

Dagstuhl Seminar
“Coordination and Fusion in Multimodal Interaction”



Media Coordination in SmartKom

Norbert Reithinger



Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Stuhlsatzenhausweg 3, Geb. 43.1 - 66123 Saarbrücken

Tel.: (0681) 302-5346

Email: bert@dfki.de

www.smartkom.org

www.dfki.de/~bert

Overview

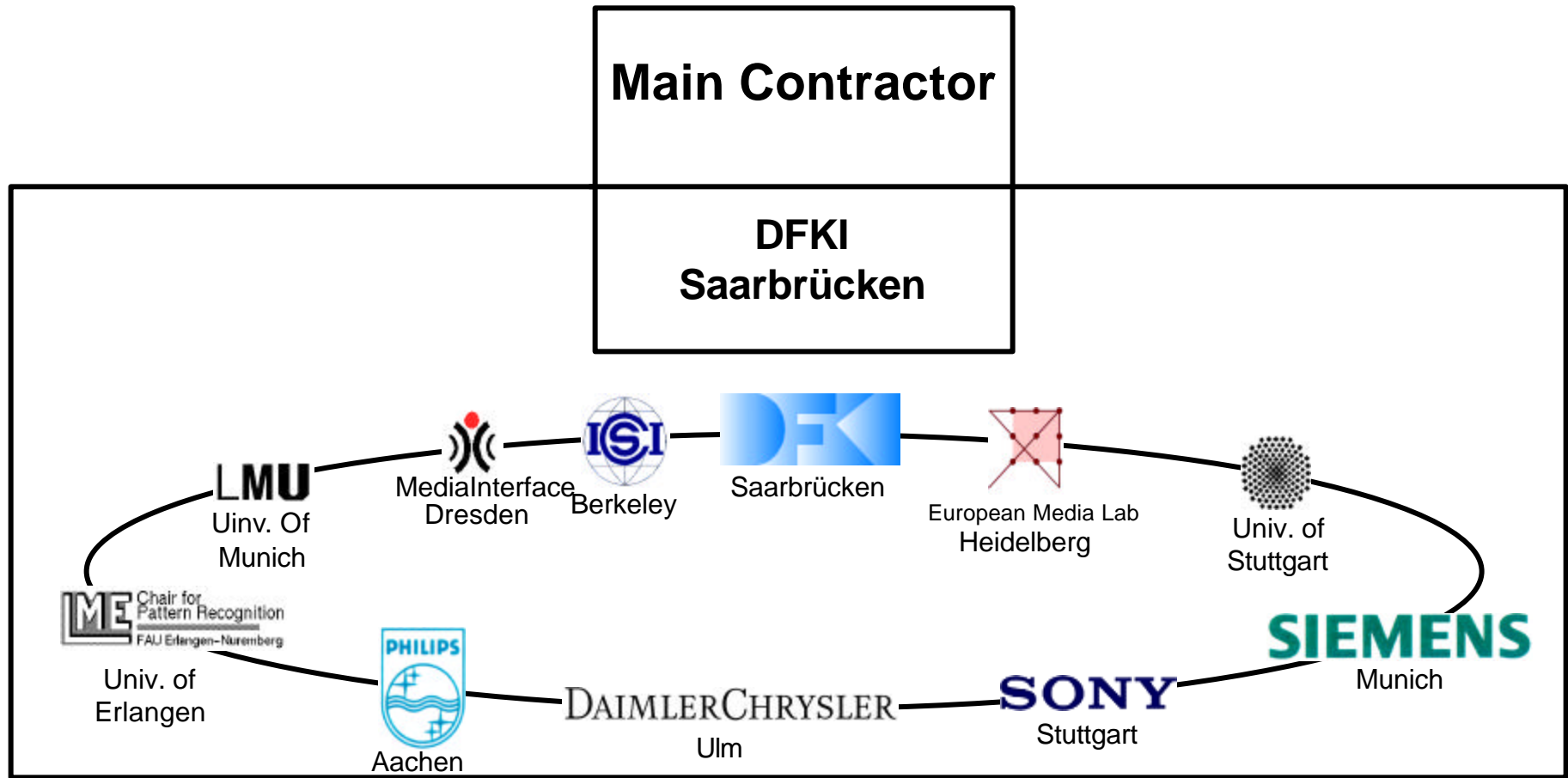
- **Situated Delegation-oriented Dialog Paradigm**
- **More About the System Software**
- **Media Coordination Issues**
- **Media Processing: The Data Flow**
- **Processing the User's State**
- **Media Fusion**
- **Media Design**
- **Conclusion**



