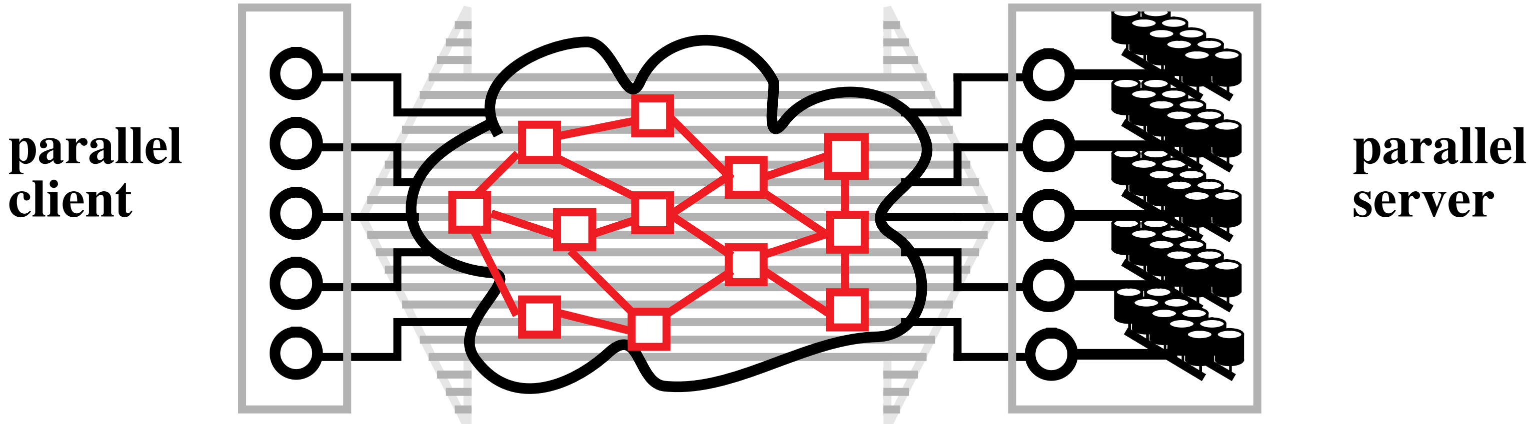
Meeting Agenda

I. Reporting progress

II. Feedback, critics, and advice

Parallel Data Movement



I/O intensive applications

- Scientific and multimedia applications
- Required bandwidth >> link bandwidth

Switched networking technology

- ATM, HiPPI, Switched-Ethernet, Switched-FDDI
- High connectivity => High aggregate bandwidth

