INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Another glance at Relay Stations in Latency-Insensitive Designs

Julien Boucaron — Jean-Vivien Millo — Robert de Simone

N° 5557

Avril 2005

Thème COM

ISSN 0249-6399 ISRN INRIA/RR--5557--FR+ENG

Another glance at Relay Stations in Latency-Insensitive Designs *

Julien Boucaron , Jean-Vivien Millo , Robert de Simone

Thème COM —Systèmes communicants
Projet Aoste

Rapport de recherche n° 5557 —Avril 2005 —19 pages

Abstract: We revisit the formal modeling of *relay stations*, which are specific connection elements used in the theory of Latency-Insensitive Design of Globally-Asynchronous/Locally-Synchronous systems. Relay stations are in charge of taking into account the physical mandatory latencies, while handling the regulation of signal/data traffic so as to avoid starvation, deadlock and congestion of local IP synchronous computation blocks. Since proposed by Carloni *et al*, the structure and behaviors of these relay stations have been amply characterized and analysed. But previous works never provided a fully formal and cycle-accurate description of these mechanisms, amenable to formal verification for instance (instead, mainly simulation models were developed). Due to the needed precision of the whole scheme we feel such a formal description might be needed. We describe such an attempt here. On its way, this work also led us to a number of (hopefully insightful) remarks on favorable and disfavorable graph topologies and initialization features, that are also reported here.

Key-words: Latency Insensitive, Relay Station, Shell, GALS, Formal Verification, Marked Graph, Synchronous, Esterel, SyncCharts

* This work is partially funded under a grant from Texas Instruments, Villeneuve-Loubet - FRANCE

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 —Télécopie : +33 4 92 38 77 65

Point de vue formel sur les stations relais dans la conception insensible aux latences de modèles GALS

Résumé : Nous revisitons la modélisation formelle des *stations relais*, qui sont des éléments de connection spécifiques utilisés dans la théorie du *Latency-Insensitive Design* (LID) pour la modélisation de type *Globalement Asynchrone, Localement Synchrone* (GALS) de systèmes sur puce (SoC). Les *stations relais* assurent la prise en charge des latences physiques imposées en décomposant les fils de connexion en sections. Elles gèrent également la régulation du traffic des signaux et des données de manière à éviter famine, blocage et congestion des composants de calcul synchrones locaux. Depuis que Carloni *et al* l'ont proposé, la structure et les comportements de ces *relay stations* ont été amplement caractérisés et analysés. Mais ces travaux restent informels. Nous proposons ici une modélisation formelle complète de ces éléments de nature synchrone. Avec la modélisation correspondante des modules d'encapsulation (*Shells*) des composants de calculs, ceci permet la vérification automatique des propriétés de correction attendues des composants et du système.
La modélisation passe par 3 étapes, dont une étape asynchrone entre la spécification purement synchrone du départ et le résultat lui aussi synchrone incluant les latences requises. Ce modèle asynchrone intermédiaire assimilé de manière abstraite à un graphe de marquage ou d'èvenement (sous-classe des réseaux de Petri), permet de résoudre certains problèmes cruciaux traitant des contraintes initiales permettant d'assurer les propriétés d'absence de famine et de congestion. D'autres conditions, portant sur la topologie des connections, sont également exhibées.

Mots-clés : Latency Insensitive, Relay Station, Shell, GALS, Vérification Formelle, Synchrone, Esterel, SyncChart

1 Introduction

Long wire interconnect latencies may induce time-closure difficulties in modern SoC designs, with propagation of signals across the dye in a single clock cycle becoming problematic. The theory of latency-insensitive design (LID), proposed originally by L. Carloni, K. McMillan and A. Sangiovanni-Vincentelli [14, 15], offers solutions for this issue. The theory can roughly be described as such: an initial fully synchronous reference specification is first desynchronized as an asynchronous network of synchronous block components (a GALS system). Then proper interconnect mechanisms are introduced to *re*synchronize the global system, but allowing specified (integer-time) latencies at interconnects, under the form of fixed-sized lines of so-called *relay stations*. These relay stations, together with *"shell"* wrappers around the synchronous *"pearl"* IP blocks, are in charge of managing the signal value flows. With their help proper regulation is performed between computation blocks that may be temporarily unable to run, either because of input data unavailability, or because of the unability of the rest of the network to store their results if they were produced. The second problem comes from the boundedness of hardware resources, and the fixed-size buffering capacity of the interconnects (the lines of relay stations).

Since their invention relay stations have been a subject of attention for a number of research groups. Extensive modeling, characterization and analysis were provided in [8, 10, 9]. Still, the modeling level never completely reached a fully formal stage, so that proofs of correctness are still informal, either based on textual proof hints, or simulation model executions. We shall somehow use a paper by Casu et Macchiarulo [18], which provides such an (excellent) modeling, as our starting point. We depart from their description on a number of features, though (for instance they do not include the output functions as part of their FSM state machines describing the control structure of each relay station).

