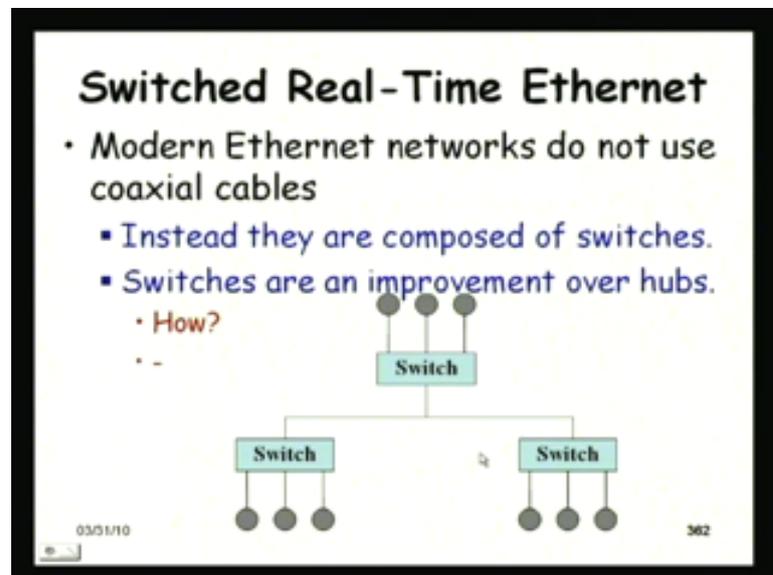


Real-Time Systems
Prof. Dr. Rajib Mall
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture No. # 36
Performance of Two Real-Time Communication Protocols

Let us get started we will continue to discuss from where we left last time we were discussing some protocols for communication of real-time messages in a Local Area Network environment and we had all most completed our discussion on that excepting a small technique that we could not discuss. We will just complete that few minutes and then we will start to discuss the performance of this protocol. How do one determine the performance of this protocol in such a situation and we will as a sample consider two protocols important Protocols .See how we can compare their performance and draw some meaningful conclusions from the performance comparison.

(Refer Slide Time: 01:27)



Let us look at the switch the real-time ethernet. We had seen that ethernet had some difficulty with respect to transmission of real-time messages can we with small modifications overcome that. We had seen that ethernet do not use coaxial cables composed of hubs and switches. So, can we do something to the switches have something implemented the switches. So, that we can have the priorities of the real-time

messages taken into account during transmission. We had said that the ethernet environment that we were discussing is not really suitable for real-time message transmission, because it does not distinguish between who is transmitting.

(Refer Slide Time: 02:32)

Switched Real-Time Ethernet

- In a star topology:
 - There is a dedicated cable in each direction.
 - Therefore, collision cannot occur on any cable.
- However, switches do not guarantee which packet will be sent first.
 - Solution:
 - Employ a switch with bandwidth reservation capabilities.

03/01/10 363

So, we had seen that in the star topology that the switches incorporate there is dedicated cable in each direction. So, using a switch collisions do not occur as compared to a hub. In a hub it is a broad cast and can be collisions in switch there is no collision and only thing is that switches do not guarantee which packet will be sent first. If we can have a switch where bandwidth reservation is possible node having a real-time message can reserve a bandwidth send a command message to the switch saying that it needs to transmit real-time messages for some duration.

If we have such a switch then we can also transmit real-time messages in the ethernet situation. There is no collisions here in a switch and if there is a reservation, see only difficulty with the switch is that if there are two or more requests come the switch does not guarantee which will be handled first. So, there is a reservation then the one that has reserved will be served. A simple extension to ethernet make it work for real-time messages. So, we will not discuss much about it neither there is too much on the simple concept.

(Refer Slide Time: 04:17)

Performance Comparison

- Performance comparison based on network utilization:
 - Bounded Access Protocol (IEEE 802.4)
 - Priority Based Protocol (IEEE 802.5)

03/31/10 364

Now, let us look at the performance comparison issue. How do we compare performance. So, we have to compare the performance based on the network utilization achieved based on whether some messages have missed out on the dead line at a specific utilization. We will see how we can compare performance of two important protocol the bounded access protocol that is 802.4 IEEE 802.4 and the priority- based protocol the IEEE 802.5. We had discuss about the working of these two protocols.

(Refer Slide Time: 05:02)

Classification of Message Sets

- Each network needs to cater to certain number and type of real-time messages:
 - Called the **real-time message set**.
- The types of real-time messages can be classified into:
 - **Unsaturated schedulable**
 - **Saturated schedulable**
 - **Unschedulable**

03/31/10 365

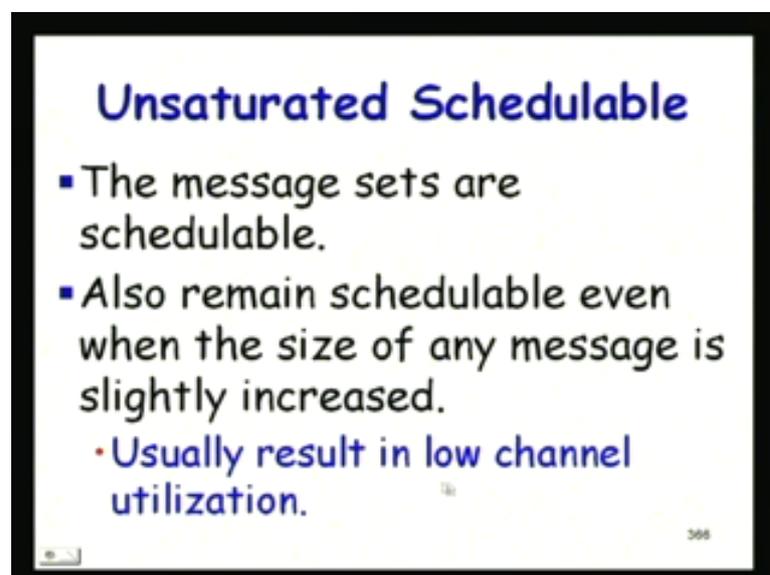
Now, before we can do the comparison first, let us try to see what kind of messages are handled the message sets for this networks. Now, we had seen that there can be real-time messages and then non real-time messages. The real-time messages will classify into two

types: The Unsaturated schedulable messages and the Saturated schedulable messages. So, talking of the message sets and we will also look at Unschedulable situation. Depending on the type of messages that are there in our problem we can classify that situation either as a unsaturated schedulable where all the messages can be easily transmitted without any impact on the dead line no messages misleads dead line.

So, if for a given situation that is there are set of messages in our system and we can design the network or in given in our a design all the messages can be transmitted without any messages missing its dead line then it is a schedulable situation is not it. So, that is the unsaturated schedulable , because even if some messages takes little bit more time then still none will miss their dead line. So, we have some amount of freedom here with respect to the messages taking little bit more time. But, if it is a saturated schedulable then for the message set the deadlines are just being met .If there is a small variation in any of the messages either in the arrival or may be in the time that need to transmit then some dead lines will be missed. So, that is the saturated set.

Unsaturated set even if there is small variations the dead line will continue to be met. Here the dead line is just met for all messages normally, but if there is a small variations even then some messages going to miss deadline to just able to meet the dead line. Unschedulable is that the situation where some message will miss dead line it was not possible to guarantee that messages will meet their dead line it is the unschedulable set. So, for a given situation if we cannot guarantee that all messages will meet their dead line that is the unschedulable set.

