

Integration of Elastic and Adaptive Streaming Flows

Thomas Bonald

France Telecom R&D

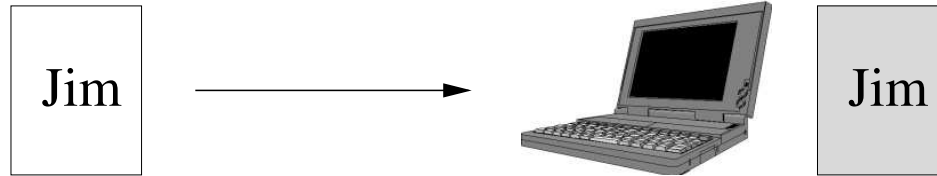
`thomas.bonald@francetelecom.com`

Joint work with Alexandre Proutière

Sigmatrics - Performance, New York City, June 2004

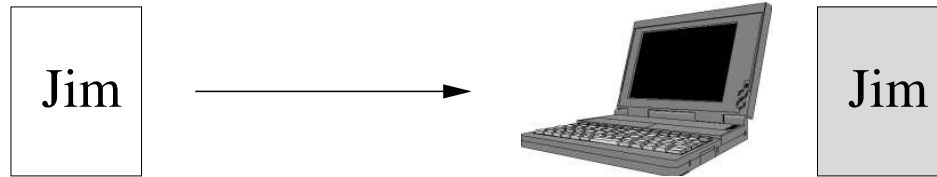
Elastic vs. streaming traffic

- Elastic traffic
 - transfer of digital documents (e.g., Web, emails)
 - a fixed **flow size** (in bits)
 - a variable flow duration

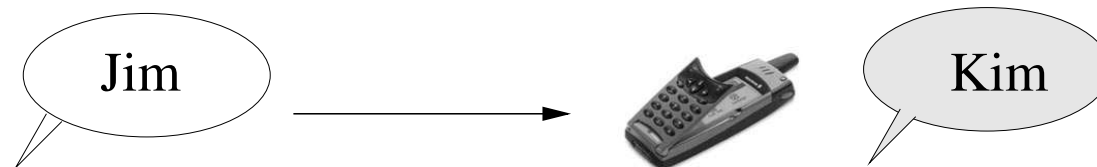


Elastic vs. streaming traffic

- Elastic traffic
 - transfer of digital documents (e.g., Web, emails)
 - a fixed **flow size** (in bits)
 - a variable flow duration

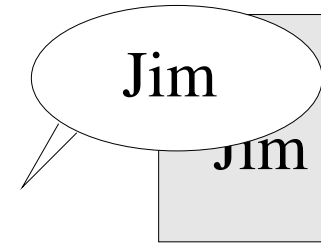


- Streaming traffic
 - real-time transfer of signals (e.g., voice, video)
 - a fixed **flow duration**
 - a variable flow size



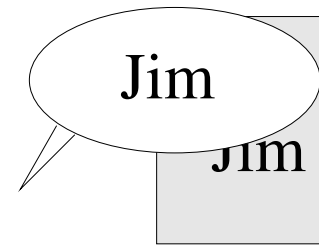
Integration

- Service differentiation
 - e.g., **priority** to streaming traffic
 - requires the marking of all packets
 - may starve elastic traffic
- cf. **Núñez Queija, 2000, Delcoigne et. al., 2004**



Integration

- Service differentiation
 - e.g., **priority** to streaming traffic
 - requires the marking of all packets
 - may starve elastic trafficcf. **Núñez Queija, 2000, Delcoigne et. al., 2004**



- Best effort
 - e.g., FIFO or fair queueing
 - requires the self **adaptation** of all flows
 - both elastic and adaptive streaming flows suffer from congestion periods

