

# The HPC-DAG Task Model for Heterogeneous Real-Time Systems

**Abstract**—Recent commercial hardware platforms for embedded real-time systems feature heterogeneous processing units and computing accelerators on the same System-on-Chip. When designing complex real-time application for such architectures, the designer is exposed to a number of difficult choices, like deciding on which engine to execute a certain task, or what degree of parallelism to adopt for a given computing function.

To help the designer exploring the wide space of design choices and tune the scheduling parameters, we propose a novel real-time application model, called HPC-DAG (*Heterogeneous Parallel Condition Directed Acyclic Graph Model*), specifically conceived for heterogeneous platforms. An HPC-DAG allows specifying alternative implementations of a software component for different processing engines, as well as conditional branches to model if-then-else statements.

We also propose a schedulability analysis for the HPC-DAG model and a heuristic allocation algorithm so that all deadlines are respected. Our analysis takes into account the cost of preempting a task, which can be non-negligible on certain processors. We show the use of our approach on a realistic case study, and we demonstrate its effectiveness by comparing it with state-of-the-art algorithms in the literature.

**Index Terms**—Real-time partitioning, heterogeneous architecture, unrelated, preemption

## I. INTRODUCTION

Modern cyber-physical embedded systems are increasingly complex and demand powerful computational hardware platforms. A recent trend in hardware architecture design is to combine high performance multi-core CPU hosts with a number of application-specific accelerators (e.g. Graphic Processing Units – GPUs, Deep Learning Accelerators – DLAs, or FPGAs for programmable hardware) in order to support complex real-time applications with machine learning and image processing software modules. Such application-specific processors are defined by different Instruction Set Architecture (ISA).

As an example of modern heterogeneous computing platform, the NVIDIA Volta GPU architecture<sup>1</sup>, couples a fairly traditional GPU architecture (hundreds of small SIMD processing units, called *CUDA cores*, grouped in computing clusters, called *Streaming Multiprocessors*) with hardware pipelines specifically designed for tensor processing (*Tensor Cores*), like *matrix multiply and accumulate* operations that are typical of neural network arithmetics. The integrated version of the NVIDIA Volta architecture is embedded within the NVIDIA Xavier SoC, which can now be found in the NVIDIA Jetson AGX board and in the NVIDIA Pegasus board. In such embedded

platforms, tensor processing can also be operated in specifically designed compute engines such as the DLA (Deep Learning Accelerator<sup>2</sup>). Moreover, another application specific engine is the PVA (Programmable Vision Accelerator), designed for executing signal processing algorithms, such as *stereo disparity* and *optical flow*.

In such platforms, the main and novel challenge for a sound timing analysis is represented by the drastic differences between the various computing engines, at the level of ISA, preemption capability, memory hierarchy and inter-connection.

When programming these platforms, the software designer is confronted with several design choices: on which processor engine should a task be executed? Should a certain sub-system be implemented in parallel or sequentially? These choices could significantly impact the timing behavior of the application and the resource utilization. The analysis is complicated by the fact that, on certain processors, the overhead induced by preempting a lower priority task can be large: for example, the overhead of preempting a graphical task executing on certain GPU architectures is in the same order of magnitude of the worst-case execution time of the task. As we will see in Section VII-C, such overhead depends on the computing engine and on the type of task.

**Contributions.** To help the designer exploring the design space, a novel model of real-time task called HPC-DAG (Heterogeneous Parallel Conditional Directed Acyclic Graph) is presented in Section II. Thanks to the graph structure, the HPC-DAG model allows specifying the degree of parallelism of real-time sub-tasks. The designer can use special *alternative* nodes in the graph to model alternative implementations of the same functionality on different computing engines to be selected off-line, and *conditional* nodes in the graph to model if-then-else branches to be selected at run-time. Alternative nodes are used to leverage the diversity of computing accelerators within our target platform. In Section III, we demonstrate the use of the HPC-DAG model by representing a module of the Autoware library.

Then, in Section IV we present a schedulability analysis that will be used in Section V by a set of allocation heuristics to map tasks on computing platforms and to assign scheduling parameters. In particular, we present a novel technique to reduce the pessimism due to high preemption costs in the analysis (Section IV-F).