(Refer Slide Time: 07:55)



Unsaturated Schedulable

- The message sets are schedulable.
- Also remain schedulable even when the size of any message is slightly increased.
- Usually result in low channel utilization.

368

This is what I was telling you the unsaturated schedulable the message sets are schedulable and also remain schedulable even when the size of any message is slightly increased. So, Just remember here that what we are calling here is a message sets are the different situations .In one situation we had some messages and then we check that network for that set that is one set then in another situation that we are asked to design or test that is another message set. So, all the message sets that are possible .All the message sets that we are considering they will consider them as schedulable all the message sets and then they remain schedulable for any small change with the messages either the size increases little bit or there is a change in the arrival etcetera.

Then still they will remain schedulable and it does not take much to reason out that these do not achieve hundred percent channel utilization .There must be something free in the channel that is how we are able to accommodate even when there is a small change in any of the message. So, we are not taking of hundred percent channel utilization here low channel utilization. So, this is the set of the unsaturated schedulable message sets.

(Refer Slide Time: 09:29)

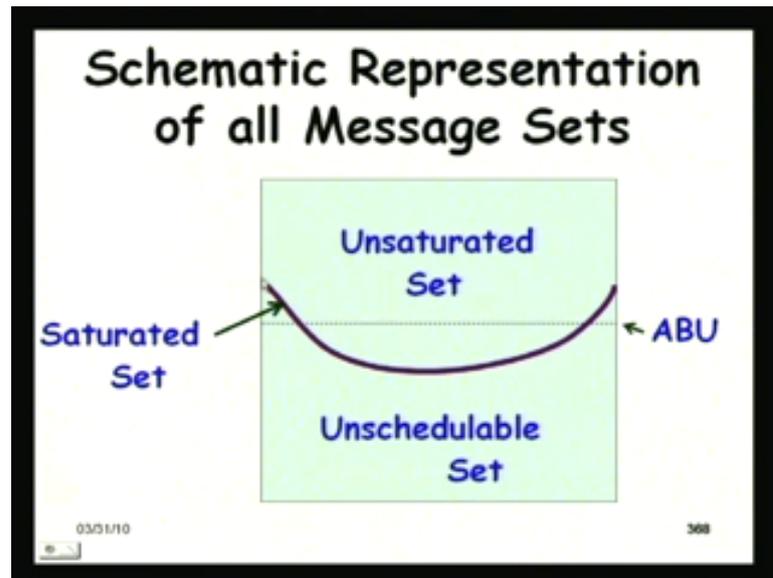
Classification of Message Sets

- **Saturated Schedulable:**
 - The message sets in this class are schedulable:
 - But any increase in the size of a message would make it unschedulable.
- **Unschedulable:**
 - Message sets for which deadlines of at least some messages would be missed.

03/01/10 367

We can have the saturated schedulable messages here the message sets are schedulable all messages meet their deadline, but whenever there is any change with any message then something will miss it is dead line. It will become unschedulable and unschedulable is that we know that some messages are missing their deadline that is the unschedulable set. So, given any scenario you can either classify it is the unsaturated schedulable, saturated schedulable or unschedulable.

(Refer Slide Time: 10:08)



Now, if we consider all possible scenarios that we might be asked to experiment all possible sets of messages in a network if we try to experiment. We will see that there is a unsaturated set for all these scenarios each point here is a scenario. Now, for all these message sets the messages are not only meet their dead line but, also any change in them do not cause any message to miss it is dead line remains schedulable these are the sets of unschedulable. We know that for this situation one or more messages miss their deadline, but ,then just see here this line it indicates those sets of messages where they have just remained schedulable any change will bring them to this site, they will become unschedulable. Do you see this line here it represents the saturated set of messages each point here is a scenario.

Message let us say node n_1, n_2, n_3 had some messages. So, that the point might be here the just meeting their dead line any small change we will make one of them these deadline another situation is this one. Does that appear and then we will define some metrics one we will call it is the Absolute Breakdown Utilization. (C) The dotted we will define a metric actually that we have not yet defined. Let me just define this metrics the absolute breakdown utilization and the guarantee probability two metrics we will define. Then we will reason that those two metrics are really study metrics or they are well formed metrics to compare performance of two protocols. Depending on the value of the ABU and GPU for two protocols we will reason out whether they are good or bad.

(Refer Slide Time: 12:42)

Utilization Metrics

- Performance comparison of different protocols would be made based on two protocols:
 - **Absolute Breakdown Utilization (ABU)**
 - **Guarantee Probability at a Utilization U ($GP(U)$)**

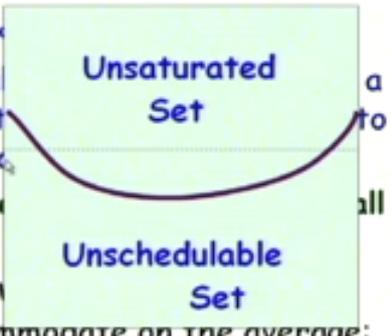
03/01/10 369

Now, let us see these two metrics. We will compare the protocols based on two metrics: one is the absolute breakdown utilization which we have just indicated in the previous diagram without really explaining. Now we will explain this. The absolute breakdown utilization ABU and the Guarantee Probability at Utilization U which we will in short form write GPU.

(Refer Slide Time: 13:12)

Absolute Breakdown Utilization (ABU)

- ABU of a network
 - The expected value of utilization due to some message set S at which the messages start to miss their respective deadline.
 - It is the average utilization of the messages in the network.
- ABU indicates how much utilization a network can accommodate on the average:
 - Without any message missing its deadline.



03/01/10 370

First let us try to understand this metric absolute breakdown utilization or ABU. To conceptually understand this metric ABU is the expected value of utilization due to some message set S at which the messages start to miss their respective dead line. As we keep on increasing the utilization of a network at some point they start missing their deadline.

So, this captures exactly that at what point as we keep on adding more and more messages at some point messages will start missing their deadline. This indicates exactly that of course, the average value, because it can occur for different scenarios. So, it is the average of utilization of all the messages in saturated set.

So, this is the another interpretation of this. If we can find out the saturated set of messages all the possible sets of messages where they are just schedulable and then find out what was the utilization for those sets of messages .Then that is exactly is the absolute breakdown utilization and this indicates how much traffic a network can accommodate on the average without a message missing its deadline. As I was saying that we keep on adding more and more messages real-time messages and at some point of time the real-time messages will start missing their deadline that is one saturated set there may be many scenarios like that. So, that is why we had drawn that line .Just see here we had computed the utilization due to all of this and the dotted line indicates the average utilization due to all the message sets in the saturated set.

(No Audio: 15:45-15:57)I think this picture is also available here. This is the average of all the messages in the saturated set.

(Refer Slide Time: 16:10)

Absolute Breakdown Utilization (ABU)

- Let $U(S)$ be the utilization of the channel due to message set S .
- Let C_i is the size of message $i \in S$, and T_i is its period.

$$U(S) = \sum_{i \in S} \frac{C_i}{T_i}$$

$$ABU = \frac{\sum_{S \in Sat} U(S)}{|Sat|}$$

- Sat is the set of all saturated message sets.

