Article

Double-Layer Energy Efficient Synchronous-Asynchronous Circuit-Switched NoC

Sandy A. Wasif 1,*, Salma Hesham 1,2, Diana Goehringer 3, Klaus Hofmann 4 and Mohamed A. Abd El Ghany 1,4

Abstract: A network-on-chip (NoC) offers high performance, flexibility and scalability in communication infrastructure within multi-core platforms. However, NoCs contribute significantly to the overall system’s power consumption. The double-layer energy efficient synchronous-asynchronous circuit-switched NoC (CS-NoC) is proposed to enhance the power utilization. To reduce the dynamic power consumption, single-rail asynchronous protocols are utilized. The two-phase and four-phase encoding algorithms are analyzed to determine the most efficient technique. For the data layer, the two asynchronous protocols reduced the power consumption by 80%, with an increase in latency when compared with the fully synchronous protocol. However, the two-phase single-rail protocol had better performance compared with the four-phase protocol by 38%, with the same power consumption and a slight increase in area of 5%. Based on this conducted analysis, the asynchronous two-phase layer had significant power reduction yet operated at a moderate frequency. Therefore, the proposed NoC is divided into two data transfer layers with a single control layer. The data transfer layers are designed using synchronous and asynchronous protocols. The synchronous layer is designated to high-frequency loads, and the asynchronous layer is confined to low-frequency loads. The switching between the layers creates a trade-off between the maximum allowed frequency and the power consumption. The proposed NoC reduces the overall power consumption by 23% when compared with recent previous work. The NoC maintains the same system performance with an 8% area increase over the fully synchronous double-layer in the literature.

Keywords: circuit switching; NoC; asynchronous designs; power consumption; power gating; dark silicon

1. Introduction

Innovative designs and novel fabrication approaches allowed for a decrease in transistor size, reaching nanometer dimensions. This permitted the integration of billions of transistors into a single chip. For example, the Apple M1 processor contains 16 billion transistors. These large systems are divided into multiple cores, reaching a thousand cores per chip [1], which are labeled as multi-processor system-on-chip (MPSOC). The on-chip communication for these large systems creates the need for efficient and competent protocols. The network on chip (NoC) concept was introduced to address the shortcomings of conventional protocols. NoCs are reconfigurable switches consisting of physical links, routers and network interfaces [2]. NoCs introduce an independent layer solely responsible for connecting different intellectual properties (IPs) inside the system. NoCs grant reliable communication, flexibility, scalability and good performance [3].
NoCs absorb a considerable amount of the chip’s power consumption [4]. There were many efforts dedicated to introducing efficient NoC designs in terms of power dissipation. Several tracks were utilized to minimize both the static and dynamic power consumption. The dynamic voltage and frequency scaling (DVFS) technique was incorporated for systems characterized with varying load frequencies to diminish dynamic power consumption [5]. The power gating approach was employed for minimizing static power consumption [6, 7]. For synchronous systems, clock gating was introduced to lessen the dynamic power consumption [8]. The globally asynchronous locally synchronous (GALS) design paradigm was established specifically for system-on-chip (SoC) [9]. It addresses the clock skew obstacle, reduces the overall power consumption and maintains high system performance. GALS systems are considered the incentive for the rise of asynchronous designs specifically for communication purposes [9, 10].

Another obstacle that arises for large systems in a small area is the limitations on the allowed thermal dissipation. As a result of the thermal design power (TDP) constraints, chips are incapable of operating at full capacity [11]. These constraints are becoming more dominant with the increasing transistor density in the new generations of technology. If these constraints are violated, it could have an adverse impact on the chip functionality and lifetime. To overcome this challenge, parts of the chip must operate at a lower frequency (dim) or remain inactive (dark) to minimize thermal dissipation. This phenomenon is known as “dark silicon”, and recent researchers are attempting to exploit it for further power reduction [12].

In this paper, the intent is to propose an efficient NoC design that reduces the overall power consumption while maintaining the performance. The proposed NoC is illustrated in Figure 1, with a single control layer and double data transfer layers. This work is an extension of the previous installment in [13], where different asynchronous protocols were explored. The contributions of this work can be presented as follows:

- Proposal of a novel design for the CS-NoC with an asynchronous data transfer layer and synchronous control layer to reduce dynamic power consumption;
- Proposal of a double-layer NoC that leverages the dark silicon phenomenon to maintain the TDP constraints along with the power efficiency;
- Specialized integration of power aware simulation flow and customized asynchronous synthesis flow for optimum evaluation of the proposed design.

![Figure 1](image-url)
This paper is divided into five main sections. Section 1 is the introduction for the work and contributions. Section 2 covers the recent related research work in the area. It highlights the different tracks utilized to reach power optimization. Section 3 covers the proposed asynchronous data subrouter design using two-phase and four-phase protocols. Section 4 covers the proposed double-layer NoC with the flow used for simulation and synthesis. It also includes discussions for the obtained results. Finally, Section 5 is for the conclusions and recommendations for future work.

2. Background and Related Work

This section will cover the basic design concepts utilized in the work. Related recent research work to this field is also discussed in detail.

