IBM Research

Mobile Speech Solutions and Conversational Multi-modal Computing

Multi-Modal Browser Architecture

Overview on the support of multi-modal browsers in 3GPP

Stéphane H. Maes, IBM
smaes@us.ibm.com

In collaboration with Chummun Ferial, SONY
ferial.chummun@ipce.eu.sony.co.jp

OUTLINE

- Motivation: Multi-modal Mobile e-Business
  - Introduction
  - Multi-channel and pain points
  - Multi-modal and value proposition
  - Definitions
- Recommended architecture
  - MVC Principle
  - DOM-Based MVC multi-modal and multi-device browsers
  - Supported configurations
- What is needed?
  - Infrastructure
  - Protocols, Interfaces and Components to be Standardize
  - DSR and Multi-modal Protocol stack for 3GPP
- Conclusions

Motivation: Multi-modal Mobile e-Business

Introduction - Multi-modal Browser

- **Modality:** A particular type physical interface that can be perceived or interacted with by the user (e.g. voice interface, GUI display with keypad etc...)
- **Multi-modal Browser:**
  - A browser that enables the user to interact with an application through different modes of intercation (e.g. typically: Voice and GUI).
  - Accordingly a multi-modal-browser provides different moadlities for input and output
  - Ideally it lets the user select at any time the modality that is the most appropriate to perform a particular interaction given this interaction and the users situation (activity, environment etc...)

- **Thesis:** By improving the user interface, we believe that multi-modal browsing will significantly accelerate the acceptance and growth of m-Commerce.

Multi-channel scenario: travel reservations

- Same application can be adapted to different channels
- Synchronization across different channels is needed but more complex.

- Multiple access mechanisms
- One interaction mode per device

PC

Flights Hotels Cars Packages Cruises Maps

EXPRESS SEARCH

Departing from: [_____]

Going to: [_____]

When are you leaving? [Dec] [31] [Noon]

When are you returning? [Jan] [_1] [Noon]

Tip: We have many more flight, hotel, and car options.

WHAT'S NEW

Ski Travel: Choose from more than 80 ski destinations

Cruise Travel: Take a virtual tour of select cruise ships

+ Standardized rich visual interface

– Not suitable for mobile use

Voice

I need a direct flight from New York to San Francisco after 7:30pm today

There are five direct flights from New York's LaGuardia airport to San Francisco after 7:30pm today: Delta flight nnn...

Book me on the United flight

+ Access from any telephone

– Output is inherently sequential

WAP

From: LGA__	From: LGA__	From: LGA__
To: _____	To: SFO__	To: SFO__
Date: _____	Date: _____	Date: 00/12/11

+ Mobile and becoming ubiquitous

– Hard to enter data

Pain points in multi-channel e-business

Most mobile device usage today is not for e-business applications

Pain points

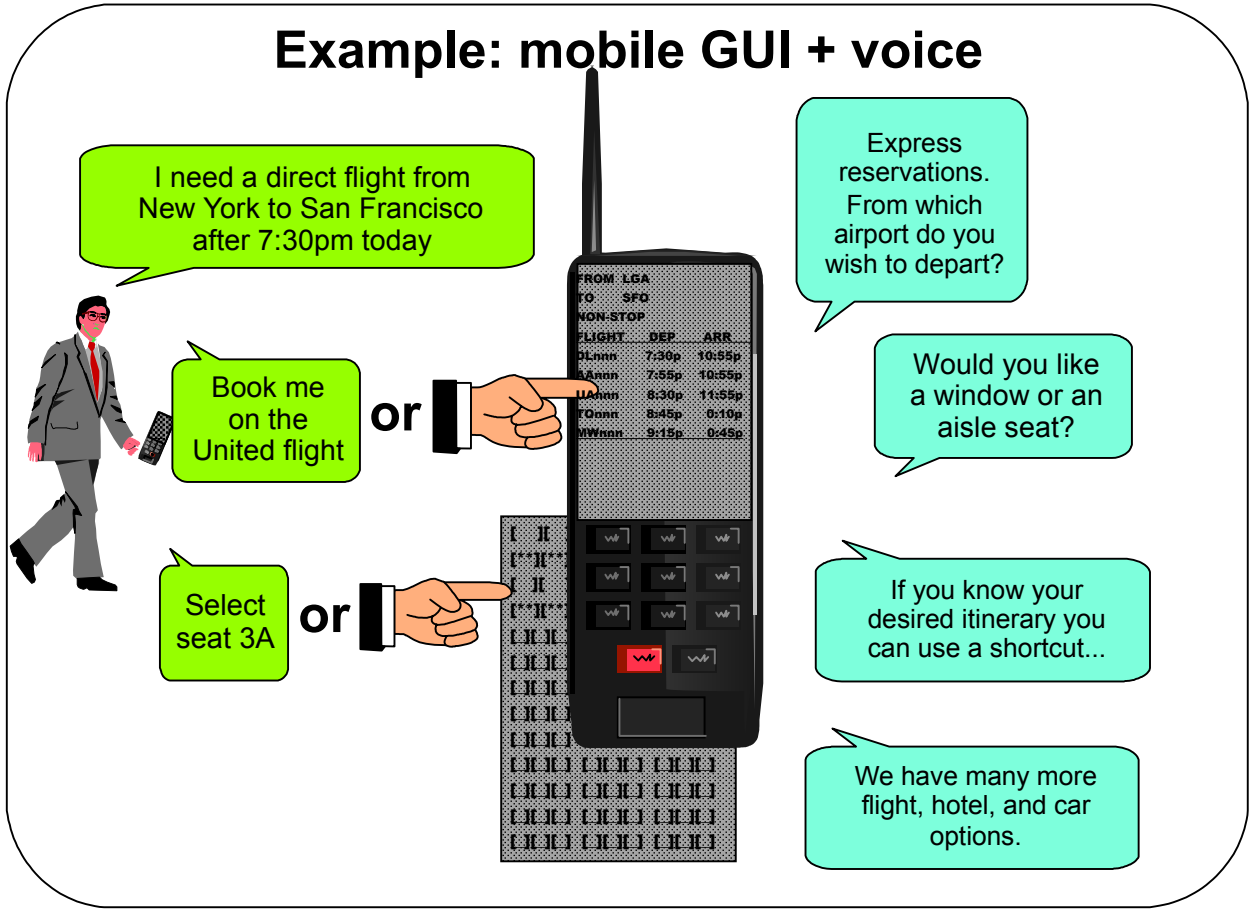
- **hard to enter and access data using small devices**
  - tiny keypads and screens
- **voice Recognition still makes mistakes**
  - blocking if repeated
- **Voice is serial**
  - difficult to manage long output
- **one interaction mode does not suit all circumstances**
  - each mode has its pros and cons
- **all-in-one devices are no panacea**
  - bulky and expensive
- **multiple devices have pros and cons**

No immediate relief is in sight:

- Devices are getting smaller, not larger
- Devices and applications are becoming more complex
- Adding color, animation, camera, etc. does not simplify or contribute to e-business
- CRMs / IVRs are mostly not yet web-centric

Multi-modal scenario: travel reservation

- User can select at any time the preferred modality of interaction
- Can be extended to selection of the preferred device (multi-device)



Additional examples:

- display seat selection chart (not simply "window or aisle")
- use voice or keys to enter PIN code and performs speaker verification
- use audio or voice for notifications
- information can be saved for later use

- User is not tied to a particular channel's presentation flow
- Interaction becomes a personal and optimized experience
- Multi-modal output is an example of multi-media where the different modalities are closely synchronized.