Dynamic multicomputers and multiple link supercomputer

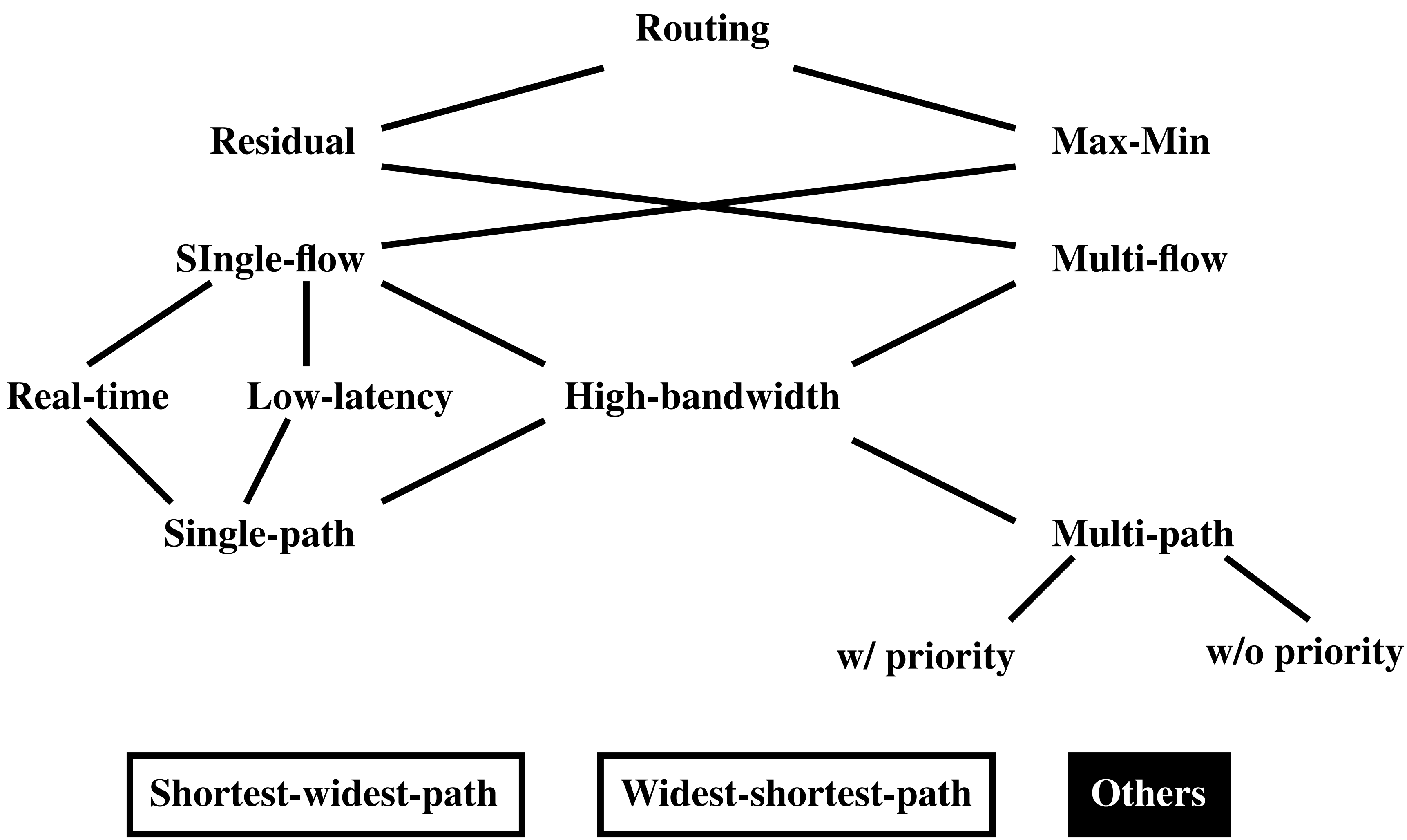
- Appl. can select byte-link mapping besed on bandwidth

Parallel Flows --- A Parallel Delivery Service

Parallel Flows: A logical data stream is split and transmitted in parallel from multiple sources to multiple destinations

- **Coordinated high-bandwidth routing**
  - **- select path based available bandwidth (single-flow)**
    - Routing based on per packet delay is not suitable for HB
  - **- allocate to group of requests at a time (multi-flow)**
    - Independent routing may cause conflicts
- **Service interface**
  - **- a network abstraction of parallel flows**
  - **- application-network interactions for bandwidth planning, change modification, and bandwidth replanning**

Working Chart (Coordinated Routing)

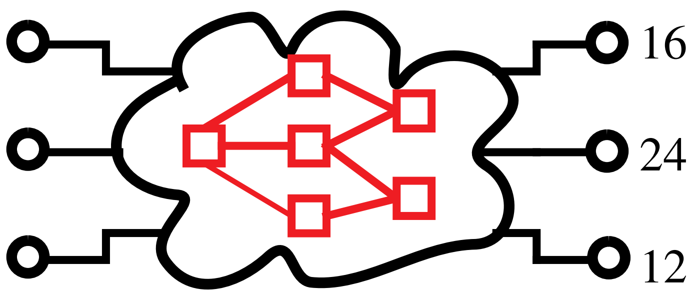


Coordinated Routing: Residual Bandwidth Model

- **Shortest-widest path**
- **Multi-path per flow**
  - **More bandwidth for applications**
  - **NP-hard problem ==> polynomial LP (linear-programming) problem**
- **Routing Algorithms**
  - **LP is not fast enough for on-line routing**
  - **Tom Leighton's fast approximation algorithm**
    - do min-max for each individual flow independently
    - put them together and reroute bottlenecks
  - **Weighted Round-Robin**
- **Evaluating Algorithms**
  - **Compare with LP and do simulation**