2.1. Background

Asynchronous designs are divided into two main categories; single-rail and dual-rail [14]. The single-rail category, also known as bundled data, is characterized by separating the control signals from the data signals. This simplifies the design at the expense of added challenges in the synthesis process to guarantee correct functionality [14]. To overcome these challenges, dual-rail protocols merge the data and request signal together. This offers high robustness against temperature and process variations, which is a crucial feature for scaled down transistors. However, this adds to the circuit complexity, potentially reducing performance and efficiency [15]. The two protocols could be realized using two-phase or four-phase encoding algorithms. Four-phase designs exhibit simpler hardware implementation at the expense of reduced performance due to the large number of transitions per cycle. Contrarily, two-phase designs are more complex with higher performance [16].

NoCs could be implemented in different formats and architectures. The most common topologies are 2D mesh, torus, folded torus and hybrid. These designs are characterized by assigning a router for each processing element or IP, referred to as direct topologies [17]. The common switching techniques are packet switching and circuit switching. In the former technique, packets traverse the network through different paths with no reservations. It offers adequate utilization of resources at the expense of complex designs and possible collisions. The latter technique operates in a different manner; the path from the source to the destination is entirely reserved before the initiation of transmission. This allows complete independence between the control logic and data logic, yet it lacks efficient resource usage [18]. Circuit switching offers predictability and fixed latency after path reservation.

2.2. Related Work

Several efforts are being dedicated toward optimizing NoC designs for overall power reduction [5,6,9]. Recently, asynchronous techniques for NoCs gained popularity, and different designs were introduced. For packet-switching NoCs, a fully asynchronous router was implemented in [19] using single-rail protocols. The control logic and data logic were implemented using four-phase and two-phase protocols, respectively. The results showed that the performance was largely dependent on the packet size; as the packet size increased, the synchronous designs became more efficient. Several other researchers added variations in the design to optimize packet switching with asynchronous protocols. These ranged from customized power reduction techniques that fit the packet styles [20] to adaptive routing algorithms that enhance the overall performance [21]. Generally, packet switching lacks predictability, making it a good candidate for asynchronous designs. Circuit-switching algorithms were not as deeply explored with asynchronous designs as for packet switching. GALS designs were also explored in detail with asynchronous designs, since they are mainly synchronous IPs communicated with asynchronous interconnects. In [2], four-phase QDI asynchronous protocols were employed, which showed promising results in terms of performance with added complexity. GALS was also explored in [22], which combined
asynchronous designs with DVFS and power gating for optimum power consumption. All these mentioned references highlight the significant contribution of asynchronous designs in terms of power optimization and maintaining the system’s performance.

Accurately setting constraints to verify the functionality from a timing analysis perspective is a major challenge for single-rail asynchronous protocols. The commercial CAD tools for synthesis flow cater to synchronous designs. Timing analysis and constraints specifically rely heavily on the existence of a clock signal. Two tracks were followed to overcome these challenges. The first one was to define new methods for asynchronous timing analysis through introducing novel specified languages. The second track was to evolve the current synthesis flow to fit asynchronous designs. This track operates with commercial CAD tools and under the normal hardware description languages (HDLs) [23]. For example, the work in [24] aimed at improving the process of identifying and setting relative timing constraints (RTCs) using the available tools. Modifications to both the logical synthesis flow and the physical implementation were implemented for Synopsys tools. These constraints are mainly evaluated by comparing the data path and the control path for single-rail protocols. The results deduce that these modifications ease the synthesis and timing analysis for asynchronous designs, making them more accessible using commercial tools.

Design constraints due to the dark silicon phenomenon were also addressed in several recent research works, ranging from managing it to leveraging from it. In [25], a multi-layer NoC was introduced with power gating, referred to as darkNoC. It can activate only one layer at a time, whereas all the other layers remain dark. It showed a promising energy-delay product with only added area utilization. Another method to overcome the challenge of maintaining TDP constraints is to fully operate but at a lower frequency. In [26], different layers operated at different voltage scales instead of being completely deactivated. This showed a reduction in the overall power consumption with less added area when compared with NoCs implementing deactivation through power gating. In [27], the independence of operation within the circuit-switched NoC enabled operation at various frequencies and supply voltages to further reduce power consumption. Finally, to fully address the thermal dissipation of the chip, the patterns used for the deactivation of chip subparts were analyzed in [12,28]. The work in [12] used folded torus topology with virtual clusters for efficient network building. The algorithm mapped tasks to routers that were virtually close yet physically apart. This ensured efficient performance, with even thermal distribution across the chip and low peak temperatures. These clusters were able to activate at full capacity or low capacity or even completely turn off based on the load. In [28], the target was to optimize the system performance and maintain the overall system temperature within the safe limits. This work introduces a dynamic mapping technique that is efficient for many-core systems, taking into consideration the caches related to each core. However, the dynamic mapping task is not a major concern for circuit-switched NoCs, as the path is preset before the initiation of transmission.

3. Proposed Asynchronous Router Design

This section will cover the phases for implementing the proposed circuit-switched router architecture design with asynchronous protocols. First, the asynchronous design for the data subrouter using two-phase and four-phase protocols is explored. Then, the flow used to simulate and synthesize the design is established. Finally, the results and comparisons among the two protocols are displayed, and a protocol is chosen for the proposed NoC based upon these results.

