Latency in Embedded Systems

Sean Simmons
Director, DSP
RIM

December 16, 2009

Latency - Definitions

"*Time delay* between *input event* being applied to a system and the associated *output action* from the system"

- ***Input event* can be things like:**
  - Change in sound - someone tells you a joke.
  - Change in voltage - interrupt line activation.
  - Arrival of a message from another thread/process/computer.
- ***Output action* can be things like:**
  - Change in sound - you laugh at a joke.
  - Change in voltage - output pin changes polarity.
  - Sending of a message to another thread/process/computer.

Latency – When is it Important?

- **Typically when there is feedback:**
  - Conversation between two people. Conversations suffer when there is a large time delay between what you say and then hearing their response.
  - Client-server protocols. Either end may have expectations on how long a response should take
  - Control systems. Car breaking systems or home heating systems.
  - User interfaces. Time from key press to displaying character on screen.
- **Typically not important when there is no feedback:**
  - Broadcast. Of course there are limits – a latency of hours/days/years might be problematic.
  - Recording something for playback at some later time.

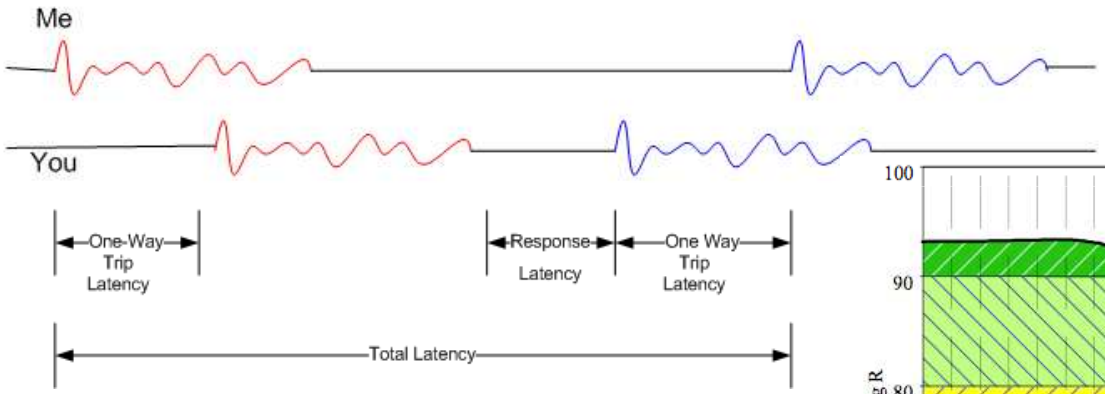
Latency – How Much?

- Tolerance to increasing latency is system dependent.
- Most systems have a point at which they cease to "function" as designed.
- "Point" of failure is very system dependent.
  Varies over many orders of magnitude
  – nanoseconds to hours.
- Failure mode (or degradation) is system dependent.
  - For most computer protocols degradation is catastrophic.
  - For control systems degradation usually results in instability.
  - For end-to-end-delay in speech tele-services, perceived degradation is gradual and "smooth".

Latency – Categorization

- Scheduling
  - Buffering to manage different clock domains (radio/packet clock vs sample clock)
  - Buffering for network queuing 'jitter' of packets
  - Buffering for task scheduling (sharing of processor resources)
  - Inherent in system design (TDM Radio Interface)
- Resource Limitations
  - Time to execute functions (processor clock rate),
  - Influence of memory hierarchies/system design
  - HW limitations (serial ports)
- Algorithmic
  - Delay in filters
  - Interleavers

Latency in Phone Calls



- One way delays >150 ms start to impact call quality.
- Delay can vary between calls.
- Delay can vary within calls.

