Finite State Machines

• The system is characterised by explicitly depicting its states as well as the transitions from one state to another.
• One particular state is specified as the initial one
• States and transitions are in a finite number.
• Transitions are triggered by input events.
• Transitions generate outputs.
• FSMs are suitable for modeling control dominated reactive systems (react on inputs with specific outputs; not much computation)

Finite State Machines (cont’d)

Extended FSM

• Variables can be associated to the FSM.
  - Changes of variables can be specified as actions associated to transitions.
  - Actions can also be specified to be performed at the entry and/or exit of a state.
  - Extended FSMs are suitable for systems which are both control and computation intensive.
• Guards (expressed as conditions) may be specified for transitions:
  The transition is performed when the associated event(s) occur and if the associated guard is true

Synchronous FSMs & Synchronous Languages

1. FSM and Extended FSM models
2. The State Explosion Problem
3. Hierarchical Concurrent FSMs
4. Time and Synchrony
5. Synchronous/Reactive Languages
FSM Example-1 (cont'd)

Elevator controller with extended FSM

We associate to the FSM a variable storing the current floor.

- You might wonder: Do we really have one single state of the system? Of course not!

  The global system state is now encoded in the FSM state plus the value of the associated variable.

FSM Example-2

Parking counter

- Input events \(\{\text{in, out}\}\):
  - \(\text{in}\): car enters;
  - \(\text{out}\): car leaves.

- Outputs \(\{1, 2, 3, \ldots \ N\}\):
  - \(i\): display value \(i\)

- States \(\{S_0, S_1, S_2, \ldots S_N\}\):
  - \(S_i\): \(i\) cars in the parking.

FSM Example-2 (cont'd)

Parking counter with extended FSM

We associate to the FSM a variable \(c\) storing the number of cars.

State Explosion

- Complex systems tend to have very large number of states. This particularly is the case in the presence of concurrency.
  The phenomenon is called state explosion.

- Every global state of a concurrent system must be represented individually ⇒ interleaving of independent actions leads to an exponential number of states.

- Expressing such a system as a FSM (or extended FSM) is very difficult.
State Explosion (cont’d)

Example
After starting the system, it waits simultaneously for event a followed by x, and event b followed by y. Events can arrive in any order, except that x follows a and y follows b. Once the events are received, output o is emitted. Then the system waits for the reset signal r to return into the initial state.

- Input events {a, b, x, y, r}
- Output {o}
- States {S₀, S₁, ..., S₈}

Hierarchical Concurrent Finite State Machines

⇐ There are two important mechanisms that reduce the size of an FSM model:
1. Hierarchy
2. Concurrency

Important
- Using Hierarchy and concurrency we only reduce the size of the graphical or textual model; the intrinsic complexity - the number of states of the actual system - cannot be reduced.
- However, the difficulty of realising the model can be drastically reduced.

Hierarchical Concurrent Finite State Machines (cont’d)

Hierarchy
- A single state S can represent an enclosed state machine F:
  Being in state S means that state machine F is active ⇒ the system is in one of the states of the state machine F (or states).

Concurrency
- Two or more state machines are viewed as being simultaneously active ⇒ the system is in one state of each parallel state machine simultaneously (and states).
Hierarchical Concurrent Finite State Machines (cont'd)

Statecharts is a graphical language for hierarchical concurrent FSMs

- System enters state Y ⇒ it will be in both A and B.
- A consists of D and C; C is initial state. D consists of E and F; E is initial state for D.
- B consists of G, I, and H; H is initial state for B.
- Entering Y, the system will be simultaneously in C and H; event a occurs ⇒ system transfers to E and I; event c occurs ⇒ system transfers to F; generates event x which triggers transition from I to G.

The example from slide 7/8 using Statecharts:

FSMs: Time and Sync

- The synchrony hypothesis:
  The time is a sequence of instants (clock ticks) between which nothing interesting occurs. In each instant, some events (inputs) occur in the environment, and a reaction (output) is computed instantly by the modelled design.
  - Computation and internal communication (between the FSMs composing the system, like event x in slide 11) take no time (compare to Discrete Event, where components can have arbitrary delays!).
  - Events are either simultaneous (occur at the same clock tick) or one strictly precedes the other (as opposed to dataflow and Petri Nets where we only have a partial order of events).

FSMs: Time and Synchrony (cont'd)
**FSMs: Time and Synchrony (cont’d)**

**Question**
Is the synchronous model sufficiently realistic to be used in practice?

**Answer**
For some applications yes!
It is the case when the following assumption is true:

The reaction time of the system (including internal communication) is negligible compared to the rate of external events.

