Analyzing Concurrency in Streaming Applications

S. Stuijk and T. Basten
Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands.
{s.stuijk,a.a.basten}@tue.nl

Abstract

We present a concurrency model that allows reasoning about concurrency in executable specifications of streaming applications. It provides measures for five different concurrency properties. The aim of the model is to provide insight into concurrency bottlenecks in an application and to provide global direction when performing implementation-independent concurrency optimization. The model focuses on task-level concurrency. A concurrency optimization method and a prototype implementation of a supporting analysis tool have been developed. We use the model and tool to optimize the concurrency in a number of multimedia applications. The results show that the concurrency model allows target-architecture-independent concurrency optimization.

Keywords. streaming, task-level concurrency, multi-processing, concurrency measures, system-on-chip design

1 Introduction

The consumer-electronics market is characterized by rapid developments in embedded multimedia systems. In recent years, we have for instance seen successful market introductions of portable MP3 players, digital cameras and set-top boxes. The pace at which new products are introduced on the market is ever increasing. Consumer-product manufacturers try to cope with this trend by decreasing their time-to-market. On the other hand, the complexity of embedded multimedia systems is growing, as users have high expectation about the functionality and quality delivered by new products. To deal with these adverse trends, the electronic-design community expects that future electronic systems re-use platforms that integrate many IP-blocks. Software can be executed concurrently on the IP-blocks in these *multi-processor systems-on-chip*. Novel programming techniques are required to use these systems. These techniques must exploit the concurrency that is present in the hardware architecture and meet with the timing-, energy-, performance-, and cost-constraints. A coarse overview of the multi-processor system-on-chip programming trajectory is shown in Figure 1. The figure shows a subdivision of the programming problem into two subsequent steps (mapping and binding). The programming of the hardware level is done from an intermediate level, called the implementation level. This step binds one (or a few) compute tasks onto one processor. In this way, we can relay on traditional compiler technology and minimize the overhead of a run-time system. The step from the specification level to the implementation level is responsible for subdividing the (executable) specification in such a way that the resulting tasks can efficiently be programmed on the hardware platform. This so-called multi-processor mapping step must consider aspects like concurrency, energy and timing. To do this, it will need information about the underlying hardware platform. This information is gradually added during the mapping trajectory.

The programming trajectory covers a system-level design methodology from the early design stages till the actual system-on-chip solution. We focus on (data-intensive) streaming applications, as we are targeting multimedia applications. Many design flows for embedded multimedia systems are based on some kind of task graph [1, 2, 22, 24, 31]. All such flows could benefit from a good initial task graph as their input. In current design practice, the programming trajectory for such applications typically starts with an executable specification of the application written by an application designer. The specification is usually given as a sequential program that describes the logical functions used in the application and it is written in a language like C or C++. Target platforms usually allow for concurrent execution of the application. For this reason, part of the mapping flow deals with the extraction of the concurrency from the application. Concurrency has a large impact on the system, which means that the extraction should be performed early in the programming flow. Concurrency analysis is to some extent independent of the (precise) architecture targeted. In other words, some transformations performed to extract concurrency from the application are valid for a large class of architectures. This enables efficient re-use of

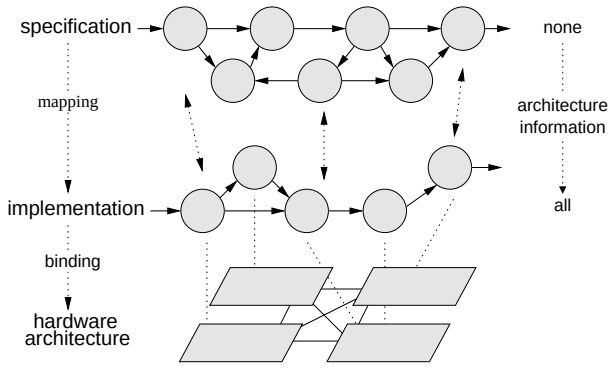


Figure 1. Multi-processor system-on-chip programming trajectory.

optimizations performed on the specification if changes are made to the hardware platform. In this paper, we propose a technique to analyze the concurrency that is available in an application independent of the exact target architecture. The result of the proposed design-space exploration is a so-called task graph, that makes data transformations and the underlying data streams explicit. It is our aim to provide a specification of the application which forms a good starting point for mapping it onto many different systems-on-chip platforms. In other words, we try to answer the question of how to come to a good initial task graph.

The next section introduces an abstract model for parallel (streaming) computations. The concurrency model is discussed in Section 3. A prototype concurrency-analysis tool implementing the concurrency measures is presented in Section 4. In Section 5, a supporting concurrency optimization method is presented. The concurrency model is used in a number of case studies to optimize the concurrency of a JPEG decoder, an H.263 video-conferencing decoder and a 3D recursive search algorithm. Methods to take aspects of the target-architecture-platform into account during the concurrency analysis are discussed in Section 7. An overview of related work on multi-processor programming and concurrency analysis is given in Section 8. Section 9 concludes.

2. Model of Computation

The model of computation introduced in this section captures the core of parallel (streaming) applications. It only specifies those aspects that are necessary for concurrency analysis. This allows for many instantiations of the model. It is, for example, sufficiently abstract to comprise a number of data-flow models like Kahn process network [12, 13], Synchronous dataflow [18] and a subclass of Petri nets, called marked graphs [3]. Our concurrency analysis can be applied to (executable) specifications in all these models.

2.1. Computational networks

We assume that a parallel computation is organized as a collection of autonomous compute nodes that are connected to each other by point-to-point connections. Compute nodes exchange information through these connections. These connections are the only way of communication. A given node computes on data coming along its input connections to produce output on some or all of its output connections. This informal definition of a parallel computation is the basis of the computational-network model. A compute node has a set of input ports and a set of output ports for connections to its environment. Input and output data is modeled using strings of data-elements, which is a good abstract model of data streams. The execution of a compute node implies the reading of input strings from its input ports and writing the appropriate output strings to its output ports. This is done following the *transformation* that describes the behavior of the node.

Definition 2.1 (Compute node) *A compute node is a tuple* (I, O, t) *where*

- I *is a set of input ports;*
- O *is a set (disjoint of* I*) of output ports;*
- t *is a transformation, for example, a function describing how a compute node computes a (tuple of) output strings using a (tuple of) input strings.*

Note that we are not interested in the exact form of a transformation and that we do not define the types of data received and sent over ports. We also do not require a specification of aspects like the number of input ports read, the amount of data elements read on each port, etc. For our purposes, these details are irrelevant, and including them in the definition of compute nodes would unnecessarily restrict and complicate matters. The only requirement is that transformations allow an operational implementation resulting in an event diagram, as further explained in the next subsection. The definition of a compute node enables us to define a *computational network*. It contains a set of compute nodes that are connected to each other using point-to-point connections that transfer data streams in order (fifo communication). We abstract from the exact capacity of connections. Some ports of compute nodes may remain unconnected. These ports allow connections to the environment.

Definition 2.2 (Computational network) *A computational network* CN *is a tuple* (N, C, I, O) *where*

- N *is a set of compute nodes;*
- C *is a set of connections;*

- *every connection in C connects an output port of a compute node to an input port of a compute node;*
- *every port of every compute node is connected to at most one connection;*
- *I is the set of input ports of the computational network, being defined as those input ports of the nodes in N not connected to a connection in C;*
- *O is the set of output ports of the computational network, being the unconnected output ports of the compute nodes in N.*

An example of a computational network is shown in Figure 2. The network contains five compute nodes. The input port of node a is unconnected. So, this port is an input port of the network and node a is therefore called an input node of the network. Nodes d and e provide output ports to the environment and are therefore called output nodes of the network.

2.2. Executions

In this sub-section, we give an abstract notion of executions of computational networks, which forms the basis for our concurrency model. A computational network consists of a set of compute nodes which together perform a computation. The nodes communicate with each other through connections. Each node performs a sequence of actions (e.g. C/C++ statements in an executable specification) which are modeled as a totally ordered sequence of events. These events are classified into the following three types:

1. **write event** Such an event models a write operation in which a compute node writes to one of its output ports.
2. **read event** Such an event models a read operation from an input port.
3. **internal event** Such an event models the execution of an action or a sequence of actions not including read or write operations.

Lamport [17] has shown that the events in such an event model form a partial order. This partial order is called the *causality relation* or *happened before relation* and is denoted by $\prec$. Lamport's *logical clocks* can be used to create an ordering that is consistent with causality for all events that occur during a computation. Lamport's system of logical clocks assumes one logical clock per compute node. This logical clock assigns to every event a time-stamp that is the logical clock value at the moment the event occurred. Every event is performed within a period corresponding to a single logical clock value. The clock of a node is incremented once between two events. Furthermore, since

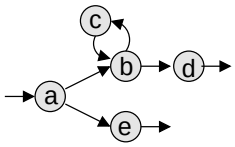


Figure 2. An example of a computational network.

communication imposes a causality relation that must be respected by the logical clocks, the clock of a reading node is updated using the time-stamp of the received event: the local clock is set to the maximum of the received time-stamp and the current clock value.

