Flexible Filters: Load Balancing through Backpressure for Stream Programs

Rebecca L. Collins and Luca P. Carloni
Technical Report CUCS-030-09
Department of Computer Science
Columbia University
[rlc2119,luca]@cs.columbia.edu

ABSTRACT

Stream processing is a promising paradigm for programming multi-core systems for high-performance embedded applications. We propose flexible filters *as a technique that combines static mapping of the stream program tasks with dynamic load balancing of their execution. The goal is to improve the system-level processing throughput of the program when it is executed on a distributed-memory multi-core system as well as the local (core-level) memory utilization. Our technique is distributed and scalable because it is based on point-to-point handshake signals exchanged between neighboring cores. Load balancing with flexible filters can be applied to stream applications that present large dynamic variations in the computational load of their tasks and the dimension of the stream data tokens. In order to demonstrate the practicality of our technique, we present the performance improvements for the case study of a JPEG encoder running on the IBM Cell multi-core processor.*

1. INTRODUCTION

Stream processing [4, 14, 18, 22] is a promising model for programming multi-core system-on-chip platforms that is applicable to a wide range of applications including high-performance embedded applications, signal processing, image compression and continuous database queries [5, 21]. The basic idea of stream processing is to decompose an application into a sequence of data items (*tokens*) and a collection of tasks (referred to as *filters* or *kernels*) that operate upon the stream of tokens as they pass through them. Filters communicate with each other explicitly by exchanging the tokens through point-to-point communication channels. This model exposes the inherent locality and concurrency of the application [14] and enables the realization of efficient implementations based on mapping the filters onto parallel processor architectures. For instance, high-performance implementations can often be obtained by mapping the filters on a pipeline of processing cores that communicate via message-passing protocol and queue buffers.

Previous works [11, 16] have shown how filters can be mapped to the actual cores to balance the load and optimally utilize the available resources. For example, if there are more filters than cores, several filters may be mapped to a single core, and if there are more cores than filters, stateless filters may be replicated on several cores. In some systems, the cores context-switch between several filters based on priority queues. However, in general it remains a challenge to achieve an optimal mapping that maximizes the stream processing throughput while accounting for the data dependencies among the filters and the available hardware resources (processing cores, memories, and interconnect). Coarse-grained distributions of the filters across the cores can result in an uneven processing load while fine-grained distributions may lead to inefficiencies due to the overhead of synchronization and data transfers among the filters.

In this paper we propose *flexible filters* as a technique to implement stream programs on distributed-memory multi-core platforms that combines static mapping of the stream program filters with dynamic load balancing of their execution. The goal is to increase the overall processing throughput of the stream program by reducing the impact of *bottleneck filters* running on particular cores. A filter can cause a bottleneck because either (a) its algorithmic characteristics make it disproportionately expensive to run on a given core with respect to the other filters running on neighboring cores or (b) at run time it may go through phases where it has to process a larger number of tokens per unit of time. When a filter becomes a bottleneck its upstream or downstream filters, or both, may start suffering a loss of throughput and, ultimately, this affects the processing throughput of the overall implementation. If the bottleneck is caused by a long-latency computation that delays the production of new tokens, the downstream filters may become idle due to the lack of inputs. But even when it continues to produce new tokens, a filter may temporarily become a bottleneck when cannot sustain the processing rate of the upstream filters. If this is the case, the input buffers of its processing core start filling up. This ultimately leads to the emission of *backpressure* signals that are sent back to the cores running the upstream filters, which are forced to become idle to avoid loss of data.

The basic idea of flexible filters is precisely to take advantage of the available cycles on these neighboring cores and use them to dynamically accelerate the execution of bottleneck filters. In other words potential bottleneck filters can be balanced by making their mapping to the underlying architecture “flexible” so that for certain periods they can run simultaneously on more than one processing core to execute different subsequences of the data stream. This is achieved with the following procedure:

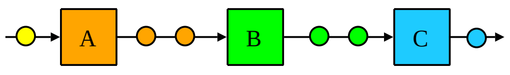

Figure 1: Example of Stream Program Structure.

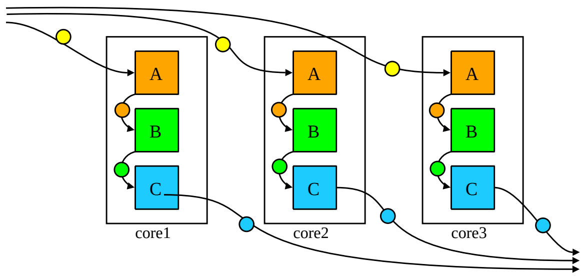

Figure 2: SPMD Mapping.

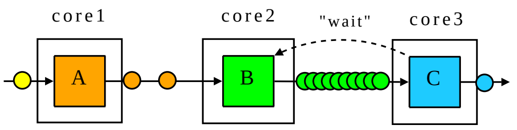

Figure 3: Baseline Pipeline Mapping.

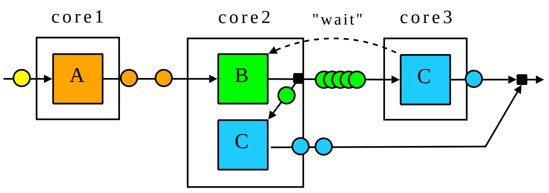

Figure 4: Flexible-Filter Mapping.

1. identification of bottleneck filters through profiling of the application with the target multi-core platform;
2. completion of a static *redundant* mapping that replicates the code of the bottleneck filters to deploy them on multiple, typically neighboring cores;
3. addition of auxiliary code that leverages the backpressure mechanism to dynamically activate the execution of the additional copies of the bottleneck filters when necessary, while preserving the correct ordering of the tokens in the data stream.

Flexible filters differ from many previous load-balancing approaches because the load balancing is based only on backpressure and the task reassignment to idle cores is guided by the data dependencies across the filters in the stream program rather than random selection. No centralized control is required, and no extra messages are sent among cores beyond backpressure messages, which are already used to prevent the communication buffers from overflowing. Since load balancing is driven by the runtime load, flexible filters can be used not only to optimize the implementation of programs whose filters have constantly unbalanced computational loads but also to adjust temporary imbalances due to spikes of activity, e.g. in Bloom filters applications.

2. FLEXIBLE FILTERS

Background. To implement a stream program on a multi-core architecture each of its filters must be mapped to at least one core. Filters may be duplicated on more than one core if they do not maintain state between two data tokens (i.e. given an input token x a stateless filter will produce the same output token regardless of what tokens came before x), One core may host several filters and rely on a scheduler so that it is time-multiplexed among them while minimizing the context-switch overhead. It may be possible to duplicate filters that have state, but additional, possibly application-dependent, bookkeeping steps would be necessary. In our approach, the ordering of tokens must be preserved in the final output of the program (though the stream may be split and joined within the program), and no tokens can be dropped without affecting the correct functionality of the program. The performance of a given implementation can be measured by its *maximum sustainable throughput (MST)*, i.e. the maximum rate at which data tokens can be processed under the assumption that the environment is always willing to produce new tokens and does not ever require the system to stall through a backpressure signal. Assuming an ideal multi-core architecture where the overhead of inter-core data communication and intra-core context switching is negligible and each core has unlimited local memory, an ideal mapping of filters would have the following properties: no core ever stalls and the MST scales linearly with the number of cores.

