PIREGRID
EFA53/08

QoS for Cloud Computing

PIREGRID THEMATIC DAY
May, 10thn 2011
University of Pau

Prof. Congduc Pham
http://www.univ-pau.fr/~cpham
Université de Pau, France

What is Quality of Service?

- Quality of service is the ability to provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance
- QoS criteria are numerous and is highly dependant of the application...
  - Throughput, Delay, jitter, Loss rate
- ... Or of the end-user
  - Image resolution, sound quality, appropriate language, ...

PireGrid
WWW.PIREGRID.EU

COOPERACIÓN COOPÉRATION TERRITORIAL TERRITORIALE
ESPAÑA-FRANCE-ANDORRA
2007-2013

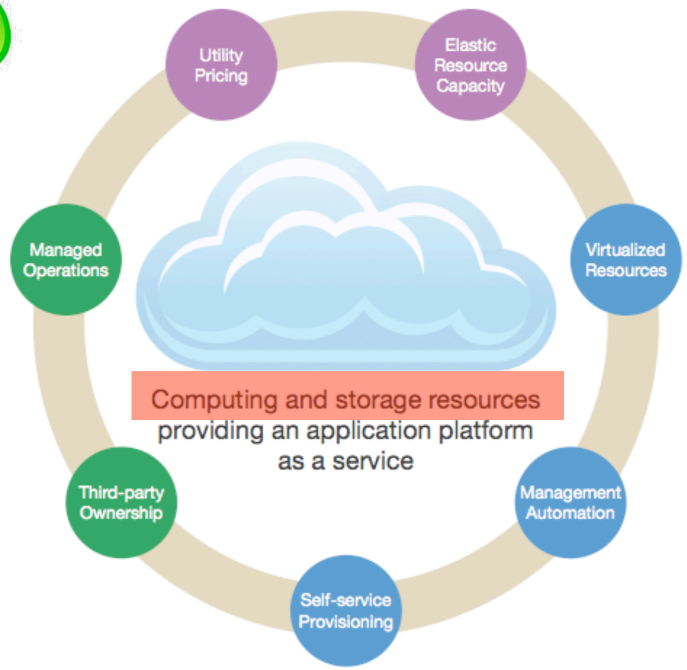
Common Service Specification

- **Loss:** probability that a flow's data is lost
- **Delay:** time it takes a packet's flow to get from source to destination
- **Delay jitter:** maximum difference between the delays experienced by two packets of the flow
- **Bandwidth:** maximum rate at which the source can send data
- **QoS spectrum:**

QoS for Cloud

- The shared cloud assumption

Many users, with various profile and different needs!



Economic Elements: Pay-as-you-go, pay-as-you-grow, no CAPEX.

Architectural Elements: Simple, abstract environment for development.

Strategic Elements: Focus on your core business, leave the rest to someone else.