Network dynamics

- A dynamic, **random** number of flows
 - cf. Erlang's model

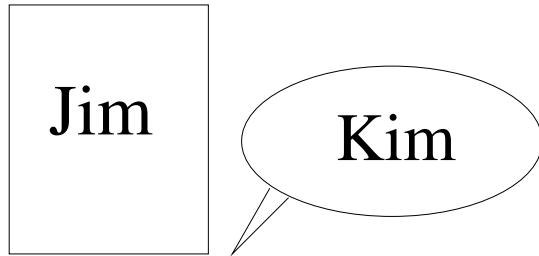
Network dynamics

- A dynamic, **random** number of flows
 - cf. Erlang's model

Jim

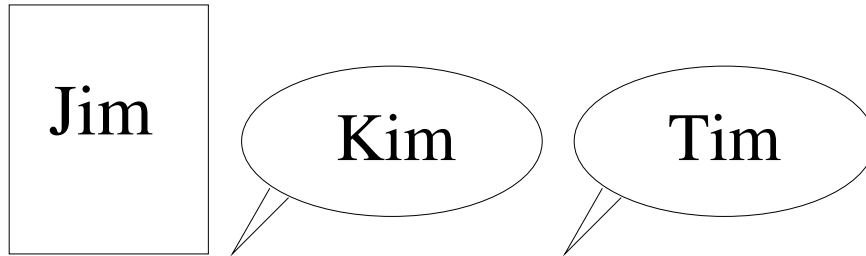
Network dynamics

- A dynamic, **random** number of flows
 - cf. Erlang's model



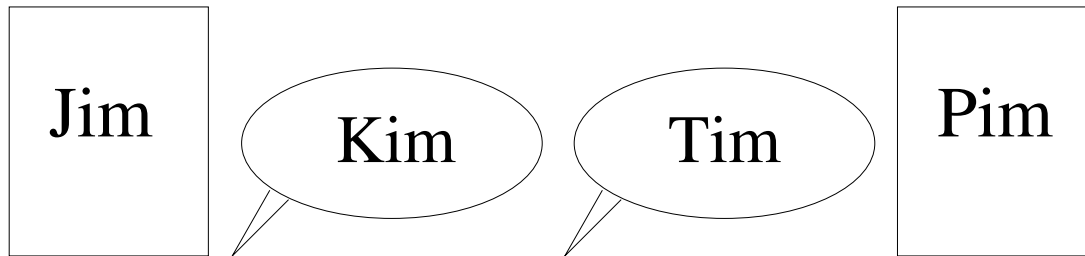
Network dynamics

- A dynamic, **random** number of flows
 - cf. Erlang's model



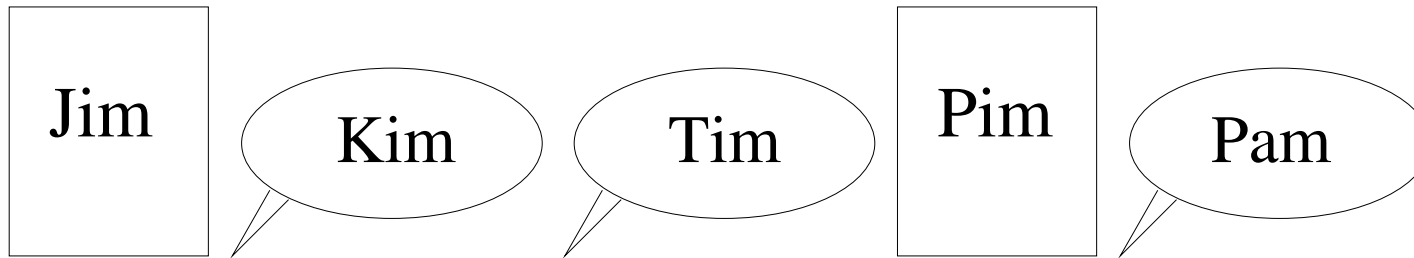
Network dynamics

- A dynamic, **random** number of flows
– cf. Erlang's model



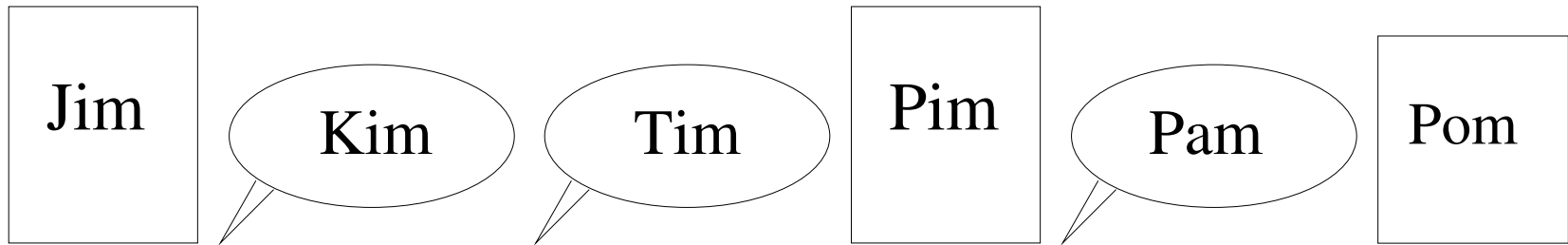
Network dynamics

- A dynamic, **random** number of flows
 - cf. Erlang's model



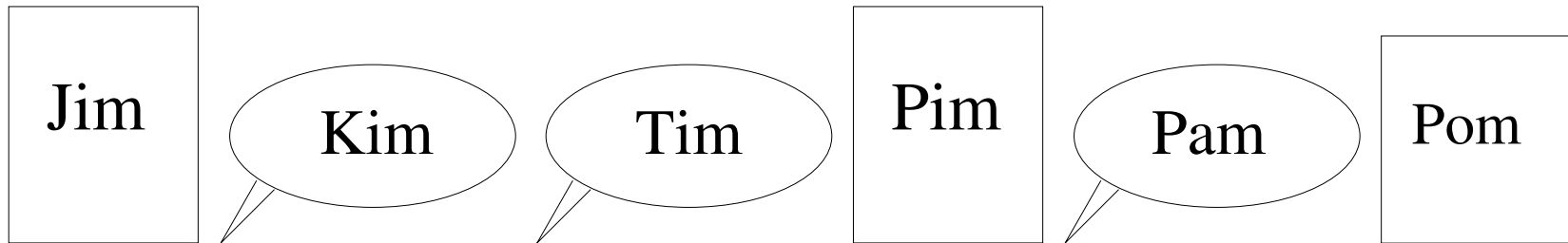
Network dynamics

- A dynamic, **random** number of flows
 - cf. Erlang's model



Network dynamics

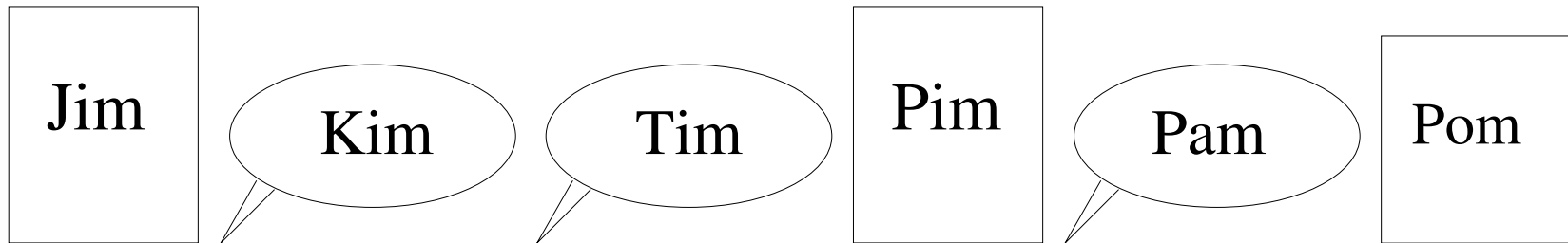
- A dynamic, **random** number of flows
 - cf. Erlang's model



- The flow-level approach
 - assumes a **perfect** rate adaptation
 - studies the evolution of the number of flows
 - deduces **user perceived** performance

Network dynamics

- A dynamic, **random** number of flows
 - cf. Erlang's model



- The flow-level approach
 - assumes a **perfect** rate adaptation
 - studies the evolution of the number of flows
 - deduces **user perceived** performance
- Related work
 - **Key, Massoulié, Bain & Kelly, 2003**
the network **stability condition** is not affected by the presence of streaming traffic

