Congestion In Large Balanced Fair Links

Thomas Bonald (Telecom Paris-Tech), Jean-Paul Haddad (Ernst and Young) and Ravi R. Mazumdar (Waterloo)

ITC 2011, San Francisco

Introduction

- File transfers compose much of the traffic of the current Internet
- Main measures of the quality of service (QoS) are the transfer rates and duration of the file transfer
- Being able to estimate congestion (when rates are below desired rates) is of great importance to dimensioning capacity to achieve QoS requirements
- Doing so that is both insensitive to traffic characteristics and tractable will lead to robust engineering rules in designing future networks

Scope Of Talk

- The main focus of this talk will be on congestion in single links that operate under a balanced fair allocation scheme for heterogeneous flows with differing maximum or peak bandwidth requirements
- Using ideas from local limit large deviations of convolution measures associated, formulas for estimating different measures of congestion that are computationally tractable for large parameters will be presented.
  A presentation of the mathematical background can be found in:

  R. R. Mazumdar, *Performance Modelling, Loss Networks and Statistical Multiplexing*, Series on Communication Networks (J. Walrand, ed.), Morgan and Claypool, 2010.

Nomenclature

- The system is a single link with M classes of traffic
- Link capacity C
- Rate limits on individual flows r_i, $i = 1 \dots M$
- Traffic intensity $\alpha_i = \lambda_i / \mu_i$, $i = 1 \dots M$
- $\beta_i = \alpha_i / r_i$, $i = 1 \dots M$
- Load $\rho = \sum_j \alpha_j / C$
- Allocated bandwidth ϕ_i, $i = 1 \dots M$

Flow Level Model

- Introduced by Roberts and Massoulié [4]
- Ignores the packet level dynamics and models the file transfers as fluid flows
- The bandwidth allocated to flows of the same class are shared equally
- In this talk, we will assume that all flows are rate limited and go through a single bottleneck link
- This can be modeled as letting each class of flow go to separate processor sharing queues but with variable capacity depending on number of flows in system

Markov Process

- Let X be the state process, where the state is the numbers of flows of each class
- X is modeled as a continuous time jump Markov process
- State transition rates: $q(\vec{x},\vec{y}) = \begin{cases} \lambda_i & \vec{y} = \vec{x} + \vec{e}_i \\ \mu_i \phi_i(\vec{x}) & \vec{y} = \vec{x} - \vec{e}_i \\ 0 & Otherwise \end{cases}$

Bandwidth Allocation

- Bandwidth allocation is a fundamental, well studied problem
- Most popular and studied class of allocations are the *Utility* based allocations
- Let $\vec{x}$ be the state vector whose components x_i are the number flows of class i

$$\max_{\phi} \sum_j x_j U(\phi_j(\vec{x})/x_j)$$

$$s.t. \ \textstyle\sum_j \phi_j(\vec{x}) \leq C$$
$$\phi_i(\vec{x}) \leq x_i r_i$$

when $U_i(x) = \log x$ it is termed *proportional fairness*.

Insensitive Allocations

- Characterized by Balance Function Φ
- Allocation is defined as $\phi_i(\vec{x}) = \frac{\Phi(\vec{x}-\vec{e}_i)}{\Phi(\vec{x})}$
- Insensitive allocations have the advantage that the stationary distribution $\pi(\vec{x})$ depends on the flow size distribution only through its mean
- $$\pi(\vec{x}) = \pi(\vec{0})\Phi(\vec{x}) \prod_{i=1}^{M} \alpha_i^{x_i}$$

Balanced Fairness

- Introduced by Bonald and Proutière [2].
- Most efficient insensitive allocation is Balanced Fairness

Lemma

Consider another positive function $\tilde{\Phi}$ such that $\tilde{\Phi}(0) = 1$ and the rate and capacity constraints are satisfied. Then

$$\tilde{\Phi}(\vec{x}) \geq \Phi(\vec{x}) \quad \forall \vec{x} \in \mathbb{Z}_+^M. \tag{1}$$

- The Balance Function for a single link is:

$$\Phi(\vec{x}) = \max\left(\frac{1}{C}\sum_{i=1}^{M}\Phi(\vec{x}-\vec{e}_i), \max_{i:\, x_i>0}\frac{\Phi(\vec{x}-\vec{e}_i)}{x_i r_i}\right)$$

- The last constraint i.e. $\phi_i(\vec{x}) \leq x_i r_i$ is a rate constraint on each flow. If $r_i = \infty$ it would reduce to processor sharing.