Weighted Round-Robin

- Treat the whole network as a single resource server
- Let every flow have a weighted equal chance to be served
- Each flow routes a certain fraction of its bandwidth
- Exponential cutoff to complete in Log N rounds



Round

1	*2*	*3*	*4*
4	4	4	4
6	6	6	6
3	3	3	3

Caught by ABR Flow Control and Max-Min Fairness

- Peter's talk on ATM in SDI Seminar
- ATM Forum ABR rate-based flow control adopted
- ATM is the state-of-art networking technology
- ABR is targeted for fast data file transfer
- Max-Min fair rate info is more accurate than estimated residual bandwidth
- PNNI supports link-state source routing
- ATM switches are available in both PDL and PSL

Routing ABR traffic: Link-state Information

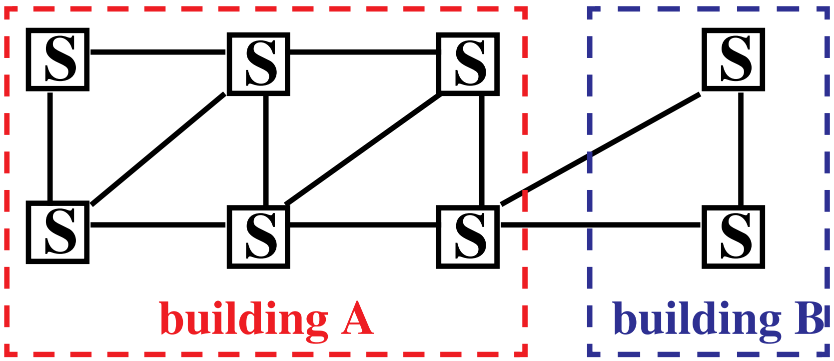
- Neither hosts nor switches have all ABR VCs info.
- A simple one: # of ABR VCs for every output port
  - balance # of VC passing a link
  - simple to calculate
  - problem: if some VCs can not use their fair share (e.g., low-PCR, or have bottlenecks somewhere else)
- Discrete rate vector of # VCs
  - A vector of # of VCs indexed by a set of discrete rates
  - Actual rates are rounded to the closest discrete rates