Three Alternative Mappings. Flexible filters can be seen as a mapping approach that sits in between two other implementation techniques: *Single Program Multiple Data (SPMD) Mapping* and *Baseline Pipeline Mapping*.

Consider the simple example of a generic stream program whose structure is shown in Fig. 1: it consists of three filters A, B, and C with data tokens traveling between them on communication channels (A, B) and (B, C). An SPMD mapping of this program on an architecture with three cores is shown in Fig. 2: each core contains the entire program and the data stream is distributed evenly among them. If the filters have *latencies*[1] $L_A = 2$, $L_B = 2$, and $L_C = 3$, respectively, then the MST obtained with this mapping is $\frac{\#\ cores}{L_A+L_B+L_C} = \frac{3}{7} = 0.429$. This value corresponds to the ideal MST for the given architecture and can be theoretically reached by the SPMD mapping because in this mapping no core receives data tokens from another core. In practice, however, the local memory of a core is not unlimited and it may not be able to accommodate the code of all of the filters. Even if it is possible, the filters' code instructions reduce the amount of buffer space for data tokens, and may reduce the ability to overlap data transfer with computation (e.g. double-buffering). Furthermore, SPMD mappings require off-chip bandwidth to scale linearly with the number of cores. Therefore, we will use the theoretical SPMD implementation as a benchmark for the stream program's ideal throughput, but we will move on to implementations that have a smaller per-core memory footprint. In particular, our

[1]The latency of a filter is the time necessary to execute it on a given core as a stand-alone task. In a heterogeneous multi-core architecture the same filter would have different latencies when executed on different cores. However, in this work we assume that cores are mapped only on homogeneous architectures.

Time Steps		t_0	t_1	t_2	t_3	t_4	t_5	t_6
core$_0$	Step	A$_{0,0}$	A$_{0,1}$	B$_{0,0}$	B$_{0,1}$	C$_{0,0}$	C$_{0,1}$	C$_{0,2}$
	Block(s)	0	0	0	0	0	0	0
core$_1$	Step	A$_{1,0}$	A$_{1,1}$	B$_{1,0}$	B$_{1,1}$	C$_{1,0}$	C$_{1,1}$	C$_{1,2}$
	Block(s)	1	1	1	1	1	1	1
core$_2$	Step	A$_{2,0}$	A$_{2,1}$	B$_{2,0}$	B$_{2,1}$	C$_{2,0}$	C$_{2,1}$	C$_{2,2}$
	Block(s)	2	2	2	2	2	2	2

Table 1: SPMD Mapping Timeline.

goal is to achieve ideal MST (like SPMD mappings) using the fewest copies of filters possible in the overall system.

A baseline pipeline mapping is shown in Fig. 3: here, each filter is mapped to a separate core. Using the same latency values as above, this mapping delivers an MST equal to $\frac{1}{3} = 0.333$ because it is bound by the latency of filter C that can process a new data token only every three time steps. Since some filters might be more computationally intensive than others, it is a challenge to keep all of the cores active when each filter is mapped to a different core. In this example, once the buffers between core$_2$ and core$_3$ (where B and C are located, respectively) fill up, core$_3$ requests core$_2$ to stall occasionally through the emission of a backpressure signal (and backpressure will continue to propagate upstream).

However, suppose that core$_2$ can also execute filter C. Then, instead of stalling, core$_2$ can "work ahead" on the data tokens in its buffers. Now the rate at which data tokens are processed by filter C is increased, and core$_3$ has fewer data tokens to process, and so the system can run faster. This is the idea of the flexible-filter mapping as illustrated in Fig. 4. Filter C is duplicated on core$_2$ so that core$_2$ can share core$_3$'s load. Besides increasing the code footprint in core$_2$ with respect to the baseline pipeline mapping, this adds also some complexity to the program because now the data stream is split and merged around core$_3$. The split and merge steps are accomplished with two *auxiliary filters*: *flex_split* and *flex_merge*. As explained in Section 4, these filters, which are represented as small black boxes in Fig. 4, can be added to the stream program without changing any of the original filters to guarantee the preservation of the order among the data tokens in the stream. Before evaluating precisely the MST of this mapping we introduce the notion of data block.

Data Blocks and Buffering Queues. Besides the filter-to-core mapping, the implementation of a stream program onto a target architecture requires also the mapping of the point-to-point communication channels between the filters. Channels are typically realized as input and output buffering queues within the private local memory associated with each core. Standard handshake protocols can be used between input and output queues to prevent buffer overflow by means of backpressure signals. Our model of input/output queues is closely related to our model of data. Each filter has an inherent input data token size which represents the minimum amount of data that the filter requires to "fire", i.e. to consume a set of input tokens and produce a set of output tokens. Communicating filters may require different data types for their tokens, and may require a different minimum number of tokens. For example, a filter that adds a constant to a stream of integers only requires a minimum of one integer to operate, while a filter that computes a matrix multiplication requires entire matrices. While sometimes a core may only have enough memory space for a single data token, it is often the case that many data tokens are stored in the queues. We abstract the capacity of input and output queues with the notion of a *data block*, which is a subsequence of data tokens. Each data block may contain many data tokens, and the blocks, like tokens, form a stream and follow an ordering that depends on their place in the bigger stream. One difference between data tokens and data blocks with respect to scheduling the flow of data is that it is possible to break a data block up into several pieces that can be executed in parallel. A data block is the input unit for *flex_split* and the output unit for *flex_merge*. The divisibility of data blocks is one factor that enables load balancing with flexible filters. But data blocks can only contain a finite number of data tokens and cannot be divided into arbitrarily sized fractions. Coarser granularity can limit the benefits of flexibility in the data stream because it puts more constraints on the possible data flow.

For the remainder of this section, we show the details of how data blocks are processed given several different mappings of our example stream program on a three core system where each core has enough buffer space for two blocks of data and can work on either block that is present in memory.

Table 1 shows the timeline for an SPMD mapping where the entire stream program is duplicated separately to each core, with no intercommunication. The table shows both the current step being executed on each core, and the contents of the core's memory. Filter A, whose latency is assumed equal to two, is computed over two time steps, which for block i are denoted $A_{i,0}$ and $A_{i,1}$, respectively. With an SPMD mapping, cores work on one block until completion and then start the next block. Though not shown in this table, depending on the method of data forwarding, each core may also be holding the next block (3, 4, and 5 for core$_0$, core$_1$, and core$_2$, respectively). As mentioned above, the MST for this mapping is $\frac{3}{7} = 0.429$ since the implementation can process three blocks every seven time steps.

