ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Multimedia Technology

Section # 21: Quality of service

Instructor: George Xylomenos

Department: Informatics

Contents

- Requirements
- Classification
- Scheduling
- Monitoring
- Integrated and differentiated services

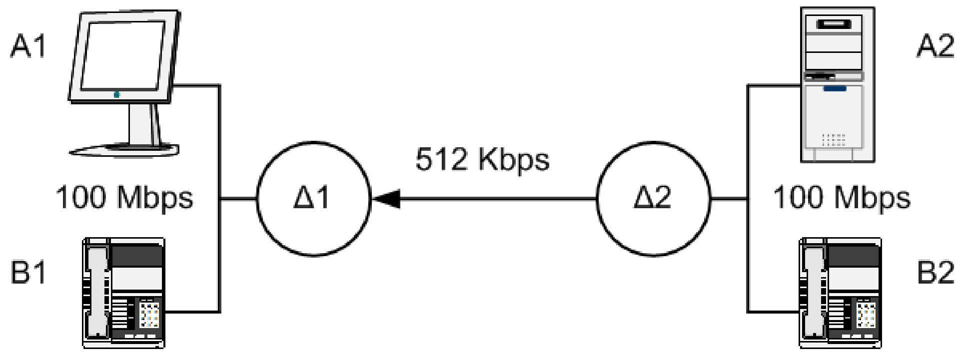
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Requirements

Class: Multimedia Technology, **Section # 21:** Quality of service

Instructor: George Xylomenos, **Department:** Informatics

Quality of service (1 of 3)



- Example of multimedia communication
  - FTP flow: A2->A1
  - Conferencing flow: B1<->B2
  - LANs connected at 512 Kbps
  - LANs are 100 Mbps
  - Routers Δ1 and Δ2

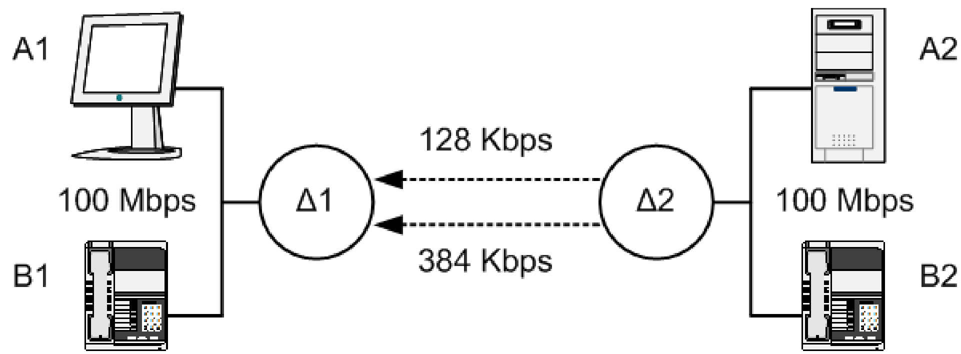
Quality of service (2 of 3)

- Focus on router Δ2
  - If offered load > 512 Kbps
  - High delay and packet loss
- Conferencing requirements
  - H.261 at 320 Kbps and G.711 at 64 Kbps
  - Total: 384 Kbps for audio/video
  - Ideally, FTP should take up the rest (128 Kbps)
    - Whatever remains from conferencing

Quality of service (3 of 3)

- Best effort environment
  - Packets are mixed in Δ2 queue
  - Scheduling is (usually) FIFO
  - FTP does not have a fixed transmission rate
    - TCP tries to grab as much as possible!
  - Possible delay and loss for AV packets
    - First delay grows as queue fills up
    - Then packets are dropped

QoE requirements (1 of 7)



- Easy solution: introduce priorities
  - AV packets sent before FTP packets
    - FTP packets sent when line is free
  - Every packet must be classified
    - To impose priorities

QoE requirements (2 of 7)

- In general: say we have different services
  - Maybe with different costs (net neutrality?)
    - The FTP user may have bought the better service
    - In this case AV packets should not have priority
    - Priorities can be based on different criteria
- <u>Packets must be classified</u>
  - Classification is a differentiation <u>mechanism</u>
  - What differentiation, is a matter of <u>policy</u>

