Carnegie Mellon

Models and Algorithms for Hierarchical Resource Management in Integrated Services Networks

Hui Zhang
School of Computer Science
Carnegie Mellon University

Carnegie Mellon

Other Contributors

- Jon Bennett
- Yanghua Chu
- Eugene Ng
- Donpaul Stephens
- Ion Stoica

- Dennis Chang
- Keng Lim
- Jianliang He
- Bo Fan

Carnegie Mellon

Outline

- Introduction
- Hierarchical Generalized Processor Sharing (H-GPS) and Hierarchical Packet Fair Queueing (H-PFQ)
- Hierarchical Fair Service Curve (H-FSC)
- Simulation & experimental results
- Conclusion

Carnegie Mellon

Introduction

- Packet network
  - statistical sharing of network resources
  - packets delayed or dropped during congestion
- Traditional congestion control technique
  - end system dynamically measures available bandwidth
  - adapts to network congestion
  - works only if all sources comply
  - cannot provide differential or guaranteed QoS
- QoS approaches today
  - end-to-end delay or bandwidth guarantee on a per flow basis
  - different QoS classes: guaranteed, statistical, predicted, controlled load, differential

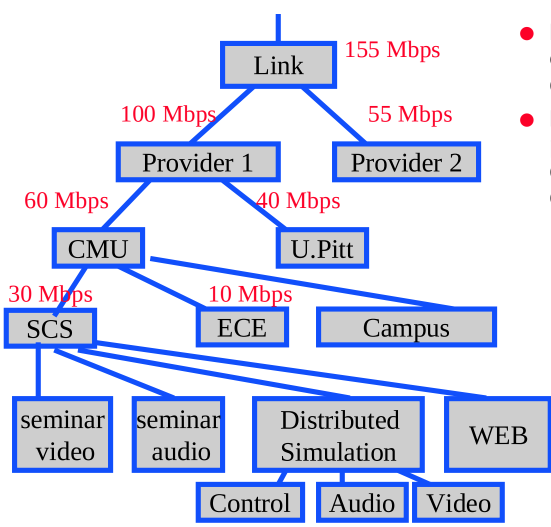
Carnegie Mellon

QoS at Different Traffic Granularities

- Service Providers
  - MCI vs. UUNET
- Organizations
  - CMU vs. University of Pittsburgh
  - SCS vs. ECE
- Applications
  - distributed simulation vs video conferencing
  - FTP vs. WEB vs. Telnet

Carnegie Mellon

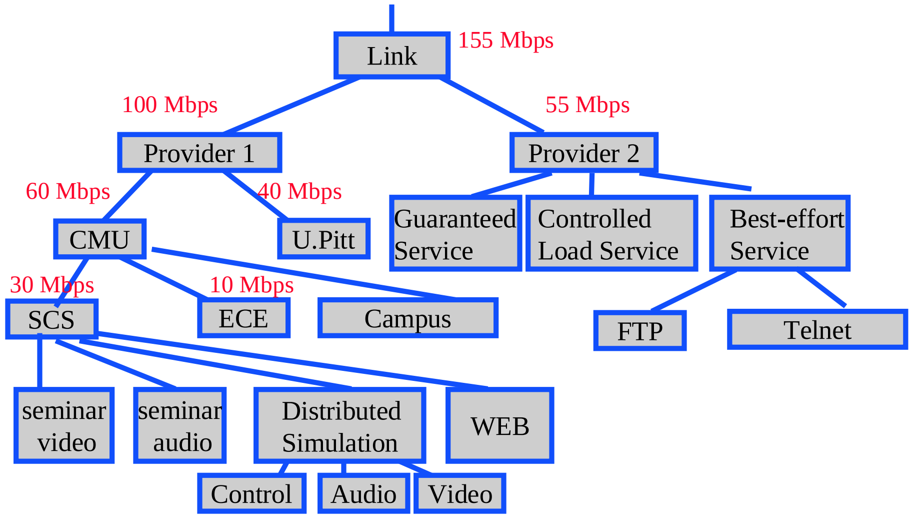
Hierarchical Resource Sharing



- Resource contention/sharing at different levels
- Resource management policies should be set at different levels, by different entities
  - resource owner
  - service providers
  - organizations
  - applications

Carnegie Mellon

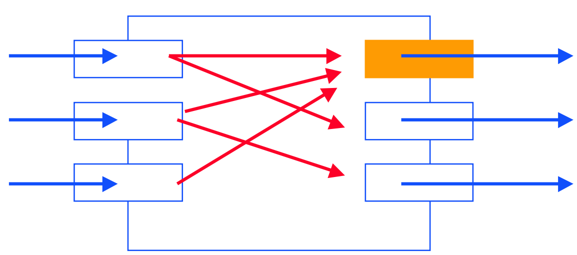
Hierarchical Resource Sharing



1998 Hui Zhang

Carnegie Mellon

Packet Scheduling Algorithm

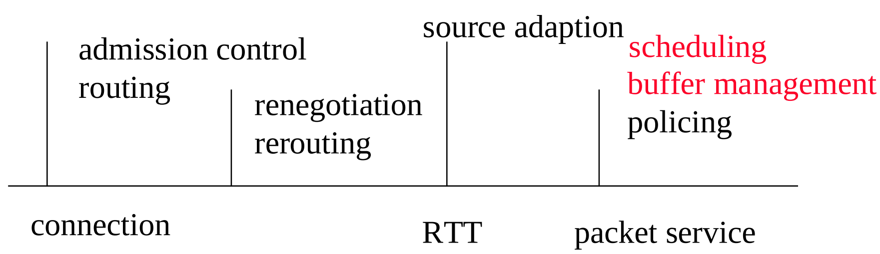


- Operate at the multiplexing point
- Control the interactions among
  - traffic streams in the same class
  - different performance-oriented traffic classes
  - different administrative traffic classes

1998 Hui Zhang

Carnegie Mellon

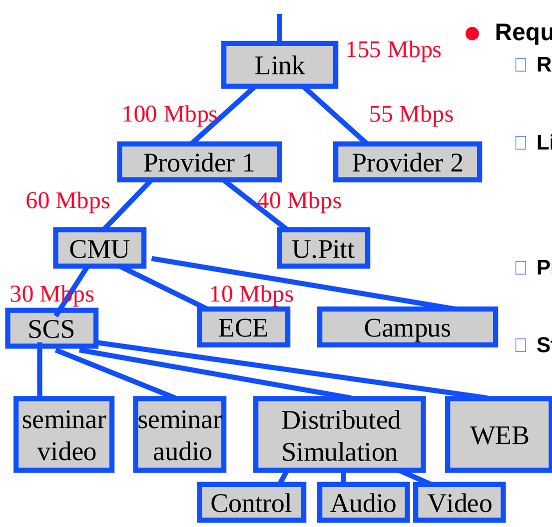
Traffic Control Mechanisms and Timescales



- **Scheduling**
  - **operates at the smallest timescale**
  - **at the multiplexing point**