Table 2 illustrates the trace when the filters are mapped with a baseline pipeline mapping, i.e. where each filter is assigned to only one distinct core. Even though the filters' latencies are not equal, the buffer capacity allows the faster filters to work ahead. However, in time step t_{18}, core$_2$ must stall. At this step, core$_2$'s memory contains blocks 6 and 7, and even though core$_1$ is ready to pass block 8, core$_3$ holds blocks 4 and 5 and will not be ready to take the next block from core$_2$ until it is done processing block 4. Therefore, core$_2$ must wait until core$_3$ is ready to accept the next block before it can make space in its queue for block 8. The state of the system is the same in time steps t_{22} and t_{25} in terms of the state of each core with respect to the blocks in that core's memory. In fact, the state of the system will begin to cycle through a pattern of states, in this case the pattern from t_{22} to t_{24}. By analyzing the behavior during one iteration of the cycle, we can have confirmation that the baseline pipeline implementation has an MST equal to $\frac{1}{3}$. Note that if the filter latencies are unbalanced, stalling will occur no matter how much buffer space is available on the cores: extra buffer space simply extends the time that it takes to initially fill up the buffers.

Time Steps		t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}
core$_0$	Step	$A_{0,0}$	$A_{0,1}$	$A_{1,0}$	$A_{1,1}$	$A_{2,0}$	$A_{2,1}$	$A_{3,0}$	$A_{3,1}$	$A_{4,0}$	$A_{4,1}$	$A_{5,0}$	$A_{5,1}$	$A_{6,0}$	$A_{6,1}$	$A_{7,0}$	$A_{7,1}$
	Block(s)	0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7
core$_1$	Step			$B_{0,0}$	$B_{0,1}$	$B_{1,0}$	$B_{1,1}$	$B_{2,0}$	$B_{2,1}$	$B_{3,0}$	$B_{3,1}$	$B_{4,0}$	$B_{4,1}$	$B_{5,0}$	$B_{5,1}$	$B_{6,0}$	$B_{6,1}$
	Block(s)			0	0	1	1	2	2	3	3	4	4	4,5	5	5,6	5,6
core$_2$	Step					$C_{0,0}$	$C_{0,1}$	$C_{0,2}$	$C_{1,0}$	$C_{1,1}$	$C_{1,2}$	$C_{2,0}$	$C_{2,1}$	$C_{2,2}$	$C_{3,0}$	$C_{3,1}$	$C_{3,2}$
	Block(s)					0	0	0,1	1	1,2	1,2	2,3	2,3	2,3	3,4	3,4	3,4

Time Steps		t_{16}	t_{17}	t_{18}	t_{19}	t_{20}	t_{21}	t_{22}	t_{23}	t_{24}	t_{25}
core$_0$	Step	$A_{8,0}$	$A_{8,1}$	$A_{9,0}$	$A_{9,1}$	$A_{10,0}$	$A_{10,1}$	$A_{11,0}$	$A_{11,1}$	-	$A_{12,0}$
	Block(s)	8	8	8,9	9	9,10	9,10	10,11	10,11	10,11	11,12
core$_1$	Step	$B_{7,0}$	$B_{7,1}$	-	$B_{8,0}$	$B_{8,1}$	-	$B_{9,0}$	$B_{9,1}$	-	$B_{10,0}$
	Block(s)	6,7	6,7	6,7	7,8	7,8	7,8	8,9	8,9	8,9	9,10
core$_2$	Step	$C_{4,0}$	$C_{4,1}$	$C_{4,2}$	$C_{5,0}$	$C_{5,1}$	$C_{5,2}$	$C_{6,0}$	$C_{6,1}$	$C_{6,2}$	$C_{7,0}$
	Block(s)	4,5	4,5	4,5	5,6	5,6	5,6	6,7	6,7	6,7	7,8

Table 2: Baseline Pipeline Mapping Timeline.

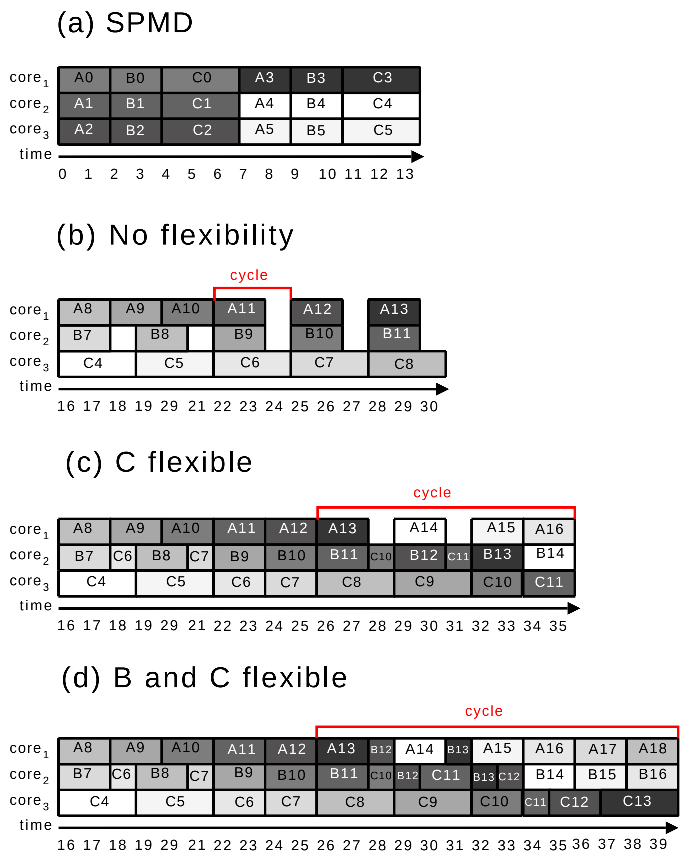

Figure 5: Flexible Filter Timelines

Fig. 5 summarizes timelines for several mapping in a more abbreviated format that does not include the current memory state. For each case other than SPMD, the timelines start during t_{16} using the same state of t_{16} in Table 2 and continue until a cyclic pattern emerges. Fig. 5(a) and (b) depict the same timelines as in Tables 2 and 1, respectively. Fig. 5(c) shows the timeline for a flexible-filter mapping where C is made flexible and is mapped to core$_2$ and core$_3$ (same as Fig. 4). The cyclic pattern for this mapping begins at time step t_{26} and continues until t_{35}. Fig. 5(d) shows an alternative flexible-filter mapping where both B and C have been made flexible. In particular, C is again mapped to core$_2$ and core$_3$, while B is mapped to core$_1$ and core$_2$. Here, the pattern goes from time step t_{26} to t_{39}.

