Improving Communication Patterns in Polyhedral Process Networks
Christophe Alias

To cite this version:

HAL Id: hal-01665155
https://inria.hal.science/hal-01665155
Submitted on 18 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Improving Communication Patterns in Polyhedral Process Networks

Christophe Alias
Abstract: Embedded systems performances are bounded by power consumption. The trend is to offload greedy computations on hardware accelerators as GPU, Xeon Phi or FPGA. FPGA chips combine both flexibility of programmable chips and energy-efficiency of specialized hardware and appear as a natural solution. Hardware design is long, fastidious and bug prone. Hardware compilers from high-level languages (High-level synthesis, HLS) are required to exploit all the capabilities of FPGA while satisfying tight time-to-market constraints. Compiler optimizations for parallelism and data locality restructure deeply the execution order of the processes, hence the read/write patterns in communication channels. This breaks most FIFO channels, which have to be implemented with addressable buffers. Expensive hardware is required to enforce synchronizations, which often results in dramatic performance loss. In this paper, we present an algorithm to partition the communications so that most FIFO channels can be recovered after a loop tiling, a key optimization for parallelism and data locality. Experimental results show a drastic improvement of FIFO detection for regular kernels at the cost of (few) additional storage. As a bonus, the storage can even be reduced in some cases.

Key-words: High-level synthesis, polyhedral compilation, FIFO, FPGA
Amélioration des schémas de communication dans les réseaux de processus polyédriques

Résumé : Les performances des systèmes embarqués sont limitées par la consommation électrique. La tendance est de déléguer les calculs gourmands en ressources à des accélérateurs matériels comme les GPU, les Xeon Phi ou les FPGA. Les circuits FPGA allient la flexibilité d’un circuit programmable et l’efficacité énergétique d’un circuit spécialisé et apparaissent comme une solution naturelle. Des compilateurs de matériels à partir d’un langage haut-niveau sont requis pour exploiter au mieux les FPGA tout en remplissant les contraintes de mise sur le marché. Les optimisations de compilateur restructurent profondément les calculs et les schémas de communication (ordre de lecture/écriture). En conséquence, la plupart des canaux de communication ne sont plus des FIFOs et doivent être implémentées avec un tableau adressable, ce qui nécessite du matériel supplémentaire pour la synchronisation. Dans ce rapport, nous présentons un algorithme capable de partitionner les communications de sorte que la plupart des FIFO puissent être retrouvées après un tuilage de boucles. Les résultats expérimentaux confirment la puissance de notre algorithme et son faible surcoût en stockage.

Mots-clés : Synthèse de circuit haut-niveau, compilation polyédrique, FIFO, FPGA
1 Introduction

Since the end of Dennard scaling, the performance of embedded systems is bounded by power consumption. The trend is to trade genericity (processors) for energy efficiency (hardware accelerators) by offloading critical tasks to specialized hardware. FPGA chips combine both flexibility of programmable chips and energy-efficiency of specialized hardware and appear as a natural solution. High-level synthesis (HLS) techniques are required to exploit all the capabilities of FPGA, while satisfying tight time-to-market constraints. Parallelization techniques from high-performance compilers are progressively migrating to HLS, particularly the models and algorithms from the polyhedral model – a powerful framework to design compiler optimizations. Additional constraints must be fulfilled before plugging a compiler optimization into an HLS tool. Unlike software, the hardware size is bounded by the available silicon surface. The bigger is a parallel unit, the less it can be duplicated, thereby limiting the overall performance. In particular, tough program optimizations are likely to spoil the performances if the circuit is not post-optimized carefully [5]. An important consequence is that the roofline model is not longer valid in HLS [8]. Indeed, peak performance is no longer a constant: it decreases with the operational intensity. The bigger is the operational intensity, the bigger is buffer size and the less is the space remaining for the computation itself. Consequently, it is important to produce at source-level a precise model of the circuit which allows to predict accurately the resource consumption. Process networks are a natural and convenient intermediate representation for HLS [4, 13, 18]. A sequential program is translated to a process network by partitioning computations into processes and flow dependences into channels. Then, the process and buffers are factorized and mapped to hardware.