Carnegie Mellon

Hierarchical Resource Scheduling



- **Requirements:**
  - **Real-time**
    - **QoS for *leaf* classes (i.e., delay, bandwidth)**
  - **Link-Sharing**
    - **service sharing among classes**
    - **"proper" distribution of excess service among sibling classes (e.g., fair)**
  - **Priority**
    - **decoupled delay/bandwidth allocation**
  - **Statistical sharing**
    - **encourage adaptation by end-system**

Carnegie Mellon

Previous Work

- Class-Based Queueing by Van Jacobson, Sally Floyd
  - ad hoc link-sharing semantics
  - no firm guarantees
- CSZ Algorithm by Clark, Shenker, Zhang
  - firm guarantee comparable to that provided by GPS
  - no fine grain link-sharing
- Neither designed to be scalable with number of classes
- Fundamental Problem
  - more difficult to provide QoS in a hierarchy

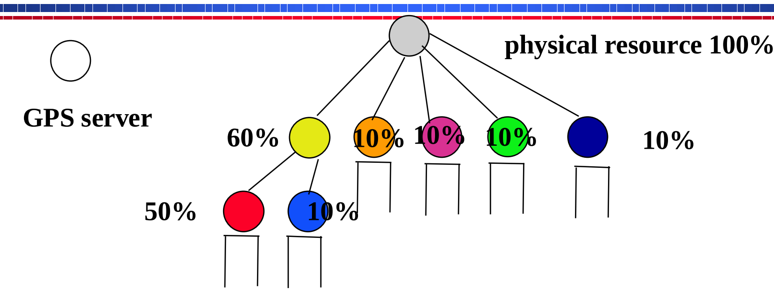
Carnegie Mellon

Our Approach

- Start with an idealized model that can capture all the requirements
- Develop practical algorithms to approximate the ideal model
- Two models and algorithms:
  - Hierarchical Generalized Processor Sharing model and Hierarchical Packet Fair Queueing algorithm
  - Hierarchical Fair Service Curve model and algorithm

Carnegie Mellon

Hierarchical Generalized Processor Sharing

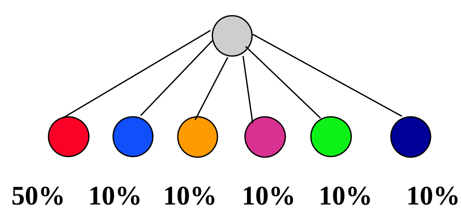


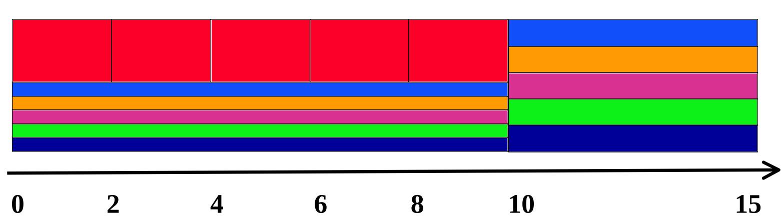
- Idealized fluid algorithm
  - service multiple queues simultaneously, in proportional to shares
  - parent distributes service instantaneously to children

Carnegie Mellon

GPS Example

- [Parekh 92]
- Red session has packets backlogged between time 0 and 10
- Other sessions have packets continuously backlogged

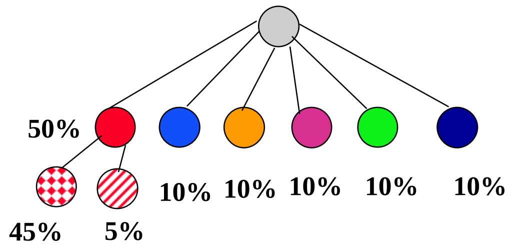


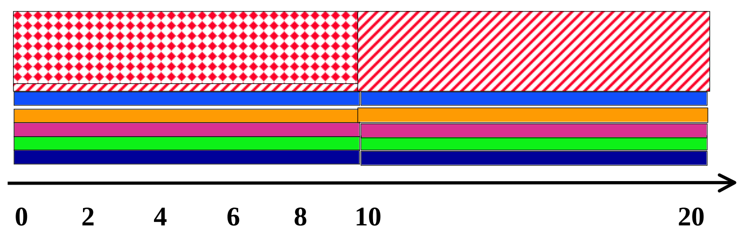


Carnegie Mellon

H-GPS Example

- session has packets backlogged between time 0 and 10
- Other sessions have packets continuously backlogged





Carnegie Mellon

GPS and H-GPS Properties

- Proven end-to-end delay bounds for guaranteed traffic
- Adaptation friendly
  - no punishment in the future for using excess service
  - scalable technique for estimating excess bandwidth for adaptive applications
- GPS: guaranteed service for each flow
- H-GPS: guaranteed service for all classes
  - leaf class: individual flow
  - interior class: traffic aggregate
  - QoS met for all traffic classes simultaneously
  - ideal semantics for hierarchical resource management

Carnegie Mellon

Packet vs. Fluid System

- GPS and H-GPS are fluid systems
  - multiple queues can be serviced simultaneously
  - no non-preemption unit
- Real system are packet systems
  - one queue is served at any given time
  - packet transmission is not preempted
- Design issue: packet algorithms approximating the fluid H-GPS algorithm
  - accuracy: delay bound, fairness, bandwidth distribution
  - complexity: practical implementation

Carnegie Mellon

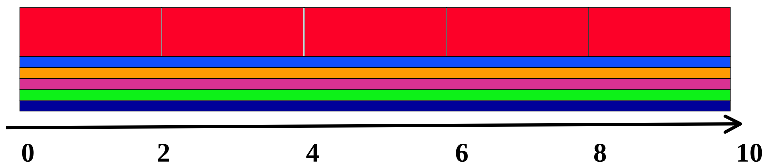
Packet Approximation of Fluid System

- Standard techniques of approximating fluid GPS
  - select packet that will finish first in GPS assuming that there are no future arrivals
- Important properties of GPS
  - finishing order of packets currently in system independent of future arrivals
- Implementation based on virtual time
  - assign virtual finish time to each packet upon arrival
  - packets served in increasing order of virtual times
- Various Packet Fair Queueing (PFQ) algorithms
  - Weighted Fair Queueing (WFQ), Self-Clocked Fair Queueing (SCFQ), Start-time Fair Queueing (SFQ)

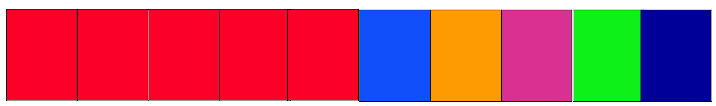
Carnegie Mellon

Approximating GPS with WFQ

- Fluid GPS system service order



- Weighted Fair Queueing
  - select the first packet that finishes in GPS



- Observation: packet finishes in WFQ no late than in GPS
  - basis for proving delay bound for WFQ
- Another observation: some packet finishes service much earlier in WFQ in GPS