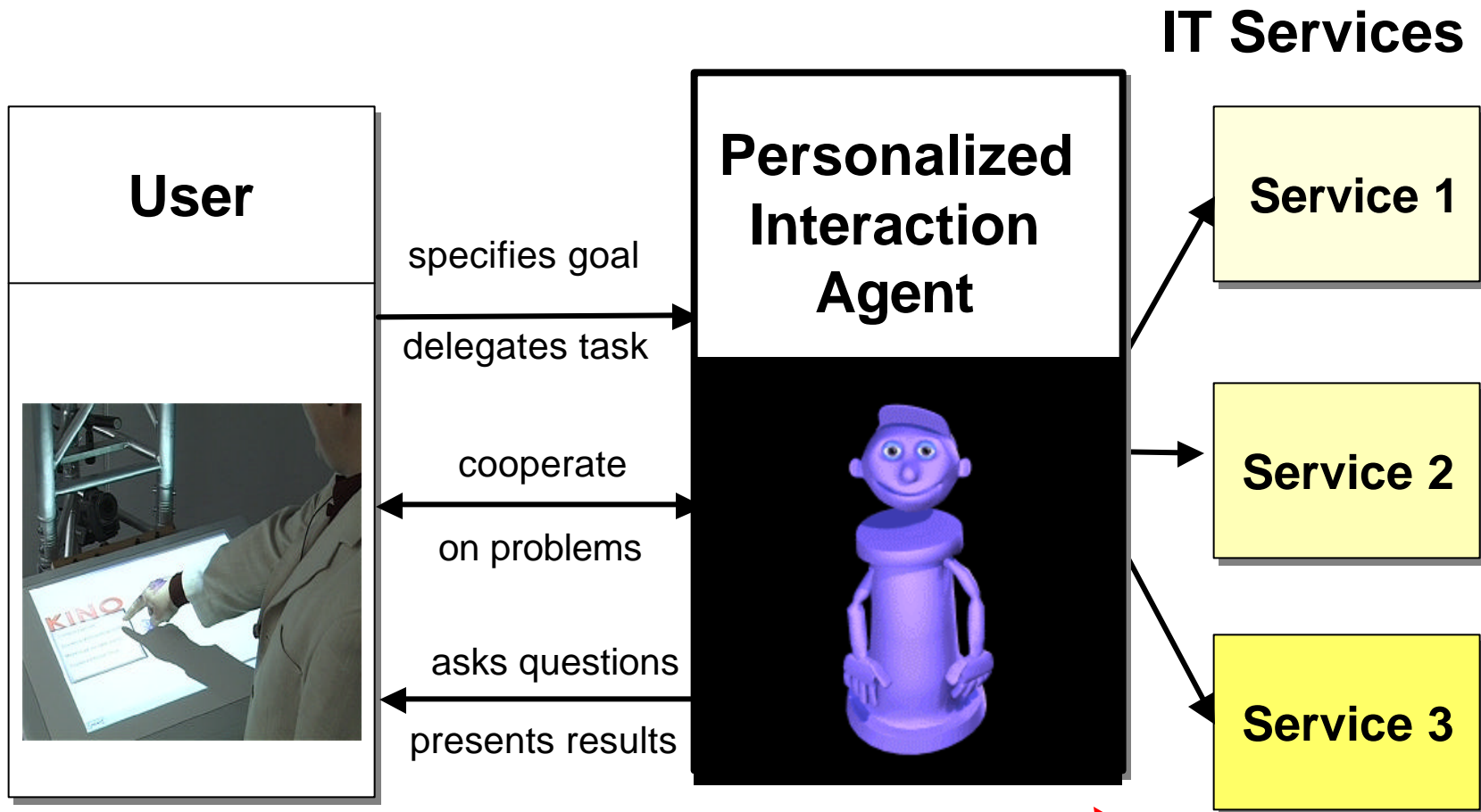
The SmartKom Consortium

Project Budget: €25.5 million

Project Duration: 4 years (September 1999 – September 2003)