Multi-modal e-business value proposition

Multi-modal e-business value proposition

- **easily enter and access data using small devices**
  - by combining multiple input & output modes
- **choose at any time the interaction mode that suits the task and circumstances**
  - input: key, touch, stylus, voice...
  - output: display, tactile, audio...
  - don't be blocked by limitation / mistakes of a given interaction mode at a given moment
- **use several devices in combination**
  - by exploiting the resources of multiple devices

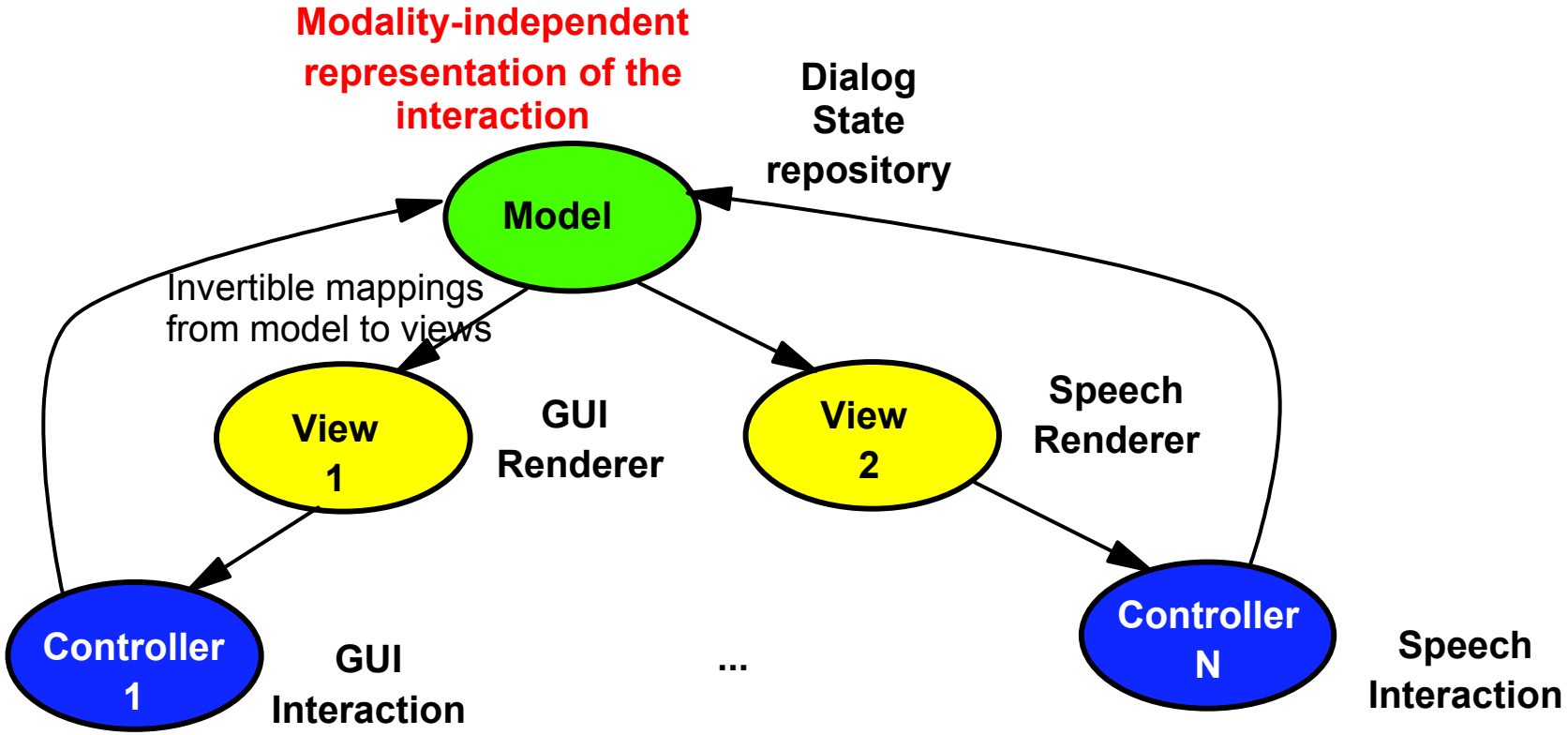
Definitions - Summary

- **Channel:** a particular user agent, device, or a particular modality. Do not confuse this is not a physical communication channel. It is rather typpically the browser / user agent used to access, browse and interact with online information
- **Multi-channel applications:** applications designed for ubiquitous access through different channels, one channel at a time. No particular attention is paid to synchronization or coordination across different channels.
- **Multi-modal applications:** multi-channel applications, where multiple channels are simultaneously available and synchronized.
  - There are no fundamental differences between **multiple devices** (**multi-device browsing**) and **multiple modalities.**

Recommended Architecture

Model View Controller Principle (MVC)

User must be able to switch channel at any unpredictable moment while interacting with the application and seamlessly continue to interact.

