QORE

## QoS and Resource Optimisation in the Internet

**T-Systems, TZD**
**Head of Department Media & Voice Solutions,**
**Dirk Hetzer**

Berlin Jan 06



QORE

## Structure

- Motivation
- QoS and Resource Optimisation Goals
- QORE Optimisation Algorithms
- QORE Architectural Design and Scenarios
- Ongoing Work – Intermon
- Summary



QORE

## Topics

- Introduction
- QoS provision in Internet
- Internet Bandwidth Allocation
- Advance Resource Reservation
- Optimisation Techniques
- Measurement Based QoS and Resource planning Framework
- Conclusion and Challenges for Future Work

# QORE

## Motivation

Requests for next generation networks fix/mobil

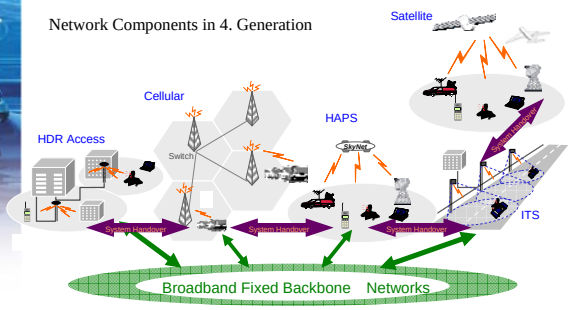
- QoS All-IP Network (QoS Signalling, IPv6, Intser/DiffServ Basis)
- Mobile Middelware-Plattform
- User Profiling
- Middelware Plattform for mobile Applications ( open API's)
- Multicast / Streaming
- Security / PKI
- Location based services
- Support of different Service Provider
- Adaptation of Applications (XML)



# QORE

## QORE Motivation

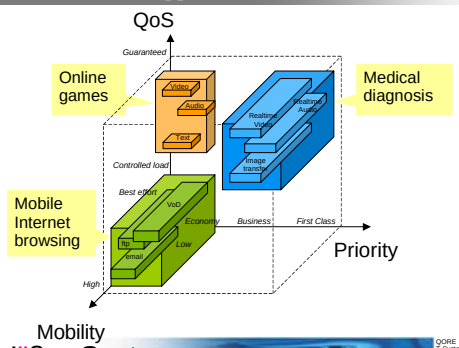
Network Components in 4. Generation






# QORE

## IP QoS for mobile Applications



QORE

## Trends mobile Communication

Isolated service sets
- sequential service for cellular (phone, fax, …) and IP (WAP, Web, email, ...)
- terminal: unified user interface, isolated stacks
- infra: radio capacity sharing, user-initiated IP connectivity, cellular push

Integrated service sets
- parallel service for cellular and IP (e.g. WAP and GSM voice)
- integrated service for cellular and IP (e.g. WTA and GSM voice)
- terminal: "pre all-IP" user interface
- infra: radio coordination, PS control plane for CS user plane

All-IP service set
- IP traffic evolution: bursty + streaming + real-time voice&video
- key requirements: fast session set-up, push, QoS (handover, delay, jitter)



QORE

## Motivation – different kind of QoS requirements

- Elasticity
- Autonomy
- Large scale
- Time criticality
- Safety criticality
- Geographical dispersion
- Mobility
- Evolution



QORE

## Motivation – need for QoS optimisation derived from

- QoS tuning and adaptation for application and traffic classes
- QoS/SLAs specification and monitoring of applications or traffic QoS parameters
- QoS based adaptive resource control for different traffic classes
- Bandwidth Brokerage (BB) for QoS and resource mapping
- DiffServ traffic classes and corresponding SLAs and QoS requirements
- Measurement Based Admission Control
- Inter-domain QoS analysis and traffic
- Backbone traffic engineering …….