Carnegie Mellon

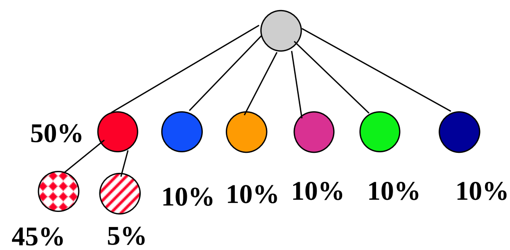
Packet Approximation of H-GPS

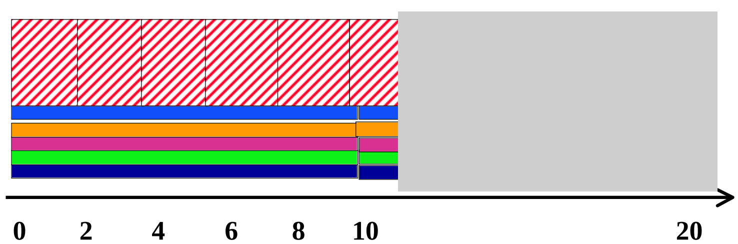
- Idea 1: use similar approach as approximating GPS
  - select packet finishing first in H-GPS, assuming there are no future arrivals
- Problem
  - relative finish order of packets dependent on future arrivals
  - can not use virtual time implementation

Carnegie Mellon

H-GPS Example

- has no packets backlogged between time 0 and 11
- Other sessions have packets continuously backlogged
- Need to make a decision at time 11

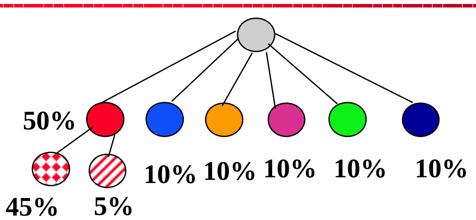




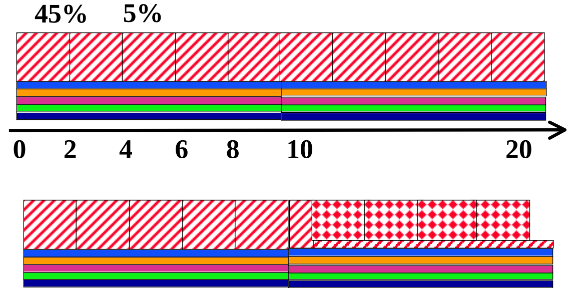
1998 Hui Zhang

Carnegie Mellon

Approximating H-GPS: Decision at Time 11



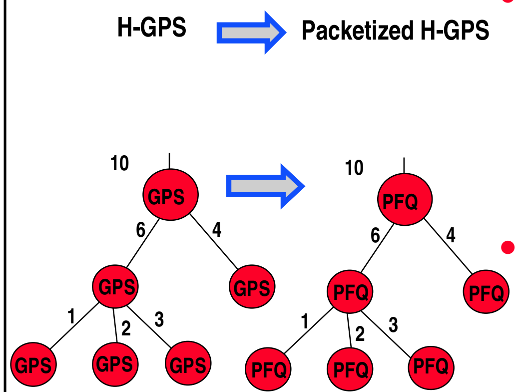
- Case 1: no future arrivals



- Case 2: active at time 11
- Relative finish order dependent on future arrivals

1998 Hui Zhang

Packet Approximation of H-GPS



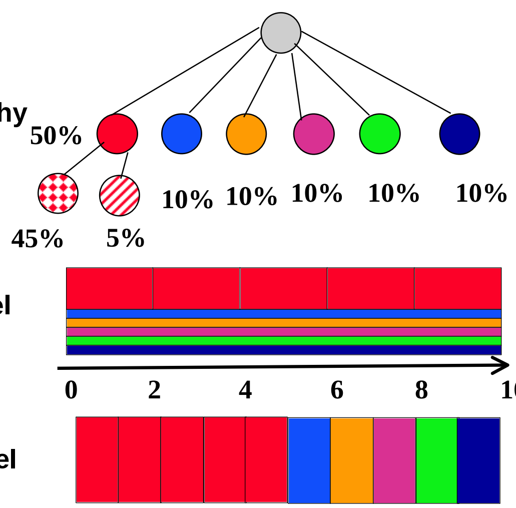
- Idea 1
  - select packet finishing first in H-GPS assuming there are no future arrivals
  - problem:
    - finish order in system dependent on future arrivals
    - virtual time implementation won't work
- Idea 2
  - use a hierarchy of PFQ to approximate H-GPS

1998 Hui Zhang

23

H-GPS and H-WFQ Example

- Two levels of hierarchy
- GPS order at top level
- WFQ order at top level



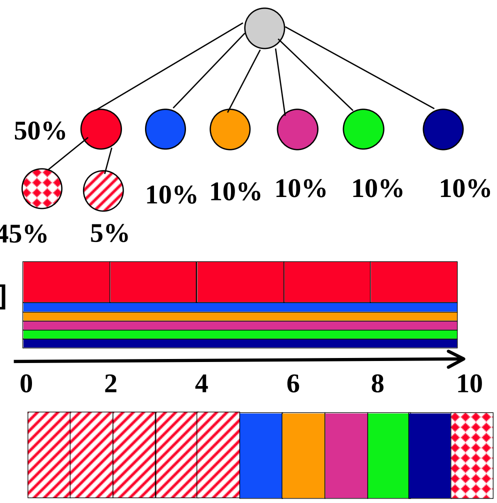
1998 Hui Zhang

24

Carnegie Mellon

H-GPS and H-WFQ Example

- Arrival process
  - active at time 5
  - all other flows continuously backlogged
- Top level GPS server
- Top level WFQ server at [0,4]
  - make decision assuming there are no future arrivals
  - select packets in class
  - only class is active
- Class active at time 5
  - has to endure a long delay



Carnegie Mellon

PFQ and H-PFQ

- Many PFQ algorithms
  - SCFQ, SFQ, FBFQ, WFQ
  - how do properties of PFQ relate to H-PFQ?
- Worst-case Fair Index
  - measuring the accuracy of PFQ in approximating GPS

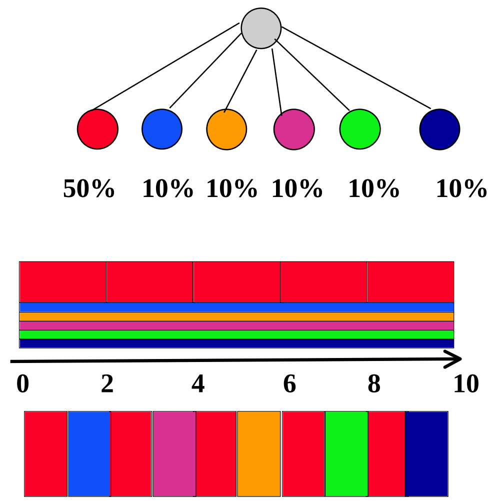
$$D_H = D + \sum_{h=1}^{H-1} \frac{\alpha^h}{r^h}$$

