NETWORK TOMOGRAPHY BASED ON FLOW LEVEL MEASUREMENTS

Dogu Arifler, Gustavo de Veciana, and Brian L. Evans

Department of Electrical and Computer Engineering
The University of Texas at Austin, Austin, TX 78712-0240 USA
{arifler, gustavo, bevans}@ece.utexas.edu

ABSTRACT

Internet traffic primarily consists of packets from elastic flows, i.e. Web transfers, file transfers (FTP), and e-mail, whose transfers are mediated via the Transmission Control Protocol. We develop a conditional sampling technique to analyze throughput correlations among elastic flow classes based on flow level measurements from current network traffic monitoring tools. The primary contributions of this paper are: (1) a demonstration of throughput correlation among temporally overlapping flows on congested resources by using analytical/simulation models, and (2) application of a multivariate statistical method (principal components) to infer network properties, such as the number of shared resources by flows in the network from non-intrusive, flow level measurements collected at a single site. Our proposal for using flow level measurements to infer network properties differs significantly from previous network tomography research that has employed end-to-end packet level measurements for making inferences.

1. INTRODUCTION

A commonly accepted definition of an Internet (IP) *flow* is a unidirectional sequence of packets between a source and a destination endpoint identified by common IP addresses, Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) port numbers, IP protocol type, type of service fields in IP headers, etc. An IP *flow class* is a collection, or aggregation, of flows having a common attribute. For example, we can refer to all flows sharing common source and destination IP address prefixes as a flow class. State-of-the-art network monitoring tools (such as Cisco's NetFlow) are capable of generating *flow records*. A flow record contains the source and destination IP addresses, source and destination port numbers, start and end times, and the size (in bytes and packets) of flows traversing that network element.

A significant portion of the IP traffic consists of packets from *elastic flows*, i.e. Web transfers, file transfers (FTP), and e-mail, whose transfers are mediated via TCP. TCP uses packet delay and loss as indicators of the available bandwidth to adjust the data transmission window at the sender. An understanding of the interactions and dependencies among elastic flow classes in the network may be critical in designing and provisioning networks. For example, on determining that two flow classes carrying Web content destined for two different customer bases are experiencing poor performance due to a bottleneck link serving them, the Web content provider might choose to replicate content at a second location to reduce the load on the bottleneck link. From a customer's perspective, on the other hand, determining whether a network provider uses a diverse set of routes (an indication of robustness) when carrying different flow classes of the customer may be valuable, especially since network providers are unwilling to disclose their backbone topology. In this case, lack of interdependence among traffic classes might indicate such a route diversity.

Our premise is that elastic flow classes that are temporally overlapping long enough on the same path, or bottleneck, will tend to have correlated throughputs. While our premise is intuitive, the extent of such correlations needs to be quantified, especially when flow classes visiting multiple resources can introduce throughput correlations among flow classes that do not necessarily share paths or bottlenecks.

In our work, we evaluate the extent of throughput correlations on analytical/simulation models for dynamic bandwidth sharing mechanisms [1], [2] that approximate TCP. This evaluation supports our premise that flow classes that temporally overlap on congested resources will have correlated throughputs. Finally, we propose to use a statistical methodology for analyzing the structure of the conditional throughput correlation matrix. The methodology can be used to identify the underlying causes for the variability in throughput observations. We argue that the variability in data can naturally be attributed to shared paths or bottlenecks. This key finding will be used in our future studies to infer which TCP flow classes share paths or bottlenecks. In contrast to previous network tomography research that makes inferences based on packet level characteristics such as packet loss and packet delay (e.g., [3], [4], [5], [6], and [7]), we employ a flow level, or user-perceived, performance measure (throughput), which is directly available from the state-of-the-art network monitoring tools, in order to infer network properties.

2. SIMULATION MODELS

The collection of elastic flows in the network is denoted by a set $\mathcal{F}$. The size, start time (time of arrival of the first packet in a flow), end time (time of arrival of the last packet in a flow), and duration of a flow will be denoted by v_f, s_f, e_f, and $d_f = e_f - s_f$ respectively. Each flow $f \in \mathcal{F}$ belongs to a flow class $c \in \mathcal{C}$. The function $\phi : \mathcal{F} \to \mathcal{C}$ determines the class of a particular flow. We let $\mathcal{F}_c(t) = \{f \in \mathcal{F} : \phi(f) = c \text{ and } s_f \leq t < e_f\}$ denote the set of flows that belong to class c and are *active* at time t.