Concept: MWD Advisors, www.mwdadvisors.com

PireGrid
WWW.PIREGRID.EU

Application's profile

Traditional scientific computing applications

Business/ financial applications & simulations

Interactive applications with user feedback

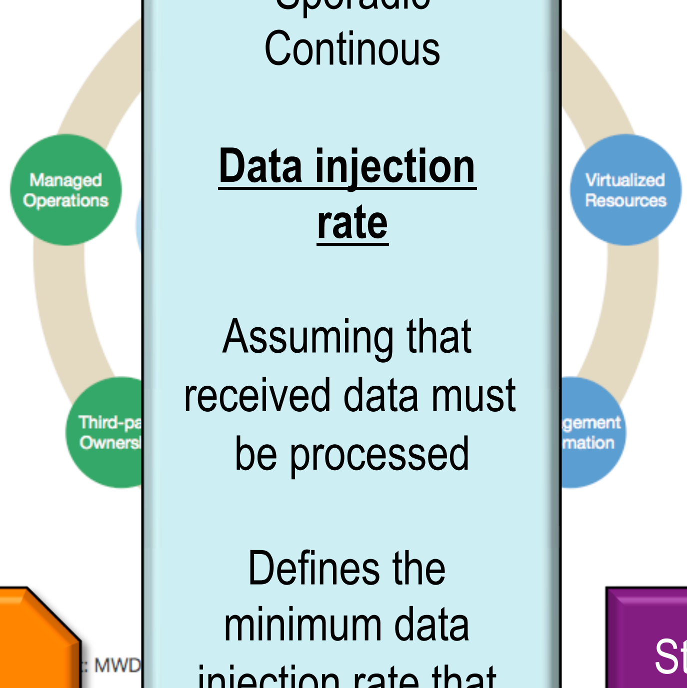
Time scale

Bounded in time
Sporadic
Continous

Data injection rate

Assuming that received data must be processed

Defines the minimum data injection rate that is required by the application



Environmental and data processing/ analysis

Urgent computing for public safety, disaster relief

Storage and databases operations/ queries

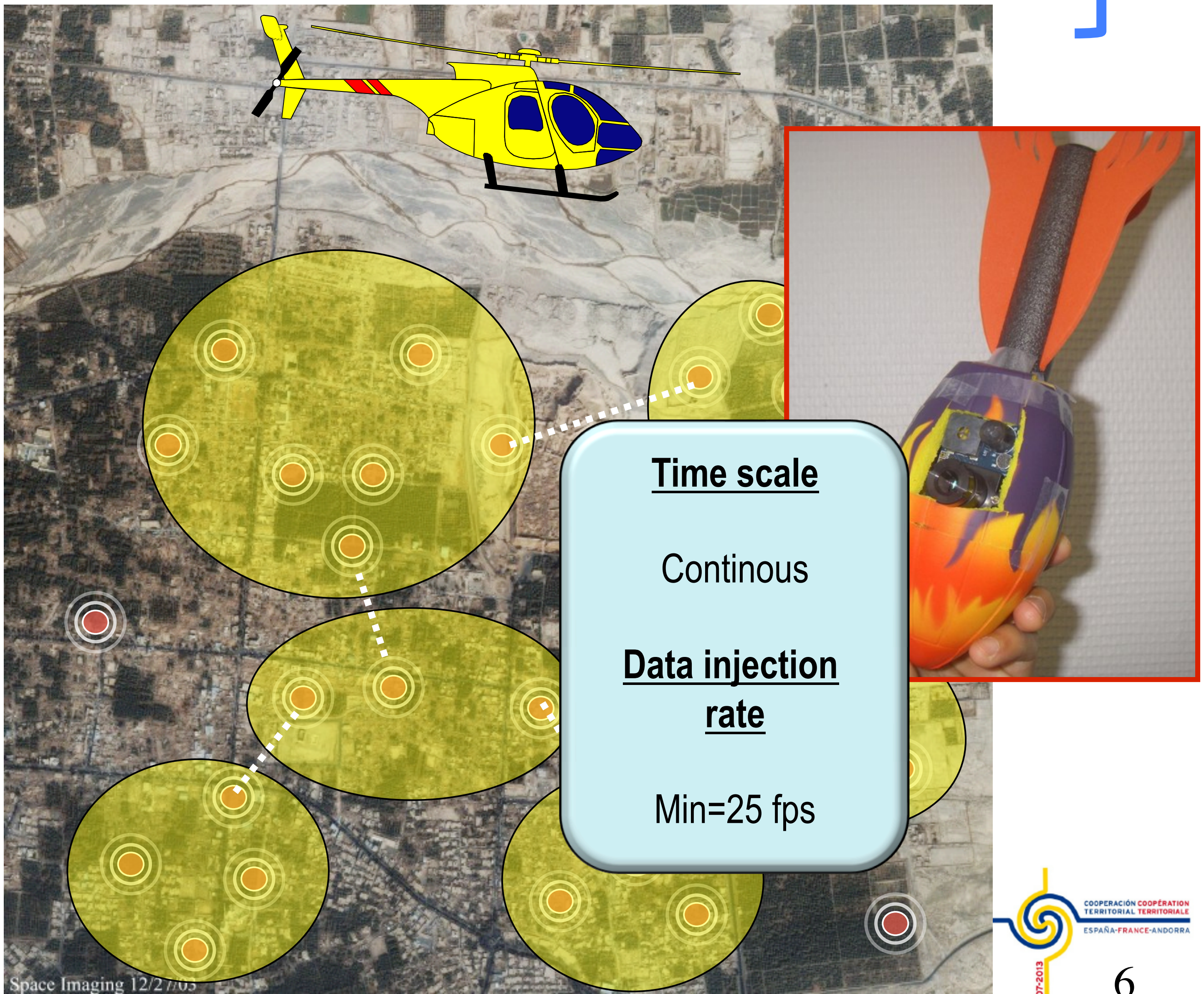
PireGrid
WWW.PIREGRID.EU

COOPERACIÓN COOPÉRATION TERRITORIAL TERRITORIALE
ESPAÑA-FRANCE-ANDORRA
2007-2013

Urgent disaster relief

Imote2

Multimedia board



How to provide QoS?

- Many ways to provide QoS
  - Scheduling, admission control, trafic control, dynamic resource provisioning, ...

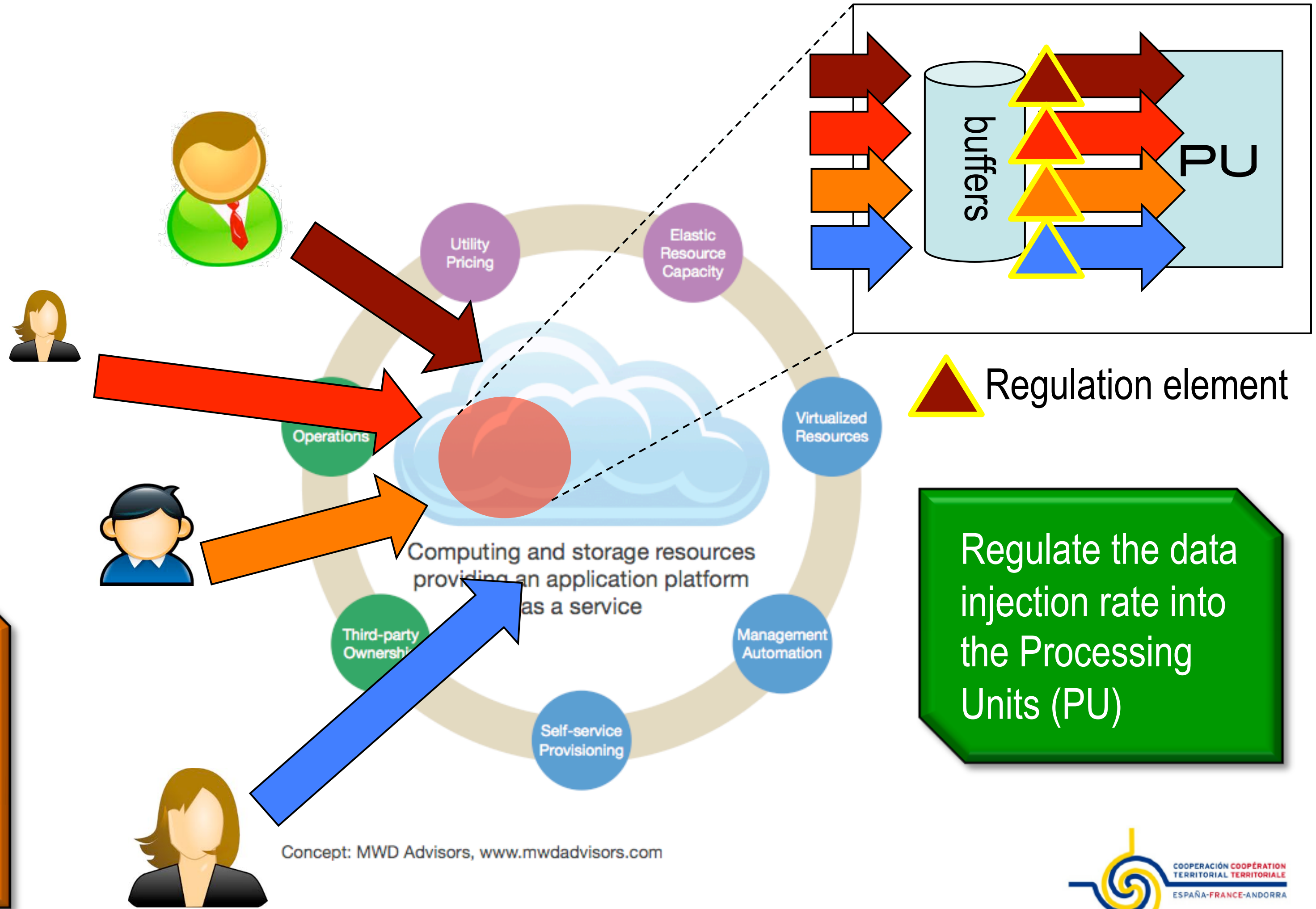
How to take into account the various application's profiles?
How to protect users from misbehaving applications?
How to handle urgent demands?

Regulate (adapt & control) the data injection rate into the computing resources

PireGrid
WWW.PIREGRID.EU

COOPERACIÓN COOPÉRATION TERRITORIAL TERRITORIALE
ESPAÑA-FRANCE-ANDORRA
2007-2013

QoS: general picture



Concept: MWD Advisors, www.mwdadvisors.com

Regulate the data injection rate into the Processing Units (PU)