371

Now, we can develop a expression for this. So, Let us say for some message set S how do you indicate the utilization due to that utilization is basically how much data is transmitted over what time C_i by T_i . The utilization due to the message set S will be indicated by $U(S)$ utilization due to message set say as $U(S)$ which is the summation of

all the message size divided by the period over which this size of data needs to be transmitted. For this set all the messages that are existing. The utilization due to each of those messages gives the utilization due to that set C_i is the size of the message and T_i is the period over which the C_i needs to be transmitted. Now, the ABU is a summation of all such utilizations for the saturated set of messages divided by how many message sets are there. The cardinality of the sat set and the utilization due to all saturated message sets this is the cardinality of the sat set.

So, the sat we are saying that the sat consists of many situations. We find that in some situation let us say we had node n_1 had got another message let us say some m_1 or let us say m_5 then it becomes unschedulable. So, that is one saturated set. Now, let us say node n_2 got another message. There can be various situations in which there can be an example of a saturation of the messages. All those situations sat is that set and this is the cardinality of the sat set and $\sigma U(S)$ is the sum of the utilization due to all the sets of messages in sat. Sat is the all the saturated message sets.

(Refer Slide Time: 18:55)

Guarantee Probability (GP)

- $GP(U)$: Probability that all deadlines of a message set with utilization U would be met.
- If utilization is lower than ABU:
 - What will be the value of $GP(U)$?
 - $GP(U)$ will be close to 1.
- If utilization is more than ABU:
 - How will the value of $GP(U)$ change?
 - $GP(U)$ will approach 0.

372

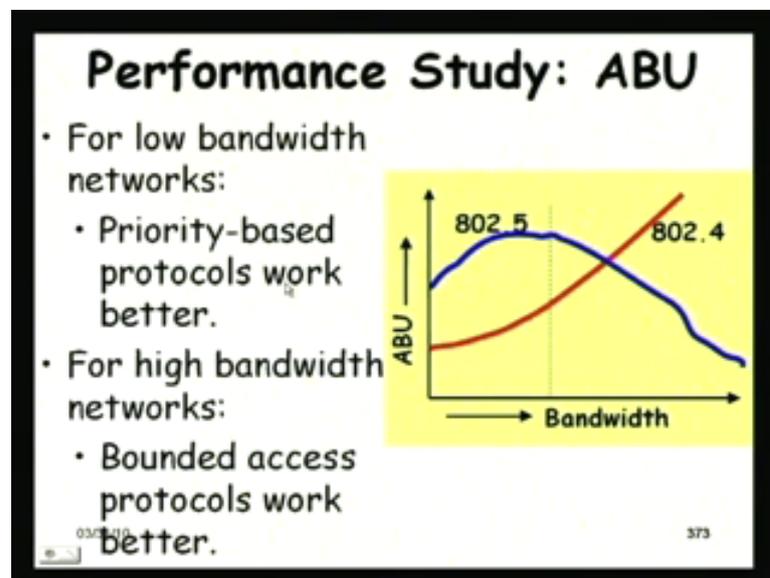
Now, let us look at the other metric that we are going to use the guarantee probability at a utilization value. So, the guarantee probability is always defined with respect to a utilization value. When the network is having certain utilization then, what is the probability that all the deadlines are met. See as you keep on increasing the utilization of the network when, there is low utilization all the deadlines will possibly be met and as the utilization you keep on increasing with real-time messages at some time or other we will see that the deadlines are getting missed. So, this indicates what is the probability if

the utilization is U then what will be the probability that the deadlines will be met. See as the utilization lower than ABU the absolute breakdown utilization.

This is the average of the utilization due to all the messages in the saturation set then what will be the value of GPU. What do you think, if the utilization lower than absolute breakdown utilization then the message set will be schedulable is not it. The utilization is lower than what was for the saturated set that is what we are saying lower than ABU. Then the guarantee probability will be almost one is not it. So, GPU will be close to one why not exactly equal to one, because you are considering the average value. See this is the average absolute breakdown utilization this is the average of utilization due to all the messages in the saturation set.

Still, some messages at that utilization can miss the deadline, but what about if the utilization is more than ABU how will the GPU change it will approach zero is not it. It is more than ABU then just see the message set was saturated and now if we increase above that the utilization is more than that then definitely something or other going to miss their dead line. So, it will be closed to the probability that the none of the message will miss their dead line probability will be almost zero.

(Refer Slide Time: 22:01)



So, that is the interpretation of the guarantee probability. Now, based on this the two metrics. There is a evaluation of the two protocols and this is for the ABU metric the 802.4 protocol as the bandwidth increases the absolute breakdown utilization increases. Is it intuitive or counter intuitive? as the bandwidth increases of the network the absolute

breakdown utilization increases. Is it intuitive or counter intuitive?(()) Does it agree with your intuition that as you keep on increasing the network bandwidth then the average breakdown utilization for that network increases its intuitive is not it. Because if your network has higher capacity then definitely it will be able to transmit more messages. So, that is about it.

The absolute breakdown utilization improves with bandwidth that is intuitive that is for the 802.4 protocol which is the bounded access protocol, but for the 802.5 just see the behavior it improves for some time which is intuitive and then suddenly starts to drop very fast here. The absolute breakdown utilization falls even below this that means, that by increasing the bandwidth of the network you have used a faster network and you are gaining nothing, actually you are losing. So, how is that possible let us examine that. When the bandwidth is low see here. When the bandwidth is low bandwidth network the priority-based protocol is working better. See here, it is providing higher absolute breakdown utilization that means, that given a message set the chances that it will work or none of the messages will miss their deadline is greater than in 802.5.

The priority protocol and it may not meet the deadline in the bounded access protocol that is 802.4. So that is what is shown in this diagram. Observe this that given a message it is likely to meet all the messages meeting the deadline is held and then after some time the 802.4 becomes better and the 802.5 the performance really drops. So, for high bandwidth networks the bounded access protocol is working better. Why is that? I mean how come this appearing like this anybody can think of any reason.

(Refer Slide Time: 26:52)

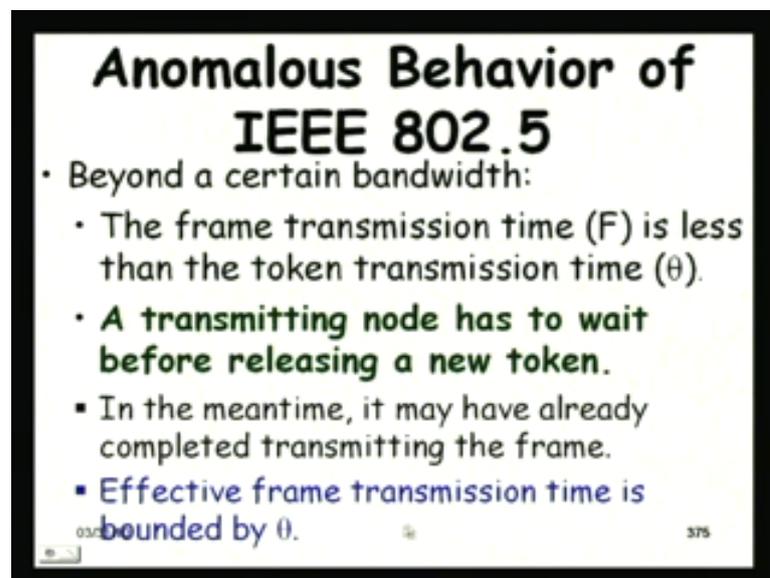
Relation between Performance and Bandwidth

- Intuitively performance of a protocol should improve with bandwidth.
 - Performance of IEEE 802.4 improves monotonically with bandwidth.
 - However for IEEE 802.5:
 - Performance initially improves, but starts to drop off beyond a certain point.
- Why this behavior?