1	2	1	3	1	1							......
5	10	15	20	25	30							

155 / 10 = 15.5 (155 - 40) / 6 = 19.2

Routing ABR Traffic: Algorithms

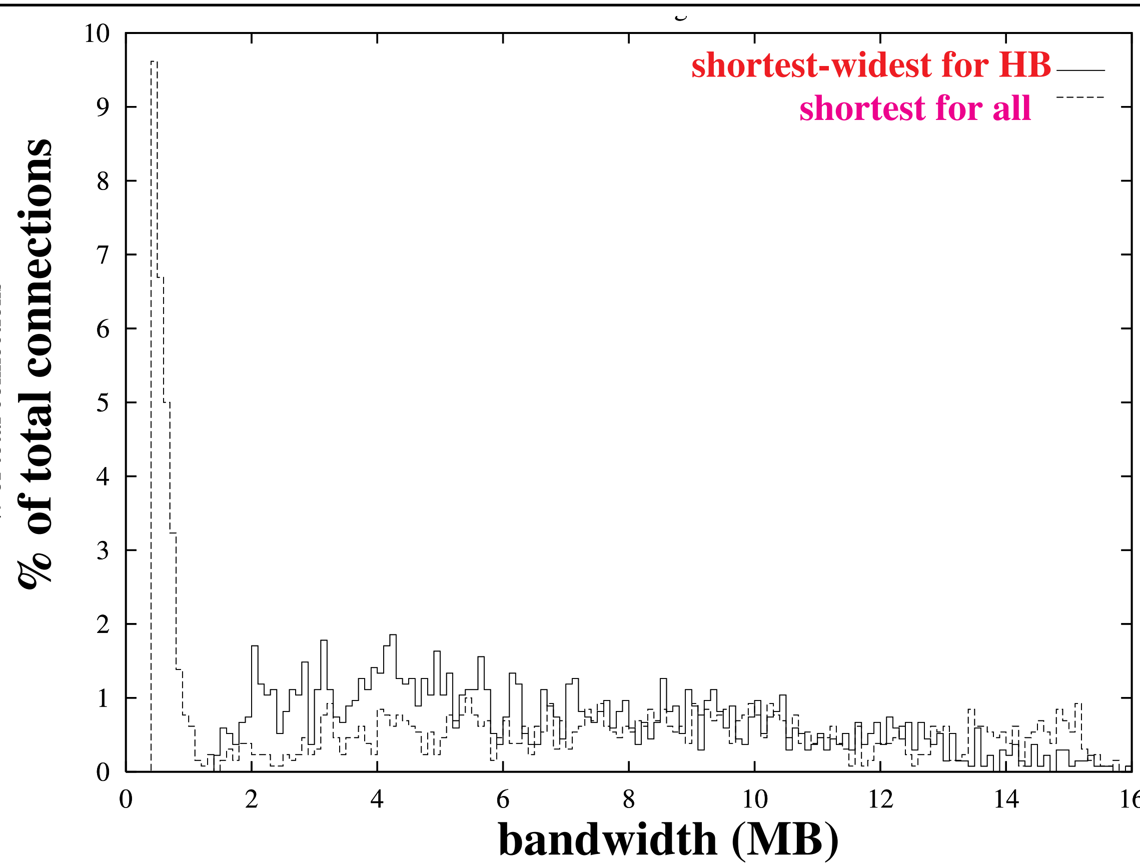
- Widest-shortest path for low latency traffic
- Shortest-widest path for high bandwidth traffic



- redundant paths
- asymmetric

	HB traffic	LL traffic	All traffic
Shortest-widest for HB traffic	4.935 MB/s	3.104 MB/s	3.810 MB/s
Shortest for all traffic	1.485 MB/s	1.132 MB/s	1.281 MB/s

Bandwidth Distribution for High-Bandwidth Traffic



- **Shortest path produces more very fast & very slow paths but much fewer intermediate speed path**

Carnegie Mellon

How to Avoid Large Variance in Achieved bandwidth

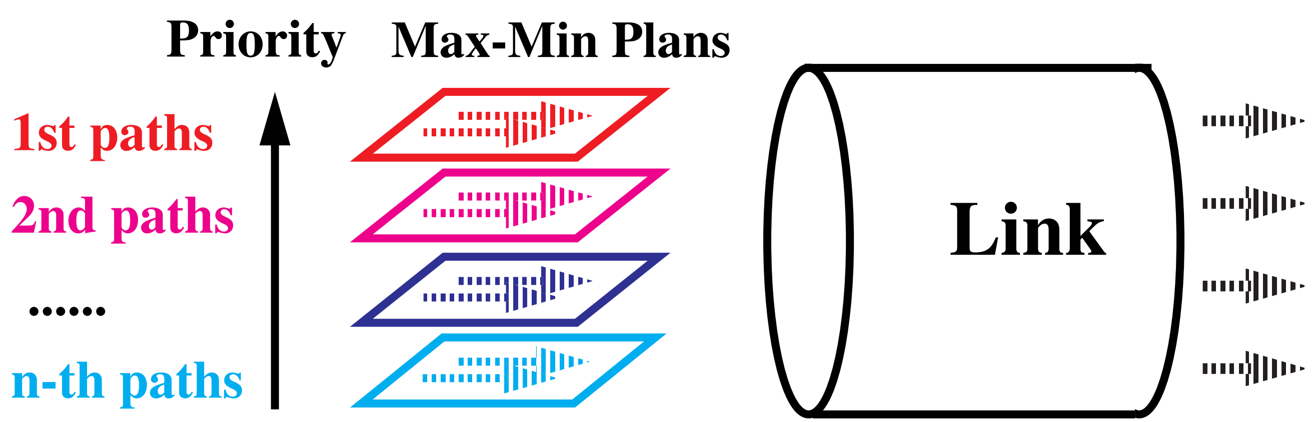
- **Initial routing plan outdated**
  - rerouting
  - using multi-path
  - coordinate a group of high-bandwidth traffic
  - end system replanning
- **Shortest-widest algorithm needs to be revisited**
  - consume more network resource
  - more sensitive to network load change
  - can be a bad choice, if
    - a very long path with just slightly higher bandwidth
  - hybrid of shortest-path and widest-path?