QoE requirements (3 of 7)

- What if the conferencing app misbehaves?
  - For example, send more than 384 Kbps
    - Due to errors in the implementation
    - Or to take advantage of priority
  - This will starve the FTP flow
- <u>Flow isolation is needed for protection</u>
  - We want 384 and 128 Kbps virtual flows
  - With each flow protected from the other ones

QoE requirements (4 of 7)

- Isolating flows leads to another issue
  - What should we do when one flow pauses?
    - For example, no speech transmission during pauses
  - Resource usage should be efficient
  - If the line is free, we should use it
    - We can service the other flows
    - They end up transmitting more than expected
    - This is good for elastic flows (FTP is elastic)

QoE requirements (5 of 7)

- Monitoring: does a flow behave as it should?
  - When it does not, we have two options
    - Either drop or delay additional packets
  - Traffic monitoring and marking
    - At the endpoints and/or at the routers
- What is marking for?
  - If there is no congestion, just mark packets
  - They will be dropped if they face congestion later

QoE requirements (6 of 7)

- Say that we have two conferencing apps
  - Fairness means giving each half the bandwidth
    - But they need 384 Kbps each
    - They will both suffer 25% loss!
- Some flows require a minimum service
  - Without it, they just cannot work
  - Admitting them is bad for everyone
    - The do not work, and take resources from others
  - Flows should ask or resources first

QoE requirements (7 of 7)

- Call admission
  - Term comes from telephony
  - Flows need to specify their requirements
  - Acceptance
    - The network may reserve resources
  - Rejection
    - The resources needed are not available
  - This makes sure everything works
    - As long as it is accepted!

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

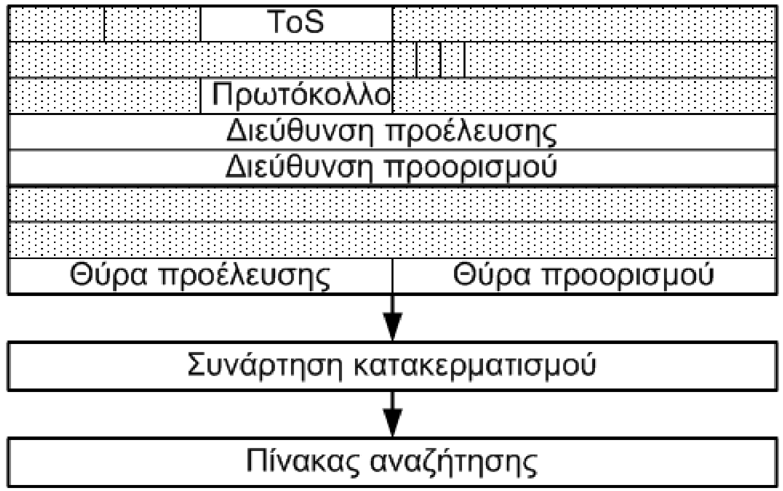
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Classification

Class: Multimedia Technology, **Section # 21:** Quality of service

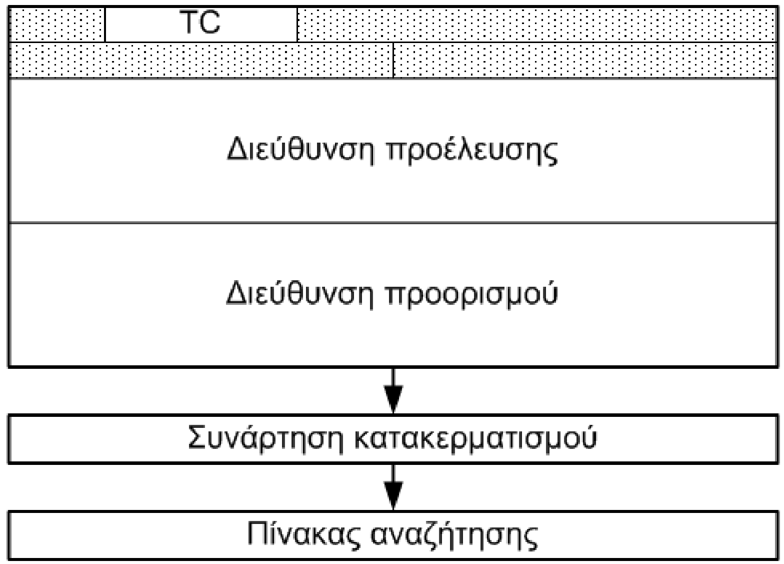
Instructor: George Xylomenos, **Department:** Informatics