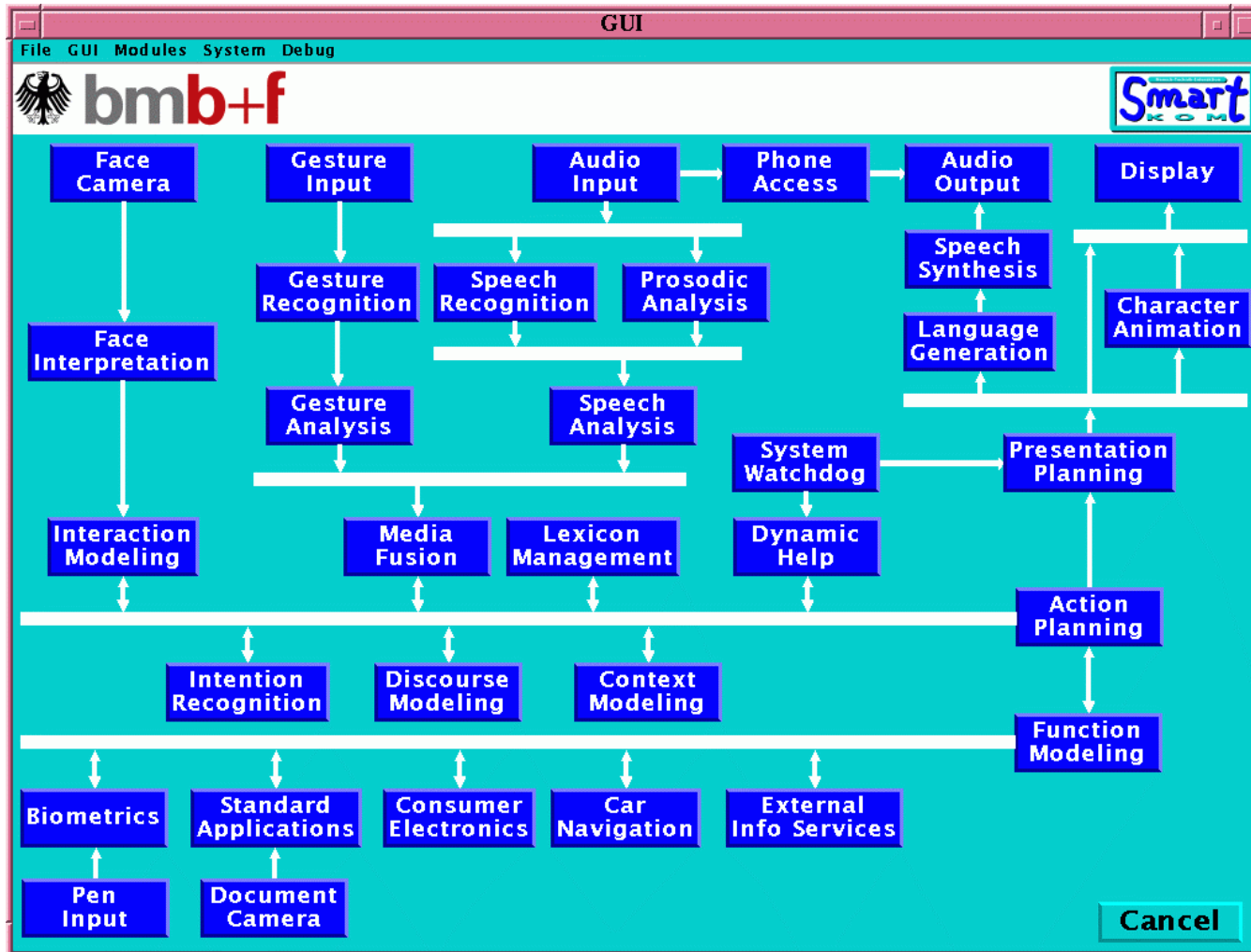
Situated Delegation-oriented Dialog Paradigm



Smartakus



More About the System



More About the System

- **Modules realized as independent processes**
- **Not all must be there (critical path: speech or graphic input to speech or graphic output)**
- **(Mostly) independent from display size**
- **Pool Communication Architecture (PCA) based on PVM for Linux and NT**
- **Modules know about their I/O pools**
- **Literature:**
 - Andreas Klüter, Alassane Ndiaye, Heinz Kirchmann: *Verbmobil From a Software Engineering Point of View: System Design and Software Integration*. In Wolfgang Wahlster: *Verbmobil - Foundation of Speech-To-Speech Translation*. Springer, 2000.
- **Data exchanged using M3L documents** <C:\Documents and Settings\bert\Desktop\SmartKom-Systeminfo\index.html>
- **All modules and pools are visualized here ...**