QORE

## Usage areas for QoS and resource optimisation

- **Operational performance management**
- **-> Optimisation of resource allocation for QoS enabled applications and Value added Virtual Private Networks at router access points (border router, IXP, peering points, core´and edge router)**
- **Network planning**
- **-> Simulation of different strategies for resource reservation based on operation research techniques considering different kinds of QoS requirements (elastic application levels, cost restrictions, delay and timing dependencies, intelligent mobile components)**



QORE

## Structure

- Motivation
- QoS and Resource Optimisation Goals
- QORE Optimisation Issues
- QORE architectural design and scenarios
- Ongoing Work – Intermon
- Summary



QORE

## Network Planning and QoS Optimisation

**Optimisation of QoS and resource reservation levels of elastic applications i.e. applications with different levels of QoS support and resource requirements**

| Elasicity and Rate type | Descriptions |
|---|---|
| Stream | Predictable delivery at a relatively constant bit rate (CBR). For example, although their rates often fluctuate, audio and video data streams are considered CBR because they have a quantifiable upper boundary. |
| Burst | Unpredictable delivery of "blocks" of data at a variable bit rate (VBR). Applications like file transfer move data in bulk that can increase data rate to use all available bandwidth (no upper bound). |

QORE

## Operative Performance Management and Optimisation

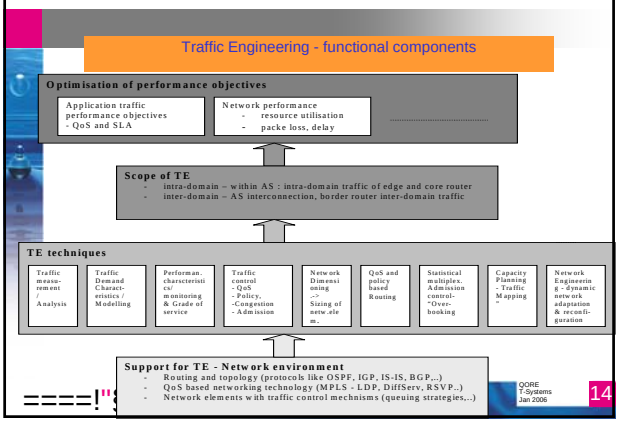
**QoS and resource optimisation combined with operative control**

- effective bandwidth
- management of bandwidth (Bandwidth Brokers)
- measurement based admission control
- traffic and QoS model predictions



QORE

## Traffic Engineering - functional components






QORE

## Admission Control

- Admission control is a methodology, whether or not new flows requesting resources can be accepted while maintaining QoS.
  - In a traditional approach to perform admission control, a flow is described by an *a priori* traffic descriptor containing parameters to characterize the properties of the flow's expected traffic.
  - The measurement-based admission control (MBAC) approach [] uses *a priori* source characterization only for incoming flows (and those recently admitted), measurements are used to characterize those flows that have been in place for reasonable duration.
- Admission control algorithms cf. Knightly and Shroff :
  - Average and peak rate combinatorics
  - Additive effective bandwidths
  - Engineering the "loss curve"
  - Maximum variance approaches

QORE

## Resource Allocation in Advance (1)

**Objective and benefits**

- Immediate reservation (for instance RSVP) are requested and granted just when the resources are actually needed -> Versus:
- Advance Reservation, i.e. resources are reserved in advance to the actual usage phase -> Benefits:
  - to overcome the blocking probability of a communication network
  - Duration and Requirements for reservation schedule and type (adaptable)
  - Increased probability of resource availability
  - Consideration of alternative strategies for resource optimisation
  - Flexible handling of applications by decoupling the starting time of the service from the time the service request is made
  - Optimisation of resource utilisation and QoS provision by



QORE

## Resource Allocation in Advance (2)

**Application requirements**

- Bandwidth reservation for Grid Computing : Advance Reservation API -> Scheduling Working Group
- Scheduling for multimedia applications and performance critical systems -> video on demand, distributed multimedia delivery
- Hard real time and embedded System applications (Worst Case delay guarantee)
- QoS based Virtual private networks -> adaptable resource reservation for optimal QoS provision
- Mobile and wireless services -> reservation in advance to continuously keep the communication service alive
- Best effort and elastic applications



QORE

## Resource Allocation in Advance (3)

**Internet Protocols and background**

- Protocols
  - ReRA (Resource Reservation in Advance)
    - on top of RSVP: evaluation and selection of an appropriate call admission control strategy, internal management of advance reservations
    - mapping to ATM
  - Differentiated Services Bandwidth Broker
- Theoretical models
  - Admission control
  - Network calculus
  - Effective bandwidth