In our example, when no filters are flexible, the MST is degraded by 22%. When only C is flexible, the MST is increased to $\frac{4}{10} = 0.400$ (only 7% degradation). When both B and C are flexible, the MST reaches its ideal limit of 0.429, thus matching the MST of the SPMD mapping. Note that we are not simply breaking up a filter to achieve data parallelism (e.g. as in [10]); instead, data parallelism is used to dynamically balance load as an alternative to stalling one of the processing cores in response to backpressure. In an implementation with optimal MST, such as the examples of Fig. 5(a) and (d), all of the cores are always busy, but by using flexible filters only one copy of filter A, and two copies of B and C are necessary instead of the three copies of each filter that are used in the SPMD implementation.

Granularity of Firing Constraints and Buffer Size. In the previous examples, we assumed that it is always possible to break one of C's data blocks up into thirds and B's data blocks up into halves. Suppose, however, that C requires such large data tokens that a data block only holds two tokens. Since data tokens are the minimum amount of data that a filter can fire on, it is now only possible to break one of C's data blocks up into two pieces. Fig. 6 repeats the mapping from Fig. 5(c) to show the timeline when C has this constraint. There are two cases shown. In Fig. 6(a), we assume buffers of size two just like the previous examples, while in Fig. 6(b) we assume that the buffer has capacity for one data block only. At t_7 in Fig. 6(b), core$_3$ must wait for Block 1 until core$_2$ is ready to send it. Similarly, core$_1$ must also wait to send Block 2 to core$_2$. When buffers have enough capacity for two blocks, the MST is $\frac{6}{15} = 0.4$, which is the same as the MST when we did not have the additional granularity constraint. However, when the buffers only hold one block, the MST is degraded to $\frac{2}{5.5} = 0.364$. From this example, we can see that buffers play a critical role in insulating performance from granularity constraints.

3. CALCULATING THE MST

In this section we describe a method to calculate the MST of a mapping of a stream program on a given multi-core architecture. We assume static filter latencies. This assumption simplifies the analysis, and allows us to compare MST across different filter configurations. However, notice that dynamic load balancing with flexible filters does not require that the filters have a fixed latency. If the latencies change at runtime, the system automatically settles into the best possible schedule given the new latencies and the static mapping.

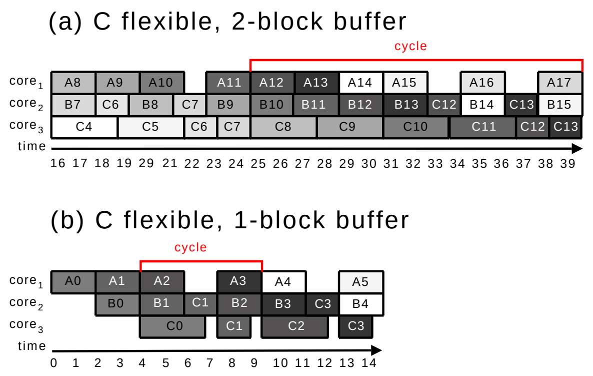

Figure 6: Time-line when filter C has a granularity of 2 tokens per block.

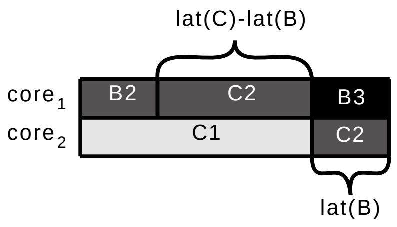

Figure 7: One flexible filter.

One Flexible Filter. When only one filter is flexible across two cores, the MST of the two cores together can be calculated as follows: given filters B and C with latencies L_B and L_C, where B is fixed and C is flexible, if $L_C > L_B$, the MST $= \frac{2}{L_B+L_C}$. If the latency of the non-flexible B exceeds the latency of flexible C ($L_B > L_C$), then there is no MST improvement because B is the bottleneck. The timeline in Fig. 7 shows how filters B and C overlap when $L_C > L_B$. If $L_B > L_C$ and B is flexible and pushed downstream, the portions would overlap similarly to achieve MST $= \frac{2}{L_B+L_C}$.

A Pipeline of Flexible Filters. Consider the case where two consecutive filters are flexible next to each other. For instance, consider the example of Fig. 5(d) with filters A, B, C, where B and C are flexible. We would like to know how close the performance of a system with this flexible configuration can come to the ideal MST for an architecture with a given number of cores. Based on which filters are flexible, we can define the following linear programming problem to determine the system's MST: For each core core$_i$, we construct a variable X_i based on the percentage of time that core$_i$ spends on each filter. In the above example, core$_1$'s variable $X_1 = X_{1A} + X_{1B} + X_{1W}$, where X_{1A} and X_{1B} correspond to the time that core$_1$ executes filters A and B, respectively, and X_{1W} corresponds to the time that core$_1$ waits. The execution time of a filter partitioned across multiple cores is defined to be equal to the sum of the execution time of the part of the filter executed on each core. For example, core$_2$ and core$_3$ may work on a filter C and the sum of time spent on filter C by both cores must equal L_C. The X_i variables correspond to the cyclical firing pattern described with Fig. 5. The complete linear programming problem for our example is the following:

Minimize $X_{1W} + X_{2W} + X_{3W}$ (wait time), subject to

$$\begin{aligned} X_{1A} &= L_A \\ X_{1B} + X_{2B} &= L_B \\ X_{2C} + X_{3C} &= L_C \end{aligned}$$

which ensure that the total time spent working on a block for a filter by the cores matches the latency of that filter, and:

$$\begin{aligned} X_1 - X_2 = X_{1A} + X_{1B} + X_{1W} - X_{2B} - X_{2C} - X_{2W} &= 0 \\ X_2 - X_3 = X_{2B} + X_{2C} + X_{2W} - X_{3C} - X_{3W} &= 0 \end{aligned}$$

which ensure that the wait times are selected so that all cores work (and wait) over the same period of time. One solution for the example program is $X_{1A} = 2$, $X_{1B} = \frac{1}{3}$, $X_{2B} = \frac{5}{3}$, $X_{2C} = \frac{2}{3}$, $X_{3C} = \frac{7}{3}$, and $X_{1W}, X_{2W}, X_{3W} = 0$.

We can now give the constraints for general applications. For each filter f,

$$\sum_i X_{if} = L_f$$

for all i such that f is mapped to core$_i$. And for each core$_i$ (except the core with the highest index),

$$\sum_{f \in filters\ on\ core_i} X_{if} \quad - \quad \sum_{f \in filters\ on\ core_{i+1}} X_{(i+1)f} = 0$$

Additional constraints could be added for granularity restrictions, and the problem could be solved as an integer linear program.