Classification (1 of 2)



- Classification in IPv4
  - Based on header fields
  - Mask used to cover unimportant fields
  - The result is hashed
  - We lookup the hash in a table

Classification (2 of 2)



- Classification in IPv6
  - In IPv4 we normally look at TCP/UDP ports
    - Had to distinguish flows
  - IPv6 has a field for that!

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

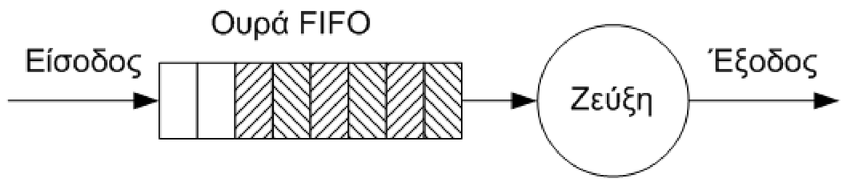
ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

Scheduling

Class: Multimedia Technology, **Section # 21:** Quality of service

Instructor: George Xylomenos, **Department:** Informatics

FIFO (1 of 4)



- Packet scheduling
  - Packets enter a queue on arrival
  - How do we select the next packet to sen?
- FIFO: the simplest option
  - Packets added to the tail of the queue
  - Packets sent from the head of the queue
  - Packets transmitted in order of arrival

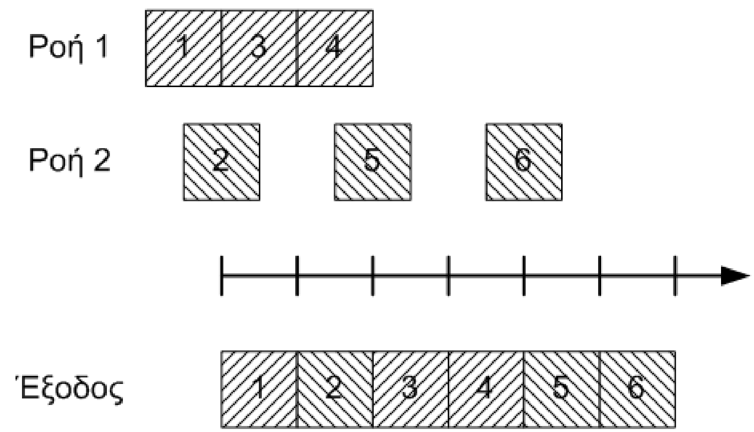
FIFO (2 of 4)

- Packet drop policy
  - What happens when the queue is full?
  - Latest packet dropped (drop tail)
    - No changes to already queued packets
    - Most common on the internet
  - Oldest packet dropped
    - Prefers newer packets
    - Drops packets that are already late

FIFO (3 of 4)