- The balance function can be simplified to:

$$\Phi(\vec{x}) = \begin{cases} \displaystyle\prod_{i=1}^{M} \frac{1}{x_i! r_i^{x_i}} & \text{if } \vec{x}^T \vec{r} \leq C, \\ \displaystyle\frac{1}{C} \sum_{i=1}^{M} \Phi(\vec{x} - \vec{e_i}) & \text{Otherwise} \end{cases}$$

- **Lemma** $\forall i = 1 \ldots M$, $\phi_i(x) = x_i r_i$ iff $\vec{x}^T \vec{r} \leq C$
  This property implies that either all classes get their max rate or none do
- **Theorem** Stable iff $\rho < 1$

Balanced Fairness and Proportional Fairness

- Assuming $r_i = \infty \ \forall\, i$, Balanced Fairness coincides with proportional fairness on many topologies and has been empirically shown to approximate Proportional Fairness well in many cases
- Massoulié [3] proved some very useful theoretical connections between Balanced Fairness and Proportional Fairness
  - **Theorem** If there exists $\tilde{\phi}$ s.t. $\phi_i^{BF}(n\vec{x}) \longrightarrow \tilde{\phi}_i(\vec{x})$ as $n \to \infty$, then $\tilde{\phi}(\vec{x}) = \phi^{PF}(\vec{x})$
  - **Theorem** $\lim_{n\to\infty} \frac{1}{n} \log \pi^{BF}(n\vec{x}) \Rightarrow -\max \sum_j x_j \log(\phi_j/\alpha_j)$ s.t. $\phi \in \mathcal{C}$
    Where $\mathcal{C}$ is the set of feasible allocations.
- **Conjecture** $\phi_i^{BF}(n\vec{x}) \longrightarrow \phi_i^{PF}(\vec{x})$ as $n \to \infty$
- Walton [5] has generalized the results of Massoulié to any max stable (ie. stability condition $\rho < 1$) insensitive allocation

Congestion Metrics

- We will look at three metrics related to the long run congestion of the system:

1. Probability of congestion P - The long run fraction of time that the system spends in a congested state.
2. Probabilities of congestion P_i - The long run probability that an arrival of class i will arrive at a congested system or cause the congestion in link.
3. F_i - Fraction of the average sojourn time that a customer of class i does not get its maximum rate while in the system.

Congestion Metrics

- From PASTA and the properties of balanced fairness, one can get a simple characterization of the first two congestion metrics:
  - $P = \sum_{\vec{x}:\, \vec{x}^T \vec{r} > C} \pi(\vec{x})$
  - $P_i = \sum_{\vec{x}:\, \vec{x}^T \vec{r} > C - r_i} \pi(\vec{x})$

Congestion Metrics

- Formally, we define

$$F_i = \frac{\mathsf{E}_i \left[\int_0^{\tau_i} 1_{\{\vec{X}(t)^T \vec{r} > C\}} dt \right]}{\mathsf{E}_i [\tau_i]}$$

  Where τ_i is the sojourn time of a class i arrival, $\vec{X}$ the stationary state process and E_i indicates the expectation with respect to the Palm probability of arrivals of class i

- For our purposes, the metric is not useful in this form and we require an alternative characterization

Congestion Metrics

- **Theorem** (Swiss Army Formula) [1]
  $$\lambda_A E_A \left[\int_0^{W_0} Z(s) dB(s) \right] = \frac{1}{t} E_\pi \left[\int_0^t X(s^-) Z(s) dB(s) \right]$$
  Where A is a point process, W_n a sequence of marks for A, X, Z non-negative processes and B a non-decreasing process
- Applying the Swiss Army Formula, we now get

$$F_i = \frac{\sum_{\vec{x}:\, \vec{x}^T \vec{r} > C} x_i \pi(\vec{x})}{\sum_{\vec{x}} x_i \pi(\vec{x})}$$

Congestion Metrics

- The congestion metrics can be written as a function of far fewer states
- **Lemma**

$$P = \sum_{i=1}^{M} \frac{\rho_i B_i}{1-\rho}$$

and

$$P_i = B_i + P$$

with

$$B_i = \sum_{\vec{x}:\, C - r_i < \vec{x}^T \vec{r} \leq C} \pi(\vec{x})$$

Congestion Metrics

- **Lemma** For all $i, j = 1, \ldots, M$, let

$$Q_{ij} = \sum_{\vec{x}:\, C - r_j < \vec{x}^T \vec{r} \leq C} x_i \pi(\vec{x}),$$

and

$$Q_i = \sum_{\vec{x}:\, \vec{x}^T \vec{r} > C} x_i \pi(\vec{x}).$$

Then

$$Q_i = \frac{\rho_i P_i}{1 - \rho} + \sum_{j=1}^{M} \frac{\rho_j Q_{ij}}{1 - \rho},$$

$$F_i = \frac{Q_i}{Q_i + \sum_{\vec{x}:\, \vec{x}^T \vec{r} \leq C} x_i \pi(\vec{x})}.$$