  - tight H-PFQ delay bound $\Longleftrightarrow$ small WFI of PFQ
- All previously proposed PFQ algorithms have large WFI's

Carnegie Mellon

A More Accurate Approximation of GPS

- Problem with WFQ
  - WFQ can be ahead of GPS too much
- Worst-case Fair Weighted Fair Queueing (WF²Q)
  - a packet is eligible if it starts service in GPS
  - among all eligible packets, select the one that finishes first in GPS
  - optimal algorithm in approximating GPS
  - normalized WFI is one maximal size packet
- WF²Q service order



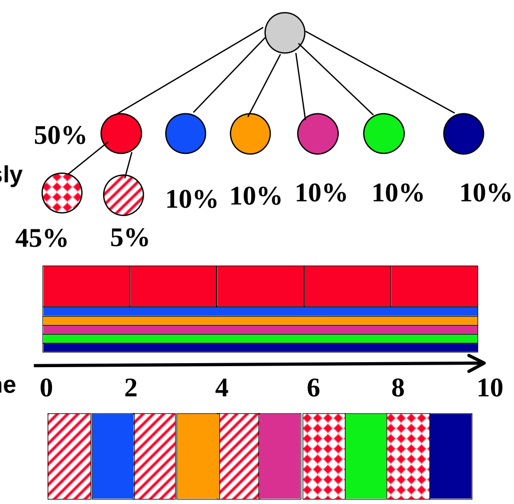
1998 Hui Zhang

27

Carnegie Mellon

H-GPS and H-WF²Q Example

- Arrival process
  - active at time 5
  - all other flows continuously backlogged
- Top level GPS server
- Top level WFQ server
  - there are no future arrival
  - never ahead of GPS by one packet size
  - class receives service immediately after becoming active



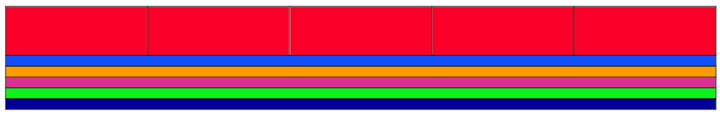
1998 Hui Zhang

28

Carnegie Mellon

WF^{2}Q+

- WFQ and WF^{2}Q
  - need to emulate fluid GPS system
  - high complexity

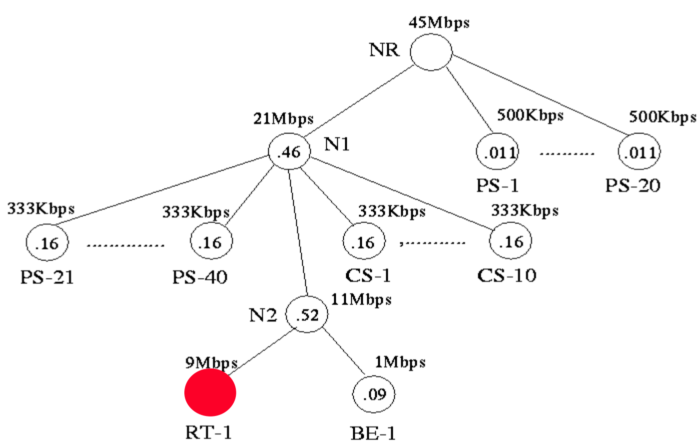


- WF^{2}Q+
  - provide same delay bound and WFI as WF^{2}Q
  - lower complexity

1998 Hui Zhang

Carnegie Mellon

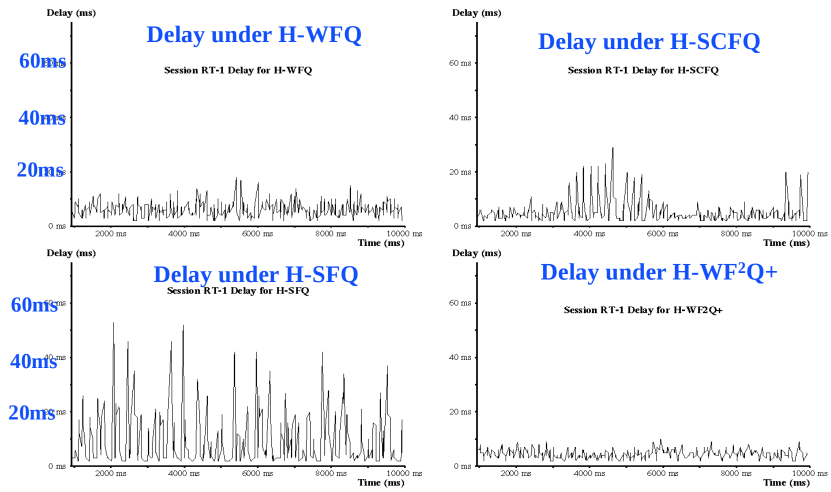
Example Hierarchy



1998 Hui Zhang

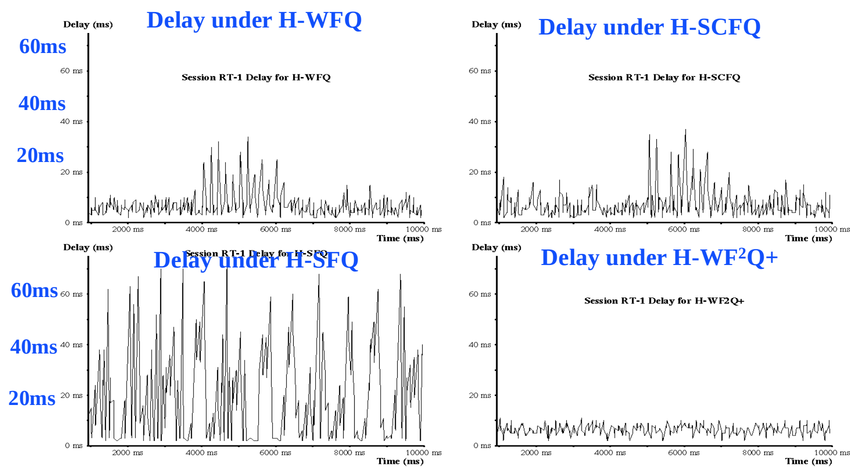
Carnegie Mellon

Uncorrelated Cross Traffic



Carnegie Mellon

Correlated Cross Traffic



Carnegie Mellon

H-GPS and H-PFQ

- Hierarchical Generalized Processor Sharing model
  - proven end-to-end delay bounds for guaranteed traffic
  - ideal semantics for hierarchical resource management
    - QoS met for all traffic classes simultaneously
  - friendly to bursty and adaptation sources
    - scalable mechanism for estimating available bandwidth by adaptive applications
    - no punishment in the future for using excess service in the past
- Hierarchical Packet Fair Queueing Algorithm
  - accurately and efficiently approximates H-GPS model
  - first algorithm to support real-time, link-sharing simultaneously

