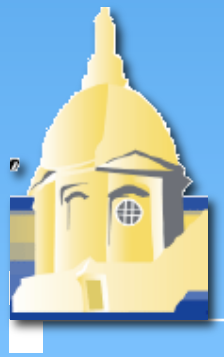
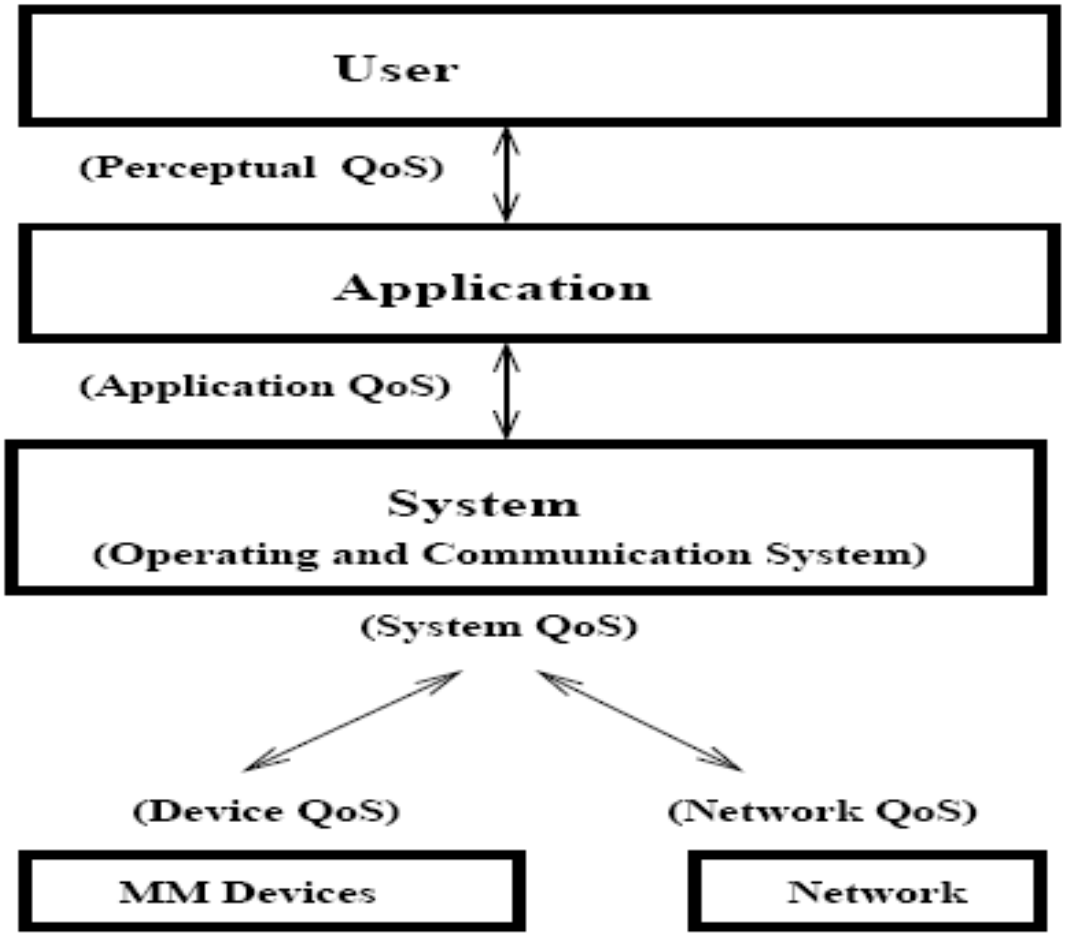
Quality of Service

- Quality of service measures the kind of service provided by the system
  - On systems that can offer flexible services, QoS allows us to compare the service received
- MM systems consist of set of services
- Examples of Multimedia QoS parameters:
  - QoS for Audio service:
    - Sample rate – 8000 samples/second
    - Sample resolution – 8 bits per sample
  - QoS for network service:
    - Throughput – 100 Mbps
    - Connection setup time – 50 ms