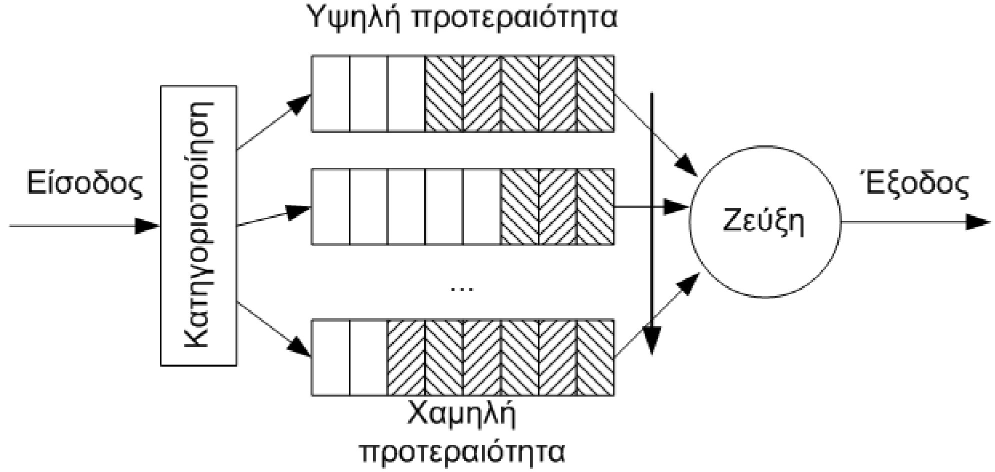
- Packet drop policy
  - Random early detection (RED)
    - We set a watermark for the queue
    - When we pass it, we start dropping packets
    - Drop probability depends on queue length
    - Probability grows with queue length
    - Sends a signal (loss) to transmitters early
    - Signals more transmitters

FIFO (4 of 4)



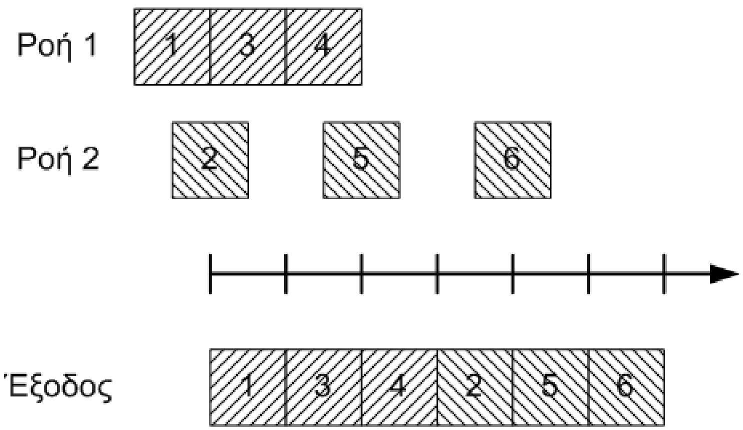
- Example of FIFO
  - Two packet flows
  - Top: arrival times
  - Bottom: departure times
  - In FIFO, only arrival times matter

Priority queues (1 of 3)



- Priority queues
  - Two or more per output link
    - Requires classifier to assign packets to queues
- Packet selected from highest priority
  - We get to priority n if 1, … , n-1 are empty

Priority queues (2 of 3)

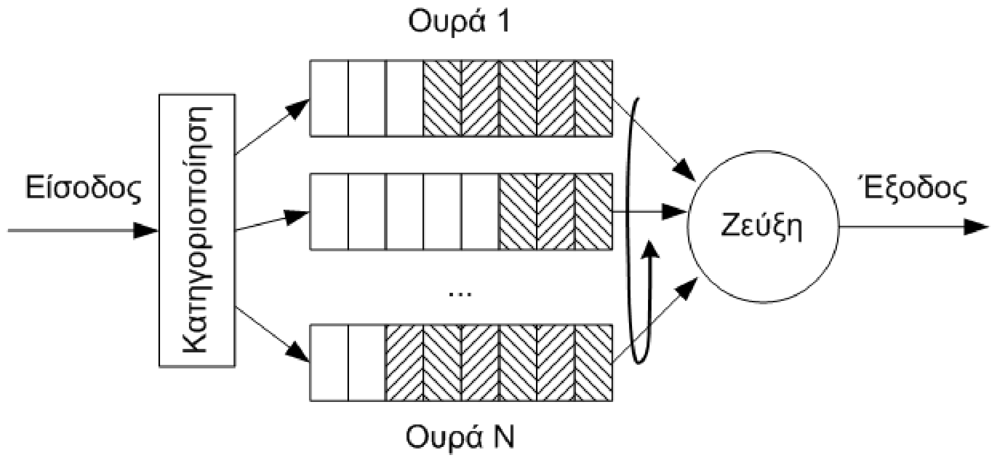


- Priority queue example
  - Flow 1 has higher priority than flow 2
  - As long as flow 1 has packets, it is served first
    - Arrival time in only important within the queue
  - Flow 2 may starve (no service)