In this paper, we focus on the translation of buffers to hardware. We propose an algorithm to restructure the buffers so they can be mapped to inexpensive FIFOs. Most often, a direct translation of a regular kernel – without optimization – produces a process network with FIFO buffers [15]. Unfortunately, data transfers optimization [3] and generally loop tiling reorganizes deeply the computations, hence the read/write order in channels (communication patterns). Consequently, most channels may no longer be implemented by a FIFO. Additional circuitry is required to enforce synchronizations [4, 19, 14, 16] which result in larger circuits and causes performance penalties. In this paper, we make the following contributions:

- We propose an algorithm to reorganize the communications between processes so that more channels can be implemented as FIFO after a loop tiling. As far as we know, this is the first algorithm to recover FIFO communication patterns after a compiler optimization.

- Experimental results show that we can recover most of the FIFO disabled by communication optimization, and more generally any loop tiling, at almost no extra storage cost.

The remainder of this paper is structured as follows. Section 2 introduces polyhedral process network and discusses how communication patterns are impacted by loop tiling. Section 3 describes our algorithm to reorganize channels. Section 4 presents experimental results. Finally, Section 5 concludes this paper and draws future research directions.

2 Preliminaries

This section defines the notions used in the remainder of this paper. Section 2.1 and 2.2 introduces the basics of compiler optimization in the polyhedral model and defines loop tiling. Section 2.3...
defines polyhedral process networks (PPN), shows how loop tiling disables FIFO communication patterns and outlines a solution.

2.1 Polyhedral Model at a Glance

Translating a program to a process network requires to split the computation into processes and flow dependences into channels. The polyhedral model focuses on kernels whose computation and flow dependences can be predicted, represented and explored at compile-time. Specifically, the control must be predictable: (for loops, if with conditions on loop counters), arrays only; and loop bounds, conditions and array accesses must be affine functions of surrounding loop counters and structure parameters (typically the array size). This way, the computation may be represented with Preburger sets (typically approximated with convex polyhedra, hence the name). This makes possible to reason geometrically about the computation and to produce precise compiler analysis thanks to integer linear programming: flow dependence analysis \[9\], scheduling \[7\] or code generation \[6, 12\] to quote a few. Most compute-intensive kernels from linear algebra and image processing fit in this category. In some case, kernels with dynamic scheduling \[7\] or code generation \[6, 12\] to quote a few. Most compute-intensive kernels from linear algebra and image processing fit in this category.

2.2 Scheduling and Loop Tiling

Compiler optimizations change the execution order to fulfill multiple goals such as increasing the parallelism degree or minimizing the communications. The new execution order is specified by a schedule. A schedule \(\theta\) maps each execution \(S, \vec{i}\) to a timestamp \(\theta_S(\vec{i}) = (t_1, \ldots, t_d) \in \mathbb{Z}^d\), the timestamps being ordered by the lexicographic order \(\ll\). In a way, a schedule dispatches each execution instance \(S, \vec{i}\) into a new loop nest, \(\theta_S(\vec{i}) = (t_1, \ldots, t_d)\) being the new iteration vector of \(S, \vec{i}\). A schedule \(\theta\) induces a new execution order \(\ll_\theta\) such that \(\langle S, \vec{i}\rangle \ll_\theta \langle T, \vec{j}\rangle\) iff \(\theta_S(\vec{i}) \ll \theta_T(\vec{j})\). Also, \(\langle S, \vec{i}\rangle \ll_\theta \langle T, \vec{j}\rangle\) means that either \(\langle S, \vec{i}\rangle \ll \theta_T(\vec{j})\) or \(\theta_S(\vec{i}) = \theta_T(\vec{j})\). When

\[\ll_\theta\]
## 2.3 Polyhedral Process Networks

Given the iteration domains and the flow dependence relation, \( \rightarrow \), we derive a polyhedral process network by partitioning iterations domains into process and flow dependence into channels. More formally, a polyhedral process network is a couple \( \langle P, C \rangle \) such that:

- Each process \( P \in P \) is specified by an iteration domain \( D_P \) and a sequential schedule \( \theta_P \) inducing an execution order \( \prec_P \) over \( D_P \). Each iteration \( \bar{i} \in D_P \) realizes the execution...
instance \(\mu_P(\tilde{i})\) in the program. The processes partition the execution instances in the program: \(\{\mu_P(D_P)\text{ for each process } P\}\) is a partition of the program computation.

- Each channel \(c \in C\) is specified by a producer process \(P_c \in P\), a consumer process \(C_c \in P\) and a dataflow relation \(\to_c\) relating each production of a value by \(P_c\) to its consumption by \(C_c\): if \(i \to_c j\), then execution \(i\) of \(P_c\) produces a value read by execution \(j\) of \(C_c\). \(\to_c\) is a subset of the flow dependences from \(P_c\) to \(C_c\) and the collection of \(\to_c\) for each channel \(c\) between two given processes \(P\) and \(C\), \(\{\to_c, (P,C) = (P_c,C_c)\}\), is a partition of flow dependences from \(P\) to \(C\).

The goal of this paper is to find out a partition of flow dependences for each producer/consumer couple \((P,C)\), such that most channels from \(P\) to \(C\) can be realized by a FIFO.

Figure 1(c) depicts the PPN obtained with the canonical partition of computation: each execution \((S,\tilde{i})\) is mapped to process \(P_S\) and executed at process iteration \(\tilde{i}: \mu_P(\tilde{i}) = (S,\tilde{i})\). For presentation reason the compute process is depicted as \(C\). Dependence depicted as \(i\) on the dependence graph in (b) are solved by channel \(i\). To read the input values in parallel, we use a different channel per couple producer/read reference, hence this partitioning. We assume that, locally, each process executes instructions in the same order than in the original program: \(\theta_{\text{load}}(i) = i\), \(\theta_{\text{compute}}(t,i) = (t,i)\) and \(\theta_{\text{nur}}(i) = i\). Remark that the leading constant (0 for load, 1 for compute, 2 for store) has disappeared: the timestamps only define an order local to their process: \(\prec_{\text{load}}, \prec_{\text{compute}}\) and \(\prec_{\text{store}}\). The global execution order is driven by the dataflow semantics: an operation is executed whenever its operands are available. The next step is to detect communication patterns to figure out how to implement channels.

**Communication Patterns** A channel \(c \in C\) might be implemented by a FIFO iff the consumer \(C_c\) read the values from \(c\) in the same order than the producer \(P_c\) write them to \(c\) (in-order) and each value is read exactly once (unicity) \([13, 15]\). The in-order constraint can be written:

\[
\forall x \to_c x', \forall y \to_c y': x \prec_C y' \Rightarrow x \preceq_P y
\]

The unicity constraints can be written:

\[
\forall x \to_c x', \forall y \to_c y': x' \neq y' \Rightarrow x \neq y
\]

Notice that unicity depends only on the dataflow relation \(\to_c\), it is independent from the execution order of the producer process \(\prec_P\) and the consumer process \(\prec_C\). Furthermore, \(\neg\text{in-order}(\to_c, \prec_P, \prec_C)\) and \(\neg\text{unicity}(\to_c)\) amount to check the emptiness of a convex polyhedron, which can be done by most LP solvers.

