Fairness in 10G Ring of Ethernet Switches

Khaled Amer

IEEE 802.17
Plenary Meeting

March 2001

Agenda

- **Objectives**
- **Simulation setup and parameters**
- **Results and analysis of various scenarios**
- **Next steps**

<u>These are preliminary incomplete results of work still in progress</u>

Objectives

- **Investigate the performance characteristics of a ring of Ethernet Switches:**
  - **Enable comparing the results with the performance characteristics of 802.17 RPR solutions**
  - **Quantify areas of strength for 802.17 solutions as compared to Ethernet switches**

Objectives …

- **Focus on fairness in:**
  - **Bandwidth utilization including locality fairness**
  - **ETE delay**

Methodology

- **Follow the methodology that the performance adhoc committee is in the process of defining**
- **Eliminate parameters of specific switches whenever possible:**
  - **Infinite buffers**
  - **Huge switching capacity rate**

Simulation setup

- **Modeling tool used for analysis: Opnet**
- **Node count: Ring of 8 nodes**
- **Ring circumference: 100Km**
- **Ring Rate: 10 Gbps**
- **Packet size: 1250 Bytes**

Simulation setup ...

- **Configurations:**
  - **Hubbing**
  - **Next hop**
- **Low traffic at the beginning to force Spanning Tree Protocol to break the ring at a predictable point**

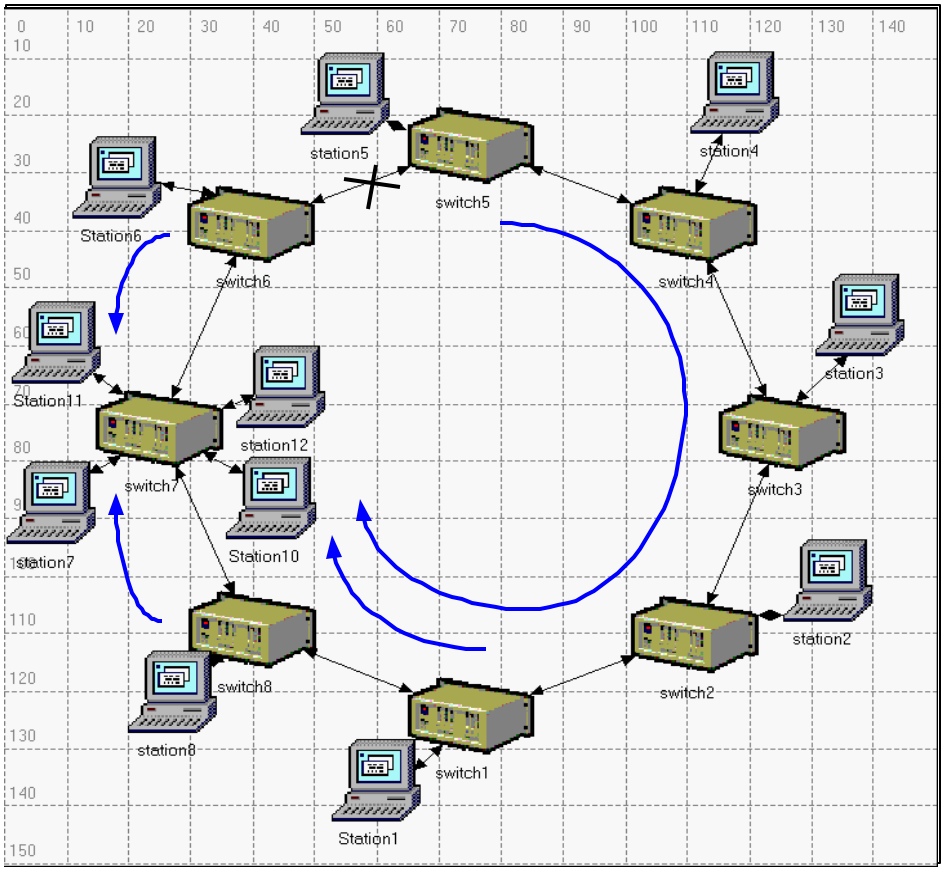
Switch Parameters

- **Generic switch**
  - **10 Gbps ports**
  - **Try to eliminate parameters of specific switches**
- **Store-and-forward**
- **Switch service rate: 10M packets/second**

Metrics

- **Throughput**
  - **In overload conditions**
  - **Per node (for now)**
- **ETE delay**

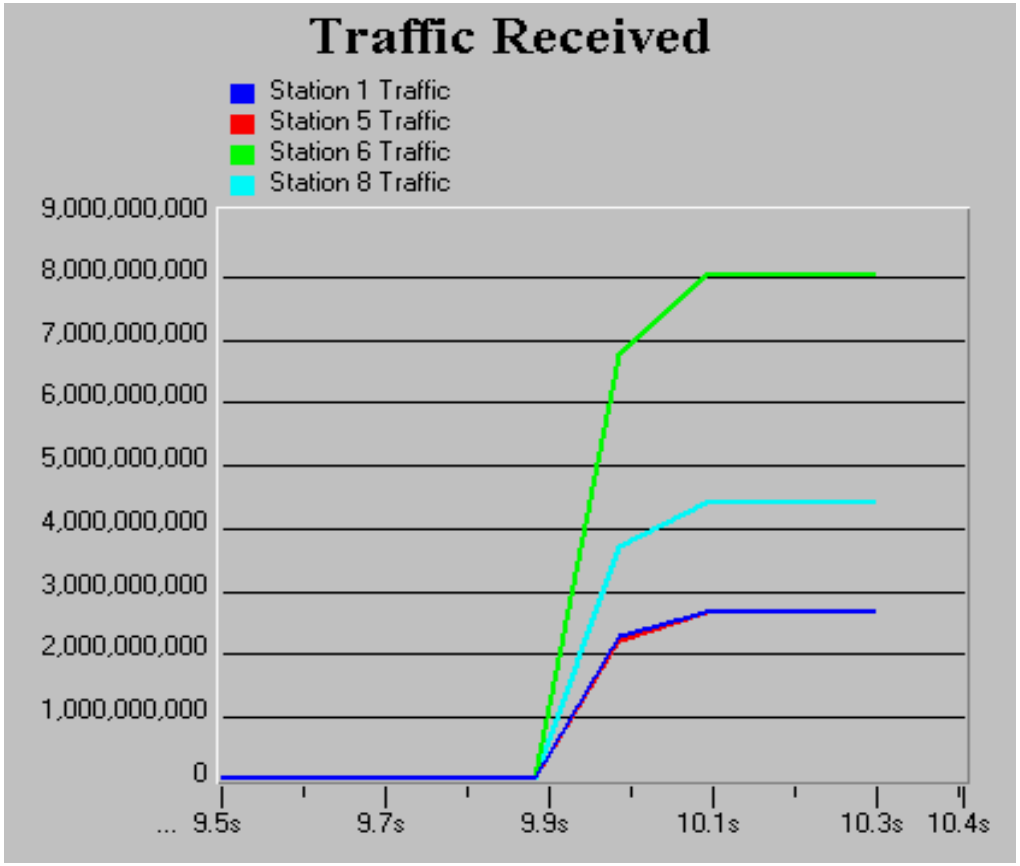
Hubbing Topology Scenario I



Setup for Hubbing Topology Scenario I

- **Station 1, 6 & 8 are transmitting to station 7**
- **Station 5 is transmitting to station 10**
- **Station 7 and 10 are attached to the same switch (switch 7)**
- **All stations generating CBR, 8Gbps**

Results: Hubbing Topology Scenario I (BW allocation)