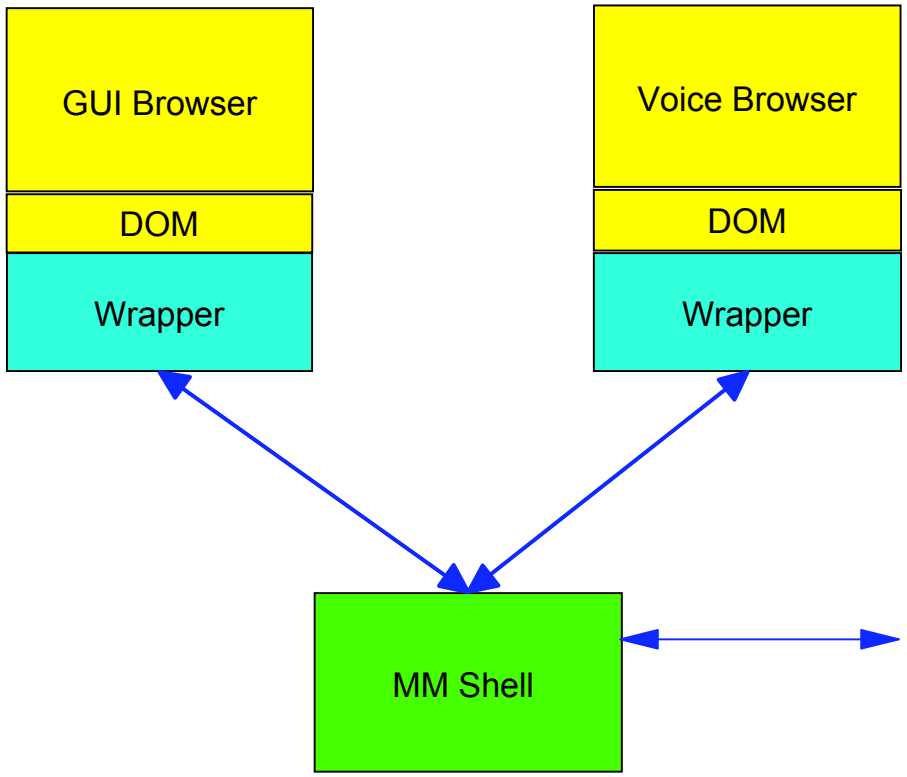


Model View Controller Architecture for Multi-modal or Multi-device Browser

- DOM: Document Object Model (http://www.w3c.org/dom).
- Adapted definitions:
- DOM L1: Interface that enables manipulation of the XML document in each browser
- DOM L2: Interface that provides access to the events associated to the user interaction within each browser