Carnegie Mellon

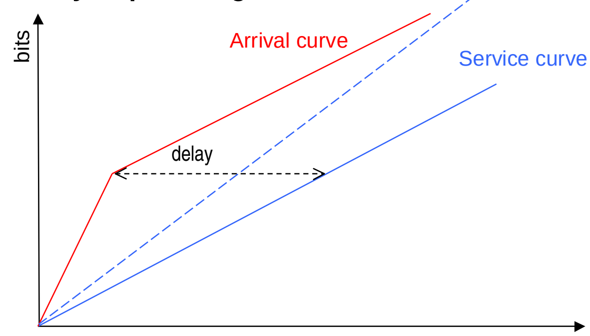
Limitation of GPS and H-GPS Model

- Service specified by one parameter: rate
- Delay bound is a function of the rate: B/R
  - B: maximum burst size
  - R: guaranteed service rate
- Coupling between delay and bandwidth allocation
  - low delay requires higher rate
  - inefficient resource utilization for low delay/low bandwidth sessions

Carnegie Mellon

Service Curve QoS Model

- Each class is assigned a service curve that specifies the minimum amount of service it should receive
- GPS guarantees a linear service curve
- Low delay requires higher bandwidth reservation



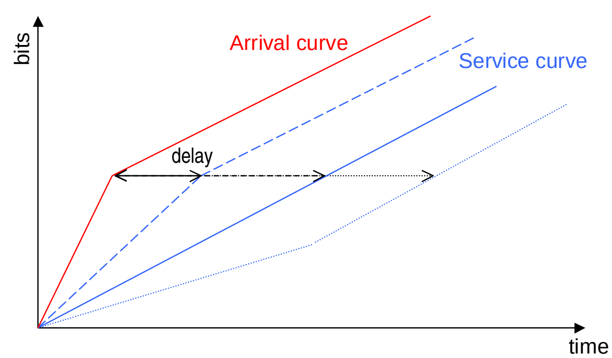
1998 Hui Zhang

35

Carnegie Mellon

Service Curve QoS Model

- Each class is assigned a service curve that specifies the minimum amount of service it should receive
- Non-linear service curve decouples the delay/bandwidth allocation



1998 Hui Zhang

36

Carnegie Mellon

Fairness

- Defined in context of Generalized Processor Sharing (GPS)
  - each session is assigned a share
  - service is instantaneously distributed in proportion to their shares
  - excess service unused by a session also distributed proportion to sessions' shares
- Important property: memoryless, non-punishment
  - a session not punished for extra service it received in the past
  - important for hierarchical link-sharing
  - encourages adaptation and statistical multiplexing

1998 Hui Zhang

37

Carnegie Mellon

Hierarchical Fair Service Curve Model

- Each class is associated a service curve
- Service curves of all classes are guaranteed
- Excess service is distributed fairly among sibling classes
- Properties
  - formally defines link-sharing, real-time, priority requirements
  - generalized fairness property
    - applicable to arbitrary-shape service curves
    - decouple excess service distribution policy from guaranteed service

1998 Hui Zhang

38

Carnegie Mellon

Fundamental Conflicts

- With non-linear service curves or priority service
  - it is impossible to guarantee service curves of all classes
  - It is impossible to simultaneously provide real-time guarantees and instantaneous fairness

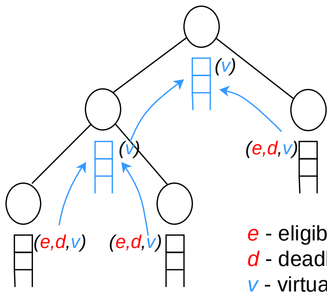
Carnegie Mellon

H-FSC Scheduling Algorithm

- Approximates H-FSC model by
  - resolving conflicts in favor of real-time requirements for leaf classes
  - minimizing discrepancy between service each class should ideally receive and the service it actually receives
- Based on two criteria
  - real-time: guarantee service curves of all leaf classes
  - link-sharing: guarantee service curves of interior classes and distribute excess service fairly
  - whenever there is a potential of conflict, real-time criteria is used

Carnegie Mellon

H-FSC Algorithm



e - eligible time
d - deadline
v - virtual time

The Algorithm

if *(there is an eligible packet)*
/* *real-time criteria* */
send *eligible packet with min. deadline d ;*
else
/* *link-sharing criteria* */
send *packet with min. virtual time v ;*

- **Eligible time (*e*): whenever there are eligible packets there are potential conflict between real-time and link-sharing goals**
- **Deadline (*d*): if all deadlines of packets of a class are met, its service curve guarantee is met**
- **Virtual time (*v*): normalized amount of service received by a class**
- ***e, d, v* are computed based on three curves E( ), D( ), V( )**

Carnegie Mellon

H-FSC Complexity

- **Theoretically, worst case complexity O(log *n*), but can be improved taking advantage of timestamp properties**
- **All service curves are two-piece linear**
  - **enough to decouple bandwidth & delay allocation**
- **Computing *e, d,* and *v* requires one comparison, division**
- **E( ), D( ), V( ) are updated *only* when session becomes active**
  - **updating D( ) and V( ) requires in the worst case three multiplications, one division and several additions, subtractions**
  - **updating *E( )* requires at most one addition**

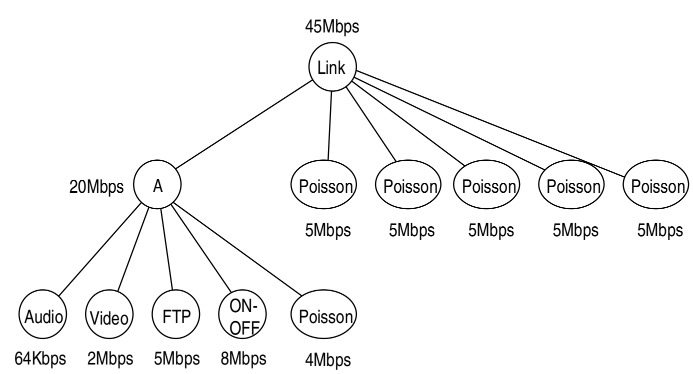
Carnegie Mellon

Delay and Fairness Properties of H-FSC

- Service curves of all leaf classes are guaranteed
- For any class, the difference between the service it should receive in the ideal H-FSC model, and the service it actually receives is bounded