03/01/10 374

Let us investigate this. Actually intuitively the performance of a protocol should improve with band width. Any protocol when the network becomes faster it should be able to even meet the deadline of messages for which it was not able to meet the dead line and we saw that 802.4 was close to this observation it was almost monotonically increasing with bandwidth as more and more bandwidth become available or the network become faster it could schedule more and more message sets. But the 802.5 the performance initially improved, but then after certain time it started to drop off from that point, but let us understand why this behavior ?

(Refer Slide Time: 26:56)

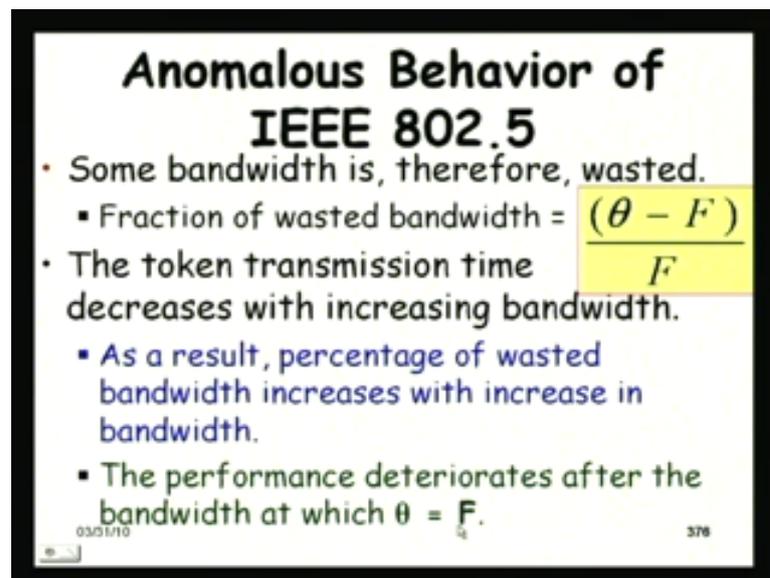


So, the idea is that if we analyze the 802.5 we will see that there are two components there: one is the frame transmission time and the token propagation time theta the token propagation time and the frame transmission time. The time to transmit a frame is maximum of F and θ , F is we had done some example also some exercises. So, F is size of the frame divided by the band width of the network is not it. So, the more and more bandwidth or the faster is the network the smaller becomes F is not it. The more is the bandwidth of the network or the faster is the network the smaller is the F , but θ remains constant for a specific type of medium.

So, beyond a certain bandwidth the transmitting node has to wait before releasing a new token, before starting to transmit a new frame if the network has high band width then it just idles it has transmitted the frame ,because the band width is more it could transmit the frame in almost no time. But the header of the token took time to arrive at the node. It was just idling there the channel was idle and there was a wastage of the bandwidth. It

transmitted that frame completed transmitting the frame and it was just waiting for the header of the token to transmit, because the network has high bandwidth could transmit very fast, but theta is dominating F and the effective transmission time is theta as a wastage of bandwidth and the fraction of wasted bandwidth we had seen this earlier also theta minus F by F. So, every F that is the frame transmission time theta minus F is wasted .See if theta was smaller than F then every F time one frame will be transmitted, but now every F it is idling for theta minus F

(Refer Slide Time: 29:33)



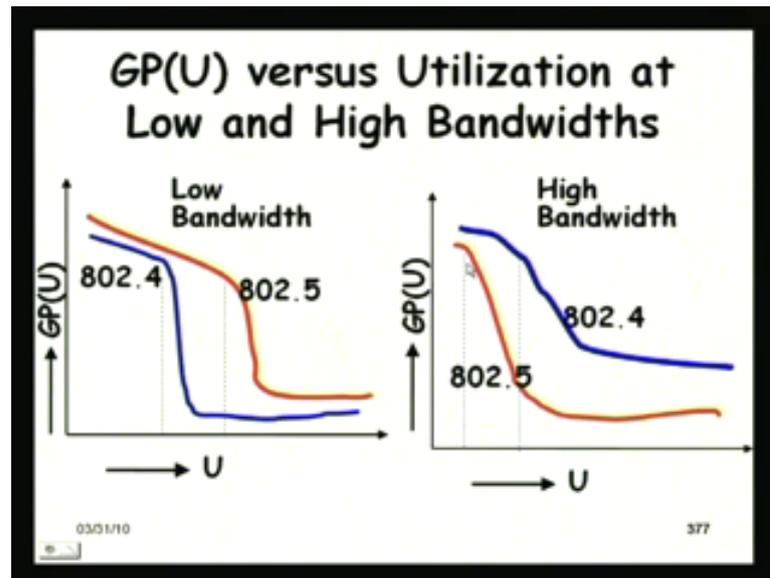
Anomalous Behavior of IEEE 802.5

- Some bandwidth is, therefore, wasted.
 - Fraction of wasted bandwidth = $\frac{(\theta - F)}{F}$
- The token transmission time decreases with increasing bandwidth.
 - As a result, percentage of wasted bandwidth increases with increase in bandwidth.
 - The performance deteriorates after the bandwidth at which $\theta = F$.

03/01/10 378

And the percentage of wasted bandwidth increases if we keep on increasing the bandwidth and F drops. The value of F keeps on dropping and theta minus F by F becomes larger and larger. So, the performance of the protocol deteriorates after the bandwidth at which the propagation time become equal to a frame transmission time. So, that we had shown with a line there after which it started dropping.

(Refer Slide Time: 30:43)



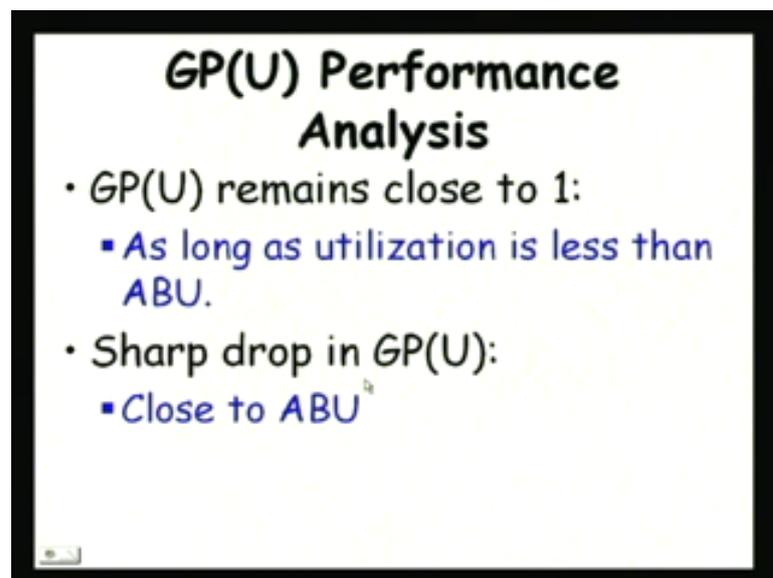
Now, let us look at the GPU the first metric we saw the ABU the performance with respect to ABU. Now, let us see the performance with respect to the guarantee probability. So, here we have two situations: one is a network the very low bandwidth and this is the network with a high bandwidth. Now, here the red line indicates the performance of the 802.5 protocol the priority protocol and the blue line indicates the 802.4 or the bounded access protocol just see here that is reversed here. Now, let us see why it is so and what it really means .See we had said that the guarantee probability at utilization at certain utilization is the probability for a given message set all deadlines will be met.

