Part III:
Call Admission Control in Integrated Services Networks

Yeali S. Sun
National Taiwan University

Outline

- Introduction
- Two approaches
  - Statistical allocation
  - Non-statistical allocation
- Issues
- Traffic characterization
- Summary

References

- H. G. Perros and K. M. Elsayed, “Call Admission Control Schemes: A Review,” IEEE Communications Magazine, pp. 82-91, November 1996.
- G. de Veciana, G. Kesidis and J. Walrand, “Resource Management in Wide-Area ATM Networks Using Effective Bandwidth,” IEEE JSAC, Vol.13, No.6, pp.1081-1090, August 1995.
- R. Guerin, H. Ahmadi and M. Naghshineh, “ Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks,” IEEE JSAC, Vol. 9, No. 7, September 1991.
- Sugih Jamin, Peter B.Danzig, Scott J. Shenker and Lixia Zhang, “A Measurement-based Admission Control Algorithm for Integrated Services Packet Networks,” IEEE/ACM Transactions on Network, February1997.
- M. Grossglauser and J-C Bolot, “On the Relevance of Long-Range Dependence in Network Traffic,” SIGCOMM’96, pp. 15-24, 1996.
- H. Fowler and W. Leland, “Local Area Network Traffic Characteristics, with Implications for Broadband Network Congestion Management,” IEEE JSAC, 9(7), pp. 1139-1149, September, 1991.
- V. Paxson and S. Floyd, “Wide-Area Traffic: The Failure of Poisson Modeling,” ACM SIGCOMM94, pp. 257-268, 1994.

Introduction

- Call Admission Control (CAC) is to handle the question:

  "Can a network/switch *accept* a *new* connection?"
- Per-connection CAC
- End-to-end CAC, per-hop CAC

- CAC decision is based on:
  - Will the new connection affect the QoS of the connections currently being carried by the node?
  - Can the switch provide the QoS requested by the new connection?

Introduction (cont'd)

- For CBR and VBR services CAC is used as a ***preventive*** scheme in congestion control
  - vs. reactive congestion control
- A preventive congestion control involves both CAC, bandwidth usage enforcement, and policing.
  - For a network providing bandwidth on demand, traffic will need to be *monitored* to verify that users *comply* with their traffic descriptors and *policed* in order to ensure fairness and individual performance.

Two Approaches in CAC

- Non-statistical resource allocation
  - simple
- Statistical resource allocation
  - more difficult to enforce quality of service
  - resource utilization vs. service agreement

Non-Statistical Resource Allocation

- A simple way is to do peak bandwidth allocation
- Suitable for CBR services
  - e.g., PCM-encoded voice, uncompressed video, very-low-bandwidth applications such as telemetry.
- Easy CAC - required bandwidth r_{new} vs. residual bandwidth

$$\sum_{i=1}^{N} r_i + r_{new} \leq C$$

  - where C is link capacity, r_i is bandwidth req. of flow I, N is total number of flows admitted on the link.

Deterministic (ρ, σ) constraint

- Traffic is *regulated with a token bucket at the user-network interface*.
- A token bucket has a constant token arrival rate, ρ, and finite token buffer size, σ.
  - It will limit the output stream to bursts of size σ and an average rate not to exceed ρ.
- Such a stream is said to satisfy a deterministic (ρ, σ) constraint.
- Based on this type of traffic characterization, the network can *reserve* an appropriate size *buffer* and minimum guaranteed *bandwidth.*
- *Deterministic end-to-end delay bounds* are satisfied with no cell loss due to buffer overflow from the output of the leaky bucket to the destination of the connection.

Non-Statistical Allocation (cont'd)

- Disadvantage
  - unless connections transmit at peak rate, the resource may be underutilized
  - over-commit resources for the worst-case scenario.

