Latency-Guided On-Chip Bus Network Design

Milenko Drinić, Darko Kirovski, Seapahn Meguerdichian, and Miodrag Potkonjak
Computer Science Department, University of California, Los Angeles, CA 90095-1596

Abstract
Deep submicron technology scaling has two major ramifications on the design process. First, reduced feature size significantly increases wire delay, thus resulting in critical paths being dominated by global interconnect rather than gate delays. Second, ultra high level of integration mandates design of systems-on-chip that encompass numerous intra-synchronous blocks with decreased functional granularity and increased communication demands. To address these issues we have developed an on-chip bus network design methodology and corresponding set of tools which, for the first time, close the synthesis loop between system and physical design. The approach has three components: a communication profiler, a bus network designer, and a fast approximate floorplanner. The communication profiler collects run-time information about the traffic between system cores. The bus network design component optimizes the bus network structure by coordinating information from the other two components. The floorplanner aims at creating a feasible floorplan and to communicate information about the most constrained parts of the network.

1 Introduction
Due to design complexity and time-to-market pressure, it is expected that future systems-on-chip are designed as network of virtual components. Virtual component (VC) is a core wrapped with logic that enables it to I/O data to the attached bus with an arbitrary bus protocol [17]. Because of high integration levels (100s of millions of transistors), the network of cores is estimated to count hundreds of VCs, where each core is smaller than 50-100K gates. The limitation on the gate count stems from the fact that larger cores would encounter synchronicity problems due to increased wire latencies [16]. Such a design paradigm does not pose significant restrictions to application design, as most of the modern multimedia, graphics, and communications applications use building blocks.

Since decreased levels of module granularity in computation result in higher communication costs, it is expected that the performance of future core-based systems is greatly affected by the inter-core communication. Communication among cores in DSM systems poses several design issues that can be classified as: (i) synchronization and (ii) performance optimization problems. While latency insensitive synchronization between cores can be resolved using relay stations and appropriate communication protocols [3], to the best of our knowledge, problems such as bus network design and core to bus assignment have not been addressed to date.

In this paper, we present a novel system-level design framework that, based on the communication profile of the modules involved, creates a single-chip bus network and assigns cores to buses such that the overall processing throughput of the system is maximized. The framework consists of three components: (i) the communication profiler, (ii) the bus network designer and (iii) the approximate floorplanner. For a given set of applications and a fixed number of pre-synthesized cores, the designer initially simulates the communication behavior of the system modules and creates a time-invariant profile of the connectivity and communication patterns among cores. The bus network design tool uses the communication profile to arrange on-chip bus structures and core connectivity. Its goal is to create a communication network which results in maximized expected throughput. Due to the time-invariance of the collected profiles, this objective is quantified heuristically. The created bus network is then fed to the approximate floorplanner which attempts to create a feasible layout. The feasibility of the layout is measured by comparing bus wirelengths to an upper bound constraint. In case of an unsuccessful search, the approximate floorplanning tool returns to the bus network designer, a list of $K$ best solutions with all unsatisfied bus constraints. In the next iteration, the bus network design tool considers these solutions, their corresponding latencies, and tries to rearrange the bus network such that at least one of these solutions can be satisfied. The goal of the synthesis process is to explore the solution space by toggling between infeasible and feasible solutions and try to find bus networks which result in higher communication throughput.

The developed synthesis framework has been deployed in optimizing performance in a number of core based designs extrapolated from several applications running on the state-of-the-art (30+ modules) systems-on-chip [15].

We present several design trade-offs involved in bus network design using the motivation example presented in Figure 1. Consider eight cores $C_1, ..., C_8$ which communicate with each other in four control cycles as presented in Figure 1(a). For simplicity and brevity, assume that each communication (control cycle) on a bus or a neighboring bus takes one bus cycle. Buses are connected using bridges. Sending a message over two bus bridges takes three clock cycles. The optimization goal is to assign cores to buses, such that no bus has more than three cores, and that all messages are delivered in as few bus cycles as possible. A possible greedy approach would identify modules that communicate most frequently among each other and assign them to a single bus. An example of such a greedy strategy would identify cores $C_1, C_5,$ and $C_6$, and assign them to a single bus. The resulting core-to-bus assignment, presented in Figure 1(b-A), would take seven clock cycles to deliver all messages.