Target: DOM-based architecture

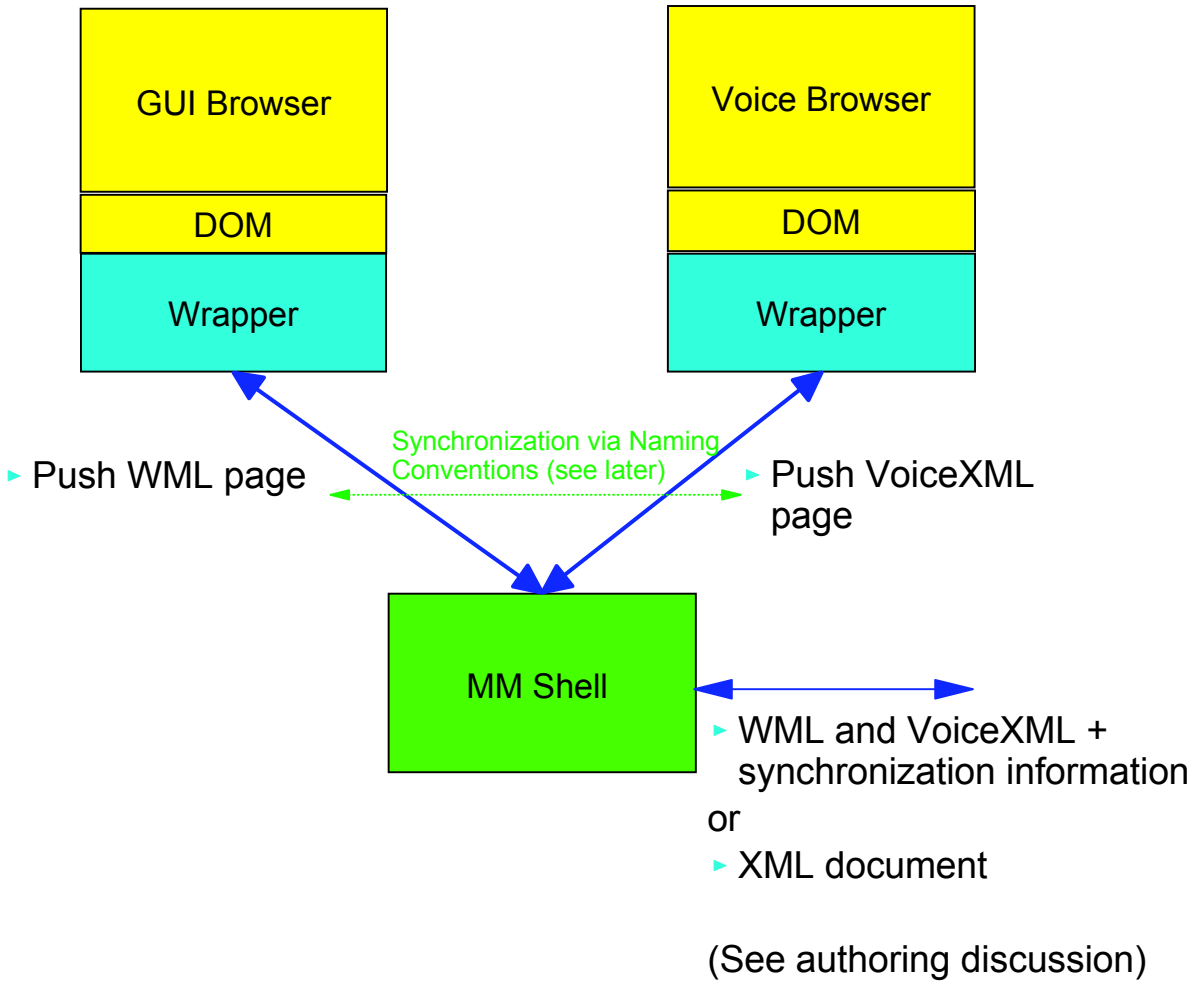
Each piece is distributable



This can be another browser in the case of multi-device browsing

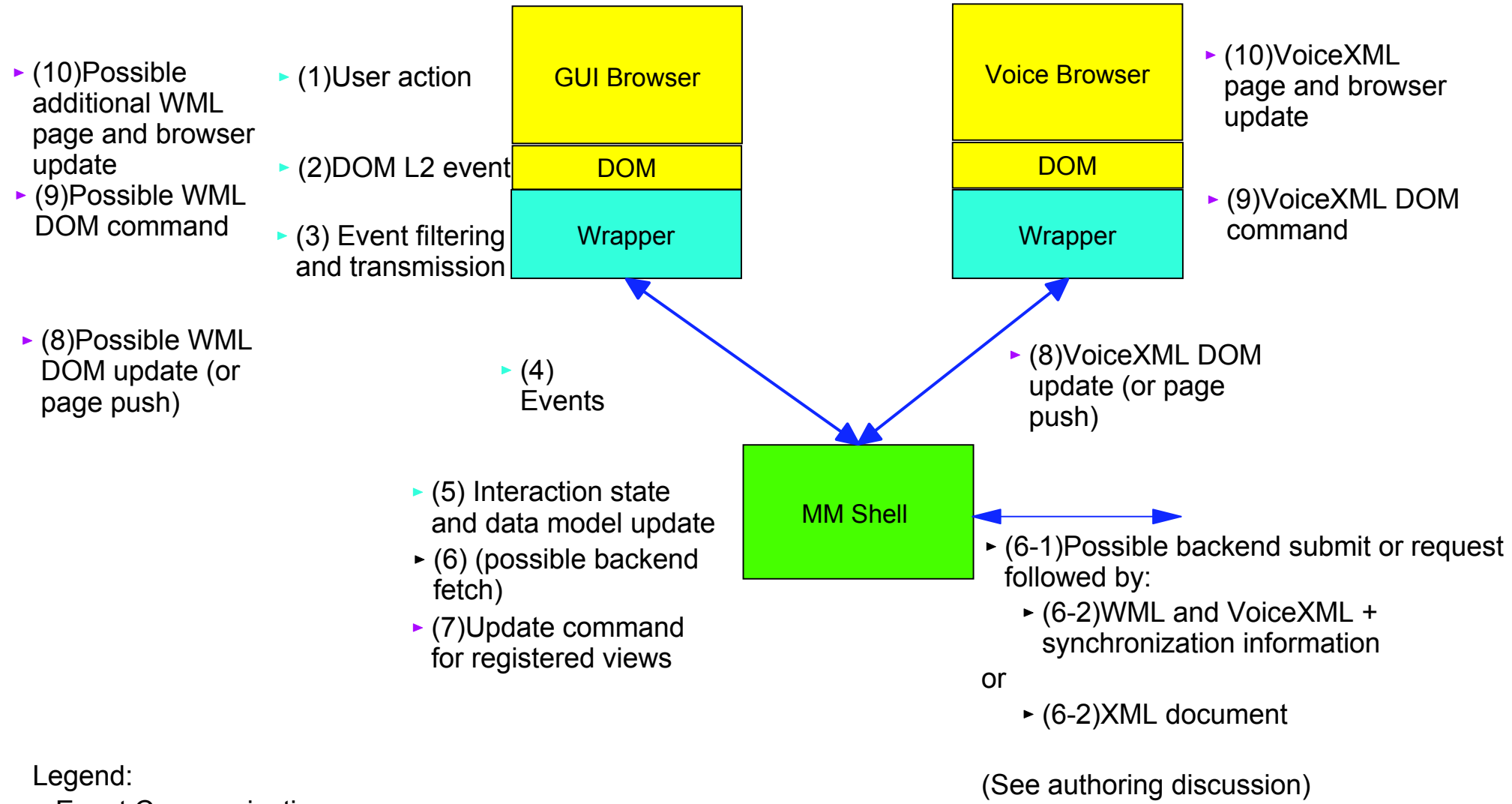
WAP MVC Multi-modal Browser

Initialization



WAP MVC Multi-modal Browser

Interaction: Assuming GUI interaction

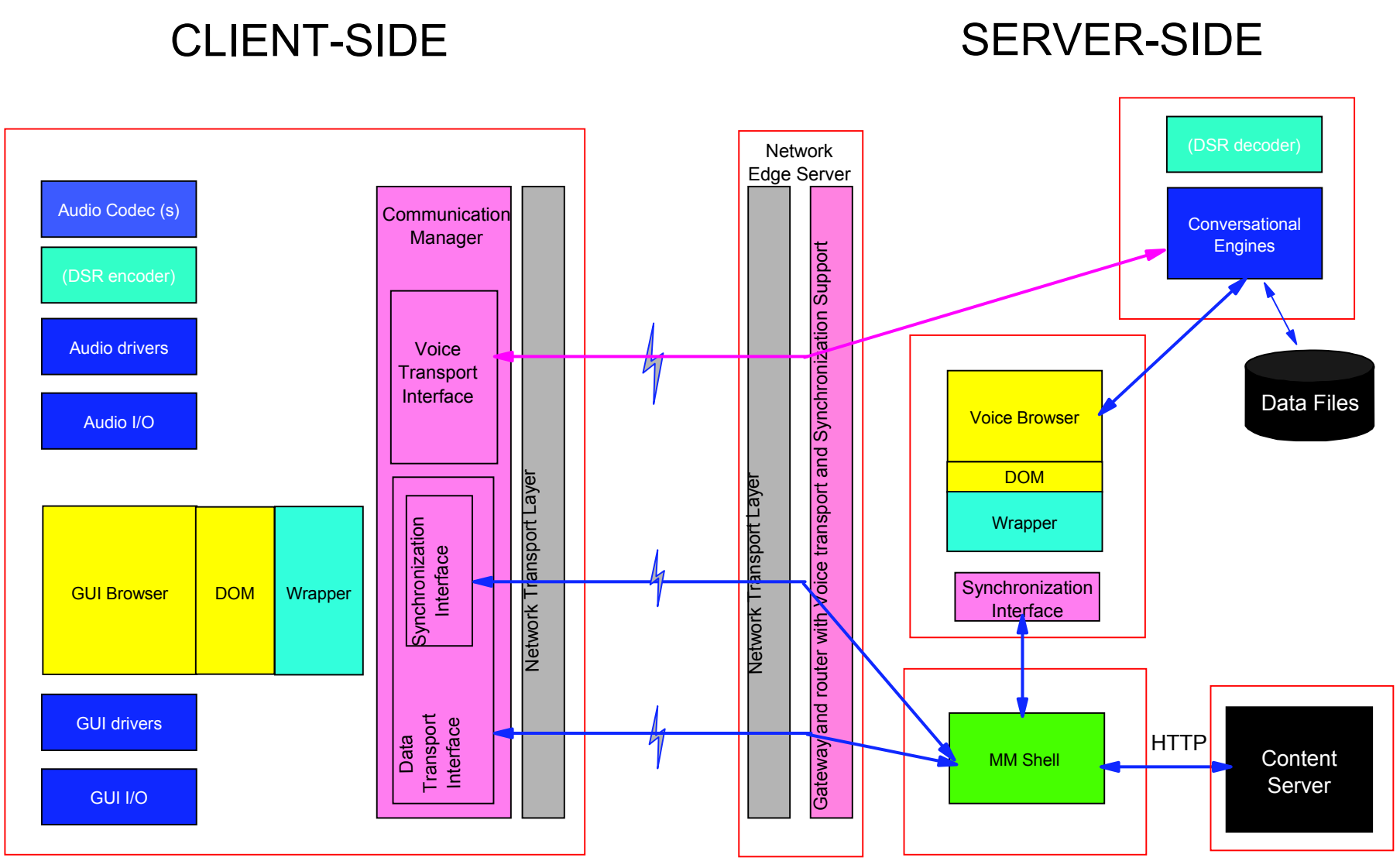


Legend:

- Event Communication
- Next Step (not always present)
- Synchronization of the views

Target Multi-modal Architecture: Thin Client

- Recommended target aarchitecture for most 3G terminals (smart phones):
  - Enables small client foot-print
  - Synchronization and voice recognition / conversational functions are on the server-side