3.1. Proposed Synchronous-Asynchronous Router

A circuit-switched NoC consists of a data layer and a control layer, as shown in Figure 2. It is characterized by total independence between the data transfer layer and the control layer due to its operating scheme. This independence allows for the use of separate synchronization protocols for each layer without the addition of synchronizers between
them. The NoC’s operation can be outlined in three consecutive actions. The first action is to reserve a path from the source to the destination with the control layer. This is done by sending a control flit that traverses the network, reserving ports at each router within the path. Then, the data layer starts the transmission of data packets through the reserved path. After successful completion of the transmission process, the control layer releases the path using another control flit. The timing diagram shown in Figure 3 is used to illustrate the operation at the interface between the control layer and the data layer. The control layer operates twice per transmission cycle—at the start and the end—to reserve and release the path, respectively. However, the data transfer layer is constantly operating between these two points. This indicates that the data layer is the primary contributor to dynamic power consumption.

![Figure 2](image_url)

**Figure 2.** The router architecture under the circuit-switching technique with a synchronous control layer and an asynchronous data transfer layer.

![Figure 3](image_url)

**Figure 3.** The timing diagram, illustrating the operation of each layer and their sequential relation.

To corroborate this indication, a power analysis was conducted on the basic fully synchronous circuit-switched NoC, and the results are shown in Figure 4. This analysis was conducted for the fully synchronous router with 65-nm technology under the same evaluation specifications mentioned in Section 3.3. The results indicate that the dynamic power consumption for the data layer was significantly larger compared with the control layer. According to the traffic variations, the power consumption percentage would also differ, yet in all realistic cases, the data transfer layer would consume additional power compared with the control layer.
The proposed circuit-switched NoC will combine different synchronization protocols to balance the power efficiency and design complexity. The data layer will follow an asynchronous handshake protocol to reduce dynamic power consumption, since the dynamic power depends on the switching activity and system frequency as shown in Equation (1). It has another benefit, as it operates at the load rate instead of the worst-case rate. The control layer will follow the synchronous protocol to reduce the design complexity, as its contribution to dynamic power consumption is significantly smaller. For the asynchronous design, single-rail encoding techniques were chosen due to their efficiency in resource utilization when compared with dual-rail techniques:

\[ P = A C V^2 F \]  

where \( A \) is the activity factor, \( C \) is the switched capacitance, \( V \) is the supply voltage and \( F \) is the system’s operating frequency.

3.2. Router Implementation

The implemented NoC is configured as a 2D mesh with an XY routing technique using a circuit-switching algorithm. The NoC design is scalable, and its size is represented as \( N \times M \). The router is divided into two main blocks: the control subrouter and the data subrouter. The control subrouter mainly consists of an arbiter, cross bar to route control flits and FSMs for the input and output ports. The control subrouter is designed using a synchronous protocol as the conventional design. The main functionality for the control subrouter is to produce the control signals used to configure the data subrouter. It is also responsible for routing the control flit throughout the network for path reservation or release.

The control subrouter’s operation is synchronized with a clock signal. The arbiter is designated to sort the requests coming from other routers. The requests are granted following a round-robin protocol. The FSMs are used to indicate the state for each input or output port. The main states for the ports are idle, reserved or active. The state for each port depends on several parameters, including (1) the grants from the arbiter based on the order of the accepted requests, (2) the acknowledgment signals coming from the succeeding routers along the path, indicating a successful reservation process, and (3) the control flit carrying information regarding the source and destination addresses. The crossbar is responsible for routing the control flits. The architectural view for the control subrouter is shown in Figure 5.
Figure 5. The internal design for the control subrouter with the arbiter, crossbar, input and output units.

The data subrouter is responsible for routing the data flits based on the control signals. It consists of five input ports and five output ports representing the four main directions and a local port for the IP connection. The design is divided into two blocks: the pipeline stage and the crossbar stage. The pipeline is responsible for completing the handshake protocol between different routers. The crossbar is responsible for the routing functionality using the control signals provided from the control subrouter. The single-rail encoding technique is implemented using four-phase and two-phase protocols.

The main difference in the pipeline stage for the four- and two-phase protocols is in the type of latch used, as shown in Figure 6. The first pipeline design is shown in Figure 6a; it implements a four-phase handshake protocol. Four-phase protocols follow four actions per each transmission process. These actions are performed through two separate signals: the request and acknowledgment signals. Two transitions are dedicated to ensuring that the cycles consistently end with the same polarity (typically zero). This reduces the complexity of the hardware design at the expense of expansion in the cycle time. The pipeline stage consists of a latch and a C-element. The C-element allows data to pass through the stage only if both the request signal of this stage and the acknowledgment signal of the next stage are equal to logic 1. This means that there are data available at the current stage and the next stage is ready to process the data. The latch is a positive latch that allows the transmission of data based on the enable value (output of the C-element).
Figure 5. The internal design for the control subrouter with the arbiter, crossbar, input and output units.

Figure 6. The asynchronous pipeline stage implemented using (a) four-phase and (b) two-phase techniques.