Analysis: Hubbing Topology Scenario I

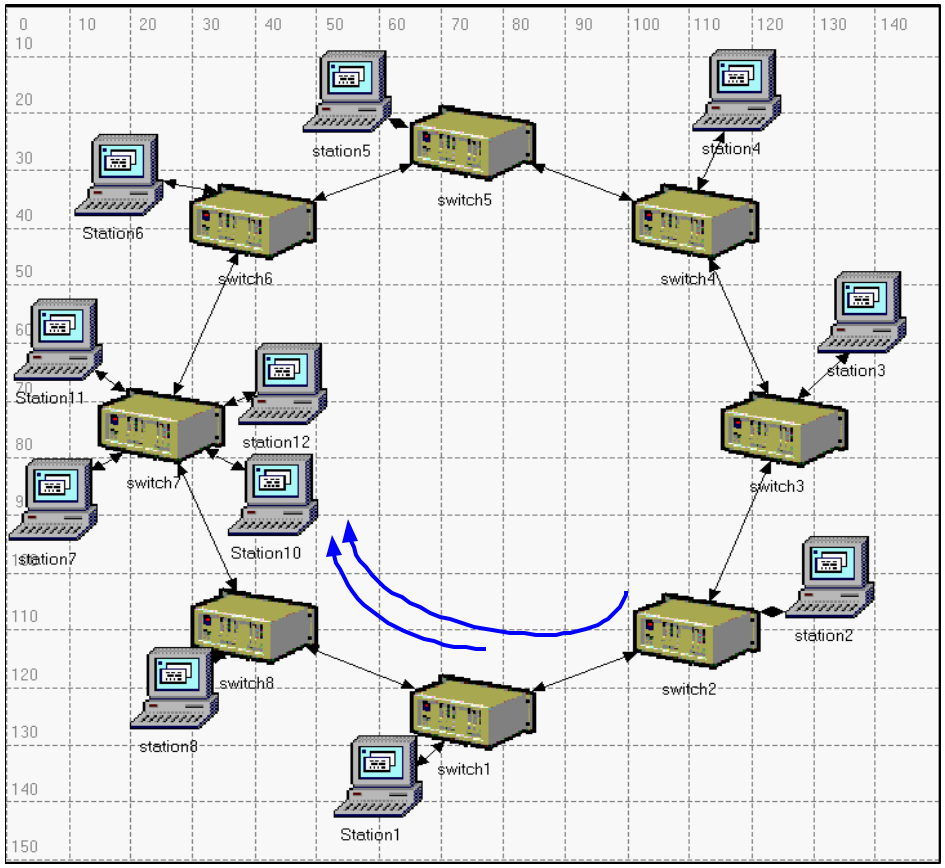
- **Station 1, 5, 6 & 8 share the BW but:**
  - **Station 1 & 5 get 2.7 Gbps**
  - **Station 8 gets 4.4 Gbps**
- **Station 8 gets all the BW it needs since it's one hop away from dest**
  - **Demonstrates location unfairness in BW use**

Results: Hubbing Topology Scenario I (ETE Delay)

Analysis: Hubbing Topology Scenario I ...

- **Station 6 has a constant ETE delay as expected**
- **Stations 1, 5, 6 have queues growing infinitely resulting in infinitely increasing ETE delays**
  - **Demonstrates location unfairness in ETE delay**

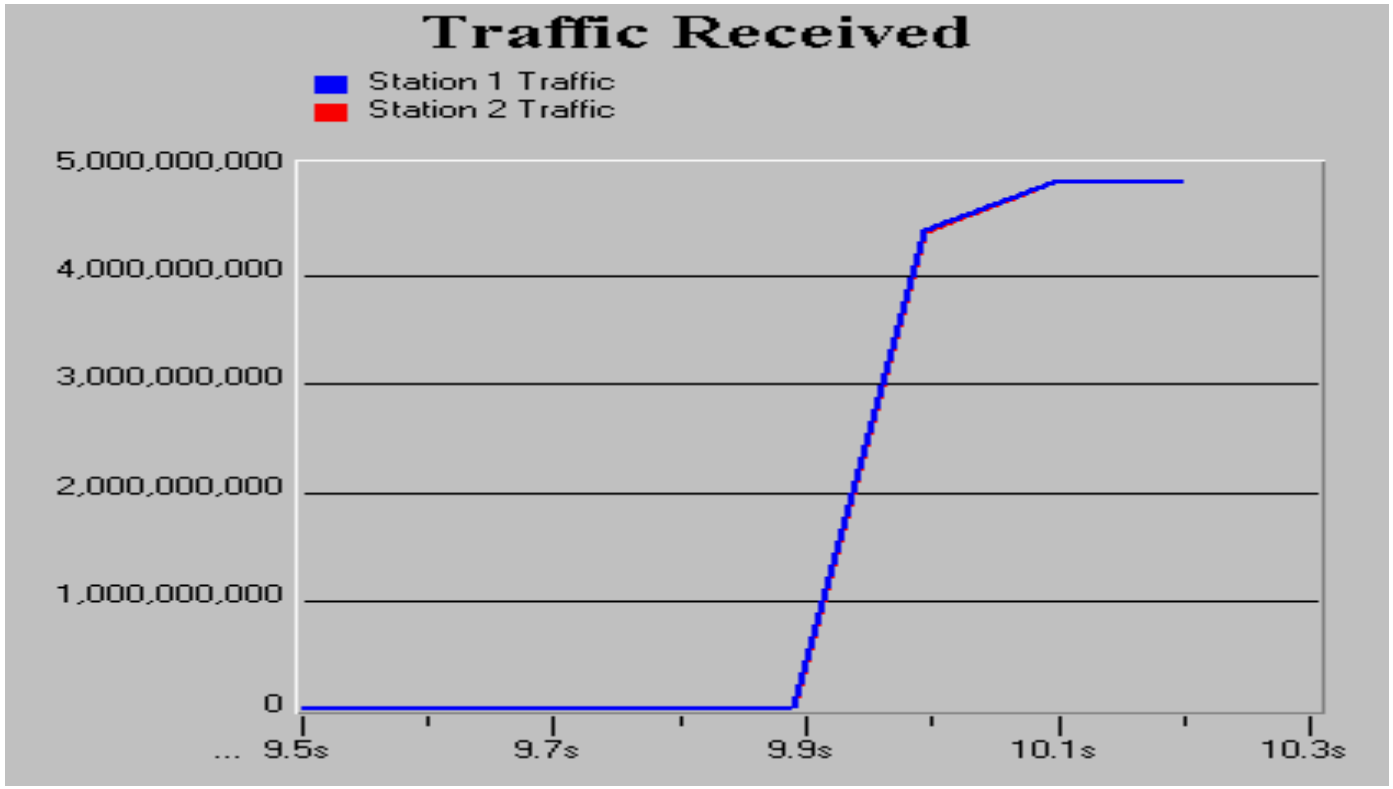
Hubbing Topology Scenario II



Setup for Hubbing Topology Scenario II

- **Station 1 is transmitting to station 10**
- **Station 2 is transmitting to station 12**
- **All stations generating CBR, 8Gbps**