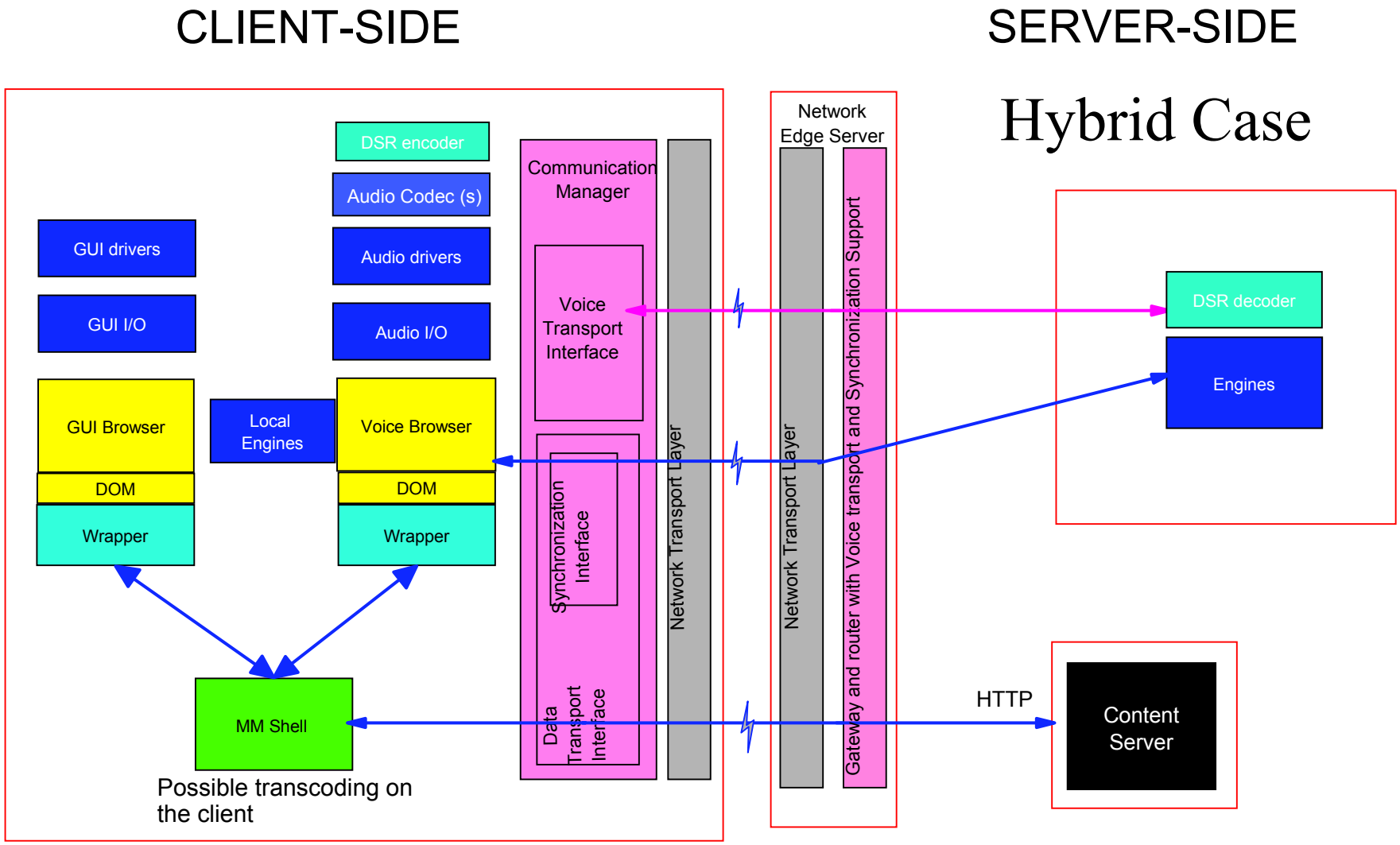


DSR is optional to improve performances of speech recognition. It can be done with existing codecs at the cost of accuracy drops

DSR - Distributed speech Recognition - See: T2-010627 (LS from SA-1: S1-010847)

Target Multi-modal Architecture: Fat Client

- Possible architecture with fatter terminals
- Requires resources to synchronize and for speech recognition / conversational engines
- Fat configuration support disconnected usage
- Hybrid case supports case where embedded client side-speech recognition capabilities are too limited for the task