Routing ABR Traffic: Multi-path

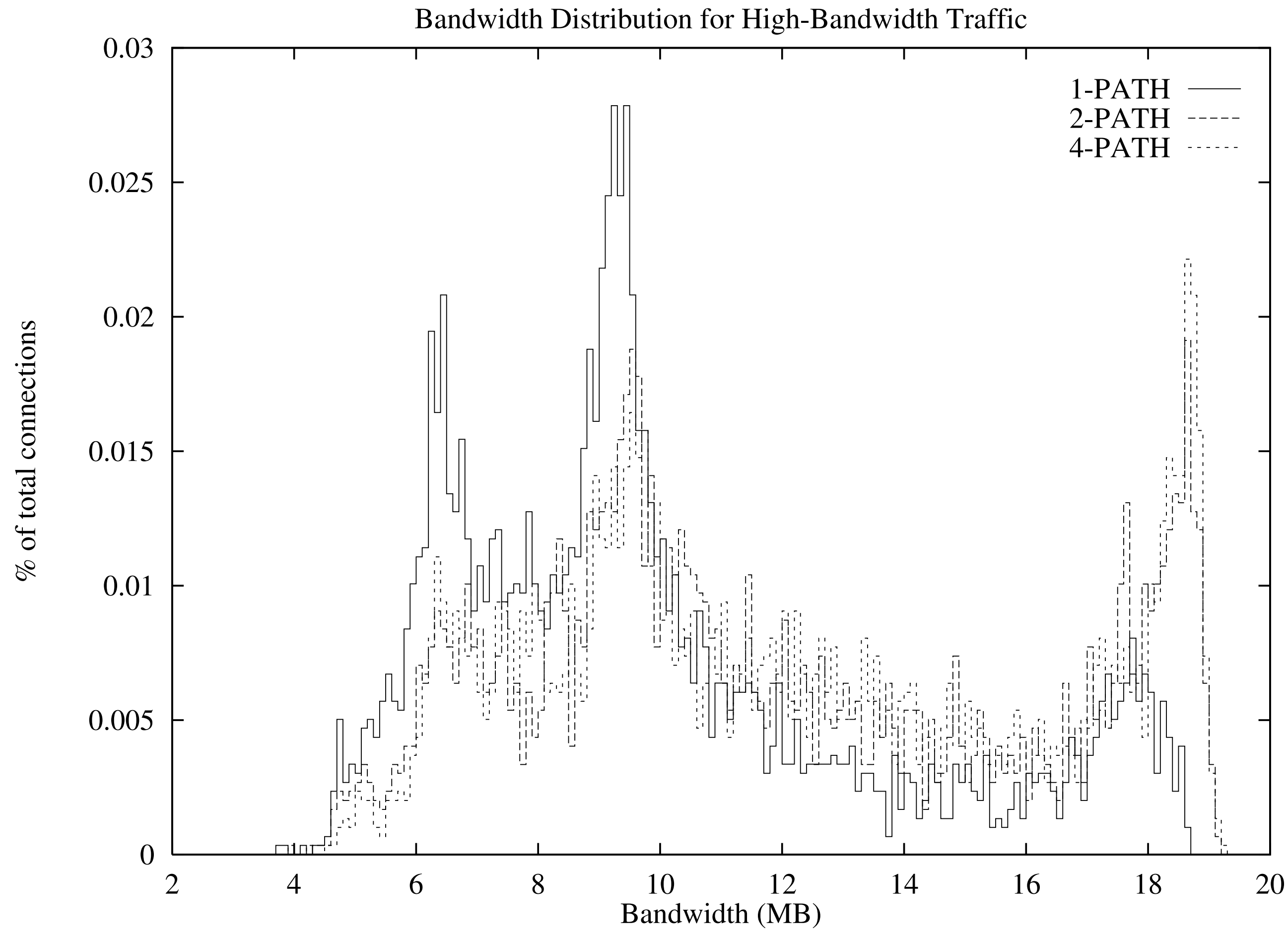
- **Goal: Try to make hungry sources happy**
- **Feasibility: high connectivity in switched networks**
- **Semantics: use unused resource or steal from others**
  - **- multi-path**
    - greedy users can set up many paths
  - **- multi-path + pricing and/or admission control**
    - hard to impose for shared networks
  - **- multi-disjoint-path**
    - A session can obtain multiple shares
  - **- multi-path, but use only one with highest rate at a time**
    - single alternative path, can be achieved by rerouting
  - **- prioritized multi-path**