Let us consider the communication overlap among modules as part of our heuristic assignment strategy. First, we define a Communication-Connectivity Graph (CCG) as an undirected weighted graph with a set of nodes $C$ representing cores and a set of edges $E$ representing existence of communication between two cores. Next, we define Weight $w_S$ of a Control Cycle $S$ as the cardinality of the set of communications that occur at control cycle $S$. Weight $w(E)$ of a CCG edge $E(C_i, C_j)$ is defined as a sum of weights: $w(E) = \sum_{C \in E} w_S$ of all control cycles $CS$ at which cores $C_i$ and $C_j$ communicate. A CCG that corresponds to the communication pattern in our motivational example is presented in Figure 1(c).

The goal of assigning a set of modules $M$ that (i) communicate frequently with each other and (ii) communicate to other cores $(C - M)$ at control cycles at which no other core communicates among $M$, can be modeled in the following way. We define a Bus $B$ as a partition of cores from $C, B \subset C$. Next, we define an Objective Function $OF(B)$ of a Bus $B$ as the sum of edge weights among cores $C_i, C_j \in B$ minus the sum of edges adjacent to nodes in $B$ and $C - B$:

$$OF(B) = \sum_{C_i \in \text{nodes} \in B} \sum_{C_j \in \text{nodes} \in C - B} w(E(C_i, C_j)) - \sum_{C_i \in \text{nodes} \in B, C_j \in \text{nodes} \in C - B} w(E(C_i, C_j))$$

We heuristically denote a particular core-to-bus assignment as $\alpha$-optimal with respect to the above mentioned optimization goal, if it represents a $K$-partitioning of $C$ into $K$ buses with maximal $OF = \sum_{i=1}^{K} OF(B_i)$. The objective function for core-to-bus assignment presented in Figure 1(b-A) equals $OF = -15$. The corresponding partitions of the associated CCG are presented in Figure 1(c-A).
Exhaustive search has identified the solution depicted in Figure 1(b-C) as α-optimal, resulting in $OF = 3$, and only four clock cycles required to deliver all messages among modules. Hence, even on such a small example we have demonstrated the feasibility of the optimal core-to-bus assignment technique, which involves a number of often counterintuitive trade-off considerations, may significantly improve on-chip communication performance.

2 Related Work

The EDA industry and academia are taking steps to address the emerging synthesis problems associated with forthcoming 0.04-micron technologies [5], [16]. DSM manufacturing technologies are expected to result in: (i) significant increase of wire latency [9], (ii) increased likelihood of signal crosstalk and delay uncertainty, (iii) current leakage, and (iv) increased power dissipation. Although most of the posed problems can be addressed at lower levels of design abstraction, the problem of increased wire latencies has significant impact on the higher level design stages. At lower design levels, researchers have explored ways of improving interconnect performance by topology optimization, buffer insertion and sizing [1], bus driver improvements, and optimal wire sizing [4]. Still, high-level and logic synthesis optimization methods, which take into account interconnect delay, have to be deployed [16]. Addressing the impact of slow interconnects at higher design stages requires interactive collaboration between the high-level design tools and placement and routing tools [16].

3 Global Design Flow

In this paper, we present a novel system-level design framework which creates a single-chip bus network and assigns cores to buses such that the overall processing throughput of the system is maximized. The framework consists of three components: (i) the communication profiler, (ii) the bus network designer, and (iii) the approximate floorplanner. The designer initially simulates the communication behavior of the system modules and creates a time-invariant profile (CCG) of the connectivity and communication patterns among given cores. The communication profiler takes into account the spatial and temporal correlation of communication patterns among cores. The design flow enters a synthesis loop which toggles between bus network design and approximate floorplanning.

The designer starts the loop with an initial bus network solution which has a feasible layout and results in relatively low-quality performance. The bus network design tool rearranges the bus network by (i) removing or adding bridges between buses or (ii) reassigning cores from one bus to another. Its goal is to create core connectivity which results in maximized expected communication throughput. This objective is estimated heuristically by considering communication delays and overlap. The created bus network is then fed to the approximate floorplanning tool which attempts to create a feasible layout. The approximation floorplanning tool memorizes a pool $\Pi(K)$ of $K$ best solutions which are returned to the bus network designer with appropriate statistics. In the next iteration, the bus network design tool considers solutions and tries to rearrange the bus network such that at least one of these solutions can be satisfied.

As the complexities of behavioral specifications increase, both design flows are becoming more vulnerable to the engineering change (EC) process due to the demand for updating design solutions. To address this issue, we have developed a generic EC methodology applicable to all design stages, which facilitates constraint manipulation to augment the design with flexibility for future changes [10].

