WHITEPAPER

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

Low Latency Systems

February 2013
Ittiam Systems

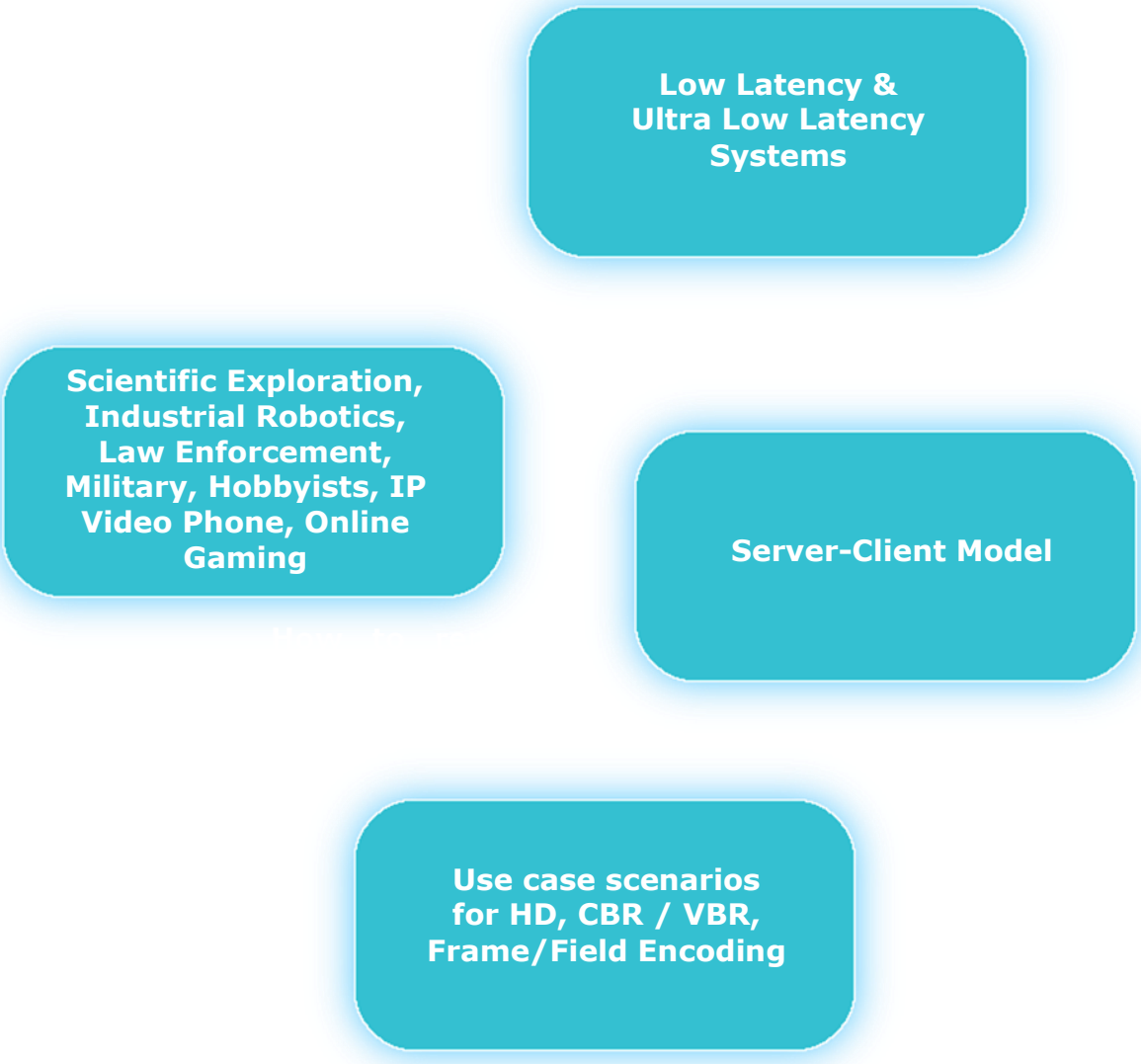
Contact us at mkt@ittiam.com

WHITEPAPER

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

Abstract

This paper discusses Low Latency systems with emphasis on understanding latency and the factors which affect it, in various use cases in a server-client configuration.



Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

Contents

Introduction 1
Target Markets and Applications 2
Generic Server-Client Model 3
Assumptions and Limitations 4
Factors which affect Latency 8
Use-case Scenarios 9
720p, 60 frames / sec, Frame encoding, CBR mode 9
720p, 60 frames / sec, Frame encoding, VBR mode 10
480p, 60 frames / sec, Frame encoding, VBR mode 10
480p, 30 frames / sec, Frame encoding, VBR mode 11
480i, 60 fields / sec, Field encoding, VBR mode 11
1080i, 60 fields / sec, Field encoding, VBR mode 12
Conclusion 13
Disclaimer 14

Figure 0-1 Generic Server-Client Model 3

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

Introduction

Low latency systems are desired in a variety of scenarios, where the low latency of the system allows the human involved to observe the events within human unnoticeable delays, and react to them quickly.

Human reaction time is about 190 milli-seconds for light stimuli, and about 160 milli-seconds for sound stimuli [HUMAN-REACTION-TIME]. Hence, about 150 milli-seconds is the expected range for interactive low latency. Some scenarios require very low latencies of 50 or 100 milli-seconds, which may be characterized as ultra-low latencies [ITTIAM-LOW-LATENCY].

This paper provides an overview of the target markets and applications, proposes a simple model for studying latency, and considers the different aspects which affect latency. In the end, we characterize the minimum achievable latency on two embedded platforms.

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

Target Markets and Applications

Table 0-1 highlights different target markets and applications where low latency is desired.

Target Markets	Applications	How Low Latency Helps
Scientific Exploration	Deep Ocean Exploration Space Exploration	Seamless real-time surveillance
Industrial Robotics	Construction Machines Working in Hazardous Environments Remote Controlled Robots	Faster reaction to the changing situation
Law Enforcement	Defusing / Detonating Explosives	Faster reaction to the changing situation
Military	Unmanned Ground Vehicles (UGV) Unmanned Aerial Vehicles (UAV)	Seamless real-time surveillance
Hobbyists	Remote Controlled Vehicles (Cars, Trucks, etc) Remote Controlled Toys	Helps quick driver responsiveness
IP Video Phone	Voice and Video over IP	Improves the conversation experience
Online Gaming	Online Games	Helps quick player responsiveness

Table 0-1 **Target Market and Applications**

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

Generic Server-Client Model

Figure 0-1 depicts a generic Server-Client model for studying the latency of the system.