As the utilization improves or as the utilization becomes more and more the guarantee probability should drop is not it, because the network is getting more and more utilized and some messages are likely to miss their deadline. But see here at low utilization the priority-based protocol had a better performance .It could give more guarantee a given message set is more likely to meet it is deadline on the 802.5 protocol and then they kept both of them kept decreasing as is expected, but at certain time at certain utilization there is a sharp drop here also there is a sharp drop and then again they almost stabilize there and this sharp drop actually indicates the absolute breakdown utilization value. So, we were saying that at certain utilization suddenly the messages set will start missing their deadline. So, this sharp drop here indicates the absolute breakdown utilization from this point onwards this is the ABU for this one. This is the ABU for this one.

So, this similar thing we had seen under ABU plot. Now, for a high bandwidth the situation is just reversed. The 802.5 is performing better from the beginning and keeps on performing better. It is not difficult to reason out that this is due to... Let me just ask you this question that why should at low bandwidth the priority-based protocol perform better and as the bandwidth increases the bounded access protocol performs better with respect to successfully scheduling a message set. So, what do you think? But why should wasted bandwidth is less then why should it perform better than 802.4. Why should the priority- based protocol perform better in the low bandwidth situation and not perform as good in the high bandwidth situation. Not hard to reason out.

Not that complicated, simple answer see the 802.5 protocol considers priority-based reservation. It does consider priority, but 802.4 does not really consider priority. It just gives in every TTRT it gives some time for each node to transmit. So, since it does not consider priority and the bandwidth is low then the TTRT is likely to be high and the chances of messages missing given a message set scenario might miss it is dead line here, because it is considering priority. So, the one that has the highest priority will try to will get the chance to transmit. So, it is able to meet the dead line. But here, the situation is reversed, because in 802.5 lot of wastage of band width is occurring. So, even if it is considering the priority of the messages, but the wastage of bandwidth is causing this to perform worse than the 802.4.

(Refer Slide Time: 36:28)

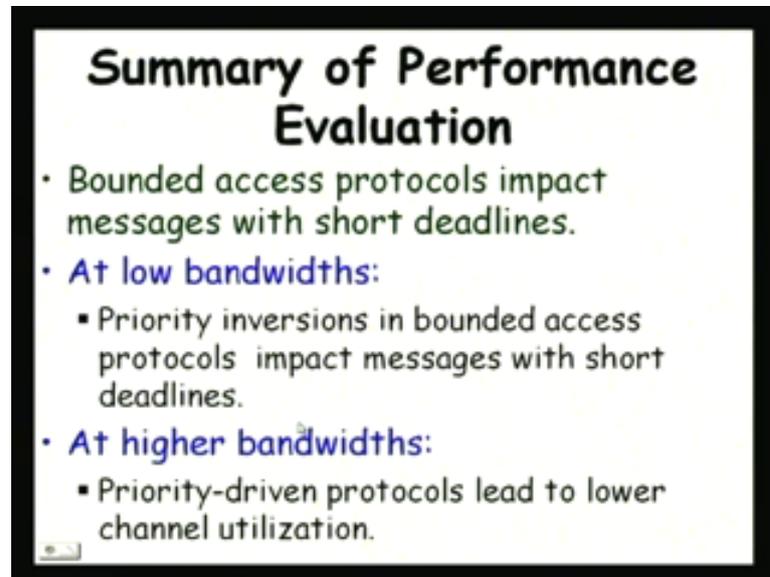


GP(U) Performance Analysis

- GP(U) remains close to 1:
 - As long as utilization is less than ABU.
- Sharp drop in GP(U):
 - Close to ABU

The GPU remains close to one when the utilization is ABU and then sharply drops close to the ABU.

(Refer Slide Time: 38:46)

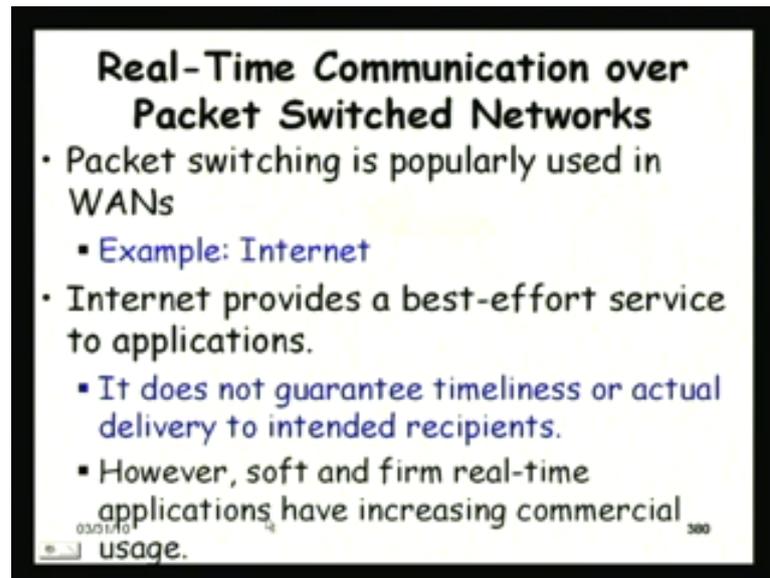


The absolute breakdown utilization and the main summary of the performance evaluation is that a bounded access protocol for short deadlines it does not work that way, if you are design situation is that you have many message with very short deadlines then, the bounded access protocol may not work well.

If the dead line is larger than bounded access protocol will be better suited at low band width the priority inversion and bounded access protocol impact the messages with short deadlines .See the Priority inversions are higher in bounded access protocol which is given by two into TTRT is not it.

Let us done some examples that the priority inversion is two into TTRT and if messages are short deadline they cannot tolerate two into TTRT, but at higher bandwidths the situation is changed, because the priority-driven protocols lead to wastage of bandwidth, the reservation field to check the reservation field they keep the channel idle.

(Refer Slide Time: 38:13)



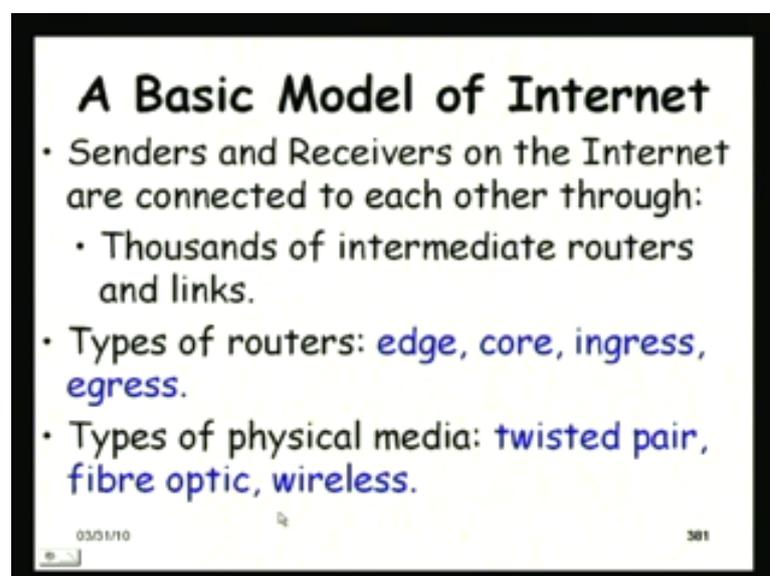
Real-Time Communication over Packet Switched Networks

- Packet switching is popularly used in WANs
 - Example: Internet
- Internet provides a best-effort service to applications.
 - It does not guarantee timeliness or actual delivery to intended recipients.
 - However, soft and firm real-time applications have increasing commercial usage.