After discussing related work in Section VI, our methodology is evaluated in Section VII by comparing it with state-of-the-art

<sup>1</sup>NVIDIA GV100 White Paper <http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf>

<sup>2</sup>Hardware specifications for the DLA available at <http://nvdla.org/>

algorithms through a set of synthetic experiments.

## II. SYSTEM MODEL

### A. Architecture model

A heterogeneous architecture is modeled as a set of *execution engines*  $\text{Arch} = \{e_1, e_2, \dots, e_m\}$ . An execution engine is characterized by 1) its execution capabilities, (i.e. its Instruction Set Architecture), specified by the engine's *tag*, and 2) its scheduling policy. An engine's tag  $\text{tag}(e_i)$  indicates the ability of a processor to execute a dedicated tasks.

As an example, a Xavier based platform such as the *NVIDIA pegasus*, can be modeled using 16 engines for a total of five different engine tags: 8 CPUs, 2 dGPUs, 2 iGPUs, 2 DLAs and 2 PVAs.

Tags express the heterogeneity of modern processor architecture: an engine tagged by dGPU (discrete GPU) or iGPU (integrated GPU) is designed to efficiently run generic GPU kernels, whereas engines with DLA tags are designed to run *deep learning inference* tasks. A deep learning task can be compiled to run on any engine, including CPUs and GPUs, however its worst-case execution time will probably be lower when running on DLAs.

In this paper, we allow the designer to compile the same task on different alternative engines with different tradeoffs in terms of performance and resource utilization, so to widen the space of possible solutions. As we will see in the next section, the HPC-DAG model supports *alternative* implementations of the same code. During the off-line analysis phase, only one of these alternative versions will be chosen depending on the overall schedulability of the system.

Communication is an important issue when considering the execution of real-time tasks on heterogeneous architectures. In modern GPUs, data transfers are performed by special engines called *copy engines*. A copy engine is a co-processor in charge of moving data between an address space visible to the CPU to an address space visible to the GPU. This translates in two physical separate memory devices in case of discrete GPUs, whereas for integrated GPUs system RAM is shared among both CPU cores and compute accelerators. We treat copy engines as processing units in which we *schedule* communication tasks.

Engines are further characterized by a scheduling policy (e.g. Fixed Priority or Earliest Deadline First), which can be *preemptive* or *non-preemptive*. In our model we allow different engines to support different scheduling policies: as we show in Section IV, in our methodology the schedulability analysis of each engine can be performed independently of the others. However, to simplify the presentation, in this paper we focus on *preemptive EDF*.

### B. The HPC-DAG task model

1) *Specification tasks*: A *specification task* is a Directed Acyclic Graph (DAG), characterized by a tuple  $\tau = \{T, D, \mathcal{V}, \mathcal{A}, \Gamma, \mathcal{E}\}$ , where:  $T$  is the period (minimum inter-arrival time);  $D$  is the relative deadline;  $\mathcal{V}$  is a set of graph nodes that represent *sub-tasks*;  $\mathcal{A}$  is a set of *alternative nodes*;

and  $\Gamma$  is a set of *conditional nodes*. The set of all the nodes is denoted by  $\mathcal{N} = \mathcal{V} \cup \mathcal{A} \cup \Gamma$ . The set  $\mathcal{E}$  is the set of edges of the graph  $\mathcal{E} : \mathcal{N} \times \mathcal{N}$ .

A sub-task  $v \in \mathcal{V}$  is the basic computation unit. It represents a block of code to be executed by one of the engines of the architecture. A sub-task is characterized by:

- A tag  $\text{tag}(v)$  which represents the ISA of the sub-task code. A sub-task can only be allocated onto an engine with the same tag;
- A worst-case execution time  $C(v)$  when executing the sub-task on the corresponding engine processor.