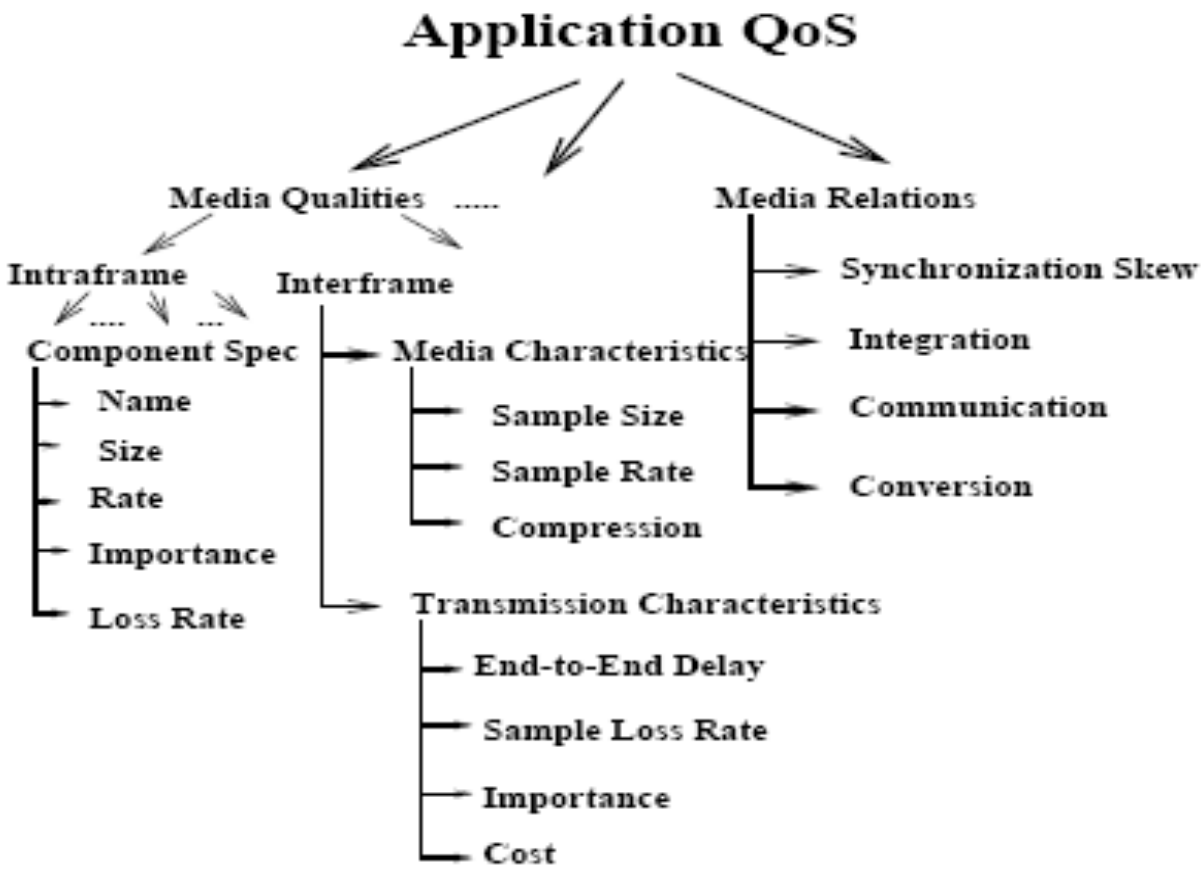


Slides courtesy Prof. Nahrstedt

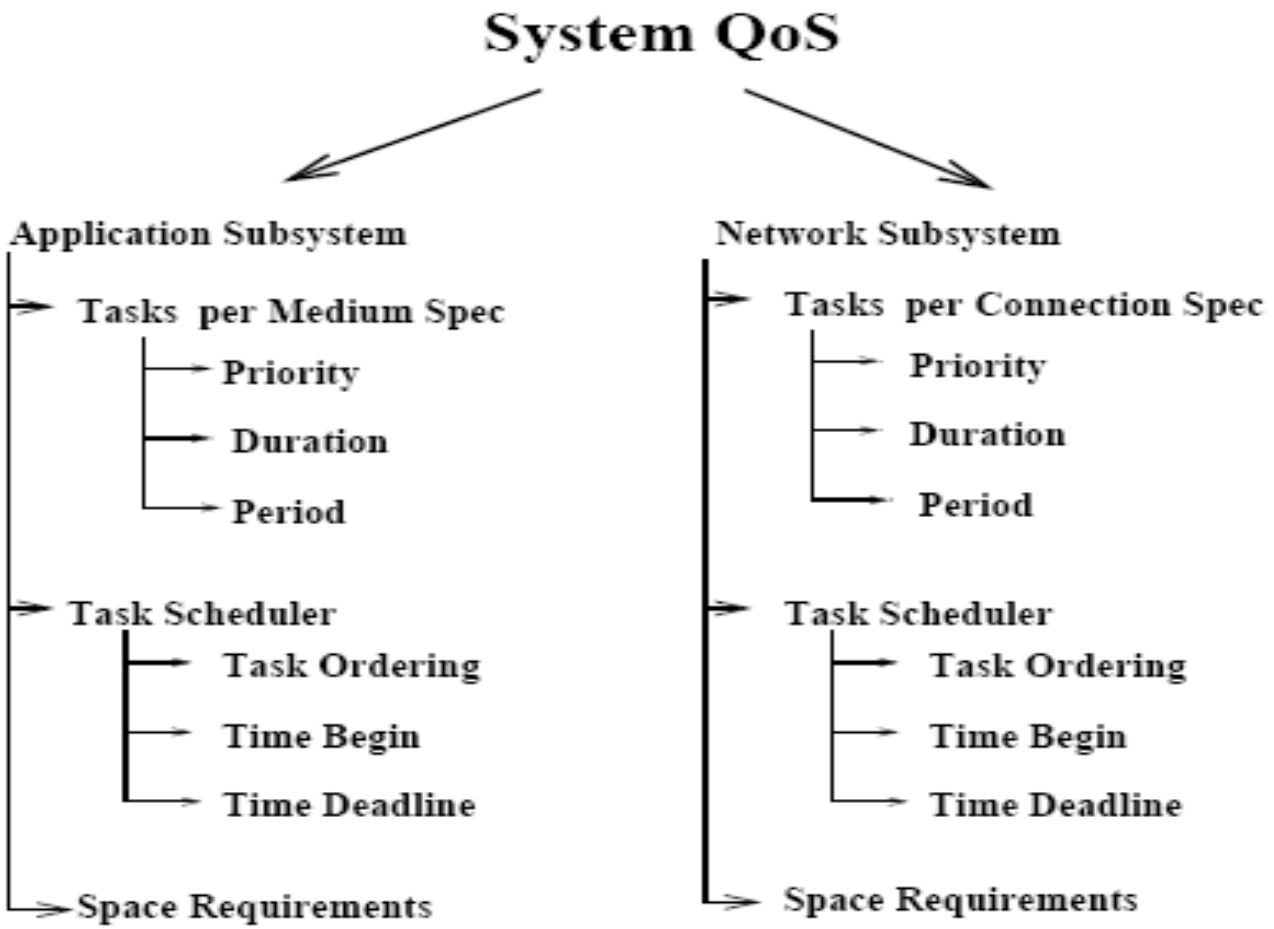
Layered Model for QoS



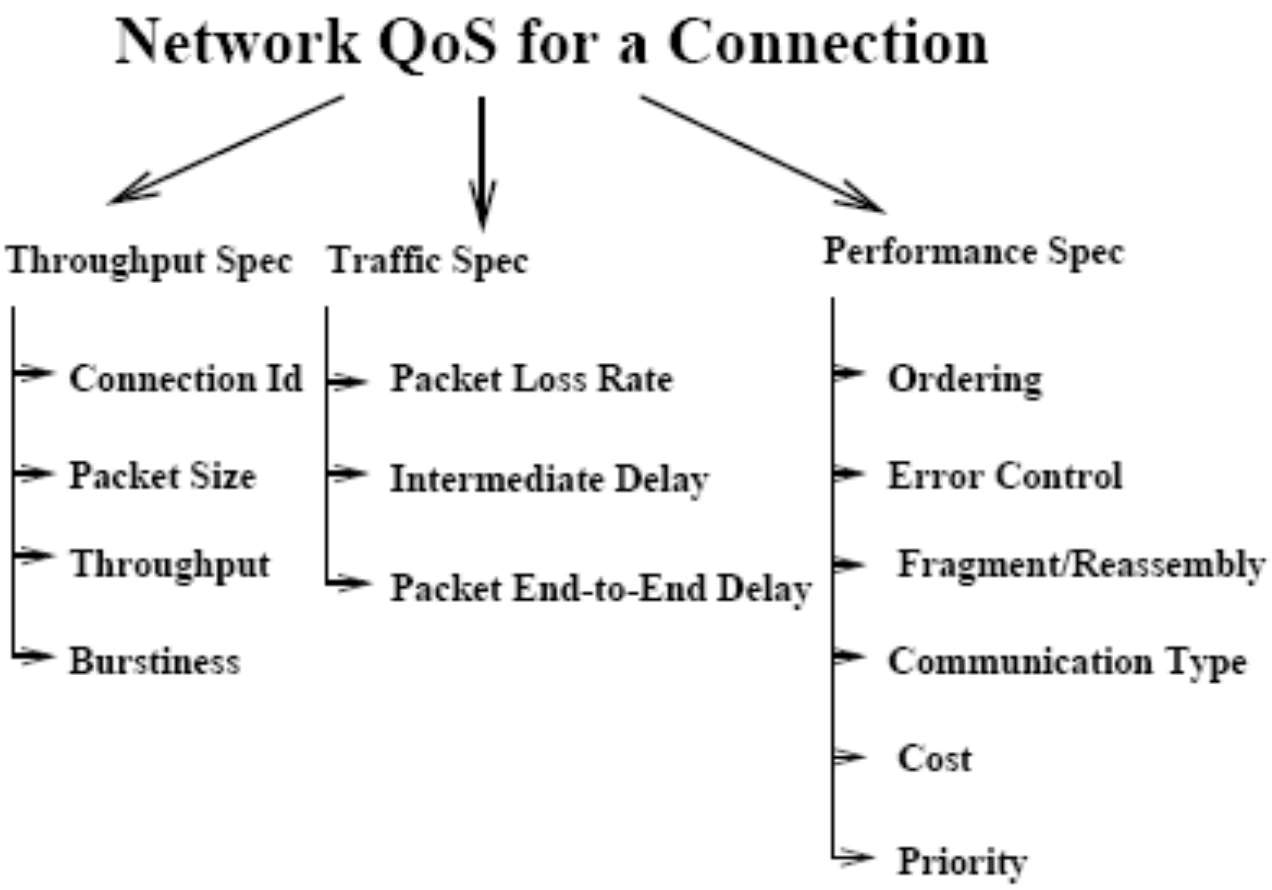
Application QoS Parameters

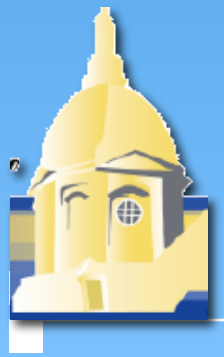


System QoS Parameters



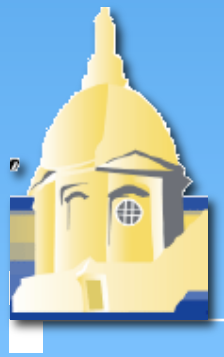
Network QoS Parameters





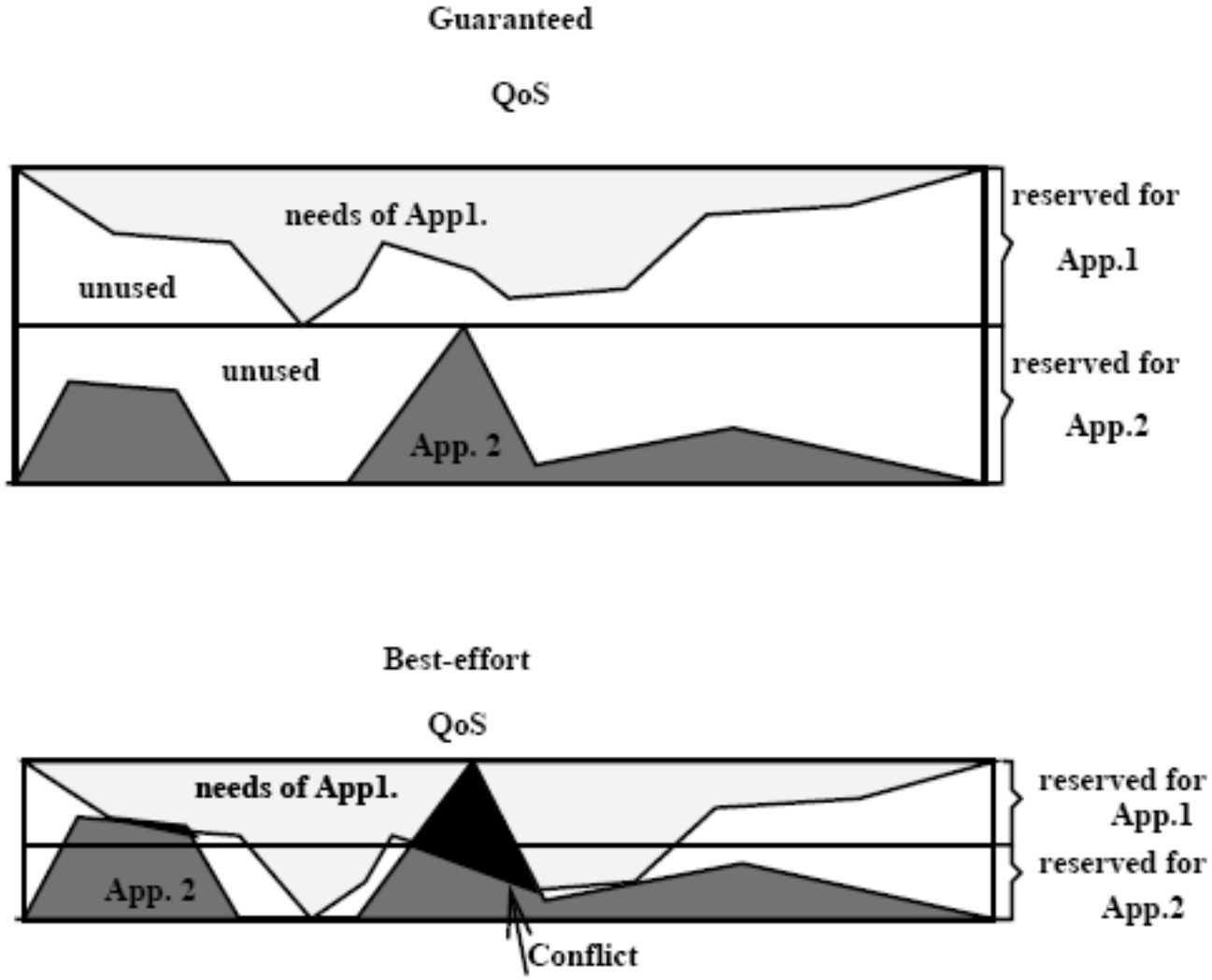
QoS Classes

- Guaranteed Service Class
  - QoS guarantees are provided based on deterministic and statistical QoS parameters
- Predictive Service Class
  - QoS parameter values are estimated and based on the past behavior of the service
- Best Effort Service Class
  - There are no guarantees or only partial guarantees are provided



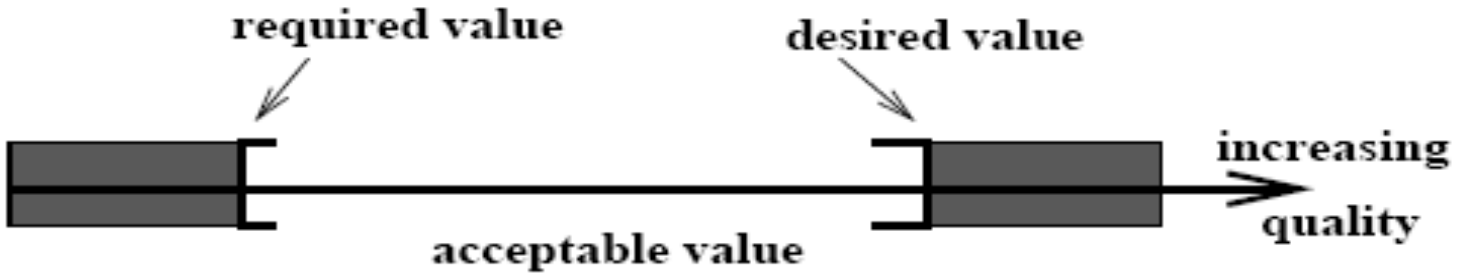
QoS Classes (cont.)

QoS Class determines: (a) reliability of offered QoS, (b) utilization of resou



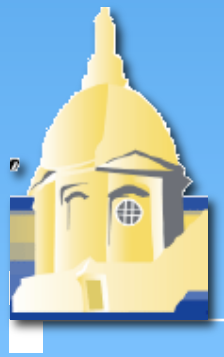
Deterministic QoS Parameters

- Single Value: QoS_1 – average (QoS_{ave}), contractual value, threshold value, target value
- Pair Value: <QoS_1, QoS_2> with
  - QoS_1 – required value; QoS_2 – desired value
  - Example: <QoS_{avg}, QoS_{peak}>; <QoS_{min}, QoS_{max}>



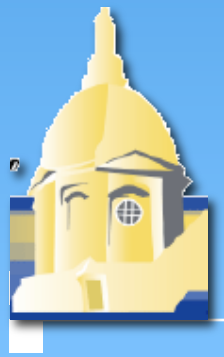
Deterministic QoS Parameter Values

- Triple of Values <QoS_1, QoS_2, QoS_3>