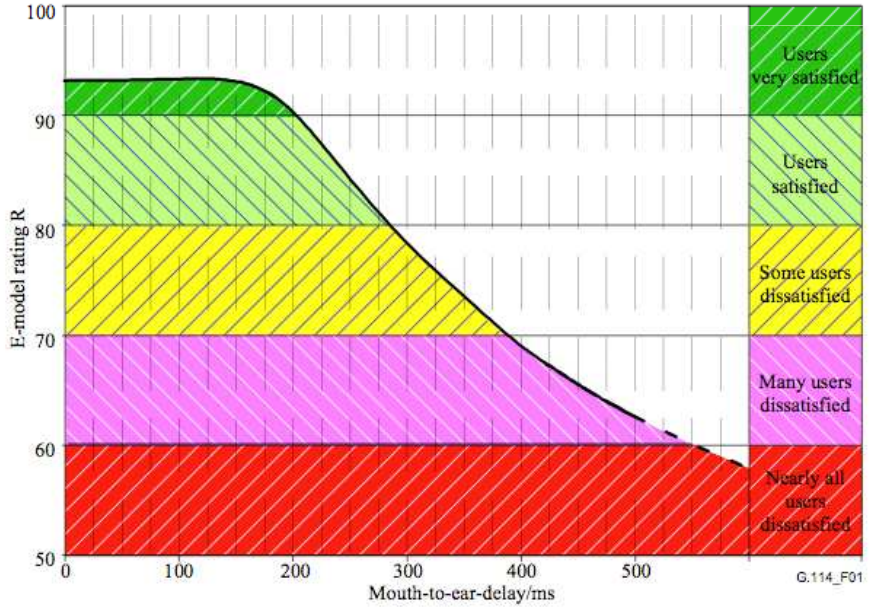
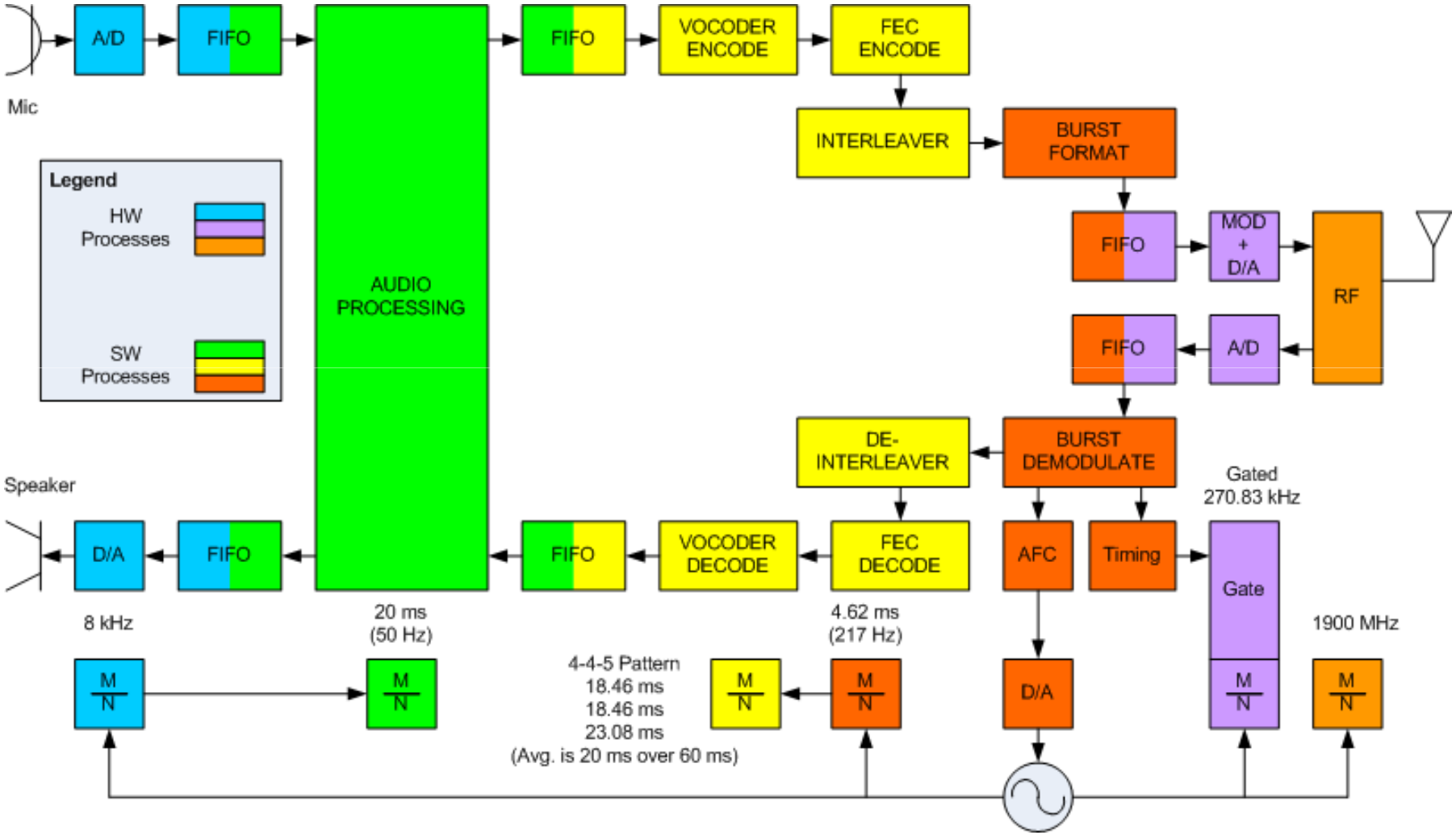