QORE

## Operation research methods for QoS and resource optimisation

**Goal function**
**optimal QoS level of applications and cost efficient resource assignment**

**Communication jobs with restrictions on**
**•Durations (time characteristics and synchronisation),**
**•Multiplexing**
**•Timing dependencies**
**•levels of QoS and resource reservation**
**•cost**



QORE

## Application of optimisation techniques and algorithms for QoS and resource optimisation

**Use of different operation research methods to solve the problem of optimal resource assignment of QoS enabled applications considering specific resource and QoS restrictions of communication jobs**

**•Assignment problem,**
**•Critical path method,**
**•Max flow problem,**
**•Shortest path,**
**•Transportation, cost flow problems**



QORE

## Restrictions for operation research methods

**Elastic stream applications**

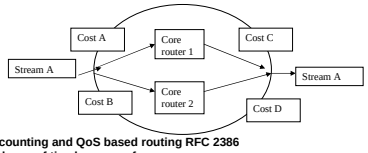
- **Delay or rate elasticity**
- **levels of bandwidth reservation**
- **Maximum Delay / minimum rate threshold -> minimum level of bandwidth reservation**
- **duration NOT dependent on bandwidth reservation and QoS level**

**Elastic data transfer (burst) applications**

- **Throughout elasticity**
- **levels of bandwidth reservation**
- **throughput threshold -> minimum level of bandwidth reservation**
- **duration dependent on bandwidth reservation and QoS level**

## QORE

### Timing and cost restrictions




**Framework**

- IP based accounting and QoS based routing RFC 2386
- Cost dependency of timely usage of resources

**Cost vector**

- characterising resource requirements of QoS based application
- Minimisation of cost based resource usage

Cost A: Cost A1, Cost A2,…… Cost An dependent on the period T1, T2, …. Tn
Cost B: Cost B1, Cost B2,…… Cost Bn dependent on the period T1, T2, …. Tn
Cost C: Cost C1, Cost C2,…… Cost Cn dependent on the period T1, T2, …. Tn
Cost D: Cost D1, Cost D2,…… Cost Dn dependent on the period T1, T2, …. Tn



## QORE

### Structure

- Motivation
- Goals
- QORE Optimisation Issues
- QORE Architectural Issues
- Ongoing Work – Intermon
- Summary



## QORE

### Optimisation considering timing requirements

***Optimal resource assignment for QoS based applications with different time requirements***

**Time requirements of multimedia and distributed application:**

- same start, end and duration: multiplexed multimedia applications – multimedia conferences
- start / end / duration relationships: distributed applications
- flexible start time and same end time: time dependent content applications
- flexible start / end: VoIP, data transfer
- flexible start /fixed end: multimedia synchronisation

QORE

## QORE Optimisation research problem

Parameter Restrictions

**1. Time Restrictions of communication jobs**

- start synchronisation
  Sk = Sl , k=1….n, l=1…...n, n <= N
  where Sk, Sl start time of job k, resp.l
- **multimedia multiplexing**
  Dk = Dl , k=1….r, l=1…...r, r <= N
  where Dk, Dl duration of job k, resp.l
- **precedence relation**
  Sk > El , k=1….n, l=1…...n, n <= N
  where Sk start time of job k, El end time of job l
- **end synchronisation**
  Ek = El , k=1….n, l=1…...n, n <= N
  where Ek, El start time of job k, resp.l



QORE

## QORE Optimisation - Parameter restrictions

**2. Elasticity restrictions**

- different QoS levels
  Qmin =< Qk =< Qmax k= 1…..N
  Where Qk – QoS level of job k,
  Qmin – minimum QoS level threshold
  Qmax – maximum QoS level threshold

- different resource levels
  Rmin =< Rk =< Rmax k= 1…..N
  Where Rk – Resource level of job k,
  Rmin – minimum resource level
  Rmax- maximum QoS level

**3. Cost vector restrictions**

- different costs for resource usage
  Cmin =< Ck =< Cmax k= 1…..N
  Where Ck – cost of job k,
  Cmin – minimum accepted job cost
  Cmax – maximum allowed job cost