Each relay station can be conceived as a cell, to be part of a line of n, then composing the sectioned wire with a latency of n clock cycles. Relay stations implement a given protocol, that will in a sense be preserved by their chaining, only increasing the mandatory latency duration. Each station can receive a valid signal data from its predecessor (either a shell around an IP block or another station), and pass it down *in the next clock cycle* to its successor. The relay station can also receive in the reverse direction a regulation signal, implementing a *"back-pressure"* feature, to indicate that the successor node is unable to accept more data. In this case the station should refrain from sending its value and keep it instead. It should also still be able to receive the next one in this cycle (as the previous node was not warned of the congestion yet), and if necessary should propagate the back-pressure congestion signal to the previous node *in the next clock cycle*. The "next-cycle delays" are needed to respect the physical latency assumption. Of course there are also times where no valid data is transmitted from the previous node because upstream computations were temporarily halted due to lack of inputs. It should thus be noted that any relay station needs a capacity to hold *two* values simultaneously, in case it cannot propagate the current one while a new one simultaneously arrives. It can also be empty, if valid data are produced more slowly than consumed.

Currently the role of relay stations is two-fold: they implement the on-line scheduling scheme requested for proper handling of congestion risks, by back-pressure mechanisms; they also provide

the temporary storage for data for as long as they cannot be forwarded further down the line. The second role is debatable: if the data were allowed to continue their route, they could be stored at the destination shell, if it would provide a dedicated buffer with the same size as the accumulated buffering capacity of all the relay stations on this line. Even better, moving all storage to a single spatial location would then ease the physical synthesis burden. This was noted in [18]. Of course the traffic regulation and the back-pressure mechanisms *should still be applied* in a mandatory fashion, since otherwise the end destination buffer could overflow. But they would only stall back data traffic and computation at shell level, not halfway through the interconnects. Back pressure mechanisms now show the net effect of retropropagating information on the congestion and traffic jam reported *"downwards"*. They to do so only when needed, but *as early as feasible*, while respecting the latencies needed to travel through the long wires.

The paper is organised as follows:
In section 2 we recall briefly the basic contextual definition of synchronous circuits (for local components) and GALS systems (as networks of local synchronous computation components connected by unbounded buffers). We mention some initialization issues, solved as in [12] by the data valueless abstraction of GALS models into Petri Net Marked Graphs. It should be noted here that the body of theoretical results developed around Marked Graphs, also called Event Graphs in the literature, can provide a number of useful analytical results for the characterisation of such systems [11, 3]. This is also true in the case of places with bounded capacity, and it provides answers to issues mentioned in previous papers on LID systems. In particular it provides sufficient conditions for proper initialization of data in lines, so as to guarantee liveness as absence of deadlock but also congestion altogether. On the other hand, Event Graphs (as all Petri Net subclasses) are inherently asynchronous as a concurrency model, and their application to scheduling and "maximal progress" remains for us to be investigated. Here again answers might already exist in the literature.
In section 3 we provide abstract requirements and formal constraints to be satisfied by relay stations models. Then we briefly recall the model of [18], which itself somehow summarizes previous works. We provide our formal model, under the form of a synchronous (cycle-accurate) Mealy machine, with regular features and output signal clear timing specification. Our model is amenable to description in Esterel [5] or SyncCharts [2], thereby allowing formal methods and model-checking techniques. Of course this could also be possible by providing a direct netlist description in `blif` format for instance, but we gain syntactic flexibility, to describe easily the combination of several relay stations into a wire of great latency for instance.
In section 3.2 we specify formally a number of correctness properties, that can be established on a line of relay stations. Of course brute-force model-checking does not allow to reason on parametric models (where here the parameter would be the latency length n of the line), so we need to instantiate several constant length values.
We describe the shell wrappers (here very close to the version of [18]) in section 4. Again we model-checked them to establish correctness properties.
Section 5 is dedicated to (modest) considerations on network topologies that can adversely impact the approach. We provide a simple family of strongly connected graphs for each no static (or off-line) scheduling *equalizing* the latencies is feasible. Equalization is a desirable property; with it

one can get rid of the on-line scheduling required for congestion control and implemented by back-pressure signaling altogether. Indeed, the purpose of the additional latencies is to ensure that all proper input data are provided *simultaneously* to the local computing block. The "non-physical" new latencies can then be shifted up and down the network, under some semantic-preserving constraint, to optimize the global cycle allocation of computation activities. Work in this direction was started in [7, 17], under the naming of *recycling* and inspired by software pipelining cycle allocation techniques. It extends and refers somehow to the paradigms of sequential circuit *retiming* [13].
We conclude with several open questions. The main topic for extension that attracts our attention is the following one: currently the design methodology starts from a monolithic synchronous specification. This is needed to retain several important synthesis techniques from commercial EDA flows. But if one can recognize that this seemingly synchronous description in fact contains informations indicating timing flexibility and potential decomposition into smaller synchronous "pearls", how could we efficiently extend the approach to use this extra knowledge ? Here we are refering to so-to-speak *asynchronous* processes (with the word "asynchronous" here applied to the *computation model*), rather than to buffered connections (where the word "asynchronous" is applied to the *communication model*. Examples of such extra information could be provided by the user (as multirate/multiclock modeling extensions, or exclusive control modes) [1, 6, 19]. It could also be extracted by dynamic semantic analysis, as is done in the *iso/endochrony* theory of Benveniste *et al* [16, 4] (to the best of our understanding).