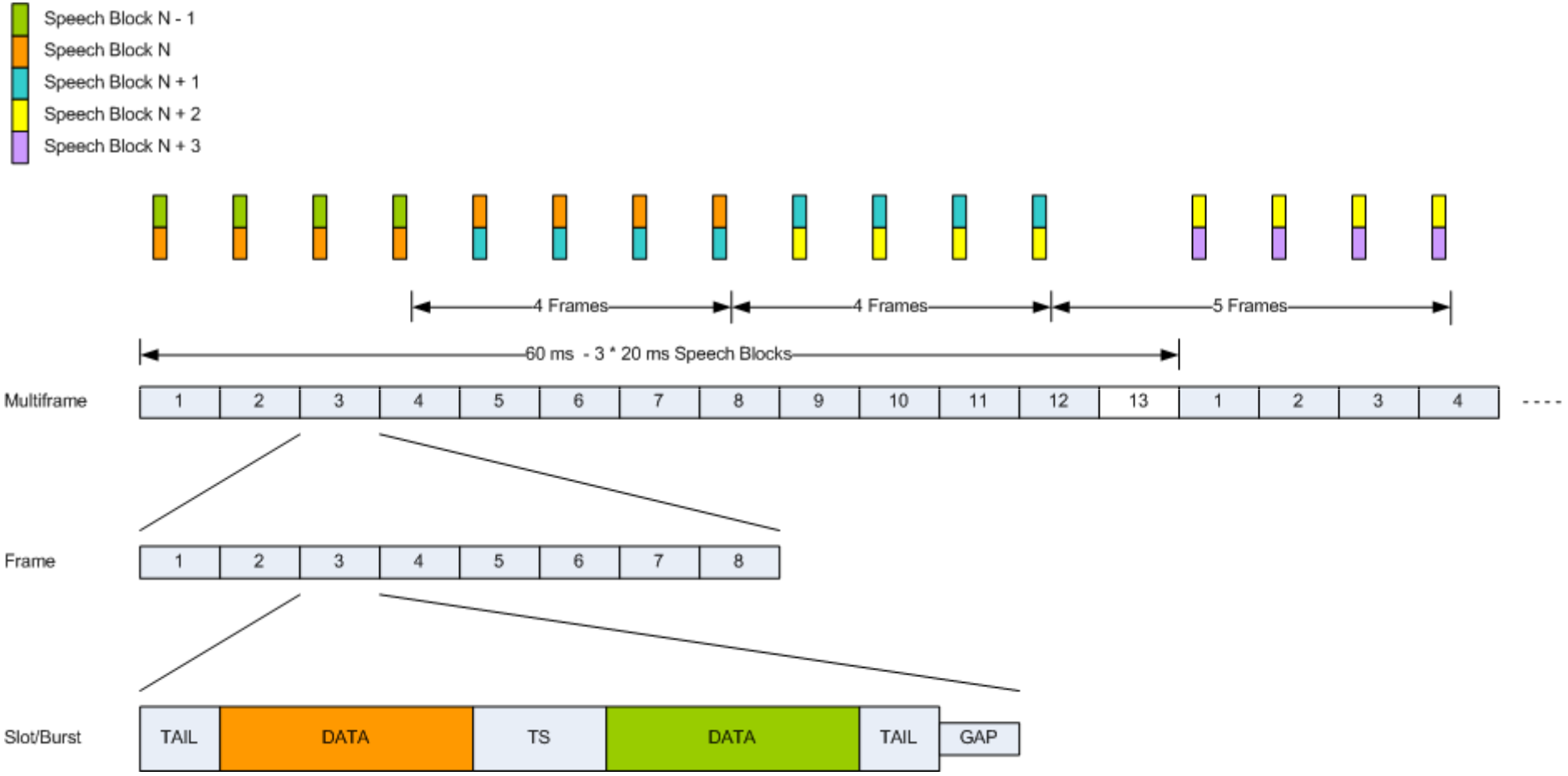
Figure 1/G.114 – Determination of the effects of absolute delay by the E-model