Priority queues (3 of 3)

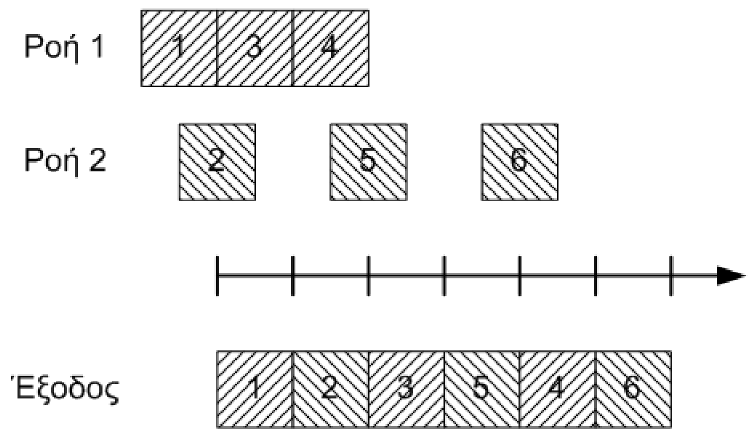
- Non-preemptive scheduling
  - A task is not interrupted after starting
    - Common in packet scheduling
    - Stopping a packet transmission would waste resources
- Preemptive scheduling
  - A task is interrupted by a higher priority one
    - Common in process scheduling
    - The interrupted process can be resumed later

Round robin (1 of 3)



- Round robin queues
  - Packets are classified into queues
  - The queues do not have a strict priority
  - Instead, we service them round robin
    - Simplest case: one packet from each queue
  - No queue can starve

Round robin (2 of 3)

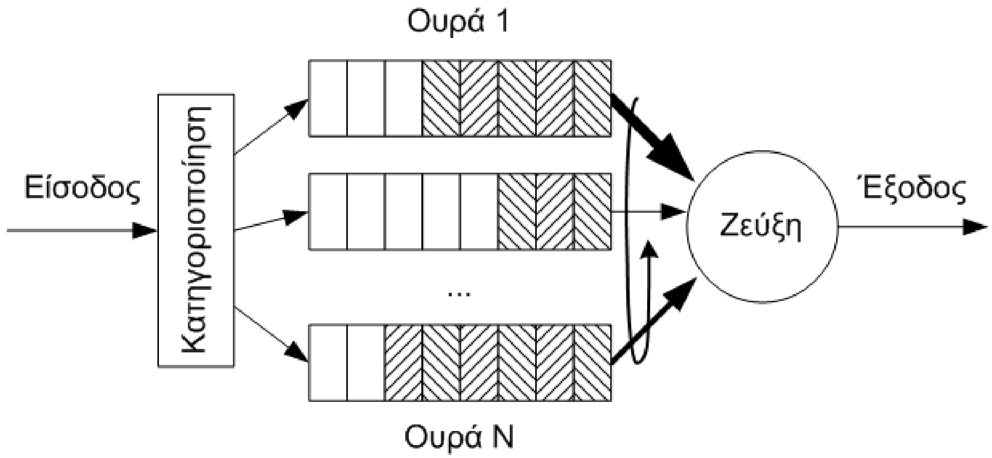


- Example of round robin
  - We constantly move from queue to queue
    - Each one gets half of the bandwidth
  - Arrival time matters only within a queue
  - What if a queue is empty?

