The University of Sydney
AUSTRALIA

School of Electrical and Information Engineering

Advanced Communication Networks

Chapter 12

ATM Traffic and Congestion Control

Based on chapter 17 of Stallings ISDN-4e book

Abbas Jamalipour

12.1 Requirements for ATM Traffic and Congestion Control

- Similar to frame relay network, traffic and congestion control techniques are required to ensure being within the capacity of the network.
- Controlling the congestion in ATM network are more difficult since there are limited number of overhead bits available.
- Traffic and congestion control are still open issues in ATM networks.
- Packet switching and frame relay networks carry non-real-time data traffic, thus typically is bursty in nature, and as a result
  - The network does not need to replicate the exact timing pattern of incoming traffic at the exit node.
  - Therefore, simple statistical multiplexing can can be used to accommodate multiple logical connections over the physical user-network interface (UNI).
  - The UNI need only be designed for a capacity somewhat greater than the sum of the average data rates for all connections.
- Congestion control mechanism in packet switching and frame relay networks are inadequate for ATM networks because:
  - The majority of traffic is not amenable to flow control (e.g. voice/video traffic).
  - Feedback is slow due to the drastically reduced cell transmission time compared to propagation delays across the network.
  - ATM networks typically support a wide range of applications requiring capacity ranging from a few kbps to several hundred Mbps.
  - Applications on ATM networks may generate very difficult traffic patterns (e.g. constant-bit-rate (CBR) versus variable-bit-rate (VBR)).
  - Different applications on ATM networks require different network services (e.g., delay-sensitive services for voice/video, and loss-sensitive for data).
  - The very high speeds in switching and transmission make ATM networks more volatile in terms of congestion and traffic control (e.g., a scheme that relies heavily on reacting to changing conditions will produce extreme and wasteful fluctuations in routing policy and flow control).
- A key issue that relates to the above difficulties is *cell delay variation.*

12.2 Cell-Delay Variation

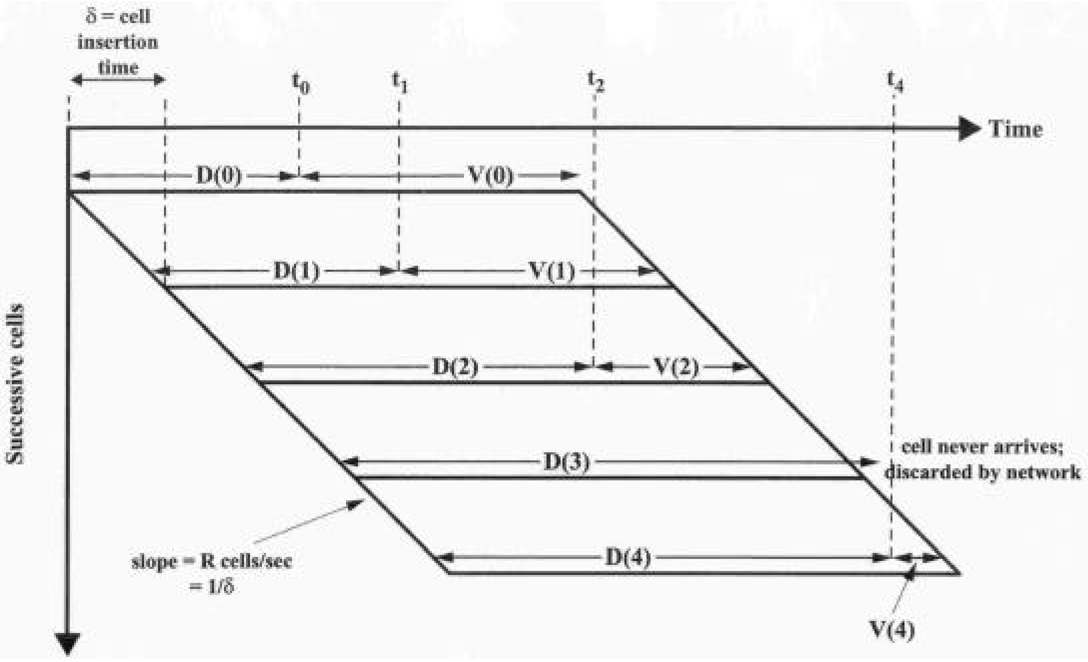
- For an ATM network, voice and video signals can be digitized and transmitted as a stream of cells → *short delay is required*
- The rate of delivery of cells to the destination user must be constant.
- The destination user might cope with variation in the delay of cells as they transmit from source user to destination user to some degree.
- A procedure for achieving constant bit rate
  - $D(i)$: end-to-end delay experienced by the i^{th} cell (unknown to destination)
  - after arriving the first cell at $t(0)$, the target user delays the cell by $V(0)$
    - $V(0)$ is an estimate of the amount of cell delay that this application can tolerate
  - subsequent cells are delayed so that all cell delivered at a constant rate of R
  - time between delivery of cells to the target application: $\delta=1/R$
  - to achieve a constant rate, $V(1)$ and other cells should be