Outline

- Model
- Traffic in isolation
 - elastic traffic only
 - streaming traffic only
- Integrated system
 - a single bottleneck
 - a common rate limit
 - multiple rate limits, multiple bottlenecks
- Conclusion

Model

- Traffic assumptions
 - Poisson arrivals, at rates λ_e, λ_s
 - i.i.d. elastic flow sizes, mean $1/\mu_e$
 - i.i.d. streaming flow durations, mean $1/\mu_s$
 - traffic intensities

$$\rho_e = \frac{\lambda_e}{\mu_e} \quad (\text{bits/s}) \quad \rho_s = \frac{\lambda_s}{\mu_s} \quad (\text{erlangs})$$

- Bandwidth sharing
 - perfect rate adaptation
 - fair sharing
 - in state $x = (x_e, x_s)$, the rate of each flow is

$$\frac{1}{x_e + x_s} \quad (\text{unit capacity link})$$

Outline

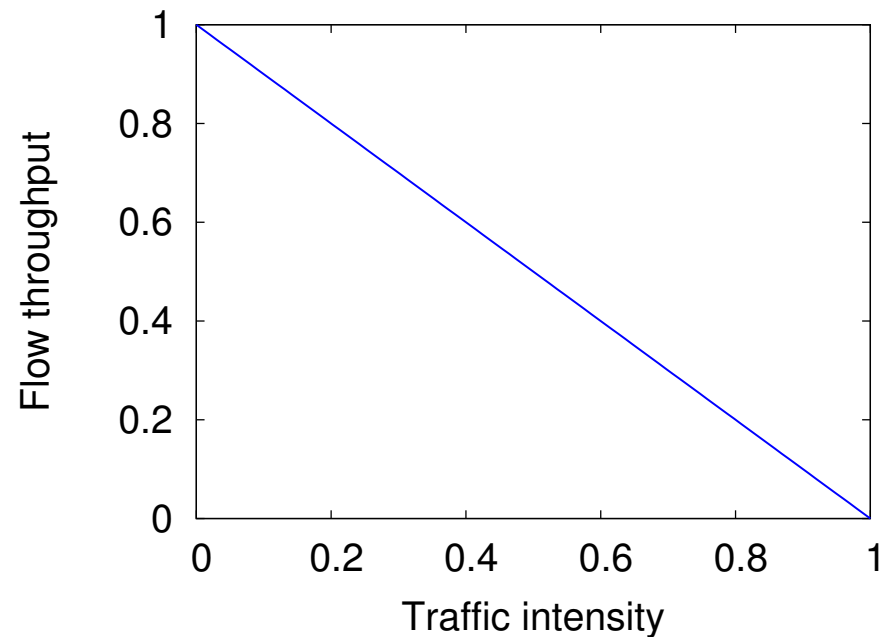
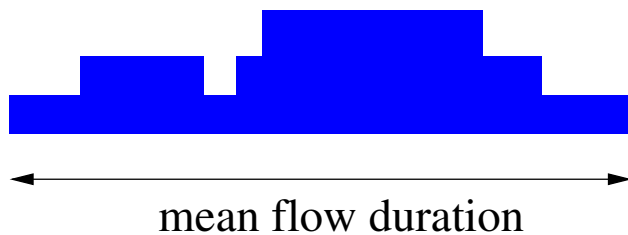
- Model
- Traffic in isolation
 - elastic traffic only
 - streaming traffic only
- Integrated system
 - a single bottleneck
 - a common rate limit
 - multiple rate limits, multiple bottlenecks
- Conclusion

Elastic traffic only

- An $M/G/1$ processor sharing queue
 - stable iff $\rho_e < 1$
 - explicit stationary distribution, $\pi_e(x) = (1 - \rho_e)\rho_e^x$
 - **insensitive** to the flow size distribution

Elastic traffic only

- An $M/G/1$ processor sharing queue
 - stable iff $\rho_e < 1$
 - explicit stationary distribution, $\pi_e(x) = (1 - \rho_e)\rho_e^x$
 - **insensitive** to the flow size distribution
- User performance