4 Preliminaries

In order to position our work, in this subsection we present the related hardware and communication model, and briefly outline our communication profiling methodology.

The generic hardware model that we have adopted assumes the following set of constraints: cores have the property of being hard, i.e. with constant layout, and hence, with specific location of their bus interface. Cores are connected using buses of limited maximal length. The bus network is switched using a topology of bus bridges, where a bus bridge is a crossover of at most four buses with associated synchronization and buffering logic [9].

Bus arbitration is performed with static priority assignment [12]. Bus latency is modeled using a second order polynomial of the wirelength adopted from [14]. Message routing across the bus network is assumed to be deadlock free [6], [8].

The communication profiler is executed as a pre-processing step with a goal of summarizing the essential statistics of communication patterns of the implemented application. The input to the communication profiler is an instance of the communication model of our target system. In our experiments, we have used traffic patterns extracted and extrapolated from the MediaBench benchmarks suite [11].

The communication profiler collects run-time information according to the definitions presented in Subsection 1.1. Weight $w(E)$ of a CCG edge $E(C_i, C_j)$ is computed as a sum of vector sums, where each vector sum is computed per $C_i - C_j$ transaction.

\[ w(E) = \sum_{E \in CCG} \frac{\sum_{i=1}^D w_{Ei} \cdot pdf(i)}{pdf(i)} \]

The vector sum is computed as a pdf-weighted sum of constraints in the preceding and succeeding $D$ control cycles to the control cycle $S$ at which the transaction occurs. The CCG built during a single run of the communication profiler, is post-processed by removing edges with small weights, typically smaller than the median of the global weight minus one standard deviation $\text{median}(w) - \text{std}(w)$.

5 Latency-Guided Design of On-Chip Bus Networks

The goal of the bus network design algorithm is to create a bus network and a core to bus assignment such that the
overall throughput of the system is maximized. In the first run of the algorithm, its input consists of a set of cores, a communication profile of the system applications represented as a CCG, and an initial ad-hoc solution (a starting bus network and bus-to-core assignment with a floorplan). In the subsequent runs, the ad-hoc solution is replaced with the information from the floorplanner which quantifies the constraints that cannot be satisfied with respect to the solution obtained in the previous run of the bus network designer.

**Problem: Balanced Partitioning of an hCCG.**

**Input:** An hCCG hypergraph and a variable a.

**Question:** Is there a balanced partitioning of hCCG into a set of hNs which results into a valid hCCG such that its OF is greater than a?

![Hypernode communication-connectivity graph](image)

**Figure 2:** An example of a hCCG and its corresponding bus structure.

We start the formal description of the problem with a series of related definitions. We first define an extension to the CCG, the Hyperedge Communication-Connectivity Graph hCCG as an undirected graph where a Hypernode hN is a collection of nodes in CCG and a Bus Hyperedge BhE is a hyperedge that encompasses at most M hypernodes. We define a relation $hN \in BhE$ when $hN$ is covered by BhE. Hypernode hN is semi-free, if it belongs to only one BhE. Hypernode hN is isolated if $hN \notin BhE_i$, and $hN \notin BhE_j$, where $BhE_i \neq BhE_j$. According to its definition, an hCCG formally describes a bus architecture. While $hN$ represents a bus segment, BhE represents a bridge and its relation to adjacent buses. A Chain of BhEs is a set of BhEs such that there exists a BhE that does not input as with another BhE in the chain. A hCCG is connected, if there exists a chain of BhEs that encloses all nodes of CCG. Hypernode hN is valid, if it satisfies the relation $|hN| \leq M$. An hCCG is valid, if and only if (i) it is composed of valid hNs, (ii) if any two BhE overlap in maximum one hN, (iii) if no hN belongs to more than two BhEs, and (iv) if hCCG is connected. An example of a valid hCCG and its corresponding bus structure is shown in Figure 2.

It is important to stress that technology-specific limitations for a desired clock cycle (parasitic capacitance of the bus segment) are reflected through two design constraints: (i) $\lambda_{seg}$ - maximum length of a bus segment and (ii) $M$ - maximum number of cores that can be attached to a bus segment (hypernode cardinality $|hN| \leq M$).