Results: Hubbing Topology Scenario II



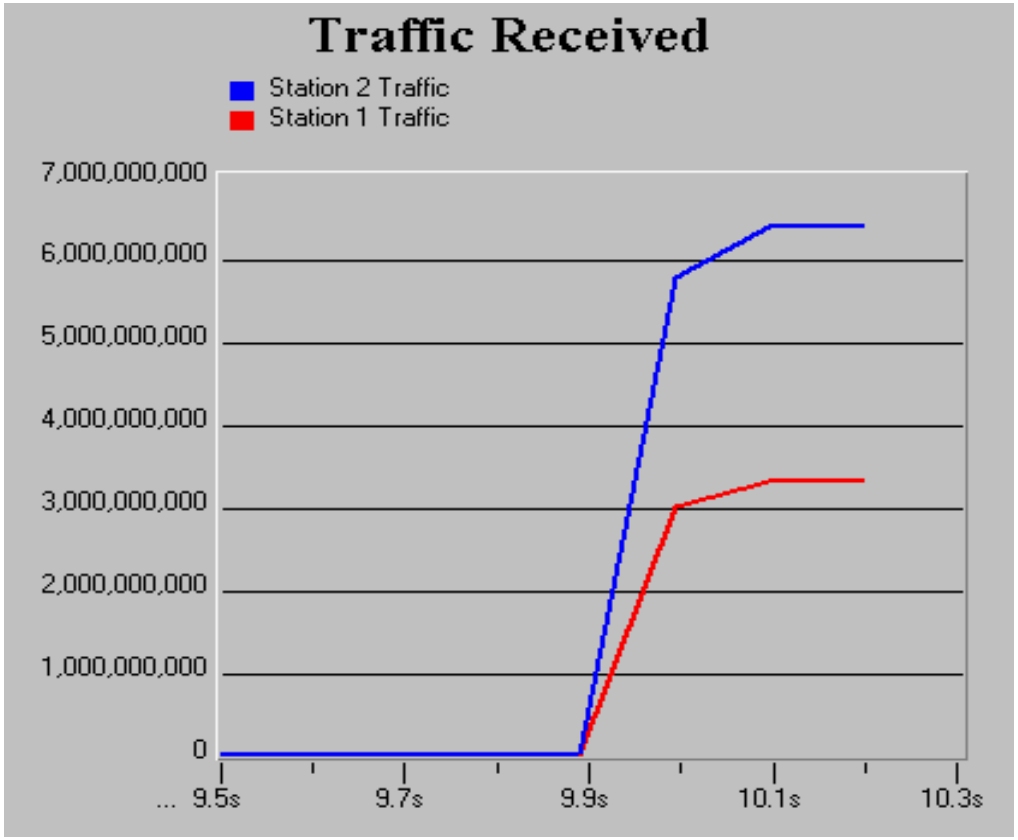
Analysis: Hubbing Topology Scenario II

- **Station 1 & 2 get the same BW in this simple case**
  - **which is good … but hold on!**

Hubbing Topology Scenario III

- **Similar to Scenario II except:**
  - **Station 2 is sending traffic twice the amount of traffic that Station 1 is sending**
    - **Station 2 is sending 10 Gbps**
    - **Station 1 is sending 5 Gbps**
  - **Hence the ring is overloaded:**

Results: Hubbing Topology Scenario III



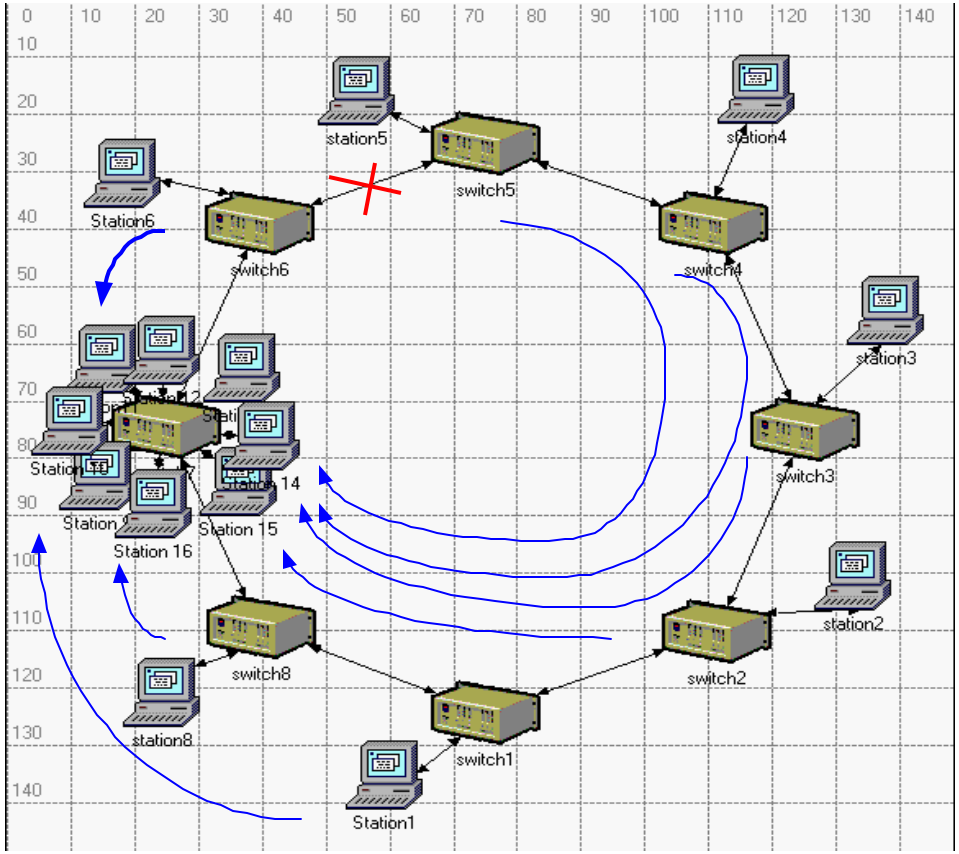
Analysis: Hubbing Topology Scenario III

- **Station 2 gets twice the BW of station1 even though both don't get all the BW needed**
- **No reservation mechanism at the MAC level to prevent a greedy station from chewing up more or most of the BW**

Intro to Hubbing Topology Scenario IV

- **More stations sending to a single hub**
  - **7 active stations**
- **Staggered input for source stations**
  - **0.4 seconds apart**
- **When all stations are on, the aggregate BW demand is 20Gbps (200% overload)**
- **All stations transmit to separate stations on the same hub switch (switch 7)**

