Switched Interconnect for Embedded System-on-Chip Signal Processing Designs

Daniel Wiklund
Dept. of Physics and Measurement Technology
Linköpings Universitet
Sweden

Abstract—**The upcoming systems for communication places new demands on the performance of the signal processing elements. With the increase in computation comes a demand for increase in communication between processing cores and memories that require a new approach for the on-chip bus system. This paper discusses the structure and design approach for such a system-on-a-chip with emphasis on the interconnect architecture and design issues.**

I. Introduction

As the feature sizes of the manufacturing processes is constantly shrinking, the possibility and demand for more functionality on a single chip goes up. This can lead to many problems, e.g. as the memory access bandwidth through the bus gets to low to cope with the demand.

When more and more functionality is to be crammed onto a single chip it is often done in the classical way by using multi-drop arbited buses to connect the processing elements, both to each other and to the memory controller. This has become inadequate for high end applications primarily in communications where the demands on processing and communication capacity has risen beyond the upper limit of the bus designs. One common way to alleviate this problem is to use several on-chip point-to-point links that constrain flexibility and reusability.

Our proposed architecture instead makes use of a two dimensional switched interconnect structure to increase the processing core to memory bandwidth and connectivity to the desirable level.

In typical communication applications the most critical aspect of all systems is latency. Therefore the internal communication and processing must be fast and relatively predictable, thus disallowing the general use of cache memories.

II. Platform architecture

A natural extension to the bus-based design is to make the bus more advanced and with higher performance, not only in terms of bus clock speed, but also in connectivity, compound bandwidth, maximum usage, and the number of concurrent data streams. This is very similar to the step taken in parallel computer design many years ago where the processing units are connected according to some network topology of higher dimension than one. This will lead to higher performance in core-to-core and core-to-memory communications. A generalized view of this can be found in figure 1.

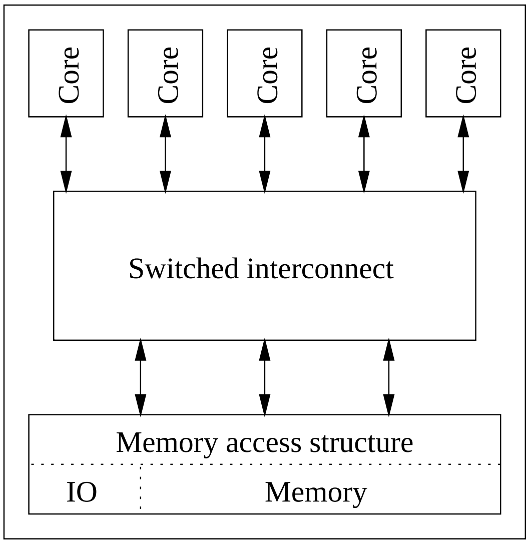

Fig. 1. Typical platform using switched interconnect

In current technology the main memory would typically be off-chip, but this is changing towards more on-chip memory for future processes.

Typical applications for this platform would be a VoIP gateway, baseband processing in a cellular phone base station, or a future MPEG terminal.

III. Interconnect considerations

The thing that constrains the performance most of a old fashion bus is the shared media in combination with the global arbitration that is commonly used to resolve the bus conflicts that can occur when two or more masters want to use the bus at the same time.

By moving to two dimensions we simultaneously get rid of both of those problems. The media is not globally shared and arbitration is only needed for local conflicts in the switching nodes and core connections (wrappers).

A more thorough description of the architecture considerations can be found in an earlier publication [1].