Carnegie Mellon

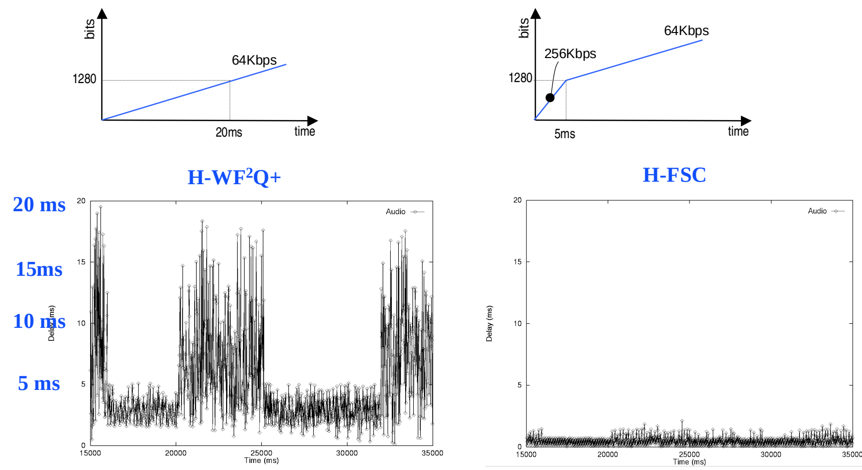
Simulations: Class Hierarchy



Carnegie Mellon

Simulations: Delay Results for Audio (H-WF^{2}Q+ vs. H-FSC)

Packet size = 1280 bits, bandwidth = 64Kbps



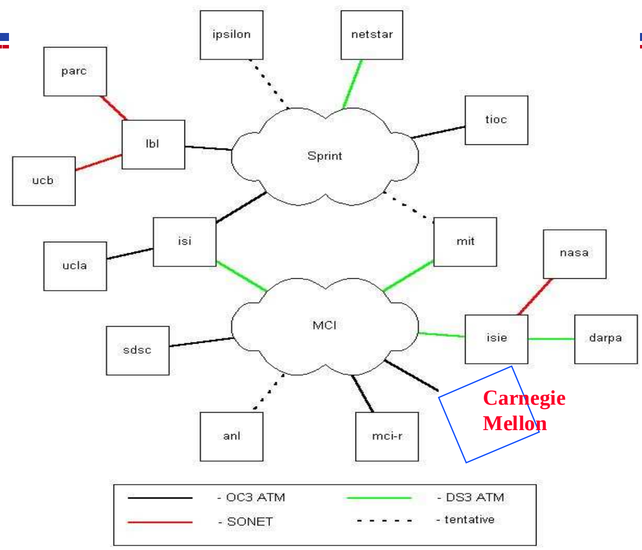
Carnegie Mellon

Implementation Status

- 622 Mbps ATM Implementation by Fore Systems
- Software IP Implementation on NetBSD/PC router
  - flat hierarchy, 1000 flows, 9 usec/packet
  - 3 levels hierarchy, 1000 flows, 12 usec/packet
  - generic interface for control software: Beagle, RSVP
- Alpha release
  - University of Utal, Washington University, Sun Microsystem
- High speed implementation
  - Ascend GRF router
    - software implementation completed in vBNS (joint work with MCI)
    - hardware implementation in process (with Ascend)
  - Cisco IOS: summer 1998