Round robin (3 of 3)

- Work conserving scheduling
  - If a queue is empty, switch to the next
    - If there is any load, the line is never idle
- Non-work conserving scheduling
  - On empty queue, line goes idle
    - Wastes transmission resources
  - But it offers tighter delay guarantees
    - Can avoid waiting in some cases
    - Example: packet arrives just after we skipped its queue

Fair queues (1 of 2)



- Fair (weighted) queues
  - Generalization of round robin
  - Queues served in round robin manner
    - Usually with work conserving scheduling
  - But each queue gets different service
    - A fraction of the lines capacity

Fair queues (2 of 2)

- How is fair queueing implemented?
  - Ideally, we emulate fluid flow
    - Each flow is a pipe going to a larger pipe
  - In practice, we transmit entire packets
    - Fluid flow requires transmitting bits
  - If all packets same size, rather easy
  - If not, we approximate fluid flow
    - We emulate an ideal queue
    - And try to keep close to it in the long run

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Monitoring

Class: Multimedia Technology, **Section # 21:** Quality of service

Instructor: George Xylomenos, **Department:** Informatics

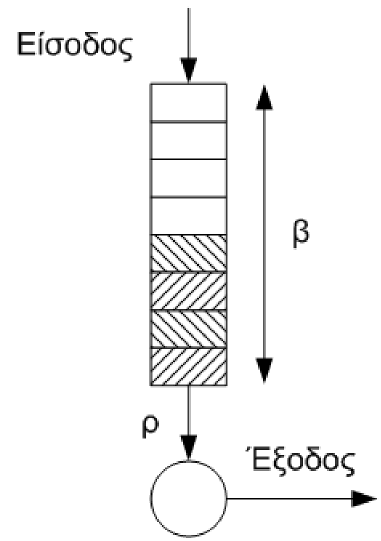
Monitoring criteria (1 of 2)

- Goals of monitoring
  - Limit packet transmission rate
  - Which time scale are we looking at?
- Average rate
  - Measured over a period of time
  - The period is critical!
    - 100 packets per second?
    - 6000 packets per minute?

Monitoring criteria (2 of 2)

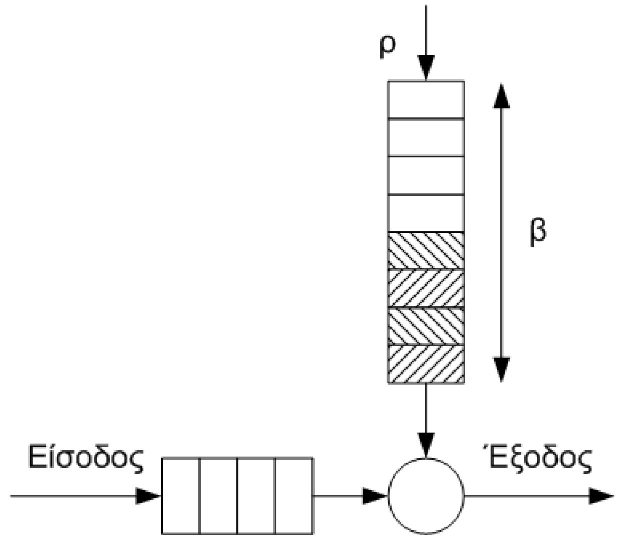
- Maximum rate
  - Maximum rate in a smaller period
    - Average 6000 packets / min
    - Maximum 1500 packets / sec
    - Limits the burstiness of the source
- Maximum burst
  - Maximum number of back-to-back packets