Using the solution to the problem we can compute the MST as function of the ideal MST as follows:

$$\begin{array}{ccc} \%\ of\ time\ doing\ useful\ stuff & * & ideal\ MST \\ \dfrac{X_1 + X_2 + X_3 - (X_{1W} + X_{2W} + X_{3W})}{X_1 + X_2 + X_3} & * & \dfrac{3}{L_A + L_B + L_C} \end{array}$$

4. SPLIT AND MERGE

As mentioned in Section 2, once the filters that will be flexible have been selected the original stream program is transformed into a flexible stream program by duplicating these filters and by adding pairs of *flex_split* and *flex_merge* auxiliary filters around the duplicated filters. No additional programming is required from the programmer for this step.

Flex_split and *flex_merge* can be provided by an application independent library. They rely on an underlying message passing API for the target architecture that supports the handshake communication protocol and backpressure mechanism between communicating cores. Before a core can send data downstream, it needs to check that there is buffer space for the data. A typical handshake protocol starts with the sending core placing a request to send data; the receiving core sends back an acknowledgement about how much data it can receive; and finally the sending core sends the data. The handshake ensures that buffers do not overflow by sending backpressure signals upstream (through the acknowledgement signals).

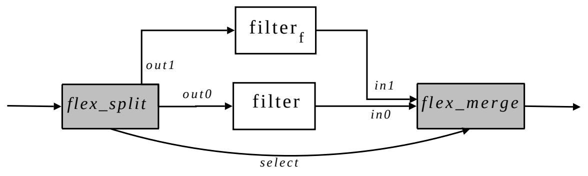

Figure 8: Making a filter flexible.

Algorithm 1 flex_split
Input: stream in; Output: streams $out0, out1, select$

```
pop data block $b$ from $in$
$n0 \leftarrow avail(out0)$
$n1 \leftarrow |b| - n0$
for $i = 0$ to $n0 - 1$ do
    $select[i] \leftarrow 0$
end for
for $i = n0$ to $|b| - 1$ do
    $select[i] \leftarrow 1$
end for
push $n0$ tokens from $b$ to $out0$
push $n1$ tokens from $b$ to $out1$
```

Flex_split (pseudocode shown in Algorithm 1) reuses the information about how much buffer space is available to direct load balancing. *Flex_split* checks how much space is available on the queue for filter f's primary copy and divides the data stream by sending as much data to f's primary copy that it can, and then sending any leftover data to the flexible copy. *Flex_merge* (pseudocode in Algorithm 2) takes input from both of f's copies along with a bitstream from *flex_split* that tells which copy has the next data token. Fig. 8 shows how the filters are connected to each other. *Flex_split* divides the data stream between $out0$ and $out1$, and produces a *select* stream that contains the information on how to reconstruct the correct ordering of the stream. *Flex_merge* takes the *select* stream directly from *flex_split* together with two input streams $in0$ and $in1$ to reassemble the stream into its original order. *Flex_split* and *flex_merge* are added to the stream program in the same way regardless of whether the flexibility will be pushed to the filter's upstream or downstream neighbor. The direction of flexibility is determined based on where the duplicate filter is mapped.

Note that another possible implementation of *flex_split* would consult the availability on $out1$. In this implementation we simply send everything leftover to $out1$ and let the underlying stream mechanisms stall the data stream in the case that this is too much data for $out1$ as well. If a flexible filter is inherently slower than the rest of the filters, then the imbalance will cause its queue to be full often, and the data flow will be redirected at regular intervals. If the flexible filter instead is usually even with its neighbors and only occasionally has spikes of activities that cause it to slow down, or if its upstream neighbor occasionally creates extra data tokens on its output, then in normal circumstances the data flow will follow the baseline pipeline behavior. But during spikes of data or activity *flex_split* will adapt the data flow accordingly. Notice also that the *select* data structure may be compressed to counts of how many of the next tokens go to $out0$ and then how many go to $out1$, instead of a separate *select* element for every token. In practice, if the data tokens are vectors or other large data structures, then using a separate select element for each token will not take up an unreasonable portion of memory.

Algorithm 2 flex_merge
Input: stream $in0, in1, select$; Output: stream out

```
pop $i$ from $select$
if $i$ is 0 then
    pop token $t$ from $in0$
else
    pop token $t$ from $in1$
end if
push $t$ to $out$;
```

Filter Mapping. The last step in our approach is mapping the entire set of filters to cores including duplicate flexible filters and auxiliary filters. Flexible filter mapping starts with a baseline pipeline mapping for the filters of the stream, and a duplicate copy of a flexible filter is mapped to the forward or backward neighbor of the core that the original filter is mapped to. If several consecutive filters are co-located on a single core, it may be useful to treat them as a single filter for flexibility purposes to cut back on the overhead of data queueing between filters. As filters are mapped to cores, one design decision is where to map the split and merge filters. All of the data stream must pass through both filters. In the case of a single flexible filter, which is likely a performance bottleneck, it is natural to map the split and merge filters to the same cores as the flexible filter's upstream and downstream neighbors, respectively. When there are several consecutive flexible filters, however, the best choice is less clear. One possibility is to map the split/merge pair between two flexible filters onto a separate core. We explore a few possibilities in our experiments in the next section.

5. EXPERIMENTS

In this section we describe a case study of flexible filters used with a JPEG encoder. We performed all of our experiments on a QS21 CellBlade [19] which hosts two IBM Cell processors. The Cell architecture is a heterogeneous multicore system-on-chip originally designed for high-performance embedded applications. It features one PowerPC processing core called the PPU, eight synergistic SIMD processing units called SPUs, and a powerful intercommunication network called the Element Interconnect Bus (EIB) capable of sustaining 205 GB/s of data transfers [2]. Each SPU has a 256 KB local memory that is shared between code and data. The Cell processor is a good architecture for testing the performance of flexible filters since it exposes the tradeoff in memory for code and buffer space when we make filters flexible.

To program the IBM Cell we took advantage of Gedae [1], a data flow language, that has provided the communication layer for our implementation, handling low-level details like direct-memory access (DMA) alignment and double buffering. Gedae provides API functions to implement the communication channels, including an `avail()` function that

[2]Notice that in our experiments we mapped the filters only to the SPUs processors.

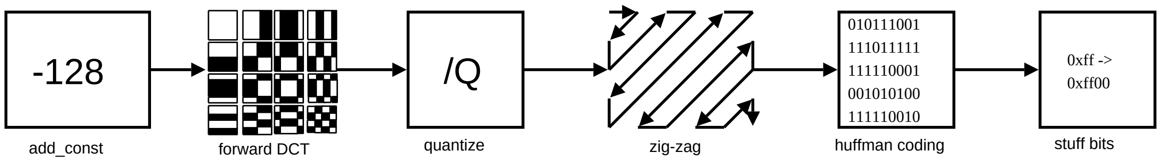

Figure 9: JPEG block diagram

filter	description
add_const	shifts values by 128
forward DCT	discrete cosine transform
quantize	normalizes results, outcome is that many values become zero and are easier to compress
zigzag	converts 2D block to 1D vector by listing block elements in zigzag order
coding	applies Huffman coding to data stream
stuff_bits	packs bits into integers and adds some padding bits

Table 3: Description of JPEG filters

gives feedback to the user about how much space is available in input and output buffers.