  - QoS_1 – best value
  - QoS_2 – average value
  - QoS_3 – worst value
- Example:
  - <QoS_{peak}, QoS_{avg}, QoS_{min}>, where QoS is network bandwidth



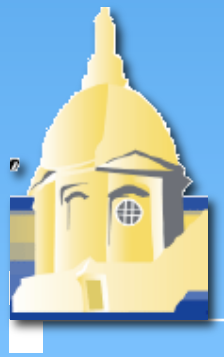
Guaranteed QoS

- We need to provide 100% guarantees for QoS values (hard guarantees) or very close to 100% (soft guarantees)
- Current QoS calculation and resource allocation are based on:
  - Hard upper bounds for imposed workloads
  - Worst case assumptions about system behavior
- Advantages: QoS guarantees are satisfied even in the worst case case (high reliability in guarantees)
- Disadvantage: Over-reservation of resources, hence needless rejection of requests



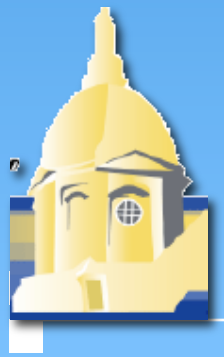
Predictive QoS Parameters

- We utilize QoS values (QoS1, ..QoSi) and compute average
  - QoSbound step at K>i is QoSK = 1/i*∑jQoSj
- We utilize QoS values (QoS1, , QoSi) and compute maximum value
  - QoSK = max j=1,…i (QoSj)
- We utilize QoS values (QoS1, , QoSi) and compute minimum value
  - QoSK = min j=1,…i (QoSj)



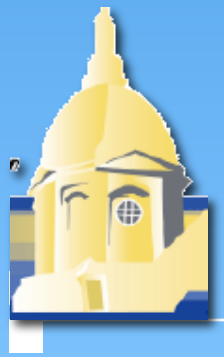
Best Effort QoS

- No QoS bounds or possible very weak QoS bounds
- Advantages: resource capacities can be statistically multiplexed, hence more processing requests can be granted
- Disadvantages: QoS may be temporally violated