$$t(1) + V(1) = t(0) + V(0) + \delta \quad \Rightarrow \quad V(1) = V(0) - [t(1) - (t(0) + \delta)]$$

or in general, $$V(i) = V(0) - [t(i) - (t(0) + i \times \delta)]$$

or $$\boxed{V(i) = V(i - 1) - [t(i) - (t(i - 1) + i \times \delta)]}$$

  - if the computed $V(i)$ is negative, then that cell is discarded
  - $V(0)$ is a function of the anticipated cell-delay variation. To minimize this delay, the subscriber may request a minimal cell-delay variation from network.



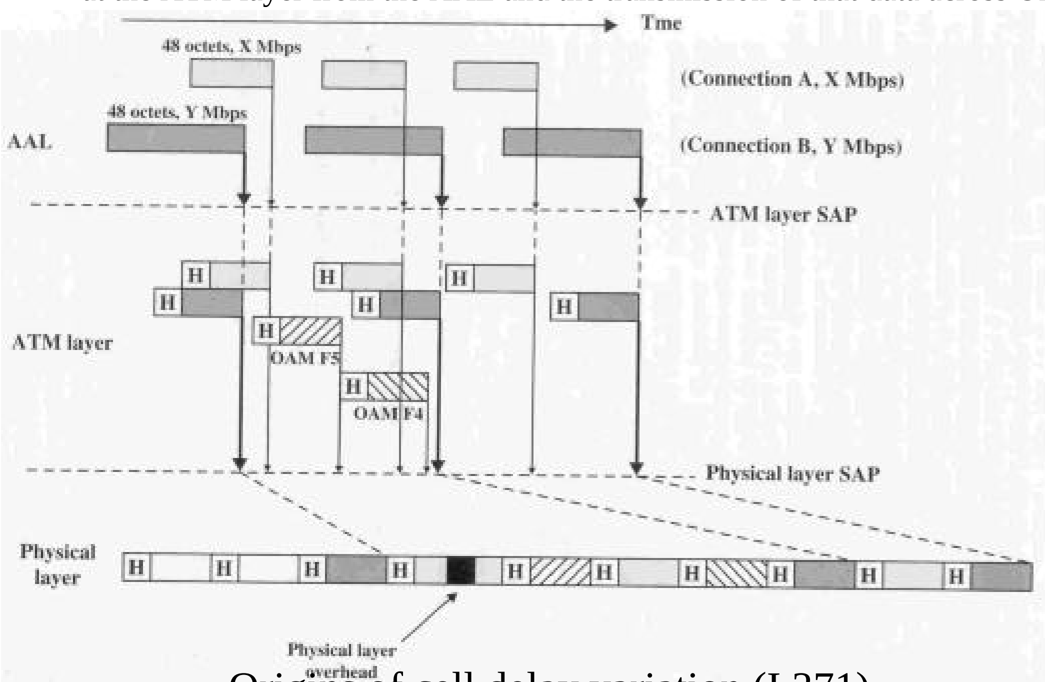
Network Contribution to Cell-Delay Variation

- Cell-delay variation is also due to events within the network
  - considerable in PSN due to queuing delay at each intermediate node
  - less in frame relay networks due to queuing delay at each node
  - minimal in the case of ATM networks because
    - Cells are fixed-size with fixed-header format and no flow & error control
    - ATM switches have extremely high throughput, low processing delay
  - Therefore the only factor in noticeable cell-delay variation within the network is the congestion.

Cell-Delay Variation at the UNI

- Even if applications generate data at CBR, cell delay variation can occur at source due to the processing that takes place at the three ATM layers
  - ATM connections A and B support user data rates of X and Y Mbps, respectiv.
  - Then the time required to generate a 48-byte block of data in μsec is

    connection A: 48*8/X   connection B: 48*8/Y
  - The cells must be interleaved and delivered to physical layer → delay
  - None of such delays can predicated and non follow any repetitive pattern
  - Thus, there iis a random element to the time interval between reception of data at the ATM layer from the AAL and the transmission of that data across UNI.