We use known analytical models to generate flow records (as would be available from commonly used flow level measurement technologies at a monitoring site) by simulation. We use fluid

D. Arifler and B. L. Evans were supported by The State of Texas Advanced Technology Program under project 003658-0614-2001.

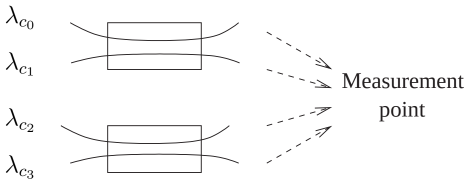

Fig. 1. Two parallel M/GI/1 processor sharing queues. Arrival rates of flows for each class are shown. We assume that all other infrastructure that flows visit is overprovisioned.

models to determine the bandwidth shares [1], [2] achieved by flows at a given time. In such models, the bandwidth allocated to a flow is adjusted *instantaneously* when the number of flows in the system changes as a result of flow arrivals and departures. The dynamic bandwidth sharing model approximates actual rate control mechanisms (such as TCP) well due to the assumption of separation of times scales: the time scale of flow durations is much longer than the time scale on which rate control mechanisms converge to equilibrium. We consider bandwidth sharing among flows first on a single link and then in a "linear" network.

The simplest model is one in which the number of flows on a single link can be modelled as an M/GI/1 (a single-server queueing system with an exponential interarrival time distribution and a general, independent service time distribution) processor sharing queue [8]. That is, if all flows share similar round trip times (RTT) and packet loss rates, the link bandwidth is shared equally among the active flows. A parallel collection of queues as shown in Fig. 1 constitutes one of our simplest test cases to investigate throughput correlations. Fig. 2 illustrates a linear network model with two links. Such a network can be used to model flows traversing several links that interact with the cross traffic on these links. The linear network model enables us to investigate the coupling effects between flows following multi-link paths (e.g., c_0) and cross traffic (e.g., c_1 and c_2) and possibly between flow classes not sharing a link (e.g., c_1 and c_2). To determine the character of bandwidth sharing among flows on a linear network, we consider proportionally fair sharing [9]. In each case, we assume that the flows in class c arrive according to a Poisson process with rate λ_c. To study the effect of flow size distribution on throughput correlations, we consider two distributions from which flow sizes are drawn independently: (1) an exponential (exp) distribution with mean $1/\mu = 1$, i.e., a distribution with pdf $f(x|\mu) = \mu \exp(-\mu x)$, and (2) a bounded Pareto (BP) distribution [10] whose pdf is given by

$$f(x|\alpha, k, q) = \frac{\alpha k^{\alpha}}{1-(k/q)^{\alpha}} x^{-\alpha-1}, \qquad k \le x \le q. \tag{1}$$

The bounded Pareto distribution that is used has an exponent of power law $\alpha = 1.3$ with minimum size $k = 0.242$ and maximum size $q = 10,000$. The particular bounded Pareto distribution has a mean equal to 1 as well. Flow sizes selected from the bounded Pareto distribution consist of a very large number of short flows and a few very long flows (as may be the case in the current Internet).

We denote the bandwidth share of a flow f at time t by $b^f(t)$. Then, the *perceived throughput* r_f for a flow f is given by

$$r_f = \frac{1}{d_f} \int_{s_f}^{e_f} b^f(t) dt,$$

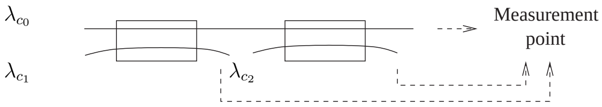

Fig. 2. A linear network. Arrival rates of flows for each class are shown. We assume that all other infrastructure that flows visit is overprovisioned.

or $r_f = v_f/d_f$. The average throughput (over the number of active flows) of a flow class $c \in \mathcal{C}$ at the measurement point (see Figs. 1 and 2) is

$$r_c(t) = \begin{cases} \frac{1}{|\mathcal{F}_c(t)|} \sum_{f \in \mathcal{F}_c(t)} r_f, & \text{if } |\mathcal{F}_c(t)| > 0, \\ 0, & \text{otherwise.} \end{cases}$$

3. CORRELATION AMONG FLOW CLASS THROUGHPUTS

We compute correlations among flow class throughputs by using *temporal throughput observations* at times when *all* of the flow classes are active. Note that the requirement of this *conditional sampling* strategy is a stringent one, especially when the offered load of a flow class under consideration is low. However, we choose to impose our stringent condition to guarantee positive definiteness of the throughput correlation matrix that is used in this section.