Prioritized Multi-level Max-Min Fairness

- Multipath is prioritized with one path per priority
- Paths with the same priority obey Max-Min fairness
- Lower priority paths yield to higher priority paths
- Pros and cons
  - + make use of unused bandwidth and help balance load
  - + no interference to other sessions using single-path
  - - may not get much bandwidth if network load is heavy



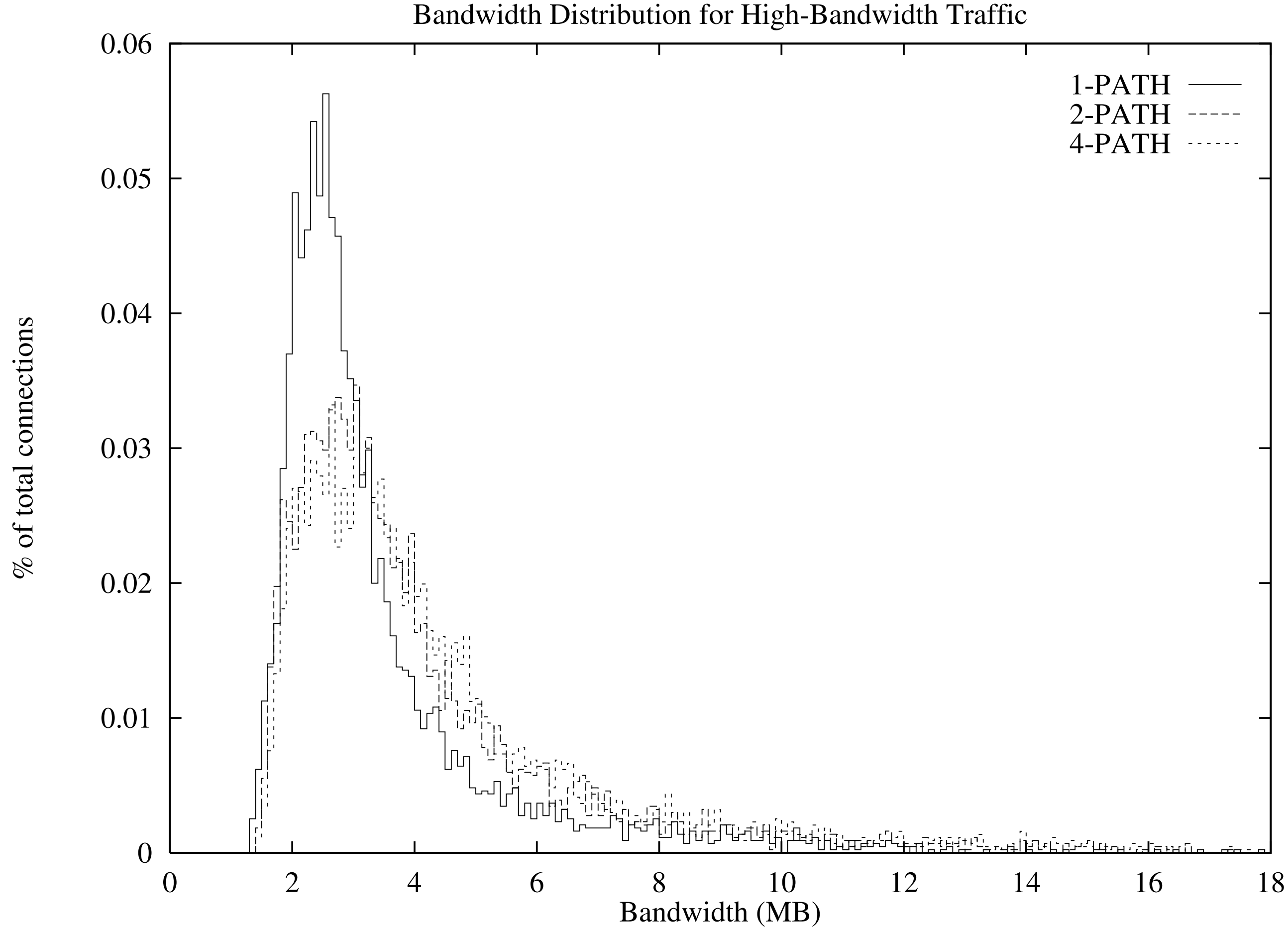
Bandwidth Distribution for HB Traffic --- Loght Load



