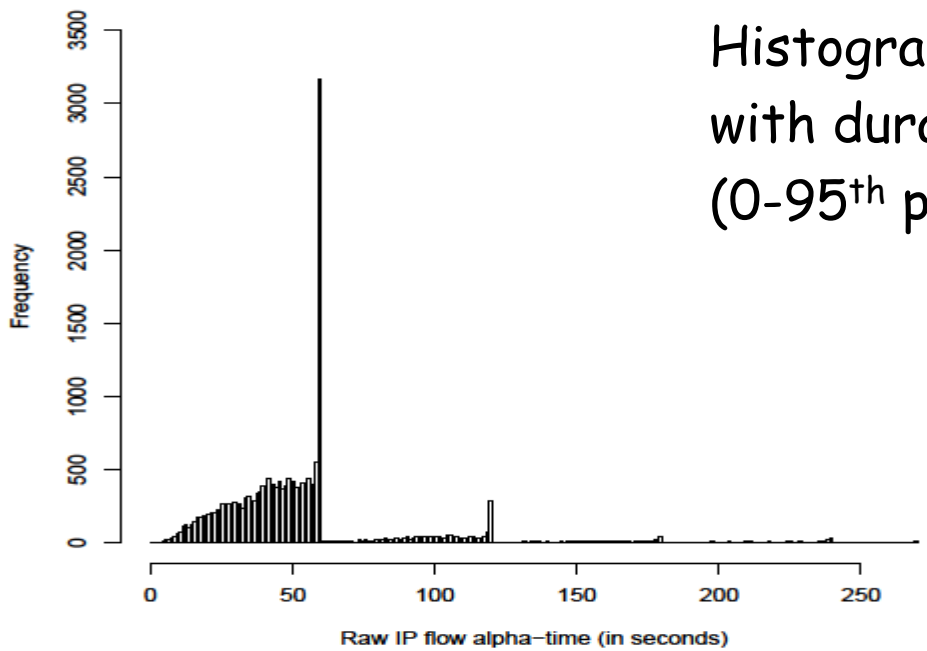
- α-bytes: 1MB + 2MB + 1MB + 1MB + 1.5MB = 6.5MB (*1000)
- α-time:



# Dataset

- NetFlow data over 7 months (May-Nov 2011) collected at ESnet site PE router
- Threshold (H) for a-flow report is 1GByte/min = 133Mbps
- 22041 raw IP flows, 125 (/24) prefix flows, and 1548 (/32) prefix flows

# Online vs. offline




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

- 89.84% a-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)

# Answer to question 2

- Should circuit setup and PBR config. be online or offline?
  - Answer: online solution does not seem feasible unless VC setup delay is reduced

# Questions for HNTES design

- Is a Flow monitoring module(FMM) that can capture all packets necessary, or is NetFlow data sufficient (given 1-in-1000 sampling)?
- Should circuit setup and PBR config. be online or offline?
- ➢ If offline, should PBRs be set for raw IP flow identifiers or prefix flow identifiers?
- But do IP addresses of nodes that create alpha flows stay unchanged? /24 or /32?
- Should prefix flow IDs added to PBR table be aged out (parameter A days)?