2 Preliminaries

Synchronous circuit: A synchronous circuit is associated with a clock. It has a signal interface consisting of two sets of (Boolean) input and output signals, and an internal state consisting of a set of (Boolean) registers (or flip-flops). On each clock *tick*, it produces current outputs and next-instant register values from the current values of inputs and registers.

Formally, a synchronous circuit is thus a structure $< \mathcal{I}, \mathcal{O}, \mathcal{R}, Out, Next >$, where

- $\mathcal{I}$ is a set of Boolean input variables $\{I_0, \ldots, I_{n-1}\}$. We call the vector $I =< I_0, \ldots, I_{n-1} > \in B^n$ an input event. It represents the valuation of all input variables at a given instant.
- $\mathcal{O}$ is the set of Boolean output variables $\{O_0, ..., O_{m-1}\}$, We call the vector $O =< O_0, \ldots, O_{m-1} > \in B^m$ an output event. It represents the valuation of all output variables at a given instant.
- $\mathcal{R}$ is the set of Boolean output variables $\{R_0, ..., R_{p-1}\}$, We call $R =< R_0, \ldots, R_{p-1} > \in B^p$ the current state. We also use the next-state $R' =< R'_0, \ldots, R'_{p-1} >$, using primed names.
- Out is a vector $< Out_0, \ldots, Out_{m-1} >$ of Boolean functions, $Out_j : (B^n \times B^p) \rightarrow B$. So each function Out_j defines the value of output variable O_j from the current values of input and register variables.

- $Next$ is a vector $< Next_0, \ldots, Next_{p-1} >$ of Boolean functions, $Next_j : (B^n \times B^p) \rightarrow B$. So each function $Next_j$ defines the next value of register variable R'_j from the current values of input and register variables.

Synchronous or asynchronous networks of synchronous circuits One can build larger circuits by setting local (IP) synchronous components in parallel, establishing desired point-to-point interconnections of inputs to outputs of different blocks. This is displayed in figure 1, if one assumes for the connections simple wires, and that all components run on the same clock. The result is then a compound netlist, homogeneous in nature with the local component synchronous circuits.

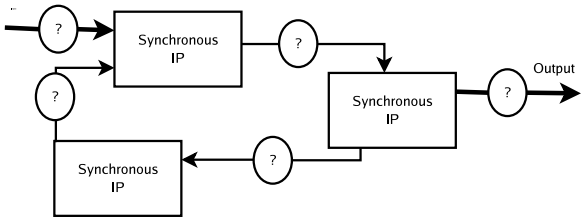


Figure 1: Network of synchronous IP blocks (synchronous or asynchronous)

On the other hand one can also assumes that local synchronous components are **not** globally synchronized, and that connections are established through "ideal" unbounded FIFO queues. This builds another interpretation of figure 1, as a global data-flow network. Now each component can be allowed to run only when all its input data values are present. The effect of its run is to consume one input value on each input channel, and to produce one output value on its output channel. It can be conceived of as a fully unrestricted GALS system. We shall use this stage of representation only as an intermediate step for conceptual modeling.

As noted in [12], the unrestricted GALS model maps directly to Event/Marked Graphs (a well-known subclass of Petri Nets) when disregarding values carried as signal data. This association helps prove that, under some careful initialization conditions, this asynchronous version is functionally equivalent to the previous, fully synchronous one (see below the discussion on initialization).

Marked Graphs Also called Event graphs in the literature, they form a specific subclass of Petri Nets where places have exactly one input transition and one output transition [11]. In our case transitions represent local synchronous components, which indeed consume one data on each input channel, and produce one on each output channel in each step. With data abstracted as "tokens" the place marking represent the number of data currently contained in the interconnect FIFO queue.

Marked/Event Graphs are "free-choice" nets. Various executions only differ in relative schedulings of firings of individual transitions, and these behaviors are *confluent*: the firing of a given transition cannot disallow the one of another if it was previously allowed. Also, the sum of all places markings in a given graph cycle remains invariant all along any execution. A Petri Net (PN) is called

live if any transition can still be executed (possibly after a number of steps) from any reachable marking. It can be proved that a Marked Graph is live if each graph cycle contains at least a token in one of its place. Picture 2 shows a Marked Graph associated with the previous GALS network (in its asynchronous form).