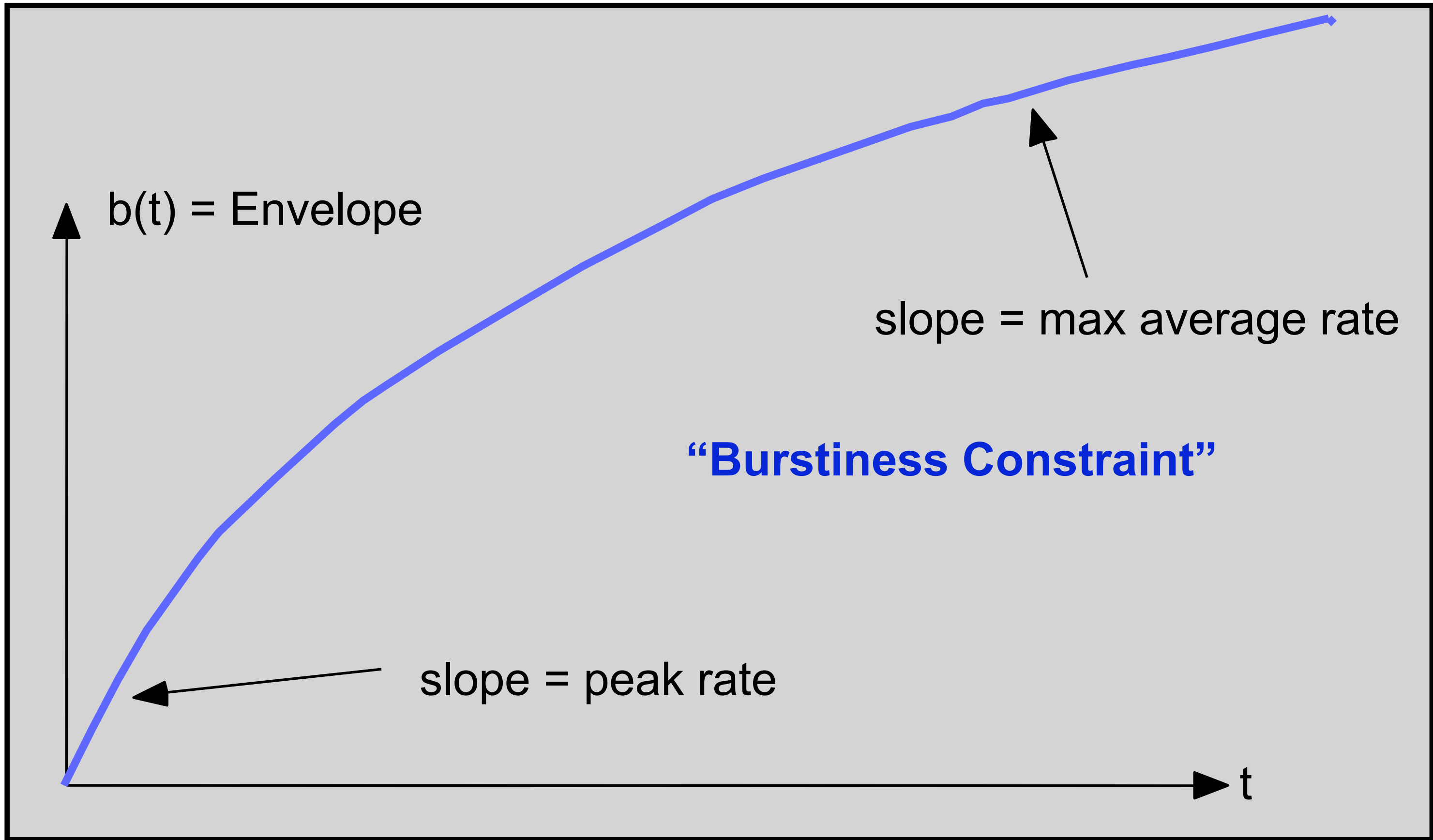
Network bandwidth is assumed to NOT BEING the bottleneck

Traffic and Service Characterization

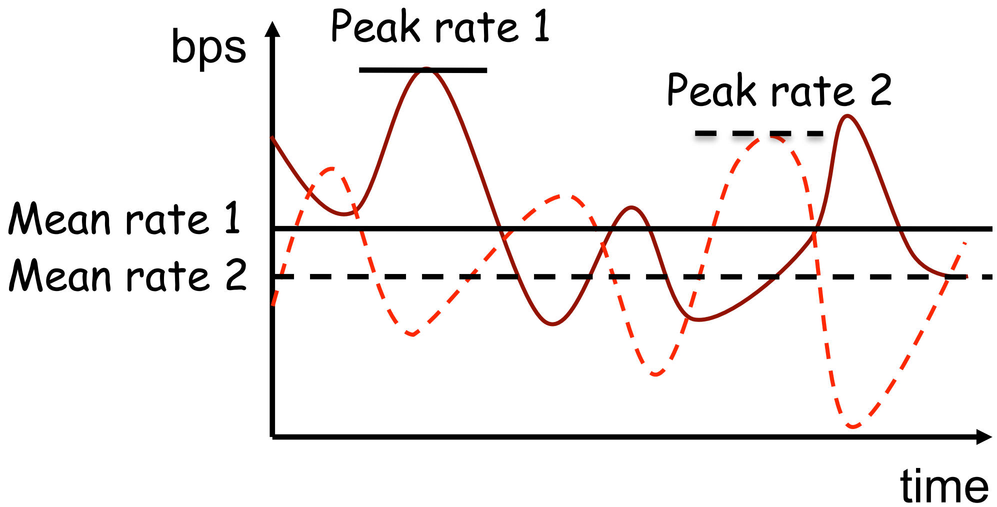
- Definitions
  - Cloud Infrastructure (CI)
  - Processing Units (PU)
- To quantify a service one has two know
  - Flow's traffic arrival
  - Service provided by the CI, i.e., resources reserved at PU
- Regulation will ne done by an envelope process, borrowed & adapted from the network community
- Ideas is to
  - Bound the data injection rate to...
  - ...isolate users from each others and...
  - ...to provide QoS enforcement at flow level

Traffic Envelope (Arrival Curve)

- Maximum amount of service that a flow can request during an interval of time t

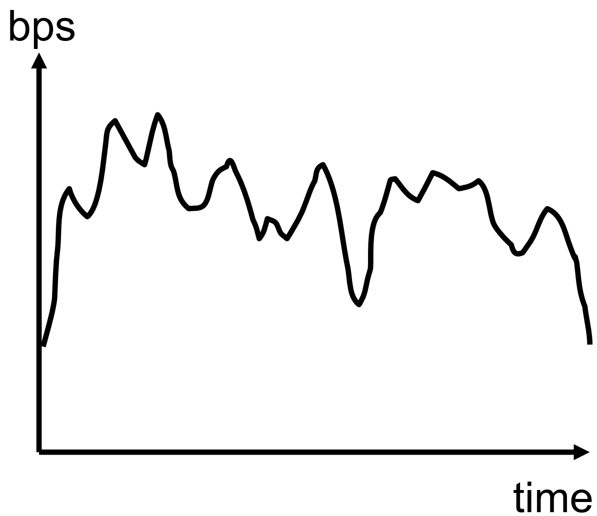


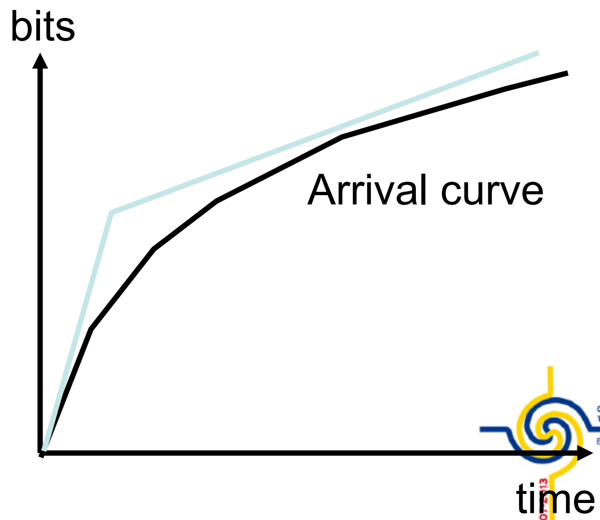
Traffic Envelope (traffic shaping)