The second pipeline stage follows the two-phase protocol as shown in Figure 6b. It requires two actions to complete the transmission process. This protocol is an enhancement over the four-phase technique. It reduces the number of actions per cycle by disregarding the polarity and eliminating the resetting at the end of the cycle. Contrary to the four-phase protocol, the cycle ends at different polarities. A transition from logic 1 to logic 0 holds the same effect as a transition from logic 0 to logic 1. This enhances the performance at the expense of added complexity to the design. The pipeline stage consists of a C-element and a capture-pass latch. This is a modified latch that operates regardless of the polarity to fit the two-phase methodology. If both the capture and pass signals have transitions, then the latch shall operate regardless of the polarity of this transition. The latch then performs one of two operations, either capturing the data inside it or passing them to the next stage. This modified latch is implemented at every port for each router within the NoC, causing an increase in the overall area. These pipeline stages follow the general designs in the literature as in [16].

The crossbar stage is responsible for routing the data across the network. For asynchronous designs, it is responsible for routing the request and acknowledgment signals as well. The request signals travel along the data in the same direction, whereas the acknowledgment signals traverse the network in the opposite direction. Data routing is completed based on the control signals produced by the control subrouter. This is implemented using a multiplexer with five possible inputs and one output. This multiplexing stage is repeated for each output port (five multiplexers). The routing for the handshake signal contains an extra design element. It integrates C-elements along with the basic multiplexing gates to ensure the stability of the control signals, which is crucial for asynchronous designs. This stage is repeated at every output port to produce the request signal entering the pipeline stage in the forward path and the acknowledgment signal entering the pipeline stage in the backward path. For example, the request signal at each output port is selected from the five possible input ports with the use of the selected lines. The modified multiplexer for routing the control signals is shown in Figure 7.
Interfacing multiple layers with different synchronization protocols requires the addition of synchronizers in between. The ports connecting the data subrouter and the control subrouter are mainly the control signals. These signals are produced during path reservation and remain constant throughout the transmission process. The values stored within these signals do not fluctuate at all. Therefore, there is no need to add extra hardware to act as synchronizers between the two layers. This is another benefit of the decoupling that occurs for circuit-switched NoCs. However, this is only valid for circuit-switched NoCs, packet-switching NoCs or any other switching technique; the design criteria may vary.

3.3. NoC Evaluation

The NoC size was chosen to be $4 \times 4$ with a total of 16 routers and a data packet size of 32 bits. The design was implemented using VHDL. To verify the NoC’s functionality under varying test cases, the traffic was randomized. The traffic was generated from MATLAB to randomly assign the source and destination routers. The constraint added to the randomization was to exclude assigning the same router in the source and destination fields simultaneously. The traffic is extracted in a format compatible with the VHDL test bench. The simulator is used to verify the functionality and extract the SAIF file. This file is used to estimate the switching activity of the different signals within the design for accurate power measurements under the random traffic. The design files along with the SAIF file and constraints were injected into the Synopsys design compiler. The design was synthesized, and the timing constraints were analyzed.

For asynchronous designs, timing analysis is more challenging specifically for single-rail protocols. The data must remain valid throughout the handshake protocol. To ensure this condition, the best-case delay for the control path must be higher than the worst-case delay for the data path. These constraints were added to the compiler tool. If these constraints were verified, matched delay elements were added to the control path. The synthesis process was repeated until all the constraints were attained. Then, the results for the power consumption, timing analysis and area utilization were extracted. The flow for simulation and synthesis is shown in Figure 8.
Electronics 2021, 10, x FOR PEER REVIEW 11 of 20

Figure 8. The flow chart used for the simulation and synthesis flow for the NoC’s evaluation.

3.4. Results and Discussions

The design was compared to the fully synchronous circuit-switched NoC presented in [27]. This work was modified to a single layer instead of two layers and tested under the same traffic for fair comparison. The synthesis process was conducted using 65-nm technology. Based on the results for the timing analysis, the data arrival time for the two-phase protocol was recorded as 0.44 ns. The clock signal period was set to a value within the range of the data arrival time in the asynchronous design for accurate dynamic power comparisons. The frequency for the control subrouter clock was 500 MHz, while the clock frequency in the data subrouter for the comparison design was 200 MHz.

The first results to examine in Figure 9 are the area comparisons among the three protocols. For the 4 × 4 NoC, the two-phase design had the highest occupied area among the three designs. The area for the two-phase design increased by 1.12% when compared with the synchronous design due to the added design complexity. The four-phase protocol had the lowest area among the designs, being 3% lower than the synchronous design. The area reduction could be due to the elimination of clock signals and their associated circuitry. Area comparisons are not crucial for the dark silicon phenomena, since a significant chip area is not fully utilized and the differences between the protocols are not consequential.

Figure 9. The area comparison among the three protocols for circuit-switched NoCs.

The results for the leakage power are shown in Figure 10a. The comparisons demonstrate that the leakage power consumption for both asynchronous protocols was lower than that of the synchronous design. The reduction in power consumption was 7% and 13% for the two-phase and four-phase protocols, respectively, when compared with the
synchronous design. The two-phase protocol had an increase in leakage power consumption over the four-phase protocol due to the added design complexity. The dynamic power consumption under the randomly generated traffic is presented in Figure 10b. The analysis was conducted with the extracted SAIF file for accurate dynamic power measurements. The same traffic was applied for all protocols for fair comparisons. The results showed a significant reduction in the consumed dynamic power for the two asynchronous protocols. The consumption was almost 80% lower than that of the synchronous design.

