Chapter-10

Real Time Communication

Introduction:

Real-time communication is a means of sharing information and interacting with people through network connections just as if they were face-to-face. Digging into the technicalities a little deeper, it’s any live (real-time) telecommunications that doesn’t have transmission delays – it’s usually a peer-to-peer connection with minimal latency, and data from a real-time communications application is sent in a direct path between the source and destination.

Examples of Real-Time Communications

There’s a difference between emailing and chatting with someone. Email is more of a time shifting form of communication – we send emails and expect to hear back from people later, and data is stored between the source and destination. Communicating through methods like email place more emphasis on delivering information reliably, not how quickly the information gets there. When chatting with someone, however, we expect responses just as if we were communicating face-to-face: in real-time. Other examples beyond instant messaging of real-time communications include:

- Video conferencing
- Presence (usually found in UC applications)
- Gaming
- File sharing
- Screen sharing
- Collaboration tools
- Machine to machine technology
- Location tracking
- Online education
- Social networking

Real-time communications applications and solutions can be used in virtually every industry: contact centers, financial services, legal firms, healthcare, education and retail can all benefit and improve processes with real-time communications applications. There are a few trends in play that are helping drive the growth of real-time communications applications.

Model of Real Time Communication:

In the model of the real time communication, end users of the message application systems as source and destination residing in different host. The network interface of each host contains input queue and output queue. Two buffer area called as input / output buffer are allocated to input and output queue to store queuing information. The queue are jointly maintained by two local servers as Transport Protocol Handler (TPH) and Network Access Control Handler (NACH). The former interfaces with local application and provides them with message transport service. The next interface with the network below and provides network access and message transmission services to the TPH. The client- server architecture produce more delay such that it is not suitable communication network architecture for real time communication.

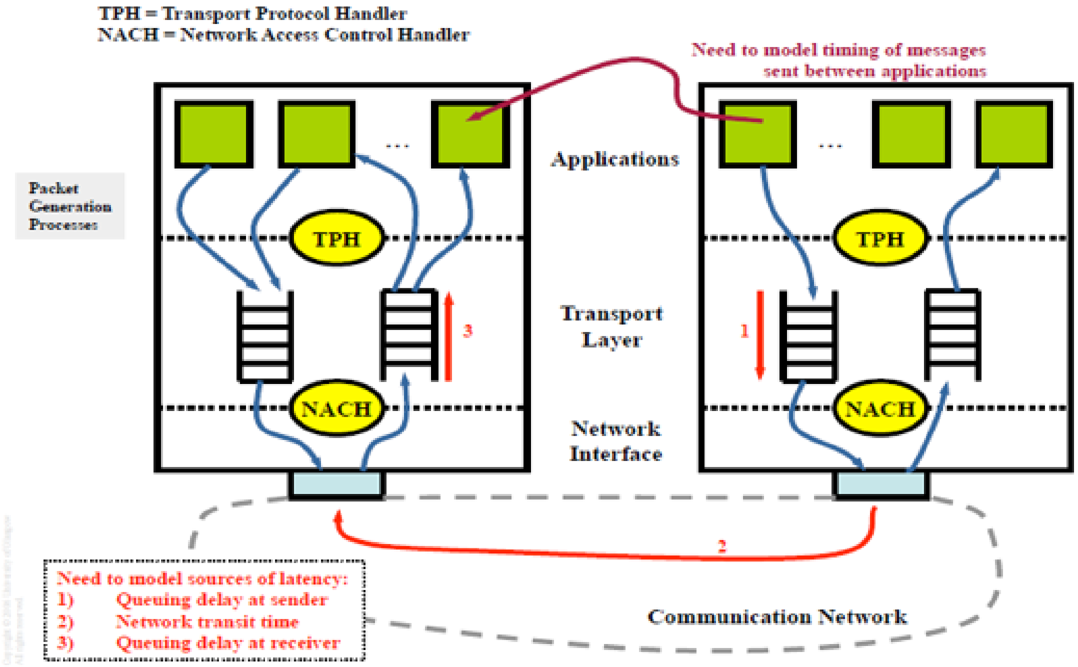

TPH = Transport Protocol Handler
NACH = Network Access Control Handler
Need to model timing of messages sent between applications
Packet Generation Processes
...
Applications
TPH
TPH
3
Transport Layer
1
NACH
NACH
Network Interface
2
Need to model sources of latency:
1) Queuing delay at sender
2) Network transit time
3) Queuing delay at receiver
Communication Network