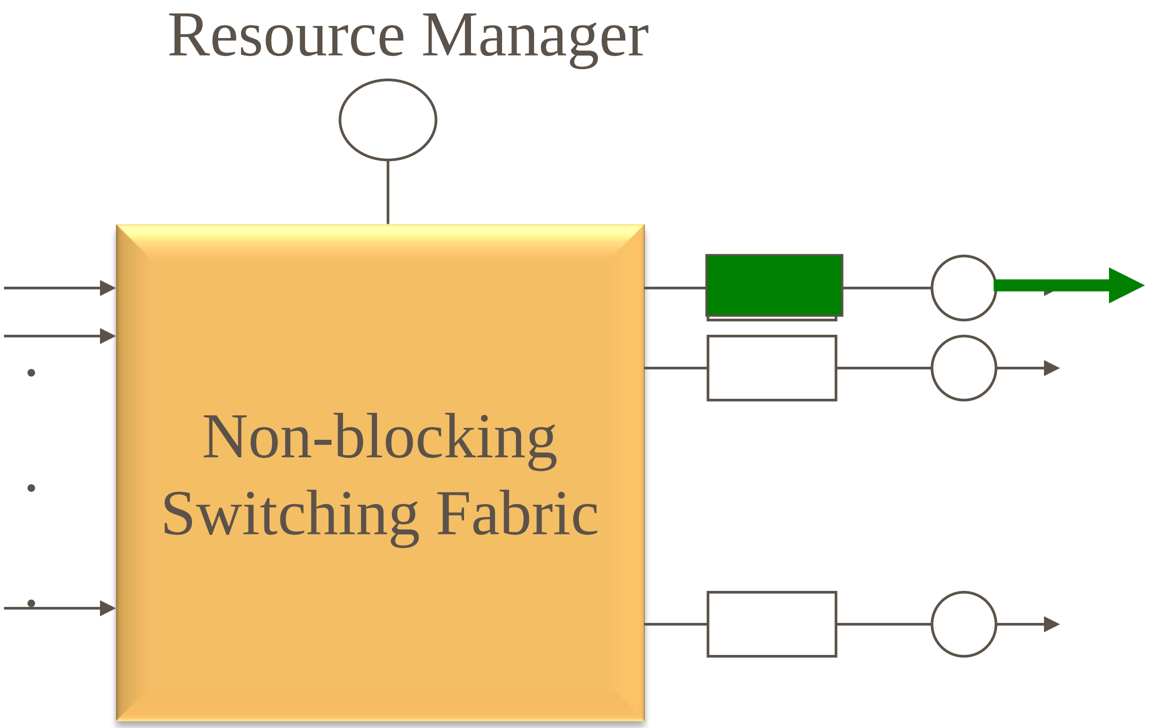
Statistical Allocation

- The goal is to increase resource utilization or efficiency.
- The idea is to take advantage of statistical gain when multiplexing a number of bursty sources on a single link.
- General approach
  - The allocated bandwidth to a connection is ***less*** than the peak rate of the source (i.e. effective bandwidth)

Average_bw_req <= Effective_bw <= Peak_bw_req

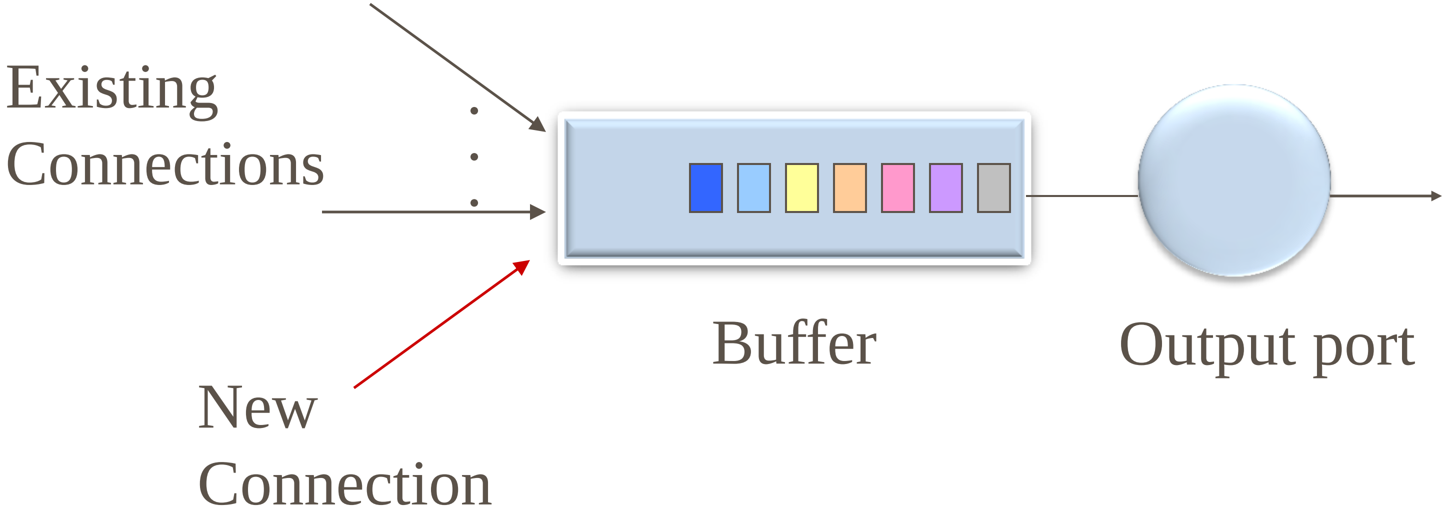
- Total bandwidth allocated may exceed the link capacity (i.e. overbooking).

A switch with output buffering



Typical Traffic Aggregation and Link Sharing

- A traffic multiplexer



Approximate statistical traffic descriptors

- Allocate resources for connections with the statistical nature of the stream of cells.
- The main advantage is to *allow the exploitation of statistical multiplexing* to increase resource utilization.
- Meantime, one still needs to guarantee QoS to individual connections.
- Connections with statistical traffic descriptors are such as ATM ABR traffic – requiring a non-zero minimum service bandwidth and being able to tolerate some cell loss.
- Effective bandwidth
  - a measure of a connection's bandwidth requirement relative to the desired QoS constraint, e.g., delay and /or loss experienced by a connection's cells.

Issues in Statistical Allocation (1/3)

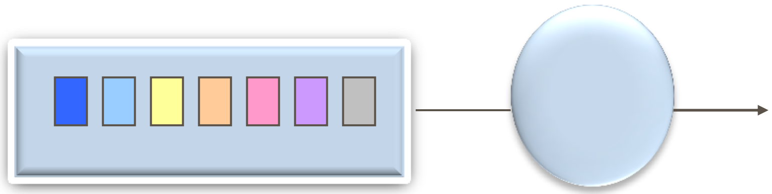
- <u>Difficult</u> to carry out effectively
  - How much one can take advantage of "multiplexing gain" depends on the characteristics of the traffic

- The difficulty is to characterize
  - Individual flow traffic arrival process, especially for the Internet applications.
  - The aggregate behavior
  - Lack of understanding as to how an arrival process is shaped deep in the network

Issues in Statistical Allocation (2/3)

- The "real-time" requirement of CAC decisions
  - Done within no more than *a few seconds*.
  - Requires a *simple* and *accurate* computation
  - May require *complete* knowledge of the entire network resource usage.
  - Must consider
    - new connection characteristics
    - existing network traffic
    - desired QoS