![Figure 10. Power comparisons among the three protocols under uniform random traffic for a circuit-switched NoC for the (a) leakage power and (b) dynamic power.](image)

Finally, the latency comparisons are shown in Figure 11. This latency was measured as the time taken to complete the transfer of data through a single router using the longest or worst path. The results indicated that the synchronous design was more efficient in the time comparison aspect. The four-phase protocol had an increase in latency of 80%. The two-phase protocol had a latency increase of 70%. This was expected, as the four-phase protocol had more transitions per cycle when compared with the two-phase protocol.

![Figure 11. The latency comparison among the three protocols under uniform random traffic for a circuit-switched NoC.](image)

Analyzing the results showed that the two-phase protocol was superior to the four-phase protocol. The two designs offered the same dynamic power reduction and comparable leakage power reduction. However, the reduction in performance was more significant for the four-phase protocol than the two-phase protocol. The only aspect where the two-
phase protocol underperformed was the area utilization. Nonetheless, the area increase was a very small percentage, and it was not a design parameter of major concern for the design of large systems under TDP constraints. Based on these results, the two-phase protocol was chosen for the implementation of the asynchronous layer in the proposed router architecture.

Table 1 presents a generic comparison with a recent asynchronous router design in [29]. To have a fair comparison, the same traffic, technology and test conditions should be applied. Since this was hard to achieve, the comparison was more of an indication for the status of the proposed work against other research. The table presents the non-scaled comparison in terms of area, latency, power and energy. The area for the proposed work significantly exceeded the one presented in [29] by almost 80%. However, there was a reduction in latency by almost 30% and 50% for the four-phase and two-phase protocols, respectively. The energy per bit was reduced for the two-phase protocol by 20% compared with the one in [29], yet it increased for the four-phase protocol by 19%.

Table 1. Comparison with a recent asynchronous router design.

| Specifications          | TaBuLA [29] | Proposed Two-Phase Protocol | Proposed Four-Phase Protocol |
|------------------------|-------------|-----------------------------|-----------------------------|
| NoC Size               | 5 × 5       | 4 × 4                       | 4 × 4                       |
| Asynchronous Protocol  | Two-Phase Bundled Data | Two-Phase Bundled Data | Four-Phase Bundled Data |
| Technology             | 40 nm       | 65 nm                       | 65 nm                       |
| Flit Size              | 32 Bits     |                             |                             |
| Area (um²)             | 24,866      | 171,399                     | 164,423                     |
| Latency (ns)           | 0.98        | 0.44                        | 0.68                        |
| Energy per Bit (pJ/bit)| 0.12        | 0.0955                      | 0.1484                      |
| Power (mW)             | -           | 6.9459                      | 6.9858                      |

4. Proposed NoC Architecture

This section will cover the proposed double-layer NoC design that leverages the dark silicon phenomena. First, the power gating implementation for obtaining an efficient design is introduced. Then, the idea and implementation for the NoC is illustrated in detail. Finally, the power-aware simulation and synthesis flow are discussed, leading to the obtained results.

4.1. Power Gating Implementation

One of the dominant techniques to reduce the overall power consumption is power gating. It allows the deactivation of any non-utilized circuitry within the system by shutting down the supply. This reduces the static power consumption of the system. However, this allows the system to experience varying power modes at the same time for different components. Extra precautionary measures should be added to maintain the functionality and system performance under the varying operating modes. Power gating could be implemented using fine grain or coarse grain patterns. For this design, the coarse grain was chosen, as the transition time effect was not relatively impactful.

For this work, power gating was applied within the router architecture. To reduce the overall power, any unreserved router should not have been operating at all. Since the control subrouter was responsible for the reservation process, gating was not applied to it, as it should have been on to receive and evaluate requests. However, the data subrouter was only enabled after the reservation process was complete. For this reason, the power gating was mainly applied to the data subrouter. The control signal responsible for activating the
data subrouter was an extra bit cascaded with the input data to the router. The activation was complete only if the router was reserved for the transaction process.

The addition of power gating was accomplished through inserting another layer on top of the VHDL design files. Unified Power Format (UPF) files were responsible for specifying power domains and modes. The router was divided into three power domains: the data domain, the control domain and the router domain that encapsulated the previous two subdomains. The gating was applied to the data domain by inserting power switches and isolation cells. The power switch was added to the net connected with the power supply. The switch was turned on based on the value of the control signal to connect the data domain with the power supply. The isolation cells were added as a precautionary measure to maintain the system’s functionality. These cells isolated the deactivated blocks from the remaining operating system, so they were added to the input and output ports. There was no need to add retention cells as the output of the deactivated blocks would not affect the operating system as a whole. The gated data power domain is shown in Figure 12.

**Figure 12.** The block diagram highlighting the power domain of the data subrouter with power gating implementation.

### 4.2. Double-Layer NoC Architecture

For large systems, the chip operation was limited due to the TDP constraints, so operating at full capacity was no longer applicable. To leverage from the dark silicon phenomena, this work proposed the double-layer NoC. The NoC was divided into two data transfer layers and a single control layer. The data transfer layers were not to function at the same time to avoid TDP constraint violation. The choice for the operating layer was solely based on the frequency of the applied load. One data transfer layer was chosen to follow the conventional synchronous protocols for loads with high frequencies, as the synchronous design had the capability to support high frequencies. The other data layer was implemented as fully asynchronous using the two-phase protocol. The asynchronous layer operated for loads with lower frequencies since the design showed lower performance in the previous section. This offered service for applications with various operating frequencies and overall power reduction as well. The control layer
would remain synchronous to minimize the design complexity, and it would control the reservation for the two data layers.