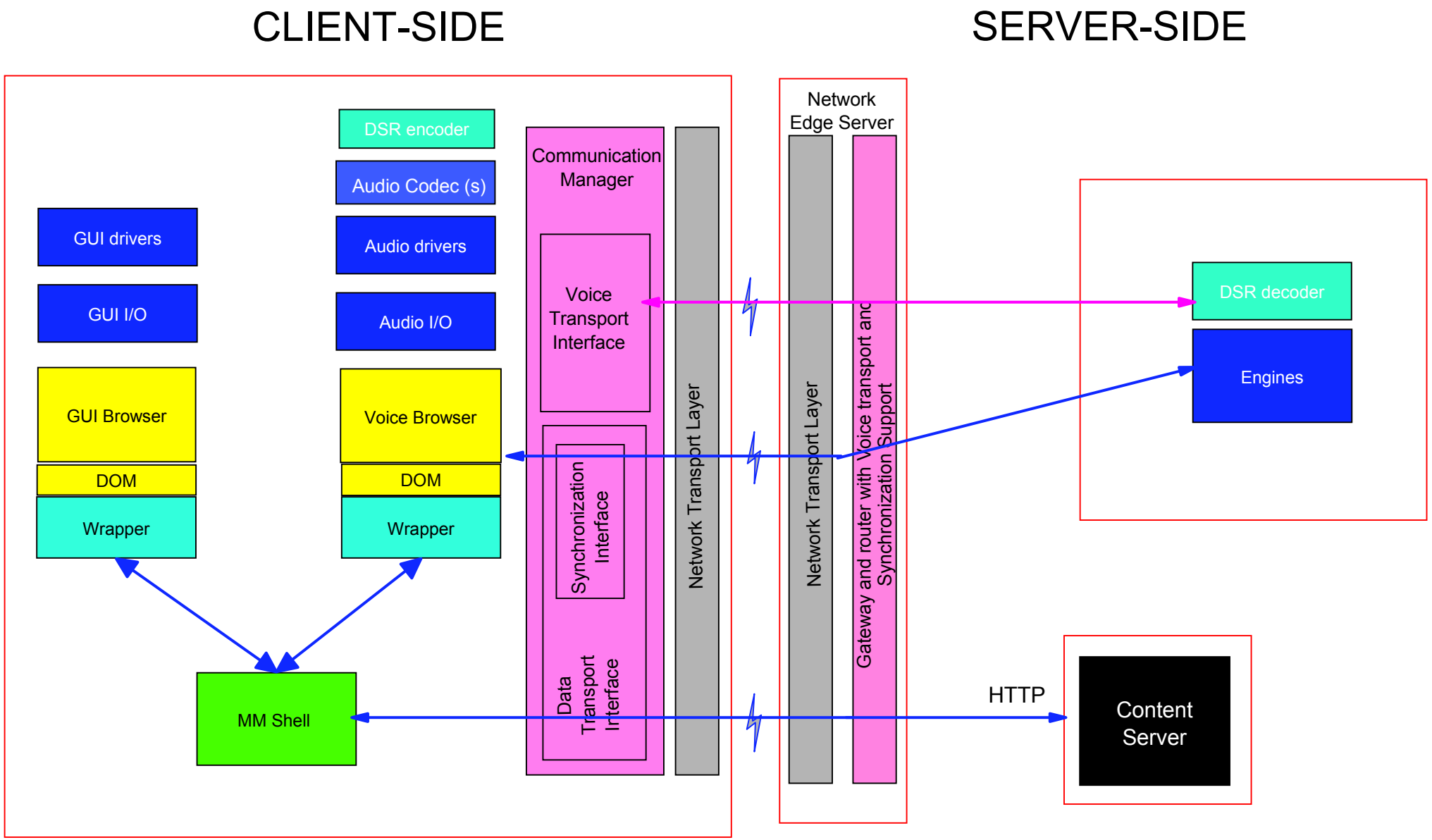


DSR is optional

DSR - Distributed speech Recognition - See: T2-010627 (LS from SA-1: S1-010847)

Variation of the fat client configuration - DSR and server side speech recognition

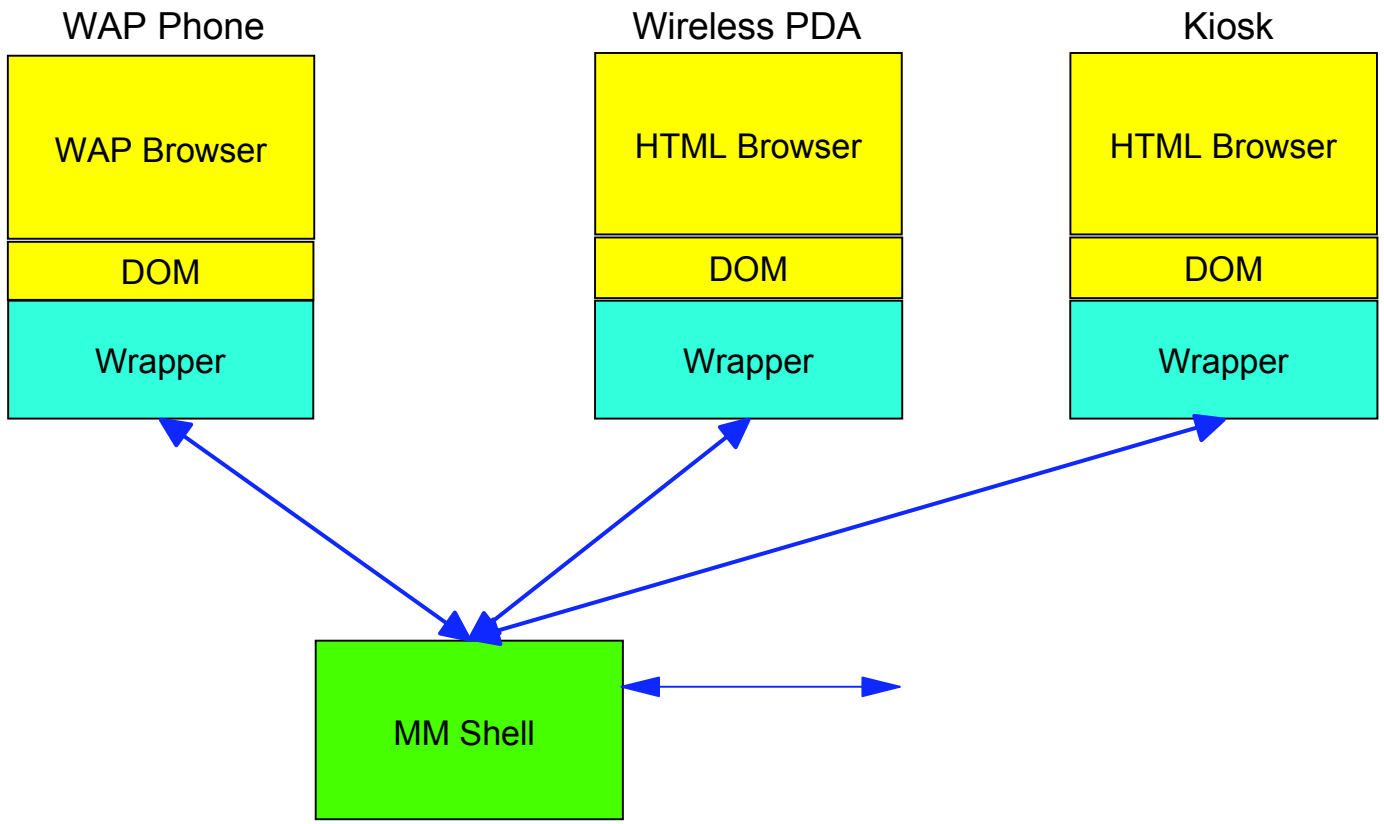
- This can address requirements to maintain the "context" on the client



DSR is optional

Multi-device Browser Configuration

- Recommended target aarchitecture for multi-device browsing
  - Related to UEsplit (3GPP Activity)