Erlang Multirate Loss System

- The states that are used to calculate the congestion measures are the same states that are used to calculate the blocking formula in an Erlang loss system
- In fact, for any state $\vec{x} : \vec{x}^T \vec{r} \leq C$, the stationary probability is proportional to the stationary of an associated loss system since $\pi(\vec{x}) = \pi(\vec{0}) \prod_i \frac{(\alpha_i / r_i)^{x_i}}{x_i!}$
- Like the loss system counterpart, when parameters are large, the computation becomes onerous
- Using ideas from local limit large deviations of convolution measures one can get an accurate approximation by scaling the traffic intensities and link capacity

Notion of a large system

The notion of a large system is obtained by scaling both the capacity and arrival rates by a factor N. Define $C(N) = NC$ and $\lambda_k(N) = N\lambda_k$. Note this notion extends to networks
In other words the *large* system can be seen as a N fold scaling of a nominal system where connections arrive at rate λ_k, allocated $\frac{\phi_k(\vec{x})}{x_k}$ units of bandwidth, and the server capacity is C.

Main Results

Theorem

$$P(N) \sim \sum_{i=1}^{M} \frac{\rho_i P_i^B(N)}{1-\rho}$$

and for all $i = 1 \ldots M$:

$$P_i(N) \sim P_i^B(N) + P(N)$$

Where:

$$P_i^B(N) \sim e^{-NI} e^{\tau d \epsilon(N)} \frac{d}{\sqrt{2\pi N}\sigma} \frac{1 - e^{\tau r_i}}{1 - e^{\tau d}}$$

d is the greatest common divisor of $r_1, \ldots, r_M$,
$\epsilon(N) = \frac{NC}{d} - \left\lfloor \frac{NC}{d} \right\rfloor$,

τ is the unique solution to the equation $\displaystyle\sum_{i=1}^{M} r_i \beta_i e^{\tau r_i} = C$,

$$I = C\tau - \sum_{i=1}^{M} \beta_i \left(e^{\tau r_i} - 1\right),$$

$$\sigma^2 = \sum_{i=1}^{M} r_i^2 \beta_i e^{\tau r_i}.$$

Theorem

$$F_i(N) \sim \frac{r_i}{NC(1-\rho)} P_i(N) + \sum_{j=1}^{M} \frac{\rho_j}{1-\rho} P_{ij}^B(N)$$

$$P_{ij}^B(N) \sim e^{-NI_i} e^{\tau_i d \epsilon_i(N)} \frac{d}{\sqrt{2\pi N}\sigma_i} \frac{1-e^{\tau_i r_j}}{1-e^{\tau_i d}}$$

Where:

d is the greatest common divisor of $r_1, \ldots, r_M$,

$\epsilon_i(N) = \frac{NC - r_i}{d} - \left\lfloor \frac{NC - r_i}{d} \right\rfloor$,

τ is the unique solution to the equation $\displaystyle\sum_{j=1}^{M} r_j \beta_j e^{\tau r_j} = C$,

$$\sigma^2 = \sum_{j=1}^{M} r_j^2 \beta_j e^{\tau r_j},$$

$\tau_i = \tau - \frac{r_i}{N\sigma^2}$,

$$I_i = \left(C - \frac{r_i}{N}\right) \tau_i - \sum_{j=1}^{M} \beta_j \left(e^{\tau_i r_j} - 1\right),$$

$$\sigma_i^2 = \sum_{j=1}^{M} r_i^2 \beta_j e^{\tau_i r_j}$$

Method of Proof

- Renormalize the congestion formulas so that they are now computed using the stationary distributions of the associated loss system
- Show that the normalization constants of the loss system and original system coincide in the limit
- Apply approximation for loss networks to the formulas for the congestion metrics

Numerical Example

- The system has $M = 3$ classes of traffic
- Link capacity $C = 10$
- Rate limits $r_1 = 1$, $r_2 = 2$, $r_3 = 5$
- Loads $\rho_1/\rho = 0.5$, $\rho_2/\rho = 0.3$, $\rho_3/\rho = 0.2$

Numerical Example

Congestion Probabilities
Medium load, $\rho = 0.6$

	Exact			Approximation		
N	$F_1(N)$	$F_2(N)$	$F_3(N)$	$F_1(N)$	$F_2(N)$	$F_3(N)$
10	9.98e-04	1.24e-03	2.36e-03	9.99e-04	1.24e-03	2.36e-03
20	5.60e-06	6.95e-06	1.32e-05	5.60e-06	6.95e-06	1.32e-05
30	3.63e-08	4.50e-08	8.57e-08	3.63e-08	4.50e-08	8.57e-08
40	2.49e-10	3.09e-10	5.89e-10	2.49e-10	3.09e-10	5.89e-10
50	1.77e-12	2.19e-12	4.18e-12	1.77e-12	2.19e-12	4.18e-12