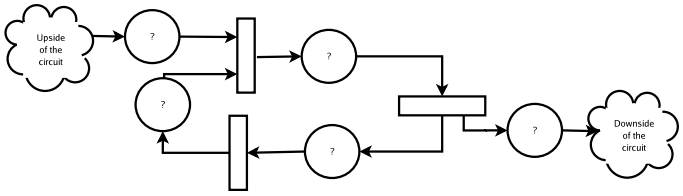


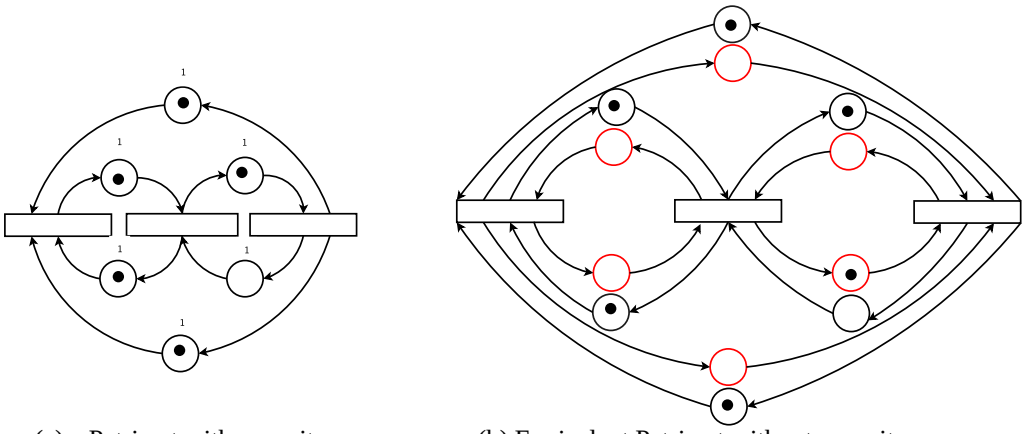
Figure 2: a live Marked Graph associated with the previous GALS picture

Marked Graphs with Place Capacities In GALS theories (such as Latency-Insensitive design and others), the purpose is usually to build a model "in between" the fully synchronous and asynchronous ones. In particular it is important in SoC design to be able to restrict interconnects so as to use only bounded space. The general philosophy is thus: first, desynchronize the fully synchronous specification; second, resynchronize it by careful scheduling mechanisms in a way that respect mandatory physical latencies, while using only bounded communication resources. At the abstract PN level, this boundedness can be modeled with place capacities (the scheduling issue will be dealt with elsewhere later).

Capacities are introduced in Petri Nets by requesting that a given place cannot hold more than n tokens, n being the capacity of that place. Capacities can be traced back to the foundation of PN history, without a clear seminal paper (see [3] for definitions). In fact it is immediate to replace a PN with capacity with another equivalent one *without* capacity by adding a new place for each existing one, with as marking as the difference betwen the original place capacity and its current initial marking. This new place is connected to transition in the reverse way as the original. Figure 3 displays a PN net with capacities (here of 1 for simplicity), and the equivalent PN with duplicated backward places.

Of course the bounded capacity raises new liveness problems, this time because of congestion and overflow instead of starvation and lack of available data tokens. Fortunately we can use the important fact that the above completion preserves the Marked Graph subclass, and inside this context solutions will be found. As will appear later, the latency-insensitive *relax-synchronized* version of our GALS system will possess a capacity of holding $2n$ data token on a connection line comprising of n relay-stations.

The final models produced in LID theory are (on first approximation) latency-bounded, resynchronized versions of marked graphs with capacities. In the sequel we shall call them *relaxed-synchronous* systems, as they combine both synchronous features (all components and interconnects run on the same clock), and user-imposed interconnect minimal latencies (a constant integer delay for the line to transmit its signal/data values. While the data are still in transit, computation parts



(a) a Petri net with capacity (b) Equivalent Petri net without capacity

Figure 3: From a Petri net with capacity (a) to an other Petri net without capacity but with the same behaviour (b)

are paused by their surrounding shells, using clock-gating mechanisms. To respect the fixed-sized bufering ability a back-pressure congestion control protocol is applied across relay-stations.

Initial and well-formedness conditions We consider here various issues of proper initialization and structural well-formedness of the networks, ensuring for each semantics, be it synchronous or asynchronous, both starvation-freedom (or PN liveness), and congestion-freedom (or PN safety). We also briefly consider the quantitative issue of production rate (or throughput).

We recall the well-known fact that any graph can be decomposed as a directed acyclic graph (DAG) of strongly connected components (SCC), with a SCC possibly containing a single node.

Concerning **synchronous networks of synchronous components**, a valid signal/data must be present on each wire at the clock tick. In order to achieve this (while assuming it from the network primary inputs), one usually imposes that there is no combinatorial loop across the network. In other words *each loop in the network graph must cross a register*, which produces its output in the next clock cycle than it received it as input. Here the network graph consists of the local dependencies inside the components plus the interconnections between components. This is a strictly weaker condition as to impose that all component outputs are *latched* (as in Moore fashion), even though the second assumption is often recommended for composite design style, and is actually implicitly adopted in some of the GALS literature. Note here that the programme of splitting up long combinatorial wires into sections is only fulfilled if not all local outputs are latched. Still, if it is the case one remains capable of turning unit delays into arbitrarily chosen delays.

Concerning **asynchronous networks of synchronous components** it is also the case that the network is live (so that all local components get fired infinitely often) iff *there is at least one token in each network cycle loop* (provided the primary inputs each provide an infinite stream of signal/data of course). This is a direct consequence of the result of Marked Graphs liveness. This

matches closely the corresponding assumption on synchronous networks, provided the register is in fact a latch on a local output (but still not all outputs need to be latched, only one in each network communication cycle). The latched output can then be, in a sense, drawn from the local component to become the seed initial value of the interconnecting FIFO queue. Of course initialization with more values in the queues is feasible with liveness preserved (the more token the better in this case). But it is problematic to figure out how to obtain these seed values in general if starting from a fully synchronous specification with which to retain functional equivalence.