Quality-aware Service Model

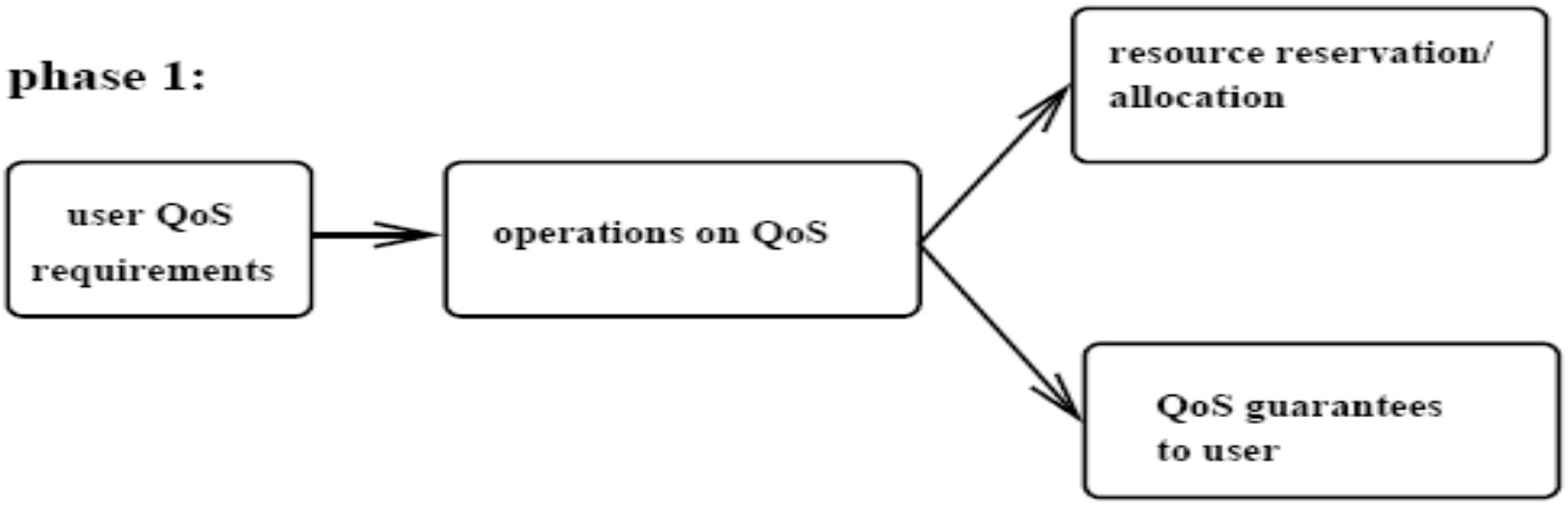
- Quality-aware Autonomous Single Service
  - Consists of a set of functions
  - Accepts input data with QoS level QoSin
    - QoSin=[q1in,..qnin]
  - Generates output data with QoS level QoSout
    - QoSout=[q1out,..qnout]
- Example: Video player service
  - Input QoS: [Recorded Video Frame Rate, Recorded Frame Size, Recorded Pixel Precision]
  - QoSin=[30fps, 640x480 pixels, 24 bits per pixel]
  - Output QoS: [Playback Video Frame Rate, Playback Frame Size, Playback Pixel Precision]
  - QoSout=[20fps, 320x240pixels, 24bits per pixel]



Quality-aware Service Model

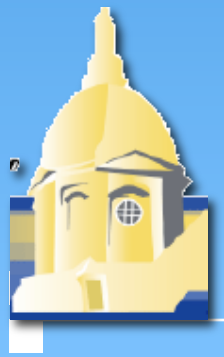
- Quality-aware Composite Service
  - Consists of set of autonomous services that are connected into a directed acyclic graph, called service graph
  - Is correct if the inter-service satisfied the following relation:
    - QoSout of Service K ‘satisfies ‘QoSin of Service M iff
    - qKjout = qMlin for qMlin being single QoS value
    - qKjout is in qMlin for qMlin being a range of QoS value
- Example:
  - Video-on-demand service, consists of two services: retrieval service and playback service
    - Output quality of the retrieval service needs to correspond to input quality of playback service, or at least falls into the range of input quality of playback service

Relation between QoS and Resources

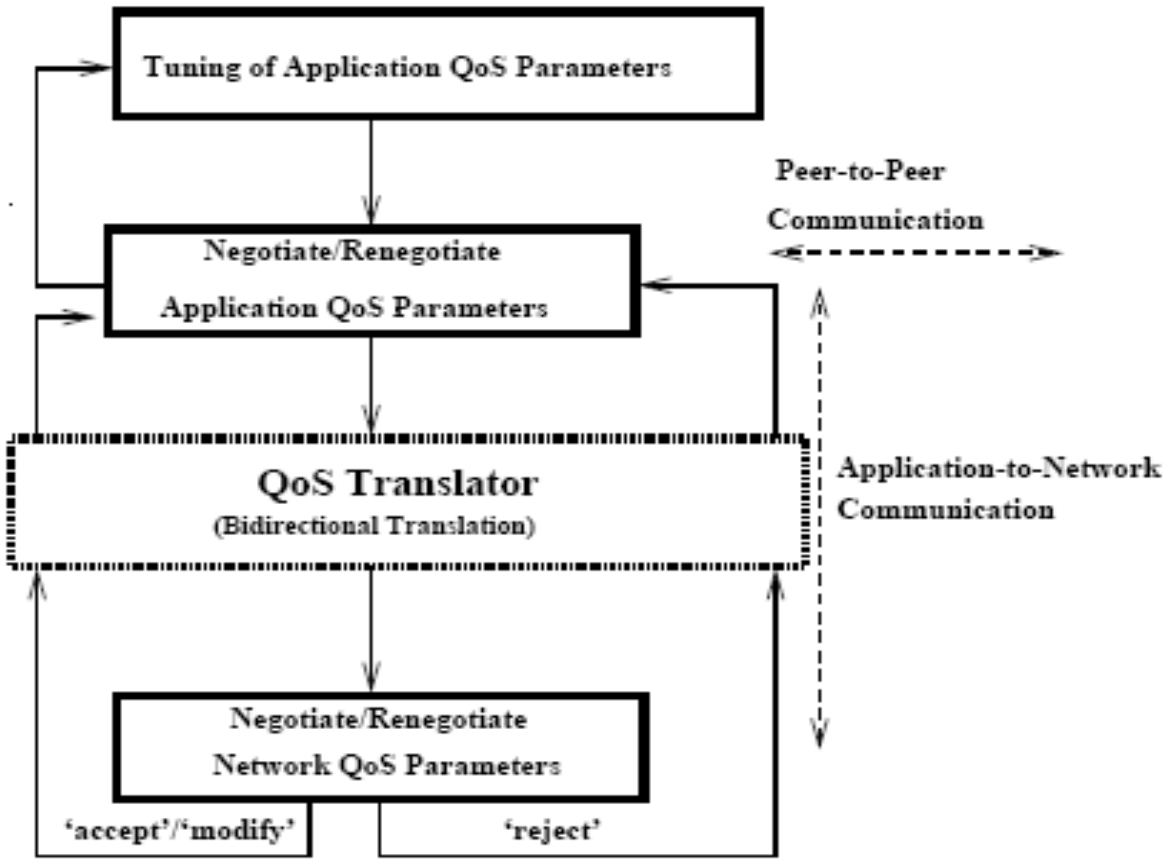


Phase 1: Establishment Phase (QoS Operations)

- QoS Translation at different Layers
  - User-Application
  - Application-OS/Transport Subsystem
- QoS Negotiation
  - Negotiation of QoS parameters among two peers/components
- QoS Routing along the end-to-end path