QORE

## QORE Optimisation -- Goal Function

**Maximisation of the total efficiency metric E as sum of efficiency metric of particular jobs**

**Efficiency metric is defined to consider QoS, resource levels and cost vector requirements of jobs (applications)**

**E = Sum of Ei ->max i= 1…..N**

Where Ei = (Di x Qi) / (Ci x Ri)

Ei – efficiency metric of job i
Di – duration of job i
Qi – QoS level of job i
Ci – Cost level of job i
Ri – Resource level of job i
N – Number of jobs

Where the global resource restriction GR is fulfilled
- R = Sum of (Ri x Di) < = GR

QORE

## Operation research problems for Optimal Resource Allocation

- Bandwidth packing problem
  - assigning calls to paths in a capacitated graph, such that capacities are not violated and the total profit is maximized.
  - Usage of Tabu Search Method with adaptation of k-shortest path
- Optimal Resource Allocation In Networks with routing (RAIN)
  - Find one route for each demand in a network, so that the bandwidth requirements of demands are satisfied within resource capacities of the bidirecional links of the network
  - NP-hard for demands subject to multiple additive or multiplicative QoS criteria
- Knapsack problem
  - capacity of edge corresponds to the size of knapsack and the calls correspond to objects to be packed in the knapsack -> NP-hard
- Call Admission Control
  - BCA (Batch Call Admission): all calls request to be established at the same time. In particular, BCA includes the case where all call instances have the same starting-time. If alternatives are not allowed, BCA is equivalent to the unsplittable flow problem
  - GCA (general call admission): every request specifies a starting time and duration



QORE

Optimisation techniques (2)

## Operation research problems for Optimal Resource Allocation

- Scheduling problem for allocating scarce resources to activities over a period of time.
  - a discrete optimization problem and therefore very hard to solve in practice (most of the scheduling problems are NP-Hard).
  - Efficient solutions including Mixed Integer Linear Programming, Branch-and-Bound
- Constraint satisfaction problem
  - defined by a triple (X, D, C), where X = {x1,, ..., xn} is a set of variables, D = {D1,, ..., Dn} a set of finite domains associated with the variables, and C = {C1,, ..., Cm} a set of constraints
  - The question is whether there exists an assignment of values to variables, so that all the constraints are satisfied
  - Using constraints to reduce the computational effort needed to solve combinatorial problems



QORE

Measurement based Resource and QoS Planning (2)

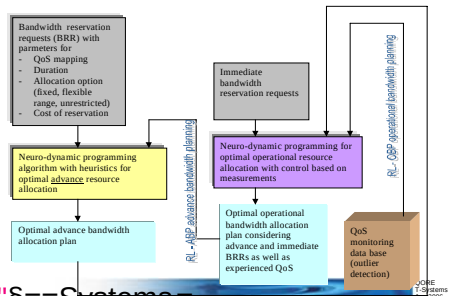
## Reinforcement Learning for adaptable bandwidth planning using measurements

- Machine learning theory that derives its roots from control principles effectively used to solve problems with in communication networks
  - Learning agent has to formulate a policy, which determines the appropriate action to take in each state in order to maximize the expected cumulative reward over time
  - The reward is derived from how favorable the outcome is of the action taken by the agent in a particular state
- Control mechanism to change the planned optimal bandwidth allocation considering actual QoS analysis in an operational environment
  - In case of QoS congestion, the current optimal advance allocation plan is adapted (recalculated) with a strategy for later allocation of the advance resource reservation

QORE

## Measurement based Resource and QoS Planning (3)

**Integration of Neuro-dynamic framework with QoS monitoring data base**



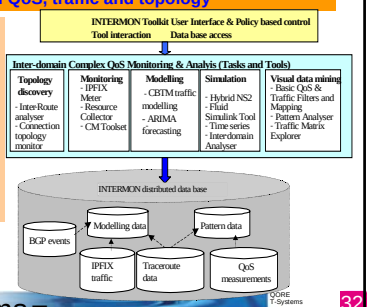



QORE

## Measurement based Resource and QoS Planning (4)

**Bandwidth optimisation with feedback from monitoring and analysis of QoS, traffic and topology**

- Data base of QoS measurement and topology tools of INTERMON provide end-to-end information per required time scale for QoS and resource planning
- Integration of INTERMON data base in a reinforcement learning strategy for QoS and resource planning of inter-domain connections