We first divide the observation time into discrete intervals. We denote the number of discretized time intervals for a measurement period by T and the number of discretized intervals over which all flow classes are active by $N(T)$. We assume that the throughput of a flow at a discretized time interval is equal to its "continuous-time" throughput if the flow is active anytime during that interval. We also assume that $r_{c_i}(n)$ and $r_{c_j}(n)$ are realizations of ergodic random processes of throughputs of flow classes c_i and c_j respectively (on discretized intervals). The *conditional* mean and variance of throughput for flow class c_i are defined as

$$m_{c_i} = \lim_{T\to\infty} \frac{1}{N(T)} \sum_{n=0}^{T} r_{c_i}(n)\, \mathbf{1}_{\{r_{c_i}(n)>0, \forall c_i \in \mathcal{C}\}},$$

$$\sigma_{c_i}^2 = \lim_{T\to\infty} \frac{1}{N(T)} \sum_{n=0}^{T} (r_{c_i}(n) - m_{c_i})^2\, \mathbf{1}_{\{r_{c_i}(n)>0, \forall c_i \in \mathcal{C}\}},$$

where $\mathbf{1}_E$ is the standard indicator function, which is equal to 1 if E is true and 0, otherwise. The conditional correlation of throughputs of flow classes c_i and c_j is defined as

$$\rho_{c_i c_j} = \lim_{T\to\infty} \sum_{n=0}^{T} \frac{(r_{c_i}(n) - m_{c_i})(r_{c_j}(n) - m_{c_j})\, \mathbf{1}_{\{r_{c_i}(n)>0, \forall c_i \in \mathcal{C}\}}}{N(T)\sigma_{c_i}\sigma_{c_j}}.$$

For p classes, we can obtain a *conditional correlation matrix* $\boldsymbol{\rho} = [\rho_{c_i c_j}]$, where $\rho_{c_i c_i} = 1$, and $i, j = 0, \ldots, p-1$.

We next consider a number of scenarios based on topologies described in Section 2. For each scenario, correlations are computed based on five (long enough) simulation runs with different

Table 1. Class throughput correlation matrix. The flows in class c arrive according to a Poisson process with rate λ_c with sizes drawn independently from an exponential (exp) and a bounded Pareto (BP) distribution in (1) with mean 1. $\lambda_{c_0} = 0.4, \lambda_{c_1} = 0.4, \lambda_{c_2} = 0.4, \lambda_{c_3} = 0.4$.

exp	Class 0	Class 1	Class 2	Class 3
Class 0	1	0.899	0.104	0.104
Class 1	0.899	1	0.106	0.106
Class 2	0.104	0.106	1	0.899
Class 3	0.104	0.106	0.899	1
BP	Class 0	Class 1	Class 2	Class 3
Class 0	1	0.860	0.051	0.061
Class 1	0.860	1	0.055	0.065
Class 2	0.051	0.055	1	0.846
Class 3	0.061	0.065	0.846	1

random seeds to obtain independent correlation values. The conditional correlation matrix is estimated by taking averages of correlations over five simulation runs. We consider flow sizes with an exponential distribution (exp) and bounded Pareto distribution (BP), and obtain two correlation matrices for each scenario.

In the case of two parallel M/GI/1 processor sharing queues (in Fig. 1), the correlations among throughputs of flow classes, each offering a load of 0.4, are shown in Table 1. The throughput correlation between flow classes sharing resources (M/GI/1 processor sharing queues) is large, while the correlation between flow classes not sharing resources is small or negligible.

The correlations among flow class throughputs in the linear network (shown in Fig. 2) that uses proportionally fair sharing are tabulated in Tables 2 – 4 for classes offering different loads. The throughputs of c_1 and c_2 may exhibit some degree of correlation via c_0, even though c_1 and c_2 are not sharing a link. We can make three observations from these correlation matrices. First, the extent of correlation between the throughputs of classes sharing a link is higher for more congested links. Second, the effect of flow class 0 offering a higher load relative to the cross traffic on the links it visited is to introduce higher degree of cross-coupling between c_1 and c_2 (see Table 4). Third, when flow class 0 visited two bottlenecks the degree of cross-coupling between c_1 and c_2 through c_0 was small (see Table 3). We also observed that the flow size distribution does not significantly affect the nature of correlation among flow class throughputs.