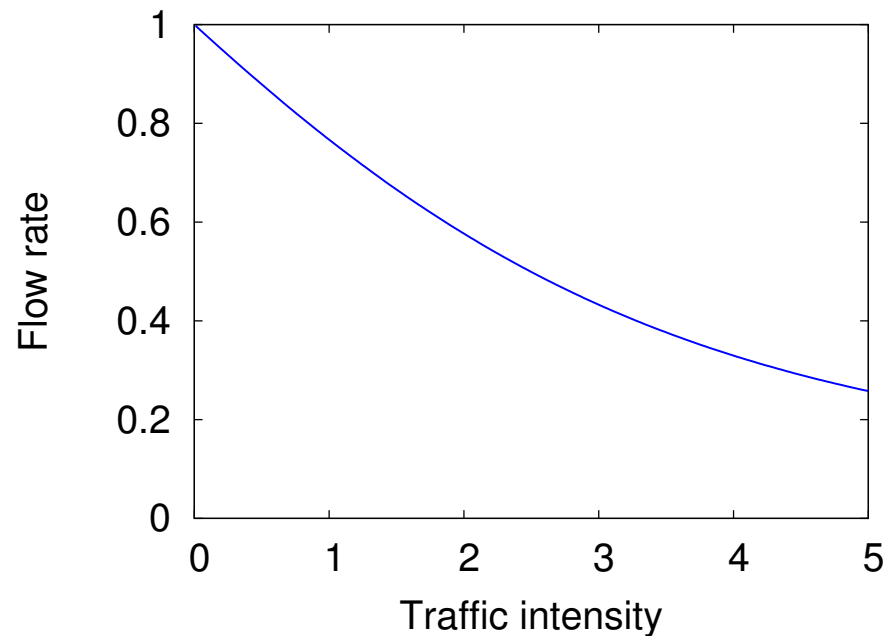
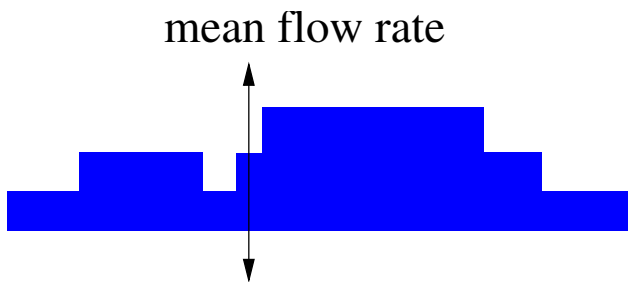


Streaming traffic only

- An $M/G/\infty$ queue
 - always stable
 - explicit stationary distribution, $\pi_s(x) = e^{-\rho_s} \frac{\rho_s^x}{x!}$
 - **insensitive** to the flow duration distribution

Streaming traffic only

- An $M/G/\infty$ queue
 - always stable
 - explicit stationary distribution, $\pi_s(x) = e^{-\rho_s} \frac{\rho_s^x}{x!}$
 - **insensitive** to the flow duration distribution
- User performance



Outline

- Model
- Traffic in isolation
 - elastic traffic only
 - streaming traffic only
- **Integrated system**
 - a single bottleneck
 - a common rate limit
 - multiple rate limits, multiple bottlenecks
- Conclusion

A single bottleneck

- Two **coupled** processor sharing queues
 - with state-dependent service rates

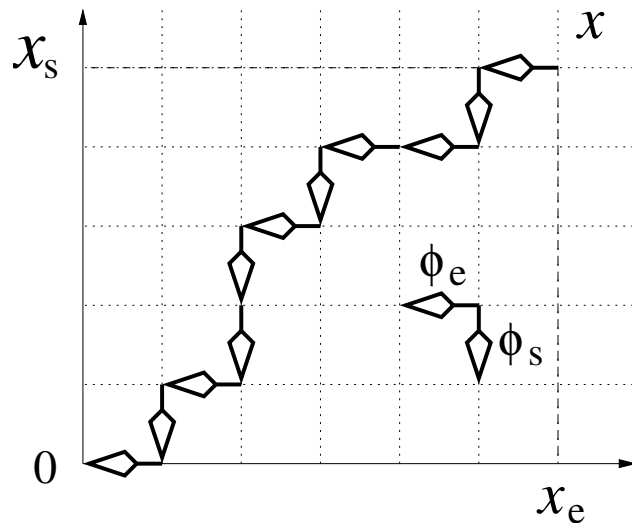
$$\phi_e(x) = \frac{x_e}{x_e + x_s} \mu_e, \quad \phi_s(x) = x_s \mu_s$$