# Raw IP flow vs. prefix flow

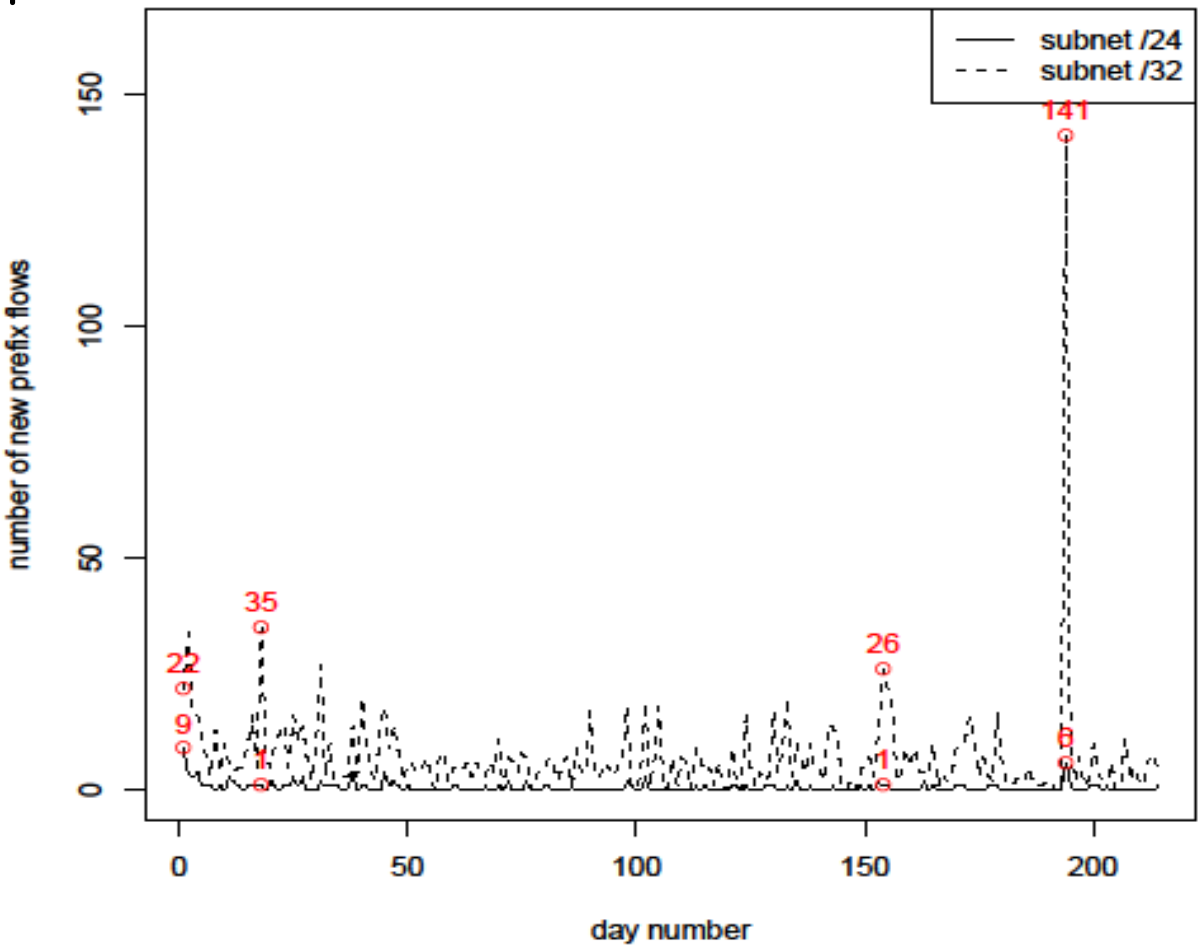
- Port numbers are ephemeral for most high-speed file transfer applications, such as GridFTP
  - Answer to Q: Use prefix flow IDs
- Hypothesis:
  - Computing systems that run the high-speed file transfer applications don't change their IP addresses and/or subnet IDs often
  - Flows with previously unseen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

# Questions for HNTES design

- Is a Flow monitoring module(FMM) that can capture all packets necessary, or is NetFlow data sufficient (given 1-in-1000 sampling)?
- Should circuit setup and PBR config. be online or offline?
- If offline, should PBRs be set for raw IP flow identifiers or prefix flow identifiers?
- ➢ But do IP addresses of nodes that create alpha flows stay unchanged? /24 or /32?
- ➢ Should prefix flow IDs added to PBR table be aged out (parameter A days)?