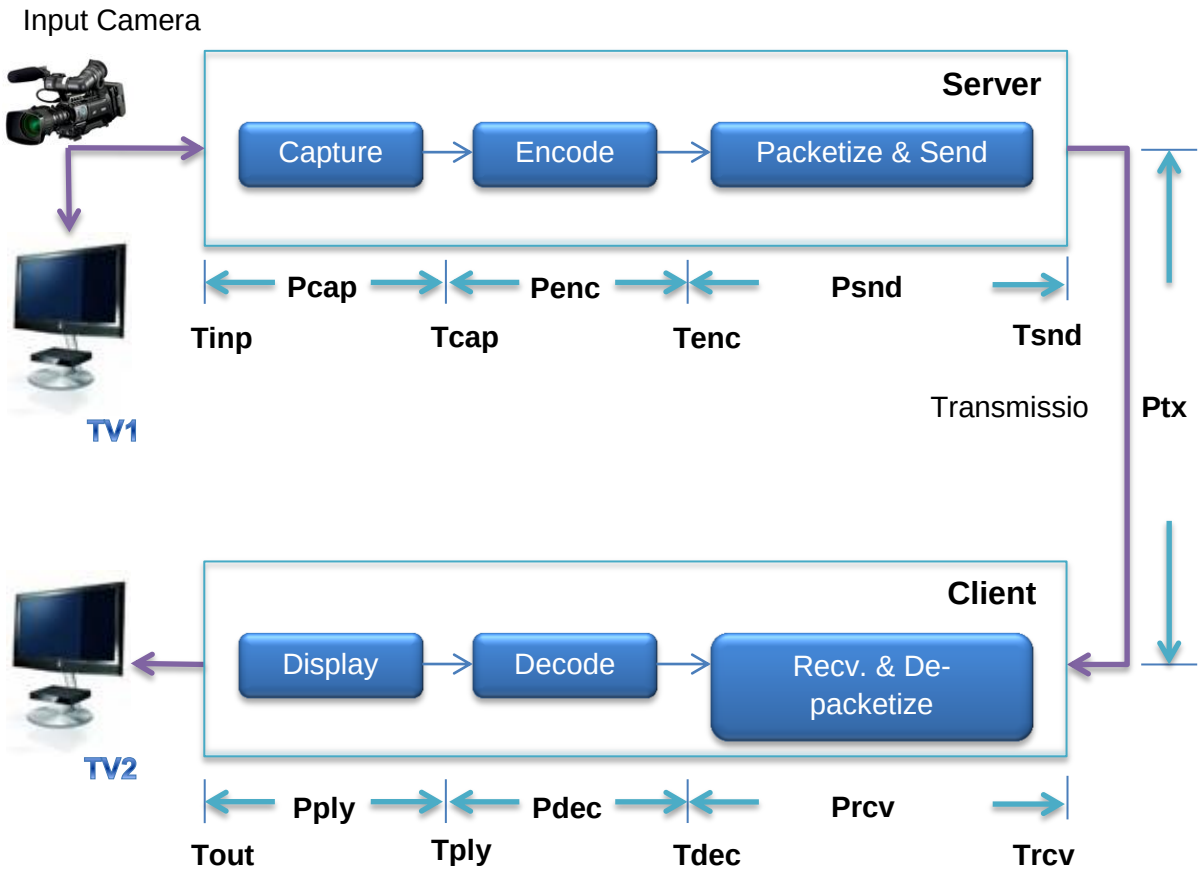


Figure 0-1 Generic Server-Client Model

The AV Server is connected to an audio-video input source (eg: camera, DVD player).

The AV Client is connected to an audio-video output device (eg: TV).

The AV Server captures the audio-video input from the input source, encodes it, and packetizes and sends the data over a network to the AV Client. The AV Client receives the data and de-packetizes it, decodes the coded audio-video, and displays the audio-video output on the output device.

The Server system consists of the following modules - capture, encode, packetize and send. The ‘capture’ module captures the analog picture over the input video port, digitizes it and copies it to memory. The ‘encode’ module accesses the captured picture from memory, encodes it and provides the encoded picture in memory. The ‘packetize and send’ module packetizes the encoded picture and sends the packetized data over network.

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

The Client system consists of the following modules - receive and de-packetize, decode, display. The 'receive and de-packetize' module receives the data over the network, de-packetizes it and provides the de-packetized data in memory. The 'decode' module accesses the encoded picture from memory, decodes it and provides the decoded picture in memory. The 'display' module copies the decoded picture from memory, converts it to analog signals and displays the picture (frame / field) on the output video port of the device.

Assumptions and Limitations

- As audio processing takes lesser time than video processing, we only consider the impact of video processing on the latency.
- For the sake of simplicity, we intentionally leave out any pre- / post-processing modules like scaler, deinterlacer, etc. in our generic system model.

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

Timing Notations

In the generic Server-Client model shown in Figure 0-1, we define the following timing notations for time instant `T` –

Time Instant Notation	Abbreviation For	Description
`Tinp`	Time instant - input	Time instant when the first pixel of the picture (frame / field) is available on the input video port of the device
`Tcap`	Time instant - capture	Time instant when the picture is fully available in digitized form (eg: YUV, RGB, etc) in memory
`Tenc`	Time instant - encode	Time instant when the encoded picture (frame / field) is available in memory
`Tsnd`	Time instant - send	Time instant when the last byte of the encoded picture is packetized and sent over the network
`Trcv`	Time instant - receive	Time instant when the last byte of the packetized encoded picture is received from the network
`Tdec`	Time instant - decode	Time instant when the encoded picture is fully de-packetized and ready to be decoded
`Tply`	Time instant - playout	Time instant when the picture is fully decoded and available in memory to be displayed
`Tout`	Time instant - output	Time instant when the first pixel of the picture is available on the output video port of the device