A single bottleneck

- Two **coupled** processor sharing queues
 - with state-dependent service rates

$$\phi_e(x) = \frac{x_e}{x_e + x_s} \mu_e, \quad \phi_s(x) = x_s \mu_s$$

- Necessary and sufficient condition for **insensitivity**
 - **balance** property: $\phi_e(x)\phi_s(x - f_e) = \phi_s(x)\phi_e(x - f_s)$



$$\Phi(x) = \frac{1}{\phi_e(x)\phi_s(x - f_e) \dots \phi_e(f_e)}$$

$$\pi(x) = \pi(0)\Phi(x)\rho_e^{x_e}\rho_s^{x_s}$$

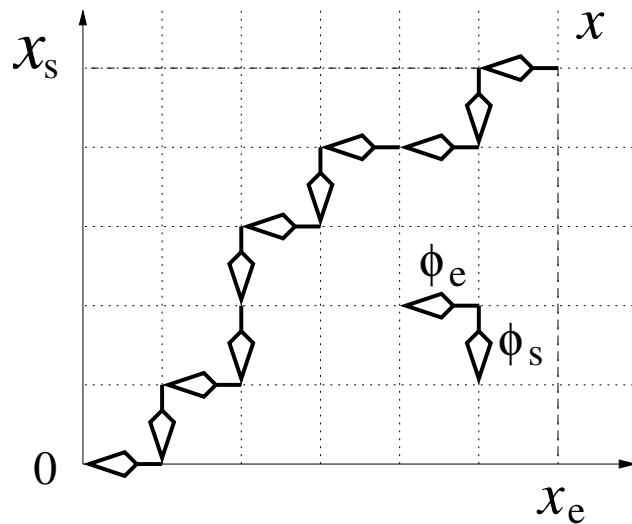
A single bottleneck

- Two **coupled** processor sharing queues
 - with state-dependent service rates

$$\phi_e(x) = \frac{x_e}{x_e + x_s} \mu_e, \quad \phi_s(x) = x_s \mu_s$$

- The balance property does **not** hold!

$$\phi_e(x) \phi_s(x - f_e) \neq \phi_s(x) \phi_e(x - f_s)$$

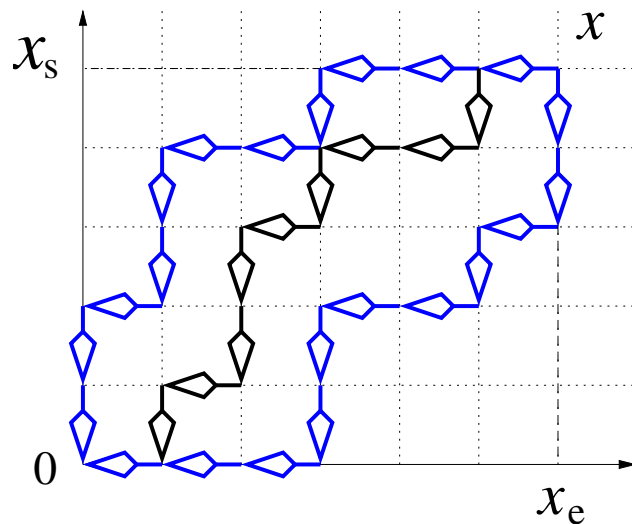


$$\Phi(x) = ?$$

$$\pi(x) = ?$$

Insensitive bounds

- Stochastic comparison with insensitive systems
 - let $\hat{\Phi}(x)$, $\check{\Phi}(x)$ be the **extreme paths** from x to 0
 - the corresponding “virtual” systems provide bounds for the original system
- Applicable to any **monotonic** queueing network!

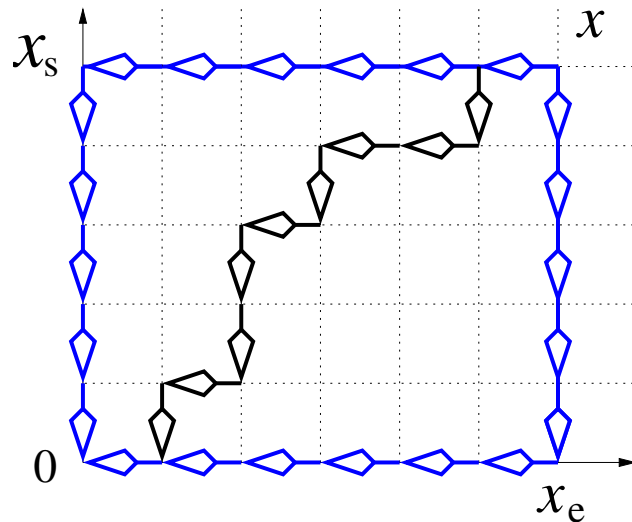


$$\hat{\pi}(x) = \hat{\pi}(0) \hat{\Phi}(x) \rho_e^{x_e} \rho_s^{x_s}$$

$$\check{\pi}(x) = \check{\pi}(0) \check{\Phi}(x) \rho_e^{x_e} \rho_s^{x_s}$$

Insensitive bounds

- Stochastic comparison with insensitive systems
 - let $\hat{\Phi}(x)$, $\check{\Phi}(x)$ be the **extreme paths** from x to 0
 - the corresponding “virtual” systems provide bounds for the original system
- Applicable to any **monotonic** queueing network!



$$\hat{\pi}(x) = \hat{\pi}(0) \hat{\Phi}(x) \rho_e^{x_e} \rho_s^{x_s}$$

$$\check{\pi}(x) = \check{\pi}(0) \check{\Phi}(x) \rho_e^{x_e} \rho_s^{x_s}$$

Another insensitive bound

- Assume **all traffic is elastic**
 - an $M/G/1$ processor sharing queue
 - stable iff $\rho_e + \rho_s < 1$
 - explicit stationary distribution

$$\tilde{\pi}(x) = (1 - \rho_e - \rho_s) \binom{x_e + x_s}{x_e} \rho_e^{x_e} \rho_s^{x_s}$$