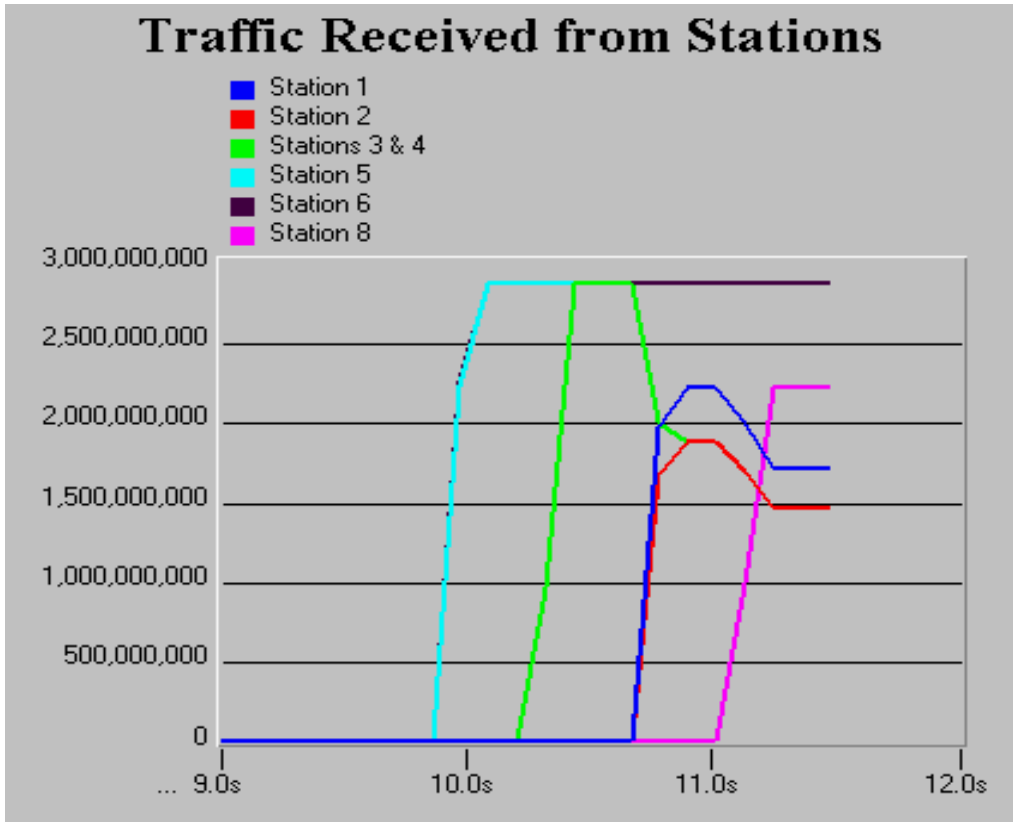
Hubbing Topology Scenario IV



Setup for Hubbing Topology Scenario IV

- **Station 5 & 6 start first**
- **Followed by stations 3 & 4**
- **Followed by stations 1 & 2**
- **Finally followed by station 8**
- **All stations generating CBR, 2.8Gbps**
  - **Gives aggregate BW demand of 20Gbps**

Results: Hubbing Topology Scenario IV (BW Allocation)



Analysis: Hubbing Topology Scenario IV

- **In underload conditions each station gets its BW demand granted**
- **As ring gets overloaded, unfairness in BW usage occurs:**
  - **Station 2, 3, 4 & 5 get 1.5 Gbps**
  - **Station 1 gets 1.7 Gbps**
  - **Station 8 gets 2.25 Gbps**
  - **Station 6 gets 2.8**

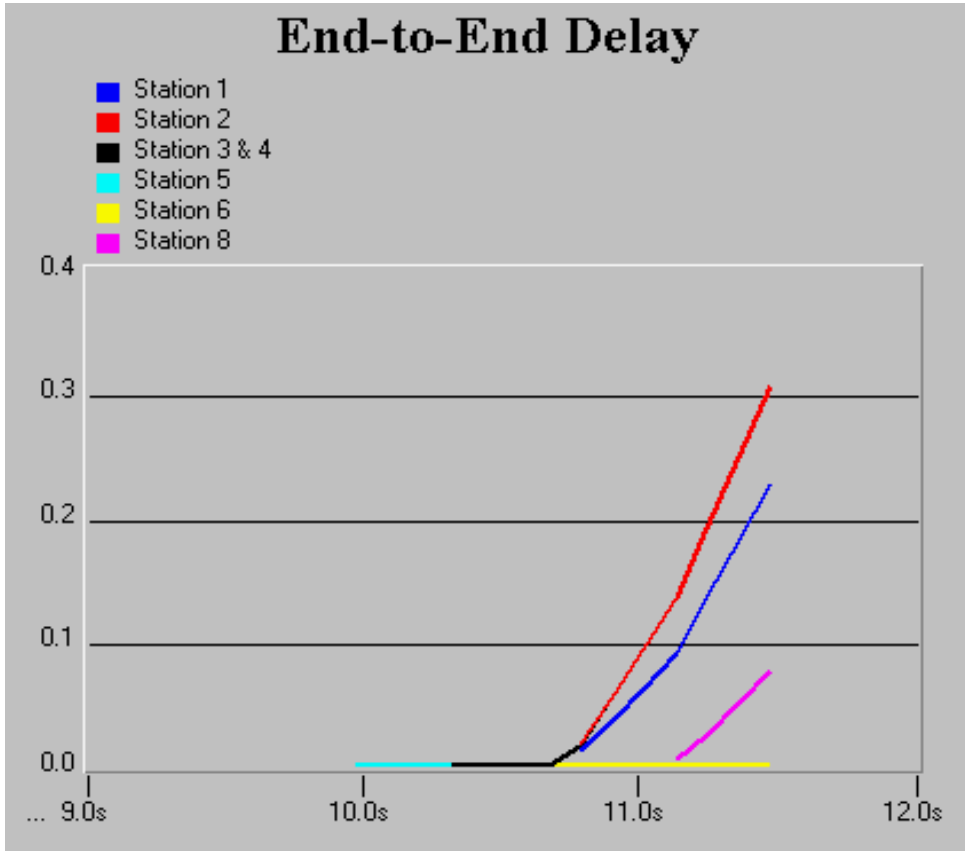
Analysis: Hubbing Topology Scenario IV ...

- **In overload unfairness is caused by Spanning Tree**
  - **Forces traffic from Station 5 to go around the ring**
  - **This competes with traffic from Stations 1, 2, 3, 4 & 8**
  - **Station 6 has no competition**

Analysis: Hubbing Topology Scenario IV ...

- **There is even unfairness between stations 1, 2, 3, 4, 5 & 8**
  - **Stations 2, 3, 4 &5 compete for BW on link between switches 1 &2 since that link is overloaded**
  - **Then they all compete with S1 for BW on link between switches 1 & 8**
  - **Then they all compete with S8 for BW on link between switches 8 & 7**

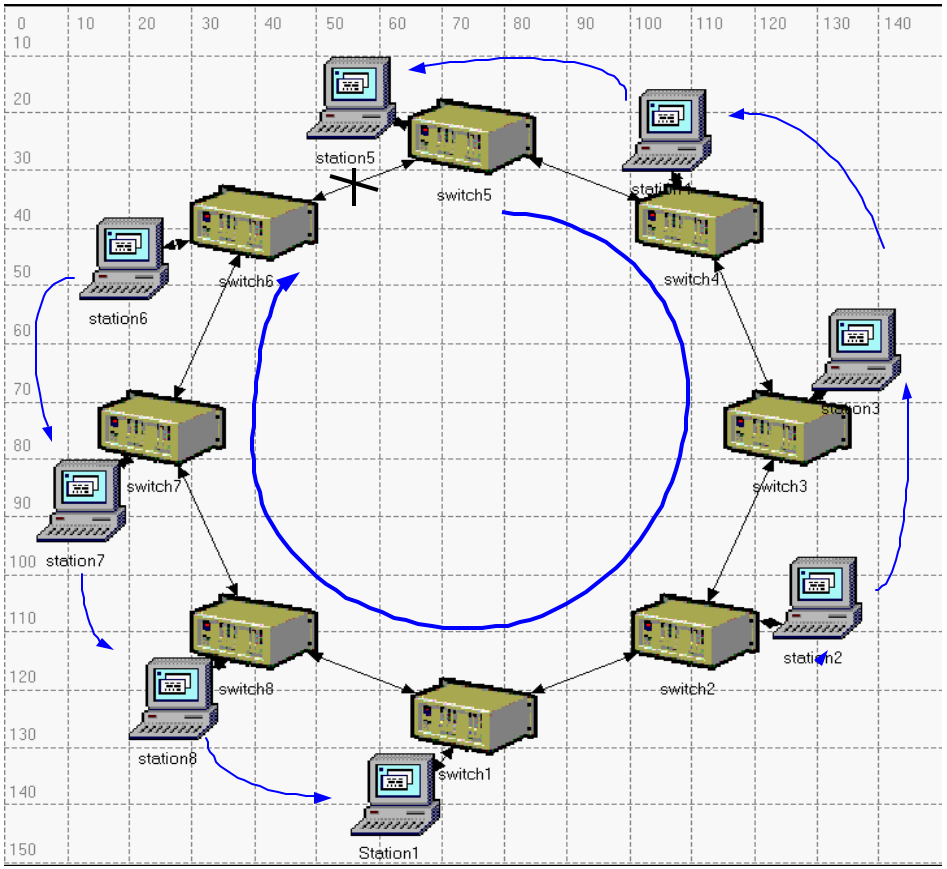
Results: Hubbing Topology Scenario IV (ETE Delay)



Analysis: Hubbing Topology Scenario IV ...