Fig. 9 shows a block level diagram of a baseline gray-scale JPEG encoder, and Table 3 describes the functionality of each filter. Each of the filters operates on 8x8 pixel blocks. The following filters are stateless between blocks: *add_const*, *forward DCT*, *quantize*, and *zigzag*. Filter *coding* does retain a small amount of state (one integer: the previous block's DC coefficient). Since there is state, it is slightly more difficult to make this filter flexible, but in this particular case, the DC coefficient of the previous block can also be captured in *zigzag* or *quantize* and then tagged on to the current block. So flexibility is possible while requiring a small amount of bookkeeping. Filter *Stuff_bits* packs the bits produced by Huffman coding into bytes so they can be written to output. In addition, *stuff_bits* checks for `0xff` words in the data stream and adds padding bits when it finds them (`0xff` is reserved as a tag marker in the bitstream). It would be more difficult to make this filter flexible because codes produced by *coding* have variable length and may span across byte boundaries in the output of *stuff_bits*.

Fig. 10 shows some profiling data that we collected empirically with a baseline pipeline implementation of the JPEG encoder. Since our first implementation of the *DCT* filter was rather slow in comparison to the other filters, we made two additional implementations. We used all three versions to evaluate how the benefits of flexibility vary depending on the relative latency of the flexible filter. Filters *Add_const* and *zig-zag* are omitted because their measured latencies are negligible. Indeed, they are lower than the latency of the function used to measure the latency on the Cell (200-300 ns, called twice).

Below is a brief description of the DCT implementations:

- *Slow.* Implements the 2-D DCT based on its standard definition, $f(x,y) = \sum_{u=0}^{n-1}\sum_{v=0}^{n-1} T(u,v)h(x,y,u,v)$ [9].

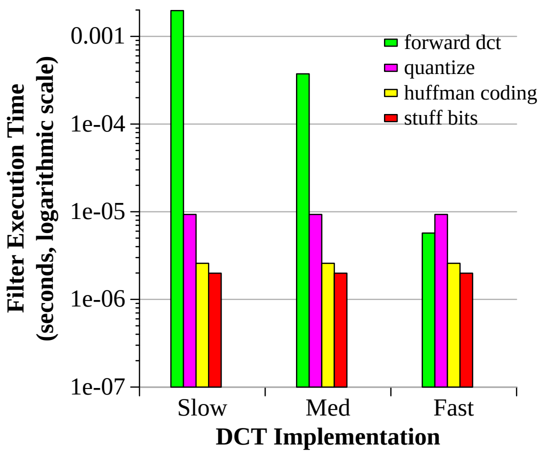

Figure 10: Profile of JPEG Filters for three DCT Implementations

- *Medium.* Computes the 2-D DCT by computing 1-D DCT on each column of the 8x8 pixel block and then on each row of the block. The 1-D DCTs are implemented with `for` loops.

- *Fast.* Similarly to the medium algorithm the 2-D DCT is computed as 1-D DCTs over the columns and rows of the block. The 1-D DCT is implemented in a pipeline fashion, based on the algorithm described in [15].

Notice that our implementation is not highly optimized for the Cell architecture; for example, we did not make use of the Cell intrinsic vector operations. A more efficient encoder could certainly be written that takes advantage of vector operations on the Cell. With regard to flexible filters, the implementation can change the relative costs of different filters with respect to each other and to the overhead of data transfer, but it would not change the algorithms or strategies of flexible filter load balancing. We stop at a basic implementation since our goal is not to make the most efficient JPEG encoder for the Cell, but to test the flexible filter idea.

Implementation with One Flexible Filter. Fig. 11 shows the speedup from using a flexible filter for *forward DCT*, using the mapping shown in Fig. 12. All other filters are not flexible. We observe that adding flexibility to the stream program can significantly improve performance and also that the benefits of flexibility are greater when there is a greater imbalance between the latency of *forward DCT* and the latencies of other filters. Performance improved nearly 85% for the slow DCT implementation and up to 20% for the fast DCT implementation. One unexpected result is that the performance shows improvement for the fast implementation where *forward DCT* is not the system bottleneck.

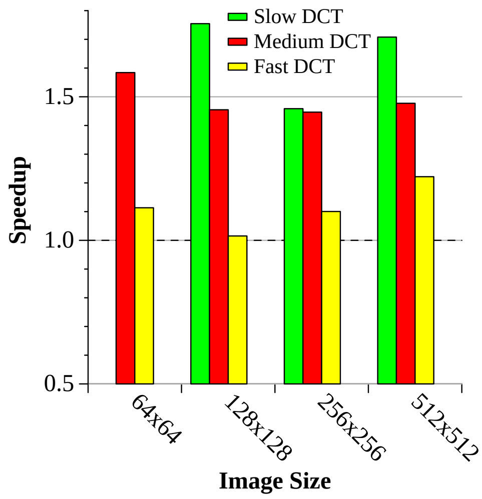

Figure 11: Speedup gained from giving the DCT filter flexibility degree 1. Image sizes are 64x64, 128x128, 256x256, and 512x512 pixels.

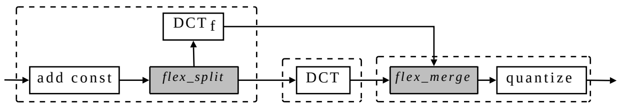

Figure 12: Mapping Flexibility to Cores.

Table 4 shows the actual time to process the images. In some cases, the medium DCT implementation with flexibility outperforms the fast DCT implementation with no flexibility.

Implementation with Two Flexible Filters. In the fast DCT implementation, filter *quantize* is the system bottleneck. Fig. 13 shows three possible mappings for the filters when we make both *forward DCT* and *quantize* flexible. The first mapping follows the convention of mapping *flex_split* to the upstream neighbor of the flexible filter, and *flex_merge* to the downstream neighbor. Since two filters in a row are flexible, the first *flex_merge* must come before the second *flex_split.* The second mapping attempts to reduce the amount of data being transferred in exchange for reduced flexibility. The third mapping offloads *flex_merge* and *flex_split* between *forward DCT* and *quantize* to a separate core. As reported in Fig. 14 the performance of all cases is roughly the same. Speedup is as high as 25% for the 512x512 image, though there is no speedup for the 128x128 image. To illustrate how the profile data translates to dynamic load we collected the following measures: for the 512x512 image, when only *forward DCT* is flexible, *flex_split* redirects around 35% of the data tokens to the secondary filter. When both *forward DCT* and *quantize* are flexible using the first mapping from Fig. 13, 47% of the data is redirected to the secondary *forward DCT* and 34% is redirected to the secondary *quantize.*

Overhead of Flexibility. If the absence of overhead from data transfer and inefficiencies from data granularity, we would expect, contrary to the results in Fig. 14, that making both *forward DCT* and *quantize* flexible should improve performance. It turns out, however, that the extra communication overhead outweighs the benefits. In addition to the overhead of moving the data, flexible filters incur some additional overhead. First, *flex_split* and *flex_merge* require a couple extra steps of data copying; in addition, streams that can push and pop arbitrary numbers of data tokens according to user-level input are more complex than streams that always push and pop the same number of tokens; finally, flexible filters require some code duplication compared to a baseline pipeline pipeline (though less code overhead than an SPMD mapping).