Recording

March
Tuesday, October 23, 2001
12:28:17

Recognition

Synthesis

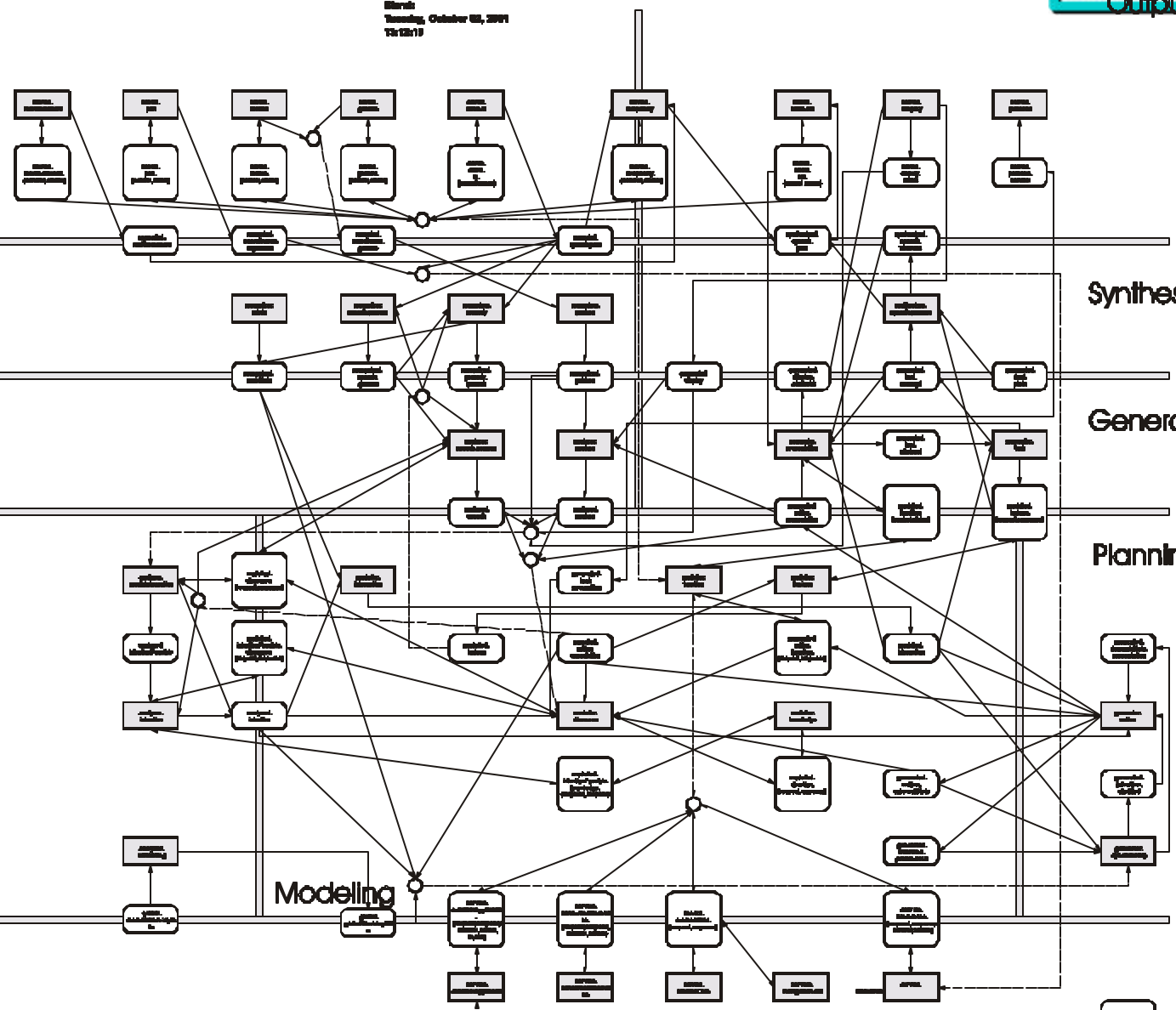
Interpretation

Generation

Understanding

Planning

Services



Modelling



Media Coordination Issues

- **Input:**
 - Speech
 - Words
 - Prosody: boundaries, stress, emotion
 - Mimics: neutral, anger
 - Gesture:
 - Touch free (scenario public)
 - Touch sensitive screen
- **Output:**
 - Display objects
 - Speech
 - Agent: posture, gesture, lip movement

