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
Introduction

- Call Admission Control (CAC) is to handle the question of whether or not a network/switch can 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
- A preventive congestion control involves both **CAC** and **bandwidth enforcement**.
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 $r_{\text{new}}$ vs. residual bandwidth

\[
\sum_{i=1}^{N} r_i + r_{\text{new}} \leq C
\]

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

\[
\text{Average}_\text{bw}_\text{req} \leq \text{Effective}_\text{bw} \leq \text{Peak}_\text{bw}_\text{req}
\]

- 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

Diagram:

- Existing Connections
- New Connection
- Buffer
- Output port
Issues in Statistical Allocation (1/3)

- **Difficult** 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
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 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 - (\(\sigma, \rho\))-regulated where \(\sigma\) is maximum burst size, \(\rho\) : mean rate.
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.
Traffic Shaping and Receiver Buffering

- Sources are described by either
  - peak and average rates or
  - using a token bucket as a traffic shaper (average rate and maximum burst size)

- Receiver buffer to accommodate transient network delay and jitter.
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 $\varepsilon$, on the rate of lost/late packets based on statistical characterization of traffic.

- 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

- Performance models of telecommunication systems were based on the assumption of Poisson arrival processes
  - call arrival process
  - call duration

- Computer Networks
  - Poisson arrival process
  - Exponential interarrival time distribution

\[ N(k, t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \]
\[ a(t) = \lambda e^{-\lambda t} \]
Traffic Characterization (cont’d)

- These models did not capture the burstiness present in traffic resulting from applications such as file transfer and packetized encoded video.

- Experimental data has 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).
Poisson processes

- In the past in telecommunication networks, Poisson processes have been widely used to model *telephone call arrivals*.
- 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 process has attractive *theoretical properties* and *analytic simplicity*. 
Internet Traffic

- Works have shown that LAN traffic is much better modeled using statistically self-similar processes.
- Self-similar processes have very different theoretical properties than Poisson processes.
  - Poisson model underestimates the packet arrivals generated by the user side of a TELNET connection;
  - interarrivals preserves 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.
Self Similarity: Long-range Dependence (LRD)

- LRD describes the rate of decay of statistical dependence.
- LRD decays *more* slowly than an exponential decay.
- 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

- A distribution is heavy-tailed if
  \[ P [X \geq x] \sim cx^{-\beta}, \quad \text{as } x \to \infty, \quad \beta \geq 0 \]
  for some \( \beta \) and some constant \( c \),
  the ratio \( P [X \geq x]/(cx^{-\beta}) \) tends to 1 as \( x \to \infty \).

- This definition includes the Pareto and Weibull distributions.

- A more general definition of heavy-tailed is that if the conditional mean exceedance (CMEx) of the random variable \( X \) is an increasing function of \( x \), i.e.,
  \[ \text{CMEx} = E[X - x|X \geq x]. \]
The Pareto distribution

- Described by two parameters: shape parameter $\beta$ and location parameter $a$.
- It has the cumulative distribution function [HK80]:

$$F(x) = P[X \leq x] = 1 - \left(\frac{a}{x}\right)^\beta$$

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

- and probability density function:

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

- If $\beta \leq 2$, the distribution has infinit variance, and
- if $\beta \leq 1$, it has infinite mean.
The Pareto distribution

- Also referred to as the power-law distribution, the double exponential distribution, and the hyperbolic distribution.

- In communications, heavy-tailed distributions have been used to model telephone call holding times [DMRW94] and frame sizes for variable-bit-rate video [GW94].

- [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.
The Heavy-tailed Distribution: discussion

- For waiting times with a light-tailed distribution such as the 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 such as the 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.

- For the Pareto distribution with \( \beta > 1 \) (with finite mean), the conditional mean exceedance is a linear function of \( x \). i.e. \( \text{CMEx} = \frac{x}{(\beta - 1)} \).
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

- Long-range dependence refers to burstiness across different time scales.
- 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, in contrast to Poisson models, “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.
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.

- 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
  - e.g., *measurement duration, sampling interval, memory window size*, etc.
- 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 (2/3)

The Measurement Process

- $T$, measurement window
- $S$, sampling interval
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.

- If 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.
References