Each piece is distributable

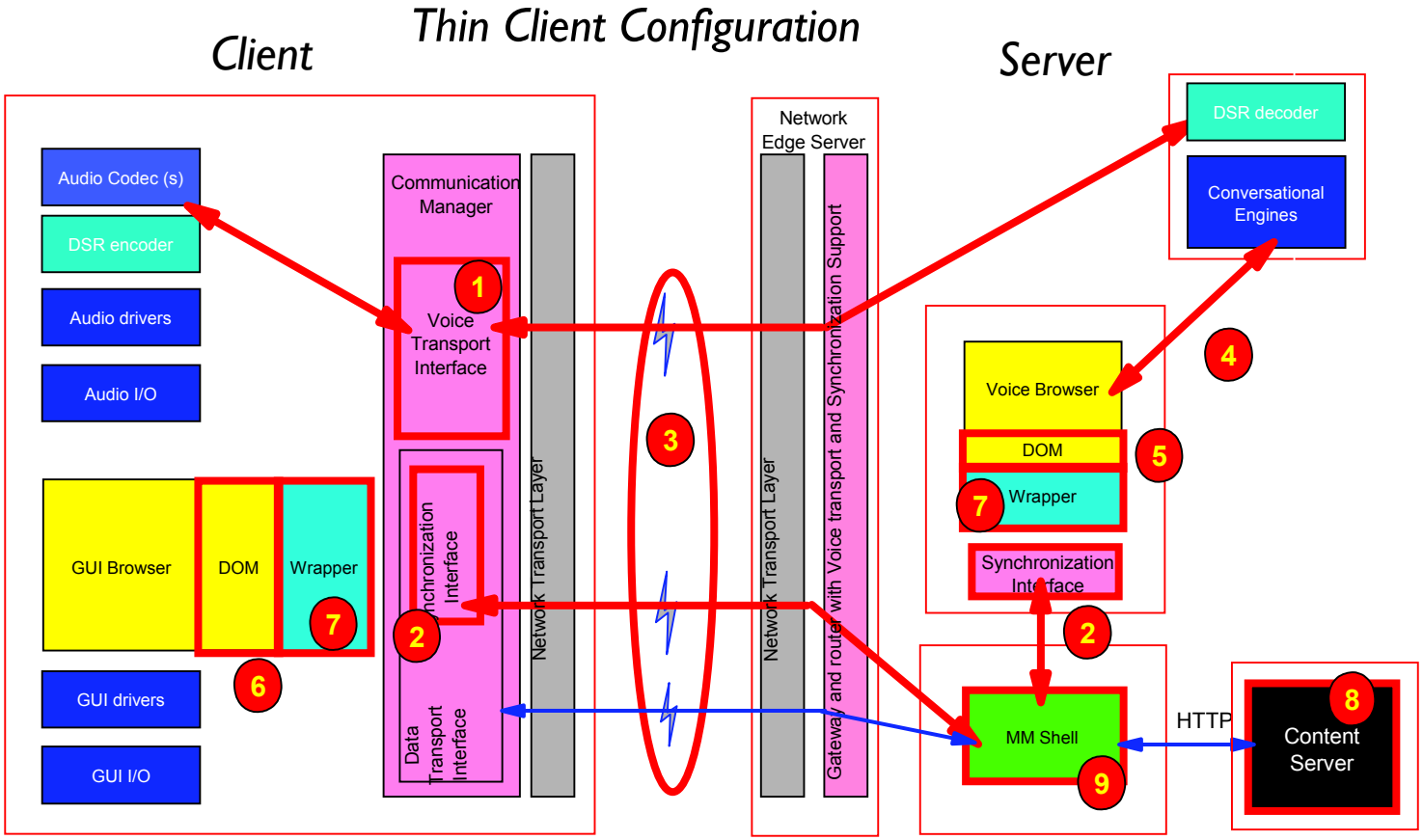
What is needed?

Infrastructure Requirements and current inhibitors

- **Client:**
  - DOM-L2 compliant browsers and Wrapper (or look alike or subset)
  - Support for synchronization protocols (e.g. SOAP)
    - SOAP (1.1) is currently defined by W3C as XML protocol
  - Support for Voice and Data (VoIP, DSR stack (SIP, SDP,SOAP, Payload),...)
  - Capabilities (audio sub-system; CPU / memory for Fat client configurations)
  - Channel / user descriptors: delivery context descriptors
  - Dynamic discosvery and bindings (later)
- **Network and gateways:**
  - Support voice and data (**DSR protocol stack** - T2-010627)
  - Support synchronization protocols (**SOAP over SIP**)
  - Support session / user information exchanges (Delivery context)
- **Server middleware:**
  - Supports:
    - voice and data (DSR protocol stack)
    - Synchronization protocols (SOAP)
    - Session / user management (delivery context)
    - Synchronization, state persistence
- **Authoring:**
  - Standards
  - Tools

These inhibitors should disappear in the next 2 to 5 years.

Protocols, Interfaces and Components to Standardize



Thin Client Configuration / Multi-device

1) **ETSI - STQ, IETF-AVT and 3GPP, ITU-SG16**
2) **W3C** (**XML Protocols** / MM), ETSI, **WAP Forum**, **3GPP**
   W3C DI for delivery context
3) IETF, **3GPP**, WAP Forum
4) **ETSI - STQ**
5) **W3C Voice Activity**
6) **WAP Forum**, W3C, ETSI-STQ, 3GPP?
7) W3C MM WG, **WAP Forum** (WAE - Mobile DOM), 3GPP?
8) **W3C** (DI, XForms, MM, etc...), **WAP Forum**
9) W3C, WAP Forum

etc…

Fat Client Configuration:

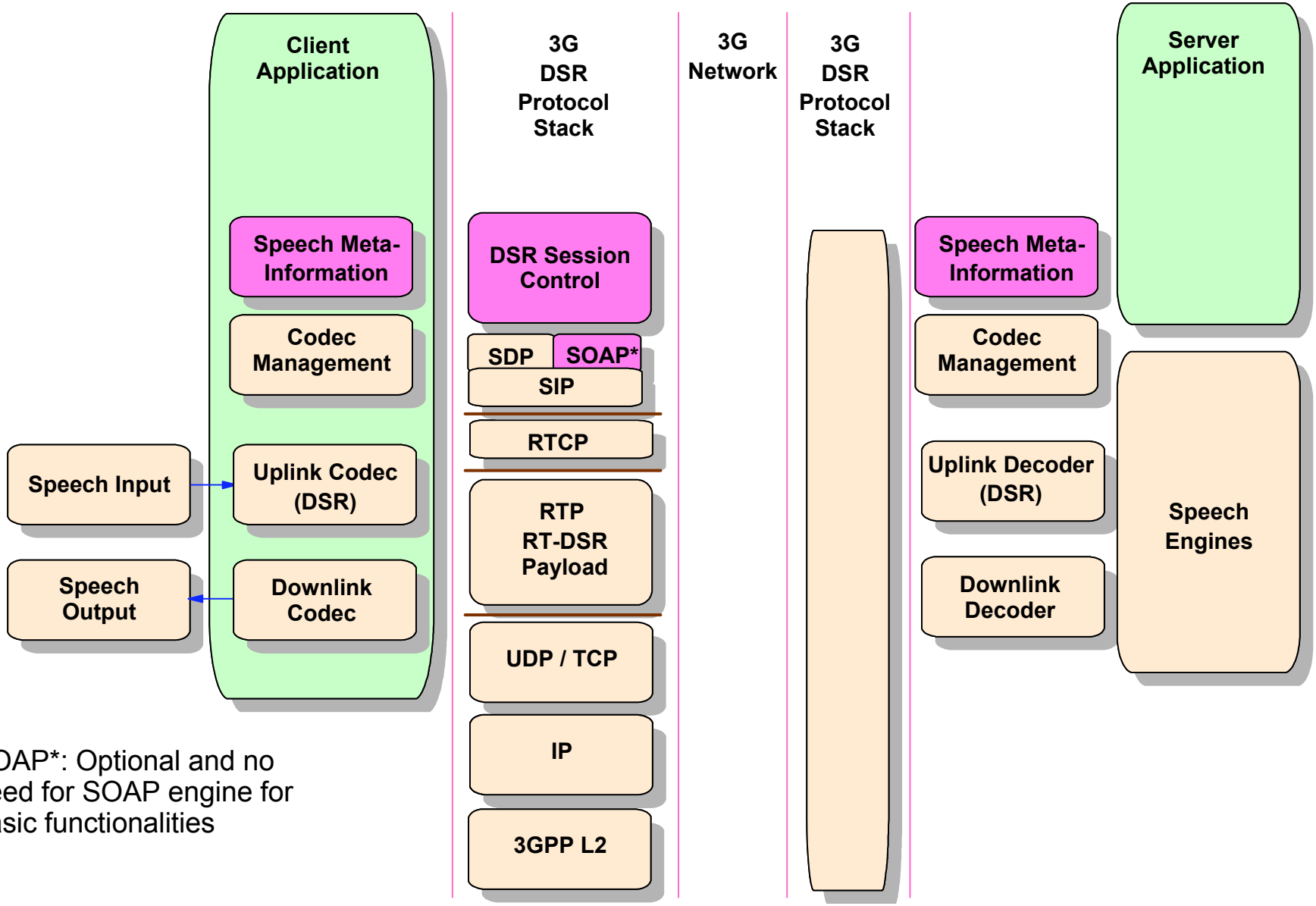
5) **W3C Voice activity**
6) **WAP Forum**, W3C, ETSI-STQ, 3GPP?
7) W3C MM WG, **WAP Forum,** 3GPP?
8) **W3C** (DI, XForms, MM, etc...), **WAP Forum**
9) W3C, WAP Forum