To reason accurately about timing aspects without referring to concrete implementations, we use Lamport's clocks but we associate a duration with the events that take place in the compute nodes and with the communication over the connections. Of course, these durations must be in some way reasonable for actual system implementations. They should be abstract but realistic. The duration for a connection can for instance be based on the amount of data communicated and the propagation delay of the on-chip interconnect. The latter is an architecture-dependent property that can easily be taken into account in the duration assignment. We introduce two *duration functions* d_e and d_c that map a set of events E of an execution plus the set of connections C of a computational network to the set of natural numbers, $\mathbb{N}$. Formally, $d_e : E \rightarrow \mathbb{N}$ and $d_c : C \rightarrow \mathbb{N}$. Lamport's original system of logical clocks can be modeled by assigning a duration of one to every event and a delay of zero to every connection.

Our time-stamping mechanism based on Lamport's logical clocks essentially is a time-stamping function t that maps the set of events E to the totally ordered set $\mathbb{N}$, formally, $t : E \rightarrow \mathbb{N}$. This mapping is such that $e \prec e'$ implies $t(e) < t(e')$. The timestamps can be computed via a set of counters, the local logical clocks. Each compute node in the computational network maintains a different counter, all initially set to 0. Let t_i denote the counter maintained by compute node n_i. When a compute node n_i executes an event, it updates first its local clock t_i and then time-stamps the event. Thus this time-stamp is the value of the local logical clock after the event is executed. The protocol used to update the clock t_i of n_i is the following:

1. When n_i executes an internal or write event e, the clock value t_i is updated to $t_i := t_i + d_e(e)$.
2. When n_i executes a read event e, where y is the time-stamp of the corresponding write event and c is the connection over which the event was received, the clock is updated to $t_i := \max(t_i, y + d_c(c)) + d_e(e)$. Note that the maximum operation synchronizes the logical clocks of the sender and receiver, taking into account the delay caused by the communication.

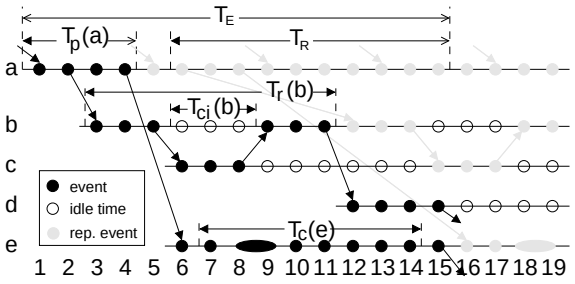


Figure 3. An event diagram.

An execution of a computational network is the set of events that occur in all compute nodes when they transform a set of strings on the input ports of the computational network. The execution can be displayed graphically in an *event diagram*, such as the one shown in Figure 3. This figure shows in black an ordering of all events that take place during an execution in the network of Figure 2 for a given input, assuming the string on the input port of the network has a finite length. As we are targeting streaming applications, we have in practice often unbounded input strings. These unbounded strings can often be abstracted appropriately into an indefinite number of repetitions of the same finite input. This leads to an unbounded repetition of the same execution pattern. The gray nodes in Figure 3 represent the repetition of this execution pattern. In practice, one must make sure to get a representative execution pattern as a basis for concurrency analysis.

Let's consider the details of the execution in Figure 3. Node a starts with reading input from the environment. At the end, d and e produce values for the environment. The connection between nodes a and e has a duration of one logical clock value, all other connections have a duration of zero logical clock values. Node e executes one event that takes two logical clock values; all other events require one logical clock value. Note that this diagram is kept simple for illustrative purposes. The annotations in the diagram are explained below.

Our time-stamping mechanism can be used to analyze the ordering and abstract timing of events that take place in a computation. We define a number of measures, illustrated in Figure 3. Note that all the measures introduced here and also most measures of the next section are defined with respect to a single execution. We do not explicitly mention this execution in all formulas because that would compromise readability. Let $CN = (N, C, I, O)$ be a computational network, E the set of events that occur during the execution of CN and $E_n \subseteq E$ the set of all events which occur in compute node $n \in N$.

Definition 2.3 (Processing time) *The processing time, $T_p(n)$, of a compute node $n \in N$ is defined as follows:*

$$T_p(n) = \sum_{e \in E_n} d_e(e)$$

Definition 2.4 (Execution time) *The execution time, $T_E(CN)$, of CN is defined as:*

$$T_E(CN) = \max_{e \in E} t(e)$$

Definition 2.5 (Computation time) *The computation time, $T_c(n)$, of a compute node n in which the set of* internal *events $I_n \subseteq E_n$ occurs, is defined as follows:*

$$T_c(n) = \sum_{e \in I_n} d_e(e)$$

The term 'computation' is mainly used in this paper for the transformations performed on the strings of data in the network and not the communication of the strings of data. The combination of computation and communication is referred to with the term 'processing'. The terms 'execution' and 'run-time' are used if also idle time is taken into account.

Definition 2.6 (Communication idle time) *The communication idle time, $T_{ci}(n)$, of a compute node n is defined as:*

$$T_{ci}(n) = (\max_{e \in E_n} t(e) - \min_{e \in E_n} t(e)) - T_p(n)$$

We explained that we are aiming at streaming applications. The execution of a network can therefore be seen as a repetition of a single execution pattern. This repetition is shown with the gray nodes in the event diagram of Figure 3. Idle times at the beginning of nodes, such as b, c, d and e, in Figure 3, and the end of nodes, such as a, b, c, can be used for operations on other inputs. This motivates Definition 2.6, as well as the following definition.

Definition 2.7 (Run time) *The run-time $T_R(CN)$ of CN and $T_r(n)$ of node n is:*

$$T_r(n) = \max_{e \in E_n} t(e) - \min_{e \in E_n} t(e);\ T_R(CN) = \max_{n \in N} T_r(n)$$

Definition 2.8 (Sequential execution time) *The sequential time of an execution, $T_{SE}(CN)$, is defined to be the sum of the processing times of all compute nodes in the computational network:*

$$T_{SE}(CN) = \sum_{n \in N} T_p(n)$$

The sequential execution time approximates the execution time of a sequential version of the computation. This approximation is in general not entirely accurate because communication in a parallel execution is in general replaced by memory accesses plus extra control statements in a sequential execution, but we have to make a trade off between accuracy and abstractness. The introduced error is acceptable as long as the measures defined above and in the remainder provide a good basis for concurrency analysis. Our experiments confirm that the accuracy is sufficient.

3. Concurrency model

A computational network that realizes a computation has certain concurrency properties. The goal of the concurrency model is to quantify the presence of the different concurrency properties in a computational network. Concurrency is influenced by many things. It is for instance influenced by the way the computation is divided over the different compute nodes in the computational network and by the communication within the network. It can also be influenced by the compute platform or the used (run-time) scheduler. This (run-time) scheduler assigns the nodes to processors on which they execute. We are interested in the concurrency properties that are determined by the computational network itself and not its implementation environment, because we want to have a computational network that has good concurrency properties for many environments in which it may operate. Only in a later design phase, we propose to consider the environment and fine tune the concurrency in the computational network to this environment. However, this phase is not considered in the core part of this paper. Section 7 briefly returns to this aspect. To leave out the effects caused by the implementation environment, we assume that a compute node can execute as soon as the required data becomes available and furthermore that nodes do not have to wait for data on the input ports of the computational network. Using logical time, these assumptions can easily be realized.

An important goal of the concurrency measures is that they provide a global direction when optimizing the concurrency. To realize this, all measures are normalized to the range $[0, 1]$, in which a value of 1 means that the measured concurrency property is optimal and a value close to 0 means that it is very bad. The measures should besides the global direction also provide enough detail to find the concurrency bottlenecks. For this reason, a detailed measure per node is defined. We next introduce all five concurrency measures and motivate their applicability. The use of all measures in a design-space exploration strategy is discussed in Section 5.

Computation load. In a parallel execution, we want to minimize the overhead of communicating data between nodes. The nodes should spend as much time as possible on computation and not on communication. The time spent by a node on the computation is expressed in the computation time. The time that a node spends on both the computation and the communication is expressed in the processing time. The ratio between computation time and processing time should be as high as possible for every node, as computation, i.e., data transformation, is the main goal of every computational network. These observations lead to the first concurrency measure, the *computation load*.

Definition 3.1 (Computation load) *The computation load of computational network CN and a compute node $n \in N$ are defined as follows:*

$$CompLd(CN) = \frac{\sum_{n \in N} CompLd(n)}{|N|}$$

$$CompLd(n) = \frac{T_c(n)}{T_p(n)}$$

The computation load of the network serves as the global measure; the computation loads of the nodes serve as the detailed measure. The nodes with low computation loads may point to concurrency bottlenecks.