In this paper, we do not model the parallelization inside the GPU. Instead, we model a GPU sub-task as a single executing context able to potentially exploit all the parallel resources of the GPU's execution engine. This is compliant with what has been disclosed with regards to the NVIDIA GPU application scheduler [1], [2] that assumes only one GPU context being resident within the GPU at a given time. As an example, if the GPU node is an image processing workload, parallelization is exploited at the level of pixels of the image and not by processing multiple images at the same time instant.

A conditional node  $\gamma \in \Gamma$  represents alternative paths in the graph due to non-deterministic on-line conditions (e.g. if-then-else conditions). At run-time, only one of the outgoing edges of  $\gamma$  is executed, but it is not possible to know in advance which one.

An alternative node  $a \in \mathcal{A}$  represents alternative implementations of parts of the graph/task. During the configuration phase (which is detailed in Section V-A) our methodology selects one between many possible alternative implementations of the program by selecting only one of the outgoing edges of  $a$  and removing (part of) the paths starting from the other edges. This can be useful when modeling sub-tasks than can be executed on different engines with different execution costs.

An edge  $e(n_i, n_j) \in \mathcal{E}$  models a precedence constraint (and related communication) between node  $n_i$  and node  $n_j$ , where  $n_i$  and  $n_j$  can be sub-tasks, alternative nodes or conditional nodes.

The set of *immediate predecessors* of a node  $n_j$ , denoted by  $\text{pred}(n_j)$ , is the set of all nodes  $n_i$  such that there exists an edge  $(n_i, n_j)$ . The set of *predecessors* of a node  $n_j$  is the set of all nodes for which there exist a path toward  $n_j$ . If a node has no predecessor, it is a *source node* of the graph. In our model we allow a graph to have several source nodes. In the same way we define the set of *immediate successors* of node  $n_j$ , denoted by  $\text{succ}(n_j)$ , as the set of all nodes  $n_k$  such that there exists an edge  $(n_j, n_k)$ , and the set of *successors* of  $n_j$  as the set of nodes for which there is a path from  $n_j$ . If a node has no successors, it is a *sink node* of the graph, and we allow a graph to have several sink nodes.

Conditional nodes and alternative nodes always have at least 2 outgoing edges, so they cannot be sinks. To simplify the reasoning, we also assume that they always have at least one predecessor node, so they cannot be sources.

2) *Concrete tasks*: A concrete task  $\bar{\tau} = \{T, D, \bar{\mathcal{V}}, \bar{\Gamma}, \bar{\mathcal{E}}\}$  is an instance of a specification task where all alternatives

have been removed by making implementation choices. In the following, the *volume*  $\text{vol}(\bar{\tau})$  denotes the total cumulative WCET of the concrete task. It is computed in linear time in the number of conditional vertices [3].

Before explaining how to obtain a concrete task from a specification task, we present an example.

**Example 1.** Consider the task specification described in Figure 1a. Each sub-task node is labeled by the sub-task id and engine tag. Alternative nodes are denoted by square boxes and conditional nodes are denoted by diamond boxes. The black boxes denote corresponding junction nodes for alternatives and conditional, they are used to improve the readability of the figure but they are not part of the task specification<sup>3</sup>.

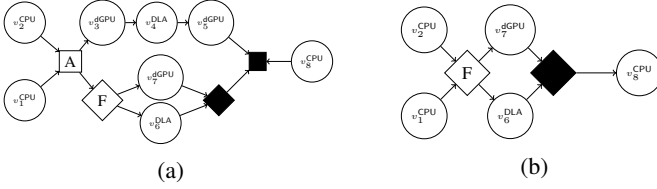


Fig. 1: Task specification and concrete tasks

Sub-tasks  $v_1^{\text{CPU}}$  and  $v_2^{\text{CPU}}$  are the sources (entry points) of the DAG.  $v_1^{\text{CPU}}$ ,  $v_2^{\text{CPU}}$  are marked by the **CPU** tag and can run concurrently: during the off-line analysis they may be allocated on the same or onto different engines. Sub-task  $v_4^{\text{DLA}}$  has an outgoing edge to  $v_5^{\text{dGPU}}$ , thus sub-task  $v_5^{\text{dGPU}}$  cannot start its execution before sub-task  $v_4^{\text{DLA}}$  has finished its execution. Sub-tasks  $v_1^{\text{CPU}}$  and  $v_2^{\text{CPU}}$  have each one outgoing edge to the alternative node A. Thus,  $\tau$  can execute either:

- 1) by following  $v_3^{\text{dGPU}}$  and then  $v_4^{\text{DLA}}$ ,  $v_5^{\text{dGPU}}$  and finishing its instance on  $v_8^{\text{CPU}}$ ;
- 2) or by following the conditional node F and select, according to an undetermined condition evaluated on-line, either to execute  $v_6^{\text{DLA}}$  or  $v_7^{\text{dGPU}}$ , and finishing its instance on  $v_8^{\text{CPU}}$ .

The two patterns are alternative ways to execute the same functionalities at different costs. Figure 1b represents one of the concrete tasks of  $\tau_i$ . During the analysis, alternative execution ( $v_3^{\text{dGPU}}$ ,  $v_4^{\text{DLA}}$ ,  $v_5^{\text{dGPU}}$ ) has been dropped.

In our model, data transfers between tasks can be modeled by special sub-tasks tagged with tag CP (*Copy-Engine*).

We consider a *sporadic task* model, therefore parameter T represents the minimum inter-arrival time between two instances of the same concrete task. When an instance of a task is activated at time  $t$ , all source sub-tasks are simultaneously activated. All subsequent sub-tasks are activated upon completion of their predecessors, and sink sub-tasks must all complete no later than time  $t + D$ . We assume *constrained deadline tasks*, that is  $D \leq T$ .

We now present a procedure to generate a concrete task  $\bar{\tau}$  from a specification task  $\tau$ , when all alternatives have been

chosen. The procedure starts by initializing  $\bar{\mathcal{V}} = \emptyset$ ,  $\bar{\Gamma} = \emptyset$ . First, all the source sub-tasks of  $\tau$  are added to  $\bar{\mathcal{V}}$ . Then, for every immediate successor node  $n_j$  of a node  $n_i \in \{\bar{\mathcal{V}} \cup \bar{\Gamma}\}$ : if  $n_j$  is a sub-task node (a conditional node, respectively), it is added to  $\bar{\mathcal{V}}$  (to  $\bar{\Gamma}$ , respectively); if it is an alternative node, we consider the selected immediate successor  $n_k$  of  $n_j$  and we add it to  $\bar{\mathcal{V}}$  or to  $\bar{\Gamma}$ , respectively. The procedure is iterated until all nodes of  $\tau$  have been visited. The set of edges  $\bar{\mathcal{E}} \subseteq \mathcal{E}$  is updated accordingly.

We denote by  $\Omega(\tau)$  the set of all concrete tasks of a specification task  $\tau$ .  $\Omega(\tau)$  is generated by simply enumerating all possible alternatives.

### III. AUTOWARE: A COMPLEX REAL-TIME FOR AUTONOMOUS DRIVING

In this section, we present how *Autoware*<sup>4</sup> [4], a complex real-time application, can be modeled using the HPC-DAG task model. Autoware is an open source autonomous driving framework which provides a set of software modules to perform sensing, computing and actuation for autonomous vehicles. Autoware processes different types of sensors data such as lidars, cameras, light detectors, etc., with the goal of creating a model of the surrounding environment that is then used to take actuation decisions. Autoware is compound of several packages ranging from vehicle localization to path planning and following.

The goal of the localization package is to provide a precise position of the autonomous vehicle. Therefore, it matches Lidar data with offline map data by means of *Normal Distributions Transform (NDT) algorithm*. Autoware provides two implementations of NDT, a *pure CPU* implementation and GPU-based implementation. These implementations are defined separately and the user can configure either CPU or GPU implementation. Therefore, only two alternative implementations are available, however more possibilities may be offered by allowing different implementations at function-call level. We have rewritten the NDT implementation to provide finer alternatives. As the task graph of the localization package is very large, for space constraints we only present the first part of the NDT in Figure 2.