Table 0-1 Timing Instant Notations

In addition, we define timing notations for time period `P` as

Time Period Notation	Abbreviation For	Equivalent Time Instant Notation
`Pcap`	Time period - capture	`Tcap - Tinp`
`Penc`	Time period - encode	`Tenc - Tcap`
`Psnd`	Time period - send	`Tsnd - Tenc`
`Ptx`	Time period - transmission	`Trcv - Tsnd`
`Prcv`	Time period - receive	`Tdec - Trcv`
`Pdec`	Time period - decode	`Tply - Tdec`
`Pply`	Time period - playout	`Tout - Tply`

Table 0-2 Time Period Notations

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

Hence, we can say that,

Latency of the Server,

L-Server = $Pcap + Penc + Psnd = Tsnd - Tinp$

Latency of the Transmission,

L-Transmission = $Ptx = Trcv - Tsnd$

Latency of the Client,

L-Client = $Prcv + Pdec + Pply = Tout - Trcv$

Therefore, the End-to-End or Total Latency of the system,

L-Total = L-Server + L-Transmission + L-Client

$= (Tsnd - Tinp) + (Trcv - Tsnd) + (Tout - Trcv) = Tout - Tinp$

For purposes of this study, we assume that the transmission time is zero, i.e.

L-Transmission = $Ptx = Trcv - Tsnd = 0$

This implies that, in our model, $Trcv = Tsnd$

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

Measuring Latency

Glass-to-glass latency measurement is a popular mechanism to measure latency across Server and Client.

A TV is connected to the video input source of the server – let us call this TV1. Another TV is connected to Client, which acts as output of Client. Call this TV2. The latency between TV1 and TV2 is glass-to-glass latency (See Generic Latency Model Diagram).

From the timing model, we know that the glass-to-glass latency is identical to the total latency of the system `(Tout – Tinp)`.

However, the perceived glass-to-glass latency (visually perceived latency between TV2 and TV1) will be in the range of

`(Tout – Tinp)` <= 'Perceived glass-to-glass latency' <= (`(Tout – Tinp)` + 1 picture duration)

This is because we may perceive the latency as the time difference between the display of the 'first' pixel of a frame on TV1 to the display of the 'last' pixel of the same frame on TV2.

Eg: If the glass-to-glass latency is 67 ms and one picture duration is 17 ms, then the perceived glass-to-glass latency will be in the range of (67 ms to 84 ms).

Note	All latency numbers mentioned henceforth shall be a range, corresponding to the 'Perceived glass-to-glass latency'.

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

Factors Affecting Latency

1. **Video Resolution** – 1080i, 720p, 480i, etc

   Higher the resolution, greater the complexity of encoding and decoding the video, resulting in higher time required for video encoding and decoding operations. This normally translates to higher latency.

2. **Encoding unit** – slice, field or frame

   Progressive picture needs to be processed as one whole frame at a time, but interlaced picture may be processed as one field at a time. While frame encoding is by far the simplest mode, field encoding allows the fields to be processed individually, thereby reducing the processing time for each field by a factor of 2 (compared to frame processing time). This allows the processing operations of one field to be done parallel to the other field's operations.

   To obtain the full benefits of field encoding mode, the capture and display drivers must support field mode of operation, and the encoder and decoder must support partial output / input mode respectively..

   Slice mode of encoding provides options of parallelizing the encode, transmit, receive, and decode operations for the slices of a picture, but requiring partial output / input mode to be supported by the video encoder / decoder respectively.

3. **Bitrate mode**[a] – Constant Bit-Rate (CBR) or Variable Bit-Rate (VBR)

   The bitrate mode being Constant Bit-Rate (CBR) or Variable Bit-Rate (VBR) has an impact on the time taken to send and receive one encoded picture. In VBR mode, the data can be sent in a burst mode, only limited by the bandwidth of the network. However, in CBR mode, the data has to be sent at the specified rate so as to maintain the average bitrate over the network in any period of time. Therefore, the latency in CBR case will be higher than the VBR case due to the transmission bit-rate constraint.

4. **Picture refresh rate** – 30 Hz or 60 Hz

   Given that a particular refresh rate is feasible to process, a higher picture refresh rate means reduced frame duration. This means reduced time to send & receive in cases of CBR transmission mode.

5. **Any pre- or post- processing**

   Any added pre- or post- processing to the generic model is an additional step in the processing, which incurs additional processing time and adds to the latency.