Example 3.1 We calculate the computation load of the computational network of Figure 2 with the event diagram of Figure 3. Let ECN denote the network. The computation time and processing time of the different nodes are found using Definitions 2.5 and 2.3 and the event diagram of the computation.

$$\begin{aligned} CompLd(ECN) &= \frac{1}{5} \cdot \left(\frac{1}{4} + \frac{2}{6} + \frac{1}{3} + \frac{2}{4} + \frac{8}{10} \right) \\ &= \frac{133}{300} \approx 0.44 \end{aligned}$$

Thus 44% of the total processing time is spent on meaningful computation. Node a has the lowest computation load, namely $1/4$. To improve this, the node should be assigned a larger computation task, or it can be merged with another node. □

Processing load. A compute node is during an execution either busy, performing events, or it is idle. It can be idle because it is waiting for data or it has finished its processing but other nodes have not finished executing. To get a balanced workload over nodes, we must balance the processing and run-times of the different nodes. This is important to optimize streaming behavior. To get a notion of the workload balance, we consider the ratio between the processing time and the run-time. The second concurrency measure, *processing load*, looks at this aspect.

Definition 3.2 (Processing load) *The processing load of computational network CN and of a compute node $n \in N$ are defined as follows:*

$$ProcLd(CN) = \frac{\sum_{n \in N} T_p(n)}{|N| \cdot T_R(CN)}; \;\; ProcLd(n) = \frac{T_p(n)}{T_r(n)}$$

For individual nodes, the processing load computes the ratio of the processing time and run-time. In other words, it calculates the ratio between the time that a node is busy and

the time that a node is either busy or waiting before it can continue processing. For the network, we assume (ideal) streaming behavior and we only consider the event diagram of a (small) part of the actual execution, typically the events caused by one or a few inputs to the network. Each node can start operating on a next input to the network with a rate that is determined by the node with the longest run-time. Therefore, the processing load measure of a network must not consider the ratio of the processing time and run-time per node, but compare the processing time of the nodes to the run-time of the network. The bottlenecks in obtaining a better processing load are the nodes with the lowest processing load and the node with the longest run-time.

Example 3.2 We continue with our running example. To calculate the processing load of computational network ECN, we need the maximum run-time of the nodes in the network. Node e requires 10 logical clock values from the logical clock value at which it starts processing. The other nodes require fewer logical clock values. The processing load is then equal to:

$$ProcLd(ECN) = \frac{4+6+3+4+10}{5 \cdot 10} = \frac{27}{50} \approx 0.54$$

The processing load for the individual nodes is 1 for nodes a, c, d and e, and $\frac{6}{9} = \frac{2}{3}$ for node b. The processing load of the network indicates that the nodes in the network are on average 54% of their time busy with computation or communication and 46% of their time idle. Potential points for improvement are nodes b (lowest processing load) and e (longest run-time). The fact that almost all nodes have a processing load of 1 whereas the overall processing load is only 0.54 indicates that the workload balance over the nodes is bad and that e is the most serious bottleneck, which brings us immediately to the next measure. □

Restart interval. The compute node with the longest run-time is determining the rate at which new computations can be started in the computational network. This node plays an important role in the throughput of the computational network. The throughput is an important property when a system designer is designing a streaming application. To get a notion of it, we introduce the *restart* measure through Definition 3.3. In general, the closer the restart measure comes to one, the higher the throughput realized by the computational network. However, note that the best restart does not guarantee the best network. Generally, good values for the restart can be obtained through very fine-grained compute nodes. However, this gives communication overhead (and possibly scheduling overhead). The restart measure should therefore be balanced with other measures. Restart is an abstract notion of throughput; it is not equal to it.

Definition 3.3 (Restart) *The restart of computational network CN and a compute node $n \in N$ are defined as follows:*

$$Restart(CN) = \frac{1}{T_R(CN)};\ Restart(n) = \frac{1}{T_r(n)}$$

The restart measure partly overlaps with the processing load, as they both point to the node with the longest run time. However, the restart measure is more fine-grained as it may point to a set of nodes which have a long run-time. The processing load points only to the node with the longest run-time, ignoring other nodes with a long run-time that are also potential throughput bottlenecks.

Example 3.3 The restart value for our example network ECN is $1/10$ with node e being the bottleneck node with the lowest restart value. One issue needs explanation. Consider two networks CN_1 and CN_2 that realize the same computation. The maximum run-time of the nodes in CN_1 is 1000 and 100 in CN_2. The restart for CN_1 is 0.001 and for CN_2 0.01. Looking at the absolute values, it is difficult to see that CN_2 is much faster than CN_1. This relative difference in restart interval must be made visible when comparing different solutions for the same application. This can be done by normalizing the values of the restart measure over a set of designs with the largest value. The disadvantage of this approach is that the 1 value for the best network (CN_1 in the example) may suggest that the restart value is optimal whereas this is obviously not always the case. Nevertheless, we choose this solution in our design optimization method, discussed later in this paper, in order to make relative differences visible. □

Synchronization. A parallel computation will in most cases be faster than a sequential implementation of that computation. This is often in the literature referred to as speed-up [21]. The realized speed-up depends on the synchronization that is required between the different nodes in the network, the introduced communication overhead, and how well the computation is balanced over the different nodes. The second and third aspect are covered by the computation load and processing load respectively. The influence of synchronization is not yet fully captured in the measures so far, although a poor synchronization does affect the processing load. Synchronization is important when considering concurrency, because synchronization is limiting the execution of compute nodes and with that the number of compute nodes that can run in parallel. Synchronization constraints may impose the restriction that two compute nodes can only execute one after another. Synchronization determines in this way the time that a computation will take in a computational network (see Definition 2.4).

Our concurrency measure, *synchronization*, is related to the speed-up. The measure is based on the inverse value of

the speed-up achieved by the computational network compared to a sequential solution (using Definitions 2.4 and 2.8). This value is subtracted from 1 to meet our objective that a value of 1 for a measure indicates a good solution from the concurrency point of view. Synchronization measures for nodes are not really meaningful. Due to communication overhead, individual nodes are usually slower in a parallel execution than in a sequential execution. For diagnosing synchronization bottlenecks, we can use event diagrams instead. In Figure 3, for example, the synchronization pattern between the nodes b and c causes idle time that may be removable.

Definition 3.4 (Synchronization) *The synchronization of CN is:*

$$Sync(CN) = 1 - \frac{T_E(CN)}{T_{SE}(CN)}$$

Assuming that an execution performs at least one event, the range of values for this measure is in $(-\infty, 1)$ where a value close to 1 indicates a good solution and 0 a solution that is as fast as the sequential execution. Negative values indicate that the execution is slower than a sequential execution. The fact that negative values are possible is not really a problem with respect to our goal that values should be in the range $[0, 1]$. Negative values will be rare and can easily be removed by taking the maximum with 0; it is more important that the optimum value is close to 1.

Example 3.4 To compute the value for the synchronization measure of computational network ECN of Example 3.1, we need the execution time of the computational network and the sequential execution time. Figure 3 shows that the execution time equals 15 logical clock values. The sequential execution time is found using Definition 2.8 and is equal to 27. The synchronization is then:

$$Sync(ECN) = 1 - \frac{15}{27} = \frac{12}{27} \approx 0.44$$

We can conclude that the parallel execution performs the computation approximately 44% faster than a sequential implementation of the computation. □

Structure. The previous measures consider the event diagram of an execution of a computational network with a given input. The structure of the network plays only an implicit role. The structure itself can already provide insight in the synchronization constraints and potential bottlenecks in the network. It reveals the chains of compute nodes that belong to the different parts of the computation taking place in the network. In other words, it reveals the different data-streams that are processed in the network. If many different data-streams go through one node, then this node may be a synchronization bottleneck for those data-streams. A measure is needed to quantify this concurrency property. Many parallel data-streams can be a sign of good utilization of data parallelism.

A data path through a network is a sequence of nodes from a network input to a network output. In the presence of cycles (feedback loops), there are infinitely many such paths. Therefore, we restrict our paths to go through at most one feedback loop. This prevents grouping of paths with different feedback loops, which are in fact different data-streams, in one path. A path p_1 is called a *sub-path* of p_2 if the nodes on path p_1 are a subset of the nodes on path p_2. Path p_2 is in that case called a *super-path* of p_1. The paths that are present in a computational network can be grouped into *computational paths*.

Definition 3.5 (Computational path) *A computational path is defined as the tuple (u, u', P) with u and u' respectively an input node of the network and an output node of the network and P a set of paths. For every path $\langle v_0, v_1, v_2, \ldots, v_k \rangle \in P$, it holds that $v_0 = u$ and $v_k = u'$. For the set of paths P, the following must hold:*

1. *For each $p_1, p_2 \in P$, p_1 is a sub-path of p_2 or p_2 is a sub-path of p_1 (i.e., P is totally ordered using the sub-path relation);*
2. *P is maximal, i.e., there is no path p not in P that can be added to P such that P is still totally ordered.*

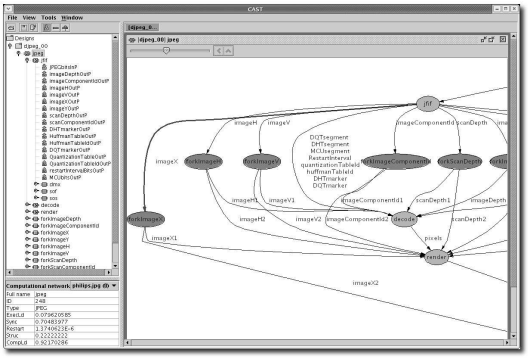
Note that requirement 1 in the above definition excludes the possibility that two feedback loops are part of one computational path; requirement 2 implies among others that it is not allowed to skip feedback loops.