- A **conservative** approximation

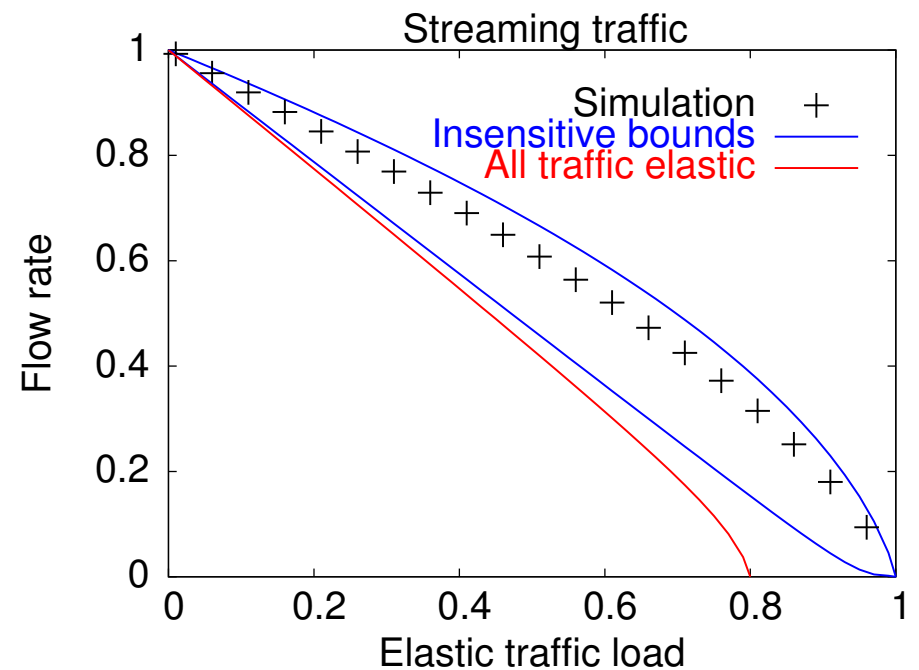
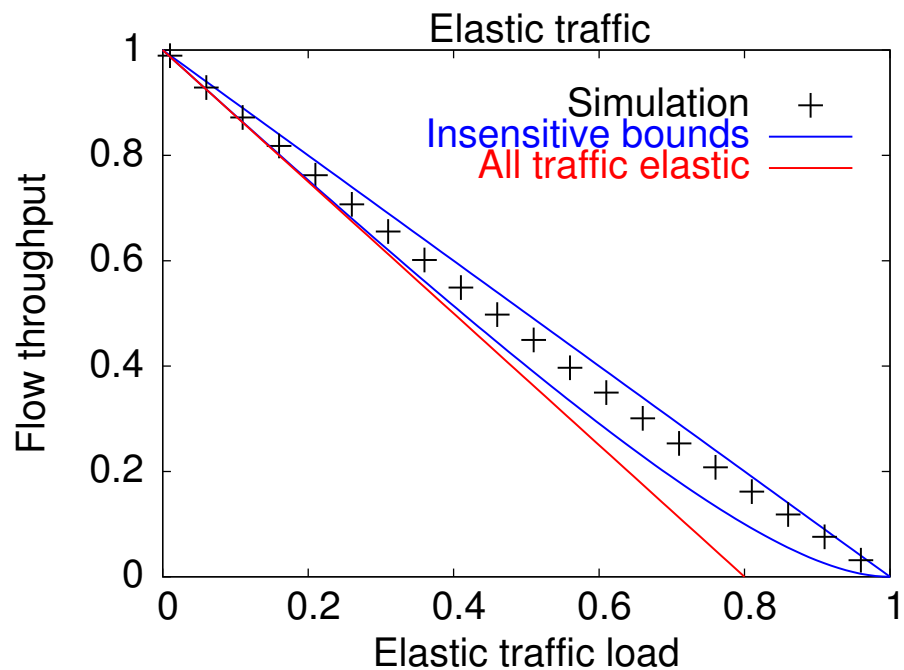
$$\tilde{\phi}_e(x) = \frac{x_e}{x_e + x_s} \mu_e = \phi_e(x), \quad \tilde{\phi}_s(x) = \frac{x_s}{x_e + x_s} \mu_s \leq x_s \mu_s = \phi_s(x)$$

A single bottleneck

- Service rates

$$\phi_e(x) = \frac{x_e}{x_e + x_s} \mu_e, \quad \phi_s(x) = x_s \mu_s$$

- User performance (20% streaming traffic)