BlackBerry

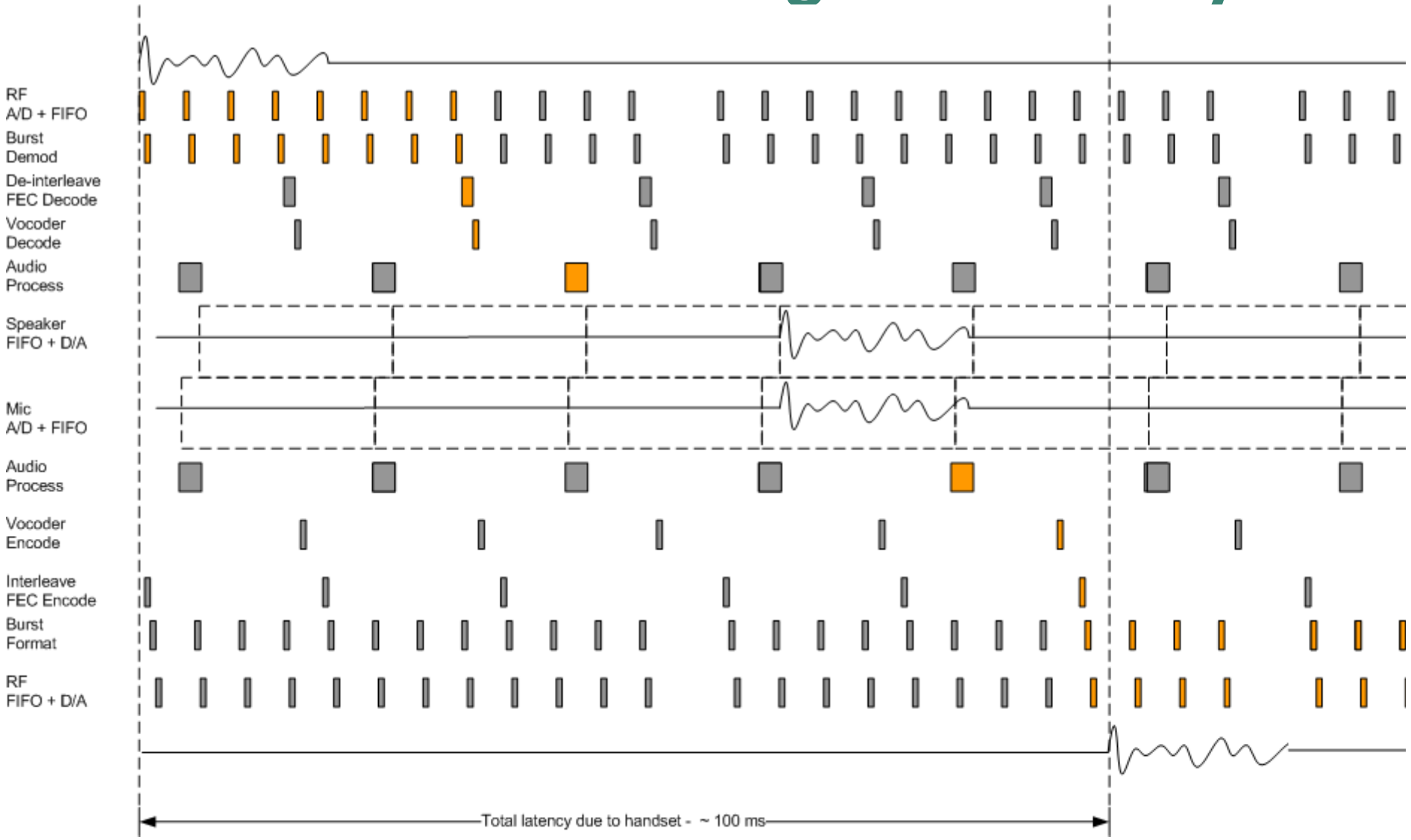
GSM Handset Block Diagram



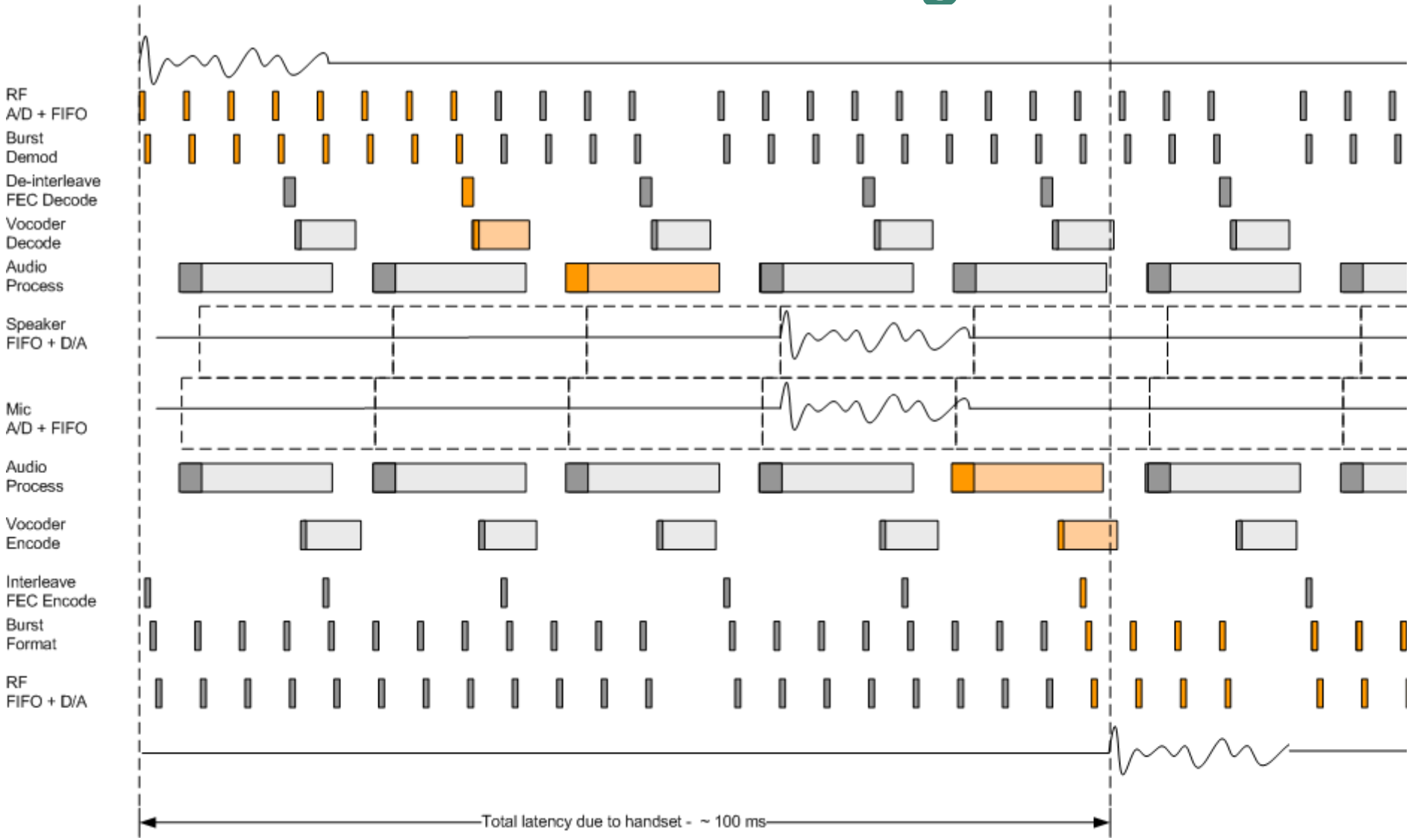