Fig: a real time communication model

The fig. above shows the data path follow to transfer the message in between two hosts. Two marked circle TPH and NACH are transport protocol handler and network access control handler. When requested to send message by a local application task, the source TPH places the massage in output queue. Each outgoing message is delivered to network under the control of source NACH. After the placement of message on network, the NACH of destination host place it on input queue of destination and notify the destination IPH. The destination IPH then moves the message to address space of the destination application task and notify the destination task of arrival of message.

Here the end to end model by chain of jobs is used to represent the message sending activity. The application task available on source and destination are modeled as the predecessor as well as successor of the chain. At the beginning and end of the chain are source and destination chain are the transport protocol processing job. In between them, each job that access the network or transmits the message becomes ready for execution after its predecessor completes. Ideally the network delivers messages to receiver with no delay. In reality there is some of the delay are associated with queuing delay at sender, network transmit time and queuing delay at receiver as well as network.

- Network is not always ready to accept a packet when it becomes available and data may be queued if produced faster than the network can deliver it such that Queuing delay arise at sender.
- Application task are not always ready to accept packets arriving from network and Network may deliver data in bursts form such that Queuing delay occurs at receiver.
- Due to cross-traffic or bottleneck links such that Queuing delay occurs in the network
- Network transit time also generates the delay.

Hence a synchronization protocol is needed to model the real time communication model.

Real Time Traffic Model:

The real time traffic means isochronous or synchronous traffic, consisting stream of message that are generated by their sources and delivered to their respective destination on continuous basis. The traffic includes the periodic, aperiodic and sporadic messages. The periodic and sporadic message are synchronous in nature and there requirement of guarantee of on time delivery whereas aperiodic message are asynchronous in nature and shows the soft timing constraint.

In real time traffic model, each message (M_i) be characterized by tuples of inter-packet spacing (Pi), message length (e_i), reception deadline (Di) as below.

$$M_i = (p_i , e_i , D_i)$$

This traffic model is called peak rate model in real time communication.

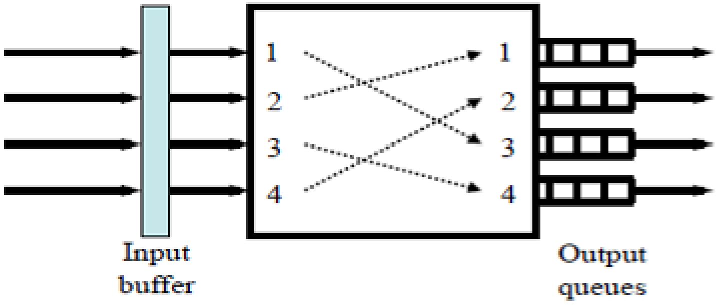
Throughput, Delay and Jitter:

- The throughput is the measure of the numbers of packets or message stream that the network can deliver per unit time.
- The delay (latency) is time taken to deliver the packet or message stream. It is fixed due to minimum propagation delay due to speed of light and varying due to queuing on path.
- The term jitter indicates variance on delay.
- Many real time communication protocol and algorithms are designed to keep not only the worst case delay and jitter as small.