Daniel Wiklund is a Ph.D. student at the Electronic Devices group, Dept. of Physics and Measurement Technology, Linköping University, Sweden. E-mail: danwi@ifm.liu.se

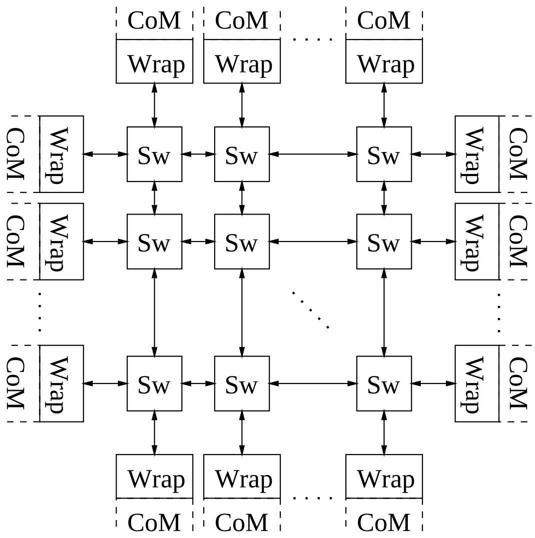

Fig. 2. Switched interconnect architecture with wrappers to encapsulate the cores. (CoM=Core or memory port)

IV. Interconnect architecture

Four principal parts make up the interconnect system (network). These are the switches, links, wrappers, and control. The switches include the routing controller and an arbiter for resolving local route conflicts. The routing algorithm currently under consideration can be labeled pseudo-dynamic since it allows only for restricted dynamic routing in case of switch conflicts. A fully dynamic routing algorithm instead allows for unlimited routability but is far more complex. A simple but efficient routing algorithm would be a modification of the west-first algorithm [2] with the additional functionality to route around busy switches by allowing "wrong" turns under some circumstances. This will make the routability significantly higher than the pure west-first algorithm while keeping simplicity.

The wrappers handle the differences in protocol and communication parameters between the individual IP cores and the network. By using configurable wrappers a lot of repeated work in design and verification can be avoided when reusing the network for new chip designs.

The control circuit handles the startup and on-the-fly configuration as well as exception handling and monitoring of the network, see section V. The links are simply unidirectional point-to-point links that connect switches and wrappers in the network.

A simulator for the system model of the interconnect network is currently being implemented in C++. The intention of the model/simulator is to answer many of the questions that are currently under consideration regarding routing algorithms, connectivity, capacity, etc. The simulator is designed to allow several different routing algorithms, arbiters, etc. to be evaluated.

V. Configuration and control

The configuration of the interconnect network is typically done on two levels. The first part is done during chip design where hardware related parameters like port widths and clocking scheme are configured. The other part is the on-the-fly configuration which sets up dynamic (software) parameters like fixed routes.

The control tasks in the network include monitoring the function of the network as well as handling the exceptions that may arise while the system is running. Exceptions can, for example, be due to an error that cause some communication hogging a route for an extremely long time.

VI. Related work

The concept of switched networks for on-chip solutions has surfaced quite recently. Because of this there is not very much work available on these kinds of systems. One of the few is the work by Guerrier and Greiner [3]. They have reached a high bandwidth, strong connectivity, and reasonable silicon area. The drawback is that they use a packet switched fat tree topology that has the inherent drawbacks of a quite high average latency and a probability to get prohibitively high latencies for some packets. The maximum usage is in practice also limited to approximately 40%-50%. This means that their solution is not very suitable for chips in communication systems.

VII. Conclusions

As can be seen from the previously presented work [1] it is clear that a switched interconnect structure will be feasible and that it will have a high performance.

VIII. Future work

Work is currently done on the C++ simulator. The results from this simulator will be used as a guide for a future implementation of the switched interconnect system. Further on, the wrappers will be investigated more thoroughly to examine the level of configurability.

IX. Acknowledgment

This work is being supported by the Swedish Foundation for Strategic Research (SSF) through the Intelect program and is supervised by associate professor Dake Liu.

References

[1] Daniel Wiklund and Dake Liu, "Switched interconnect for system-on-a-chip designs," in *Proc. of the IP 2000 System-on-Chip for a connected world conference*, 2000, pp. 187–192.

[2] David E. Culler, Jaswinder Pal Singh, and Anoop Gupta, *Parallel computer architecture, A hardware/software approach,* Morgan Kaufmann Publishers Inc., 1999.

[3] P Guerrier and A Greiner, "A generic architecture for on-chip packet-switched interconnections," in *Proc of the Design, Automation and Test in Europe Conference and Exhibition*, 2000, pp. 250–256.