IMPACT OF HALF-DUPLEX AND FULL-DUPLEX DMA IMPLEMENTATIONS ON NoC PERFORMANCE

Francesca Palumbo, Danilo Pani, Alessandro Pilia and Luigi Raffo

EOLAB - Microelectronics Lab
DIEE - Dept. of Electrical and Electronic Eng.
University of Cagliari - ITALY
OUTLINE

- Scenario: migration to multi-processors architectures and communication issues
- The Higher Level Modules Impact Investigation: the targeted problem and the adopted architecture
- Deadlock Discussion and Experimental Results
- Final Remarks
DIGITAL DEVICES ARE ALL AROUND US ...

TRANSPORTATION

CONSUMER ELECTRONICS

MEDICAL SYSTEMS

DOMOTICS

MILITARY APPS

Francesca Palumbo, PhD – Dept. Of Electrical and Electronics Eng. – University of Cagliari (Italy)
TOWARDS MULTI-PROCESSORS ERA - 1

- Modern software applications overcome the limits of standard architectures:
  - computational hungry;
  - unpredictable workloads fluctuations.
- Migration towards architectures able to provide:
  - adequate computational support;
  - discrete flexibility level.
A common approach to withstand standard architectural limits and to provide appropriate computational support is to develop more complex systems integrating more cores.

**PARALLEL ACCESS TO THE RESOURCES**

**SCALABILITY**

NOCS 2010 – Grenoble (France) 3-6 May

Francesca Palumbo, PhD – Dept. Of Electrical and Electronics Eng. – University of Cagliari (Italy)
COMMUNICATION ISSUES

- Common MPPs workloads are poorly served by classical packed switched NoC (Princeton University).

- Hybrid switching commutation techniques:
  - represent a solution to withstand the heterogeneity of the traffic, providing the appropriate level of QoS;
  - require to fulfil more stringent requirements.

- MPPs targeting multithreaded applications:
  - have been designed to foster computational performance;
  - processors have to be relieved of communication management.

- Direct Memory Access (DMA) engines:
  - handle memory accesses;
• Scenario: migration to multi-processors architectures and communication issues
• The Higher Level Modules Impact Investigation: the targeted problem and the adopted architecture
• Deadlock Discussion and Experimental Results
• Final Remarks
The performance of the modules in the Network Layer can be influenced by those of the components in the System/Transport layers.

Hybrid Switching NoC: Packet Switching and Circuit Switching handled in parallel over different links.

The composition of the processing core, the local memory, the DMA module, and the NoC modular elements (Network Interface, router) could lead to deceptive conclusions improving and testing NoC performance and functionalities stand alone, since the higher level communication layers components impact can not be overseen in this way.

HALF-DUPLEX DMA

**Blocking:** it can not transmit and receive data to/from the NoC in parallel.

The Half-Duplex DMA is composed of:

- one FSM handling all the control to dispatch/receive data;
- a unique Memory Mapped Register (MMR) bank storing all the control information to handle the communication;
- one Address Generator, controlled by the FSM, to properly access the memory;
- one down counter, programmed as soon as a transaction occurs, to access the proper location of the MMR bank.
**FULL-DUPEX DMA**

*Non-blocking*: it can transmit and receive data to/from the NoC in parallel.

The Full-Duplex DMA is composed of:

- two different FSM to handle respectively data dispatch and data receive;
- two different MMR banks storing all the control information to handle the communication in both ways;
- two different Address Generator, controlled by the proper FSM, to properly access the memory;
- two different down counter to respectively access the proper location of the related *MMR* bank, during send and receive;
- some shared logic to handle coordination.
• Scenario: migration to multi-processors architectures and communication issues
• The Higher Level Modules Impact Investigation: the targeted problem and the adopted architecture
• Deadlock Discussion and Experimental Results
• Final Remarks
• Depending on the DMA adopted, hybrid switching can cause the onset of deadlocks.
• Even if the architecture is able:
  – to meet the *consumption assumption*: all the messages forwarded through the routers can be accepted and delivered if the destination node is able to consume them (Song et al.)
  – to provide separated NI buffers per message type, which is a necessary but not sufficient condition to avoid deadlocks (Song et al.)
  – to guarantee that the NoC and IP dependency graphs are cycle-free in isolation (Dally et al.)

due to protocol interactions between the NI and the higher level communication layers deadlock could still arise.
DEADLOCK ANALYSIS - 2

T_12 – T_02 circuit
T_02 - T_00 denied circuit
(T_01, T_10, T_11) – T_00 lower priority packets.

The Half-Duplex DMA is blocking then lower priority packets will not be consumed by T_00.

Such a situation can not take place adopting a Full-Duplex non-blocking DMA.

Considering a common wormhole strategy, the ACK from T_02 to T_00, might be stalled indefinitely by not digested lower priority packets.

Note: Half-Duplex DMA can still be adopted but require to implement additional deadlock lower priority packets avoidance strategy.
SIMULATION ENVIRONMENT

- 8 x 8 2-D mesh
- SysCgrid Simulation Framework enables multi-parametric HDL cycle-accurate simulations in parallel over a grid. It allows to change:
  - spatial configuration of the allocated resources
  - define the traffic to be injected: e.g., volume of the circuits and packets, temporal distribution of the injected traffic items…
  - architectural parameters

**PERFORMANCE – INJECTION DELAY**

**Injection Delay**: time that lasts between the theoretical scheduled time for a packet to be injected and its actual injection time during execution.

**Full-Duplex non-blocking DMA**: better absolute values, predictability and robustness to traffic heterogeneity.
Full-Duplex non-blocking DMA: besides the better absolute values, still more robustness is appreciable.
Total Transmission Time: ensemble of statistics related to the time necessary to each packet to go from source to the destination.

Full-Duplex non-blocking DMA: better absolute values, more predictability and more robustness.
OUTLINE

- Scenario: migration to multi-processors architectures and communication issues
- The Higher Level Modules Impact Investigation: the targeted problem and the adopted architecture
- Deadlock Discussion and Experimental Results
- Final Remarks
CONCLUSIONS

• This research demonstrated the effect of two different DMA models, Half-Duplex and Full-Duplex, on the performance of a hybrid switching NoC.

• Our analysis demonstrated that the Full-Duplex DMA is able to better sustain the underlying communication medium:
  – Helping in preventing the onset of deadlock
  – Providing better performance in terms of simulation results.

• It has been demonstrated also that optimal NoC performance can not be achieved overlooking the effects that, in an ISO OSI stack representation of the end-to-end communication process, the higher level communication layers can have. Therefore, stand-alone NoC testing could lead to draw conclusions that could not hold in general.
IMPACT OF HALF-DUPLEX AND FULL-DUPLEX DMA IMPLEMENTATIONS ON NoC PERFORMANCE

Francesca Palumbo, PhD
francesca.palumbo@diee.unica.it

EOLAB - Microelectronics Lab
DIEE - Dept. of Electrical and Electronic Eng.
University of Cagliari - ITALY