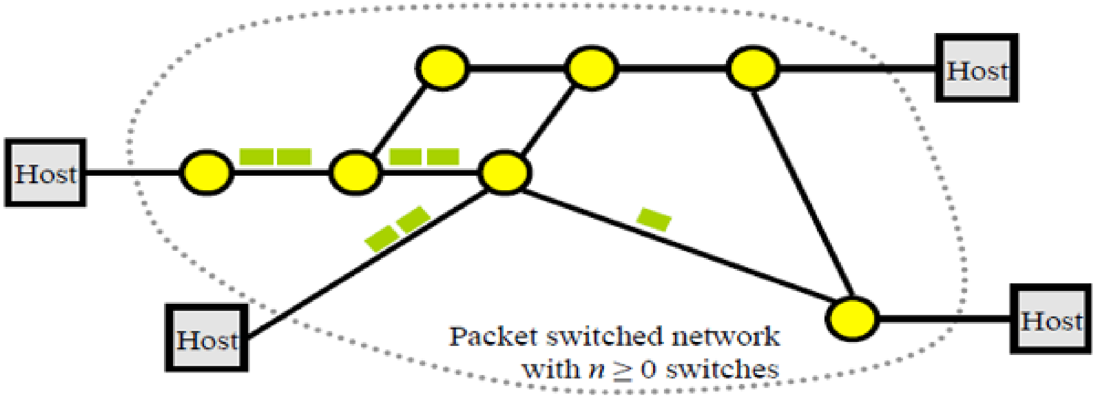
Real Time Connections and Service Disciplines:

The connection oriented approach is used for real time traffic. According to this approach, a simplex logical connection is set up for transmission of each message stream between source and destination. The fixed routing is used to route the packets and chosen route is fixed until torn down of system or invocation of adaptation mechanism. The new connection is established if the existing network meets quality of service parameter like end to end delay, jitter etc. Generally packet switching network is preferred for transmission of message streams.

- Fig below shows the packet switching network along with multi hop switch for multi hop network and circle of the network represents a switch.



- Switches are buffered I.e. there is a buffer pool for each output links, holding the packets that are queued for transmission on the link. Once the switch route the packets to the queue , the packets waits in the queue until schedulaer schedules it for transmission and then it is transmitted to next hop at the other end of the output link.
- The amount of time the switch takes to route the packets is small but the time a packets takes passing trough a swith is equal to its output time at output queue plus packet transmission time and this is called as hop delay o the packet.



- The end to end delay of each packet trough a switched network is equal to the sum of the per hope delays it suffers passing through all switches in the route plus total time it takes to propagate along the all links between the switches.

- The combination of acceptance test and admission control protocol , a synchronization protocol and a scheduling algorithms used for the purpose of rate control (jitter) and scheduling of packets transmission is called a service discipline
- Service disciplines are divided into two categories as rate allocating and rate controlled.
- Rate allocating disciplines allows the packets on each connection to be transmitted at higher rates than guaranteed rate.
- A service disciplines is said to rate controlled if it each connection guaranteed rate but never allows packet to send above guaranteed rate.

Priority – Based Service Disciplines For Switched Network:

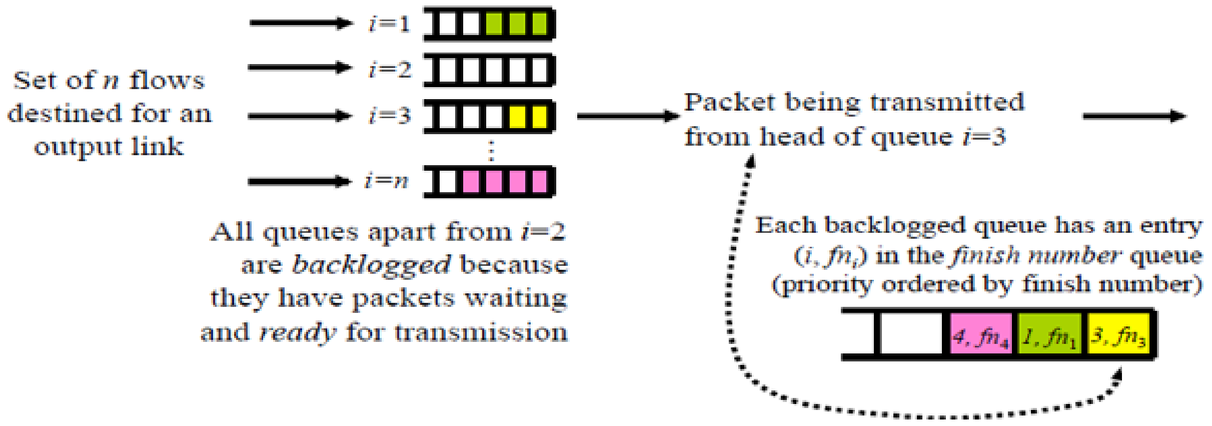
According to a priority based service disciplines, the transmission of ready packets are scheduled in priority driven manner. Waited fair queuing (WFQ) and Waited round robin scheduling are common approach for scheduling the packets in real time communication network.