Traffic are variable by nature, must take into account burstiness constraints

Use an envelope process to bound the data injection rate while allowing for variable, bursty traffic

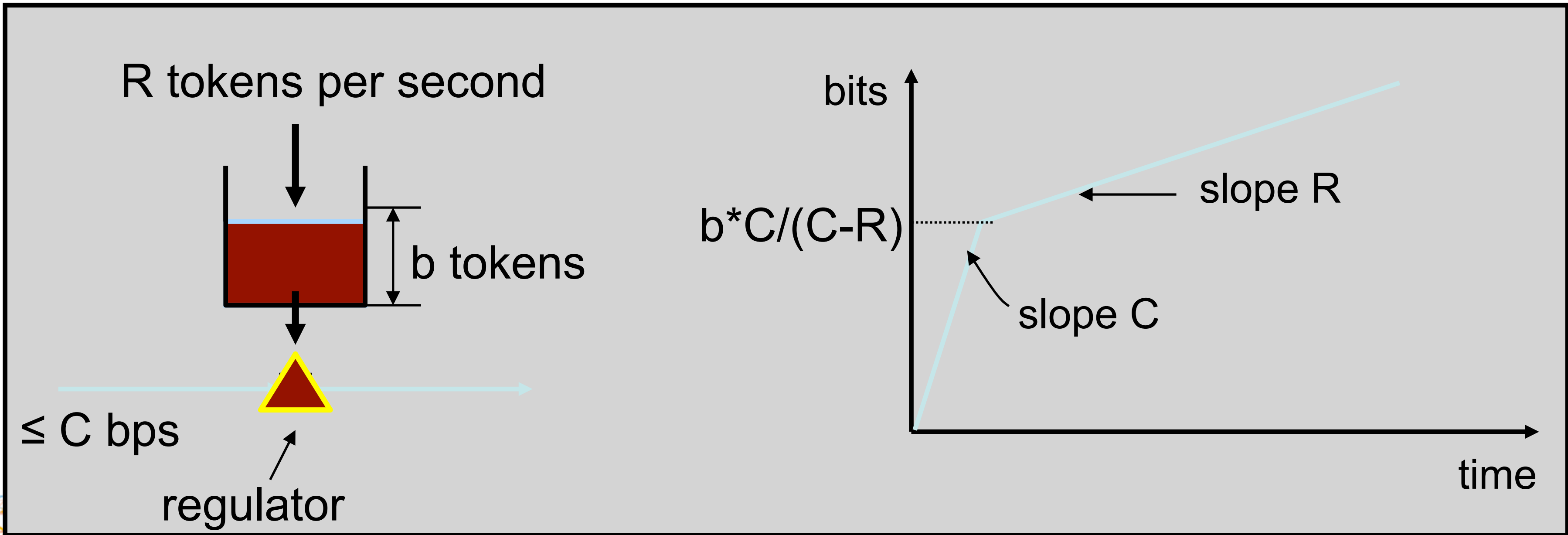




PireGrid
WWW.PIREGRID.EU

Ex: Token Bucket (1)

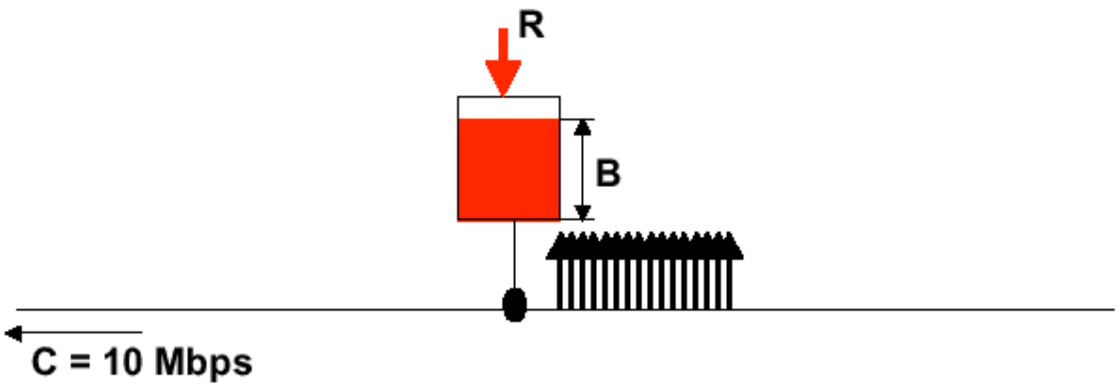
- Characterized by three parameters (b, R, C)
  - b – token depth
  - R – average arrival rate
  - C – maximum data injection rate
- A bit is transmitted only when there is an available token
  - When a bit is transmitted exactly one token is consumed



Token Bucket (2)