Carnegie Mellon

CAIRN High Bandwidth Physical Links
01/08/97

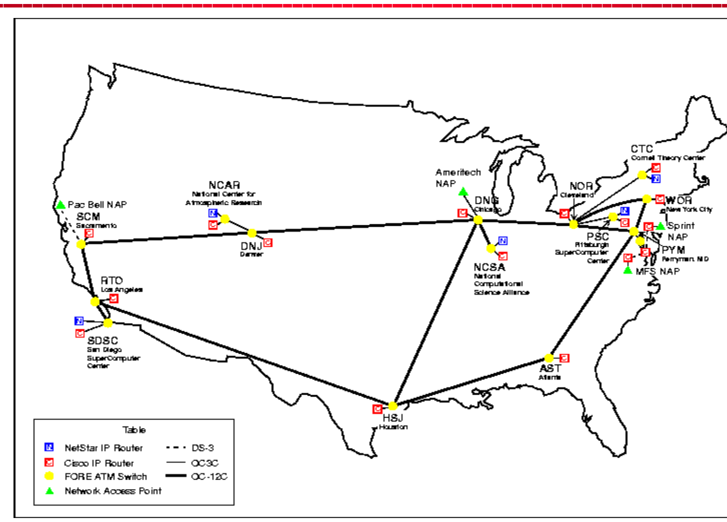


Carnegie Mellon

1998 Hui Zhang

Carnegie Mellon

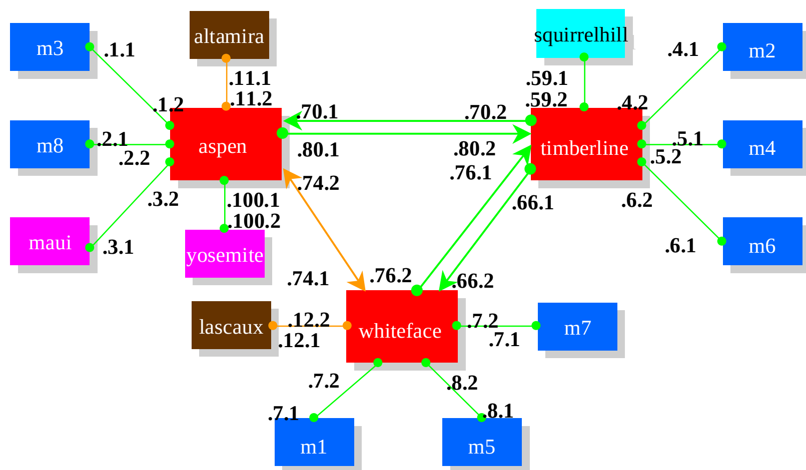
NSF vBNS Network



1998 Hui Zhang

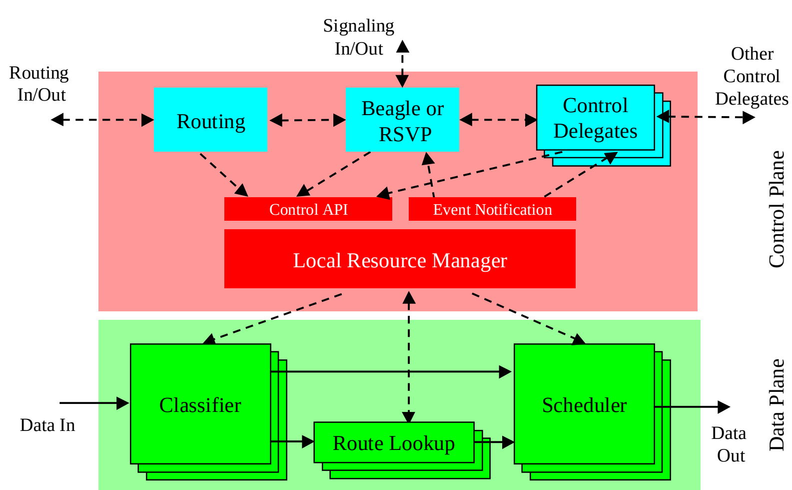
Carnegie Mellon

Local Darwin Testbed



Carnegie Mellon

Router Node Architecture



Carnegie Mellon

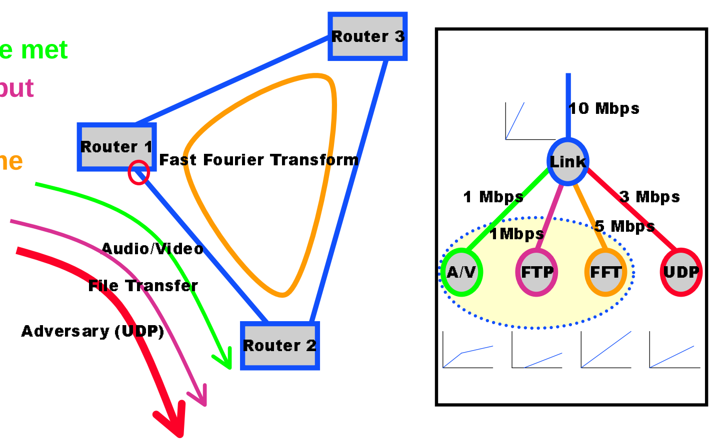
Simple Virtual Network Experiment

- Experiment set-up
  - application virtual mesh with
    - FFT distributed across three nodes
    - Audio/video transmission
    - FTP application
  - background UDP traffic
- Base-line senario
  - reservation for individual flows in application mesh
  - no virtual mesh specific resource management policy
- Test case senario
  - two level of resource management hierarchy
  - virtual mesh specific resource management policy

Carnegie Mellon

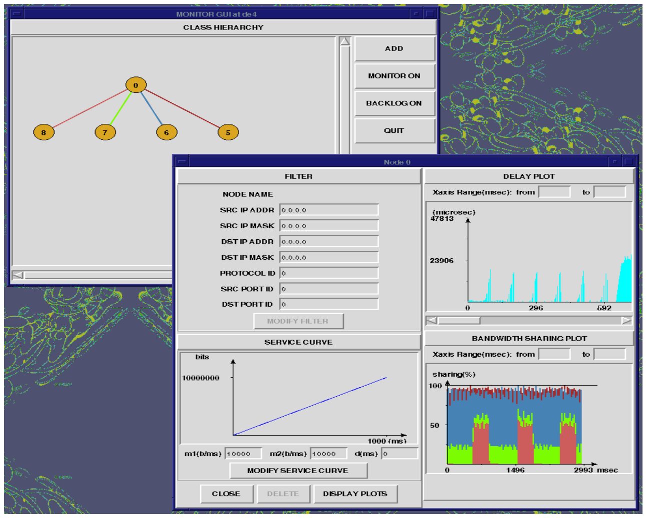
Base-line Case: No Hierarchy

- A/V guarantee met
- FTP throughput
  - 1.7 Mbps
- FFT finish time
  - 29 second



Carnegie Mellon

Link-sharing at Root Node



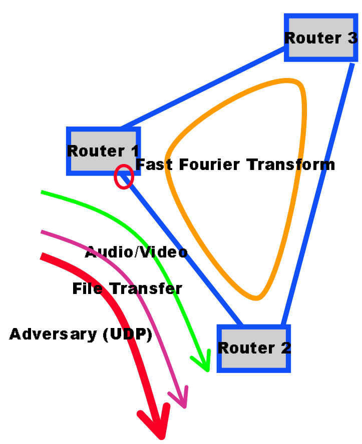
1998 Hui Zhang

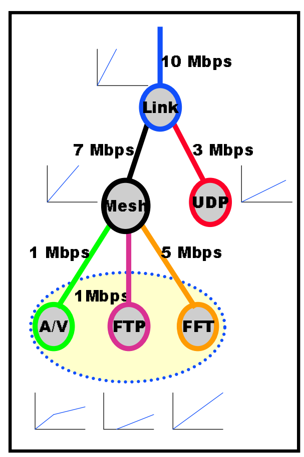
53

Carnegie Mellon

Test Case: Two Levels of Resource Hierarchy

- A/V guarantee met
- FTP throughput
  - 4 Mbps
  - factor of 2.4 improvement
- FFT finish time
  - 27 second
- Same amount of total resource reserved



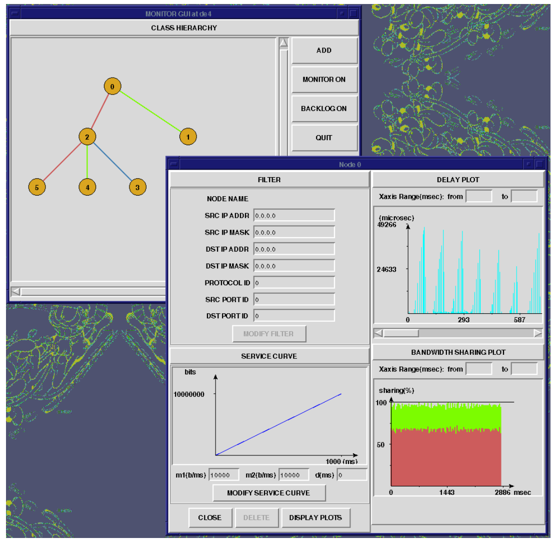


1998 Hui Zhang

54

Carnegie
Mellon

Link-sharing at Root Node

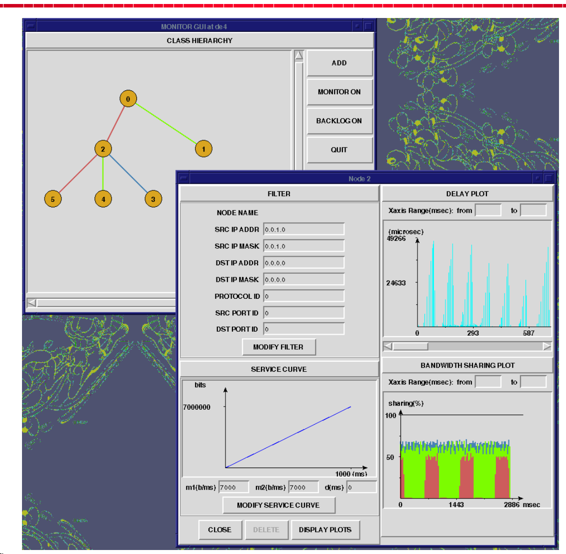


1998 Hui Zhang

55

Carnegie
Mellon

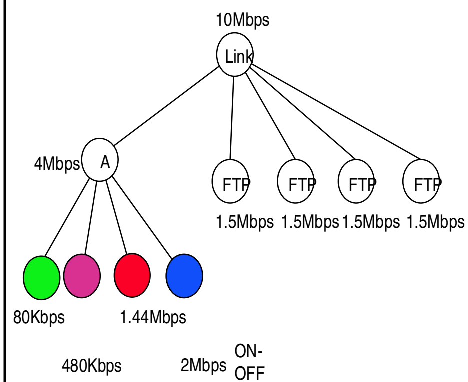
Link-sharing at Interior Application Node