QoS Operations within Establishment Phase



User/Application QoS Translation

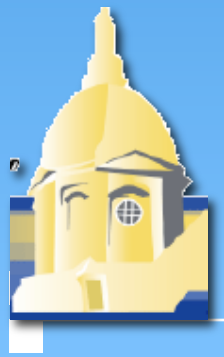
Overlay P2P QoS Negotiation

Application/Transport QoS Translation

QoS Negotiation/ QoS Routing in Transport Subsystem

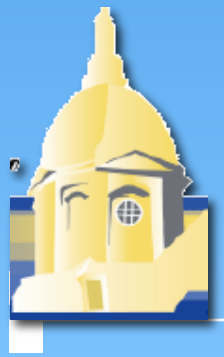
Operations on QoS in Phase 1 (QoS Translations)

- Layered Translation of QoS parameters (must be bidirectional)
  - Human (user QoS) – application QoS
  - Application QoS – system QoS
  - System QoS – network QoS
- Media Scaling
  - Transparent scaling
  - Non-transparent scaling

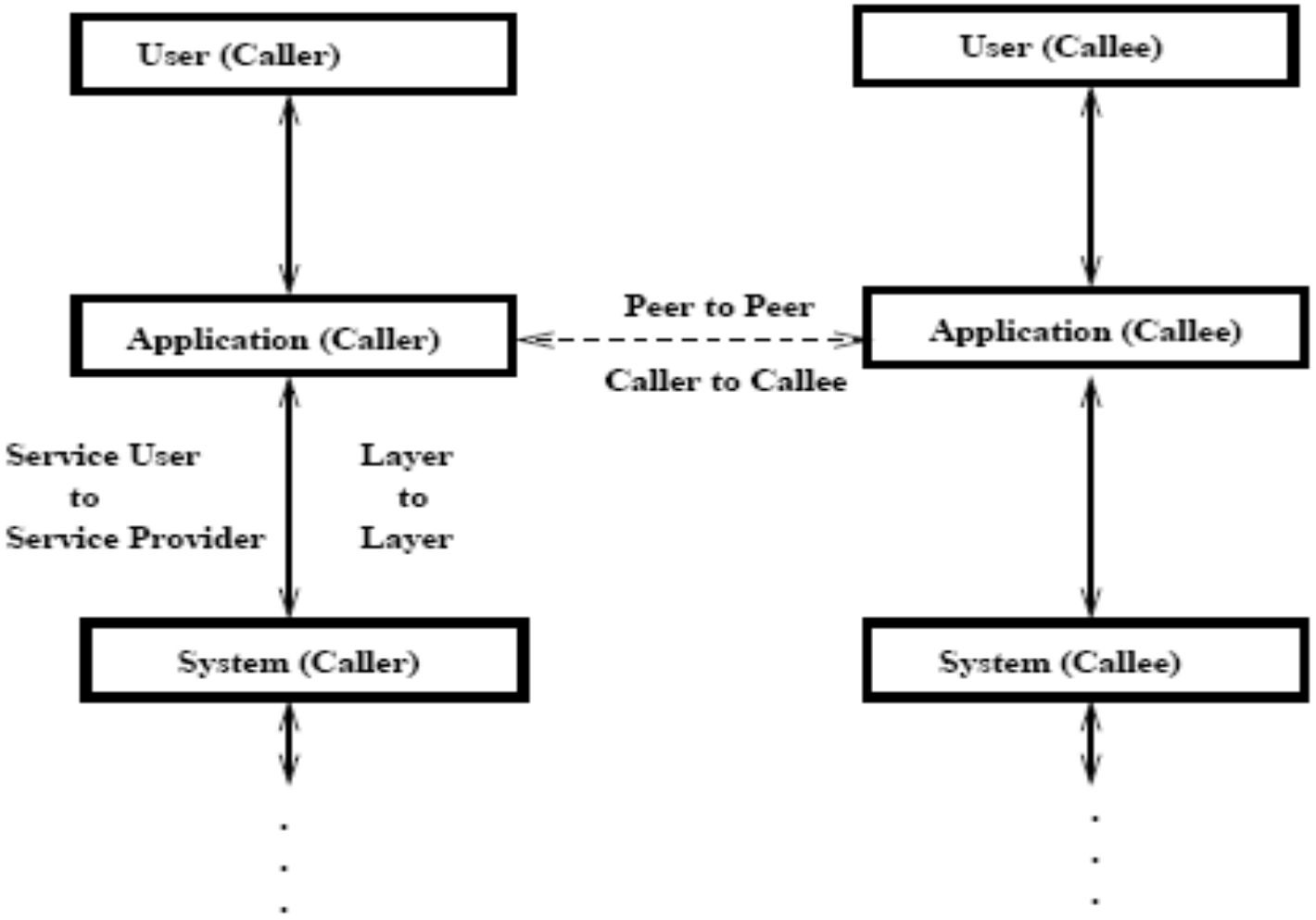


Media Scaling (Examples)

- Audio
  - Transparent scaling difficult (one hears the quantization noise)
  - Non-transparent scaling should be used
- Video
  - Temporal scaling
  - Spatial scaling
  - Color space scaling (reduction of number of entries in color space)

