SoC interconnect architecture considerations


The SoC interconnect architecture has a huge impact on what a given SoC can deliver. This is a huge topic of interest for the SoC designers. This is hardly surprising, with the SoC designs nowadays getting more and more communication-centric. Most of the SoCs nowadays consist of multiple processors, hardware accelerators for specific tasks, on-chip memories, several standard interfaces to connect to real world devices and custom Intellectual Property (IP) blocks.

Given that many SoCs nowadays are comprised of a plethora of such blocks, the interconnect architecture can be a key differentiator among the SoCs. Traditional communication architectures were bus-based architectures with a central crossbar responsible for arbitration among the various masters connected on the bus and the protocol specific communication between the masters and slaves on the bus. However, given the number of increasing IP blocks that need to be integrated on an SoC, there has been a shift towards Network-On-Chip (NoC) communication architectures. These NoC architectures use the packet switched network concepts for interaction among the various initiators and targets.

Application specific NoC architectures are needed to support applications that require high performance and low power consumption. Complex applications, such as in-car navigation systems combined with dual rear seat entertainment systems, in-car forward camera analytics running with a number of vision algorithms, Ethernet surround view application, etc, are now implemented in consumer and automotive electronics, which must provide high performance with low energy consumption. Optimizing for these applications early in the design flow can potentially yield better optimization results. The NoC architecture is typically customized for each application or a limited range of target applications, in order to achieve optimal performance and power trade-off. Figure 1 below depicts a typical L3 SoC interconnect.
While designing a NoC interconnect, this brief list of considerations will serve to guide the SoC interconnect architect:

- Interconnect rationale
- Topology selection/organization of arbitration blocks
- Connectivity matrix considerations
- Bus protocol considerations
- Clocking and power considerations
- Physical design considerations
- Quality of service (QoS) considerations
- Micro-architectural considerations
- Cache coherency considerations
- Functional safety considerations
- Security considerations

**Interconnect rationale**

Having a clearly defined interconnect rationale serves as a guiding principle to the SoC architect and enables the architect to choose the appropriate topology that is best suited for the interconnect.

The interconnect rationale can potentially be categorized as follows:

- High throughput concurrency goals
- Non-blocking objectives/real time objectives
- Low latency objectives

**Topology selection/organization of arbitration blocks**

One of the first steps in designing an interconnect architecture is the topology selection. In the application-specific scenarios, an approach based on customization can potentially yield better area/performance/power trade-offs.

Broadly speaking, there are two kinds of topologies to choose from. A unified initiator-target switch based micro architecture, where the initiators and targets connect to the same single switch. This
topology avoids multiple switch hops within the SoC interconnect and as a result, ends up with less latencies within the interconnect. Figure 2 represents an example of such micro architecture.

![Unified initiator-target switch based micro architecture.](image)

Figure 2 Unified initiator-target switch based micro architecture.

In a separate initiator and target switch based micro-architecture, the initiators connect to one switch and the targets connect to another switch. These two switches are then interconnected using one or more links depending upon the application needs. Though the additional switches come with an area/performance penalty vs the unified switch approach, they aid in terms of avoiding loops/multiple possible route generations within the interconnect and thus, help simplify the micro architecture. Figure 3 represents an example of such micro architecture.
Additionally, one could use both approaches in a NoC micro-architecture, depending upon the specific needs of the targeted application(s).

**Connectivity matrix considerations**

A typical interconnect comprises of several initiators and targets e.g. an interconnect for an SoC catering to the automotive infotainment space comprised of 61 initiators and 55 targets on the main interconnect and around 200 targets on the peripheral interconnects. Specifying/reviewing the connectivity manually across each individual initiator and target pair is extremely challenging and prone to errors. It has been found that it may be helpful to classify several initiators and targets across multiple categories and then establish connectivity across these categories.

Here are some example categories for initiators:
- Processors/host
- Video/graphics
- General purpose DMA, etc.
- IO DMA etc.
- Debug/test

Here are some example categories for targets:
- Memories
- General CFG space
- IO targets
- Security
- Debug/test

**Bus protocol considerations**
A typical SoC may consist of several IPs sourced from multiple vendors. The IPs may use various bus protocols like AXI, AHB, OCP or custom protocols. SoC Interconnect needs to either provide the protocol conversion features or comprehend the existence of protocol converters outside the interconnect fabric.