Approximate statistical traffic descriptor (1/6)

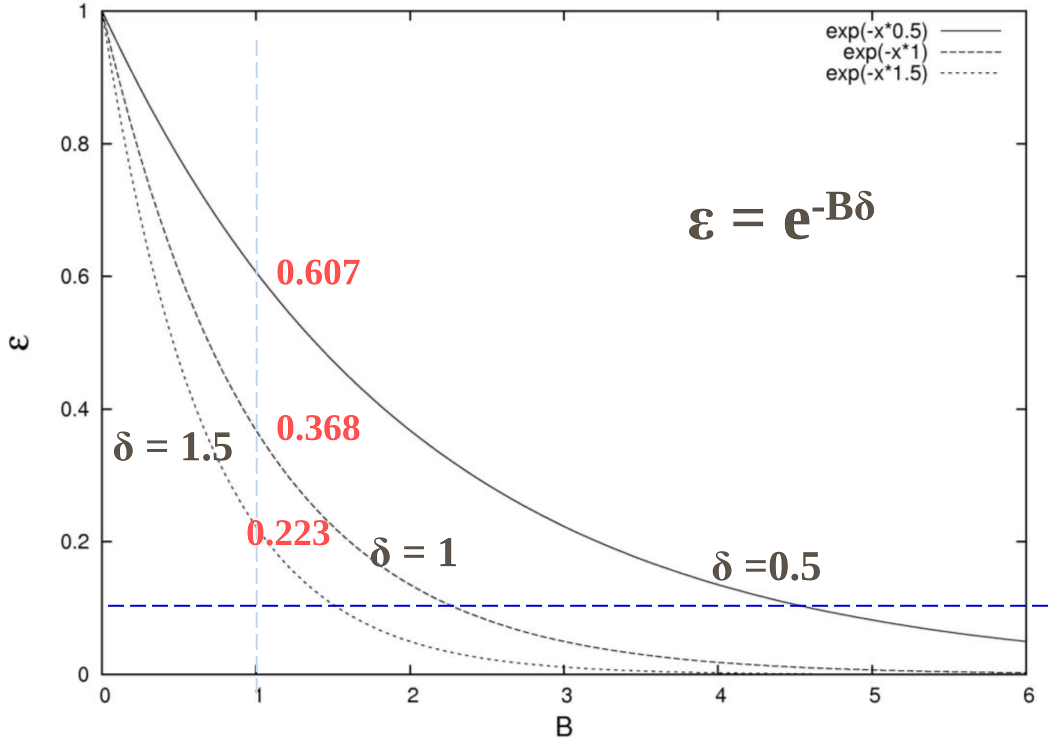


Assume

- a buffered link with capacity c cell/sec
- an ergodic arrival packet stream A(t)
- X denotes the buffer's stationary workload
- QoS goal: *limit the likelihood of large delays* or *ensure that cell loss probabilities at the link are small*, i.e.,

$$P\{X \geq B\} \leq \varepsilon := e^{-B\delta} << 1$$

where δ is the parameter used to determine the stringency of the QoS constraint.



G. de Veciana, G. Kesidis and J. Walrand, "Resource Management in Wide-Area ATM Networks Using Effective Bandwidth," IEEE JSAC, Vol.13, No.6, pp.1081-1090, August 1995.

Approximate statistical traffic descriptor (2/6)

- It has been shown that, for both continuous and discrete-time arrival processes, for all $\delta > 0$ the source effective bandwidth

$$\alpha(\delta) < c \Leftrightarrow \lim_{B\to\infty} \frac{\log P\{X > B\}}{B} \le -\delta$$

- Equivalently,

$$P\{X > B\} = \exp[-B\alpha^{-1}(c) + o(B)]$$

$$where \quad o(B) \; satisfies \; \lim_{B\to\infty} o(B)/B = 0$$

Approximate statistical traffic descriptor (3/6)

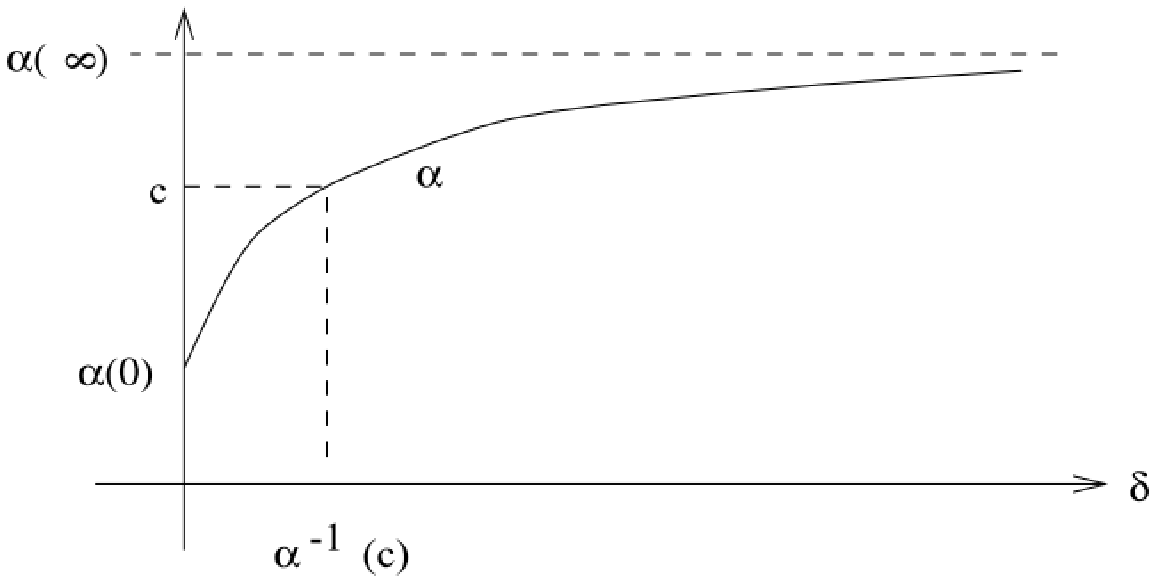
- The source's effective bandwidth is an non-decreasing function in δ with mean rate α(0) and peak rate $\alpha(\infty)$ (*log moment generating function of A(.) and δ)