Origins of cell delay variation (I.371)

12.3 ATM Service Categories

- ATM is intended to carry different type of traffic simultaneously incl. real-time flow (e.g., voice/video) and bursty TCP flows.
- Service categories defined by the ATM Forum:
  - **Real-Time Services**
    - Constant Bit Rate (CBR)
    - Real-Time Variable Bit Rate (rt-VBR)
  - **Non-Real-Time Services**
    - Non-Real-Time Variable Bit Rate (nrt-VBR)
    - Available Bit Rate (ABR)
    - Unspecified Bit Rate (UBR)
- **Real-Time Services** concern about the amount of delay and the variability of delay (*jitter*). These applications typically involve a flow of information to a user that is intended to reproduce that flow at a source (e.g., voice or audio transmission).
- **Non-Real-Time Services** are intended for applications that have bursty traffic characteristics and do not have tight constraints on delay and delay variation (more flexibility for the network to handle traffic and use of statistical multiplexing).

Constant Bit Rate (CBR)

- used by applications that require a fixed data rate that is continuously available during the connection lifetime
- used commonly for uncompressed audio and video information
  - videoconferencing
  - interactive audio (e.g. , telephony)
  - audio/video distribution (e.g., TV, distance learning, pay-per-view)
  - audio/video retrieval (e.g., video-on-demand, audio library)

Real-Time Variable Bit Rate (rt-VBR)

- intended for time-sensitive applications; i.e. those requiring tightly constraint delay and delay variation
- data rate varies with time (different from CBR)

Non-Real-Time Variable Bit Rate (nrt-VBR)

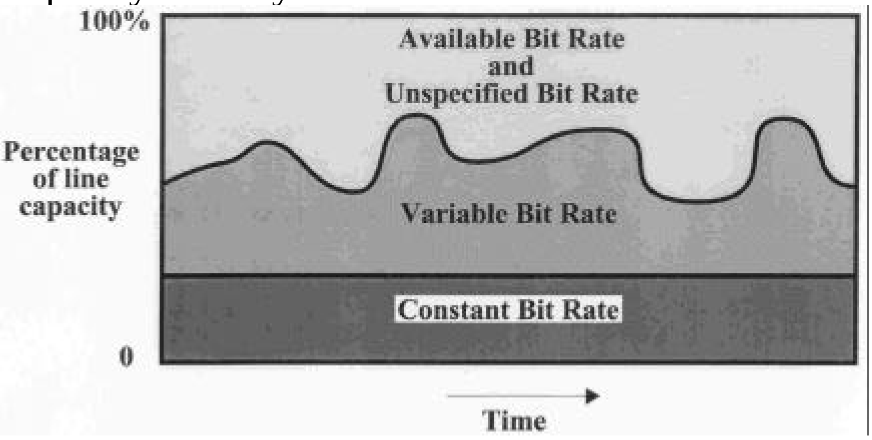
- can be used for data transfers that have critical response time requirements (e.g., airline reservations, banking transactions)
- The end system specifies a peak cell rate, a sustainable or average cell rate, and a measure of how bursty or clumped the cells may be, with which the network can allocate resources.

Unspecified Bit Rate (UBR)

- suitable for applications that can tolerate variable delays and some cell losses (such as TCP-based traffic).
- Cells are forwarded on a FIFO basis using the "left capacity"
- No initial commitment is made to UBR source and no feedback concerning congestion is provided: ***Best-Effort Service***

Available Bit Rate (ABR)

- defined to improve service provided to bursty sources
- specifies a peak cell rate (PCR) and a minimum cell rate (MCR)
- network allocates at least MCR to an ABR source
- unused capacity is shared fairly among all ABR and then UBR sources
- ABR mechanism uses explicit feedback to sources that assure that capacity is fairly allocated.