The layers were activated and deactivated through the use of UPF files as well. The router power domain then contained three sub domains for the three layers. The power switches were added to the data subdomains beside the previous switch used for gating implementation. Isolation cells were also added to ensure correct system functionality under varying power modes. The layers were mutually exclusive, which means that they did not operate at the same time. The signal controlling the switching between different layers was an input signal provided based on the frequency of the input load. A multiplexer was added to select the output data for each router. The same control signal was used as the select line for the multiplexer. The design for the double-layer router is shown in detail in Figure 13.

![Figure 13](image_url)

**Figure 13.** The proposed double-layer circuit-switched NoC with two data transfer layers and a single control layer. Power switches were added to activate or deactivate the corresponding layers.

### 4.3. Power-Aware Simulation Flow

To measure the effect of the power gating implementation, the flow for simulation and synthesis was modified. Power-aware simulations should be utilized to capture accurate power results. These simulations detect the power domains and their relations to the design files. They are able to simulate the design behavior, taking into consideration the ON and OFF power domains and the power activity in general. For the synthesis process, the technology library used was changed to a 32/28 nm package, as it contained power-aware cells that were mapped to represent special gates as the isolation cells. This package was provided by Synopsys for the purpose of research. The tools used in this flow for simulation and synthesis were Questasim, Synopsys DC and Synopsys Prime-Time.

First, the design VHDL files, constraints, UPF files and technology library were injected into Synopsys DC. The synthesis process checked for all the timing constraints to add any necessary delay elements, and a VHDL net-list was produced. The net-list, along with the UPF files and technology library, were injected into the simulator. The simulator performed the verification for the functionality under different power modes. It
also produced the SAIF file under the applied traffic and the switching in power modes for each power domain. Finally, the extracted SAIF file was inserted into Prime-Time with the design net-list, the constraints and the 32/28 nm technology files. In addition, the extracted parasitic and timing relations from Synopsys DC were also provided for accurate results. Prime-Time produced the final results for the utilized area, consumed power and the timing analysis as well. This tool provided more accurate results specifically for the static timing analysis. The flow is illustrated in Figure 14.

![Diagram of power-aware simulation and synthesis flow](image)

**Figure 14.** The power-aware simulation and synthesis flow.

### 4.4. Results and Discussions

The NoC size used for evaluation was 4 × 4 with the 32/28 nm technology library. The data packet size was 32 bits, with one extra bit as a flag for activating the powered-off data subrouters. There were two separate sets of analysis conducted. The first power analysis was dedicated to measuring the effect of power gating. This analysis was conducted with a single-layered NoC with and without power gating. The analysis was conducted using different traffic points and varying load frequencies to examine the impact of power gating. As shown in Figure 15, the layer implemented with power gating had smaller power consumption when compared with the layer implemented without power gating. This effect took place regardless of the traffic case or the frequency of the system load. This highlights that the effect of the added overhead due to the extra hardware used in power gating implementation was smaller than the power saved. Based on this conclusion, the power gating was implemented in the double-layer NoC for power optimization.
After confirming the efficiency of power gating, the proposed NoC was evaluated. The same flow used for power-aware simulation and synthesis was applied for the NoC evaluation. The proposed double-layer NoC with mixed synchronization protocols was compared with the fully synchronous NoC in [27]. The timing analysis showed that the asynchronous layer had an overall timing per router 1.5 times that of the synchronous layer. Based on the timing analysis, the fast synchronous layer was operating at a 1 ns clock period, while the slow asynchronous layer was able to support loads with 2 ns of latency or slower. This performance was comparable to the one presented in the work used for comparison.

The area comparisons, which were indicators of the cost of the design, are presented in Figure 16. As is shown, the proposed design had a slight increase in area of 8%. This increase in area is logical due to the added complexity of the asynchronous design in one of the layers. However, this increase was not significant, and the area was not the impactful aspect when it came to the design specifications for large systems. The power and latency were the major concerns, since the area was not fully utilized under the dark silicon phenomena.

The power analysis was conducted with Prime-Time under the same traffic case as shown in Figure 17. The detailed power consumption in hierarchical form is presented in
Table 2. This indicated that the asynchronous data layer had the smallest power contribution within the overall structure, whereas the synchronous layer had the largest power contribution. This is the tradeoff between the fast layer with high power consumption and the slow layer with efficient power consumption. The overall power consumption of the proposed architecture was compared with the design presented in [27]. The results showed that the overall power consumption for the proposed design was reduced by 23% when compared with the fully synchronous design.