The computational paths in a computational network represent the different data flows that go through the network. Exploiting parallelism implies that it is tried to maximize the number of different data flows. They must share as little compute nodes as possible to avoid synchronization bottlenecks. This observation leads to the definition of the *structure* measure. The measure is zero if all computational paths go through all nodes, which implies that there is no structural parallelism in the structure. This is for example the case for a pipeline structure. A value close to one indicates that the structure of the computational network is very parallel. The bottleneck for the structure are thus the nodes through which the most computational paths go.

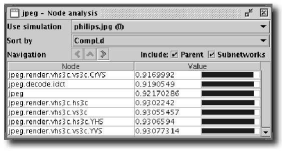
Definition 3.6 (Structure) *The structure of computational network CN and a compute node $n \in N$ are defined as follows:*

$$Struct(CN) = \frac{\sum_{n \in N} Struct(n)}{|N|}$$

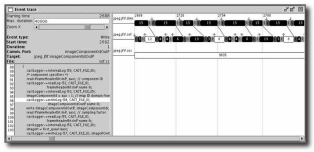
$$Struct(n) = 1 - \frac{|comp.\ paths\ through\ n|}{|comp.\ paths\ in\ CN|}$$



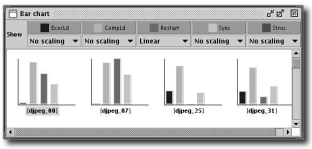
(a) Main window of CAST.



(b) Analyze details of nodes.



(c) Event diagram.



(d) Compare solutions.

Figure 4. Screen-shots of CAST.

Example 3.5 We continue with our running example. The network ECN contains three paths: $p_1 = \langle a, b, d \rangle$, $p_2 = \langle a, b, c, b, d \rangle$, $p_3 = \langle a, e \rangle$. These paths can be grouped in the computational paths $cp_1 = (a, d, \{p_1, p_2\})$ and $cp_2 = (a, e, \{p_3\})$. Two computational paths go through node a, whereas all other nodes belong to one computational path. Thus node a is a potential bottleneck. The structure for the overall network is then:

$$Struct(ECN) = 1 - \frac{2+1+1+1+1}{2} \cdot \frac{1}{5}$$
$$= 1 - \frac{6}{2 \cdot 5} = 0.4$$

So, on average a node performs transformations to 60% of the data-streams in the computational network. □

The different concurrency measures analyze different concurrency properties. Omitting one measure may result in a non-optimal solution. Consider, for example, the situation in which we would ignore the computation load. Assume now that the nodes in the network are split into a set of nodes in which each node contains only a single internal event and the required read and write events. The solution would have a very good restart measure and synchronization measure. Depending on the data-dependencies in the application it will also have a good structure measure. Typically most internal events have almost the same duration - i.e. the processing load of the network will also be good. So, these measures all indicate that this is a good solution. However, this solution is not good as there is an enormous communication overhead. This overhead is indicated by the computation load. Similar cases can be constructed for all other measures. So, none of the five measures can be omitted from the model without introducing the risk that a concurrency optimization process ends in some local optimum, focusing too much on one or a few concurrency aspects.

4. Prototype implementation

4.1. Time-stamping mechanism

Section 2.2 defines the execution of a computational network using a time-stamping mechanism based on Lamport's logical clocks. To implement this time-stamping mechanism, we need duration functions that associate a duration with the communication of data over a connection and with each event that occurs in the compute nodes. The duration functions are implemented in three parts:

Internal events. The time-stamping mechanism should allow reasoning about causality and some timing aspects on a relatively high level of abstraction without referring to implementations/physical time. Therefore, in our implementation, we define the duration of internal events (C++ statements) as the number of assembler instructions needed to execute these events on a processor using a standard compiler. We assume that the influence of a specific instruction set does not have too much influence on the results. Our experiments show that the proposed notion of time is both accurate and abstract enough to perform optimization independent of the exact processor chosen from a class of processors. As an alternative, we allow to do statistical analysis over a number of compilers/instruction sets taking for example the average number of instructions as the duration.

Read/write events. The duration function for read/write events must assign a duration to each read or write event that represents in some way the time required to communicate the data to the connection. In other words, it represents the actual time needed to call the read or write function and to transfer the data to the connection interface. This duration can be approximated using a linear function $ax + b$ with x the number of data elements communicated. The constant b approximates the time needed to call the communication primitives; the constant a approximates the time needed to read/write one data element.

Connections. The duration function for connections is

implemented as a linear function $cy + d$ with y the size of the data elements going through a connection. The constant d approximates the access time of the communication medium; the constant c approximates the time needed to transport one data-element. This linear function models all relevant aspects of communicating data over a connection. One may only say that sharing of communication resources is not taken into account. However, this does not need to be taken into account if we have reserved connections, e.g., independent virtual connections multiplexed over one physical connection. Alternatively, at the targeted abstraction level, one can take the average penalty introduced by sharing into account through the constants of the linear function.

4.2. CAST

The previous section introduced a concurrency model that allows analysis of five different concurrency properties of a computational network. The concurrency model uses the structure of the computational network and the event diagrams that can be obtained by executing the network. To construct an event diagram, a list of all events that occur in all nodes during an execution with a given input is needed. The definition of a compute node leaves the possibility open that the behavior of a node is data-dependent. This implies that there need not to be a single, unique event diagram for a computational network. So, to construct an event diagram, the computational network must be executed (i.e. simulated) with an input. Simulating the network does not contradict with our desire for abstraction. To allow abstraction from a single input, we can use multiple simulations and perform statistical analysis on them.

The software tool CAST can be used to compute the concurrency measures of a computational network. It can create and analyze an event diagram of a network which is simulated with a given input. The tool can also perform statistical analysis on the results of multiple simulations. CAST takes as input a set of C++ files which describe the computational network as a Kahn process network in YAPI [5]. It starts with annotating the original source code of the computational network with the functions for tracing all events during a simulation. It also maps each internal event onto a duration as described in Section 4.1. Next, CAST simulates the annotated computational network with the supplied input to trace all events which occur in the network. An event diagram is constructed from the traced events. After that, CAST has enough information to compute the values of the different concurrency measures. The concurrency measures can be visualized through the graphical user-interface of CAST, via bar charts and coloring and sizing of nodes. CAST provides a direct coupling between the events in the event diagram and the corresponding source code. Figure 4 shows a few screen-shots of CAST (operating on a JPEG decoder). The user-interface helps the designer to identify and resolve concurrency bottlenecks and to compare the impact of different design decisions. It supports a design-space exploration method, as explained in the next section.

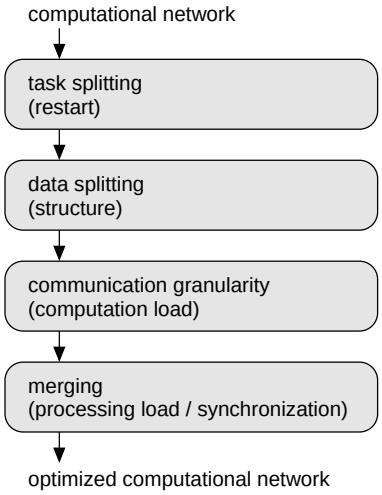


Figure 5. Exploration method.

5. Design-space exploration

A prototype concurrency-analysis tool should not only implement the computational-network model and the concurrency measures, but it should also provide support for design-space exploration, i.e., it should support the designer in finding a computational network with optimal (good) concurrency properties. This section presents a generally applicable design-space exploration method consisting of four steps which uses the concurrency model to realize this goal. Each step optimizes one or two concurrency measures. Figure 5 shows an overview of the design-space exploration method and the measures optimized in each step. The different steps of the method are explained by deriving in a structured way an implementation of a JPEG decoder [9] that has a balanced workload and good communication behavior. The basic idea of the design-space exploration method is to first identify and extract all the available concurrency in an application and then design a network that optimally exploits this concurrency. The concurrency measures are used in the design-space exploration method to identify concurrency bottlenecks. A designer must resolve these bottlenecks by modifying the computational network, i.e., a designer is responsible for the extraction of the concurrency from the computational network. The (partial) automation of concurrency extraction is left for future work.