---

**Why Do We Like Synchronous Models?**

- A set of communicating, concurrent FSMs behaves exactly like one equivalent FSM (see hierarchical concurrent FSM in slide 12 vs. FSM in slide 8).

  Models are deterministic. It is possible to formally reason about models and to formally check certain properties of the model. This is important for certain class of applications (e.g. safety critical systems)

- It is possible to efficiently synthesise (compile) synchronous models to hardware or software.

---

**Why Do We Not Like Synchronous Models?**

- Reasoning, verification and synthesis based on synchronous models are meaningful and correct only as long as:
  1. A completely synchronous implementation of the whole system is possible (the whole system acts similar to one single FSM).
  2. We are sure that for the implemented system the assumption on slide 15 is true.

- Implementing large models as synchronous systems is expensive and often technically impossible.

---

**Zero Delay Loops in Synchronous Models**

Zero delay loops: Introduce similar problems as discussed with Discrete Event models (Fö 5, slides 25 to 28).

- Problems which could occur:
  - Nondeterministic behavior
  - Unstable systems (the output cannot be determined)
Zero Delay Loops in Synchronous Models (cont'd)

☞ The compiler detects the potential of such problems and rejects the model.

☞ We do not have this problem if the FSM is implemented as a Moore machine (output is delayed by one clock cycle with respect to input). But this is not real synchronous FSM semantic!

Synchronous/Reactive Languages

☞ Synchronous/reactive languages describe systems as a set of concurrently executing synchronized activities.
  • Communication is through signals.
  • Signals are either present or absent at a certain tick.
  • The presence of a signal is called an event and possibly carries a value.
  • These language are semantically equivalent to the (extended hierarchical concurrent) FSM model !!!

☞ Esterel is a well known synchronous/reactive language. Every Esterel model can be compiled to an extended FSM.

Esterel Example

The Esterel program corresponding to the problem described on slide 7 and represented as an FSM on slide 8 and in Statecharts on 12:

```esterel
module Example
  input A, X, B, Y, R;
  output O;
loop
  [await A; await X || await B; await Y]
  emit O;
  await R
end loop
end module
```

How to implement a synchronous system?

A synchronous model (concurrent FSMs):

- Signals are propagated instantaneously through the system.
- all FSMs react instantaneously to events.
- No buffering.
How to implement a synchronous system? (cont'd)

**In hardware:**
- System described as single FSM:
  - implementation as a state machine.
- System described as several FSMs:
  - several communicating synchronous state machines or
  - implement the equivalent single (very large) state machine

*But if the system is large:*
- How do you distribute the clock signal on a large chip, in order to keep synchrony?
- If there are several chips, keeping synchrony is even more difficult.

**In software:**
- One single FSM or several FSMs:
  Generate a sequential program which emulates the equivalent state machine.

Problems:
- For large concurrent systems ⇒ state explosion ⇒ difficult to compile large programs.
- It is practically impossible to implement the software on a multiprocessor system (extremely inefficient to keep the global synchrony of such a multiprocessor software).

**In hardware/software:**
- If part of the application is implemented as hardware (ASICs) and another part as software running on one or several microprocessors

Practically impossible to implement a globally synchronous system.
- In addition to problems highlighted before with pure software or hardware implementations, we have the very different behavior of software and hardware in terms of execution and communication timing.

*If the model is impossible (or very difficult and expensive) to implement, there is no use that it is elegant, simple, and can be formally verified. We get a correct verified model but we cannot implement it correctly!*
Summary

- FSMs are the typical state based model. System states are explicitly depicted. As response to input events, an FSM reacts with a transition to a new state and also generates output events.
- FSMs are suitable for modeling control dominated reactive systems.
- FSMs can be extended, in order to enhance the expressive power, by associating variables. Actions related to the variables can be specified, as well as guards on the transitions.
- Complex systems can have an extremely large number of states: state explosion. This is critical in the case of concurrent systems.
- In order to make the models more concise, the mechanisms of hierarchy and concurrency can be used. Languages like Statecharts support such a modelling approach.