8.85 MB/s
10.38 MB/s
10.65 MB/s

- **17 % increase from 1 to 2-path, 3% from 2 to 4-path**

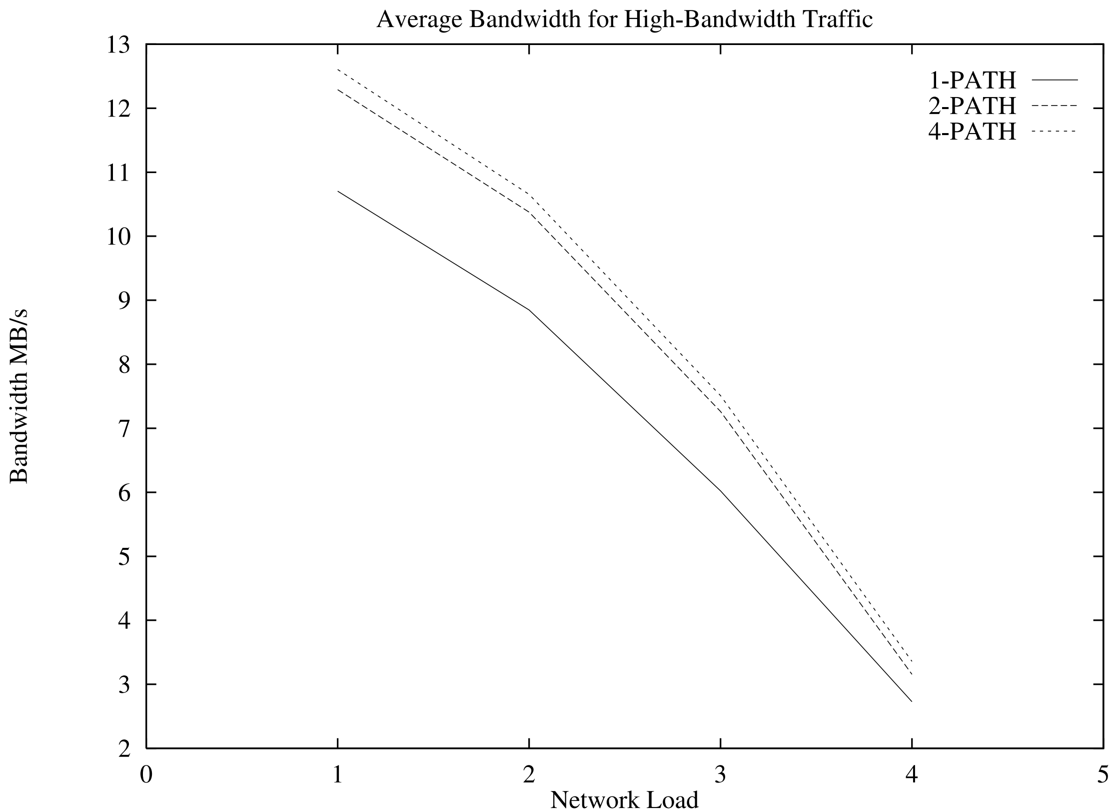
Bandwidth Distribution for HB Traffic --- Heavy Load