Finally, a channel may be implemented by a FIFO iff it verifies both in-order and unicity constraints:

\[
\text{fifo}(\to_c, \prec_P, \prec_C) := \text{in-order}(\to_c, \prec_P, \prec_C) \land \text{unicity}(\to_c)
\]

When the consumer reads the data in the same order than they are produced but a datum may be read several times: \(\text{in-order}(\to_c, \prec_P, \prec_C) \land \neg\text{unicity}(\to_c)\), the communication pattern is said to be in-order with multiplicity: the channel may be implemented with a FIFO and a register keeping the last read value for multiple reads. However, additional circuitry is required to trigger the write of a new datum in the register \([13]\): this implementation is more expensive than a single FIFO. Finally, when we have neither in-order nor unicity: \(\neg\text{in-order}(\to_c, \prec_P, \prec_C) \land \neg\text{unicity}(\to_c)\), the communication pattern is said to be out-of-order without multiplicity: significant hardware
resources are required to enforce flow- and anti- dependences between producer and consumer and additional latencies may limit the overall throughput of the circuit [4, 19, 14, 16].

Consider Figure 1.(c), channel 5, implementing dependence 5 (depicted in (b)) from \( \langle s, t-1, i \rangle \) (write \( a[t, i] \)) to \( \langle s, t, i \rangle \) (read \( a[t-1, i] \)). With the schedule defined above, the data are produced \( \langle s, t-1, i \rangle \) and read \( \langle s, t-1, i \rangle \) in the same order, and only once: the channel may be implemented as a FIFO. Now, assume that process `compute` follows the tiled execution order depicted in Figure 2(a). The execution order now executes tile with point \( (4,4) \), then tile with point \( (4,8) \), then tile with point \( (4,12) \), and so on. In each tile, the iterations are executed for each \( t \), then for each \( i \). Consider iterations depicted in red as 1, 2, 3, 4 in Figure 2(b). With the new execution order, we execute successively 1, 2, 4, 3, whereas an in-order pattern would have required 1, 2, 3, 4. Consequently, channel 5 is no longer a FIFO. The same hold for channel 4 and 6. Now, the point is to partition dependence 5 and others so FIFO communication pattern hold.

![Figure 2: Impact of loop tiling on communication pattern](image)

Consider Figure 2(c). Dependence 5 is partitioned in 3 parts: red dependences crossing tiling hyperplane \( \phi_1 \) (direction \( t \)), blue dependences crossing tiling hyperplane \( t+i \) (direction \( t+i \)) and green dependences inside a tile. Since the execution order in a tile is the same than the original execution order (actually a subset of the original execution order), green dependences will clearly verify the FIFO communication pattern. As concerns blue and red dependences, source and target are executed in the same order because the execution order is the same for each tile and dependence 5 happens to be short enough. In practice, this partitioning is effective to reveal FIFO channels. In the next section, we propose an algorithm to find such a partitioning.

### 3 Our Algorithm

Figure 3 depicts our algorithm for partitioning channels given a polyhedral process network \((P, C)\) (line 5). For each channel \( c \) from a producer \( P = P_c \) to a consumer \( C = C_c \), the channel is partitioned by depth along the lines described in the previous section (line 7). \( D_P \) and \( D_C \) are assumed to be tiled with the same number of hyperplanes. \( P \) and \( C \) are assumed to share a schedule with the shape: \( \theta(\phi_1, ..., \phi_n, \vec{i}) = (\phi_1, ..., \phi_n, \vec{i}) \). This case arise frequently with tiling schemes for I/O optimization [4]. If not, the next channel \( \rightarrow_c \) is considered (line 6). The split is realized by procedure `split` (lines 1–4). A new partition is build starting from the empty set. For each depth (hyperplane) of the tiling, the dependences crossing that hyperplane are filtered...
and added to the partition (line 3): this gives dependences $\rightarrow_{c_1,\ldots,c_n}$. Finally, dependences lying in a tile (source and target in the same tile) are added to the partition (line 4): this gives $\rightarrow_{c_1,\ldots,c_n}$. $\theta_{P}(x) \approx^n \theta_{C}(y)$ means that the $n$ first dimensions of $\theta_{P}(x)$ and $\theta_{C}(y)$ (tiling coordinates $(\phi_1,\ldots,\phi_n)$) are the same: $x$ and $y$ belong to the same tile. Consider the PPN depicted in Figure 1(c), with the tiling and schedule discussed above: process $compute$ is tiled as depicted in Figure 2(c) with the schedule $\theta_{compute} = (\phi_1, \phi_2, t, i)$. Since processes $load$ and $store$ are not tiled, the only channels processed by our algorithm are $4, 5$ and $6$. $\text{SPLIT}$ is applied on the associated dataflow relations $\rightarrow_{4_i}, \rightarrow_{5_i}$ and $\rightarrow_{6_i}$. Each dataflow relation is split in three parts as depicted in Figure 2(c). For $\rightarrow_{5_i}$: $\rightarrow_{5_i}$ crosses hyperplane $t$ (red), $\rightarrow_{5_i}$ crosses hyperplane $t+i$ (blue) and $\rightarrow_{5_i}$ stays in a tile (green).