We formally define the trade-offs involved in generation of bus networks and core to bus assignment using an objective function $OF$ (its simplified version is presented in Subsection 1.3).

$$OF = \sum_{i=1}^{CCG(N)} \left( \sum_{C_j \in N_i} w(E(C_j, C_k)) - \sum_{C_j \in N_i} \sum_{C_k \in C_j \cap CCN - hN} (w(E(C_j, C_k)) - p(C_j, C_k)) \Phi_{j,k} \right)$$

where $p(C_j, C_k)$ represents the length of the routing path between cores $C_j$ and $C_k$ (latency is modeled proportional to the number of bridges between $C_j$ and $C_k$). More formally, $p(C_j, C_k)$ is equal to the cardinality of BhEs in the routing chain that connects the $hN$s which contain $C_j$ and $C_k$. $\Phi_{j,k}$ is defined as the sum of bus segment workloads on $p(C_j, C_k)$.

$$\Phi_{j,k} = \sum_{hN \in E(C_j, C_k)} \phi(hN)$$

A workload $\phi(hN)$ of a bus segment $hN$ is defined as percentage of time when the bus segment is in use by cores associated to it. $\phi(hN)$ is determined from the statistical behavior of each core for the given application workload.

The objective function has been derived with the following set of incentives: (i) cores which communicate frequently should be placed on the same bus, (ii) cores attached to different buses should communicate through as few bridges as possible, and (iii) congestion of individual buses should be minimized. Important features of the objective function are its global system performance characterization and empirical correlation to system’s throughput (see Section 6). The search for an $\alpha$-optimal design (maximized OF) can be abstracted as follows.

**While** ($T > T_r$)

**currentPartition = bestPartition**

**While** ($\text{cumulative improvement} \geq \sigma$)

**Case** ($a > c_1$) : $\text{currPartn. SwapNodes(a \to rand())}$

**Case** ($a < c_1$ & ($a > c_2$) : $\text{currPartn. MoveNode(rand())}$

**Case** ($a < c_2$) : $\text{currPartn. MoveBE(rand())}$

$\delta = \text{currentSum} - \text{OF}$

If ($\delta < 0$) then Accept modified currentPartition

else Accept modified currentPartition with probability $p = e^{-\delta}$

End while

*Increase temperature*

End while

![Pseudo-code of the simulated annealing algorithm](image)

**Figure 3:** Pseudo-code of the simulated annealing algorithm for balanced partitioning of hCCG.

The problem of Balanced Partitioning of an hCCG is computationally intractable as it can easily be restricted by imposing $M = 2$ and simplifying OF as presented in Subsection 1.1, to the NP-complete balanced partitioning problem [7]. Therefore, to address this difficult problem, we have developed a variant of a traditional simulated annealing (SA) algorithm. The algorithm is illustrated using the pseudo-code in Figure 3.

The developed components of the algorithm that have been augmented into a traditional SA search engine are: a selection of atomic solution alterations, moves, and an engineering change pre-processing. We first describe the set of move functions. Moves can be classified into two categories: moving CCG nodes to hypernodes and modifying hCCG hyperedges BhEs. Moves do not affect the underlying CCG structure.

In the first category, we distinguish two different move actions: (i) $\text{SwapNodes}(C_i, hN_i, C_j, hN_j)$ node swapping between hypernodes and (ii) $\text{MoveNode}(C_i, hN_i, hN_j)$ node transfer from one hypernode $hN_i$ to another $hN_j$. If $hN_i$ is equal to one before the $\text{MoveNode}(C_i, hN_i, hN_j)$ operation, hypernode $hN_i$ can be removed from the list of hypernodes. Both $\text{SwapNodes}()$ and $\text{MoveNode}()$ are not performed if the resulting hCCG is not valid.

The second category of moves $\text{MoveBhE()}$ modifies hyperedges and thus, the bus network structure. There are three possible alternations of this move. The first one is a removal of a selected BhE. The BhE on an $e$-distance equal to one pseudo-randomly acquire the semi-free $hN \in BhE$. We define $e$-distance of a hypernode $hN$, $hN \in BhE$, as a set of hypernodes covered by BhEs which are included in all chains of length $e$ starting from BhE. The second variant of $\text{MoveBhE()}$ is regrouping. Here, the set of all seized $hN \in BhE$ is released from BhE and a semi-free $hN_e$ from another BhE, is seized by $hN$ to create a new hyperedge $BhE_e = hN_e \cup hN$. The third variant of $\text{MoveBhE()}$ creates two new hyperedges $BhE_e$ and $BhE_f$ from a parent hyperedge $BhE$ by pseudo-random bipartitioning of BhE. If there exists only one seized $hN \in BhE$, one of the partitions with only semi-free hypernodes is regrouped with a pseudo-randomly selected hyperedge $BhE, \neq BhE$. $\text{MoveBhE()}$ is only performed on higher search temperatures since it significantly changes the structure of the hCCG. $\text{MoveBhE()}$ by construction preserves the validness of hCCG.