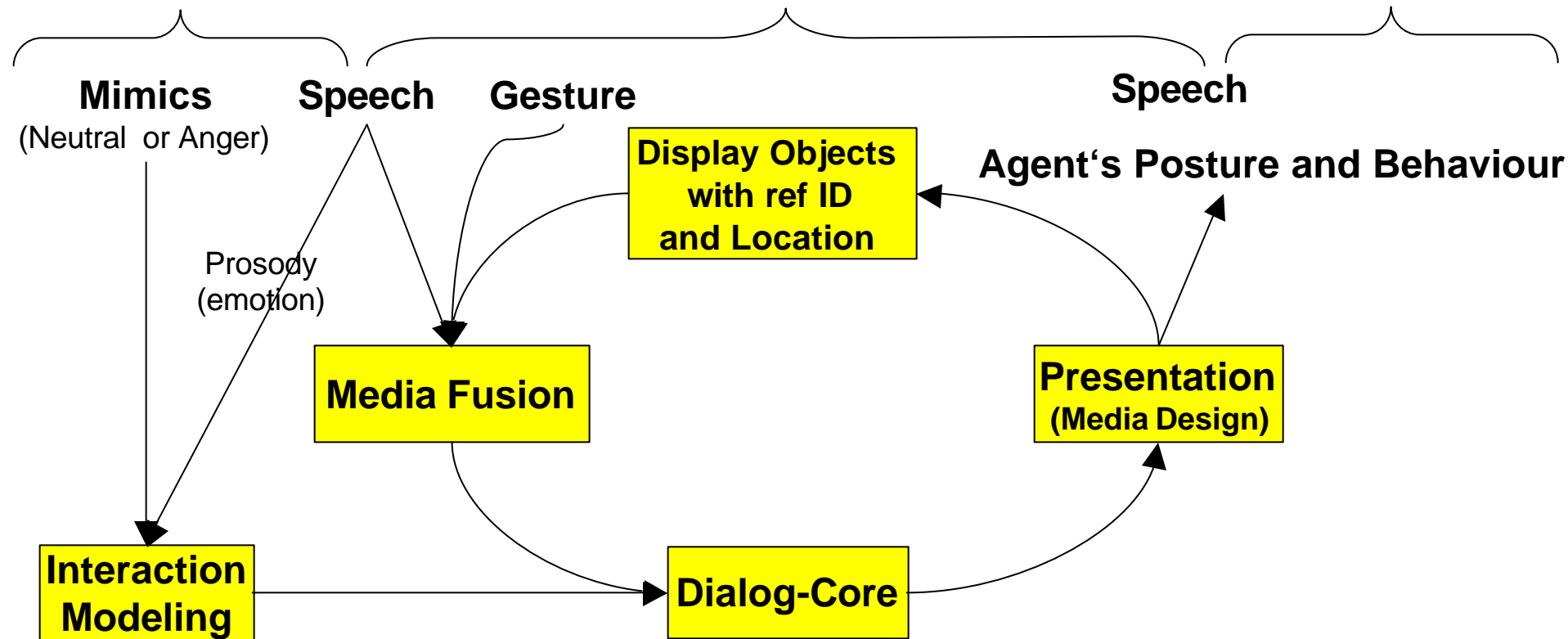


Media Processing: The Data Flow

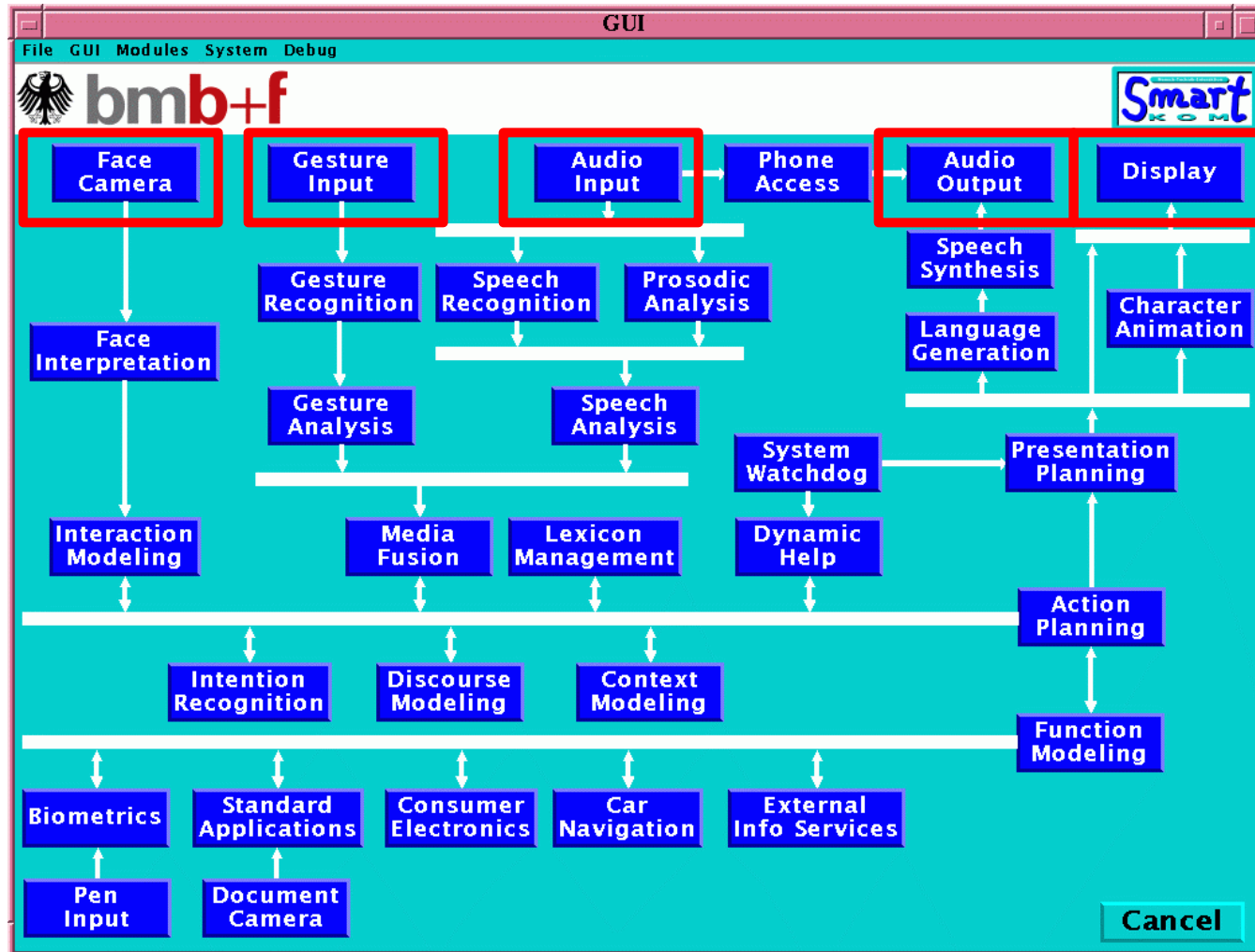
User State

Domain Information

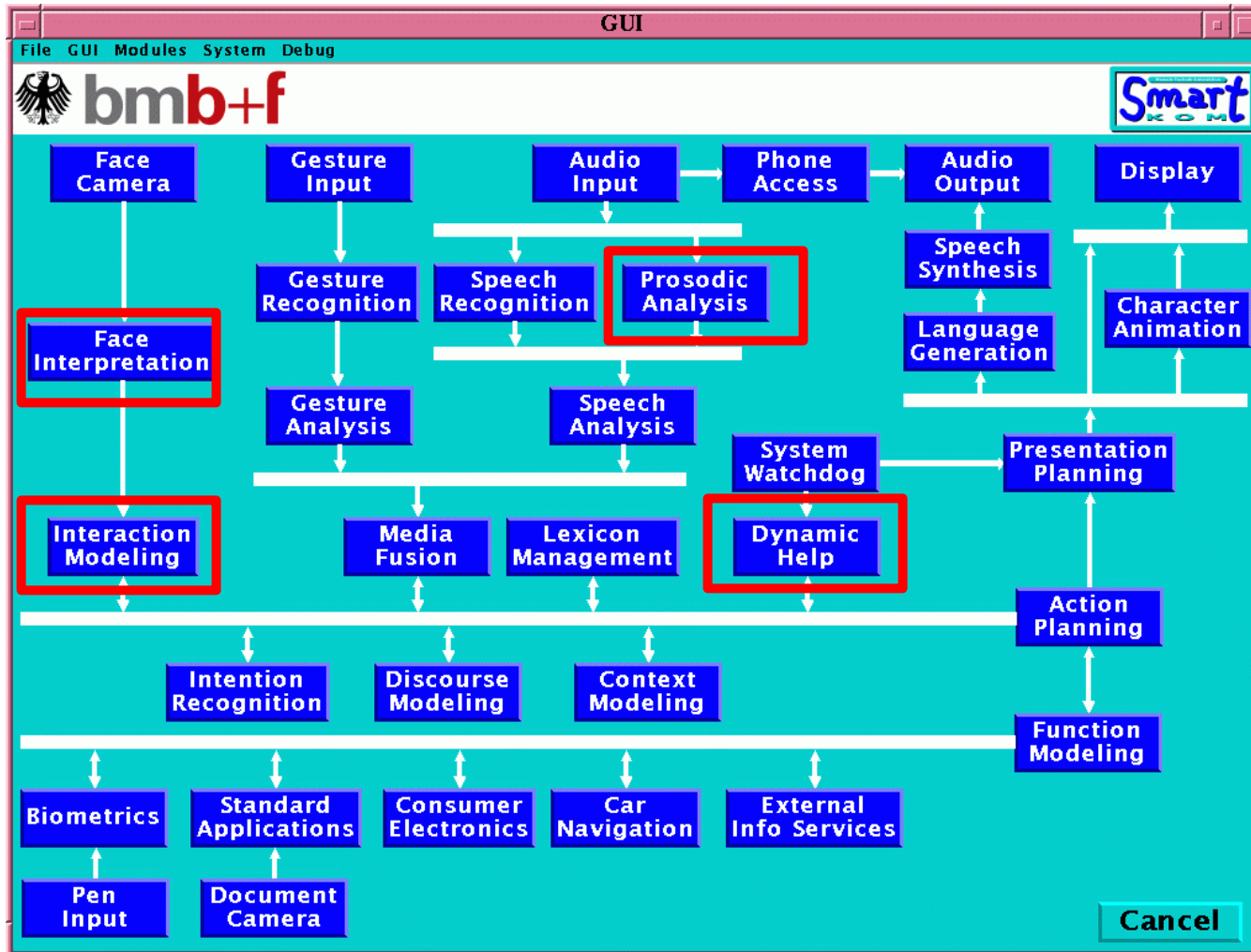
System State



The Input/Output Modules



Processing the User's State



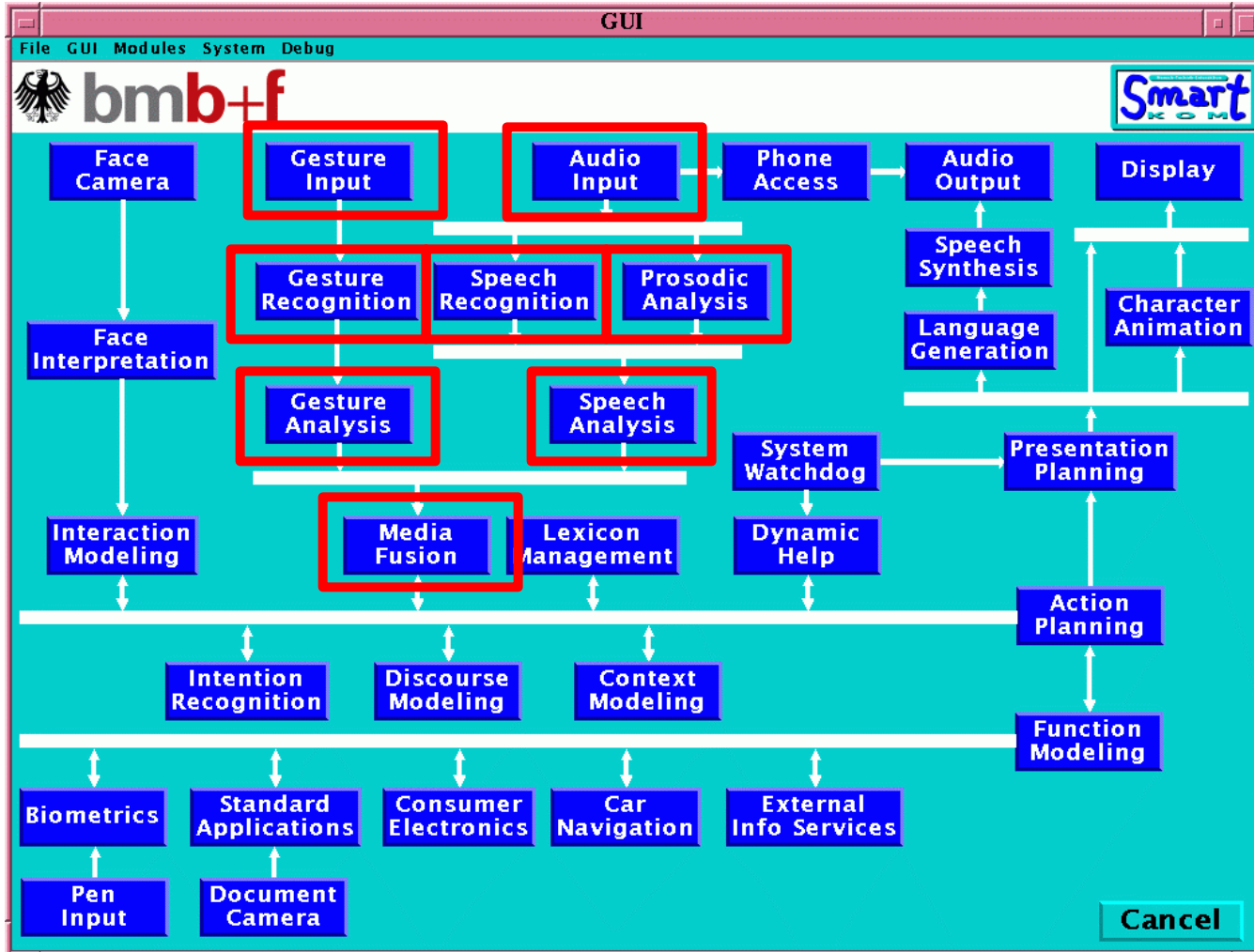
Processing the User's State

- **User state: neutral and anger**
- **Recognized using mimics and prosody**
- **In case of anger activate the dynamic help in the Dialog Core Engine**

- **Elmar Nöth will hopefully tell you more about this in his talk *Modeling the User State - The Role of Emotions***



Media Fusion



Gesture Processing

- **Objects on the screen are tagged with IDs**
- **Gesture input**
 - Natural gestures recognized by SIVIT
 - Touch sensitive screen
- **Gesture recognition**
 - Location
 - Type of gesture: pointing, tarrying, encircling
- **Gesture Analysis**
 - Reference object in the display described as XML domain model (sub-)objects (M3L schemata)
 - Bounding box
 - Output: gesture lattice with hypotheses