$$\alpha(\delta) := \frac{1}{\delta} \lim_{t\to\infty} \frac{\log E[e^{\delta A(0,t]}]}{t}$$

  - A(0,t]: the number of cell arrivals to the buffer in the interval of time (0, t]

- The effective bandwidth is the minimum bandwidth required by the connection to accommodate its desired δ-constraint.

Approximate statistical traffic descriptor (4/6)



Approximate statistical traffic descriptor: FIFO buffer (5/6)

- The key relationship needed for resource management is that between the traffic descriptor(s) and the resources necessary to supp ort
- Consider a FIFO buffer with deterministic service rate of c cells/s and arrivals consisting of the superposition of N independent sources with effective bandwidths, α_1, α_2, …, α_N, for the individual packet streams $A_i(0, t]$.

$$\sum_{i=1}^{N} \alpha_i(\delta) < c \Leftrightarrow \lim_{B\to\infty} \frac{\log P\{X > B\}}{B} \le -\delta$$

- That is,

$$\mathrm{P}\{X > B\} = \exp[-BI(c) + o(B)] \quad \text{where} \quad I^{-1}(\delta) := \sum_{i=1}^{N} \alpha_i(\delta).$$

Approximate statistical traffic descriptor: FIFO buffer (6/6)

Key characteristics of this result:

- The additivity of the individual effective bandwidths makes checking whether the QoS constraint is satisfied simple.
- The result holds for a large class of traffic streams, such as Markov modulated fluids, or Markov-modulated Poisson sources, and most reasonable stationary and ergodic traffic models.
- In the case of a shared buffer, the δ-constraint should be interpreted as a performance constraint on the buffer, e.g., cell loss.
- In case that traffic streams are statistically identical, each stream individually experiences this QoS constraint.

Issues in Statistical Allocation (3/3)

- CAC for video sources
  - Video encodings tend to be VBR-encoded.
  - Characterizing the behavior of the output process of an encoder is an open, difficult research problem
    - Compression algorithms
    - Content-dependent
- Common approach – *traffic shaping and receiver buffering*

Bounded delay packet delivery service …

- Target at the support of real-time applications
- Admission control is *mandatory* to regulate network load.
- Works have been focusing on admission control algorithms that compute the worst case theoretical queueing delay to guarantee an absolute delay bound for all packets.
- The network must calculate the worst-case behavior of all the existing flows in addition to the incoming one.
- An assumption is that at service request stage, a flow requesting real-time service must specify its traffic for network in admission control - **(ρ, σ)-regulated** where ρ denotes the mean rate and σ is maximum burst size.

Priori Traffic Characterization

- It is quite difficult to provide accurate and tight statistical models for each individual flow.
  - e.g., the *average bit rate* of a given codec in a teleconference depends on the participant's body movements; it can't possibly be predicted in advance.
- Therefore, *priori* traffic characterizations handed to admission control are often *loose upper bounds*.
- When flows are bursty, guaranteed service usually results in *low* utilization.

Higher network utilization is possible if …

- Weakening the reliability of the delay bound.
- The *probabilistic* service model
  - does NOT provide for the worst-case scenario,
  - it guarantees **a bound ε, on the rate of lost/late packets** based on statistical characterization of traffic.

$$P\{Y \geq C\} \leq \varepsilon$$

- Approach
  - Each flow is allotted an *effective bandwidth* that is larger than its average rate but less than its peak rate.

The Concept of Effective Bandwidth

- A variety of algorithms have been proposed in the literature
  - based on different *approximations* or types of bandwidth allocation schemes

Traffic Characterization: Poisson Process (1/2)

- In the past in telecommunication networks, Poisson processes have been widely used to model *telephone call arrivals*.
- ***Performance modeling*** *and* ***evaluation*** of telecommunication systems were based on the assumption of Poisson arrival processes
  - call arrival process
  - call duration

$$N(k,t)=\frac{(\lambda t)^k}{k!}e^{-\lambda t}$$

$$a(t)=\lambda e^{-\lambda t}$$

Traffic Characterization: Poisson Process (2/2)