Numerical Example

Congestion Probabilities
Heavy load, $\rho = 0.9$

	Exact			Approximation		
N	$F_1(N)$	$F_2(N)$	$F_3(N)$	$F_1(N)$	$F_2(N)$	$F_3(N)$
10	3.65e-01	3.83e-01	4.43e-01	4.38e-01	4.59e-01	5.32e-01
20	2.22e-01	2.33e-01	2.70e-01	2.41e-01	2.53e-01	2.93e-01
30	1.43e-01	1.54e-01	1.78e-01	1.53e-01	1.61e-01	1.86e-01
40	1.01e-01	1.06e-01	1.22e-01	1.03e-01	1.08e-01	1.25e-01
50	7.07e-02	7.42e-02	8.60e-02	7.18e-02	7.54e-02	8.73e-02

Numerical Example

Time-average congestion rates
Heavy load, $\rho = 0.9$

	Exact			Approximation		
N	$F_1(N)$	$F_2(N)$	$F_3(N)$	$F_1(N)$	$F_2(N)$	$F_3(N)$
10	3.87e-01	4.26e-01	5.37e-01	4.81e-01	5.49e-01	7.74e-01
20	2.31e-01	2.50e-01	3.12e-01	2.53e-01	2.78e-01	3.59e-01
30	1.51e-01	1.62e-01	2.00e-01	1.58e-01	1.71e-01	2.14e-01
40	1.03e-02	1.10e-02	1.34e-02	1.06e-01	1.14e-01	1.40e-01
50	7.20e-02	7.69e-02	9.30e-02	7.32e-02	7.83e-02	9.52e-02

Network Case

- In general, network case is very difficult to analyze
- For specific topologies, the techniques from the single link analysis can be applied
- Of practical interest is a structure occurring in access networks referred to as a parking lot network.

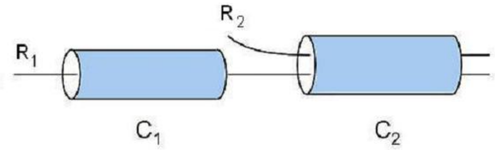

Figure: Two Link Parking Lot Network

- The network has 2 links and 2 routes
- Route $\mathtt{R}_1$ goes through both links and route $\mathtt{R}_2$ goes through the second link only
- Each of the M classes of traffic follow one of the two routes
- Only the case that the capacities of the links satisfy $C_1 < C_2$ is of interest otherwise, the problem reduces to single link case

Numerical Example

We conclude the presentation with a numerical example for a parking lot example:

- The system has $M = 4$ classes of traffic, two on each route
- Link capacities $C_1 = 5$ and $C_2 = 9$
- Rate limits on route R_1 are $r_1 = 1$, $r_2 = 2$
- Rate limits on route R_2 are $r_3 = 1$, $r_4 = 2$
- Traffic intensities on route R_1 are $\alpha_1 = 2$, $\alpha_2 = 1$
- Traffic intensities on route R_2 are $\alpha_3 = 2$, $\alpha_4 = 1$

Numerical Example

Congestion Probability $P(N)$

	Exact	Approximation
N		
10	7.41e-04	9.04e-04
20	4.67e-06	5.20e-06
30	3.29e-08	3.51e-08
40	2.43e-10	2.52e-10

Concluding Remarks

- Extension to tree networks is possible
- Balanced fairness is a good model for *insensitive* bandwidth sharing in cloud computing
- Close parallels with VCG auctions
- Large system means we can approximate balanced fairness via proportional fairness for which a mechanism design exists (primal-dual).

References I

- F. Baccelli and P. Brémaud.
  Elements of Queueing Theory, volume 47 of *Stochastic Modelling and Applied Probability*.
  Springer, Berlin, 2nd edition, 2003.

- T. Bonald and A. Proutière.
  Insensitive bandwidth sharing in data networks.
  Queueing Syst. Theory Appl., 44(1):69–100, 2003.

- L. Massoulié.
  Structural properties of proportional fairness: Stability and insensitivity.
  Ann. Appl. Probab., 17(3):809–839, 2007.

- J. W. Roberts and L. Massoulié.
  Bandwidth sharing and admission control for elastic traffic.
  Telecommunication Systems, 15:185–201, 2000.

References II

N. Walton.
Insensitive, maximum stable allocations converge to proportional fairness.
Preprint, 2010.