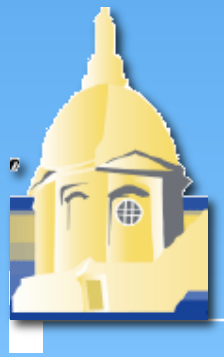


QoS Negotiation

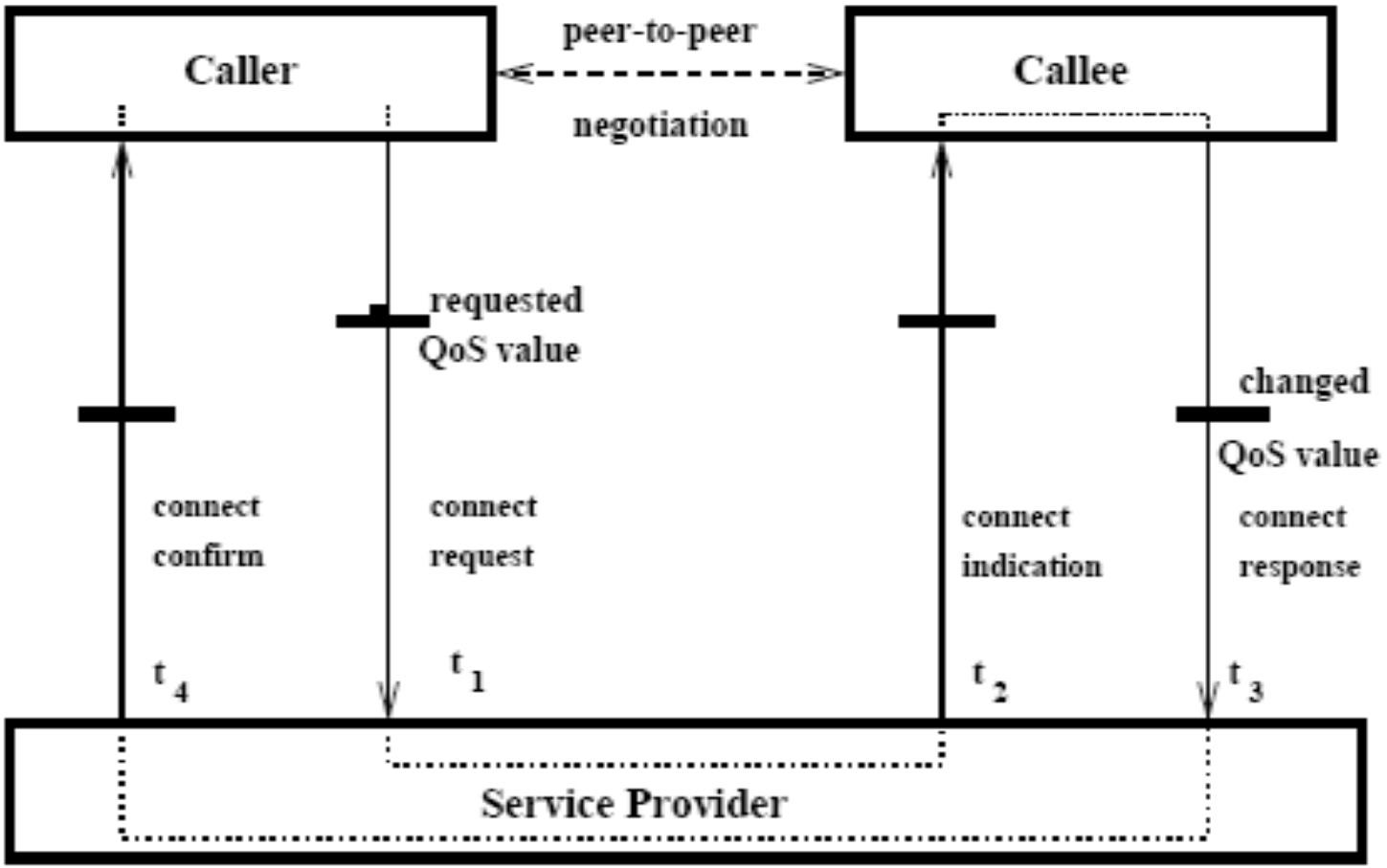


Different Types of Negotiation Protocols

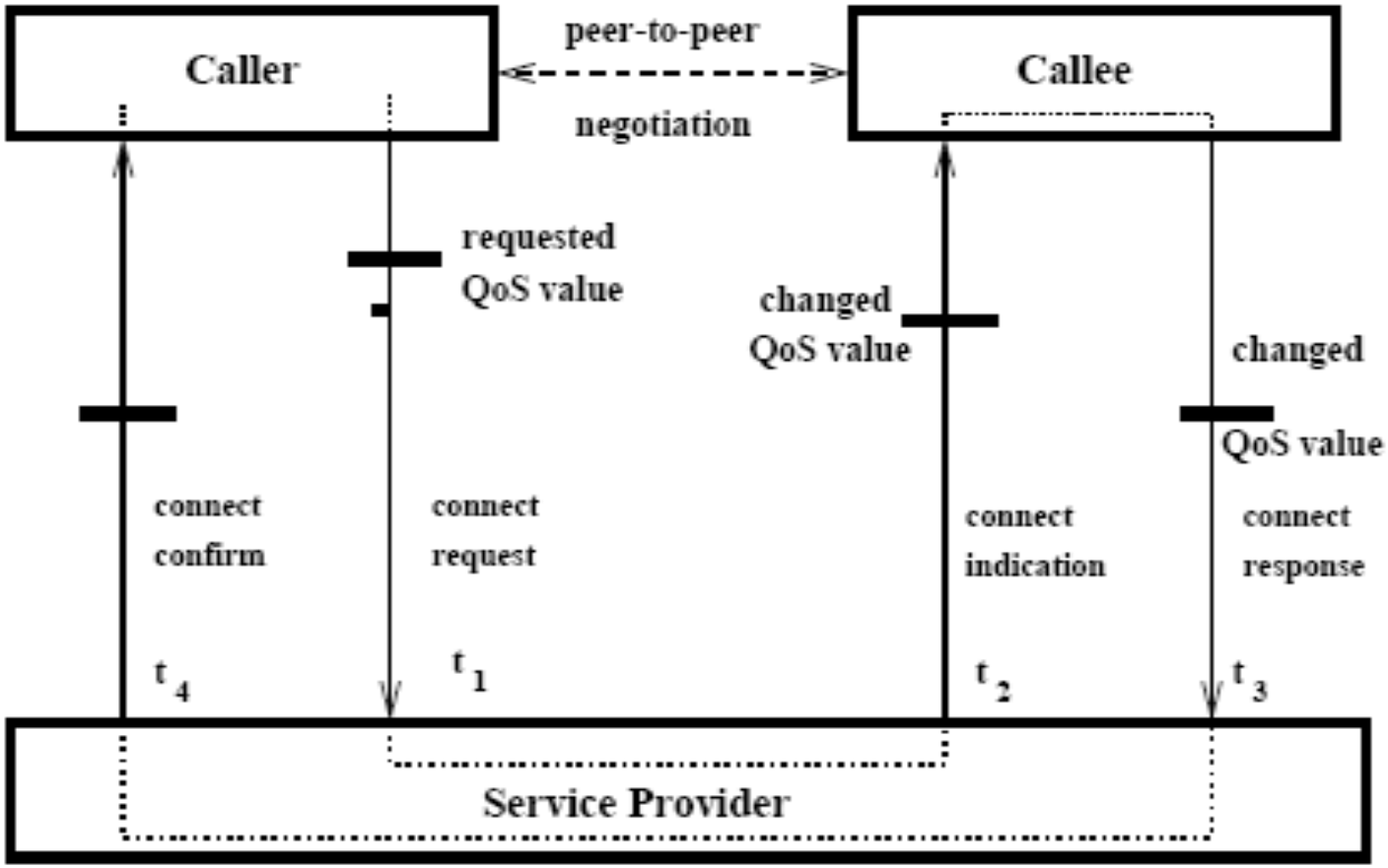
- Bilateral Peer-to-Peer Negotiation
  - Negotiation of QoS parameters between equal peers in the same layer
- Triangular Negotiation
  - Negotiation of QoS parameters between layers
- Triangular Negotiation with Bounded Value