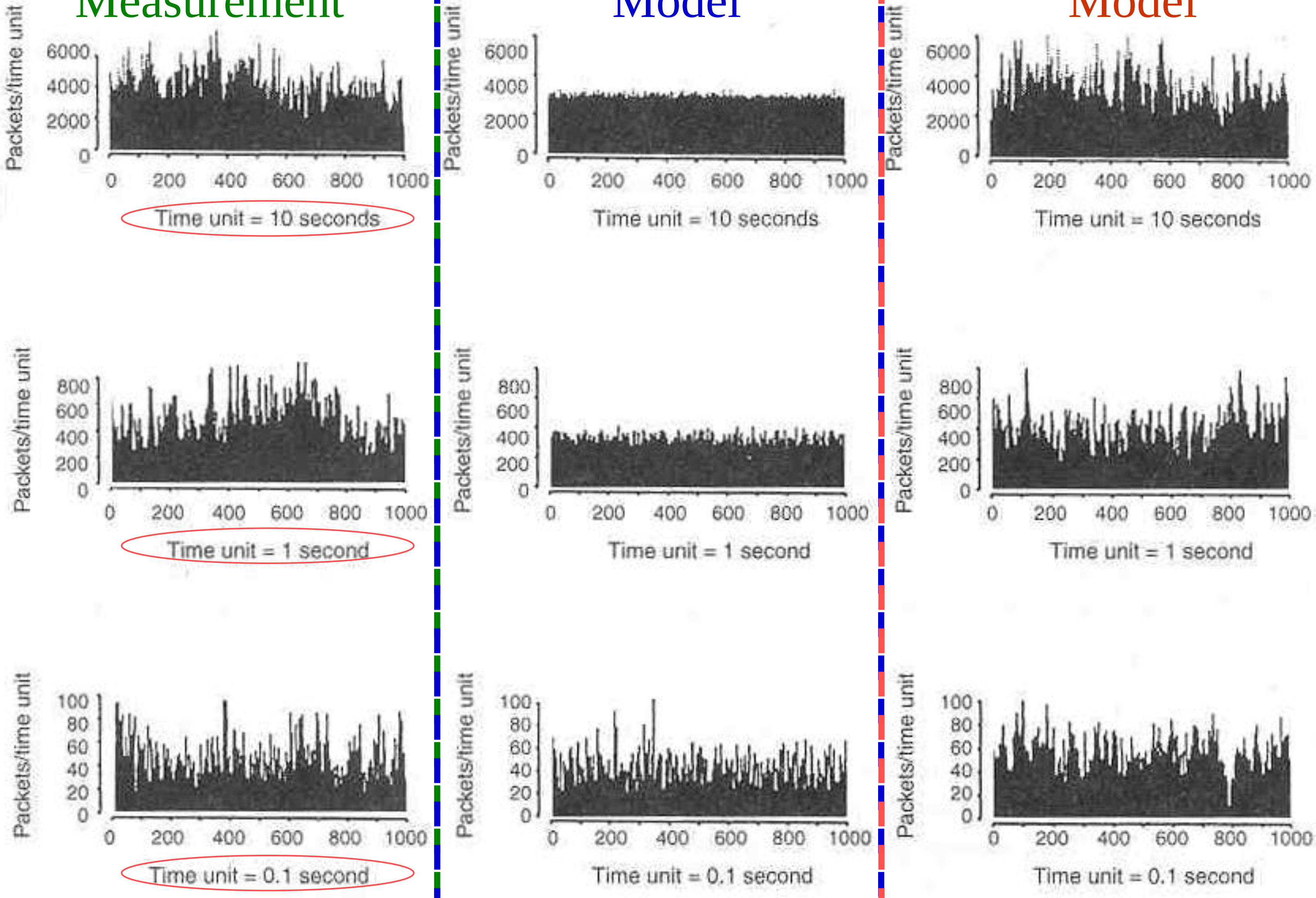
- In *data* communications, data have shown that Poisson process models are also good for modeling such as *user-initiated* TCP session arrivals, such as remote-login (telenet) and file-transfer (ftp)
  - Poisson arrival process
  - Exponential interarrival time distribution

- Poisson process has attractive *theoretical properties* and *analytic simplicity*.
- These models did *not* capture the burstiness present in traffic resulting from applications such as file transfer and packetized encoded video.

Internet Traffic

- Initially (1989), work shows that *LAN traffic* is much better modeled using statistically self-similar processes.
- Later, *more* experimental data have shown that **Internet traffic processes** exhibit properties of *self-similarity* and *long-range dependence (LRD)* (i.e. of correlations over a wide range time scales).
- Fractal - repeating geometric pattern

Actual Measurement | Poisson Model | Self-Similar Model



(a) Actual measurement

(b) Synthetic, Poisson model

(c) Synthetic, self-similar model

A self-similar process: properties

- Self-similar processes have very *different* theoretical properties than Poisson processes.
- Interarrivals preserve burstiness over *many time scales* (self-similarity)
- High degrees of multiplexing do *not* help while Poisson arrival processes are quite limited in the burstiness, especially when multiplexed to a high degree.
- Long-range Dependence (LRD)
  - It describes the rate of decay of statistical dependence.
  - LRD decays *more* slowly than an exponential decay.
- Note that some self-similar processes may exhibit *long-range dependence*.
  - But… not all processes have long-range dependence are self-similar.

The heavy-tailed distribution: definitions (1/2)

- A distribution is heavy-tailed
  if $P[X \geq x] \sim cx^{-\beta}$, as $x \to \infty$, $\beta \geq 0$, for some β and some constant c.
  Or, $P[X \geq x]/(cx^{-\beta})$ tends to 1 as $x \to \infty$.
- This definition includes the *Pareto* and *Weibull* distributions.

- A more general definition of heavy-tailed:
  if the *conditional mean exceedance* (CMEx) of the random variable X is an *increasing* function of x, i.e.,
  $CME_x = E[X - x | X \geq x]$.

The heavy-tailed distribution: discussion (2/2)

Consider X is waiting time,

- For waiting times with a *light-tailed* distribution (e.g., uniform distribution), the conditional mean exceedance is a decreasing function of x.
  -> The *longer* you have waited, the *sooner* you are likely to be done.