Based on these experiments, we conclude that the flow classes that temporally overlap on congested resources have correlated throughputs. The coupling effect of flows traversing several links that interact with the cross traffic on these links is not significant unless the offered load of such flow classes is higher than loads offered by cross traffic.

4. REDUCING THE DIMENSIONALITY OF DATA

In a realistic scenario, the number of flow classes p under consideration is often large, which results in correlation matrices that are hard to interpret. When analyzing multivariate observations, it is often possible to reduce the number of variables that account for the variability in data. A common practice in exploratory studies is to apply spectral decomposition to the correlation matrix:

$$\boldsymbol{\rho} = e_0\boldsymbol{\xi}_0\boldsymbol{\xi}_0^T + e_1\boldsymbol{\xi}_1\boldsymbol{\xi}_1^T + \ldots + e_{p-1}\boldsymbol{\xi}_{p-1}\boldsymbol{\xi}_{p-1}^T,$$

where $(e_i, \boldsymbol{\xi}_i)$ are the eigenvalue-eigenvector pairs such that $e_0 \geq e_1 \geq \ldots \geq e_{p-1} \geq 0$ since $\boldsymbol{\rho}$ is positive definite. In statistical signal processing theory, the eigenvectors are known as the *principal components* whose corresponding eigenvalues are variances of these components. The percentage of the total normalized variance, tr($\boldsymbol{\rho}$)=p, explained by the first, say $m\,(< p)$, variances is given by $\sum_{i=0}^{i=m-1} e_i/p$. We say that the percentage of normalized variance explained is “significant” when it is greater than a given threshold. If the percentage is significant, then it is possible to explain the variability in data with m variables.

Table 2. Class throughput correlation matrix. The flows in class c arrive according to a Poisson process with rate λ_c with sizes drawn independently from an exponential (exp) and a bounded Pareto (BP) distribution in (1) with mean 1. $\lambda_{c_0} = 0.2, \lambda_{c_1} = 0.6, \lambda_{c_2} = 0.1$.

exp	Class 0	Class 1	Class 2
Class 0	1	0.803	-0.034
Class 1	0.803	1	-0.145
Class 2	-0.034	-0.145	1
BP	Class 0	Class 1	Class 2
Class 0	1	0.769	0.005
Class 1	0.769	1	-0.092
Class 2	0.005	-0.092	1

Table 3. Class throughput correlation matrix. The flows in class c arrive according to a Poisson process with rate λ_c with sizes drawn independently from an exponential (exp) and a bounded Pareto (BP) distribution in (1) with mean 1. $\lambda_{c_0} = 0.4, \lambda_{c_1} = 0.4, \lambda_{c_2} = 0.4$.

exp	Class 0	Class 1	Class 2
Class 0	1	0.554	0.624
Class 1	0.554	1	0.072
Class 2	0.624	0.072	1
BP	Class 0	Class 1	Class 2
Class 0	1	0.527	0.602
Class 1	0.527	1	0.049
Class 2	0.602	0.049	1

We argue that m corresponds to *the number of shared resources* among flow classes. We list the eigenvalues of the correlation matrices reported in Tables 1 through 4, and the percentage of normalized variance explained by the first two principal components in Table 5. Since more than 90% of normalized variance is captured by the first two principal components, we can conclude that there were two shared resources in the network for the classes under consideration. This inference is based only on measurements made at a single point *without* knowing the topology of the network.

Table 4. Class throughput correlation matrix. The flows in class c arrive according to a Poisson process with rate λ_c with sizes drawn independently from an exponential (exp) and a bounded Pareto (BP) distribution in (1) with mean 1. $\lambda_{c_0} = 0.6, \lambda_{c_1} = 0.2, \lambda_{c_2} = 0.1$.

exp	Class 0	Class 1	Class 2
Class 0	1	0.810	0.631
Class 1	0.810	1	0.391
Class 2	0.631	0.391	1
BP	Class 0	Class 1	Class 2
Class 0	1	0.759	0.629
Class 1	0.759	1	0.456
Class 2	0.629	0.456	1

Table 5. Eigenvalues and the percentage of normalized variance captured by the first two principal components of class throughput correlation matrices.