Considering **relaxed-synchronous versions**, where bounded capacity channels are replacing the unbounded buffering capacity of FIFO queues, a new kind of liveness problem is raised. Because of potential congestion, local computation blocks can now get blocked because their output hannels are not ready to accept their results, which they could not store without overflow. This issue is theoretically solved by requesting that *the completed PN net do not allow any blank cycle*. Here the PN completion consists in adding the backward places to play the role of capacities. In other words each graph cycle in the completed graph should contain at least a token mark in one of its places. The net of figure 3 is a typical counterexample of this: with places each of capacity one, the net on the left is blocked; this is made explicit as blank cycles in the completed net on the right.

As we shall see later, a channel of n relay stations has a buffering capacity of $2n$ signal/data values. In the (frequent) case where the line is assumed to be initialized with only one value, then the virtual backward places all contain at least a token, thereby definitely disallowing blank cycles.

It has often been remarked in the GALS literature that, ultimately, a (simply connected) relaxed-synchronous network could run *no faster* than the speed of its slowest simple cycle loop. First, any SCC is restricted to the speed of its slowest cycle (after perhaps an initial phase where enough internal tokens can allow some parts to take “almost one lap in advance”). Then, whenever the part located upfront from a SCC starts running ahead, tokens accumulate at the entrance of this SCC until the bounded buffer gets filled, after which point there is no choice but to run the SCC behavior part. Similarly for the downstream parts, which needs data production from the upstream and SCC parts to be fed to run. It was established that the rate of the slowest loop was computed as the ratio of the number of data/token over the overall buffering ability over the loop.

3 Relay Station

We now come to the main part of this article. The purpose is to implement fixed-size communication channels that divide the long wires into sections, such that a signal/data can be propagated from one section to the next only in the next clock cycle. Similarly the signals needed to implement the congestion control back-pressure must also respect these traveling delays. To this end, *relay stations* were introduced in [14]. They are specific hardware elements that provide the proper interface between sections (and also the shells at the channel’s ends). These elements must have some buffering activities, to store data “on route” of course, but also to park these additional data which might discover that because of congestion, the cannel downstream cannot accept them.

3.1 Relay Station Modeling

Despite the number of publications describing relay stations in the literature, they are usually informally characterized. Neither their precise constraints representing the physical time requirements (in clock cycles), nor their formal model and their proper satisfaction is full described. The paper that comes nearest to this is [18]. However they do not use a pure synchronous modeling in their FSM (Finite State Machine). We shall deal now with all these issues.

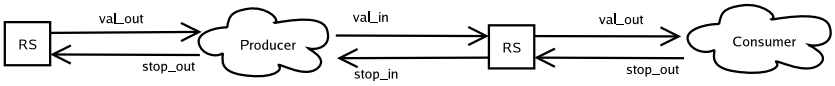


Figure 4: Relay Station - Block Diagram

We borrow from [18] the interface of input/output signals. It is depicted in figure 4. The data reception is represented by an input signal val_in being raised (it corresponds to $\neg\tau$ in the former articles on LID). It is a pure boolean signal (we can abstract the data values). Then the RS passes the data with a corresponding val_out signal. Concerning back-pressure, the RS can receive an halting order with the signal. The relay station receives input data with a valid signal $stop\_out$ being raised. It then transmits it with a $stop\_in$ signal (so $stop\_out$ is an input, and $stop\_in$ is an output).

Pseudo-physical requirement: It is important to note that signal/data cannot be propagated combinatorially from one section to the next:

- $val\_in \hookrightarrow_{next} val\_out$.
- $stop\_out \hookrightarrow_{next} stop\_in$.

On the other hand, there *can* be combinatorial relations between $stop\_in$ and val_in (resp. between $stop\_out$ and val_out, as they belong the same section.

So relay stations need registers (flip-flops for instance) to retain the signal betwen reception and propagation. In fact, as shown in [14], they need two such slots, in case a new data arrives while the current one cannot be propagated. Then, the congestion mechanism is supposed to guarantee that no further data can be received (and thus lost), because they are retained elsewhere upstream. This provides the abtract architecture of figure 5.

3.1.1 Relay Station - FSM

We represent in figure 6 the relay station as a Mealy machine, with explicit states, handling thus both the output and next state functions. Now we show a synchronous Mealy FSM using for state encoding the number of registers free within the relay station.

The FSM contains 3 states, corresponding to the occupation of the registers:

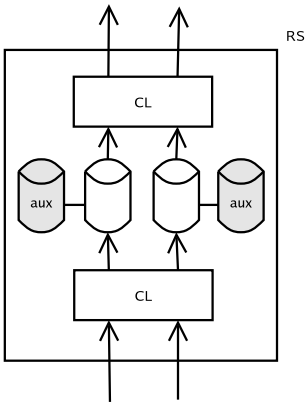


Figure 5: Relay Station structure

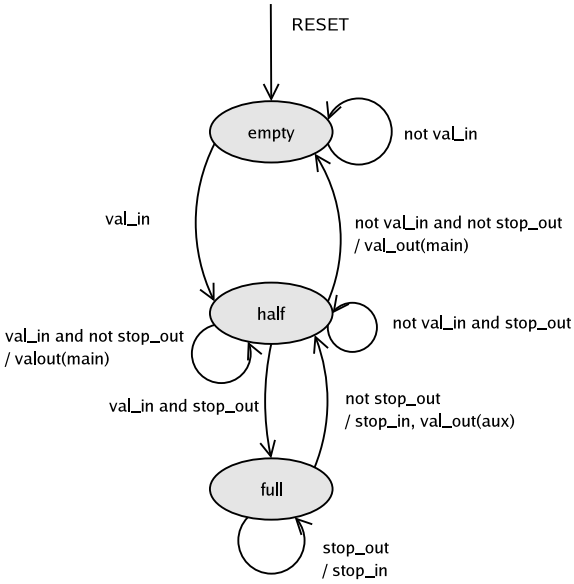


Figure 6: Relay Station FSM

`empty` *when no data are currently buffered in the RS*; in this state the RS simply wait for a valid input data, and store it in its `main` register (goes to state `half`). $stop\_out$ signals are ignored, and not propagated upstream, as this cell can absorb traffic.