**Clocking and power considerations**

Depending upon the target application requirements, some parts of the interconnect may be clocked differently, so the debug sub-system could be clocked differently to be able to independently disable/clock gate this sub-system in an end production system. Similarly, the interconnect could be partitioned such that one could power on only a subset of the interconnect for a targeted application, thus providing power savings. **Physical design considerations**

**Physical design considerations**

Several peripheral interconnects (typically called L4 interconnects) might need to be architected/instantiated and appropriate care must be chosen while partitioning these targets into various next level interconnects. The architect needs to factor in the inputs from the physical design team in terms of various sub-chips that are being planned as part of the overall SoC design and development. For example, all IO related IPs could be instantiated in one peripheral interconnect instance and all non-IO ones on a separate instance.

**Quality of service (QoS) considerations**

The architect needs to ensure that appropriate QoS schemes are chosen that work in tandem to maximize the performance of the device and fulfill the needs of the intended use case scenarios. Typically, there are certain IPs that are capable of dynamically driving the required priority, e.g. IPs that meet average bandwidth requirement could fail to meet their peak bandwidth requirement and as a result could need dynamic priority escalation and de-escalation. Additionally, there could be programmable monitors instantiated within the interconnect switch fabric to guarantee QoS to the various initiators involved in an application use case scenario. Furthermore, there could be provisions built into the device configuration to allow users to statically configure the priority for each initiator in an end system. There are certain sideband signals that can influence the arbitration scheme in the interconnect and in the external memory controller - where most of the QoS challenges exist. For a system with multiple ports to external memory, appropriate care must be taken to ensure load balancing on the multiple ports based upon the targeted applications.

**Micro architectural considerations**

Among several things, the main micro architectural considerations are:

- Rate adaptation
- Performance monitoring and control
- Identification of parallel paths for low latency requirements and non-blocking requirements

One of the key micro architectural considerations from a performance standpoint is rate adaptation. Rate adaptation becomes necessary wherever there is a width conversion or a clock conversion happening on any link. Appropriate storage needs to be factored in to avoid blocking the link due to the packets that arrive at much higher speeds than what the endpoints can consume.

Another key consideration is to identify the location of QoS knobs that enable performance monitoring and control to guarantee QoS to a specific initiator. Such components ensure that one
particular initiator does not end up consuming too much bandwidth of a link, or a target, that is shared between several data flows. Also, such knobs ensure that each initiator in the use case is guaranteed the required bandwidth needed by that initiator.

Additionally, in order to achieve low latencies during concurrent traffic scenarios, it may be desirable to have parallel paths dedicated to certain targets across certain identified switches.

**Cache coherency considerations**

Multicore SoCs use cache coherency to establish data concurrency across the cores. Coherency can be established by using a directory-based approach where a directory is maintained for every cache line that is fetched from main memory, directory-based interconnect snoops the core cache and provides up-to-date data to requesting core, or a broadcast interconnect approach where interconnect broadcasts cacheable accesses to cores providing coherency, then provides up-to-date data to requesting core.

There are area, power, and design trade-offs that need to be considered for either of these approaches.

**Functional safety considerations**

SoCs that need to fulfill functional safety requirements may need to implement either single bit redundancy, multi bit redundancy or complete hardware redundancy depending upon the Automotive Safety Integrity Level (ASIL) requirements for the safety critical applications targeted for the SoC e.g. several safety critical ADAS applications may require the SoC interconnect to implement appropriate diagnostic measures to detect failures in information transfer. In addition to the techniques mentioned above, such applications may additionally require features like timeout implementation, information redundancy, and transmission redundancy.

**Security considerations**

SoCs requiring security features may need to implement appropriate security firewalls (security core) for each target and provide the users a programmable configurability to define target memory space regions with distinct permissions for different initiators into each of those target regions. Additionally, support for appropriate side-band and/or in-band signaling may need to be factored in as part of the interconnect micro architecture in order to support several security requirements.

**References**

2. ISO 26262 International Standard – Road Vehicles – Functional Safety

More about authors Rahul Gulati, Prashant Karandikar, Vasant Easwaran, Prithvi Shankar, and Mihir Mody.

Also see:

- Slideshow: Nordic engineers explain new wireless SoC architecture
- Design planning for large SoC implementation at 40nm: Guaranteeing predictable schedule and
first-pass silicon success