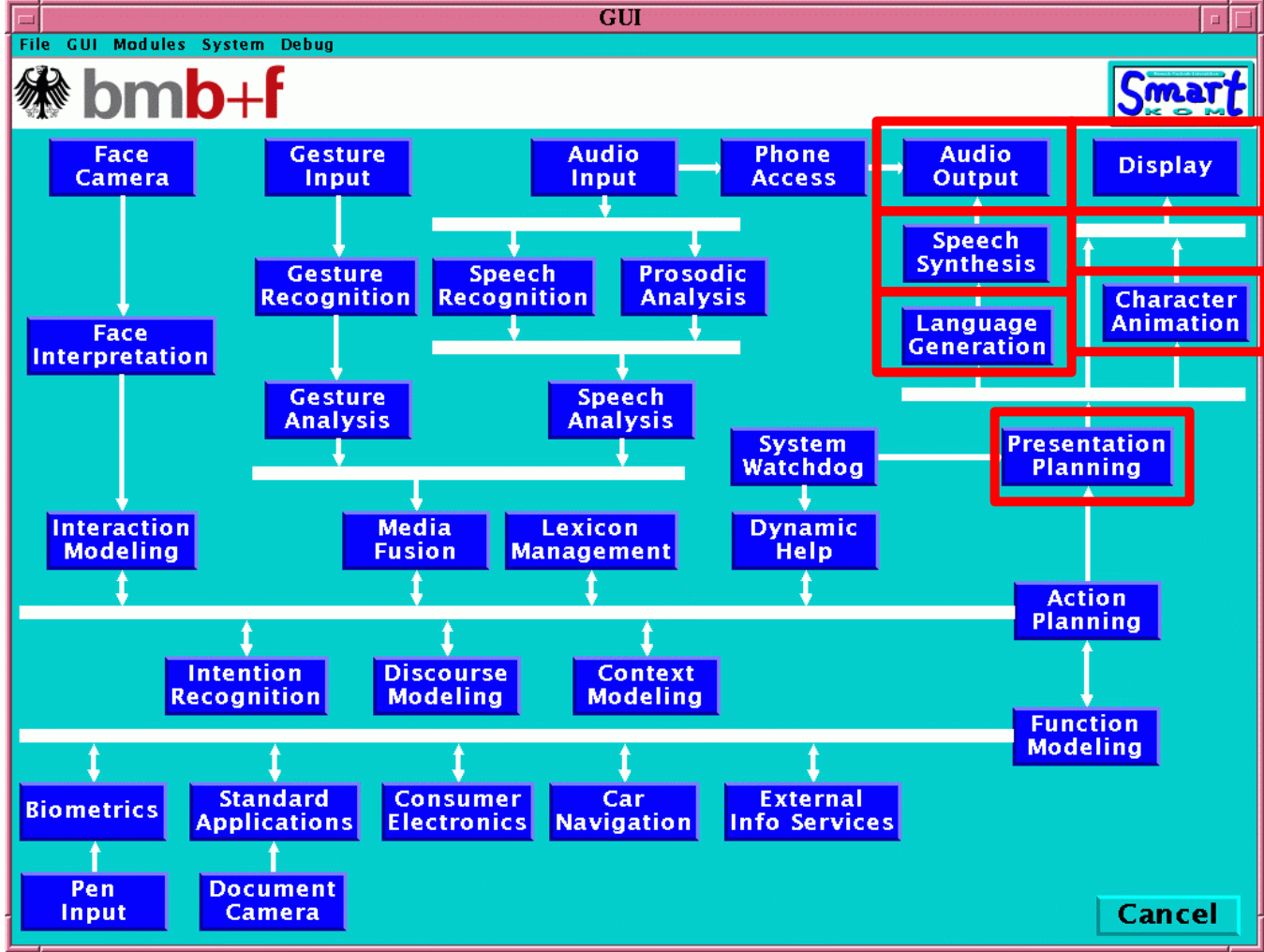


Media Fusion

- **Integrates gesture hypotheses in the intention hypotheses of speech analysis**
- **Information restriction possible from both media**
- **Possible but not necessary correspondence of gestures and placeholders (deictic expressions/ anaphora) in the intention hypothesis**
- **Necessary: Time coordination of gesture and speech information**
- **Time stamps in *ALL* M3L documents!!**
- **Output: sequence of intention hypothesis**



Media Design (Media Fission)



Media Design

- **Starts with action planning**
- **Definition of an abstract presentation goal**
- **Presentation planner:**
 - Selects presentation, style, media, and agent's general behaviour
 - Activates natural language generator which activates the speech synthesis which returns audio data and time-stamped phoneme/viseme sequence
- **Character Animation realizes the agent's behaviour**
- **Synchronized presentation of audio and visual information**



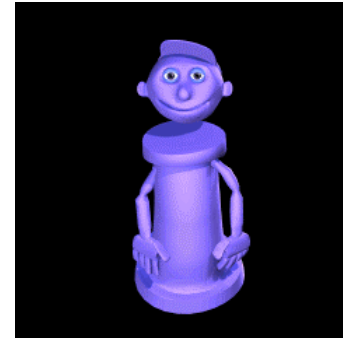
Lip Synchronization with Visemes

- **Goal:** present a speech prompt as natural as possible
- **Viseme:** elementary lip positions
- **Correspondence of visemes and phonemes**
- **Examples:**



Behavioural Schemata

- **Goal: Smartakus is always active to signal the state of the system**
- **Four main states**
 - Wait for user's input
 - User's input
 - Processing
 - System presentation
- **Current body movements**
 - 9 vital, 2 processing, 9 presentation (5 pointing, 2 movements, 2 face/mouth)
 - About 60 basic movements



Conclusion

- **Three implemented systems (Public, Home, Mobile)**
- **Media coordination implemented**
- **„Backbone“ uses declarative knowledge sources and is rather flexible**
- **Lot's remains to be done**
 - Robustness
 - Complex speech expressions
 - Complex gestures (shape and timing)
 - Implementation of all user states
 -
- **Reuse of modules in other contexts, e.g. in MIAMM**