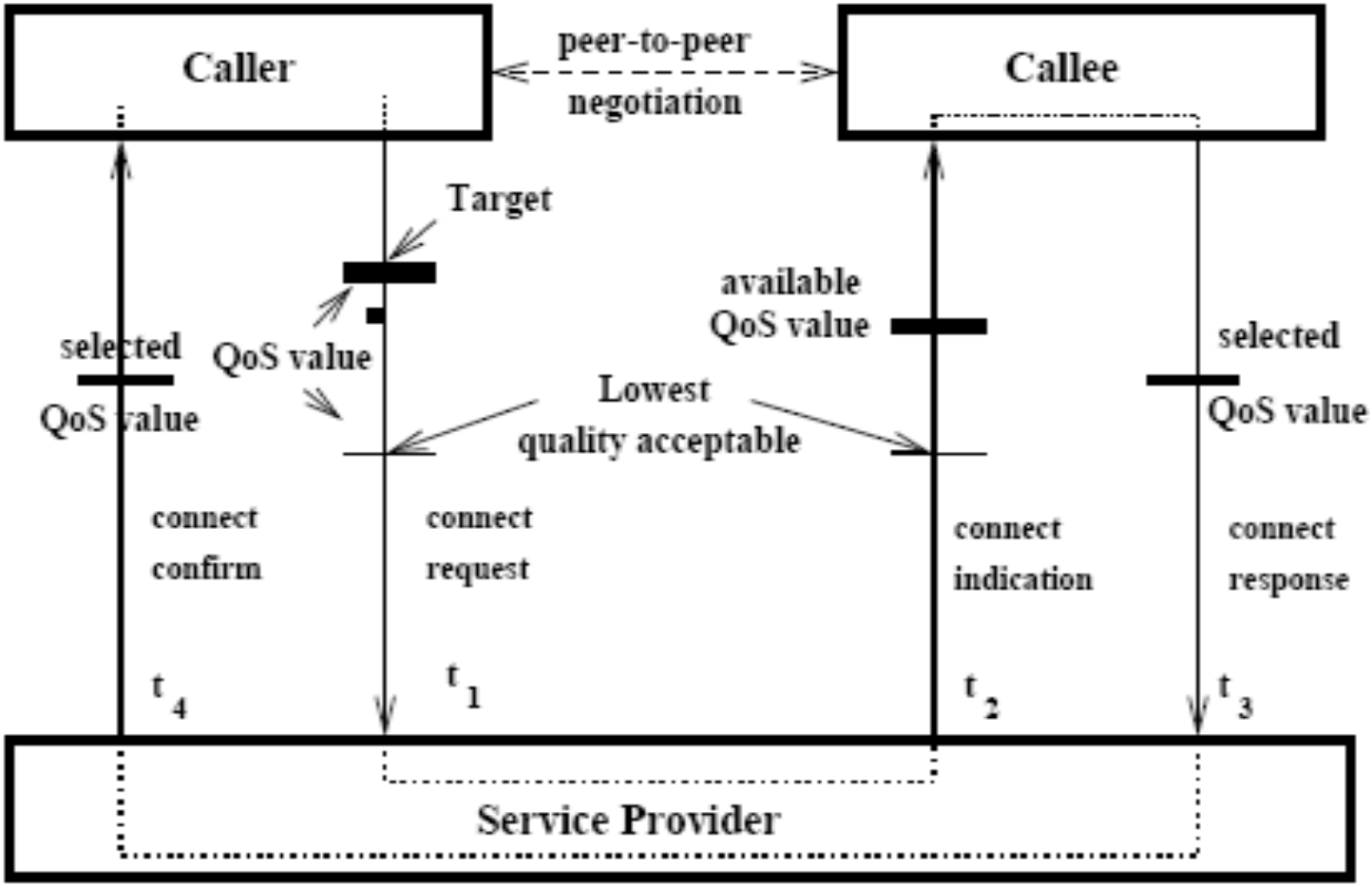
Bilateral QoS Negotiation



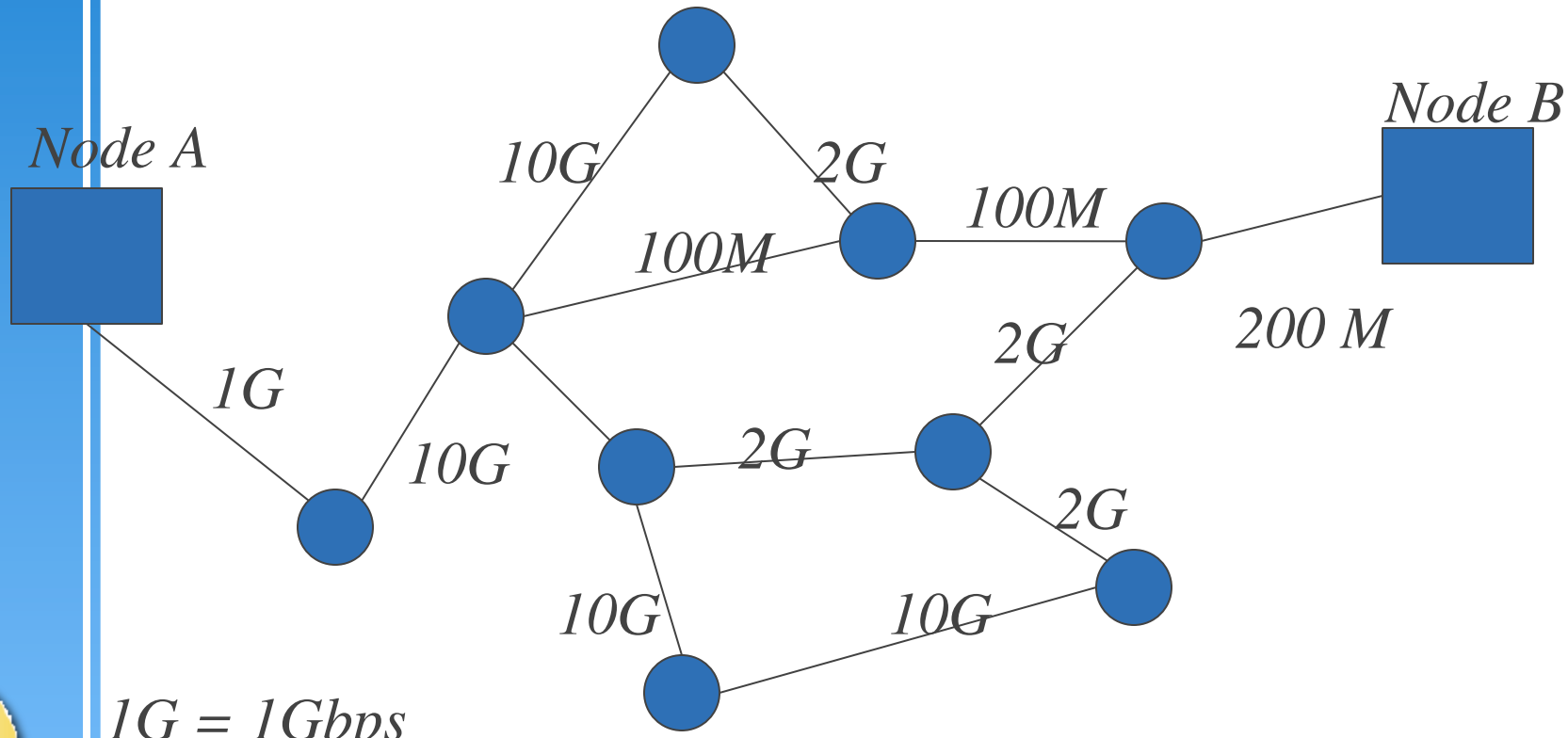
Triangular QoS Negotiation



Triangular Negotiation with Bounded Value



QoS Routing

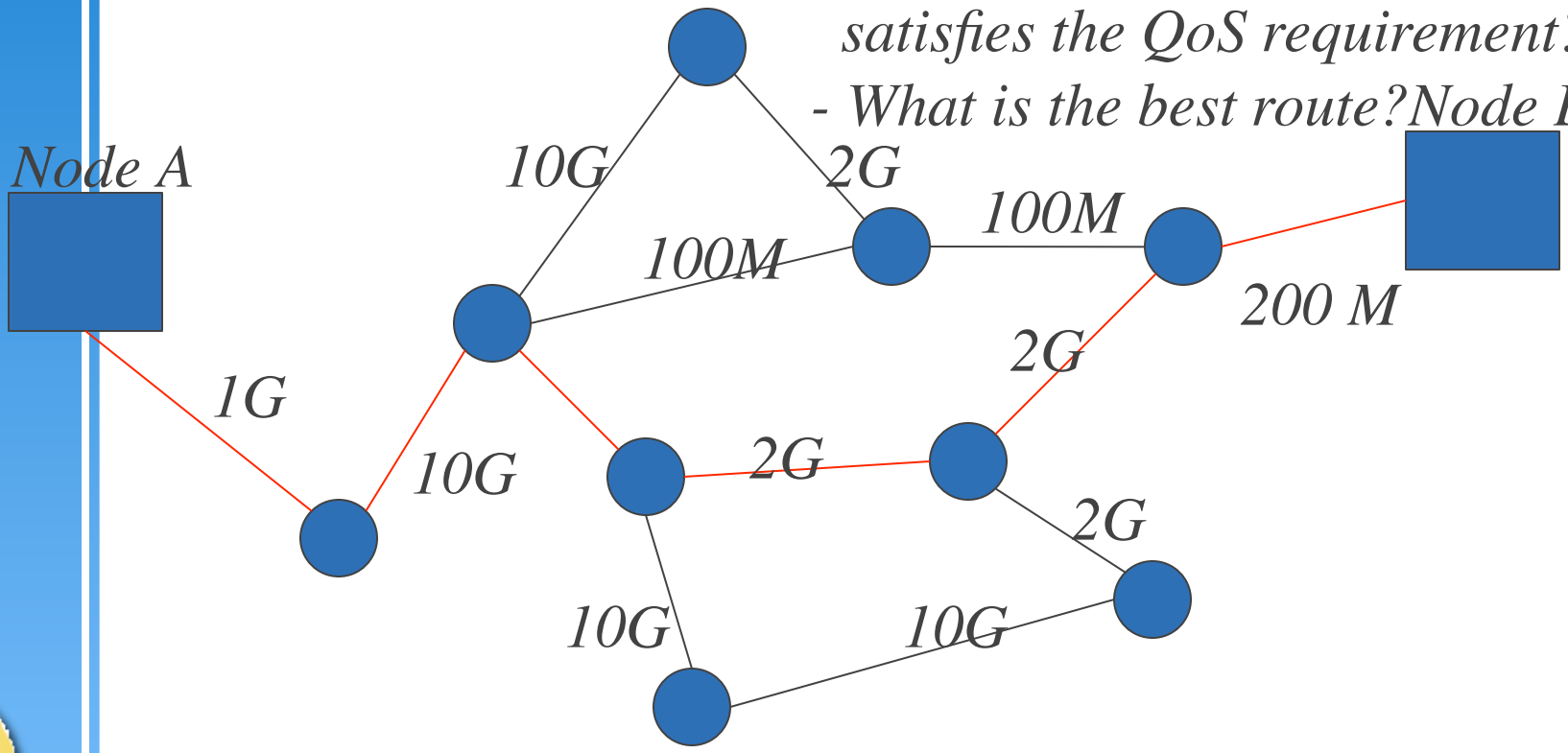

Node A
Node B
10G
2G
100M
100M
1G
10G
2G
2G
2G
200 M
10G
10G
1G = 1Gbps
10G = 10 Gbps
100 M = 100 Mbps
End Node
Network Router