Table	Eigenvalues	% Variance
1 (exp)	(2.108, 1.689, 0.102, 0.101)	94.9
1 (BP)	(1.969, 1.737, 0.154, 0.140)	92.6
2 (exp)	(1.823, 0.988, 0.189)	93.7
2 (BP)	(1.774, 1.001, 0.225)	92.5
3 (exp)	(1.871, 0.929, 0.201)	93.3
3 (BP)	(1.824, 0.952, 0.224)	92.5
4 (exp)	(2.237, 0.625, 0.138)	95.4
4 (BP)	(2.238, 0.558, 0.205)	93.2

5. PRELIMINARY VALIDATION AND CONCLUSION

In our preliminary studies with actual TCP flow measurements collected over a one-hour period at the border router at The University of Texas at Austin, we analyzed the class throughput correlation matrix of four inbound Web traffic classes (whose class throughputs were assumed to be stationary over one hour) by using a systematic method, namely *factor analysis* [11]. By using factor analysis, we were able to associate flow classes with shared resources as well. For validation purposes, we selected flow classes (with source IP addresses associated with HotMail, MSN, AOL, and CNN) whose providers were known with reasonable certainty (HotMail and MSN from Microsoft Corporation, AOL and CNN from AOL Transit Data Network). We assumed that the selected flow classes experienced congestion at their source due to high demand for their content.

Based only on collected flow measurements, the method successfully identified the classes originating from the same infrastructure after establishing a proper threshold for omitting flows with short durations when computing the class throughputs. Due to space constraints, in the present paper we omit the discussion of the determination of the threshold for short flows and factor analysis results, and outline only the determination of the number of shared resources among flow classes.

For the class throughput correlation matrix of four selected classes, we identified two significant principal components based on the following (heuristic) criteria: We assumed that a principal component was significant if it contributed more than a “variance” of 1 to the total normalized variance. The two principal components accounted for 67.5% (which was deemed significant for our exploratory purposes) of the total normalized variance with a 95% confidence interval $[66.3\%, 68.7\%]$ that was computed by using the bootstrap bias-corrected and accelerated (BC$_a$) method [12]. Hence, we concluded that the variability in four class throughputs could be explained by two principal components, which corresponded to two different providers. Note that the distribution of throughputs of short TCP flows is generally widely dispersed [2]. The omission of short flows was necessary in order to “filter out” the noise introduced by short flows into class throughputs.

6. REFERENCES

[1] L. Massoulié and J. W. Roberts, “Bandwidth sharing and admission control for elastic traffic,” *Telecom. Sys.*, vol. 15, pp. 185–201, June 2000.

[2] S. Ben Fred, T. Bonald, A. Proutiere, G. Régnié, and J. W. Roberts, “Statistical bandwidth sharing: a study of congestion at flow level,” in *Proc. ACM Conf. on Appl., Tech., Arch., and Protocols for Comp. Comm.*, Aug. 2001, pp. 111–122.

[3] R. Cáceres, N. G. Duffield, J. Horowitz, and D. F. Towsley, “Multicast-based inference of network-internal loss characteristics,” *IEEE Trans. on Info. Theory*, vol. 45, no. 7, pp. 2462–2480, Nov. 1999.

[4] F. Lo Presti, N. G. Duffield, J. Horowitz, and D. Towsley, “Multicast-based inference of network-internal delay distributions,” *IEEE Trans. on Networking*, vol. 10, no. 6, pp. 761–775, Dec. 2002.

[5] D. Rubenstein, J. Kurose, and D. Towsley, “Detecting shared congestion of flows via end-to-end measurement,” *IEEE/ACM Trans. on Networking*, vol. 10, no. 3, pp. 381–395, June 2002.

[6] M. Rabbat, R. Nowak, and M. Coates, “Network tomography and the identification of shared infrastructure,” in *Proc. IEEE Asilomar Conf. on Signals, Sys. and Comp.*, Nov. 2002, pp. 34–38.

[7] Y. Tsang, M. Coates, and R. D. Nowak, “Network delay tomography,” *IEEE/ACM Trans. on Signal Processing*, vol. 51, no. 8, pp. 2125–2136, Aug. 2003.

[8] L. Kleinrock, *Queueing Systems: Computer Applications*, vol. 2, Wiley-Interscience, 1976.

[9] F. P. Kelly, “Charging and rate control for elastic traffic,” *European Trans. on Telecom.*, vol. 8, pp. 33–37, 1997.

[10] M. E. Crovella, M. Harchol-Balter, and C. D. Murta, “Task assignment in a distributed system: Improving performance by unbalancing load,” in *Proc. ACM Conf. on Measurement and Modeling of Comp. Sys.*, June 1998, pp. 268–269.

[11] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, 5th edition, 2002.

[12] D. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, Inc., 1993.