The five concurrency measures provide a global direction when optimizing the concurrency. They also provide insight in the concurrency bottlenecks of the network. In addition to those five measures, we use a *global concurrency measure* to steer the design-space exploration. It considers all concurrency measures of the computational network

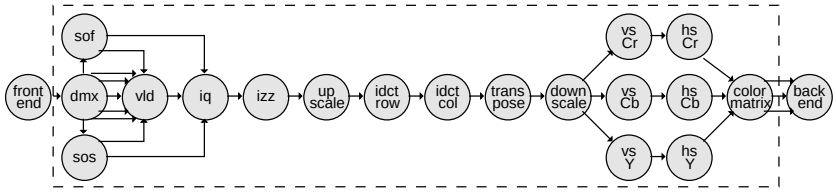


Figure 6. A JPEG decoder (Design 0).

CN that is obtained after i steps of the design-space exploration. The measure is defined as follows:

$$Conc(CN, i) = \frac{\sum measures\ optimized\ till\ step\ i}{5} \quad (1)$$

The design-space exploration remains in the same step as long as new transformations to the network are possible that improve the global concurrency measure. When such transformations are no longer possible, the design-space exploration advances to the next step. The exploration ends when the global concurrency measure does not provide any further improvement in the final step of the method.

Starting point. An experienced designer of Philips research optimized the JPEG decoder for a given multiprocessor architecture [4]. We started with the same computational network as this designer used as a starting point. Using the same starting point gives us a fair comparison between the end results. The computational network of this JPEG decoder is shown in Figure 6 and is referred to as design 0. The frontend and backend nodes model the environment of the network we want to optimize. These nodes are not taken into account in the analysis. The details of JPEG are not relevant for the remainder. Concurrency measures for this design are calculated using the statistical option of CAST and a set of five different images. Figure 7 shows the results for design 0 of the JPEG decoder, as well as for some other designs discussed further-on. It also shows the development of the global concurrency measure. For the starting point, this measure is, by definition, 0.

Task splitting. The design-space exploration method starts with task splitting. The goal of this step is to extract the available task-parallelism from the application by splitting compute nodes as far as possible. The step must optimize the *restart measure*. The candidates for improvement are the compute nodes with a low restart. For the JPEG case, we selected and modified the four nodes with the lowest restart in design 0. This modification consists of identifying sub-tasks in each node that can be split-off from the node into new nodes. For example, the *idct col* and *idct row* nodes can both be split into a sequence of four nodes which all perform part of the IDCT conversion. This increases the pipelining of data in the network. The design resulting from the modification of all four nodes is labeled design 1. Figure 7 shows the result of the transformation on the concurrency measures. It shows that the restart has indeed improved. Observe that the absolute restart values computed during the exploration are all very low. Figure 7 shows normalized values as explained in Example 3.3. The global concurrency measure for this design is also shown in Figure 7. Note that we only show the global concurrency measure at the end of each step of the exploration method. However, also if transformations in one step are done individually in sequence, the measure increases monotonically.

Data splitting. The data-splitting step aims at extracting coarse-grained data-parallelism in the application. The amount of data-parallelism in the network is visible in the *structure measure*; this step should improve this measure. Nodes that are a bottleneck according to the structure measure are considered in this step. Figure 6 shows that the data is processed in three parallel streams (i.e., three color components) between the *downscale* and *color matrix* nodes. It is possible to create these parallel streams already after the *vld* node. This will increase the data-parallelism. The *vld* node has to process the bitstream sequentially. Hence, it contains no data-parallelism. In the JPEG decoder of design 1, we created separate computational paths for the three color components (design 2). The concurrency measures (see Figure 7) show that the goal of this step is realized as both the value of the structure measure and the value of the global concurrency measure have increased. The increase of the global measure indicates that the increase of the structure measure outweighs the negative effect on the restart measure.

Extraction of data-parallelism requires typically duplication of nodes in the network. By performing data splitting after task splitting, the network is kept relatively small in the first step of the exploration method. This helps the designer in keeping an overview of the network and it avoids that changes have to be made to several copies of the same node.

Communication granularity. The cost of communication is ignored during the extraction of concurrency from an application. However, this cost will play an important role in the granularity of communication used in the final implementation. We observe that calling a function that implements the communication primitives is more expensive than a normal memory operation that is part of a normal internal event. To respect this observation, we assigned a

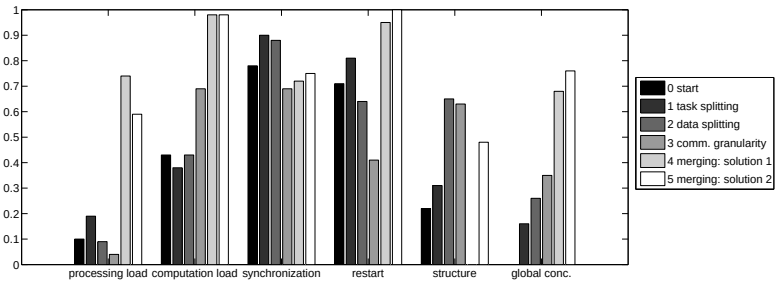


Figure 7. Concurrency measures for the JPEG decoder.

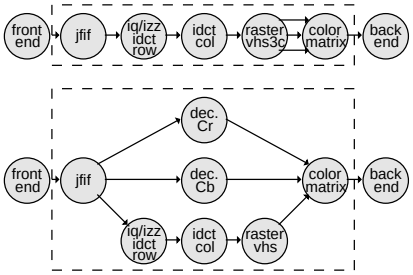


Figure 8. JPEG decoder (Design 4 and 5).

constant delay of 30 logical clock values to each read/write event. Note that the exact costs are not important; they must only respect the above observation. The statistical analysis function of CAST has been used to verify the concurrent behavior for different cost functions, showing that a delay of 30 is reasonable. Other values, possibly depending on the size of events, do not really affect the relative values of the measures.

The communication granularity step adjusts the size at which data is communicated between the nodes in the network. The objective is that nodes can execute in parallel, but do not spend too much time on communication. This step tries to optimize the *computation load* measure. The nodes in design 2 communicate single pixel values. In design 3, which is the end result of this design-space-exploration step, the nodes communicate blocks of pixels at once. Figure 7 shows that this transformation improved the computation load of the network. It also impacts the other concurrency measures. Despite the decrease in their value, the global concurrency measure indicates that this transformation is good, as its value increases from 0.26 to 0.35.

A trade-off between the communication granularity and the concurrency in the network can only be made after the concurrency has been made explicit in the network. So, the communication granularity step should be performed after the concurrency extraction (first two steps of the method).

Merging. Some of the available parallelism is removed in the merging step with the objective to obtain a more balanced workload. This is done by merging nodes in the network and removing synchronization bottlenecks. It should result in a processing load close to one. The objective of this step is to optimize the *processing load* and *synchronization* measure simultaneously. To allow a good workload balance, this step should be performed after all task-level parallelism has been made explicit in the network and the communication granularity has been decided.

There are three solutions possible to realize a balanced workload for the JPEG decoder. They differ in the amount of data-parallelism that is preserved in the final solution. First, we can remove all data-parallelism and then remove some of the task-parallelism (solution 1). This results in the computational network shown topmost in Figure 8 (design 4). Another solution is to preserve all data-parallelism and remove only the functional task parallelism, design 5, shown at the bottom in Figure 8. The third solution would be to remove some of the data-parallelism, but not all. This third solution is not further explored in this case study; we focus on the two extreme solutions. Figure 7 shows that both designs 4 and 5 meet the goal of the merging step, namely a high processing load. The structure measure shows that design 5 still contains data-parallelism, while design 4 does not. Both designs have similar values for synchronization and computation load. Synchronization does not play an important role in this case study because of the regular communication patterns. Figure 7 shows that all seven designs give a similar result when compared to a purely sequential version. However, this does not mean that they are all equally good. The synchronization measure can be interpreted as the speed-up which can be achieved if maximal parallelism can be exploited, for example by using one processor per compute node. So, designs 4 and 5 are equally fast when compared to the sequential solution, but 5 may need more resources. The latter is also visible in the lower processing load for design 5. Note that the precise resource usage and throughput in an implementation depend on mapping and scheduling decisions. As we will see in the next section, designs 4 and 5 perform equally well when dynamically scheduled on a homogeneous multiprocessor. The nodes in design 4 have less idle-time during the execution than the nodes in design 5. Design 5 has however a higher restart measure. This implies that it may have a higher throughput than design 4 in an implementation.

6. Case studies

In this section, we present three case studies in which CAST and the design-space exploration method are used to derive in a structured way a computational network with optimal (good) concurrency properties.