03/01/10 380

Now, let us we had discussed some important protocols for real-time message communication in a local area network environment. Now, let us see how real-time communication can occur over a packet switched network wide area network basically and example of most popular example of a wide area network is the internet. We had seen earlier that it provides the best effort service to applications .It does not distinguish between the senders does not guarantee timeliness or actual delivery to their intended recipient. But, we have seen that soft and firm real-time applications are increasing every day. Commercial usage of internet for handling soft and firm real-time applications are increasing very fast including banking, VOIP and so on.

(Refer Slide Time: 39:23)



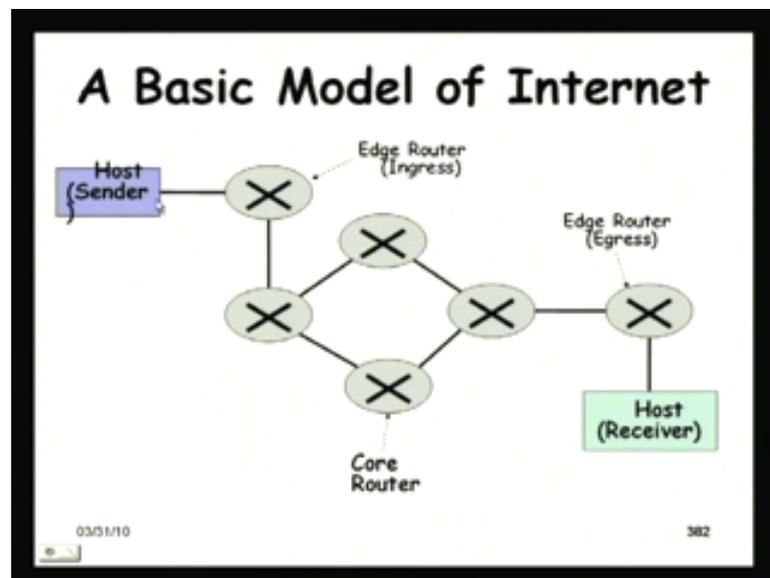
A Basic Model of Internet

- Senders and Receivers on the Internet are connected to each other through:
 - Thousands of intermediate routers and links.
- Types of routers: edge, core, ingress, egress.
- Types of physical media: twisted pair, fibre optic, wireless.

03/01/10 381

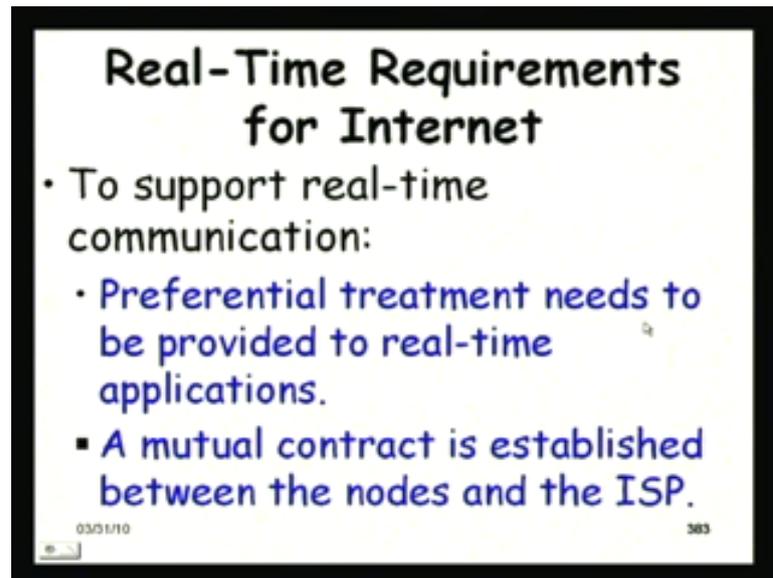
Now, before we discuss protocols for supporting real-time communication on the Internet. Let us look at a basic model of the internet. If we try to analyze the internet we will see that the senders and receivers are connected by thousands of intermediate routers and links. If we try to classify these different types of routers. We will see that there are three main Type of routers :the edge routers ,the core routers, the ingress and the egress routers and the physical media in a internet can be different some place it may be just twisted pair let us say LAN environment it can be fiber optic for a back bone. Let us say the gate way from India to United States of America or Singapore. You know that there are fiber optic cables laid under sea it can be wireless in a LAN environment.

(Refer Slide Time: 40:47)



So, if we look at the situation here see here the host this is the sender host and then this is the receiver. Then the ingress router we will distinguish between two forms of edge routers, the ingress and the egress routers then the core routers here. These two have special responsibilities .The router connected to the sender and the router connected to the receiver.

(Refer Slide Time: 41:31)



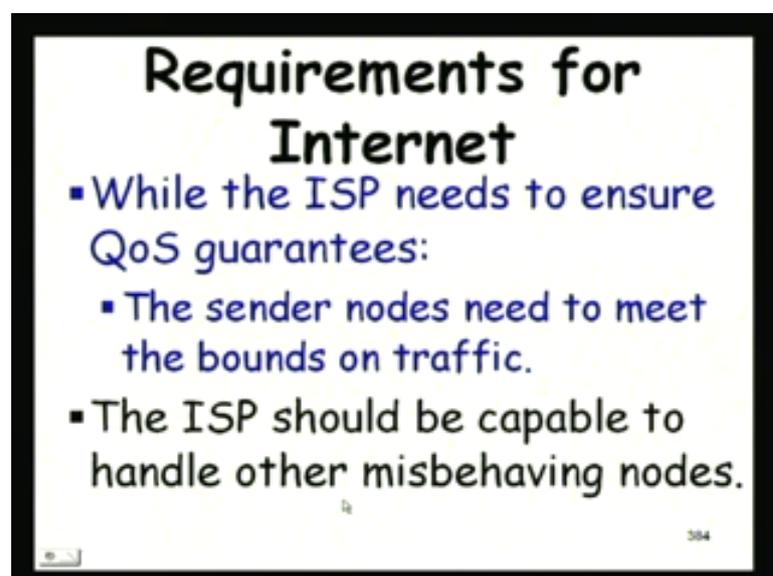
**Real-Time Requirements
for Internet**

- To support real-time communication:
 - Preferential treatment needs to be provided to real-time applications.
 - A mutual contract is established between the nodes and the ISP.