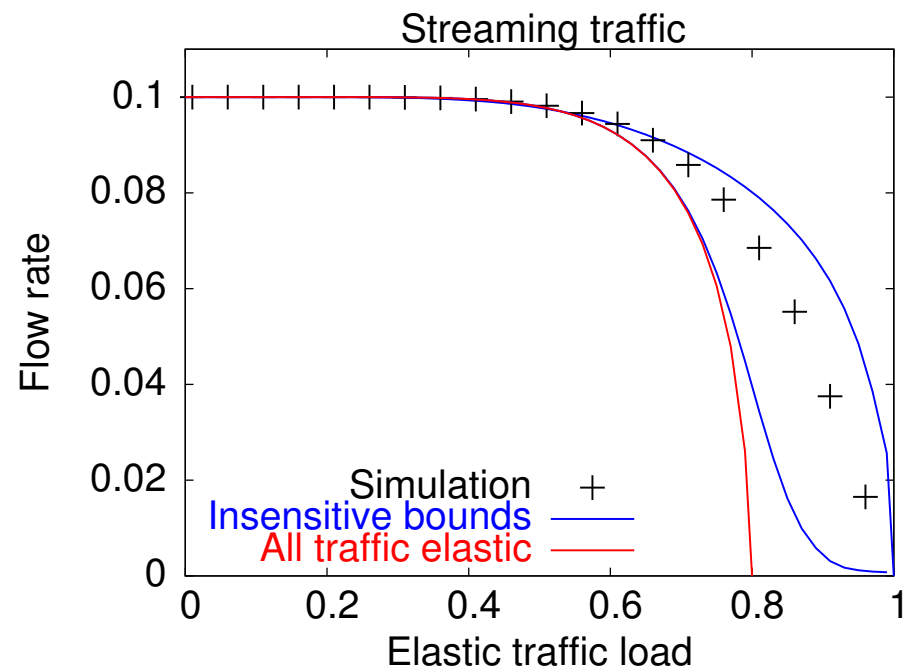
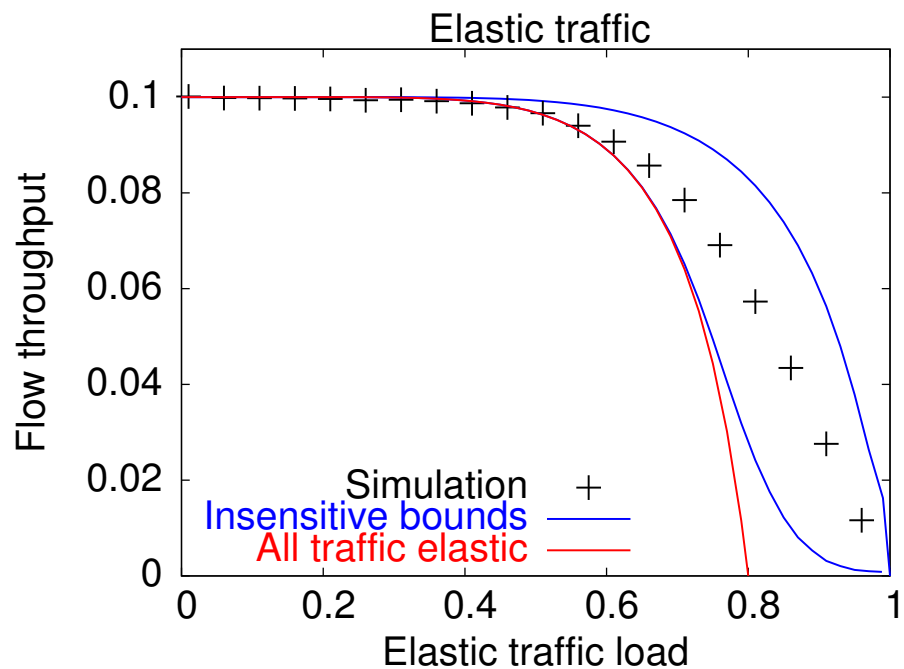


A common rate limit

- Additional per-flow rate limit $a < 1$

$$\phi_e(x) = \min(x_e a, \frac{x_e}{x_e + x_s}) \mu_e, \quad \phi_s(x) = x_s \mu_s$$

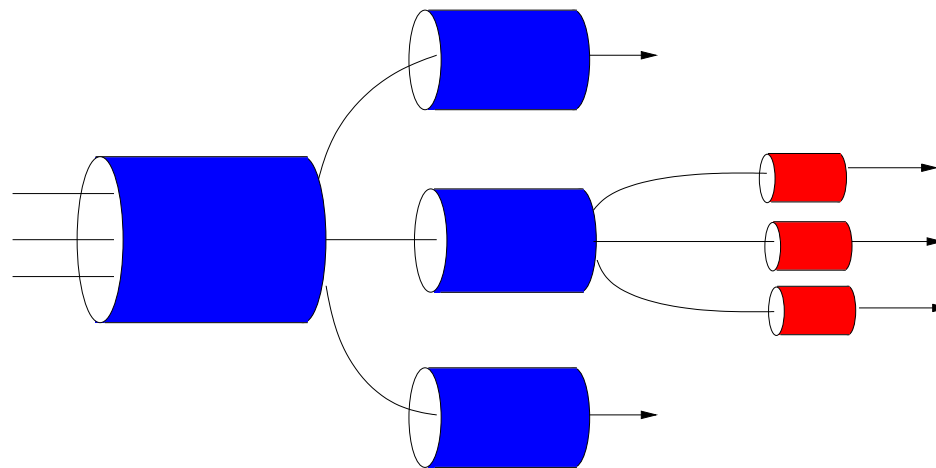
- User performance ($a = 0.1$, 20% streaming traffic)



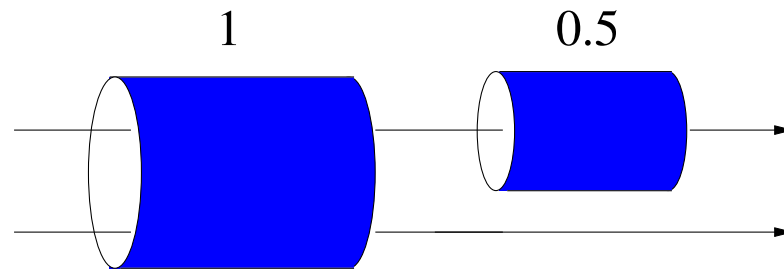
Extension

- A general model
 - multiple rate limits (e.g., different access lines)
 - successive multiplexing stages
- Assume **balanced fair** sharing
 - class- k service rates