6. **System overheads**

   Software systems which handle the processing incur additional overheads in the processing (such as framework overheads) and Operating System overheads (due to thread scheduling latency).

[a] Note that this does not relate to the encoding mode of CBR / VBR. In fact, encoding in CBR mode and sending in VBR mode is expected to yield the least latency due to lesser variation in the encoded picture sizes in CBR mode.

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

Use-case Scenarios

This section describes a few popular use-case scenarios, and details the low latency achievable on two embedded platforms (based on TI DM6467 and TI DM816x processors), using Ittiam software middleware which is highly optimized to minimize the system overheads.

CASE-1: 720p, 60 frames / sec, Frame encoding, CBR mode

Latency (in milli-sec)	**TI DM6467**	**TI DM816x**
L-Server (Tsnd – Tinp)	50 ms (Pcap = 17[b], Penc[c] = 16, Psnd[d] = 17[e])	44 ms (Pcap = 19[f], Penc = 8[g], Psnd = 17)
L-Client (Tout - Trcv)	20 ms (Prcv = 4[h], Pdec = 16, Pply = 0)	15 ms (Prcv = 2, Pdec = 7, Pply = 6[i])
System Overheads	10 ms	10 ms
L-Total (Tout - Tinp)	80 ms	69 ms
Glass-to-glass Latency[j]	**80 – 97 ms**	**69 – 86 ms**

Table 0-1 **Latency for 720p, 60 frames / sec, Frame encoding, CBR mode**

[b] 'End-to-end latency' is defined as time taken for first pixel from input to be available as first pixel of output. It could as well have been defined in terms of last pixel of the picture. As 'end-to-end latency' is defined in terms of first pixel, the picture duration is accounted in the capture. Instead, if 'end-to-end latency' is defined in terms of last pixel, the picture duration can be accounted in the display

[c] Assuming H.264 video standard for encoding / decoding in all use-cases

[d] Assuming encoded bitrate of 10 Mbps for HD resolutions and 5 Mbps for SD resolutions (with a channel bandwidth of 100 Mbps), in all use-cases

[e] Assuming 'CBR send' takes same time as the frame duration in CBR mode

[f] On DM816x platform, capture driver is not interrupt based, but involves polling (with minimum polling interval of 1 ms). In addition, the captured frame requires about 1.3 ms to be available to ARM A8. This leads to an overall overhead of about 2.3 ms for capture, over and above the picture duration

[g] On DM816x platform, there are additional overheads in providing / receiving data to encoder / decoder. These overheads are subsumed in the 'System Overheads' category

[h] Prcv step includes the time taken for de-packetization and extraction of the decodable unit from the transmitted stream

[i] On DM816x platform, there is an additional processing step required for converting the decoded picture (from YUV 420 SP to YUV 422 ILE format conversion) before display. Also, the display of the picture takes about 1.3 ms after being scheduled for display from ARM A8

[j] This corresponds to the 'perceived glass-to-glass latency'. See Page 7 for detailed explanation

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

CASE-2: 720p, 60 frames / sec, Frame encoding, VBR mode

Latency (in milli-sec)	**TI DM6467**	**TI DM816x**
L-Server (Tsnd – Tinp)	37 ms (Pcap = 17, Penc = 16, Psnd = 4)	29 ms (Pcap = 19, Penc = 8, Psnd = 2)
L-Client (Tout - Trcv)	20 ms (Prcv = 4, Pdec = 16, Pply = 0)	15 ms (Prcv = 2, Pdec = 7, Pply = 6)
System Overheads	10 ms	10 ms
L-Total (Tout - Tinp)	67 ms	54 ms
Glass-to-glass Latency	**67 – 84 ms**	**54 – 71 ms**

Table 0-2 **Latency for 720p, 60 frames / sec, Frame encoding, VBR mode**

CASE-3: 480p, 60 frames / sec, Frame encoding, VBR mode

Latency (in milli-sec)	**TI DM6467**	**TI DM816x**
L-Server (Tsnd – Tinp)	27 ms (Pcap = 17, Penc = 8, Psnd = 2)	25 ms (Pcap = 19, Penc = 4, Psnd = 2)
L-Client (Tout - Trcv)	10 ms (Prcv = 2, Pdec = 8, Pply = 0)	8 ms (Prcv = 2, Pdec = 3, Pply = 3)
System Overheads	10 ms	10 ms
L-Total (Tout - Tinp)	47 ms	43 ms
Glass-to-glass Latency	**47 – 64 ms**	**43 - 60 ms**