Example

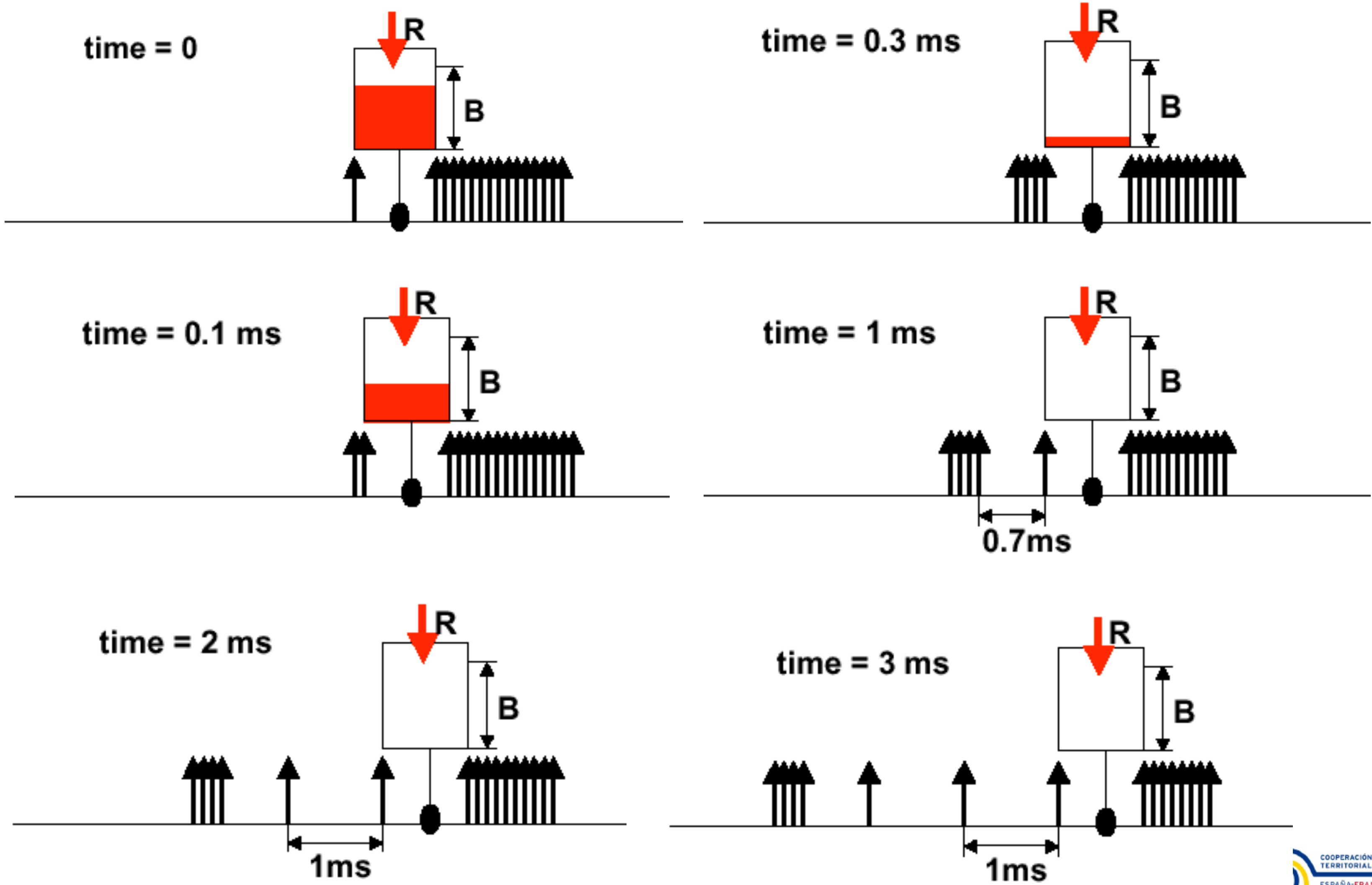
- B = 4000 bits, R = 1 Mbps, C = 10 Mbps
- Packet length = 1000 bits
- Assume the bucket is initially full and a "large" burst of packets arrives



istoica@cs.cmu.edu

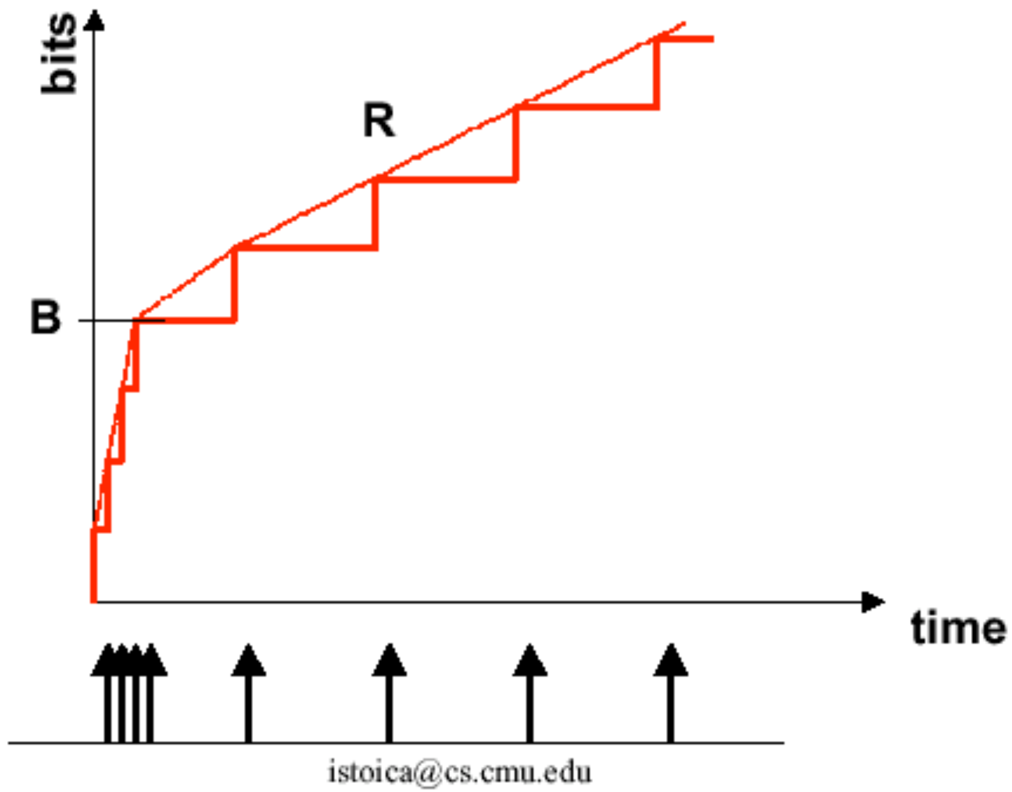
PireGrid
WWW.PIREGRID.EU

Token Bucket (3)