6.1. JPEG decoder

The actual design-space exploration of the JPEG decoder was performed in the previous section. The concurrency

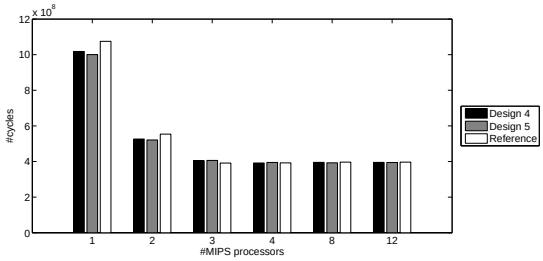


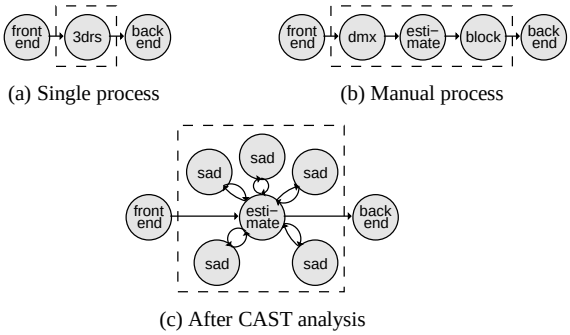
Figure 9. Mapping on multi-processor platform.

measures of the resulting solutions, shown in Figure 7, indicate that designs 4 and 5 are good solutions. To verify this, we mapped these solutions and a reference solution from [4] on the same multi-processor system [30] as used in [4]. This system consists of a set of MIPS processors and a set of memories that communicate through a snooping interconnection network. All processors in the system operate on a single queue of runable tasks. A small operating system dynamically assigns tasks to processors.

To compare the performance of designs 4 and 5 and the reference design of [4], we simulated these designs for different numbers of MIPS processors. The results of these simulations are shown in Figure 9. The figure shows that the solutions derived in our case study have the same performance characteristics as the reference design, which is a good result considering that our analysis and design-space exploration has been done independent of the multi-processor architecture. Our designs are slightly faster when one or two processors are used. The reference design has the best performance when 3 processors are used. However, the difference is not really significant. Designs 4 and 5 have similar performance for three or more processors.

6.2. 3D recursive search

This section presents a second case study that demonstrates the effectiveness of the concurrency model; it also shows the need for a higher level of abstraction than cycle-accurate simulations. The case study implements a parallel version of a sub-pixel accurate motion estimator using a 3D recursive search algorithm (3DRS). This case study started from an implementation of the 3DRS algorithm written in C. The first step involved separating the actual algorithm from the code that is needed to simulate the environment (e.g., read and write files to disk). This resulted in the computational network shown in Figure 10(a). The whole 3DRS algorithm is implemented as sequential code in a single compute node. A designer then parallelized the algorithm by hand. This resulted in the computational network shown in Figure 10(b), which was considered optimal by the designer. Analysis of the concurrency properties using CAST showed that the motion estimator, node *estimate*, is a bottleneck. A study of this node revealed that the motion estimator has to compute five sums-of-absolute differ-



(a) Single process (b) Manual process

(c) After CAST analysis

Figure 10. 3D recursive search.

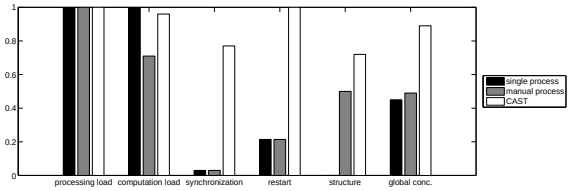


Figure 11. Concurrency measures for the 3DRS implementations.

ences (SADs) on the same data-set. These computations can be performed in parallel. The CAST analysis showed also that the *dmx* and motion *block* nodes should be integrated with the *estimate* node after extraction of the SADs. These changes were implemented by the designer and led to the design shown in Figure 10(c). The concurrency measures for the three different solutions are shown in Figure 11. The measures indicate that the largest gain is achieved in the change from the manual solution to the solution found using CAST.

It is not possible to evaluate the three designs using the same multi-processor architecture as used for the JPEG decoder case-study. Simulation of a small movie sequence consisting of 6 frames takes 4 hours to complete for a single design using a 1GHz P3 with 4GB of memory. The cycle counters used in the cycle-accurate simulator overflow during this simulation. They are too small to hold the actual cycle count. This makes it impossible to obtain, in this way, performance measures for our designs.

A notion of the speed-up of the design can be obtained from CAST. It computes the execution time of the computational network (see Definition 2.4). This time is equal to the time needed to execute the computational network on a multi-processor system if each node is mapped onto a different processor. This can be used as an estimate of how long the computational network will run on a system that contains as many processors as there are nodes. Since the times computed by CAST are based on instruction counts, they are quite accurate. We can also calculate the time that the system will need when it is executed on a single pro-

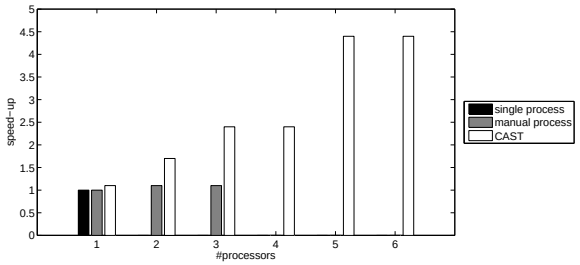


Figure 12. Speed-up of 3DRS implementations.

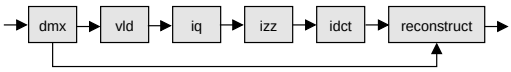


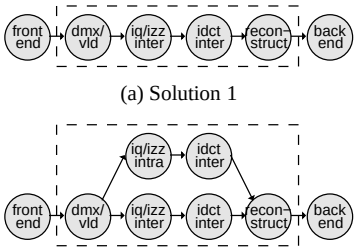
Figure 13. H.263 decoder.

cessor; this time is equal to the sequential execution time (see Definition 2.8). An estimate of the required execution time for a system that contains more than one processor, but less than the number of nodes in the network can also be made. For this estimate, we use the most optimal mapping and scheduling of the nodes on the processors. The schedule determines which nodes can execute in parallel and which nodes have to execute in sequence due to their data-dependencies. Combining the schedule with the execution times of the individual nodes, as computed by CAST, the total execution time of each mapping is estimated. To obtain all required time measures, we simulated the designs with the movie sequence and analyzed the execution using CAST. This required approximately 5 minutes for each design. The resulting performance numbers are shown in Figure 12. These results are normalized with respect to the execution time for the sequential design (Figure 10.a). The results show that both parallel implementations have a speed-up when they are executed on a multi-processor system. They show also that the design found using our concurrency model has a considerably higher speed-up than the design found by the designer.

This case study shows that our concurrency model helps in finding task-level concurrency in an application that can be extracted by a designer. The case study shows also that CAST is useful in getting fast and still accurate performance estimates at a relatively high-level of abstraction.

6.3. H.263 decoder

H.263 is a standard video-conferencing codec optimized for low data rates and relatively low motion [10]. The codec was used as a starting point for the development of the MPEG-II codec which is optimized for higher data-rates. The structure of an H.263 decoder is shown in Figure 13. The H.263 decoder supports three types of frames: I-frames, P-frames and PB-frames. To decode a PB-type of frame, the reconstruct uses the previous and next decoded frame and the decoded blocks. For a P-type frame, the reconstruct uses the previous decoded frame and the decoded



(a) Solution 1

(b) Solution 2

Figure 14. H.263 decoder

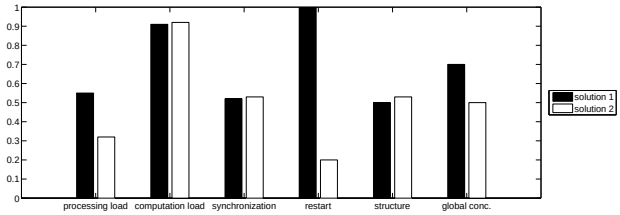


Figure 15. Concurrency measures for the H.263 decoder.

blocks. For an I-frame, only the decoded blocks are used as it has no dependencies with data from other frames.

A designer, with no background in video coding, was asked to extract the available concurrency from a given sequential C specification of the H.263 decoder. First, the designer split the decoder into the tasks shown in the block diagram of Figure 13. The designer analyzed this computational network with CAST to identify potential concurrency bottlenecks. This led to a number of transformations on the network, which were analyzed using CAST. This process was repeated until two final solutions were found. The first solution implements the decoder as a pipeline of nodes (see Figure 14(a)). The second solution exploits the option that I-frames can be processed independent of P- and PB-frames (see Figure 14(b)). The concurrency measures for both solutions are shown in Figure 15. The synchronization measure shows that both solutions are estimated to be about twice as fast as the original sequential solution. The synchronization measure can be interpreted as the speed-up which can be achieved if maximal parallelism can be exploited, i.e., usually one processor per compute node. So, solution 1 and 2 are both twice as fast, but 2 needs more resources which is visible in the lower processing load. The solutions differ on the estimated throughput, measured by the restart measure. The reason for this is that the merging of the two data streams in the reconstruct node of the second solution adds additional complexity to this bottleneck, making it only slower.

Unfortunately, it is impossible to benchmark these solutions using a cycle-accurate simulator as it has problems with the required simulation length. It is also not possible

to calculate the speed-up in a similar manner as done in the previous case study. The reason for this is that it is practically impossible to find out when which nodes can execute in parallel. However, the case study does show that a designer with no background in the application domain is able to quickly identify different sources of concurrency using our concurrency model and CAST.