GSM Handset – Radio Timing



GSM Handset – Timing and Latency

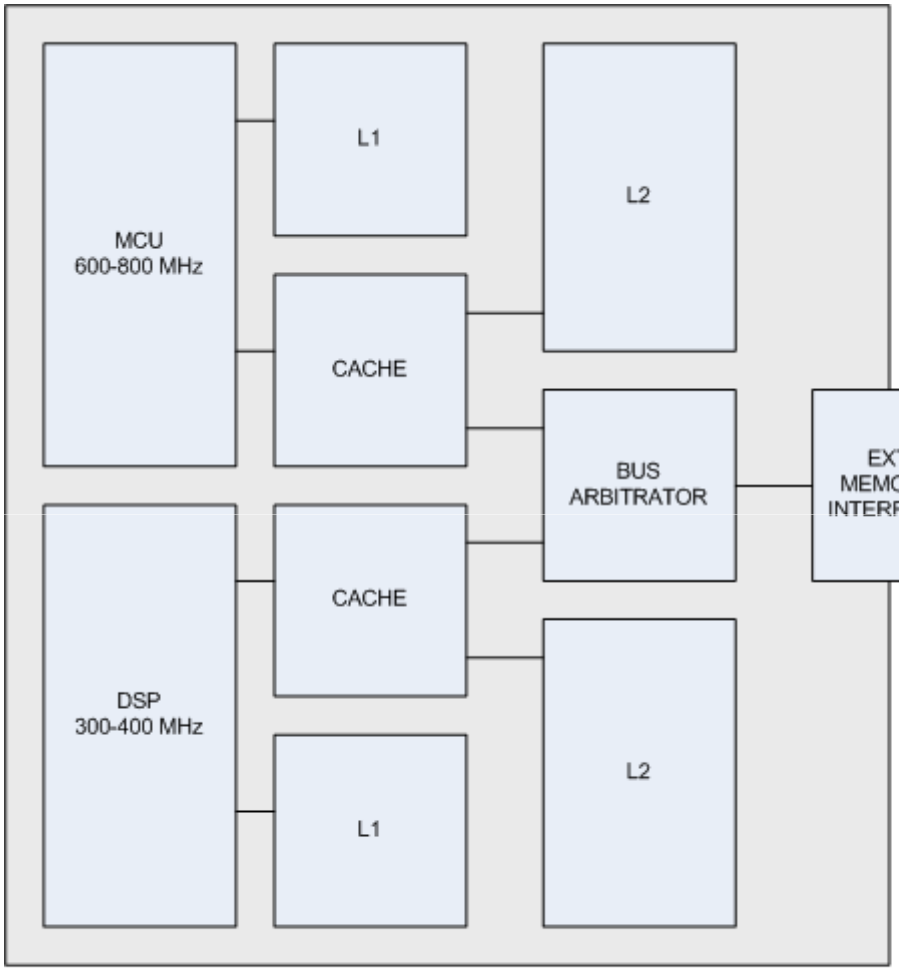


GSM Handset – Scheduling Jitter