03/01/10 383

(()) Why they are called as ingress is it. So, that is basically, because from there the packets originate the egress is from where the packet exit. (()) The other ones which are neither connected to the source nor to the destination are the core routers. Now, we will just short while see what are the roles of those routers now to support real-time communication preferential treatment needs to be provided to real-time application that is clear. It cannot just distinguish I mean treat all the messages the similar way and to do this to provide a preferential treatment to the real-time applications there has to be a contract established between the nodes and the ISP that the what is the required service quality based on that the services will be provided

(Refer Slide Time: 42:39)



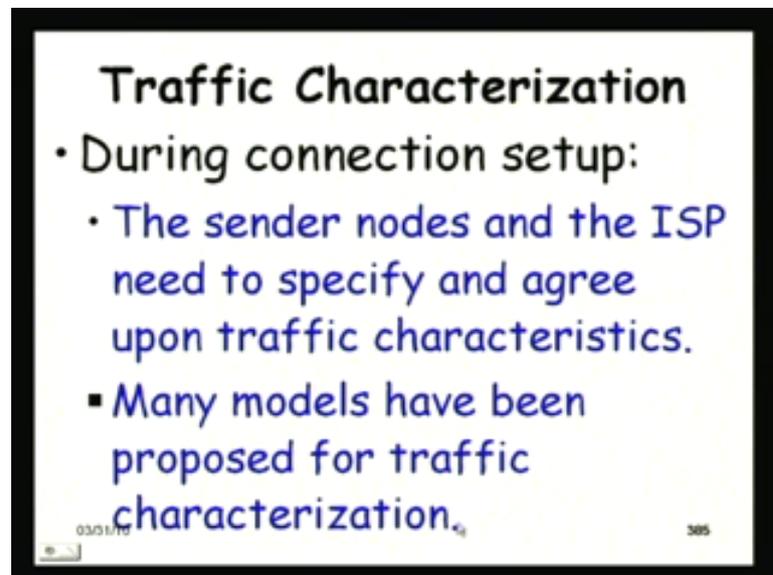
**Requirements for
Internet**

- While the ISP needs to ensure QoS guarantees:
 - The sender nodes need to meet the bounds on traffic.
 - The ISP should be capable to handle other misbehaving nodes.

384

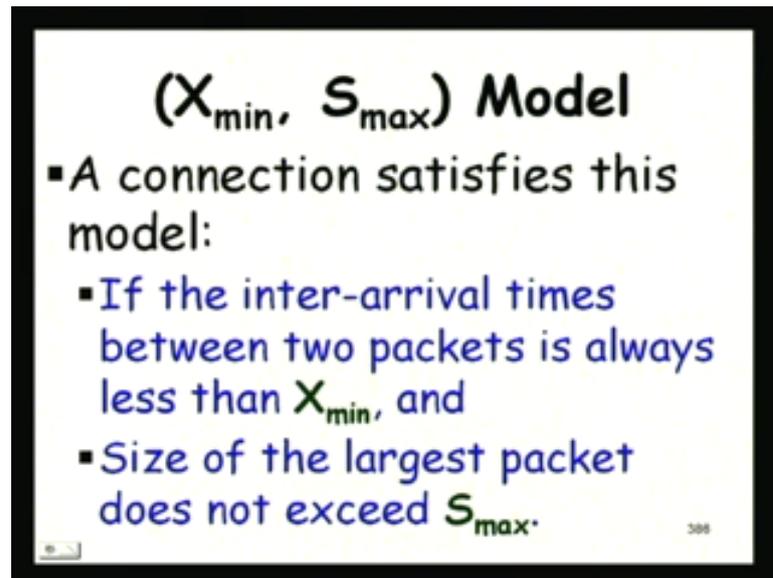
And the ISP once the contract it agrees to the contract then it has to try to meet the specified contract and the sender node on its part once the contract is signed it will try to not to exceed what it had really demanded in the contract. The contract is basically signed based on what is the kind of traffic that will be sent and then the Quality of Service parameters and. In this case if it exceeds the bound let us say a sender had requested for some service quality based for certain type of traffic .Now, if it misbehaves that is the sender is transmitting too much of data either or too burst to a data. Then the ISP would be able to handle the misbehaving node, because otherwise the other nodes will suffer. If a node does not transmitting node does not adhere to the contract. Then it should be able to penalize that possibly disconnect that.

(Refer Slide Time: 44:07)



Now, before we look further into the support in the internet the roles of the different types of routers. Let us just have a basic idea about how does one characterize the traffic, because based on this the contract will be signed between the ISP or the router and the sender. During the connection setup the quality of service that is required and the traffic characteristics will be agreed upon for this traffic characteristic the quality of service will be provided.

(Refer Slide Time: 44:55)



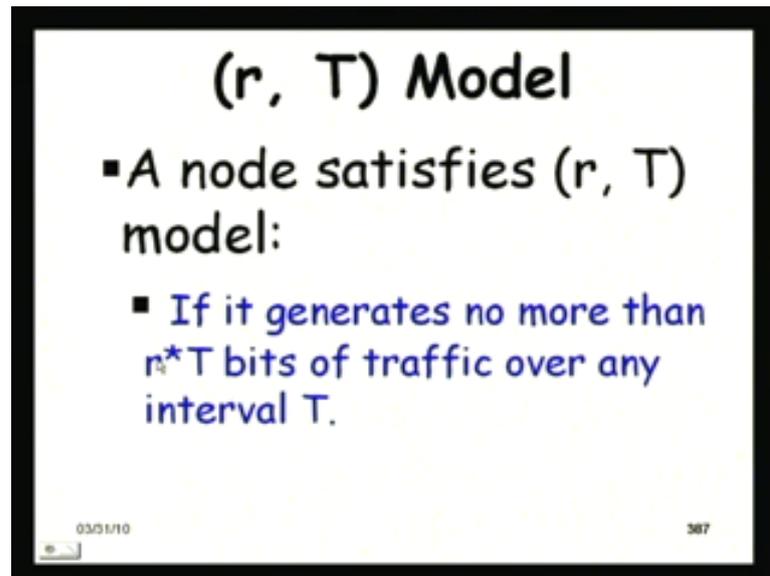
(X_{\min}, S_{\max}) Model

- A connection satisfies this model:
 - If the inter-arrival times between two packets is always less than X_{\min} , and
 - Size of the largest packet does not exceed S_{\max} .

Now, how does one characterize the traffic. There are several models of traffic characterization which are applicable to different situations or different types of traffic. Some models work well for certain type of traffic and some other model will work well for other types of traffic. Now, let us look at the simplest model the X min S max Model. Now, a traffic will satisfy this model, if the inter arrival time between two packets is always less than X min and the size of the packets is less than S max. See here we are just bounding it by a straight line, if we look at the packet arrival and the size. The data that is getting generated per unit time is less than S max in every X min time.

The data that is getting generated is less than or equal to S max every X min time or we are actually limiting the average data transmission. The X min S max model is actually limiting the average data transmission. So, at no time the average data transmission will exceed S max by X min, but what is the traffic is bursty. It might meet the X max by S min, but certain time there are spikes and the rest of the time no data. So, that will also meet their X min S max characterization, but it does not really model the bursty traffic well. It is a good model for constant traffic.