Table 2. Hierarchical power analysis for the proposed double-layer NoC.

| Hierarchy       | Switch Power (W) | Internal Power (W) | Leakage Power (W) | Total Power (W) |
|-----------------|------------------|--------------------|-------------------|-----------------|
| NoC             | 1.28 × 10⁻⁴      | 1.94 × 10⁻³        | 1.01 × 10⁻³       | 3.08 × 10⁻³     |
| Asynchronous    | 3.13 × 10⁻⁵      | 5.61 × 10⁻⁴        | 5.59 × 10⁻⁵       | 6.47 × 10⁻⁵     |
| Synchronous     | 1.22 × 10⁻³      | 1.63 × 10⁻³        | 5.18 × 10⁻⁴       | 2.27 × 10⁻³     |
| Control Layer   | 1.78 × 10⁻⁵      | 3.01 × 10⁻⁴        | 3.38 × 10⁻⁴       | 6.41 × 10⁻⁵     |

5. Conclusions

An efficient double-layer circuit-switched NoC with mixed synchronization protocols was proposed. Analysis was conducted to select the appropriate single-rail asynchronous protocol. The two single-rail schemes offered a significant power reduction of 80% when compared with the fully synchronous approach. The two-phase protocol was chosen as it maintained reasonable performance over the four-phase protocol of 38%. Power reduction techniques were utilized to further reduce the overall power consumption. To leverage the dark silicon phenomena, the NoC was modified to contain two data layers instead of one. The first data layer was fully synchronous for high-frequency loads, and it consumed the largest percentage of power. The second data layer was fully asynchronous for low-frequency loads and efficiency in power consumption. Based on the results, the proposed NoC offered comparable performance with a slight area increase of 8% and reduction in power consumption of 23% over the work in the literature.

In the future, other asynchronous protocols can be utilized to investigate their effect on the performance and power savings. Dual-rail protocols in particular may also be explored to achieve high robustness as well as power reduction. The physical implementation and layout of these designs could be used as an indicator for the overall cost and performance metrics. Finally, the proposal could be tested using different network topologies other than the 2D mesh topology.
**Author Contributions:** Conceptualization, S.A.W. and M.A.A.E.G.; methodology, S.A.W.; resources, K.H.; supervision, S.H., D.G. and M.A.A.E.G.; validation, S.A.W.; writing—original draft, S.A.W.; writing—review and editing, S.H. and M.A.A.E.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Alexander-von-Humboldt Foundation-Research Group Linkage Programme.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Borkar, S. Thousand Core ChipsA Technology Perspective. In Proceedings of the 44th ACM/IEEE Design Automation Conference, San Diego, CA, USA, 4–8 June 2007; pp. 746–749. [CrossRef]
2. Lattard, D.; Beigne, E.; Clermidy, F.; Durand, Y.; Lemaire, R.; Vivet, P.; Berens, F. A Reconfigurable Baseband Platform Based on an Asynchronous Network-on-Chip. *IEEE J. Solid-State Circuits*, **2008**, *43*, 223–235. [CrossRef]
3. Carloni, L.P.; Pande, P.; Xie, Y. Networks-on-chip in emerging interconnect paradigms: Advantages and challenges. In Proceedings of the 3rd ACM/IEEE International Symposium on Networks-on-Chip, La Jolla, CA, USA, 10–13 May 2009; pp. 93–102. [CrossRef]
4. Salihumand, P.; Jain, S.; Jacob, T.; Kumar, S.; Erraguntla, V.; Hoskote, Y.; Vangal, S.; Ruhl, G.; Borkar, N. A 2 Tbps 6 × 4 Mesh Network for a Single-Chip Cloud Computer with DVFS in 45 nm CMOS. *IEEE J. Solid-State Circuits*, **2011**, *46*, 757–766. [CrossRef]
5. Cremona, L.; Fornaciari, W.; Marchese, A.; Zanella, M.; Zoni, D. DENA: A DVFS-Capable Heterogeneous NoC Architecture. In Proceedings of the 2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Bochum, Germany, 3–5 July 2017; pp. 489–494. [CrossRef]
6. Zoni, D.; Canidio, A.; Fornaciari, W.; Englezakis, P.; Nicopoulos, C.; Sazeides, Y. BlackOut: Enabling fine-grained power gating of buffers in Network-on-Chip routers. *J. Parallel Distrib. Comput.*, **2017**, *104*, 130–145. [CrossRef]
7. Zhu, D.; Li, Y.; Chen, L. On Trade-off Between Static and Dynamic Power Consumption in NoC Power Gating. In Proceedings of the 2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), Lausanne, Switzerland, 29–31 July 2019; pp. 1–6. [CrossRef]
8. Chindhu, S.T.; Shanmugasundaram, N. Clock Gating Techniques: An Overview. In Proceedings of the 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, India, 2–3 March 2018; pp. 217–221. [CrossRef]
9. Weber, I.; de Oliveira, L.; Carara, E.; Moraes, F.G. Reducing NoC Energy Consumption Exploring Asynchronous End-to-end GALS Communication. In Proceedings of the 2020 33rd Symposium on Integrated Circuits and Systems Design (SBCCI), Campinas, Brazil, 24–28 August 2020. [CrossRef]
10. Vijayalakshmi, P.; Sathiya, K. An Efficient Prototype for Gals Systems in Asynchronous Network-on-Chips through Multiclocking. *Int. J. Sci. Res.*, **2017**, *6*, 1179–1183. [CrossRef]
11. Shafique, M.; Garg, S.; Mitra, T.; Parameswaran, S.; Henkel, J. Dark silicon as a challenge for hardware/software co-design. In Proceedings of the 2014 International Conference on Hardware/Software Co-design and System Synthesis, New Delhi, India, 12–17 October 2011. [CrossRef]
12. Yang, L.; Liu, W.; Jiang, W.; Li, M.; Chen, P.; Sha, E.H.M. FoToNoC: A folded torus-like network-on-chip based many-core systems-on-chip in the dark silicon era. *IEEE Trans. Parallel Distrib. Syst.*, **2017**, *28*, 1905–1918. [CrossRef]
13. Wasif, S.A.; Hesham, S.; Goehringer, D.; Hofmann, K.; el Ghany, M.A.A. Energy Efficient Asynchronous—Asynchronous Switched NoC. In Proceedings of the 2020 9th International Conference on Modern Circuits and Systems Technologies (MOCAST), Bremen, Germany, 7–9 September 2020; pp. 1–4. [CrossRef]
14. Sparso, J.; Furber, S. *Principles of Asynchronous Circuit Design—A Systems Perspective*; Kluwer Academic Publishers: Berlin, Germany, 2007, p. 337.
15. Beigné, E.; Vivet, P.; Thonnart, Y.; Christmann, J.F.; Clermidy, F. Asynchronous Circuit Designs for the Internet of Everything: A Methodology for Ultralow-Power Design with GALS Architecture. *IEEE Solid-State Circuits Mag.*, **2016**, *8*, 39–47. [CrossRef]
16. Sparso, J. Asynchronous Circuit Design—A Tutorial. Available online: https://orbit.dtu.dk/files/2775719/imm855.pdf (accessed on 10 June 2021).
17. Cota, É.; Amory, A.D.; Lubaszewski, M.S. Reliability, Availability and Serviceability of Networks-on-Chip; Springer: New York, NY, USA, 2012.
18. Sadawarte, Y.A.; Gaikwad, M.A.; Patrikar, R.M. Comparative study of switching techniques for network-on-chip architecture. In Proceedings of the 2011 International Conference on Communication, Computing & Security, Rourkela Odisha, India, 12–14 February 2011; pp. 243–246. [CrossRef]
19. Miorandi, G.; Balboni, M.; Nowick, S.M.; Bertozzi, D. Accurate Assessment of Bundled-Data Asynchronous NoCs Enabled by a Predictable and Efficient Hierarchical Synthesis Flow. In Proceedings of the 2017 23rd IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC), San Diego, CA, USA, 21–24 May 2017; pp. 10–17. [CrossRef]
20. Fairouz, A.; Abusultan, M.; Elshennawy, A.; Khatri, S.P. Comparing leakage reduction techniques for an asynchronous network-on-chip router. *J. Low Power Electron.*, **2018**, *14*, 414–427. [CrossRef]
21. Pontes, J.J.H.; Moreira, M.T.; Moraes, F.G.; Calazans, N.L.V. Hermes-AA: A 65nm asynchronous NoC router with adaptive routing. In Proceedings of the 23rd IEEE International SOC Conference, Las Vegas, NV, USA, 27–29 September 2010; Volume 1, pp. 493–498. [CrossRef]