12.4 Traffic and Congestion Control Framework

- Objectives of ATM layer traffic and congestion control (I.371)
  - ATM layer traffic and congestion control should support a set of ATM layer QoS classes for all network services.
  - ATM layer traffic and congestion control should not rely on AAL protocols that are network-service specific, nor on higher-layer protocols that are application specific.
  - The design of an optimum set of ATM layer traffic controls and congestion controls should minimize network and end-system complexity while maximizing network utilization.
- ITU-T defined a collection of traffic and congestion control functions

Response time	Traffic control functions	Congestion control functions
Long Term	• Network resource management	
Connection Duration	• Connection admission control	
Round-trip Propagation Time	• Fast resource management	• Explicit notification
Cell Insertion Time	• Usage parameter control • Priority control	• Selective cell discarding

- Four level of timing considered
  - **Cell insertion time**: At this level, functions react immediately to cells as they are transmitted.
  - **Round-trip propagation time**: At this level, the network responds within the lifetime of a cell in the network.
  - **Connection time**: At this level, the network determines whether a new connection at a given QoS can be accommodated and what performance levels will be agreed to.
  - **Long term**: These are controls that affect more than one ATM connection and are established for long-term use.
- The essence of the traffic-control strategy is based on
  - determining whether a given new ATM connection can be accommodated
  - agreeing with the subscriber on the performance parameters that will be supported

12.5 Traffic Control

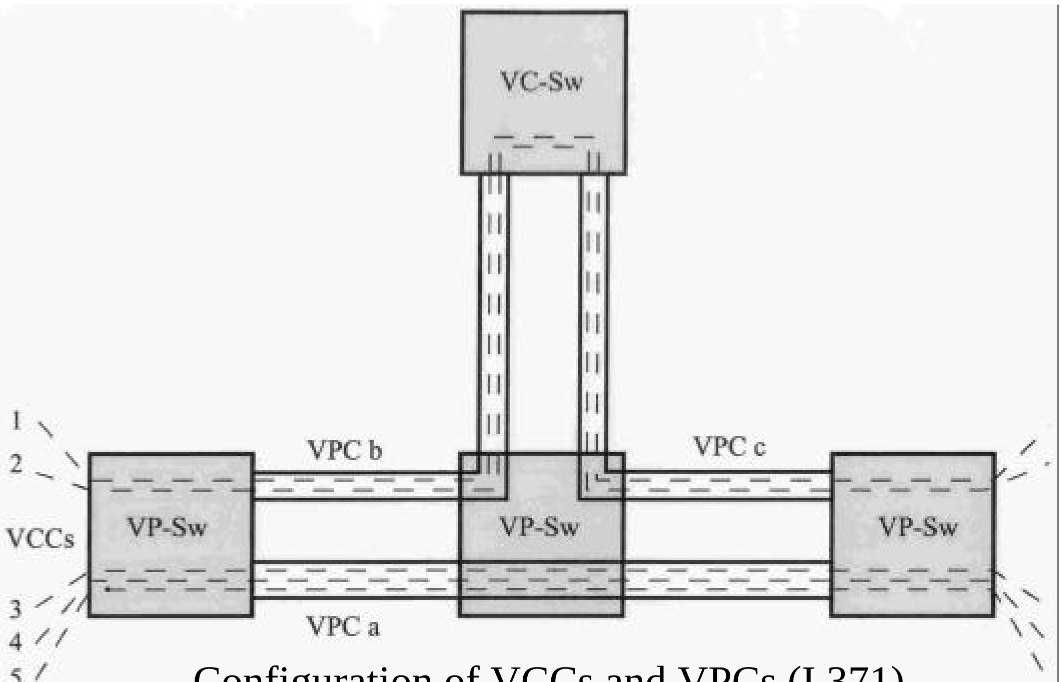
- Traffic control functions defined to maintain QoS
  - Network resource management
  - Connection admission control
  - Usage parameter control
  - Priority control
  - Fast resource management

Network Resource Management

- The essential concept
  - to allocate network resources in such a way as to separate traffic flows according to service characteristics, e.g. use of virtual paths
- VPC provides a convenient means of grouping similar VCCs
- Three cases to be consider
  - **User-to-user application:** VPC extends between a pair of UNIs. Network has no knowledge of the QoS of the individual VCCs within a VPC. It is the user's responsibility to assure that.
  - **User-to-network application:** The VPC extends between a UNI and a network node. Network is aware of the QoS of the VCCs within the VPC and has to accommodate them.
  - **Network-to-network application:** The VPC extends between two network nodes. Again the network is aware of the QoS of the VCCs within the VPC and has to accommodate them.
- Primary QoS parameters in network resource management
  - cell loss ratio
  - cell transfer delay
  - cell delay variation
- The performance of each VPC depends on the capacity of that VPC and the traffic characteristics of the VCCs contained within the VPC.
- The performance of each VCC-related function depends on the switching/ processing speed at the node and on the relative priority for each cell.