(Refer Slide Time: 46:58)



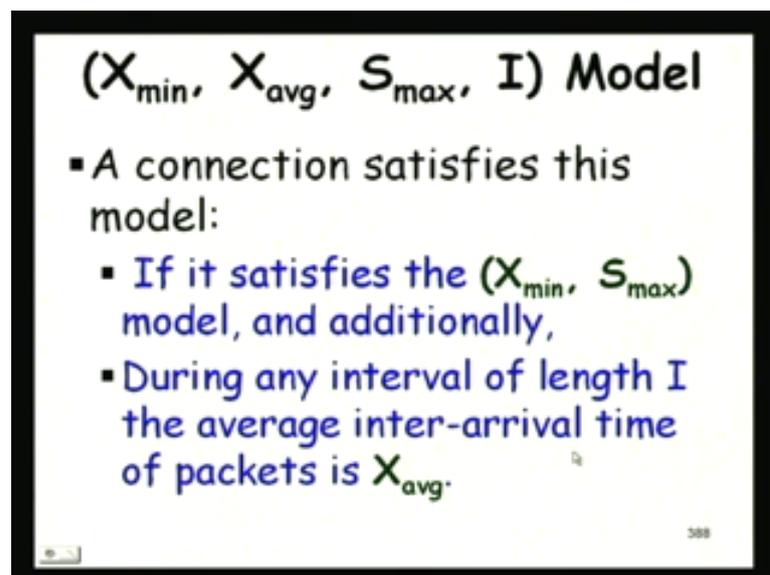
(r, T) Model

- A node satisfies (r, T) model:
 - If it generates no more than $r \cdot T$ bits of traffic over any interval T.

03/01/10 387

Now, let us look at the r comma T model. In the r comma T Model the r into T bits of traffic are generated over every interval T and r we can consider an average data that is being generated over a interval T very similar to X_{\max} S_{\max} X_{\min} r is the average data and r into T bit of traffic over any interval T , but this is a more restrictive form than S_{\min} X_{\max} model. Because for the interval T it does not generate more than r bits of data.

(Refer Slide Time: 47:25)



(X_{\min} , X_{avg} , S_{\max} , I) Model

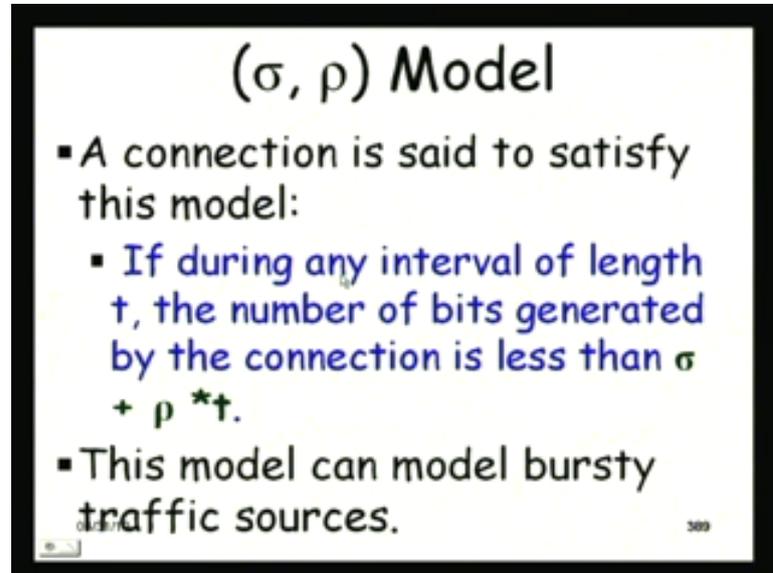
- A connection satisfies this model:
 - If it satisfies the (X_{\min} , S_{\max}) model, and additionally,
 - During any interval of length I the average inter-arrival time of packets is X_{avg} .

388

Now, let us look at the X_{\min} X_{avg} S_{\max} and I model. This can specify bursty traffic to some extent. So, here it should satisfy the X_{\min} S_{\max} that is the average data this is the bound on the average data X_{\min} S_{\max} . So, it model the X_{\min} S_{\max} and in

additional during any interval of length I the average inter-arrival of packet is X avg. So, then it limits the burst size here see here. So, for burst traffic this may be a more appropriate model not only models the average data, but also it bounds the burst size.

(Refer Slide Time: 48:49)



(σ, ρ) Model

- A connection is said to satisfy this model:
 - If during any interval of length t , the number of bits generated by the connection is less than $\sigma + \rho * t$.
- This model can model bursty traffic sources.

309

We can have the sigma comma rho model here a connection would satisfy this model if during any interval of length t , the number of bits generated by the connection is less than sigma plus rho t . So, just see here any interval of length t . So, once we have sigma and rho specified even if you take t it will be very small it should be within sigma plus rho t . You can consider rho to be the average data and sigma is the size of the maximum burst that can occur at any instantaneously. Sigma is the instantaneous burst size and rho is the average see here the two components here. The sigma is the max it is restricting the maximum instantaneous burst size and rho is the average traffic. So, this is a good model for bursty traffic to characterize bursty traffic.

(Refer Slide Time: 50:00)

Multiple Rate Bounding

- Bursty traffic sources can be characterized:
 - By bounding the traffic over multiple averaging intervals.
- A traffic would satisfy:
 - $\{(r_1, T_1), (r_2, T_2), \dots\}$ if $T_1 < T_2 < T_3, \dots$,
 - Over any interval I the number of bits generated is bounded by $r_i * T_i$, if $T_{i-1} < I < T_i$.

We can even use a multiple rate bounding. So, this is an extension of the r comma T model. So, the r comma T model was actually bounding the rate over a specific interval t . So, we are saying that if you look at the interval T the traffic will never exceed r 1 any interval T if you look, but what about if the interval we are considering is less than T will it generate bursty traffic in that interval and the rest of the time it will generate nothing. This if you just look r comma T it does not exclude that possibility. Just a single r comma T it guarantees that within the every T 1 the traffic is not bursty .It will be less than r 1, but within T 1 it can be bursty.

So, we can use multiple bounding intervals not only that over a large interval it is not bursty. But even for smaller intervals it is not Bursty, but can we just use just very small interval r 1 let us say T 1 is the smallest here among this can we just use r 1 comma T 1 and leave out the others will that serve the purpose or do we need all these intervals. See these are multiple bounding intervals .Over T 2 it generates certain amount of the data over T 1 it generates it is guaranteed to generate certain amount of data. The question that I am asking is that if we just restrict the amount of data over a small time T the smallest T here will that serve the purpose do we are these all redundant. What do you think?(())

These are not redundant see this is telling about the burst size here the smallest interval what is the maximum data that can occur, but over a longer and longer period how much data will be transmitted the maximum. So, this give more indication of the average data that will be transmitted over larger and larger periods. So, we need both. So, here this is a

more accurate characterization of a bursty traffic where not only we say .So the we restrict the instantaneous burst size, but also over a any given interval we characterize what will be the behavior of the traffic source. Possibly this is the most accurate characterization of a bursty traffic if we use multiple bounding intervals and then bound the traffic for different time press.

So, we just looked at we are trying to discuss how real-time communication can be established in a packet switched environment and then we saw that there has to be a contract between the router and the sender the router will guarantee the network will guarantee that certain quality of service will be met provided that the sender restricts itself to a signed or to a agreed traffic characterization suddenly start sending too bursty traffic even though it is on the average is the same then other sources might be missing their deadlines. So, from this point onwards we will build up and we will discuss how real-time messages can meet their deadline in a internet situation we will stop here.

Thank you...