QoS Routing

If QoS Request on a connection from Node A to B is 150 Mbps, the QoS Routing question is

-Does a route from A to B exist that satisfies the QoS requirement?

- What is the best route?



1G = 1Gbps

10G = 10 Gbps

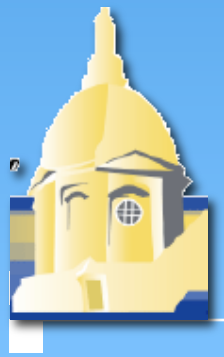
100 M = 100 Mbps

End Node

Network Router

QoS Routing

- Performed during establishment phase mostly, but also during transmission phase to adapt a route if needed
- Need to discover route (path) that meets QoS requirements such as throughput, end-to-end delay, loss rate
  - End-to-end Throughput is a min-based metric
  - End-to-end Delay is additive metric

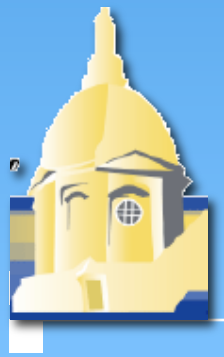


Unicast QoS Routing

- Problem Formulation:
  - Given a source node A, destination B, a set of QoS constraints C, and possibly an optimization goal, we aim to find the best feasible path from A to B which satisfies C.
- Bandwidth-optimization problem: to find a path that has the largest bandwidth on the bottleneck link (widest path)
- Bandwidth-constrained problem: to find a path whose bottleneck bandwidth is above a required threshold value
- Delay-optimized problem: to find a path whose total delay is minimized
- Delay-constrained problem: to find a path whose delay is bounded by a required value.

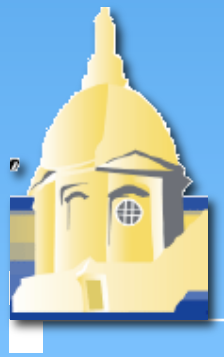
QoS Routing Strategies

- Source Routing
  - Each node maintains global state and feasible path is locally computed at the source node
- Distributed Routing
  - Control messages exchanged among nodes and the state information kept at each node is collectively used for the path search
- Hierarchical Routing
  - Nodes are clustered into groups creating multi-level hierarchy
  - One can use source routing within a cluster and distributed routing among clusters



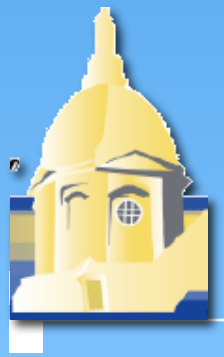
Multimedia Resource Management

- Resource managers with operations and resource management protocols
  - Various operations must be performed by resource managers in order to provide QoS
- Establishment Phase
  - Operations are executed where schedulable units utilizing shared resources must be admitted, reserved and allocated according to QoS requirements
- Enforcement Phase
  - Operations are executed where reservations and allocations must be enforced, and adapted if needed



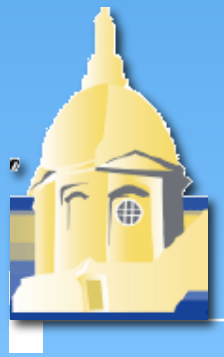
Establishment Phase Operations

- QoS to Resource Mapping
  - Need translation profiles
- Resource Admission
  - Need admission tests to check availability of shared resources
- Resource Reservation
  - Need reservation mechanisms along the end-to-end path to keep information about reservations
- Resource Allocation



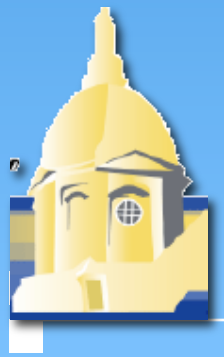
Continuous Media Resource Model

- One possible resource utilization model for multimedia data – Linear Bounded Arrival Process Model (LBAP)
- LBAP models message arrival process:
  - M – maximum message size (in bytes)
  - R – maximum message rate in messages per second
  - B – maximum burstiness (accumulation of messages)



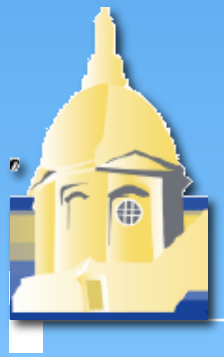
LBAP Resource Model

- If we have (M,R,B), we can predict utilization of resources:
  - Maximum number *N* of messages arriving at the resource: $N = B + R \times TimeInterval$
  - Maximal Average Rate *R' (*in bytes per second): $R' = M \times R$
  - Maximal Buffer Size (BS in bytes): $BS = M \times (B+1)$



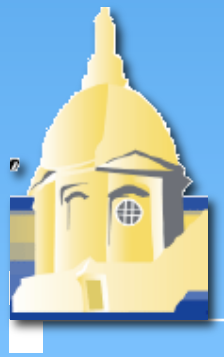
Example of LBAP

- Consider M = 1176 Bytes per message, R = 75 messages per second, B = 10 messages
- During a time interval of 1 second, the maximum number of messages arriving at a resource must not exceed N = 10 messages + (75 messages/ second * 1 second) = 85 messages
- Maximum average data rate in bytes per second is R' = 1176 bytes * 10 messages/second = 88200 bytes/second
- Maximum buffer size in bytes in BS = 1176 bytes * (10 messages + 1) = 12936 bytes



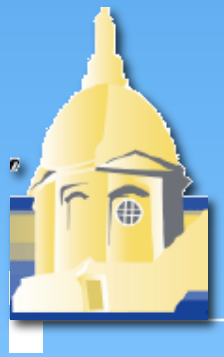
Admission Tests

- Task schedulability tests for CPUs
  - This is done for delay guarantees
- Packet schedulability tests for sharing host interfaces, switches
  - This is done for delay and jitter guarantees
- Spatial tests for buffer allocation
  - This is done for delay and reliability guarantees
- Link bandwidth tests
  - This is done for throughput guarantees



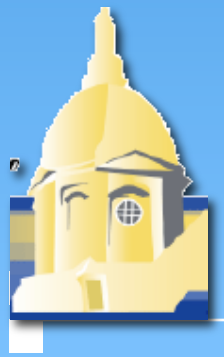
Resource Reservation and Allocation

- Two types of reservations
  - Pessimistic approach - Worst case reservation of resources
  - Optimistic approach - Average case reservation of resources
- To implement resource reservation we need:
  - Resource table
    - to capture information about managed table (e.g., process management PID table)
  - Reservation table
    - to capture reservation information
  - Reservation function
    - to map QoS to resources and operate over reservation table



Resource Reservation

- Two types of reservation styles:
  - Sender-initiated reservation
  - Receiver-initiated reservation



Conclusion – Current State

- Lack of mechanisms to support QoS guarantees
  - Need research in distributed control, monitoring, adaptation and maintenance of QoS mechanisms
- Lack of overall frameworks
  - Need QoS frameworks for heterogeneous environments (diverse networks, diverse devices, diverse OS)