- **ETE delay varies:**
  - **S6 has constant delay**
  - **Others have delays growing infinitely due to infinitely growing queues**

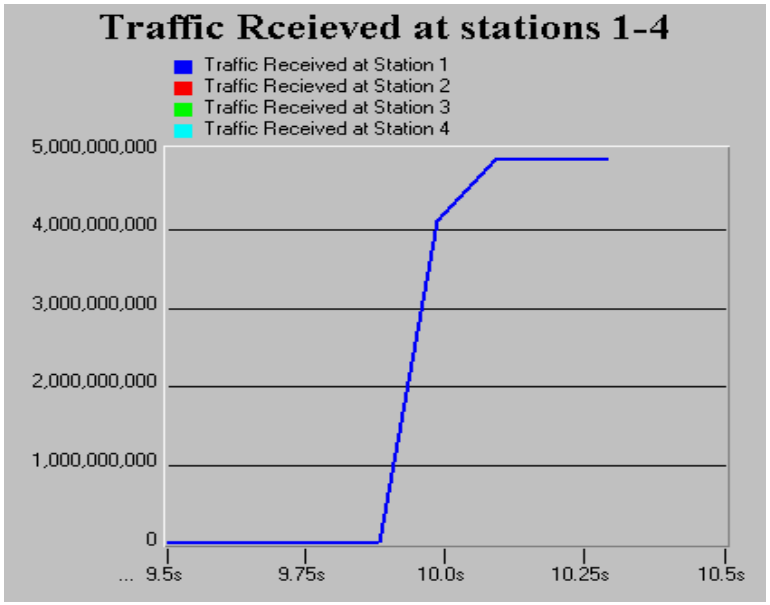
Next Hop Topology Scenario I

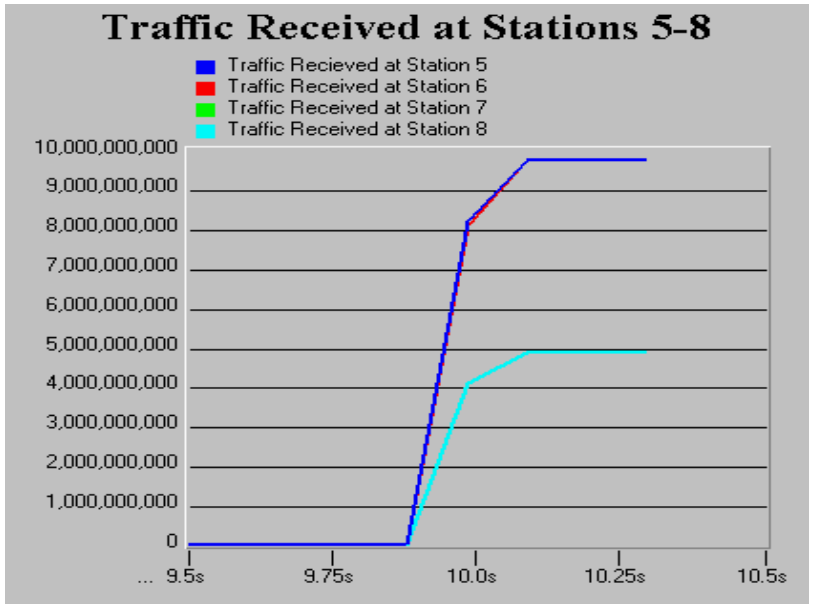


Setup for Next Hop Topology Scenario I

- **Station n sends to station n+1 (8Gbps CBR)**
  - **Station 8 sends to station 1**
  - **Station 5 sends to station 6 around the whole ring**

Results: Next Hop Topology Scenario I





Reminder of Ethernet Switch Behavior

- **Ethernet switches learn MAC address by monitoring traffic received and mapping MAC addresses to switch ports**
- **If a switch does not know on what port to send a packet, it will send to all ports**

Analysis: Next Hop Topology Scenario I

- **Stations 5, 6 & 7 get the full 8Gbps as requested**
  - **They face no contention for BW**
- **Stations 1, 2, 3, 4 & 8 get only 5Gbps!**
  - **The only two switches that learned about S6 location are switches 6 & 7**
  - **All others will send packets destined to S6 to all the switch ports**

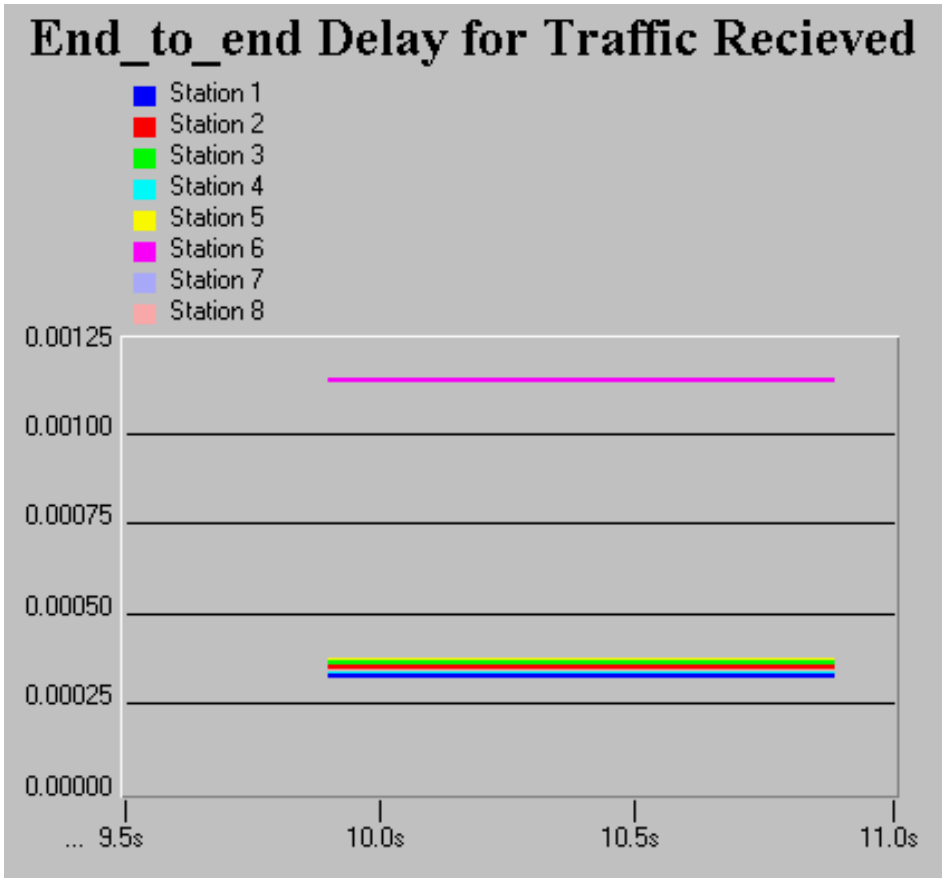
Analysis: Next Hop Topology Scenario I ...