Implementation	Encoding Time per Image (seconds) 64x64	128x128	256x256	512x512
Slow DCT	0.099	0.388	1.542	6.253
Slow DCT, flex	0.059	0.231	0.871	3.395
Med DCT	0.022	0.086	0.347	1.334
Med DCT, flex	0.014	0.059	0.240	0.903
Fast DCT	0.013	0.057	0.236	0.954
Fast DCT, flex	0.012	0.056	0.215	0.781

Table 4: Performance of JPEG Encoder

6. RELATED WORKS

Flexible filters balance load using a version of work stealing for stream programs. Work stealing is a technique used in a variety of parallel systems [8, 3, 13] to balance load by allowing idle cores to "steal" tasks from busy cores. Most work stealing techniques go through stages of load evaluation, reassignment and task migration; and their "victim" processors (from whom tasks will be stolen) are selected randomly. In contrast, flexible filters do not steal randomly, but use the knowledge that neighbors of a bottleneck filter will be idle because they dependent on this filter to continue processing data tokens. Items are never migrated between queues of different processors, instead when queues become full new items are redirected elsewhere. With flexible filters, tasks are not "stolen" per-se but rather the data flow is re-routed when a bottleneck arises. Flexible filters are identified when filters are mapped to the system and determine the available routes for data during runtime. In addition, we focus on the case of a distributed memory system where the code for tasks is also distributed. Flexible filters are specialized to stream programs because dependencies in the stream allow us to narrow down good candidates for redundant-code placement.

Load balancing approaches specific to stream programs can be categorized depending whether the stream models rely on data parallelism or pipeline parallelism (in practice both approaches can be used simultaneously [10]) In data parallel stream systems [2, 20], there can be many producers that feed many consumers, and there may be many instances of producer and consumer functions. Load balancing is achieved by routing data to different instances of consumers based on their current load and productivity. On the other hand, in pipeline parallel stream systems, the data may need to flow through a series of pipelined filters where each filter can be viewed as a producer and consumer of input and output data. The order of filters constrains the order in which tasks may be executed.

Flexible filters are a solution for load balancing of pipeline parallelism in stream programs. Many related works [21, 7, 23] for balancing load of pipeline parallel stream programs

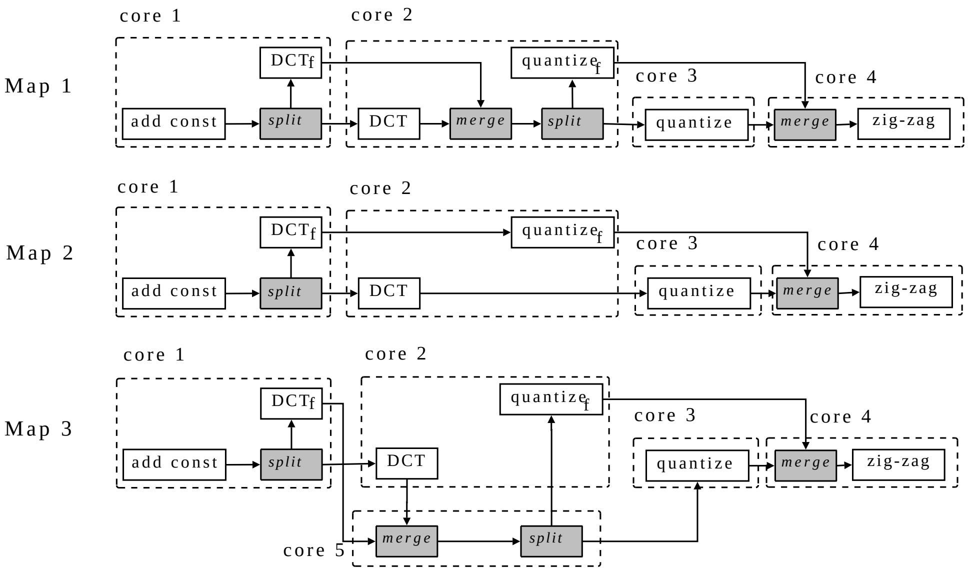

Figure 13: Mapping Pipeline Flexibility.

involve a central control and/or phases where the compute nodes collect statistics which are used by the control to direct reorganization. The number of filters is designed to outnumber the number of cores, and load balancing is typically achieved by moving filters from nodes with heavy loads to nodes with lighter loads. Flexible filters simulate filter migration by duplicating some filters on the cores and invoking duplicates when the load becomes unbalanced. Chen *et al.* [6] perform load balancing for stream programs by compiling several alternative filter mappings. During runtime, the system can "context-switch" between the alternatives based on the properties of the data. Flexible filters, on the other hand, dynamically adapt to the current flow behavior of the system. In the DIAMOND system developed by Huston *et al.* [12], data tokens are forwarded based on threshold values in the input and output queues. Load balancing with flexible filters similarly is an outcome of the state of the queues, but the difference is that flexible filters balance load based on backpressure. Moreover, DIAMOND is optimized for distributed search which relaxes several constraints of stream programs - namely that the filters need not be executed in a particular order because they are used to eliminate unwanted data (rather than transform the data) and that data can be processed in any order.

Synchronous Data Flow (SDF) [17] is a well studied model of computation which like the stream programming model defines a computation as a network of processing nodes through which data flows. We support a less restrictive set of applications than SDF with flexible filters because flexible filters tolerate different exchange rates and adapt to improve performance in spite of them.