`half` *when it holds one data*; Then the RS cell only transmits its current, previously received signal data if ever it does not receive an halting $stop\_out$ signal (remember this combinatorial relation is ok, being inside a section). If halting, it retains its data, but must also accept a potential new one from upstream (as it has not sent any back-pressure holding signal yet). In the second case it becomes full, with the second value occupying its "emergency" `aux`iliary register. If the RS can transmit ($stop\_out$ false), it either goes back to `empty` or retrieve a new valid data in, remaining then in the same state. On the other hand it still makes no provision to propagate back-pressure (in the next clock cycle), as it is still unnecessary due to its own buffering capacity.

`full` *when it contains two data*; then it raises in any case the $stop\_in$ signal, propagating to the upstream section the hold-out $stop\_out$ signal received in the previous clock cycle. If it does not itself receive a new $stop\_out$, then the line downstream was cleared enough so that it can transmit its data; otherwise it keeps it and remains halted.

Discussion With such a precise, cycle-accurate model, one can for instance wonder whether it would be feasible to improve the design to be able, while `full`, to both propagate its current data and accept a new one, remaining full. Of course this should be useless in practice, because the val_in signal could not be received (since the previous cell, when warned of its $stop\_out$, blocks its val_out to become the current RS's val_in). But if the RS is connected to another element, the shell for instance, the constraint $stop\_out \Rightarrow \neg val\_out$ has to be checked and guaranteed, or at least appropriate behavior must be checked. This can easily be done using trivial model-checking on our formal description.

3.2 Correctness properties and formal verification

Keeping with the kind of remarks of the previous discussion, one can phrase a number of correctness properties to hold on a relay station, or a line of relay stations (or later, a network comprising shells and pearls). Remember that correctness criteria for liveness (seen as freeness from both deadlock and congestion) were already established as PN graph markings conditions, linked to data initialisation in section 2. Instances of additional properties are:

- relay stations cannot overflow;
- data cannot be lost nor overwritten;
- data order is preserved;
- at any point in time, the number of valid data produced from a line is bounded relative to the number entered:

  $$\#(val\_in) + Init\_line \leq \#(val\_out) \leq \#(val\_in) + Init\_line + 2 \times length\_line$$

  where $Init\_line$ is the number of data initially residing in the line of RSs, and $length\_line$ is the number of RSs.

- a line of n relay station cannot notify congestion to its source unless it receives enough similar back-presure signals, given its initial content;
- conversely, a line receiving enough back-pressure hold-out signals and data will eventually get filled and notice congestion.

The first property can be checked by adding a new state to the relay station, named `overfull`, which can be attained by a transition for the `full` state triggered by the `val_in and stop_out` signal combination. The check will then consist in proving that such states are unreachable in all RSs. The second and third properties could be modeled in a restricted case by "tagging" the successive data signals with indices, and then checking that these indexes are returned by the line in the same order as they were entered in the other end. The simplest scheme is to alternate 0 and 1 tags, providing an *alternated bit protocol* type verification.

We checked these properties by model-checking, with (low-range) constants replacing the integer parameters, and observers built from these formulas.

4 Shell wrappers

4.1 Shell modeling

Here our model follows rather closely the one of Casu and Macchiarulo [18]. It is depicted in figure 7.

As mentioned in section 2, one can consider the case where shells and pearls have potential zero-delay propagation (as long as there is no combinatorial loop involving only shells, without crossing a relay station). The shells will need the ability to store data that have already arrived, awaiting others still missing.

The Shell works as follows:

- The internal pearl's $clock$ and all val_out_i valid output signals are generated once we have all val_in, while $stop$ is false. The internal $stop$ signal itself represents the disjunction of all incoming $stop\_out_j$ signals from outcoming channels;
- the buffering register of a given input channel is used meanwhile as long as not all other input data are available;
- so, internal pearl's $clock$ is set to false whenever a backward $stop\_out_j$ occurs as true, or a forward val_in_i is false. In such case the registers already busy hold their *true* value, while others may receive a valid data "just now";
- $stop\_in_i$ signals are raised towards all channels whose corresponding register was already loaded (a data was received before, and still not consumed), to warn them not to propagate any value in this clock cycle. Of course such signal cannot be sent in case the data is currently received, as it would raise a causality paradox (and a combinatorial cycle).