Summary (cont’d)

- FSMs are synchronous models. The whole system reacts instantaneously to a set of simultaneous input events. Events are either simultaneous or strictly ordered.
- This is a realistic model for systems where the reaction time is negotiable compared to the rate of external events.
- FSM models are deterministic. Formal reasoning is possible, as well as efficient synthesis.
- Synchronous/Reactive languages (like Easterly) are based on the same semantics as extended synchronous concurrent Fuses.
- Implementing synchronous systems can be done efficiently under certain circumstances. However, it is practically impossible for large systems and, in particular, distributed systems and, even more, hardware/software systems.

Globally Asynchronous Locally Synchronous Systems

1. Globally Asynchronous Locally Synchronous Systems
2. Globally Asynchronous Locally Synchronous System Models

Globally Asynchronous Locally Synchronous Systems

Globally asynchronous and locally synchronous (GALS) models:

- Each FSM individually behaves like a synchronous system → reacts instantaneously on a set of available inputs and generates output.
- The global system is asynchronous → communication time is finite and non-zero; reaction time of each FSM, as viewed by other FSMs is finite and non-zero.
- With global asynchrony, buffering of signals could be needed.
With a GALS model, the set of FSMs is not any more equivalent with a single FSM (as was the case for the synchronous model).

Several nice features are gone:
- With synchronous FSMs we had the nice theoretical background and the possibility of formal verification of the whole system. Not the case with GALS.
- Every implementation of a synchronous FSM model is guaranteed to be functionally equivalent to the initial model and behave exactly and deterministically like the model (in the case we are able to produce an implementation!). Not the case with GALS.

The GALS model is not deterministic, in the sense that its behavior depends on the amount of time taken for a certain communication or transition.

Two different implementations of the same GALS model can behave differently.

A GALS model in which FSM₁ and FSM₂ communicate through a single-slot buffer.

- FSM₁ outputs a signal (writes into the buffer) every 2 ms (we neglect communication time).
- 1. If the reaction time of FSM₂ is 6 ms, every third signal from FSM₁ will be reacted on.
- 2. If we have a faster implementation of FSM₂, with reaction time 2 ms, every signal from FSM₁ will be captured.

Each individual FSM can be still verified and even formal methods can be used.

However, individual correctness of each FSM does not guarantee the correctness of the whole system. The system behaves correctly only if, in addition, certain assumptions regarding the timing of components and of communications are satisfied.

Example on previous slide:
- Each FSM can be functionally verified individually.
- The global system will be correct (no signal is lost) if FSM₂ has a reaction time which is smaller than the production rate of FSM₁.
- Estimation and simulation can be used in order to verify that a certain implementation (like FSM₁ as software on a certain µprocessor, and FSM₂ as an ASIC) satisfies this assumption.
A GALS system is modeled as a network of FSMs:

- Each FSM has a locally synchronous behavior: it executes a transition by producing a single output reaction based on a single, snapshot input assignment in zero time.
- A system has a globally asynchronous behavior: each FSM reads inputs, executes a transition, and produces outputs in a finite amount of time as seen by the rest of the system.

FSMs communicate through signals:

- A signal, in general, carries an event and associated data.
- A signal is communicated between two FSMs via a connection that has an associated input buffer.
- A sender can communicate a signal to several receivers; each receiver buffers the signal in its own input buffer (of a certain size) associated to the connection.
- Communication is asynchronous and has undefined (finite) delays. Each input buffer stores the most recently received events and values.
- In general, the transmitter sends without waiting for the receiver; nothing prevents the transmitter from sending a new event before the last one was consumed and, thus, potentially, overwriting it.

A FSM reacts when at least one event is available on any of its inputs; in this case the FSM

- reads and consumes the available input signal(s);
- identifies the matching transition and performs the corresponding state transition with the associated action set;
- writes the outputs associated to the transition.

The reaction takes a certain, finite, amount of time. After executing a certain transition, the FSM will be ready to react to new inputs (according to the rules above).

Question: When? Immediately, just after it finished the current transition?
Answer: Not necessarily!

When a certain FSM is ready to check inputs and react, depends on the particular scheduling policy used at implementation.

- The scheduling policy has to be considered when checking if the timing of a certain implementation is correct (see Fö 10).
Summary

• For many systems and, in particular, larger distributed hardware/software systems, only GALS is a realistic approach for implementation.

• Some of the nice features of synchronous FSMs are gone in this case. Formal reasoning about the global system is not possible any more.

• GALS systems can be represented as a network of FSMs. A FSM has a locally synchronous behavior; however, each FSM reads inputs, executes a transition, and produces outputs in a finite amount of time as seen by the rest of the system. FSMs communicate through signals in an asynchronous manner.