Leaky bucket



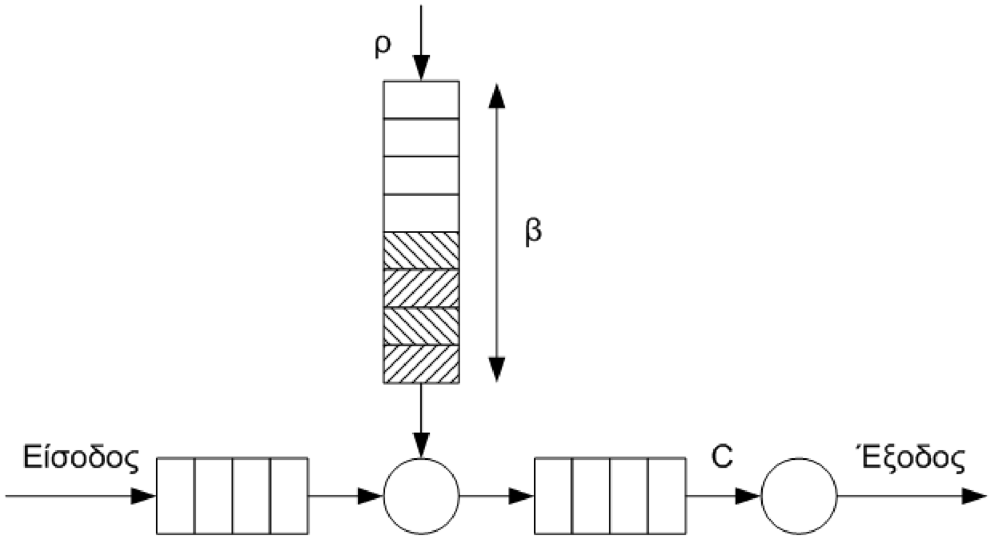
- Shapes traffic to average rate
  - Packets are inserted to the bucket on arrival
  - The bucket empties at rate ρ (average)
  - The bucket has size β (max arrival burst)
    - But we always send no more than ρ

Token bucket



- Shapes traffic to average rate and burst
  - Each packet needs tokens to go
  - Tokens are gathered at rate ρ (average)
  - Can hold up to β tokens (max output burst)
    - Can send these at top speed

Combined buckets



- The two buckets connected in series
  - We first have a token bucket
    - Limits average (ρ) and burst (β)
  - We then have a leaky bucket
    - Limits maximum (C > ρ)

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Integrated and differentiated services

Class: Multimedia Technology, **Section # 21:** Quality of service

Instructor: George Xylomenos, **Department:** Informatics

Integrated (1 of 3)

- Integrated services (IntServ)
  - Three classes of service
    - Implemented at each router
  - Guaranteed: limited by token bucket
    - Must go through admission control
    - Guarantees end to end delay
  - Controlled load: statistical guarantees
  - Best effort: Everything else

Integrated (2 of 3)

- Resource Reservation Protocol (RSVP)
  - Resource reservation in one direction
  - Phase 1: sender sends probe to receiver
    - Packet gathers resource info on the way
  - Phase 2: receiver sends reservation to sender
    - This packet reserves resources in routers
  - Reservations periodically refreshed
    - Adaptation to routing changes
  - Can have shared reservations for multicast

Integrated (3 of 3)

- IntServ issues
  - Must be implemented everywhere
    - At least in a large part of the Internet
    - Same services implemented everywhere
    - But nobody wants to start first!
  - Not scalable
    - Resources are reserved per flow
    - The core network handles millions of flows
  - Only three classes of service
    - Who wants controlled load?

Differentiated (1 of 3)

- Differentiated services (DiffServ)
  – Flows are groups into classes
    - On entering the network
    - On moving to a different AS
  – Flow shaping only at these points
    - Other routers only implement something simple
  – Long term resource reservations
    - Based on agreements between ASes
  – Complexity only at the edges of the network

Differentiated (2 of 3)

- Per hop behavior (PHB)
  - How each class is handled
  - Looks at DSCP field (ToS or TC)
    - Set when entering the network/AS
- Expedited forwarding PHB (EF)
  - Reserved bandwidth at each router
    - Flow limited when entering the network
  - Very low delay and loss

Differentiated (3 of 3)

- Assured forwarding PHB (AF)
  - Fraction of bandwidth for each service class
  - Different drop priorities within each class
  - Flows marked when sending more
    - Dropped if congestion appears
- DiffServ issues
  - There are no Internet-wide services
  - Had to create end-to-end services
  - Appropriate for long term traffic shaping

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

End of Section # 21

Class: Multimedia Technology, **Section # 21:** Quality of service

Instructor: George Xylomenos, **Department:** Informatics