Embedded multiprocessors

Timo Töyry

Department of Computer Science and Engineering
Aalto University, School of Science
timo.toyry(at)aalto.fi

16.9.2011
Multiprocessors

- Typical traditional multiprocessor systems are symmetrical (SMP)
- Most of embedded multicore processors are heterogenous
- Typical variables in multicore processors:
  - Type and number of processing elements
  - Interconnects between processing elements
  - Distribution and amount of memory among processing elements
Why embedded multiprocessors

- **Requirements**
  - Real-time processing
  - Low energy/power
  - Cost-effectiveness

- **Specialization, like:**
  - Heterogenous multiprocessors
  - Heterogenous memory system
  - Networks-on-chips (NoCs) support heterogenous architecture

- **Flexibility and efficiency**
Multiprocessor design

- Analysis of the workloads
- Platform-independent optimizations make the software comply with restrictions of the embedded system
- Platform design
- Optimization of the software for the platform

Figure: Multiprocessor emebedded system design, (C) Wayne Wolf
Multiprocessor modeling and simulation

- Communicating simulators: Many multiprocessors are systems of communicating simulators
- Parallel computation simulators: They exploit parallelism support of the underlying OS and hardware to simulate multiprocessor systems
- Traditional multiprocessor simulators are usually designed for symmetric systems
- SystemC is a C++ framework for transaction-level design/simulation of heterogeneous multiprocessors
Processing elements

- A processing element can be CPU, DSP, fixed function ASIC, etc.
- In design of embedded multiprocessor two questions about processing elements will raise:
  - How many processing elements are needed?
  - What types of processing elements to use?
- Use the intended application as a design guideline:
  - Analyze the applications to determine performance and power requirements for each process of the application
  - Select suitable processor type for each process
  - Determine which processes can share a processor to determine the number of required processing elements
Interconnect networks

- Interconnect networks - Terms and metrics
- Network models
- Network topologies
- Routing and flow control
- Network-on-Chips
Interconnect networks - Terms and metrics

Terms

▶ Client
▶ Port
▶ Link
▶ Half-/full-duplex
▶ Topology

Metrics

▶ Throughput
▶ Latency
▶ Energy consumption
▶ Area
Network models

- Used to evaluate networks
- Simple generic model of network have three parts: source, line and termination
- Logic is on source and termination parts, line does not contain any logic
- No multipacket buffers on source and termination
- Some network characteristics can be measured directly from link
  - Throughput and latency
  - Transmission energy consumption, typically some circuit models are used for this
  - Physical length length of the link, longer link slower transmissions and higher energy consumption
Network models cont.

- Traffic models
  - Characterize the behaviour of traffic over the network
  - Poisson model is widely used model for traffic in telecommunications and other systems in which the interesting value is events per unit of time
  - Streams: The data is produced periodically. Described by rate and burstiness.
Network topologies

- **Bus**
  - Simplest interconnect
  - A set of wires that connect all clients
  - Area required by bus is smaller than other interconnects
  - Only one client can transfer data to another client
  - Performance not as good than with some other interconnects
  - Possibly higher energy consumption

- **Crossbar**
  - Fully connected network
  - Any input can be connected to any output
  - Supports broadcast and multicast
  - Complex structure require a large area from chip

- **Buffered Crossbar**
  - Scaled down crossbar to save space
  - Have several queues feeding data to the crossbar
  - The queues are fed from multiple sources of traffic
Network topologies cont.

- Clos network
  - “Low-cost” crossbar
  - Constructed by interconnecting smaller crossbar into multiple stages
  - Offers non-blocking point-to-point connections like full crossbar
  - Blocking multicast
  - Smaller size than full crossbar

- Mesh network
  - All nodes are connected to their neighbours
  - Balances between connectivity and link cost
  - Rich set of connections and multiple paths for data
  - Area required is somewhere between bus and smallest size crossbar variants

- Application specific
  - No specific topology
  - Adapted to the application
Routing and flow control

Routing

- Determines the path for a packet when it travels from its source to its destination
- Routing algorithm can be deterministic to adaptive
- Deterministic routing algorithms base its routing decisions on static characteristics of the network
- Adaptive algorithms adjust routes based on current condition of the network
- Connection-oriented of connectionless
- Guaranteed packet delivery or best effort

Flow control

- How links and buffers are allocated along the packets path through the network
Network-on-Chips

- Currently a lot of research on going on NoCs
- Not in many commercial products yet
- Important design factor is quality of service (QoS)
Memory systems

- Models for system memory
- Traditional parallel memory systems
- Heterogeneous memory systems
- Consistent parallel memory systems
Models for system memory

- Same models can be used as for cache memory simulation
- Delay is nonlinear function of memory size
- Delay and energy consumption is nonlinear function of the number of ports
Traditional parallel memory systems

- All processing elements have access to all memory
- Memory is installed in banks
- Each of N memory banks can be accessed independently
- Peak access rate when all banks are accessed in parallel
- Average access rate is lower, since it is hard to keep memory bus full all the time
Heterogenous memory systems

- All processing elements do not have access to all memory
- Heterogenous memory systems have some features useful in embedded systems:
  - They improve real-time performance
  - They decrease energy and power consumption
Consistent parallel memory systems

- The consistency of system memory is usually handled in software. If a program is executed on multiple processors shared variables must be explicitly protected with synchronization primitives.
- Snooping cache is a widely used technique to keep processors local caches consistent with system memory in multiprocessor environment.
- Memory consistency is not very big problem for multiprocessor embedded systems.
Multiprocessor architectures

- MPSoC (Multiprocessor system-on-chip) most common architecture in current mobile embedded devices
- MPSoC are usually heterogenous
- Common MPSoCs: Texas instrument’s OMAP, Nvidia’s Tegra