Conclusions

- Multi-modal and Multi-device browsing can accelerate the growth of wireless internet and m-Commerce
- Inhibitors will disappear in next 2 to 5 years
- Standards are key to eliminate the inhibitors and seed the market
- We have proposed a standard-based flexible, modular and extensible architecture and associated programming model
- Numerous items could be addressed by 3GPP:
  - Support for DSR and Multi-modal protocols stack (client, network, server and gateways):
    - **SOAP over SIP**
      - SOAP is currently defined by W3C as XML Protocol.
        - Currently 1.1 version exists with bidnings over HTTP for example
      - SOAP over SIP: to be done. Different proposals exist.
        - IBM has a simple implementation proposal
    - To support the stack will defacto enable multi-modal and multi-device deloyments when **User Agent** offers DOM L1/L2 appropriate interface (e.g. WAP)
  - Support for architecture, authoring and standardization elsewhere
  - Inclusion of compatibility requirements in current standardization activities
  - **Client components (DOM L1/L2, wrapper, SOAP support)**

Background Material

Distributed Speech Recognition - for 3GPP Release 5 - A first step - MM angle

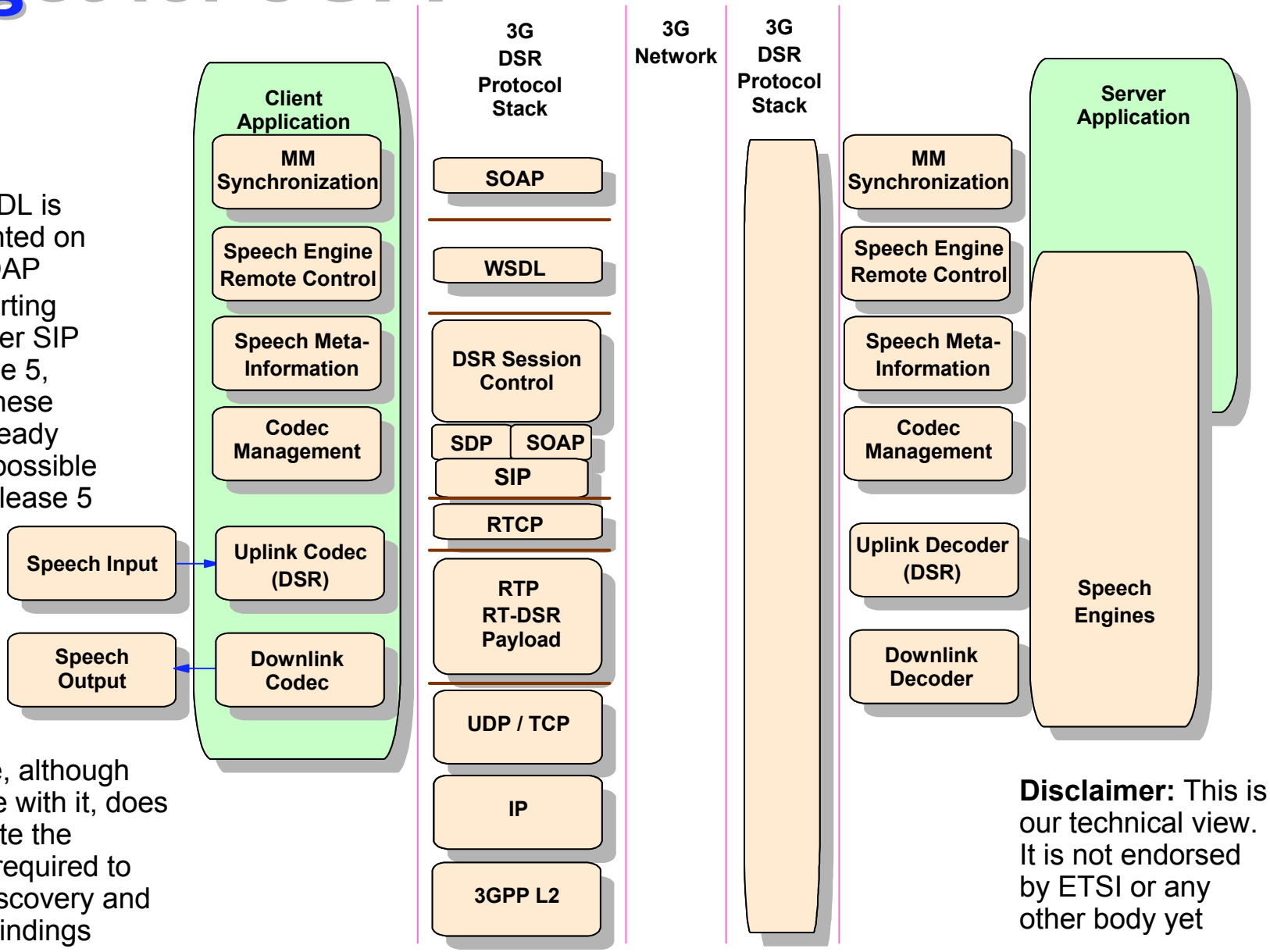


SOAP*: Optional and no need for SOAP engine for basic functionalities

See: T2-010627 (LS from SA-1: S1-010847) - In magenta: items under discussion at ETSI STQ DSR A&P: IBM Proposal to ETSI for 3GPP submission (not a ETSI endorsed item at this stage)

Distributed Speech Recognition and Multi-modal Protocol stack - Possible Target for 3GPP

- Note WSDL is implemented on top of SOAP
- By supporting SOAP over SIP in Release 5, most of these would already become possible within Release 5



- This figure, although compatible with it, does not illustrate the protocols required to support discovery and dynamic bindings

Disclaimer: This is our technical view. It is not endorsed by ETSI or any other body yet