# Number of new prefix flows daily

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



# Effectiveness of offline design

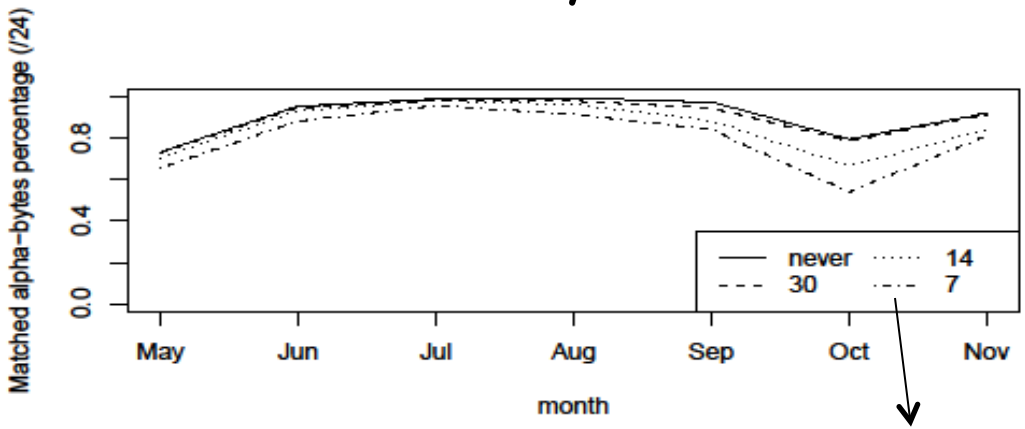
TABLE II: Number of days during which the percentages of $\alpha$-bytes, and number of raw IP flows, that are not redirected to MPLS LSPs exceed different thresholds. The total number of days is 214.

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

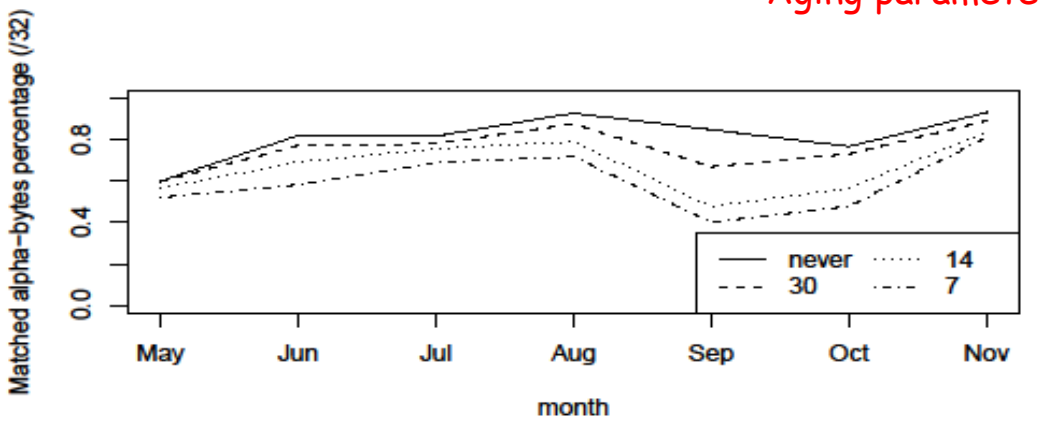
- 94.4% of the days, at least 50% of the alpha bytes would have been redirected.
- For 89.7% of the days, 75% of the alpha bytes would have redirected (aging parameter = never; prefix identifier is /24)

# Matched a-bytes percentage

Monthly:




Aging parameter



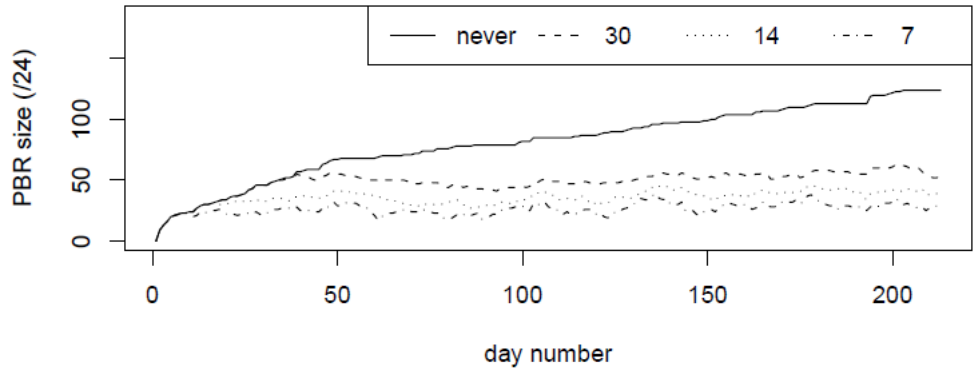

All 7 month:

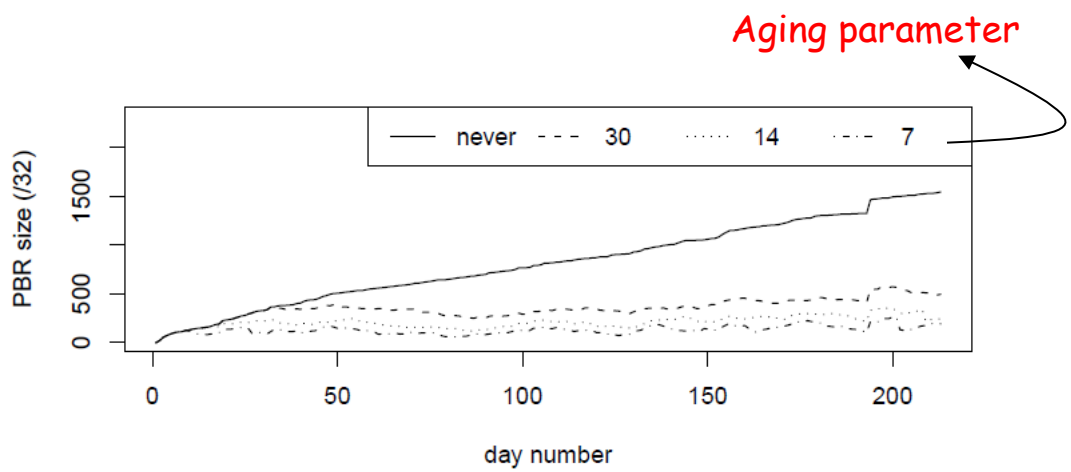
| Aging parameter | /24 | /32 |
|---|---|---|
| 7 | 82% | 67% |
| 14 | 87% | 73% |
| 30 | 91% | 82% |
| never | 92% | 86% |

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

# Effect of aging parameter on PBR table size

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



# Full mesh of LSPs required or just a few?

Number of super-prefix flows per month:

| Month | May | Jun | July | Aug | Sep | Oct | Nov |
|---|---|---|---|---|---|---|---|
| total | 13 | 15 | 16 | 16 | 18 | 18 | 18 |
| repeated | 0 | 13 | 15 | 16 | 16 | 18 | 18 |
| new | 13 | 2 | 1 | 0 | 2 | 0 | 0 |

Represents number of LSPs needed from ESnet site PE router to indicated numbers of egress routers

# Outline

- Problem statement (What?)
- Solution approach for designing HNTES (How?)
  - Formulate questions
  - Test hypotheses through
    - Analyses of ESnet NetFlow data
    - Analyses of GridFTP logs
- Why? (work from Oct. 2010-Jan. 2012)
- HNTES Project 1 planned work: Jan.-Aug. 2012