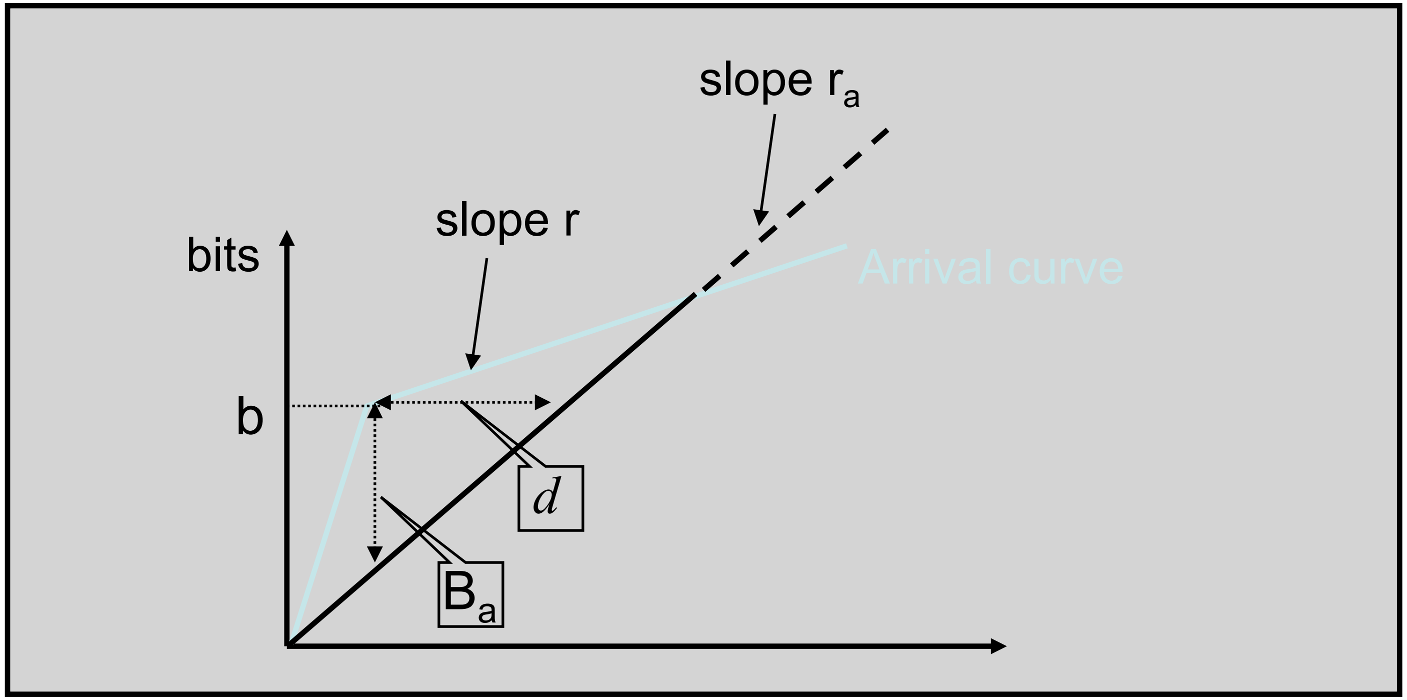
Arrival curve

A(t) – number of bits received up to time t

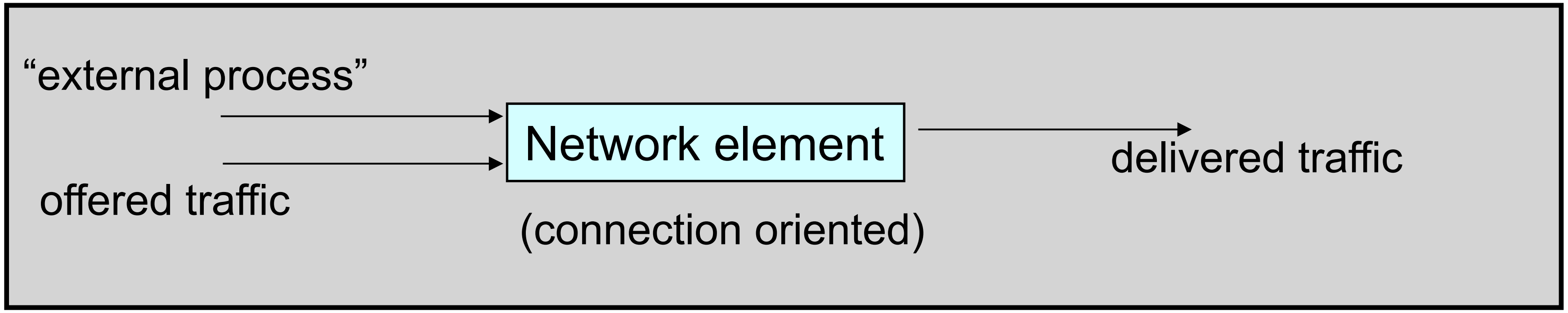


Per-hop Reservation with Token Bucket

- Given b,r,R and per-hop delay d
- Allocate bandwidth r_a and buffer space B_a such that to guarantee d



Service Model

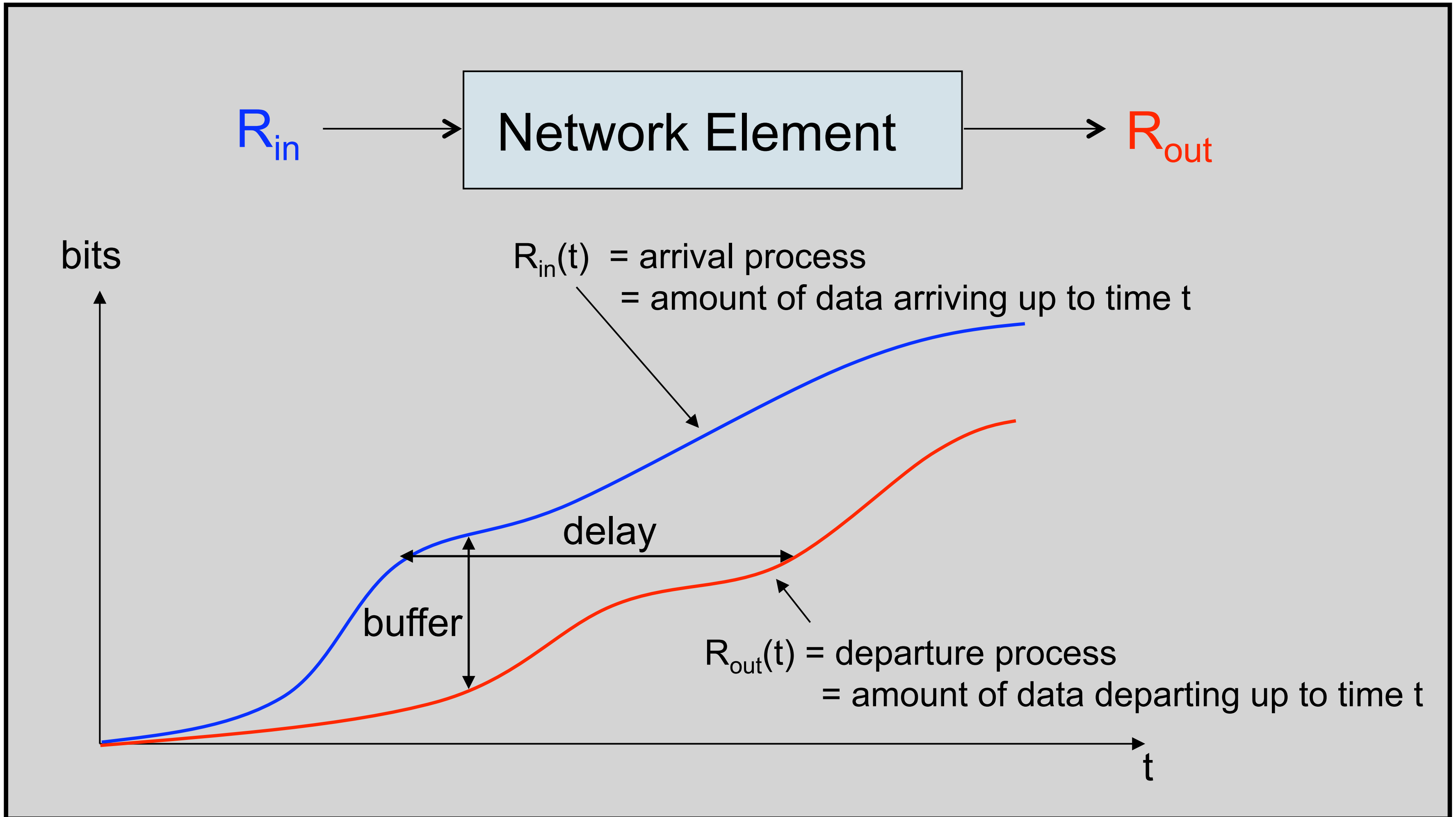


- The QoS measures (delay,throughput, loss, cost) depend on offered traffic, and possibly other external processes.
- A service model attempts to characterize the relationship between offered traffic, delivered traffic, and possibly other external processes.