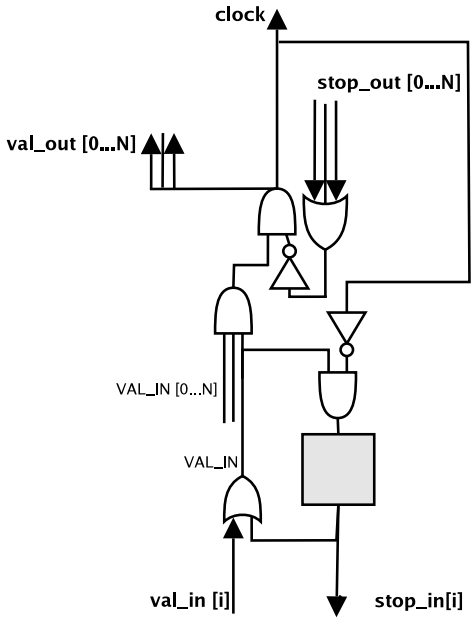


Figure 7: Shell - Circuitry

- flip-flop registers are reset when the pearl's clock is raised, as it consumes the input data. Following the previous remark, the signal $stop\_in_i$ holding back the traffic in channel i is raised for these channels where the data have arrived before the current instant, even in this case.

We should remember the constraint demanded by the relay stations for proper functioning, namely that on each output channel from the producer (is this case the shell), one has $stop\_out_j \Rightarrow \neg val\_out_j$, which holds here.

4.2 Correctness properties and formal verification

Keeping in mind relay stations, we want to show some properties such as:

- data cannot be accepted before the previous one is processed.
- data order is preserved.
- a shell cannot dead-lock.

The first property can be checked simply because the shell is connected *synchronously* to a relay station (or another shell) and thus the relay station cannot send any data to the shell when the shell

is holding a data. The shell can have only one datum from each channel as said before then it cannot overwrite or loose this data until all neeeded datum are present to react. The data order is preserved, because by hypothesis the interconnection network is only point to point, cannot loose data or alter data ordering, the shell is waiting for all datum and then react, thus partial order of the desynchronized design is compatible with the synchronous one. We can also apply the alternated bit protocol verification in this case. The Shell is dead-lock free because we already established it as PN markings conditions.

5 Graph shapes and static scheduling

So far the scheduling mechanism ensuring regulation inside the relaxed-synchronous system is dynamic (or *on-line*). One can also consider the case where a precise computation on latencies, possibly adding extra delays, would force all valid data to arrive *simultaneously* at the same computing location. Then the whole back-presure mechanism would be made useless because of a static (or *off-line*) scheduling. Of course the extra latency delays would be optional, unlike the former one that were installed to respect some physical constraints. They could be moved and displaced to some extend across the system, while preserving the "equal length" constraint on data trips. In a more general scheme one can even consider that the delay figures obtained would allow resynthesis of the local synchronous blocks, absorbing some of the delays to run more slowly but with less resource consumption, or even a floorplan redesign to redistribute critical long wires.

Figuring criteria on network graph shapes that allow efficient and useful latency equalization is an important problem. It gets rifd of the need for back-pressure signals, but can also reduce greatly the rate of τ symbol (that is, $\neg val\_in$) emitted. Equalization in a Marked Graph strongly connected component would amount in adjusting all cycle lengths to the longest one. Nevertheless this equalization is not always feasible, as shown on the counter-example of figure 8.

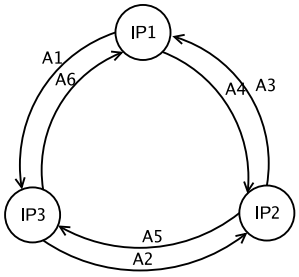


Figure 8: Not Equalizable, No Static Scheduling

There are three cycles of length 2 :
{A1, A6}
{A2, A5}
{A3, A4}

and two cycles of length 3 :
{A1, A2, A3}
{A4, A5, A6}
It is impossible to put integers on edges to get at the same time equal lengths on both exetrnal cycles and on the "local ones". The same would be true of any doubly-linked ring structures with more than three nodes.

6 Further topics

So far the theory developed here only consider the case where local synchronous components all consume and produce data on all input and output channels in each computation step, and where they all run on the same clock. In this favorable case functional determinacy and confluence are guaranteed, with latencies only impacting the relative ordering of behaviors. So it can be proven that the relaxed-synchronous version produces the same output streams from the same input streams as the fully synchronous specification (indeed the rank of a data in a stream corresponds to its time in the synchronous model, thereby reconstructing the structure of successive instants. Several papers considered extensions in the context of GALS systems, but then ignored the issue of functional correspondance with an initial well-clocked specification, which is our important correctness criterion.

This strong assumption can be weakened in a number of ways. Some are related to the various relative speeds and cadences of components in clock cycle rates, some are borne in the extension of Marked/Event graphs to more general subclasses in Petri nets in the asynchronous setting, and the most important ones are linking the two.

One can extend the framework by allowing different cadences (so that various processing blocks run at different speeds, expressed as integer multiples of the master clock). More generally, each component can be assigned its own clock, with the assumption that all clocks are subclocks of a master clock, but not necessarily periodic. One can then build multi-rate/multi-clock systems. But, unless global rates are perfectly equalized around each loops, this might require fact component with different clocks be fed streams of data of unequal lengths. Usually the link with a fully synchronous specification is attempted by introducing a specific *absent* value for every interconnection signal, so that subclocks are defined as ticking only during the instants where a given triggering signal is *not* absent.