This algorithm works pretty well for short uniform dependences $\rightarrow_{c_1}$: if $\text{fifo}(c)$ before tiling, then, after tiling, the algorithm can split $c$ in such a way that we get FIFOs. However, when dependences are longer, e.g. $(t, i) \rightarrow (t, i + 2)$, the target operations $(t, i + 2)$ reproduce the tile execution pattern, which prevents to find a FIFO. The same happens when the tile hyperplanes are “too skewed”, e.g. $\tau_1 = (1, 1)$, $\tau_2 = (2, 1)$, dependence $(t - 1, i - 1) \rightarrow (t, i)$.

Figure 3: Our algorithm for partitioning channels

4 Experimental Evaluation

This section presents the experimental results obtained on the benchmarks of the polyhedral community. We demonstrate the capabilities of our algorithm at recovering FIFO communication patterns after loop tiling and we show how much additional storage is required.

Experimental Setup We have run our algorithm on the kernels of PolyBench/C v3.2. Tables 1 and 2 depicts the results obtained for each kernel. Each kernel is tiled to reduce I/O while exposing parallelism and translated to a PPN using our research compiler, Dcc (DPN C Compiler). Dcc actually produces a DPN (Data-aware Process Network), a PPN optimized for a specific tiled pattern. DPN features additional control processes and synchronization for...
I/O and parallelism which have nothing with our optimization. So, we actually only consider the PPN part of our DPN. We have applied our algorithm to each channel to expose FIFO patterns. For each kernel, we compare the PPN obtained after tiling to the PPN processed by our algorithm.

**Results**  Table 1 depicts the capabilities of our algorithm to find out FIFO patterns. For each kernel, we provide the channels characteristics on the original tiled PPN (Before Partitioning) and after applying our algorithm (After Partitioning). We give the total number of channels (#channel), the FIFO found among these channels (#fifo), the number of channels which were successfully turned to FIFO thanks to our algorithm (#fifo-split), the ratios #fifo/#channel (%fifo) and #fifo-split/#channel (%fifo-split), the cumulated size of the FIFO found (fifo-size) and the cumulated size of the channels found, including FIFO (total-size). On every kernel, our algorithm succeeds to expose more FIFO patterns (%fifo vs %fifo-split). On a significant number of kernels (11 among 15), we even succeed to turn all the compute channels to FIFO. On the remaining kernels, we succeed to recover all the FIFO communication patterns disabled by the tiling. Even though our method is not complete, as discussed in section 3, it happens that all the kernels fulfill the conditions expected by our algorithm (short dependence, tiling hyperplanes not too skewed).