Waited Fair Queuing (WFQ) Discipline:

It is a rate allocating service discipline and provides each flow with at least its proportional fair share link capacity and isolates the timing between flows such that it is also called as packet by packet generalized processor sharing algorithms. It is define as a packet switch has several inputs, feeding to an output link shared by n established flows where each flows i is allocated to u_i of the link such that total bandwidth allocated to all connections is:

$$U = \sum_{i=1}^{n} ui \quad \text{Where } U<=1.$$

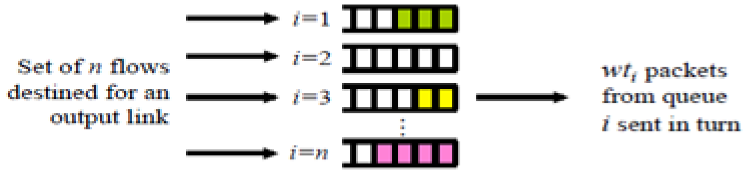
In WFQ, the output buffer comprise of two queues as FIFO and shortest finish number (SFN) for scheduling of packets. FIFO is used to scheduled n flows whereas priority ordered SFN is used to schedule the job obtained from the head FIFO queue on the basis of finish number. The finish number specifies the number of ready packets for transmission. The structure of queue is shown below.


Set of *n* flows destined for an output link
i=1
i=2
i=3
i=*n*
Packet being transmitted from head of queue *i*=3
All queues apart from *i*=2 are *backlogged* because they have packets waiting and *ready* for transmission
Each backlogged queue has an entry (*i*, *fn*ᵢ) in the *finish number* queue (priority ordered by finish number)
4, fn₄
1, fn₁
3, fn₃

- A packet becomes ready on FIFO queue, a finish number is calculated and SFN queue updated. The newly arrived packet is placed on the head of SFN queue without preempting the current transmission. When packet completes its transmission then it is removed from head of SFN and FIFO queue.

Waited Round Robin Scheduling:

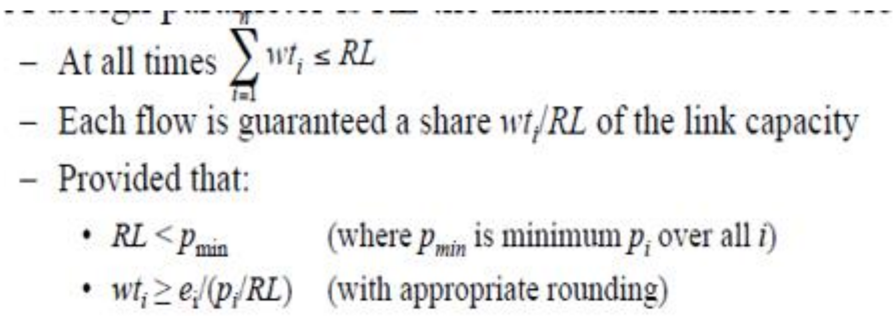
In this approach jobs are placed in FIFO queue. The job at the head of queue executes for one time slice. If it doesnot complete with in the time slice, it is preempted and put back into the queue. There are n jobs in the queue, each job gets one slice every n time slots in a round. The weighted round robin schedule extends this to give each job with weight Wi such that ith job gets Wi lenth time slice for execution. In this method each packets obtained from set of n flows are arranged in FIFO order and gets the connection for Wi time after connection and preempted and backlogged to the end of the queue if it is not completely transmitted otherwise removed from the queue. This service disciplines is shown below.