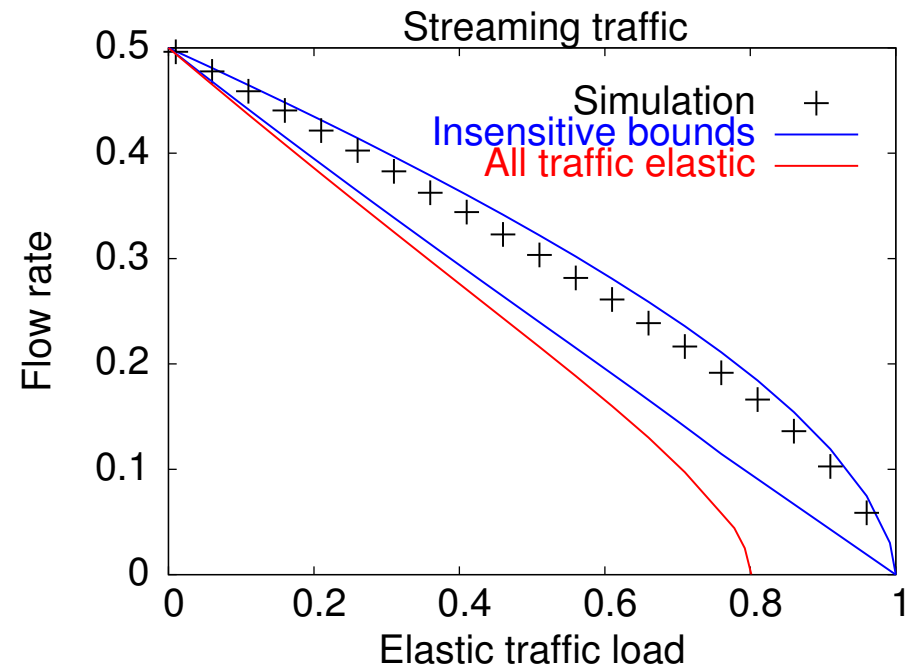
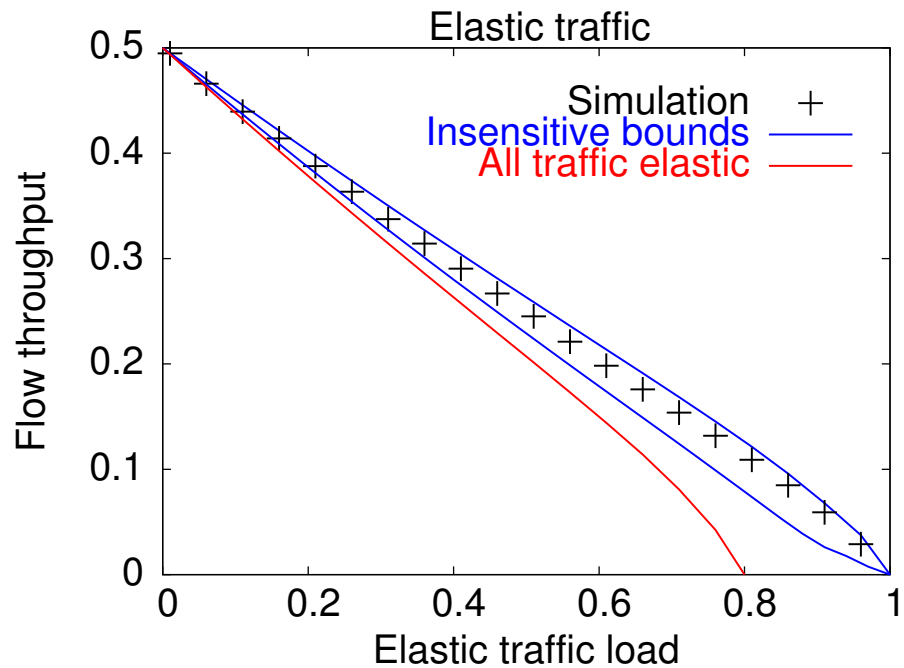
$$\phi_{e,k}(x) = \frac{x_{e,k}}{x_{e,k} + x_{s,k}} \phi_k^{\text{BF}}(x_e + x_s) \mu_{e,k}, \quad \phi_{s,k}(x) = x_{s,k} \mu_{s,k}$$



Two multiplexing stages



- User performance (20% streaming traffic)



Conclusion

- Flow-level modeling
 - the random evolution of the number of flows **drives** the overall network dynamics
 - this **must** be taken into account in performance studies

Conclusion

- Flow-level modeling
 - the random evolution of the number of flows **drives** the overall network dynamics
 - this **must** be taken into account in performance studies
- Integration of elastic and adaptive streaming traffic
 - **explicit** stochastic bounds
 - **insensitive** to the distributions of elastic flow sizes and streaming flow durations
 - valid for multiple rate limits and successive multiplexing stages

Conclusion

- Flow-level modeling
 - the random evolution of the number of flows **drives** the overall network dynamics
 - this **must** be taken into account in performance studies
- Integration of elastic and adaptive streaming traffic
 - **explicit** stochastic bounds
 - **insensitive** to the distributions of elastic flow sizes and streaming flow durations
 - valid for multiple rate limits and successive multiplexing stages