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Before Partitioning</th>
<th>After Partitioning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#channel</td>
<td>#fifo</td>
</tr>
<tr>
<td>trmm</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>gemm</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>syrk</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>symm</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>gemver</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>gesummv</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>syr2k</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>lu</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>cholesky</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>stax</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>doitgen</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>jacobi-2d</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>sedel-2d</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>jacobi-1d</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>heat-3d</td>
<td>20</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Detailed results

Table 2 depicts the additional storage required after splitting channels. For each kernel, we compare the cumulative size of channels split and successfully turn to a FIFO (size-fifo-fail) to the cumulative size of the FIFOs generated by the splitting (size-fifo-split). The size unit is a datum e.g. 4 bytes if a datum is a 32 bits float. We also quantify the additional storage required by split channels compared to the original channel (△ := size-fifo-split - size-fifo-fail / size-fifo-fail). It turns out that the FIFO generated by splitting use mostly the same data volume than the original channels. Additional resources are due to our sizing heuristic [1], which rounds channel size to a power of 2. Surprisingly, splitting can sometimes help the sizing heuristic to find out a smaller size (kernel gemm), and then reducing the storage requirements. Indeed, splitting decompose channel into channels of a smaller dimension, for which our sizing heuristic is more precise. In a way, our algorithm allows to finds out a nice piecewise allocation function whose

RR n° 9131
footprint is smaller than a single piece allocation. We plan to exploit this nice side effect in the future.

<table>
<thead>
<tr>
<th>kernel</th>
<th>Size-lifo-fail</th>
<th>Size-lifo-split</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>trmm</td>
<td>256</td>
<td>257</td>
<td>0%</td>
</tr>
<tr>
<td>gemm</td>
<td>512</td>
<td>288</td>
<td>-44%</td>
</tr>
<tr>
<td>syrk</td>
<td>8192</td>
<td>8193</td>
<td>0%</td>
</tr>
<tr>
<td>symm</td>
<td>800</td>
<td>801</td>
<td>0%</td>
</tr>
<tr>
<td>gemver</td>
<td>32</td>
<td>33</td>
<td>3%</td>
</tr>
<tr>
<td>gesummv</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>syr2k</td>
<td>8192</td>
<td>8193</td>
<td>0%</td>
</tr>
<tr>
<td>lu</td>
<td>528</td>
<td>531</td>
<td>1%</td>
</tr>
<tr>
<td>cholesky</td>
<td>273</td>
<td>275</td>
<td>1%</td>
</tr>
<tr>
<td>atax</td>
<td>1</td>
<td>1</td>
<td>0%</td>
</tr>
<tr>
<td>doitgen</td>
<td>4096</td>
<td>4097</td>
<td>0%</td>
</tr>
<tr>
<td>jacobi-2d</td>
<td>8320</td>
<td>8832</td>
<td>6%</td>
</tr>
<tr>
<td>seidel-2d</td>
<td>49952</td>
<td>52065</td>
<td>4%</td>
</tr>
<tr>
<td>jacobi-1d</td>
<td>1152</td>
<td>1174</td>
<td>2%</td>
</tr>
<tr>
<td>heat-3d</td>
<td>148608</td>
<td>158992</td>
<td>7%</td>
</tr>
</tbody>
</table>

Table 2: Impact of splitting on storage requirements

5 Conclusion

In this paper, we have proposed an algorithm to reorganize the channels of a polyhedral process network to reveal more FIFO communication patterns. Specifically, our algorithm operates producer/consumer processes whose iteration domain has been partitioned by a loop tiling. Experimental results show that our algorithm allows to recover most of the FIFO disabled by loop tiling with almost the same storage requirement. Our algorithm is sensible to the dependence size and the loop tiling chosen. In the future, we plan to design a reorganization algorithm provably complete, in the meaning that a FIFO channel will be recovered whatever the dependence size and the tiling used. We also observe that splitting channels can reduce the storage requirements in some cases. We plan to investigate how such cases can be revealed automatically.

References


Contents

1 Introduction 3

2 Preliminaries 3
   2.1 Polyhedral Model at a Glance 4
   2.2 Scheduling and Loop Tiling 4
   2.3 Polyhedral Process Networks 5

3 Our Algorithm 7

4 Experimental Evaluation 8

5 Conclusion 10