The interactivity of the bus network designer and the approximate floorplanner is enabled through an engineering change (EC) procedure $\text{EngineeringChangeProcedure()}$. The goal of this procedure is to mark the feasible portion of
the bus network as unchangeable and to reduce the solution space and thus, search time for the new iteration of the SA process. Due to brevity, we do not present the developed EC technique as the key concepts have been adopted from a generic EC methodology [10].

The developed EC process restricts certain subdomains of the solution space by adding constraints based on report from the floorplanner. This report contains quantitative information about the satisfiability of constraints posed by the output of the balanced partitioning of hCCG. The EC technique defines the likelihood $p(\text{object})$ to participate in a move action for each node $C \in hCCG$ and hyperedge $Be \in hCCG$. A core and/or a bus object, that often violated the $\lambda_{max}$ technology dependent parameter, is set with high $p(\text{object})$.

The bus network design tool outputs the final bus connectivity and core-to-bus assignment (the set of hyperedges $hN \in hCCG$, hyperedges $Be \in hCCG$, and nodes $C \in hCCG$ of the partitioned hCCG) to the approximate floorplanner for technology parameter evaluation.

The output of the bus network designer is fed to the approximation floorplanning tool. This tool solves the tooling problem by means of modified simulated annealing algorithm. It outputs the following important statistics to the network design tool: (i) the $K$ best solutions encountered; (ii) for each solution $\Pi_i \in \Pi(K)$, the area and the violated bus constraints; (iii) the overall percentage of instances that each bus constraint has violated. To estimate bus lengths, we use 1/2 the perimeter of the smallest bounding box of each bus $hN_i \in hCCG$. The objective function that the SA process optimizes is a linear combination of the area and bus length requirements. The standard SA process is augmented with solution transformation actions, moves: greedy and enabling. In the greedy move, we select the longest bus and try to improve the placement of, while in the enabling move, we select the shortest bus and try to relax the placement of its modules. We calculate the center-of-mass $C_m$ and a force vector $V$ for the module that has the longest/shortest Manhattan distance from $C_m$. The selected module is moved in the direction of $V$ in proportion to the magnitude $|V|$. More details about this procedure can be found on [http://www.cs.ucla.edu/~milenko/latency_guided_design.pdf](http://www.cs.ucla.edu/~milenko/latency_guided_design.pdf)

### 6 Experimental Results

We have demonstrated the effectiveness of our synthesis paradigm on a set of synthetic designs. The design benchmark has been assembled using a library of intellectual property cores from [13]. The area for each core was estimated based on the core gate count. Communication traffic patterns were extrapolated from a number of software multimedia benchmarks profiled at the function level [11].

The experimental results are presented using Table 1. Columns 1–4 describe the tangible properties of each system: application emphasis, estimated number of gates [13], number of buses, and number of bridges. Columns 5 and 6 quantify the run-time properties of the synthesis framework: the number of complete iterations of the bus network design and floorplanning loop and the elapsed total run-time of the semi-automated process. The last two columns represent the following properties of the obtained solution: (i) optimized system throughput as a multiple of the throughput obtained by the initial ad-hoc system which is fed as a starting solution to the bus network designer and (ii) the median ratio of idle cycle time on the system buses. The throughput improvement for various designs ranged from 46% to 216%.

### 7 Conclusion

In order to address the two major ramifications of DSM on the design process: increased interconnect delay and design reuse using blocks with decreased functional granularity and increased communication demands, we have developed a methodology for the design of on-chip bus network structures. The design methodology closes the synthesis loop between system and physical design by performing the following three procedures. First, the communication among cores is profiled to obtain run-time information about the traffic. Next, the bus network design tool creates and optimizes the bus network structure by coordinating information from the profiler and the approximate floorplanner. The latter, as the final component of the design, aims at creating a feasible floorplan. In the case of an infeasible solution, the approximate floorplanner communicates the information about the most constrained parts of the network back to the bus network design tool.

### References