Set of n flows destined for an output link
i=1
i=2
i=3
i=n
wt_i packets from queue i sent in turn

When in each round, if more than Wi packets are backlogged on queue i then Wi packets are transmitted such that

- Each flow is guranted Wi slots in each round
- Rate allocating may send more , if nothing to transmit

Then such WRR scheduling scheme is called Greedy WRR scheduling only when there must be a design parameter (RL) satisfies the following conditions.



- At all times $\sum_{i=1}^{n} wt_i \le RL$
- Each flow is guaranteed a share wt_i/RL of the link capacity
- Provided that:
  - $RL < p_{\min}$ (where p_{min} is minimum p_i over all i)
  - $wt_i \ge e_i/(p_i/RL)$ (with appropriate rounding)

Medium Access Control Protocols of Broadcast Networks:

- The transmission medium of the broadcast network is the processor.

- A MAC protocol is a discipline for scheduling this type of processor.
- Scheduling of the transmission medium is done distributedly by network interfaces of hosts in the network.

Medium Access Protocol in CAN (Controller Area Network)

- Controller area network are very small network. CANs are used to connect components of embedded controllers. An example is an automotive control system whose components control the engine, brake, and other parts of an automobile.
- The end to end length of CAN must not exceed 100 meters. This means that within the fraction of a bit time after a station starts to transmit, all stations on the network can hear the transmission.
- The output of all stations are wire-ANDed together by the bus i.e. the bit on the network during bit time is a logical 0 if the output of any station is 0 and logical 1 when the output of all stations is 1.
- CAN MAC protocol is similar to the CSMA/CD (carrier sense Multiple Access/ Collision Detection)
- A station with a packet to send waits until it hears that the network is idle and then commences to transmit the ID number of the packet. At the same time, the station listens.
- Whenever it hears a 0 on the network while it is transmitting 1, it interrupts its own transmission.
- Network connection is resolved in favor of the packet with the smallest ID among all contending packets.

MAC in IEEE 802.5 Token Ring:

In a token ring network, packets are transmitted in one direction along a circular transmission medium. A station transmits a packet by placing its packet on the output link to the network. As the packet circulates around the network, the stations identified by the destination address in the header copies the packet. When the packet returns to the source station, the station removes the packet.

Prioritized Access in IEEE 802.5 Token Ring:

P**olling**

Network contention is resolved by a polling mechanism called token passing. For the purpose of polling, each packet has in its header an 8-bit Access Control (AC) field. One of the bits in an AC field is called the token bit. By examining this bit in the current packet on the network, a station can determine whether the network is busy. If the network is free the packet is polling packet. As a polling packet circulates around the ring, the stations are polled in a round robin manner in order of physical locations on the ring. When a free token reaches a station that has outgoing packets waiting, it can seize packets if it has the highest priority at that time.

Priority Scheduling:

Prioritized access is made possible by using the two groups of 3 bits each in the AC field: Their values represent the token priority πT and the reservation priority πR. Token priority bits give the priority of the token. A station can seize the free token only when its outgoing packet has an equal or higher priority than the token priority πT.

Reservation bits in the outgoing packets is used to make reservation for future use of the network. When the station seizes the token, it leaves the token priority unchanged but sets the reservation priority to the lowest priority of the network. It then marks the token busy and puts the token in the header of the packet and transmits the packet.

When a source station removes its own packet from the network, it saves the reservation priority carried in the packet. Suppose that when the source station transmit a free token, it sets the token priority of the token to this reservation priority or the highest of its outgoing packets, whichever is higher. In this case the priority arbitration mechanism allows the stations to jointly carry out any fixed scheduling algorithm.

Schedulability Analysis:

The amount of time (execution time) each packet occupies the network is equal to its transmission time plus the round trip delay it takes to return to the source station. The delay is usually on the order of 10-2 or less of the packet transmission time. In addition to this following three factors should be taken account for:

- Context switching: A context switch time is equal to the amount of time required to transmit a free token, plus the round trip delay of the network, which is an upper bound of the time the token takes whose outgoing packets has the highest priority among all outgoing packets during the transmission of the latest data packet.
- Blocking: Since packets are transmitted non-preemptively, we also need to take account the blocking time due to nonpreemptivity. Moreover, a higher priority packet that arrives at the station just after the header of the current data packet passed the station need to wait for a lower priority packet.
- Limited Priority Levels: since the network provides only eight priority levels resulting in schedulability loss.