22. Beigné, E.; Clermidy, F.; Lhermet, H.; Miermont, S.; Thonnart, Y.; Tran, X.T.; Valentian, A.; Varreau, D.; Vivet, P.; Popon, X.; et al. An asynchronous power aware and adaptive NoC based circuit. *IEEE J. Solid-State Circuits* 2009, 44, 1167–1177. [CrossRef]

23. Nowick, S.M.; Singh, M. Asynchronous design-part 2: Systems and methodologies. *IEEE Des. Test* 2015, 32, 19–28. [CrossRef]

24. Gibiluka, M.; Moreira, M.T.; Calazans, N.L.V. A bundled-data asynchronous circuit synthesis flow using a commercial EDA framework. In Proceedings of the 2015 Euromicro Conference on Digital System Design, Madeira, Portugal, 26–28 August 2015; pp. 79–86. [CrossRef]

25. Bokhari, H.; Javaid, H.; Shafique, M.; Henkel, J.; Parameswaran, S. Darknoc: Designing energy-efficient networkonchip with multivt cells for dark silicon. In Proceedings of the 2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 1–5 June 2014. [CrossRef]

26. Zhan, J.; Ouyang, J.; Ge, F.; Zhao, J.; Xie, Y. DimNoC: A dim silicon approach towards power-efficient on-chip network. In Proceedings of the 2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 8–12 June 2015. [CrossRef]

27. Hesham, S.; Goehringer, D.; el Ghany, M.A.A. A call-up for circuit-switched NoCs in the Dark-Silicon Era. In Proceedings of the 2017 IEEE Nordic Circuits and Systems Conference (NORCAS): NORCHIP and International Symposium of System-on-Chip (SoC), Linköping, Sweden, 23–25 October 2017; pp. 1–6. [CrossRef]

28. Mohammed, M.S.; Al-Kubati, A.A.M.; Paraman, N.; Ab Rahman, A.A.-H.; Marsono, M.N. DTaPO: Dynamic Thermal-Aware Performance Optimization for Dark Silicon Many-Core Systems. *Electronics* 2020, 9, 1980. [CrossRef]

29. Bertozzi, D.; Miorandi, G.; Ghiribaldi, A.; Burleson, W.; Sadowski, G.; Bhaward, K.; Jiang, W.; Nowick, S.M. Cost-Effective and Flexible Asynchronous Interconnect Technology for GALS Systems. *IEEE Micro* 2021, 41, 69–81. [CrossRef]