An Example

- In the figure below, VCCs 1 and 2 experience a performance that depends on VPCs *b* and *c* and how these VCCs are handled by intermediate nodes.
- The situation is different for VCCs 3 and 4.
- If all of the VCCs within a VPC are handled similarly, then they should experience similar expected network performance.
- When different VCCs within the same VPC require different QoS, the VPC performance objective agreed upon the network and subscriber should be suitably set for the most demanding VCC requirement.
- Network capacity allocation options to the VPC:
  - *Agreed peak demand*—The network may set the capacity of the VPC equal to the total of the peak data rates of all of the VCCs within the VPC.
  - *Statistical multiplexing* —If the network sets the capacity of the VPC to be greater than or equal to the average data rates of all VCCs but less than the aggregate peak demand, then a statistical multiplexing service is supplied.
    - VCCs experience greater cell-delay variation and greater cell-transfer delay



Configuration of VCCs and VPCs (I.371)

VPC = Virtual path connection
VCC = Virtual channel connection
VP-Sw = Virtual path switching function
VC-Sw = Virtual channel switching function

Connection Admission Control

- The first line of defense for the network against excessive load
- When a user request a new VPC or VCC, he must specify the traffic characteristics in both directions for that connection.
- The network accepts the connection only if it can commit the resources necessary to support that traffic level while at the same time maintaining the agreed QoS of existing connections.
- Four parameters in current specifications of traffic contract

Parameter	Description	Traffic type
Peak Cell Rate (PCR)	An upper bound on the traffic that can be submitted on an ATM connection.	CBR, VBR
Cell Delay Variation (CDV)	An upper bound on the variability in the pattern of cell arrivals observed at a single measurement point with reference to the peak cell rate.	CBR, VBR
Sustainable Cell Rate (SCR)	An upper bound on the average rate of an ATM connection, calculated over the duration of the connection.	VBR
Burst Tolerance	An upper bound on the variability in the pattern of cell arrivals observed at a single measurement point with reference to the sustainable cell rate.	VBR

- For a given VPC or VCC connection, the four parameter may be specified in several ways:

	Explicitly specified parameters		Implicitly specified parameters
	Parameter values set at connection-setup time	Parameter values specified at subscription time	Parameter values set using default rules
	Requested by user/NMS	assigned by network operator	
SVC	signaling	by subscription	network-operator default rules
PVC	NMS	by subscription	network-operator default rules

SVC = switched virtual connection
PVC = permanent virtual connection
NMS = network management system

Usage Parameter Control

- Once a connection has been accepted by the Connection Admission Control function, the *Usage Parameter Control* (UPC) function of the network monitors the connection to determine whether the traffic conforms to the traffic contract.
- The main purpose of this function is to protect network resources from an overload on one connection that would adversely affect the QoS on other connections.
- The control can be done at both the virtual path and virtual channel levels, but VPC-level control is more important.
- Two separate functions by the UPC
  - control of peak cell rate and associated cell-delay variation (CDV)
  - control of sustainable cell rate and associated burst tolerance

Priority Control

- Priority control comes into play when the network, at some point beyond the UPC function, discards (CLP=1) cells.
- The objective is to discard lower priority cells in order to protect the performance for higher-priority cells.
- *Note:*The network has no way to discriminate between cells that were labeled as lower-priority by the source and cells that were tagged by the UPC function.

Fast Resource Management

- Fast resource management functions operate on the time scale of the round-trip propagation delay of the ATM connection.
- It is a potential tool for traffic control (further study).
- One example of this function is the ability of the network to respond to a request by a user to send a burst.
  - Temporarily exceeding the current traffic contract by a user to send a relatively large amount of data.
  - If the network determines available resources along the route, then the network reserve those resources and grants permission.
  - Following the burst, the normal traffic control is enforced.

12.6 Congestion Control

- The set of actions taken by the network to minimize the intensity, spread, and duration of congestion, with two functions

Selective Cell Discarding

- similar to priority control
- in the priority control function (CLP=1) cells are discarded to avoid congestion.
- Unless the network is congested, discarding the cells are limited so that the performance objectives for the (CLP=0) and (CLP=1) flows are still met.
- To recover from congestion, network is free to discard any (CLP=1) cell and may even discard (CLP=0) cells on ATM connections that are not complying with their traffic contract.

Explicit Forward Congestion Indication

- works essentially in the same manner as for frame relay
  - An ATM node that is experienced congestion may set an explicit forward congestion indication in the payload type field of the cell header of cells on connections passing through the node.
  - The indication notifies the user that congestion avoidance procedure should be initiated for traffic in same direction as the received cell.
  - The user then may invoke actions in higher-layer protocols to adaptively lower the cell rate of the connection.