7. Architectural properties

The computational-network model and the concurrency measures neglect most architecture properties. It is important to observe that the essential part of the model underlying our approach is the event diagram. Any architecture aspect that can be taken into account in the event diagram can be handled by our approach without any modifications. This includes aspects like scheduling strategies, heterogeneous processing elements, buffer sizes, etc. In this section, we discuss two examples of how to take architecture properties into account in the concurrency analysis.

Heterogeneous architecture. The assignment of a duration to an internal event (see Section 4.1) assumes implicitly that a homogeneous platform is used as it uses the same compiler for all nodes to relate their internal events to a duration. In practice, the used multi-processor system may be a heterogeneous system. We see two different solutions to take this heterogeneity into account in our concurrency measures. The first solution assumes that a mapping of nodes to processor types is made. In that case a different compiler for the different node mappings can be used. So, for each node the mapping of an internal event onto a duration is based on the compiler that comes with the processor to which this node is mapped. The second solution would be to use scaling factors for the durations of the internal events of the different nodes. For example, assume that a node *a* is mapped onto a processor which is twice as fast as the processor to which a node *b* is mapped. Then the duration of each internal event in *b* should be multiplied with two to take this difference in speed into account. The second approach can also be used to model the effect of hardware accelerators, i.e., when a node is not mapped onto a programmable processor but directly implemented in hardware.

Buffer sizes. The computational-network model assumes implicitly that the connections used between the nodes have infinite capacity. This implies that the execution of a compute node can never be blocked on a full connection. In practice, a connection will be assigned a buffer of finite size to store data as the amount of available memory in a system is limited. As a result, a node which produces data may have to wait until there is enough space in the connection buffer. In other words, the producing node must wait until the consuming node has read enough data elements from the connection. This dependency between

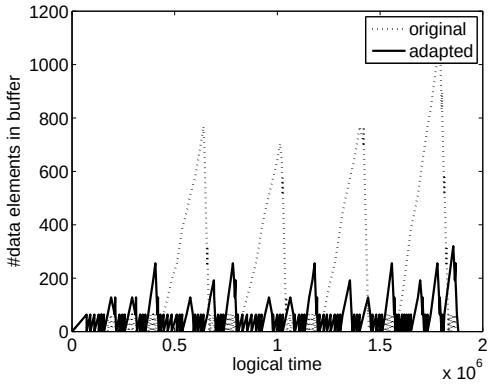


Figure 16. Buffer size requirements

the producing and consuming node will affect the event diagram and thus the concurrency measures. If a separate buffer with fixed size is assigned to each connection, then we can analyze this impact in the following way. During the construction of the event diagram the used buffer size of each connection is counted over time. If insufficient space is available, a node stalls the execution of a write event (i.e. it inserts idle time) till there is enough space on the connection to write the data elements. The resulting event diagram can then be analyzed in the normal way and the impact of the buffer size on the concurrency measures becomes visible. Note that this allows a fast exploration of the effect of buffer sizes on the concurrency properties of a design, without the need for re-executing the application.

It may also be interesting to study the number of data elements that are stored over time in the buffer and to take this information into account in the concurrency optimization. For example, the dotted line in Figure 16 shows the number of data elements stored in the connection between two of the compute nodes in a JPEG decoder design over time. Typically only 64 or 128 data elements are stored in the connection at the same time. However, there are a few points at which many more data elements need to be stored. Furthermore, it is clear that the node which produces the data elements does this at a more or less constant rate. So, the consuming node causes the large buffer requirement. Analysis of the source code of this node revealed that each time after it had received a certain amount of data a transformation was applied on it. While this transformation was executing, the producing node continued filling the connection. This transformation, however, could also be performed after a small amount of data was received - i.e. the transformation could be distributed more evenly over the execution time of the consuming node. Applying this transformation to the source code of the consuming node led to the buffer size requirements shown in Figure 16 as the ‘adapted’ version. The required buffer size is reduced with a factor three while the concurrency properties of the design are preserved.

8. Related work

Multi-processor systems inherently contain concurrency. This concurrency must be exploited in the programming trajectory. This requires a model of computation which allows the specification of concurrency in an application and the application should fit easily in it. Furthermore, the description of the application should be at the correct abstraction level to perform the required analysis [14]. A comprehensive survey of models for parallel computations used in different application domains and at different abstraction levels is given in [28]. The most interesting models for our application domain, streaming multimedia systems, are the dataflow models. Examples of these are Kahn process networks [12, 13] and Synchronous dataflow [18]. In [19], a framework is presented to compare the notions of concurrency, communication, and time between different dataflow models. Based on the desired notions, a designer can pick a model of computation that fits best with the application domain to describe the software and hardware behavior of a system component. For an entire system, these system components might be described in different models of computation. Interaction of different models of computation is captured by the Ptolemy framework [2]. It enables analysis of a system composed out of components that are described in different models of computation. Our model of computation is an abstraction that generalizes all commonly used models, thus enabling system-level concurrency analysis.

Concurrency optimization is studied in the field of systems-on-chip design as part of the problem of multi-processor programming. Typically, concurrency analysis is (an implicit) part of a design flow that maps an application onto a multi-processor system. An extensive overview of the different design-flow approaches and a classification according to the used optimization criteria and abstraction level can be found in [7]. Artemis [24] is one of the projects mentioned in this paper. It is based on a Kahn process network description of the application and incorporates the ideas from SPADE [20], i.e., system-level co-simulation is performed by using symbolic instruction traces generated and interpreted at run-time through manually defined abstract performance models. In [6] a technique to perform a multi-objective design-space exploration with Artemis is presented. Our approach provides an automatic way to derive performance numbers simplifying the analysis. We also provide a design-space exploration method to which helps in extracting concurrency from an application. Other design-space exploration methods are Milan [23], Mescal [22] and Metropolis [1]. Milan combines tools for design-space pruning with simulations at different levels of abstraction. Simulators include trace-driven, task-level evaluation tools as well as cycle-accurate third-party simulators. Mescal provides a correct-by-construction mapping flow targeting heterogeneous, application-specific, programmable (multi-) processors. Applications can be specified in any combination of models of computation that is natural for the application. The Metropolis framework allows the description and refinement of a design at various levels of abstraction. Applications are modeled as a set of communicating processes. The performance numbers generated by the simulations tools of Metropolis are based on user-specified annotations. All three approaches focus on performing a design-space exploration. None of them aims explicitly at analyzing concurrency from an application and identifying concurrency bottlenecks in the application, while this is the explicit goal of our work. Our work is complementary to these design-space exploration tools and it can be integrated into their flow.

A number of approaches exist in the field of system-on-chip design that explicitly aim at concurrency analysis and extraction. A good example of this is the Compaan tool [15]. It automatically transforms nested loop programs into a process network which makes the concurrency explicit. The tool finds all possible concurrency contained in the application. Abstraction of the functional behavior of an application is often required as Compaan expects that an application is described as a looped structure. Compaan requires a designer to make this abstraction without support from the tool. Our concurrency analysis technique requires the same abstraction of the application. However, the required abstraction can be obtained by simulating the application in our tool with a set of representative inputs. No changes have to be made to the description of the application. Also, the most concurrent program is not always the best program, e.g., due to communication overhead. Our work defines a concurrency model and exploration method that takes all relevant concurrency aspects into account. Another approach to concurrency analysis is found in [29]. The authors present a technique to identify task-level concurrency independent of the target architecture. The approach is defined on JAVA program constructs and lacks a more formal underlying model. Our work defines such a model. One concurrency measure that considers the longest path in a task graph is defined. This measure is similar to our synchronization measure, but we show that more measures are needed to avoid optimization of one single aspect.

Analysis and extraction of concurrency from applications has also been studied extensively in the field of distributed computing. Most techniques analyze the data-dependencies between different parts of the application. These data-dependencies are expressed with a partial order [25] or a message sequence chart [8]. Ravindran et al. describe in [26] the different sources of concurrency that can be identified by looking at the causality relations between events that occur in an application. A message sequence chart is used to express these causality relations and to define a concurrency measure. This measure considers the

ratio between the number of orders in which messages can be communicated in the graph (all possible ways to interleave the messages while obeying the precedence relations) and the number of orders in which messages can be communicated in the graph if there are no precedence relations. Raynal [27] developed another concurrency measure, that is closely related to our processing load. The difference is that the processing load takes all communication idle-time into account and not only the idle-time caused by the causality relations as done by Raynal's measure. A second important difference between the two approaches is that the model of computation used by Raynal neglects the time needed to communicate data between tasks. This is not a valid assumption in the domain of streaming multimedia applications. In these applications, large amounts of data are communicated between tasks. Different mappings of the tasks onto the platform result in a different timing behavior of the application as these mappings may require different use of the on-chip interconnect. For this reason, communication time is taken into account in our concurrency model. One of the important goals of Raynal's paper and our work described in this paper is to provide analysis techniques independent of real time effects, such as system load and processor speed (i.e., independent of the hardware architecture). This is different from most other concurrency optimization techniques in the field of distributed computing. Typically, concurrency optimization is performed for a given system architecture [21]. A good example of this is the research to performance analysis and design optimization for systolic processors [11, 16]. Commonly considered optimization criteria are the latency, throughput and the number of processors. The synchronization and restart measure in our concurrency model focus on the first two aspects. The third aspect is not considered as we assume that dimensioning of the platform is done in a later stage of the design-flow. An important contribution of our work compared to all the discussed approaches is that we provide a model that covers - at the targeted abstraction level - all relevant concurrency aspects for streaming applications, and not only a subset of these aspects as all other approaches, and that we provide a supporting design-space exploration method.