- For waiting times with a *medium-tailed* distribution (e.g., exponential distribution (memoryless)),
  -> the expected future waiting time is *independent* of the waiting time so far.

- For waiting times with a *heavy-tailed* distribution
  -> the *longer* you have waited, the *longer* is your expected future waiting time.

The Pareto distribution (1/2)

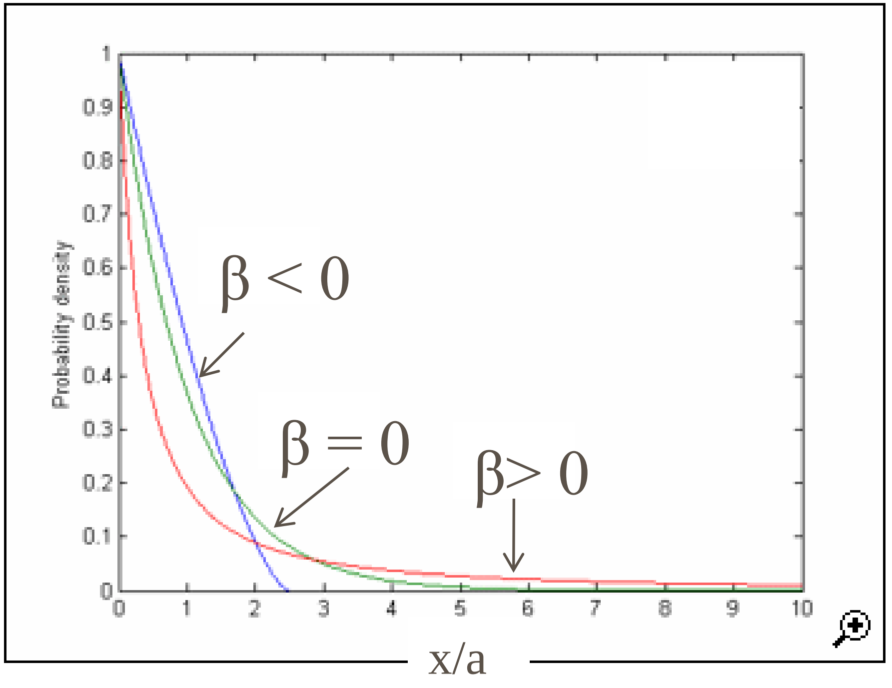
- A heavy-tailed distribution
- Described by two parameters: *shape parameter β* and *scale parameter a*.
- The cumulative distribution function:

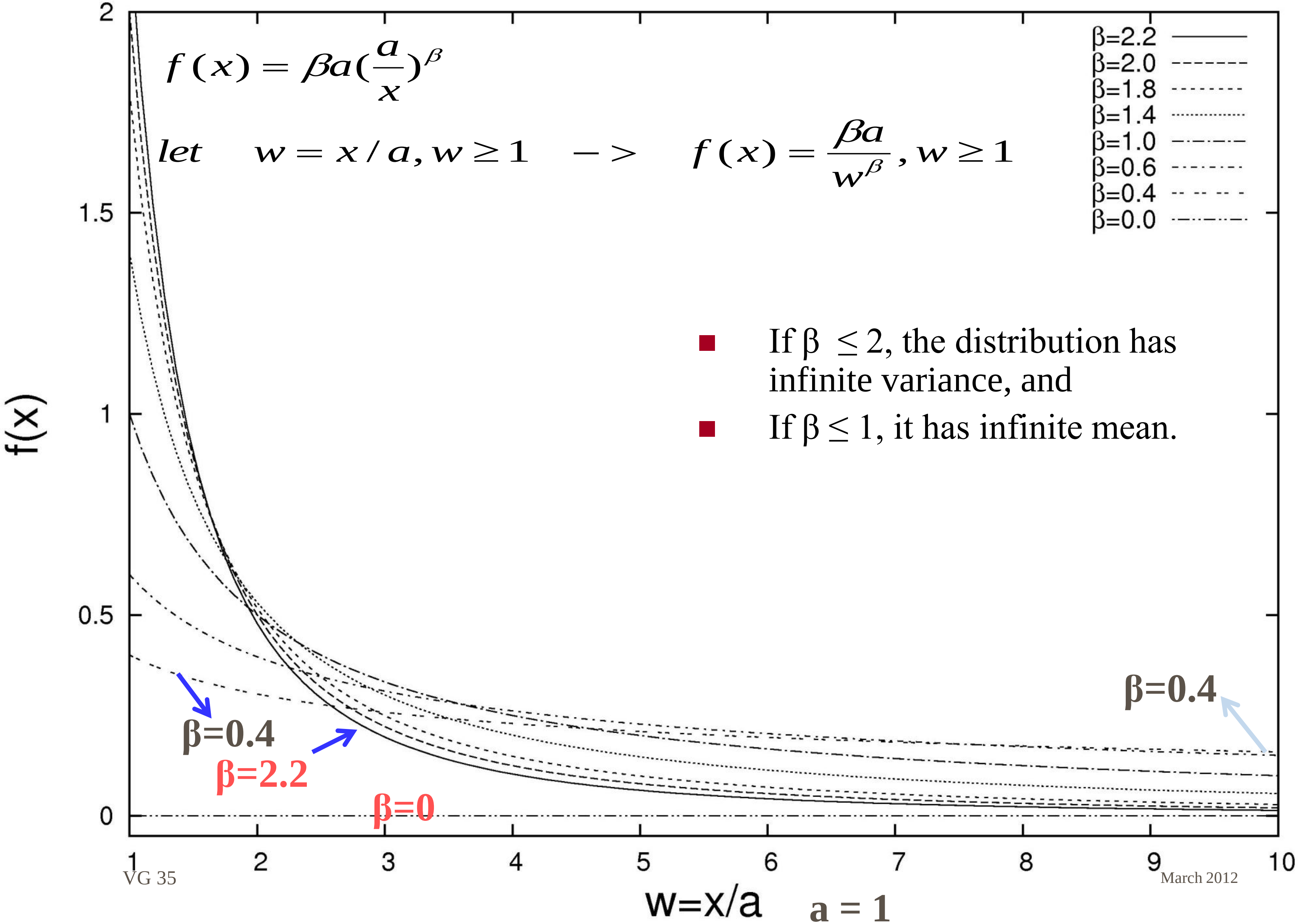
$$\mathbf{F(x) = P\,[X \leq x] = 1 - (a/x)^{\beta}}$$