2.73 MB/s
3.15 MB/s
3.36 MB/s

- **15 % increase from 1 to 2-path, 6% from 2 to 4-path**

Bandwidth for HB Traffic Over Different Traffic Load



- **Multi-path gain bandwidth for all load conditions**

Added Implementation Overhead

- Source
  - Assign priorities to VCs
- Switch
  - Scheduling discipline: no extra overhead
  - Calculate VC rate for multiple Max-Min plans
  - Advertise link information for all priority classes
- Destination
  - No added overhead
- Routing algorithms
  - Independent of the routing algorithm for single path
  - Repeat the same algorithm on reduced graph

Coordinated Routing: Max-Min Fairness Model

- **Network model**
  - multiple flows get multiple shares if passing through the same link
- **Optimality**
  - minimize total transfer time needed for all flows
- **Integral MCFP (multi-commodity flow prob.)**
  - At each priority level, only one path is used for each flow
- **Very hard problem**
  - Obtainable bandwidth from a link depends on how many flows being routed through the link

Some Known Methods for MCFP

- **Raghavan-Thompson Approach**
  - **- Solving a non-integral MCFP**
  - **- Path striping**
  - **- Randomized path selection**
- **Lang-Rao-Plotkin-Leighton Algorithms**
  - **- Find an initial solution**
    - do SPF one-by-one (Lang & Rao)
    - do Min-max flow independently (Leighton)
  - **- Repeatedly reroute flows passing through bottlenecks**
    - Define distance functions: exponential one used by Leighton and polynomial one by Lang
    - Change the distance function after each iteration, find new bottlenecks and continue

Finding An Initial Solution

- No coordinating: route each flow independently
- Naive coordinating: route one by one
  - increasing order of the # of bytes to be sent
  - decreasing order
  - random order
- Backward-coordinating: route one by one, find paths to achieve minimal transfer time for all flows have been routed so far
  - increasing order
  - decreasing order
  - random order
- We do backward-coordinating for all three orders and pick a best one as an initial solution

Rerouting Flows Passing Through Bottleneck Links

- Find the set of slowest flows, identify the links that are bottlenecks
- Reroute all flows passing these bottlenecks in
  - increasing order
  - decreasing order
  - random order
- Repeat the above steps until
  - either a best solution has been found
  - or the solution can not be improved any further
  - or some other conditions (e.g., time, or distance functions)

Looking Backward

- Data traffic (e.g., file transfer)
  - more fundamental and important, but less attended
  - understand single flow as the first step
- Initial evidences
  - more than one, i.e., best-effort, data service is needed
  - routing is an important vehicle
  - other support may help (multiclass max-min for HB & LL)
- Combine flow control and routing
  - well-known one: routing can help balance network load and reduce congestions
  - existing one: alternative path, fast reservation, residual bandwidth model
  - first one: using rate info and max-min model in routing
  - new one: prioritized max-min fairness, some algorithms

Looking Forward

- Two more months to wrap up routing part, including writing something
- Parallel flow interface design from Mar. 1, 96
- Implement the interface and simple routing on ATM by the end of summer
- Bandwidth planning and replanning afterwards
- Write thesis