- **This includes the link directly connecting these stations to the corresponding switch!**
- **Hence BW for the links between these stations and corresponding switches is split 50/50**

Next Hop Topology Scenario II

- **Similar to scenario I except:**
  - **Each station generates Poisson traffic at a 1 Gbps rate.**
  - **This was done to ensure that none of the links will be overloaded to allow evaluation of end-to-end delay**

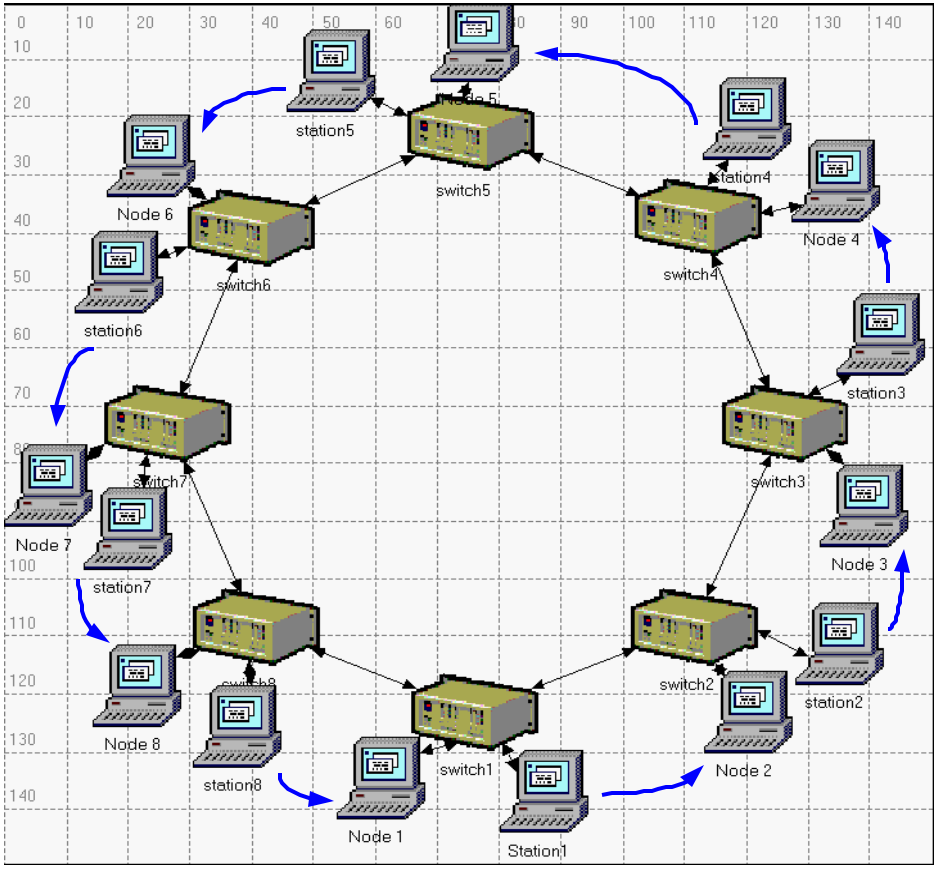
Results: Next Hop Topology Scenario II



Analysis: Next Hop Topology Scenario II

- **ETE delay:**
  - **Large for traffic received by station 6**
  - **Small for all other stations**

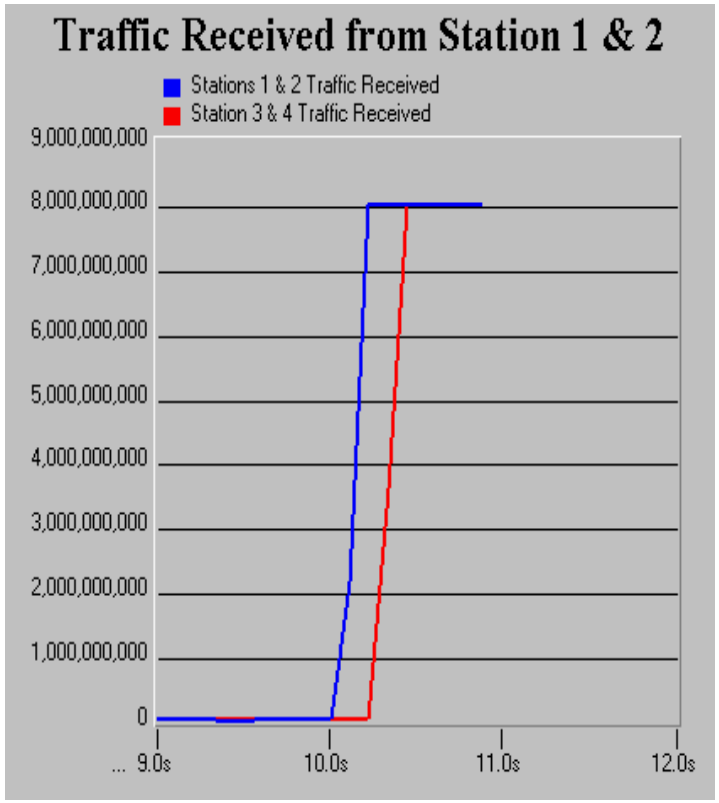
Next Hop Topology Scenario III

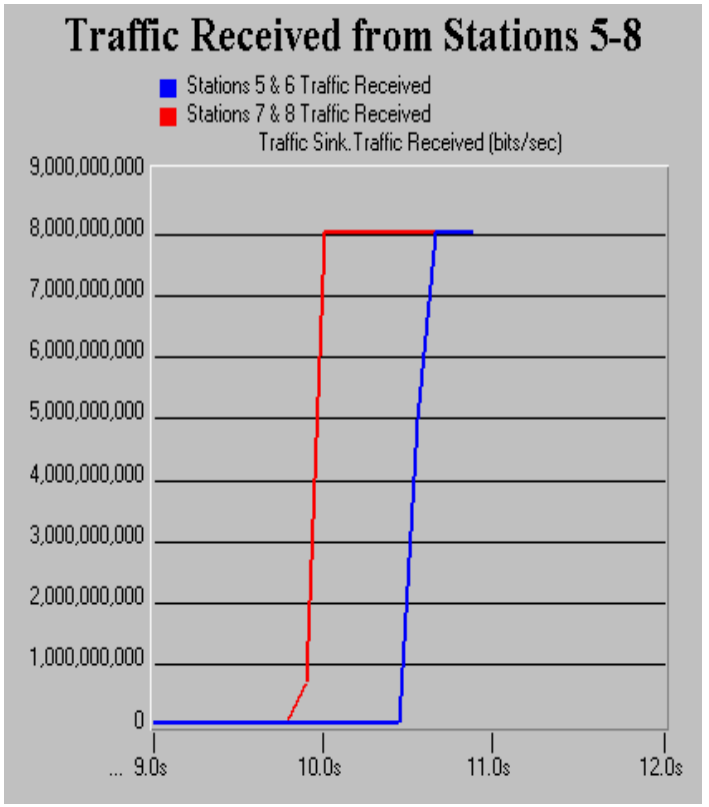


Setup for Next Hop Topology Scenario III

- **Similar to scenario I & II except:**
  - **Attached to each switch is a Station and a Node**
  - **Stations are sending to Nodes**
  - **Nodes 1 - 8 start by transmitting 100 packets/sec for a short period of time to allow the switches to learn their MAC addresses**
  - **Staggered as in hubbing scenario IV**

Results: Next Hop Topology Scenario III





Analysis: Next Hop Topology Scenario III

- **All stations get the BW of 8 Gbps as requested**
- **This is because provisions were made to allow the switches to learn the MAC addresses of receiving nodes**

Conclusions

- **Ethernet switches for RPR configurations exhibit fairness problems in:**
  - **BW allocation**
  - **ETE packet delay**

Suggestions for future simulations

- **Bursty traffic**
- **TCP and UDP apps (and combinations)**
  - **ftp, http, video-conferencing, voice, video streaming**
- **Multiple rings?**
- **Mesh of rings?**
- **RPR Ring vs. Mesh of switches**
- **Performance behavior when Link fails**

Suggestions for future simulations ...