7. CONCLUDING REMARKS

Flexible filters can significantly improve the performance of stream programs. They are especially effective in cases where one filter has a relatively high execution latency compared to other filters in the program. Since flexible filters automatically adapt the data flow to the latencies of the filters,

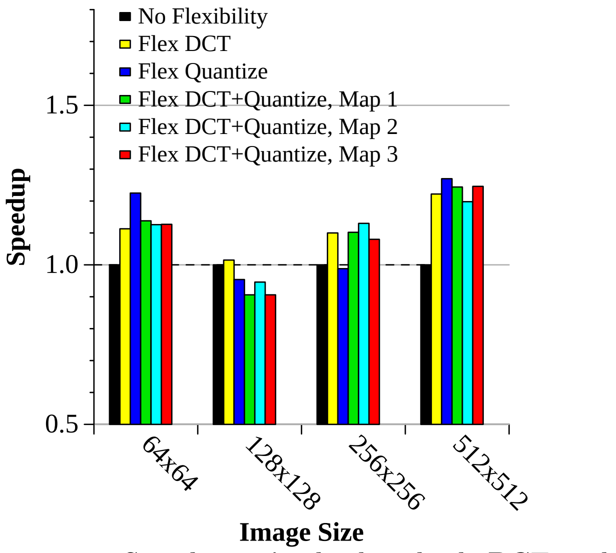

Figure 14: Speedup gained when both DCT and Quantize are flexible using the Fast DCT Implementation and mappings shown in Fig. 13.

they can reduce the need to break large filters up by hand. In addition, load balancing is determined solely by backpressure signals, and can be applied both to systems with static filter latencies and systems with dynamically varying data flow.

Understanding the overhead of data transfer and granularity and how the communication aspects of stream programs interact with their maximum sustainable throughput is an important area of future work. Communication overhead can be tied together with flexible filter mappings to take an optimization approach that aims at determining the best mapping for a stream program on a given architecture. Flexible filters have also the potential to improve the throughput of more general stream programs than those that were explored in this work, e.g. programs with stateful filters and with more complex graph structure including split/join filters. The cost of bookkeeping for these cases remains to be studied and quantified.

8. REFERENCES

[1] Gedae, http://www.gedae.com/.

[2] R. H. Arpaci-Dusseau, E. Anderson, N. Treuhaft, D. E. Culler, J. M. Hellerstein, D. Patterson, and K. Yelick. Cluster I/O with River: Making the fast case common. In *IOPADS '99: Proceedings of the sixth workshop on I/O in parallel and distributed systems*, pages 10–22, New York, NY, USA, 1999. ACM.

[3] M. A. Bender and M. O. Rabin. Online scheduling of parallel programs on heterogeneous systems with applications to Cilk. *Theory of Computing Systems Special Issue on SPAA*, 35:2002, 2002.

[4] I. Buck, T. Foley, D. Horn, J. Sugerman, K. Fatahalian, M. Houston, and P. Hanrahan. Brook for GPUs: Stream computing on graphics hardware. In *SIGGRAPH '04: ACM SIGGRAPH 2004 Papers*, pages 777–786, 2004.

[5] S. Chandrasekaran, O. Cooper, A. Deshpande, M. J. Franklin, J. M. Hellerstein, W. Hong, S. Krishnamurthy, S. Madden, V. Raman, F. Reiss, and M. Shah. TelegraphCQ: Continuous dataflow processing for an uncertain world. In *CIDR Conference*, 2003.

[6] J. Chen, M. I. Gordon, W. Thies, M. Zwicker, K. Pulli, and F. Durand. A reconfigurable architecture for load-balanced rendering. In *HWWS '05: Proceedings of the ACM SIGGRAPH/EUROGRAPHICS Conference on Graphics hardware*, pages 71–80, New York, NY, USA, 2005. ACM.

[7] Eric T. Fellheimer. Dynamic load-balancing of StreamIt cluster computations. Master's thesis, Massachusetts Institute of Technology, 2005. Department of Electrical Engineering and Computer Science.

[8] M. Frigo, C. E. Leiserson, and K. H. Randall. The implementation of the Cilk-5 multithreaded language. In *In Proceedings of the SIGPLAN Conference on Program Language Design and Implementation*, pages 212–223, 1998.

[9] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001.

[10] M. I. Gordon, W. Thie, and S. Amarasinghe. Exploiting coarse-grained task, data, and pipeline parallelism in stream programs. *SIGOPS Oper. Syst. Rev.*, 40(5):151–162, 2006.

[11] J. Gummaraju and M. Rosenblum. Stream programming on general-purpose processors. In *MICRO 38: Proceedings of the 38th annual IEEE/ACM International Symposium on Microarchitecture*, pages 343–354, Washington, DC, USA, 2005. IEEE Computer Society.

[12] L. Huston, A. Nizhner, P. Pillai, R. Sukthankar, P. Steenkiste, and J. Zhang. Dynamic load balancing for distributed search. In *HPDC '05: Proceedings of the High Performance Distributed Computing, 2005. HPDC-14. Proceedings. 14th IEEE International Symposium*, pages 157–166, Washington, DC, USA, 2005. IEEE Computer Society.

[13] P. Kakulavarapu, O. Maquelin, and G. R. Gao. Dynamic load balancing in multithreaded multiprocessor systems. In *Parallel Processing Letters*, pages 169–184, 2001.

[14] U. J. Kapasi, S. Rixner, W. J. Dally, B. Khailany, J. H. Ahn, P. Mattson, and J. D. Owens. Programmable stream processors. *Computer*, 36(8):54–62, 2003.

[15] M. Kovac and N. Ranganathan. Jaguar: A fully pipelined VLSI architecture for JPEG image compression standard. *Proceedings of the IEEE*, 83(2):247–258, 1995.

[16] M. Kudlur and S. Mahlke. Orchestrating the execution of stream programs on multicore platforms. *SIGPLAN Not.*, 43(6):114–124, 2008.

[17] E.A. Lee and D.G. Messerschmitt. Synchronous data flow. *IEEE*, 75(9):1235–1245, September 1987.

[18] M. D. McCool. Data-parallel programming on the Cell BE and the GPU using the RapidMind development platform. In *GSPx Multicore Applications Conference*, Santa Clara, October 2006.

[19] A. K. Nanda, J. R. Moulic., R. E. Hanson, G. Goldrian, M. N. Day, D. B. D'Arnora, and S. Kesavarapu. Cell/B.E. blades: Building blocks for scalable, real-time, interactive, and digital media servers. *IBM J. Res. Dev.*, 51(5):573–582, 2007.

[20] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis. Evaluating MapReduce for multi-core and multiprocessor systems. In *Proc. of the Symposium on High Performance Computer Architecture*, February 2007.

[21] M. A. Shah, J. M. Hellerstein, S. Chandrasekaran, and M. J. Franklin. Flux: An adaptive partitioning operator for continuous query systems. In *International Conference on Data Engineering (ICDE)*. IEEE, 2003.

[22] W. Thies, M. Karczmarek, M. Gordon, D. Maze, J. Wong, H. Ho, M. Brown, and S. Amarasinghe. StreamIt: A compiler for streaming applications, December 2001. MIT-LCS Technical Memo TM-622, Cambridge, MA.

[23] Y. Xing, S. Zdonik, and J.-H. Hwang. Dynamic load distribution in the Borealis stream processor. In *International Conference on Data Engineering (ICDE)*, pages 791–802, Washington, DC, USA, 2005. IEEE Computer Society.