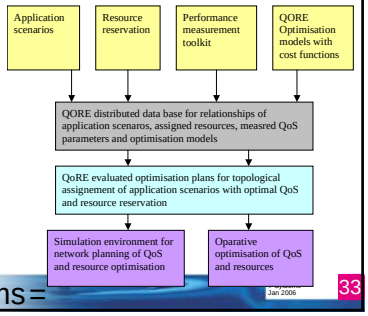



QORE

## Measurement based Resource and QoS Planning (5)

**QORE system for bandwidth optimisation based on QoS monitoring and modelling data - design**

- Bandwidth optimisation strategies for different kind of applications considering Grid, VoIP and streaming multimedia
- Optimisation with feedback from QoS measurement and patterns



# QORE

## Structure

- Motivation
- Goals
- QORE Optimisation Issues
- QORE Architectural Design and Scenarios
- Ongoing Work – Intermon
- Summary



# QORE

## QORE architecture goals

**Measurement based QoS and resource optimisation based on operation research approach**

**QoS and resource optimisation as input for**

- modelling and simulation of QoS and resource environments
- operative management and control of resources and application traffic in Internet

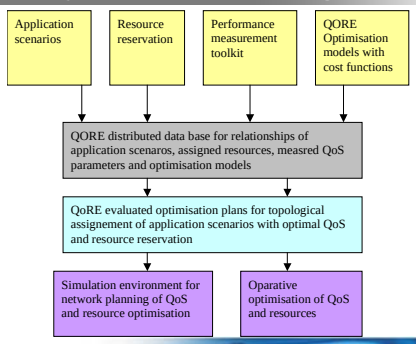
**Integration of tools using distributed data base**

- QoS parameter measurement dependent on applications
- Resource assignment using access channel approach
- Scenario (set) configuration with resource, QoS and cost level requirements
- Optimisation specifications (plans)
- Operation research algorithms dependent on optimisation specifications and scenario (set) configurations



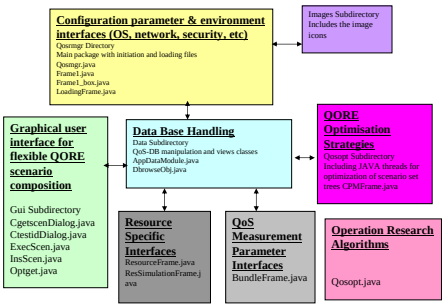
# QORE

## QORE System – Architectural Design



# QORE

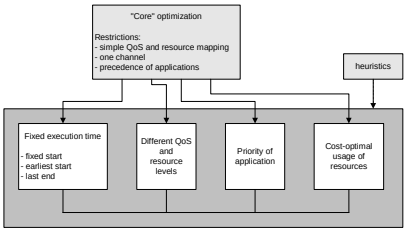
## QORE Components






# QORE

## Concept of QORE optimisation algorithms






# QORE

## QORE Optimisation

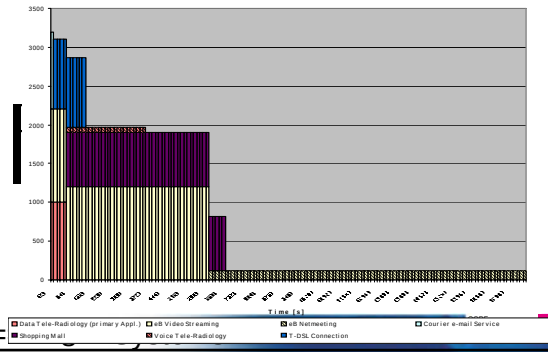
**Optimisation algorithms** consider the restrictions
- QoS levels and QoS tuning. The mapping of application QoS levels to network resources of the access channels is done considering the measurements stored in the data base.
- Time-dependent costs for usage of access channels
- Scenario characteristics
- Timing relationships (precedence)
- Priorities of applications and scenarios.

**Optimization plans",** for scenario sets:
- QoS and resource restrictions of the scenarios,
- available resources for the scenario set.