$$a,\ \beta \geq 0,\quad x \geq a$$

- For the Pareto distribution with $\beta > 1$ (with finite mean), the conditional mean exceedance is a linear function of x. i.e. $CME_x = x/(\beta - 1)$.

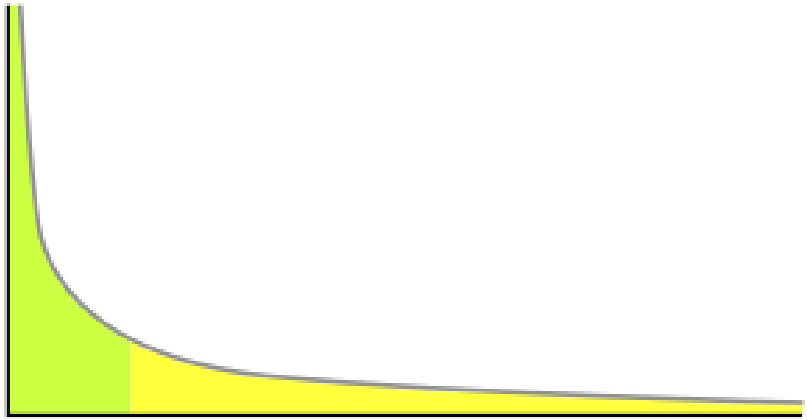
$$f(x) = \frac{\beta a^{\beta+1}}{x^{\beta}}$$





$f(x) = \beta a (\frac{a}{x})^{\beta}$
$let \quad w = x/a, w \geq 1 \quad -> \quad f(x) = \frac{\beta a}{w^{\beta}}, w \geq 1$
β=2.2
β=2.0
β=1.8
β=1.4
β=1.0
β=0.6
β=0.4
β=0.0
If β ≤ 2, the distribution has infinite variance, and
If β ≤ 1, it has infinite mean.
β=0.4
β=0.4
β=2.2
β=0
f(x)
2
1.5
1
0.5
0
1
2
3
4
5
6
7
8
9
10
w=x/a
a = 1

Power-law distribution



- To the right is the long tail
- To the left are the few that dominate (also known as the 80-20 rule)

The Pareto distribution (2/2)

- In communications, heavy-tailed distributions have been used to model *telephone call holding times* [DMRW94] and frame sizes for *variable-bit-rate video* [GW94].
  - Traditionally, telephone call holding times (CHTs) have been modeled using exponential distributions (e.g., Erlang 1918).
  - Such an approximation seriously underestimates the actual numbers of very long calls (e.g., data calls that last for many hours).
- [LO86] found that a Pareto distribution with $1.05 < \beta < 1.25$ is a good model for the amount of *CPU time* consumed by an arbitrary process.

- [DMRW94] D. Duffy, A. McIntosh, M. Rosenstein, and W. Willinger, "Statistical Analysis of CCSN/SS7 Traffic Data from Working CCS Subnetworks," IEEE JSAC, 12(3), pp. 544-551, April, 1994.
- [GW 94] M. Garrett and W. Willinger, "Analysis, Modeling and Generation of Self-Similar VBR Video Traffic," SIGCOMM '94, pp. 269-280, September, 1994.
- [LO86] W. Leland and T. Ott, "Load-balancing Heuristics and Process Behavior," PERFORMANCE '86 and ACM SIGMETRICS 1986 Joint conference on Computer Performance Modelling, Measurement and Evaluation, pp. 54-69, May 1986.

What are the challenges of self-similar traffic on resource allocation, congestion control, and network/application performance?

Implications of Long-range Dependence in Network Traffic Control and Network Capacity Planning

- Long-range dependence refers to *burstiness across different time scales*.
- Measurement data have shown that *TCP traffic* has the long-range dependence property.
- Modeling TCP traffic using Poisson or other models with no long-range dependence will result in simulations and analyses that *significantly underestimate* performance measures such as *average packet delay* or *maximum queue size*.

Implication of Self-Similar Traffic on Network Congestion Control (1/2)

- Self-similar traffic "spikes" (which cause losses) ride on longer-term "ripples".
- Congested periods can be quite long with *losses that are heavily concentrated*.
- Research results show that
  - *linear* increases in buffer size, in contrast to Poisson traffic models, do *NOT* result in large decreases in packet drop rates;
  - a *slight increase* in the number of active connections can result in *a large increase* in the *packet loss rate*.

Implication of Self-Similar Traffic on Network Congestion Control (2/2)

- Because the level of busy period traffic is *not predictable*, it would be *difficult* to efficiently *size networks* to reduce congestion adequately.
- It makes congestion control even more difficult!