Internet and Resource Reservation Protocols (see on books)

Issues in Resource Reservation

- Multipoint to Multipoint Communication
- Heterogeneity of Destinations
- Dynamic multicast group membership
- Relation to routing and admission control

Requirements for Multimedia Traffic:

- In order to ensure playback timing and jitter removal timestamps are required.
- In order to ensure presence and order of data a sequence number is required.
- Most of real-time multimedia applications are video conferencing where several clients receive data therefore multicast mode is preferred.
- In order to deal with congestion mechanism for sender notification and change of encoding parameter must be provided.
- In order to display streams generated by different standards the choice of encoding must be provided.
- In order to display audio and videos within a single A/V session mixer are required.
- In order to use high bit rate streams over a low bandwidth network, translator are required.

Why real time data cannot use TCP?

- TCP force the receiver application to wait for transmission in case of packet loss which causes large delay.
- TCP cannot support Multicast.
- TCP congestion mechanism decreases the congestion window when packet loss are detected (slow started). Audio and videos on the other hand have natural rates that cannot be suddenly decreases.
- TCP headers are larger than UDP header (40 bytes for TCP compared to 8 bytes for UDP).
- TCP doesn't contain necessary timestamp and encoding information needed by receiving application.
- TCP doesn't allow packet loss. In A/V however loss of 1-20% is tolerable.

Protocols:

There are several related protocols which support real time traffic over the internet. Some of the important protocols are

RTP (Real Time Protocol): used for real time data transport developed by extending UDP and sits between UDP and application.

RTCP (Real Time Control Protocol): used to exchange the control information between sender and receiver and works conjunction with RTP.

SIP (Session Initiation Protocol): provides the mechanism for establishing calls over IP.

RTSP (Real Time Streaming Protocol): allows user to control display as rewind, pause, forward etc.

RTP (Real Time Protocol):

There was a dilemma weather to implement RTP as a sub layer of transport layer or as a part of application layer. At this point it is common that RTP is implemented as application library, which executes in user space rather in kernel space like all protocol layers bellow RTP. RTP doesn't ensure the real time delivery itself but it provides the means for

- Jitter elimination / reduction
- Synchronization of several audio or video streams that belong to same multimedia session.
- Multiplexing of audio/video streams that belong to different session.
- Translation of audio/video streams from one encoding type to another.
- With the help of RTCP, RTP also provides hooks for adding reliability and flow/congestion control which is implemented within multimedia application. This property sometimes called as application level framing.
- RTP is a protocol that provides the basic transport layer for real time application but doesn't provide any mechanism for error and flow control , congeston control, quality feedback and synchronization. For that purpose RTCP is added as a companion to RTP to provide end to end monitoring, data delivery and QOS.

RTCP (Real Time Control Protocol):

It is responsible for three functions.

- Feedback on performance of application and network.
- Correlation and synchronization of different media streams generated by same sender for example combined audio/video.
- The way to convey the identify sender for display on a user interface.

The volume of RTCP traffic may exceeds the RTP traffic during a conference session involving larger number of participates. Normally only on participant talk at a time while other participants are listing. In mean while RTCO message are not periodically regardless if the participant is taking or not. Therefore the RTCP packet transmission are done dynamically on the basis of participants.

Standard dictates that 20% of the session bandwidth is allocated to RTCP. In other words RTCP RR SDES packets are sent every 5 RTP packets transmitted.

In addition, 5% of the RTCP bandwidth is allocated to particular participants (CNAME). The RTP transmission in the interval of participants is the function of total number of participants in the RTP session that ensures 5% of bandwidth allocation. For that purpose each participants has to continuously estimate the session size. The transmission interval is randomized to avoid synchronization effects.

- RTCP message are stackable.
- To amortize header overhead multiple RTCP message can be combined and send in a compound RTCP message.

A packet is loss if:

- Packet never arrived
- Packet arrived but corrupted.

- Packet arrived after its play out time.