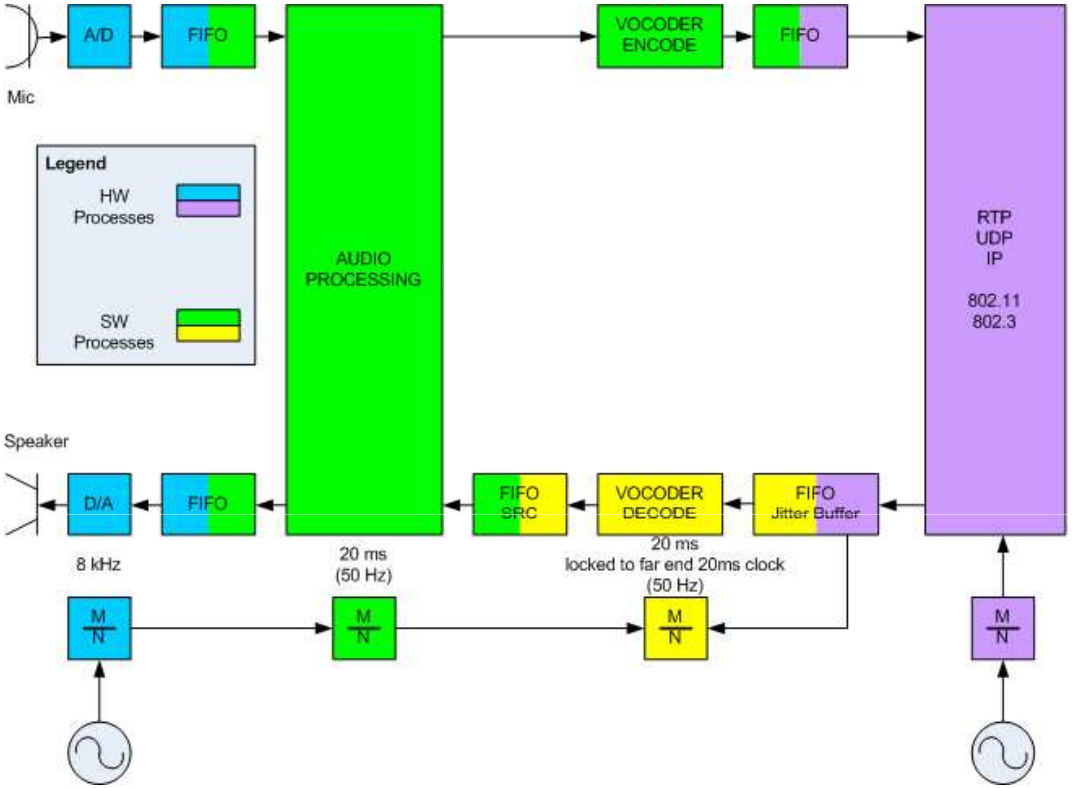
BlackBerry

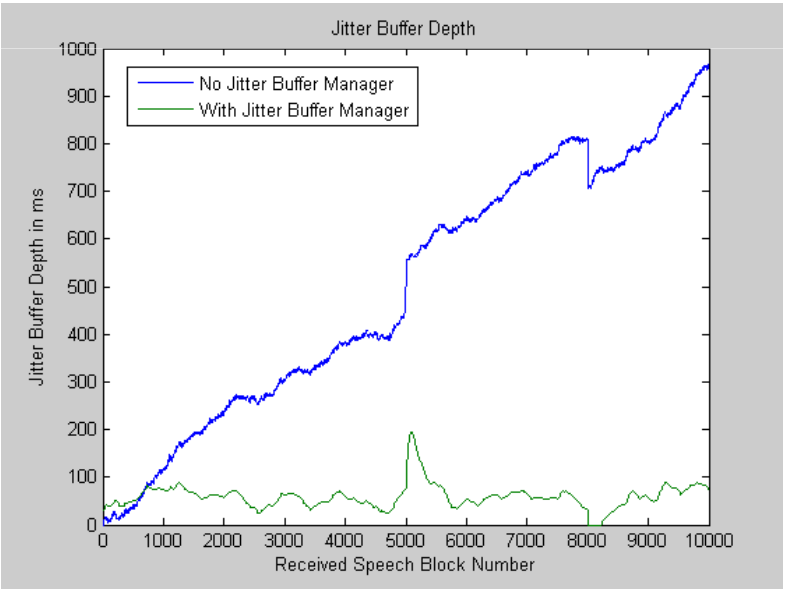
GSM Handset – Memory Hierarchy



Latency for first access	<=1 Wait State	10-20 Wait States	50-100 Wait States
Throughput	Processor Rate	Processor Rate	10-100 Mega Words/s

VoIP





BlackBerry