- **Throughput per flow and per class**
- **More scenarios for next hop and hubbing (?)**
- **Simulations for the random configuration**
  - **Stations send to random stations**
- **Packet size distributions (if needed)**
- **More scenarios with various traffic generation distributions:**
  - **Nodes generating traffic move around ring**

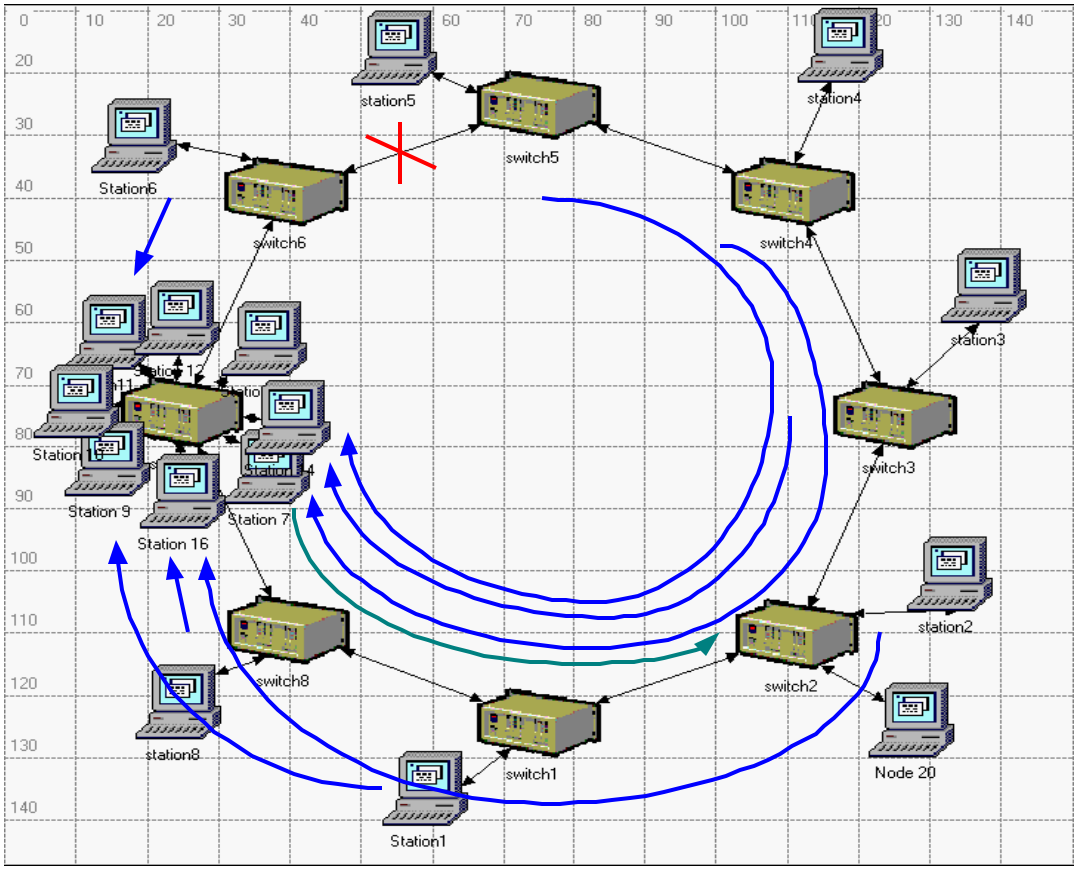
Suggestions for future simulations ...

- **Other metrics:**
  - **Same analysis for jitter**
  - **Packet loss (?)**
  - **Congestion control**
  - **Fault recovery**