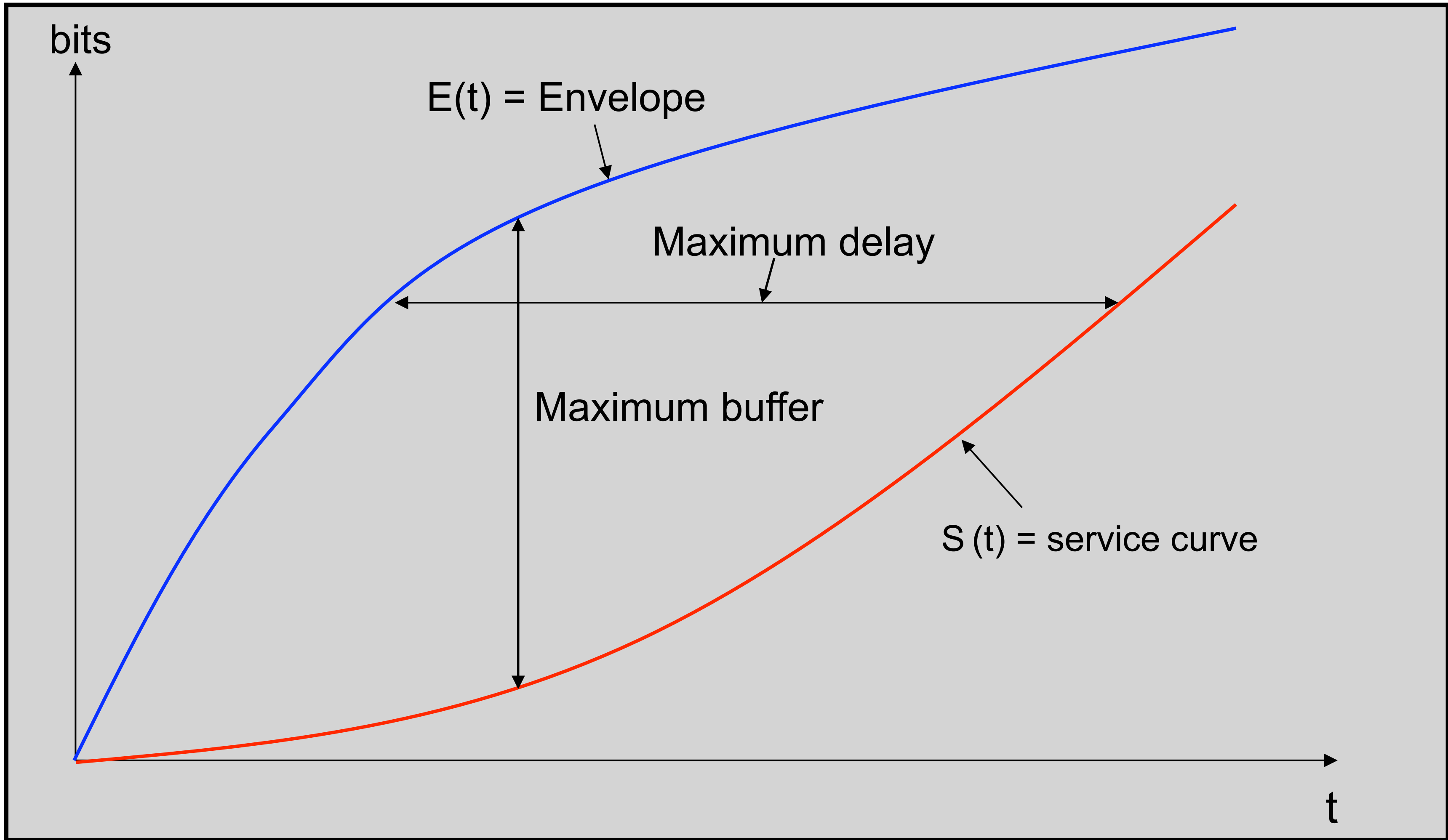
PireGrid
WWW.PIREGRID.EU

COOPERACIÓN COOPÉRATION TERRITORIAL TERRITORIALE
ESPAÑA-FRANCE-ANDORRA
2007-2013

Arrival and Departure Process

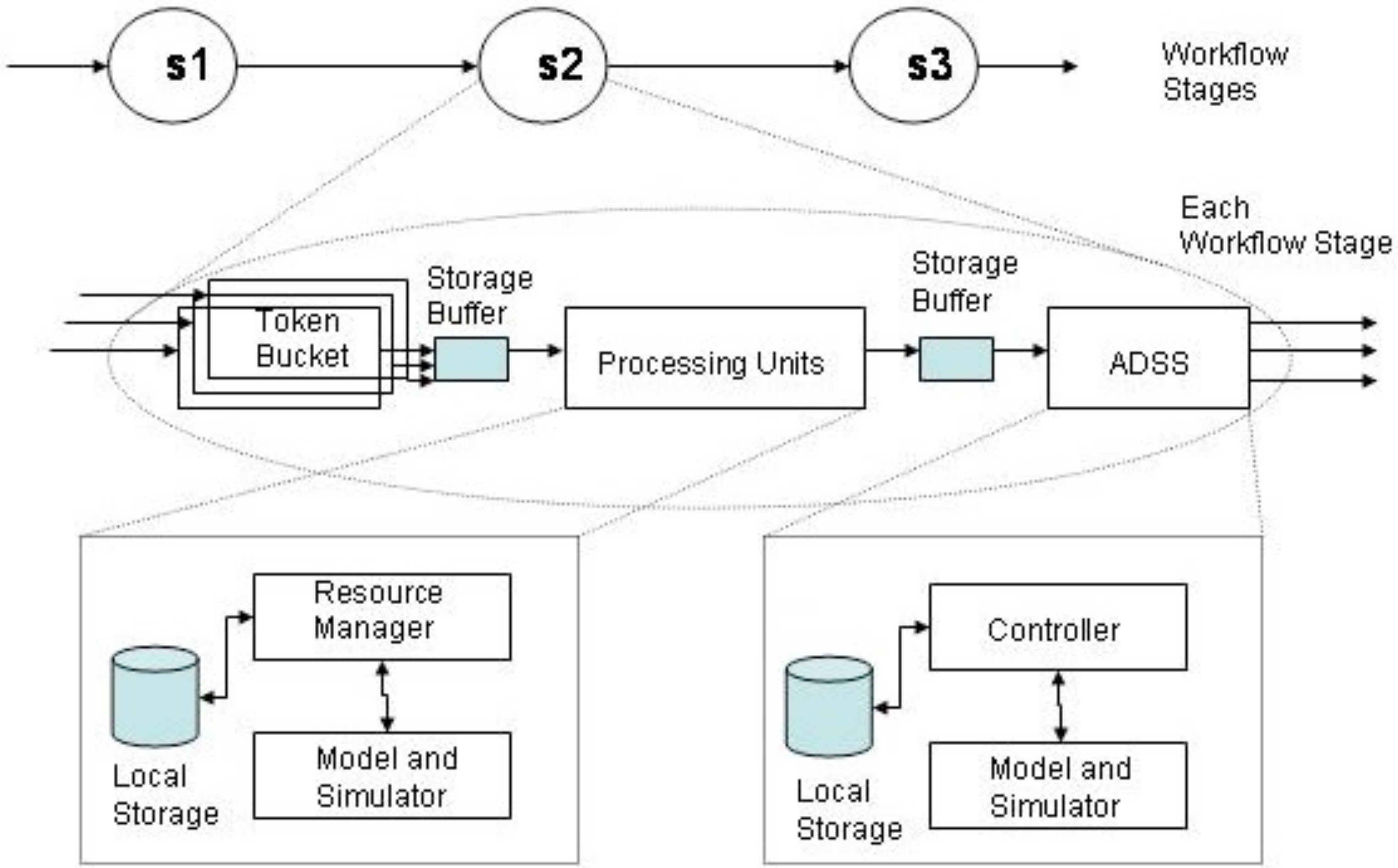


Delay and Buffer Bounds



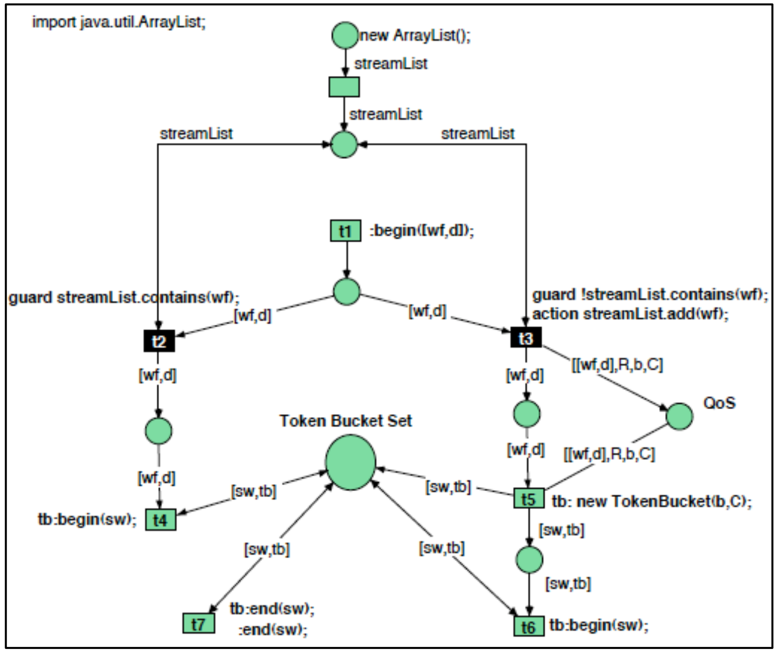
Token Bucket support in workflows (1)

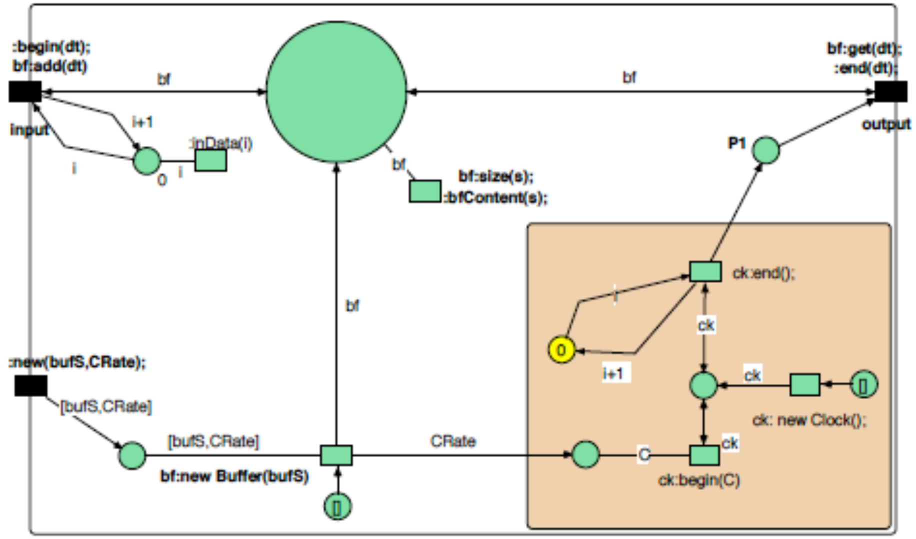
Support of superscalar pipeline models

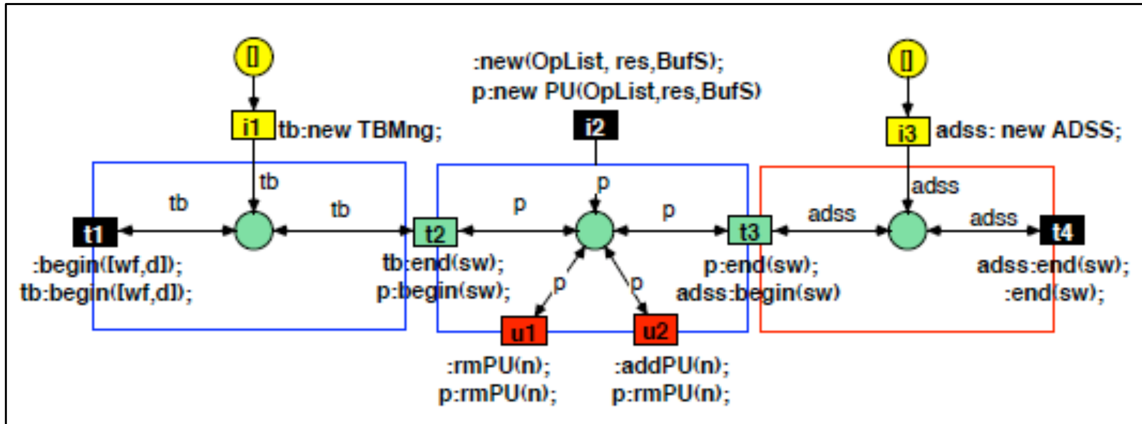


Work done with R. Tolosana, J. Banares and O. Rana

Token Bucket support in workflows (2)







TB QoS is introduced seamlessly into workflow specifications with the Renew tools

PireGrid
WWW.PIREGRID.EU

Conclusions

- Clouds will be shared clouds driven by economical constraints
- For some applications, availability of resources and isolation are of prime importance (urgent computing)
- QoS for clouds is already a necessary and hot topic in research community

PireGrid
WWW.PIREGRID.EU

COOPERACIÓN COOPÉRATION TERRITORIAL TERRITORIALE
ESPAÑA-FRANCE-ANDORRA
2007-2013

Perspectives

- Add more parameters to the TB model
  - Excess burst size
  - Advanced mark vs. drop policy
- Dynamic configuration of TB parameters at each stage of the processing path
- Take into account data inflation behaviors
- Generalized the usage of envelope processes, comparison,...

PireGrid
WWW.PIREGRID.EU

COOPERACIÓN COOPÉRATION TERRITORIAL TERRITORIALE
ESPAÑA-FRANCE-ANDORRA
2007-2013