Table 0-3 **Latency for 480p, 60 frames / sec, Frame encoding, VBR mode**

CASE-4: 480p, 30 frames / sec, Frame encoding, VBR mode

Latency (in milli-sec)	**TI DM6467**	**TI DM816x**
L-Server (Tsnd – Tinp)	44 ms (Pcap = 34, Penc = 8, Psnd = 2)	42 ms (Pcap = 36, Penc = 4, Psnd = 2)
L-Client (Tout - Trcv)	10 ms (Prcv = 2, Pdec = 8, Pply = 0)	8 ms (Prcv = 2, Pdec = 3, Pply = 3)
System Overheads	10 ms	10 ms
L-Total (Tout - Tinp)	64 ms	60 ms
Glass-to-glass Latency	**64 – 98 ms**	**60 - 94 ms**

Table 0-4 **Latency for 480p, 30 frames / sec, Frame encoding, VBR mode**

CASE-5: 480i, 60 fields / sec, Field encoding, VBR mode

Latency (in milli-sec)	**TI DM6467**	**TI DM816x**
L-Server (Tsnd – Tinp)	23 ms (Pcap = 17, Penc = 4, Psnd = 2)	23 ms (Pcap = 19, Penc = 2, Psnd = 2)
L-Client (Tout - Trcv)	6 ms (Prcv = 2, Pdec = 4, Pply = 0)	25 ms (Prcv = 2, Pdec = 2, Pply = 21[k])
System Overheads	10 ms	10 ms
L-Total (Tout - Tinp)	39 ms	58 ms
Glass-to-glass Latency	**39 – 56 ms**	**58 – 92 ms**

Table 0-5 **Latency for 480i, 60 fields / sec, Field encoding, VBR mode**

[k] DM816x display driver requires both fields to be provided for display

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

CASE-6: 1080i, 60 fields / sec, Field encoding, VBR mode

Latency (in milli-sec)	**TI DM6467**	**TI DM816x**
L-Server (Tsnd – Tinp)	36 ms (Pcap = 17, Penc = 17, Psnd = 2)	29 ms (Pcap = 19, Penc = 8, Psnd = 2)
L-Client (Tout - Trcv)	19 ms (Prcv = 2, Pdec = 17, Pply = 0)	40 ms (Prcv = 2, Pdec = 8, Pply = 30)
System Overheads	10 ms	10 ms
L-Total (Tout - Tinp)	65 ms	79 ms
Glass-to-glass Latency	**65 – 82 ms**	**79 - 113 ms**

Table 0-6 **Latency for 1080i, 60 fields / sec, Field encoding, VBR mode**

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

Conclusion

Low latency is a required condition for few critical application areas, which involve a human being responding to the conditions of some remote location, in real-time. The low latency helps different application scenarios in different ways, and is desired to be in the range of 150 milli-seconds in the normal case, and in the range of 50 to 100 milli-seconds in ultra-low latency cases.

Using timing terminologies in a simple server-client model, we have understood the glass-to-glass latency and the factors which affect it. Using two popular embedded platforms, we have studied the lowest latency feasible under different configurations.

It is clear that the lowest latency feasible in various configurations is different, due to the play of the various factors which affect latency. A detailed knowledge and understanding of the factors affecting latency is a pre-requisite to appreciate and understand the feasibility under different configurations.

In general, understanding the timing model and the factors affecting latency, offer a sound theoretical basis for understanding low latency systems. In particular, it helps in estimating the lowest possible latency achievable on any platform for any specific configuration.

Ittiam Systems Pvt Ltd
The Consulate, 1, Richmond Road
Bangalore - 560025, India

Disclaimer

This white paper is for informational purposes only. Ittiam makes no warranties, express, implied or statutory, as to the information in this document. The information contained in this document represents the current view of Ittiam Systems on the issues discussed as of the date of publication. It should not be interpreted to be a commitment on the part of Ittiam, and Ittiam cannot guarantee the accuracy of any information presented after the date of publication.