Discussions

Additional Slides

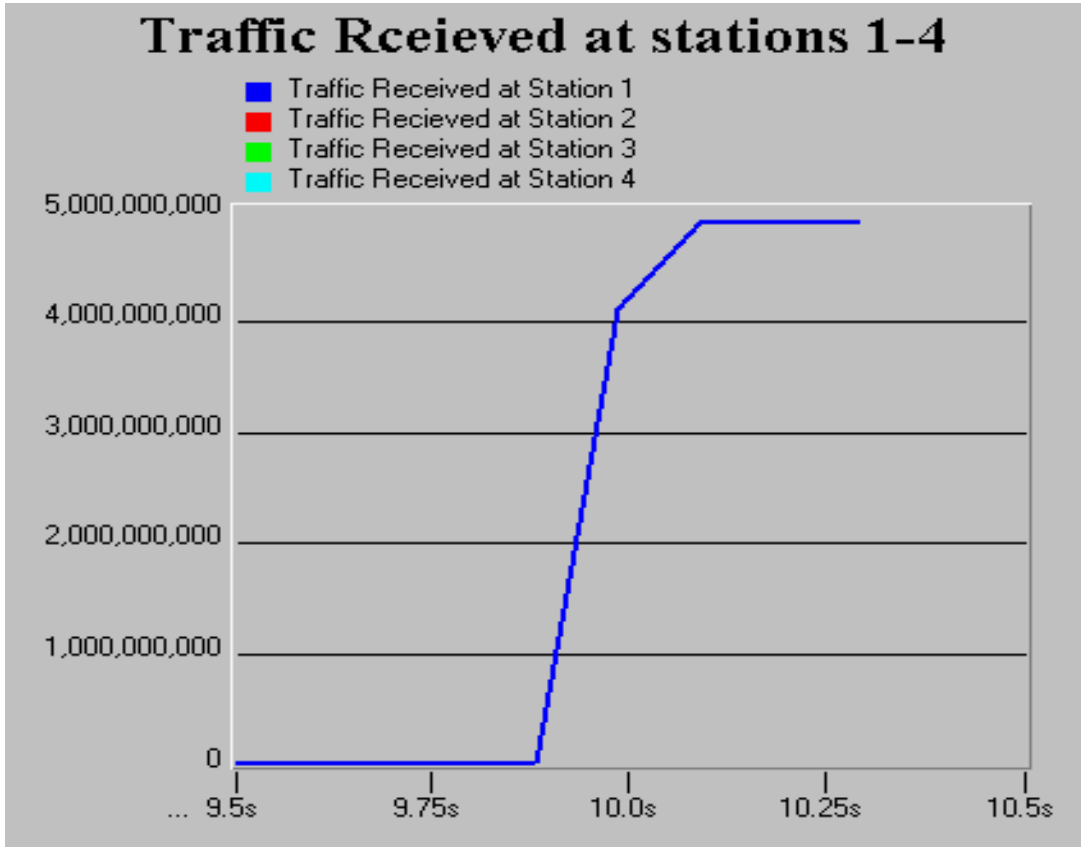
Hubbing Topology Scenario V



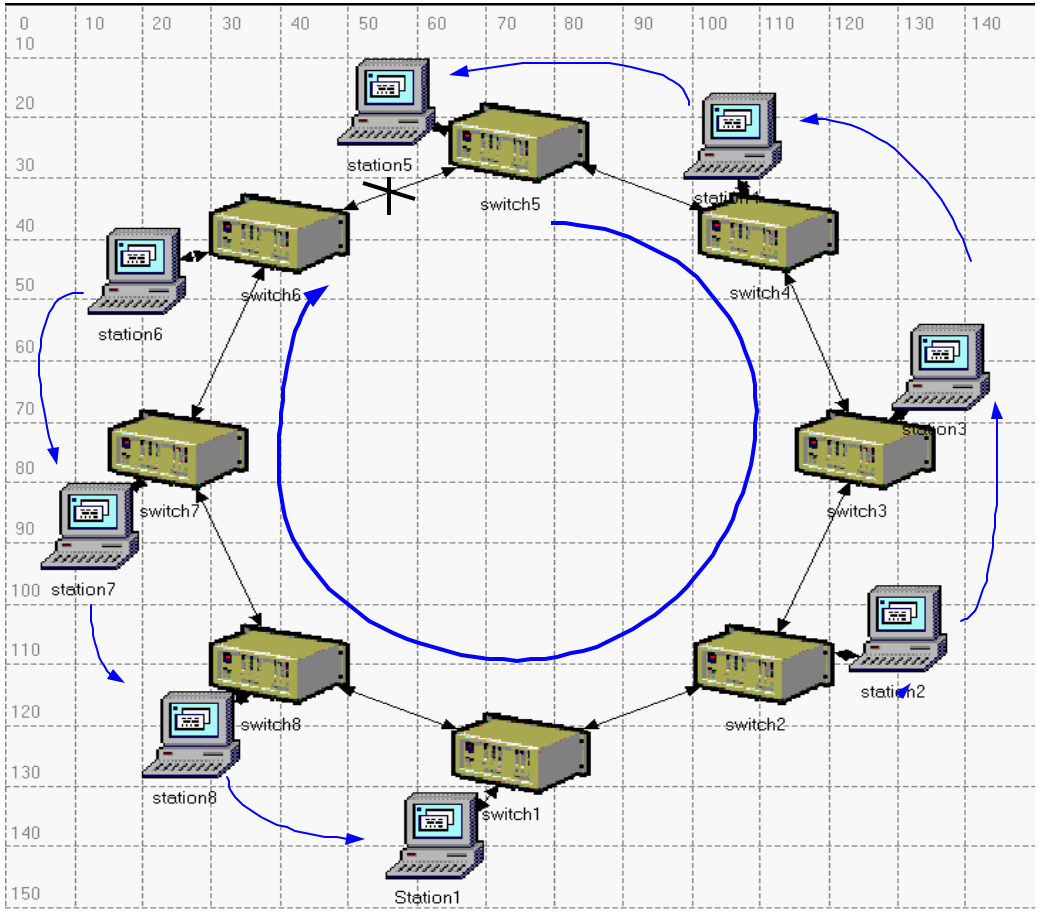
Setup for Hubbing Topology Scenario V

- **Similar to Scenario IV except:**
  - **Station 7 is sending CBR traffic to Node 20 at 9.5 Gbps**
  - **what Station 1 is sending and the ring is overloaded:**
    - **Station 2 is sending 10 Gbps**
    - **Station 1 is sending 5 Gbps**

Results: Hubbing Topology Scenario V



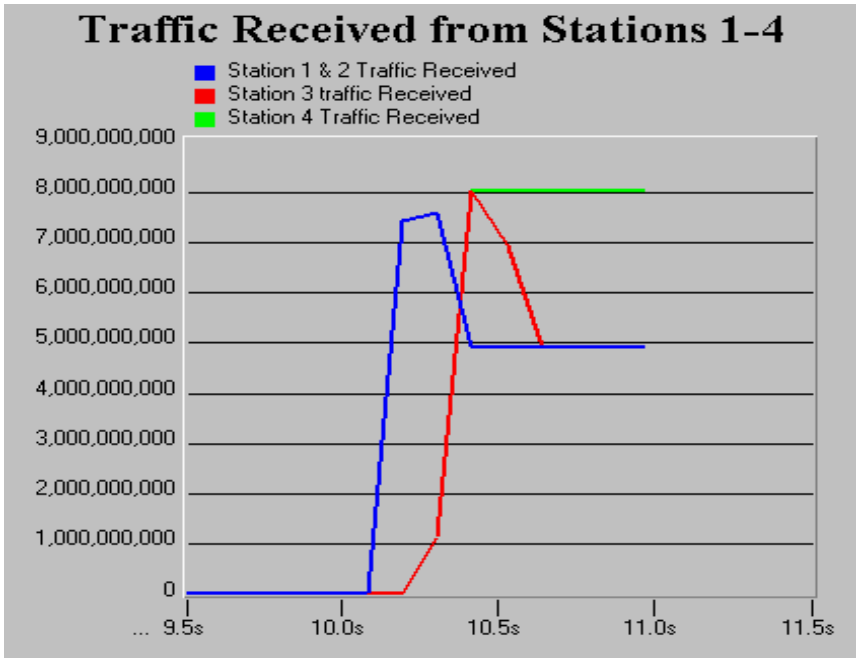
Next Hop Topology Scenario III

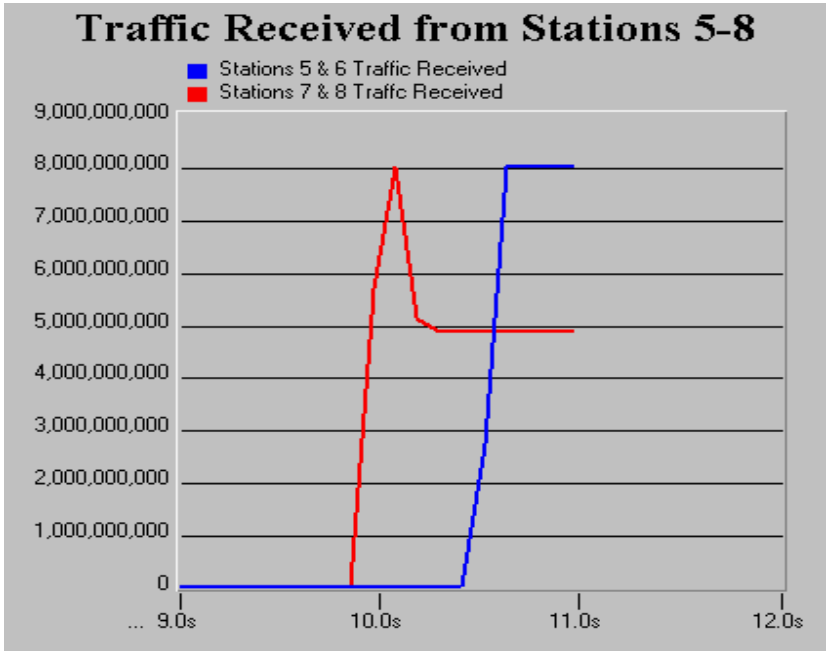


Setup for Next Hop Topology Scenario III

- **Station n sends to station n+1 (8Gbps CBR)**
  - **Station 8 sends to station 1**
  - **Station 5 sends to station 6 around the whole ring**
- **Staggering:**
  - **Station 7 & 8 start first**
  - **Station 1 & 2 next**
  - **Station 4 & 5 next**
  - **Finally station 6 & 7 next**

Results: Next Hop Topology Scenario III (BW Allocation)





Results: Next Hop Topology Scenario III (ETE Delay)