Implication of Long-range Dependence (LRD) Property on Traffic Performance

- Consider a link with *priority scheduling* between *classes of traffic*
- The higher-priority class has *no* enforced bandwidth limitations.
  - e.g., for interactive traffic such as TELNET might be given priority over bulk-data traffic such as FTP.
- If the higher-priority class *has* LRD and a high degree of variability over long time scales
- The bursts from the higher-priority traffic could *starve* the lower- priority traffic for long periods of time.

Modeling of sources with self-similarity property

- A lot of works in different areas, studied trace records and found that the same phenomenon, i.e., self-similarity property.
- A number of works focus on fitting a particular model to the observed distribution.
- Many results are mathematically complex and are not practically feasible!

Measurement-based Admission Control

Introduction (1/2)

- Guaranteed bounded delay packet delivery service
  - When a flow requests real-time service, it must characterize its traffic so that the network can make its admission control decision.
  - Typically, sources are described by either *peak and average rates* or *a filter like a token bucket*, *i.e.*, *(σ, ρ)-regulated*.
- Admission control algorithms for guaranteed service use the a *priori* characterizations of sources to calculate the worst-case behavior of *all the existing flows* in addition to the *incoming* one.

Introduction (2/2)

- Network utilization under this model is usually acceptable when flows are smooth.
- When flows are bursty, guaranteed service may result in low utilization.
- To achieve *higher network utilization*, one may need to weaken the reliability of the delay bound, e.g., the probabilistic service [*].
  - It guarantees a bound on the rate of lost/late packets based on statistical characterization of traffic.

Motivation

- Many real-time applications developed for packet-switched networks *adapt* to actual packet delays.
- They can *tolerate* occasional delay bound violations; they do not need an absolutely reliable bound.
- They are called *delay-tolerant* applications.

Predictive Service

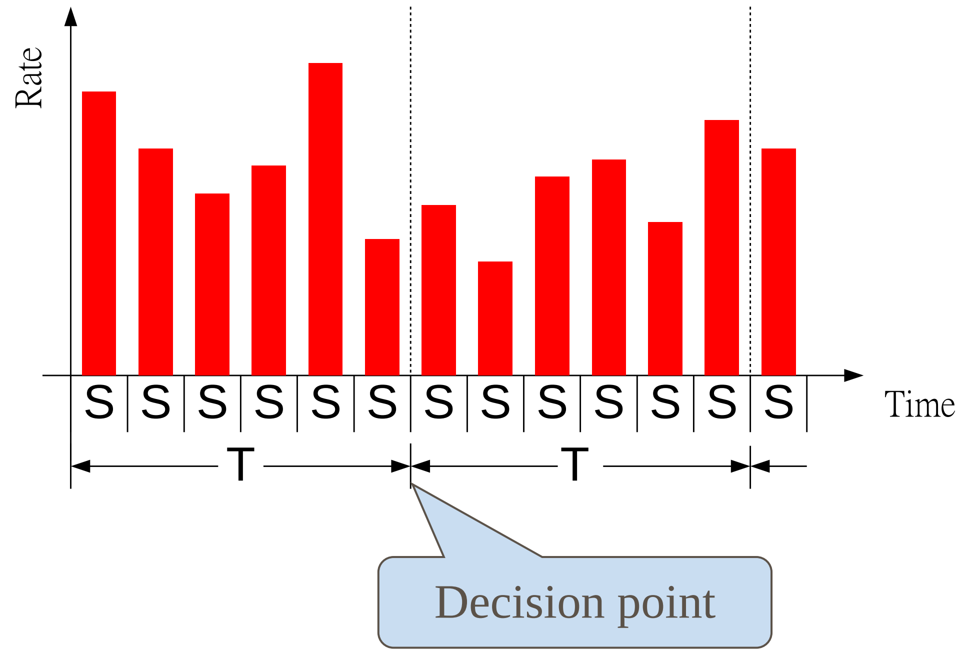
- The goal is to offer a fairly, but not absolutely, reliable bound on packet delivery times.
- Note that it does *not* specify an acceptable level of delay violations.
- The advantage is that it gives admission control a great deal more flexibility.

Measurement-based Admission Control: approach

- Target for predictive service and other more relaxed service commitments.
- The sources are characterized by token bucket filters *at admission time*.
- The behavior of existing flows is determined by *measurement* rather than by a priori characterizations.

The Measurement-based Admission Control: details (1/3)

- Measure the "characteristics" of **aggregated** behavior of **existing** flows at a queueing point
- Measurement process
  - e.g., *measurement duration, sampling interval, memory window size*, etc.



- *T, measurement window*
- *S, sampling interval*

The Measurement-based Admission Control: details (2/3)

- Use performance **prediction mechanisms** to complement current-state measurement
  - the sensitivity of the input traffic dynamics to the changes of the queue size.
- Replace the worst-case parameters with **measured quantities**.
- Use admission control algorithm at each switch to enforce the queueing delay bound at the switch.
- Leave the satisfaction of end-to-end delay requirements to the end systems.

The Measurement-based Admission Control: details (3/3)

- Sources requesting service must specify the worst-case behavior of their flow.
  - Use token bucket filter to assure traffic conformance.
- Use some *reservation protocol* to allow end systems to communicate their resource requirements to the network.
- Note that considering only recent traffic could be easily mislead, following a long period of fairly low traffic rates.

Summary

- CAC is used to decide whether or not a network/switch ***can*** *accept* a *new* connection
- CAC is often used for CBR and VBR services as a *preventive* scheme in congestion control.
- CAC is hard because it is hard to characterize individual traffic sources as well as traffic aggregates.
  - Self-similarity, long-range dependency
  - Theoretic approach
  - Measurement + prediction approaches