In general PN theory a place can be supplied tokens (here abstracting the data put in a FIFO channel) from various transitions (here processing elements). It thus merges the two flows (as a *mux*). The place can also offer its tokens at the other end to various consumers, thus operating a fork (or a *demux*) of the data flow carried through the channel. In other words tokens are shared. It gets difficult then to imagine that the rank of a data in a channel stream will recall the instant it was exchanged in a fully synchronous specification. Still, one can design a "locally-synchronous" version of places (we consider here the case of two producers and two consumers to this place): it has a main running clock, and two subclocks (one for input and one for output), so that data are taken from one input channel when the input subclock is raised, from the other otherwise (and similarly for output).

Of course the two kinds of extensions are linked, since channel sharing imposes that multiple productions or consumptions do not clash, so that it can be established that they are mutually exclusive (by being driven on exclusive subclocks). The issue of success is to guarantee liveness and throughput in the global system. This should be attained by devising the proper scheduling, which should generate the clock pulses at proper rates (in latency and cadence), so that data flow in the system smoothly. Several steps exist in this direction, with the notion of multiclock systems and clock calculus in synchronous languages [1, 6].The correctness criterion is that no component should ever require the presence of a signal data that is absent, and that signal data are not inappropriately lost (sometimes it is ok to ignore and discard them). Studies were also conducted to as when the seemingly monolithic synchronous specification in fact exhibited asynchronous behaviors based on independent clocks underneath [19]

Finally, the goal would be to define a general GALS modeling framework, where GALS components cold be put in GALS networks (to this day the framework is not compositional in the sense that local components need to be synchronous). A system would consits again of computation and interconnect communication blocks, this time each with appropriate triggering clocks, and of a scheduler providing the subclocks computation mechanism, based on their outer main clock and several signals carrying information on control flow. .

References

[1] P. Amagbedon, L. Besnard, and P. Le Guernic. Implementation of the data-flow synchronous language signal. In *Proceedings PLDI'95*, 1995.

[2] C. André. Representation and analysis of reactive behaviors: A synchron ous approach. In *Computational Engineering in Systems Applications*, pages 19–29, 1996.

[3] F. Baccelli, G. Cohen, G.J. Olsder, and J.-P. Quadrat. *Synchronization and Linearity*. Wiley, 1992.

[4] A. Benveniste, B. Caillaud, and P. Le Guernic. From synchrony to asynchrony. In *Proceedings CONCUR'99*, volume 1664 of *LNCS*, 1999.

[5] G. Berry and G. Gonthier. The Esterel synchronous programming language: Design, s emantics, implementation. *Science of Computer Programming*, 19(2):87–152, 1992.

[6] G. Berry and E. Sentovich. Multiclock Esterel. In *Proceedings CHARME'01*, volume 2144 of *LNCS*, 2001.

[7] Luca P. Carloni and Alberto Sangiovanni-Vincentelli. Combining retiming and recycling to optimize the performance of synchronous circuit. In *The Proceedings of the 16th Symposium on Integrated Circuits and System Design*, 2003.

[8] Luca P. Carloni and Alberto L. Sangiovanni-Vincentelli. Performance analysis and optimization of latency insensitive systems. In *Design Automation Conference*, pages 361–367, 2000.

[9] Ajanta Chakraborty and Mark R. Greenstreet. A minimalist source-synchronous interface. In *Proceedings of the 15th IEEE ASIC/SOC Conference*, pages 443–447, September 2002.

[10] Tiberiu Chelcea and Steven M. Nowick. Robust interfaces for mixed-timing systems with application to latency-insensitive protocols. In *Design Automation Conference*, pages 21–26, 2001.

[11] F. Commoner, Anatol W.Holt, Shimon Even, and Amir Pnueli. Marked directed graph. *Journal of Computer and System Sciences*, 5:511–523, october 1971.

[12] J. Cortadella, A. Kontratyev, L. Lavagno, and C. P. Sotiriou. A concurrent model for desynchronization. In *12th International Workshop on Logic and Synthesis*, 2003.

[13] C.E. Leiserson and J.B. Saxe. Retiming synchronous circuits. *Algorithmica*, 6, 1991.

[14] Luca P.Carloni, Kenneth L.McMillan, and Alberto L.Sangiovanni-Vincentelli. Theory of latency-insensitive design. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2001.

[15] Luca P.Carloni, Kenneth L.McMillan, Alexander Saldanha, and Alberto L.Sangiovanni-Vincentelli. A methodology for correct-by-construction latency insensitive design. In *THE BEST OF ICAD*, 200x.

[16] Dumitru Potop-Butucaru, Benoît Caillaud, and Albert Benveniste. Concurrency in synchronous systems. In *Proceedings ACSD'04*, 2004.

[17] François R.Boyer, El Mostapha Aboulhamid, Yvon Savaria, and Michel Boyer. Optimal design of synchronous circuits using software pipelining. In *Proceedings of the ICCD'98*, 1998.

[18] Mario R.Casu and Luca Macchiarulo. A detailed implementation of latency insensitive protocols. In *FMGALS 2003 Proceedings*, 2003.

[19] Montek Singh and Michael Theobald. Generalized latency-insensitive systems for single-clock and multi-clock architectures. In *DATE'04*, 2004.

Contents

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifi que
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399