**Optimization criteria**:
- minimization of costs (assignment of applications and scenarios to the lowest cost usage of the access channels)
- earliest execution time (assignment of applications and scenarios to the earliest usage of the access channels).

## QORE

### QORE Optimisation using mixed bandwidth algorithm




==



## QORE

### Structure

- Motivation
- Goals
- QORE Optimization Issues
- QORE Architectural Design
- Ongoing Work – Intermon
- Summary

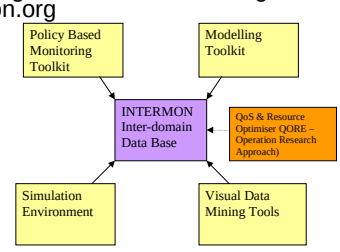
====!"§==Systems=



## QORE

### QORE Integration in European IST projects

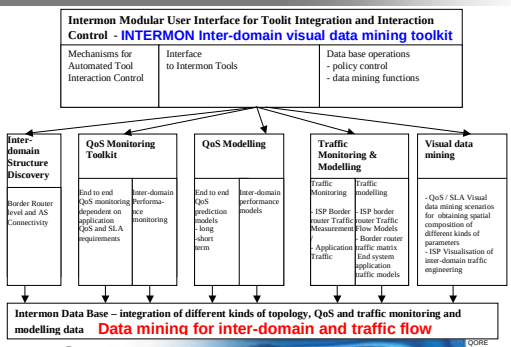
Intermon IST Project for Inter-domain monitoring , modelling and visual data mining : www.ist-intermon.org




====!"§==Systems=

QORE

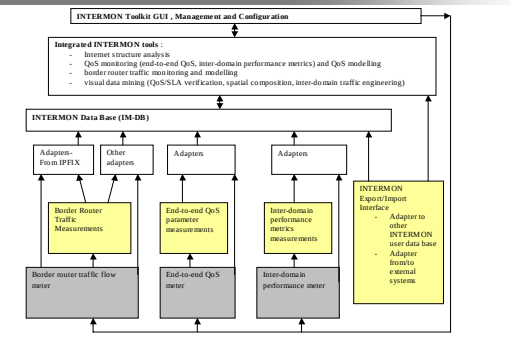
## Intermon Structure






QORE

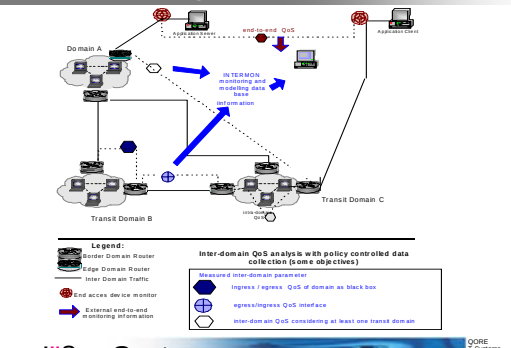
## Intermon Architecture





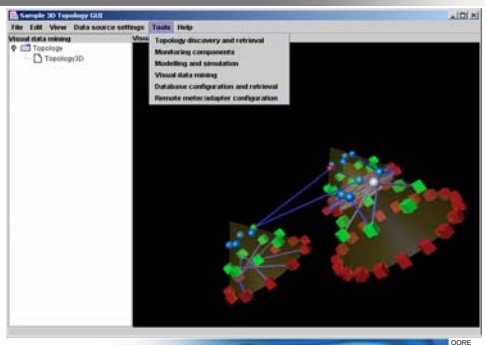

QORE

## Intermon Test Configuration



QORE

## Intermon Visualisation






QORE

## Structure

- Motivation
- Goals
- Optimisation Issues
- QORE Architectural Issues
- Ongoing Work – Intermon
- Summary



QORE

## Summary

**•Optimal resource modelling for QoS based Virtual Private Networks**
**•Optimisation of resource scheduling for QoS based applications with different time requirements**
**• Optimal delay tolerance**
**• Optimisation of QoS and resource allocation levels of elastic applications**
**• Optimal resource assignment for cost restricted QoS based applications**
**• Flexible resource allocation for QoS support of mobile access networks**