9. Conclusion

In this paper, we presented a concurrency model with a supporting design-space exploration method and an analysis tool that allow reasoning about concurrency in streaming applications at the executable-specification level. The model consists of a set of five system-level measures that provide guidance when optimizing the concurrency, a set of detailed measures that provide insight in concurrency bottlenecks and one overall global measure that takes into account the steps in the design-space exploration methods and that steers the exploration. The presented examples and case studies show that these measures are meaningful, do not (fully) overlap, allow reasoning about concurrency and are sufficient for obtaining good results. Our method still gives results with analysis times in the order of minutes when cycle-accurate simulations are no longer feasible due to long simulation times and extremely high cycle counts. The JPEG decoder case study furthermore shows also that the concurrency model and accompanying design-exploration method allow concurrency optimization independent of the exact target architecture; when the end result is implemented on a homogeneous system of MIPS processors, the performance is similar to an optimized design implemented on that architecture by an experienced designer. The 3D recursive search case study illustrates that our concurrency model can be useful in getting fast and accurate performance estimates at a relatively high-level of abstraction. The H.263 decoder case study shows that also for inexperienced designers it is possible to quickly identify different sources of concurrency using our concurrency model and CAST.

Future work includes more experiments with different applications and architectures to verify the assumptions made in the concurrency model and to fine-tune the model. We also plan to extend the concurrency model to take architecture information into account for the architecture-dependent step of the design process. Costs of communication may have a large impact on system performance, meaning they must be estimated accurately. We want to study the modeling of these costs in more detail to get a model that provides abstract but accurate information about these costs. Another direction of future work is tool support for the automatic extraction of concurrency from a computational network.

Acknowledgment. We want to thank Erwin de Kock for developing YAPI and giving access to all results of his JPEG case study, Paul Stravers and Jan Hoogerbrugge for providing the CAKE multiprocessor-architecture simulator and Jef van Meerbergen, Henk Corporaal, Johan Lukkien and the anonymous reviewers for their comments and suggestions, Jan Ypma for his work on the graphical user interface, and Andre Carmo and Rik Kneepkens for their work on the experiments. Part of this work was supported by the Dutch Science Foundation NWO, project 612.064.206, PROMES.

References

[1] F. Balarin, Y. Watanbe, H. Hsieh, L. Lavango, C. Paserone, and A. Sangiovanni-Vincentelli. Metropolis: An integrated electronic system design environment. *IEEE Computer*, 36(4):45–52, April 2003.

[2] J. Buck, S. Ha, E. Lee, and D. Messerschmitt. Ptolemy: A framework for simulating and prototyping heterogeneous systems. *Int. Journal of Computer Simulation*, 4(2):155–182, April 1994.

[3] F. Commoner, A. Holt, S. Even, and A. Pnueli. Marked directed graphs. *Journal of Computer and System Sciences*, 5(5):511–523, October 1971.

[4] E. de Kock. Multiprocessor mapping of process networks: A jpeg decoding case study. In *ISSS'02, 15th Int. System Synthesis Symposium, Proc.*, pages 68–73. ACM, 2002.

[5] E. de Kock, G. Essink, W. Smits, R. van der Wolf, J.-Y. Brunei, W. Kruijtzer, P. Lieverse, and K. Vissers. YAPI: Application modeling for signal processing systems. In *37th Design Automation Conference, Proc.*, pages 402–405. IEEE, 2000.

[6] C. Erbas, S. C. Erbas, and A. D. Pimentel. A multiobjective optimization model for exploring multiprocessor mappings of process networks. In *CODES/ISSS'03, Int. conf. on hardware/software codesign and system synthesis, Proc.*, pages 182–187. ACM, 2003.

[7] M. Gries. Methods for evaluating and covering the design space during early design development. *Integration, the VLSI Journal, Elsevier*, 38(2):131–183, December 2004.

[8] L. Hélouët and P. Le Maigat. Decomposition of message sequence charts. In *SAM'00, 2nd Workshop on SDL and MSC, Proc.*, pages 46–60, 2000.

[9] ITU. Information technology - digital compression and coding of continuous-time still images. ITU Recommendation T.81, September 1992. ITU, 1992.

[10] ITU. Video coding for low bit rate communication. ITU Recommendation H.263, February 1998. ITU, 1998.

[11] K. T. Johnson, A. R. Hurson, and B. Shirazi. General-purpose systolic arrays. *IEEE Computer*, 26(11):20–31, November 1993.

[12] G. Kahn. The semantics of a simple language for parallel programming. In *Information Processing 74, Proc.*, pages 471–475. Amsterdam, The Netherlands: North-Holland, 1974.

[13] G. Kahn and D. MacQueen. Coroutines and networks of parallel processes. In B. Gilchrist, editor, *Information Processing '77, Proc.*, pages 993–998. Amsterdam, The Netherlands: North-Holland, 1977.

[14] K. Keutzer, A. R. Newton, J. M. Rabaey, and A. Sangiovanni-Vincentelli. System-level design: orthogonalization of concerns and platform-based design. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 19(12):1523–1543, December 2000.

[15] B. Kienhuis, E. Rijpkema, and E. Deprettere. Compaan: deriving process networks from matlab for embedded signal processing architectures. In *Int. Conf. on Hardware Software Codesign, Proc.*, pages 13–17. ACM, 2000.

[16] S. Kung. *VLSI array processors*. London, UK: Prentice Hall, 1998.

[17] L. Lamport. Time, clocks, and the ordering of events in a distributed system. *Communications of the ACM*, 21(7):558–565, July 1977.

[18] E. Lee and T. Parks. Dataflow process networks. *Proceedings of the IEEE*, 83(5):773–801, May 1995.

[19] E. A. Lee and A. Sangiovanni-Vincentelli. A framework for comparing models of computation. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 17(12):1217–1229, December 1998.

[20] P. Lieverse, P. van der Wolf, E. Deprettere, and K. Vissers. A methodology for architecture exploration of heterogeneous signal processing systems. *Journal of VLSI Signal Processing*, 29(3):197–206, November 2001.

[21] A. Mazzeo, N. Mazzocca, and U. Villano. Efficiency measurements in heterogeneous distributed computing systems: From theory to practice. *Concurrency: Practice and Experience*, 10(4):285–313, May 1998.

[22] A. Mihal and K. Keutzer. *Networks on Chip, Mapping concurrent applications onto architectural platforms*, chapter 3, pages 39–59. Dordrecht, The Netherlands: Kluwer Academic Publishers, January 2003.

[23] S. Mohanty, V. K. Prasanna, S. Neema, and J. Davis. Rapid design space exploration of heterogeneous embedded systems using symbolic search and multi-granular simulation. In *Conf. on languages, compilers and tools for embedded systems, Proc.*, pages 18–27. ACM, 2002.

[24] A. Pimentel, P. van der Wolf, E. Deprettere, L. Hertzeberger, J. van Eijndhoven, and S. Vassiliadis. The Artemis architecture workbench. In *Progess Workshop on Embedded Systems, Proc.*, pages 53–62. Utrecht, The Netherlands: STW Technology Foundation, 2000.

[25] V. Pratt. Modelling concurrency with partial orders. *International Journal of Computer and Information Sciences*, 15(1):33–71, 1986.

[26] K. Ravindran and A. Thenmozhi. Extraction of logical concurrency in distributed applications. In *ICDCS'93, International Conference on Distributed Computing Systems, Proc.*, pages 66–73. IEEE, 1993.

[27] M. Raynal, M. Mizuno, and M. Neilsen. Synchronization and concurrency measures for distributed computations. In *12th International Conference on Distributed Computing Systems, Proc.*, pages 700–707. IEEE, 1992.

[28] D. Skillicorn and D. Talia. Models and languages for parallel computation. *ACM Computing Surveys*, 30(2):123–169, June 1998.

[29] R. Stahl, L. Rijnders, D. Verkest, S. Vernalde, R. Lauwereins, and F. Catthoor. Performance analysis for identification of (sub)task-level parallelism in java. In *Int. Workshop on Software and Compilers for Embedded Systems, Proc.*, pages 313–328. Springer, 2003.

[30] P. Stravers and J. Hoogerbrugge. Homogeneous multiprocessing and the future of silicon design paradigms. In *International Symposium on VLSI Technology, Systems and Applications, Proc.*, pages 184–187. IEEE, 2001.

[31] F. Thoen and F. Catthoor. *Modeling, verification and exploration of task-level concurrency in real-time embedded systems*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2000.