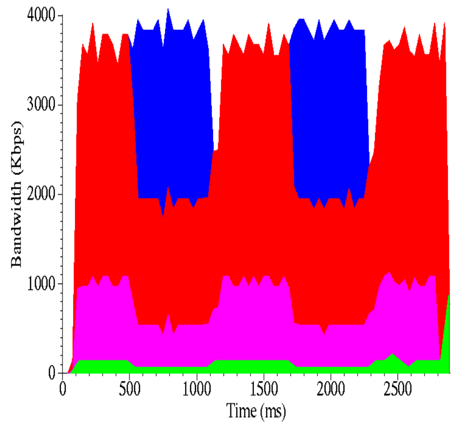


1998 Hui Zhang

56

Experiment: Link-Sharing





1998 Hui Zhang

57

Ongoing Work

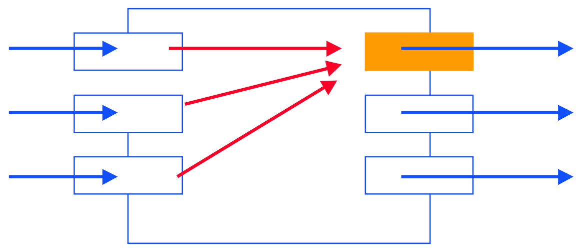
- High performance Combined-Input-Output buffered switches or routers
- Implementation with no per flow state
- Wireless networks with location-dependent errors

1998 Hui Zhang

58

Carnegie Mellon

High Speed Implementation



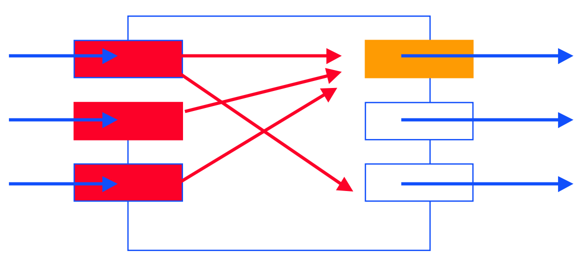
- All existing QoS scheduling/queue management algorithms assume output buffering
- Output-buffered switches/routers have fundamental limits on scalability

1998 Hui Zhang

59

Carnegie Mellon

High Speed Implementation (2)

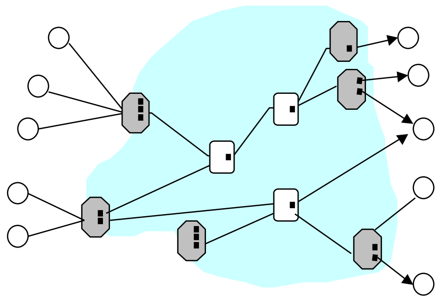


- High performance switches/routers use Input-Output buffered architecture
  - multiple contention points
  - QoS support more difficult

1998 Hui Zhang

60

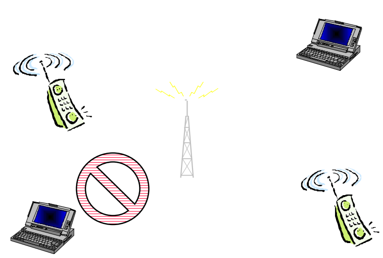
Core-Stateless Implementation



- QoS algorithms require switch to maintain per flow state
  - expensive to install and maintain
  - less robust and fault tolerant
- Core-stateless implementation architecture
  - only edge maintains per flow state
  - core routers do not maintain state

Wireless Networks

- Location dependent errors
- GPS and H-GPS do not apply well
  - backlogged flows in error cannot transmit

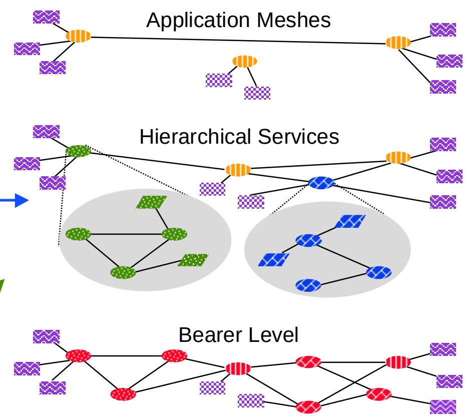


Darwin: Electronic Value-added Services Market

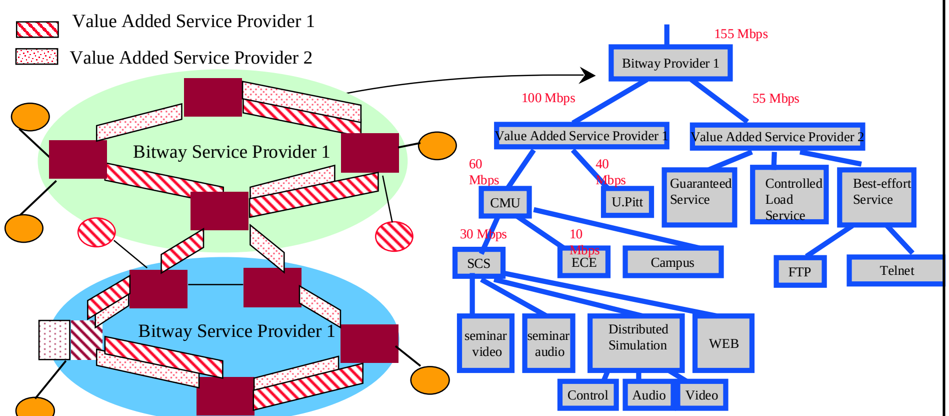
Application combines local resources with network services to create end-product

Value-added services integrate communication with computation and storage services

Electronic services market has a hierarchical structure.



Resource Virtualization



Carnegie Mellon

Summary

- Physical resource can be dynamically partitioned as a set of virtual resources with a hierarchical relationship
- Different entities (organizations, service providers, and applications) can independently specify entity-specific policies for their portions of the virtual resource hierarchy
- Admission control and hierarchical scheduling ensure policies are met for all entities
  - QoS provided for both individual traffic streams and traffic aggregates of different entities
- Collaborative and competitive resource sharing seamlessly integrated
- Abstraction applicable to other resources (CPU) also

Carnegie Mellon

References

- J.C.R. Bennet and H. Zhang, "Hierarchical Packet Fair Queueing Algorithms", IEEE/ACM Transactions on Networking, 5(5): 675-689, Oct 1997. Also in SIGCOMM'96.
- I. Stoica, H. Zhang, E. Ng, "A Hierarchical Fair Service Curve Algorithm for Link-sharing, Real-Time and Priority Services", SIGCOMM'97.
- D. Stephens, H. Zhang, "Implementing Distributed Packet Fair Queueing in a Scalable Switch Architecture", INFOCOM'98.
- I. Stoica, S. Shenker, H. Zhang, "Core-Stateless Fair Queueing: A Scalable Architecture to Approximate Fair Bandwidth Allocation in High Speed Networks", SIGCOMM'98.
- URL: http://www.cs.cmu.edu/~hzhang/publications.html