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

# GridFTP data analysis findings

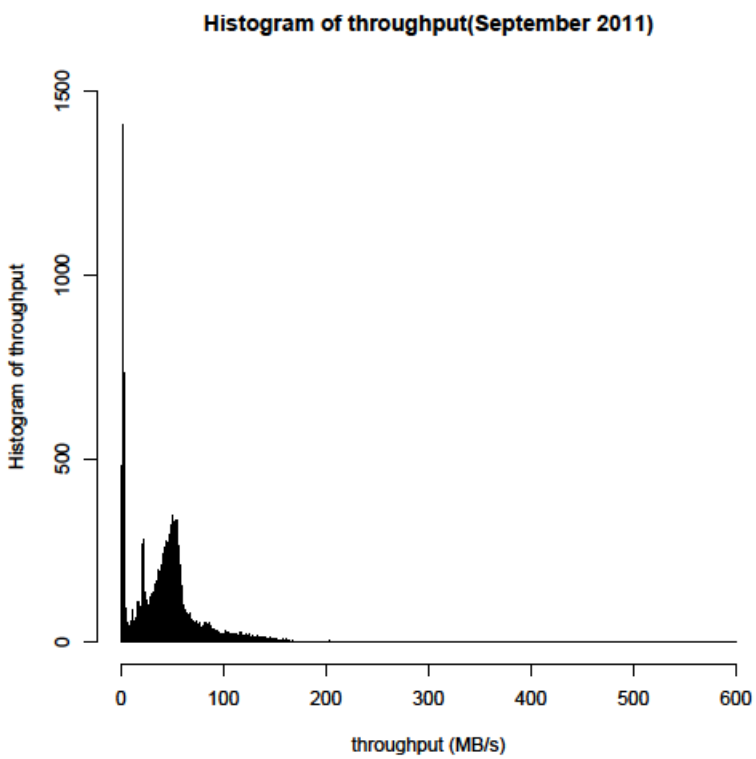
- All GridFTP transfers from NERSC GridFTP servers that > 100 MB: one month (Sept. 2010)
- Total number of transfers: 124236
- GridFTP usage statistics

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

# Throughput of GridFTP transfers




- Total number of transfers: 124236
- Most transfers get about 50 MB/sec or 400 Mb/s

# 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 > 5 mins

# Transfers longer than 5 mins NERSC (100MB dataset)

| | Min | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| Duration (sec) | 600.1 | 683.7 | 793.1 | 1167 | 1156 | 9952 |

- Number: 328 (0.0026% of total number of transfers)
- 50% of this set had a throughput< 11 Mbps
- 75% had a throughput < 17.05 Mbps
- 95% had a throughput < 34.5 Mbps
- 4 transfers had a duration > 4000 sec (incl. 9952sec max duration transfer)
  - Three had throughput of ~ 2 Mbps
  - One had throughput of 30.3 Mbps (size: 18 GB)

# Key points for HNTES 2.0 design

- From current analysis:
  - Online infeasible with current VC setup delay
  - 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
- Aging parameter:
  - 30 days: tradeoff PBR size with effectiveness
  - /24 better than /32 (negatives?)

# Outline

- Problem statement (What?)
- Solution approach for designing HNTES (How?)
  - Formulate questions
  - Test hypotheses through
    - Analyses of ESnet NetFlow data
    - Analyses of GridFTP logs
- ➢ Why? (work from Oct. 2010-Jan. 2012)
- HNTES Project 1 planned work: Jan.-Aug. 2012



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

# Why move science flows?

- Quantify negative impact of science flows on general-purpose flows
  - Simulations
  - OWAMP raw data analysis
  - SNMP data analysis
- Oct. 2010-Jan. 2012
  - Fairness issue studied in simulations
  - OWAMP analysis: surges in delay characterized for raw I2 measurements, but not explained
  - SNMP raw data downloaded and GridFTP flow correlations found

# HNTES project 1 planned work

- Jan. 2012 – Aug. 2012:
  - Complete NetFlow data analysis
  - Answer "why move" question
    - Simulation study
    - OWAMP and SNMP analyses
  - ANI testbed experimentation
    - Rate-unlimited MPLS LSPs (3rd queue)
    - NetFlow sufficiency under different conditions
  - HNTES 2.0 software prototype