The entry point of NDT is subtask `cpu_sync_CD_0`. Then, an online condition is tested to either continue the NDT processing or exit. If the condition succeeds, task will continue onward to execute `computePointsGradients`: this subtask can be executed on CPU engines by using 4 parallel threads in a fork-join fashion (right side of the node ALT17). Alternatively, it can also be executed on a GPU, therefore a memory copy operation (subtask `mem_cpy_1`) is executed before launching the `computePointsGradients_gpu` subtask on GPU. Further, the task may continue onward by testing an online condition for the execution of subtask `computeHessian`. This subtask has two implementations, on the CPU and on the GPU. If the CPU is selected for execution, then the task must also execute a memory copy (i.e. subtask `mem_cpy`). Alternatively,

<sup>3</sup>In fact, it is not always possible to insert junction nodes for an arbitrary specification.

<sup>4</sup><https://github.com/autowarefoundation/autoware>

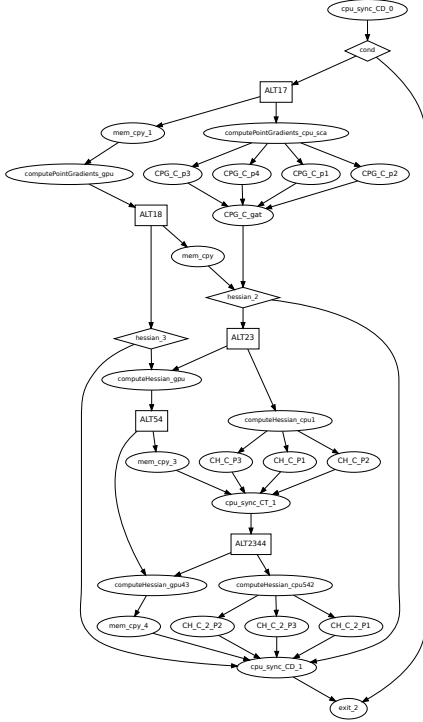


Fig. 2: The NDT algorithm of the Autware localization package modeled using HPC-DAG.

the task can continue its execution on a GPU by executing the `computeHessian_gpu` subtask, in this case no memory copy is needed since the data is already available on the GPU. To avoid cluttering the figure, we do not add an alternative node for each GPU call as each node tagged by GPU may be allocated on a discrete or an integrated GPU. We highlight that for the example presented in Figure 2, closing blocks can not be inserted because the graph is not properly nested.

	ComputePointsGradients	ComputeHessian
CPU	10.31	0.31
GPU	0.53	0.75
GPU-M	4.91	6.521
GPU-M-M	11.65	13.76

TABLE I: Execution times (milliseconds) of two sub-tasks on different engines.

Table I represents the measured execution times of the main functions represented in Figure 2 executed on an NVIDIA JETSON TX2. Table I contains the average values among 1000 executions using synthetic input data<sup>5</sup>. The CPU (resp. GPU) row represents the execution time on the CPU (resp. GPU). GPU-M includes in addition the memory copy execution time and GPU-M-M including both memory copy and `cudaMalloc` (memory allocation) execution times. When concrete tasks are selected, all memory allocation operations are completed before execution. Using the GPU

<sup>5</sup>The source code of the function is still the same as in the autware platform

can help to reduce sub-task execution time compared to CPU implementation, especially when no memory copy is needed for both `ComputePointsGradients` and `ComputeHessian`. For example when selecting the concrete task on ALT 18, we may select to execute `ComputeHessian` on GPU, therefore the data to process is already available to the GPU as `ComputePointsGradients`, its immediate predecessor, was executed also on the GPU. However, if we include memory allocation and copying, the CPU alternative costs less in terms of total execution time. In the considered computing platform, there are more CPUs available than GPUs. Therefore, the GPUs are powerful but *scarce resources*, and the sub-task allocation must be carefully planned to avoid surcharging the GPUs and miss some deadlines. Exploring the space of alternatives is not easy without an analysis to support it.

Similar scenarios can easily be found in several Autware packages and other complex applications where hardware acceleration is needed.

#### IV. SCHEDULING ANALYSIS

In this work, we consider partitioned scheduling. Each engine has its own scheduler and a separate ready-queue. Sub-tasks are allocated (partitioned) onto the available engines so that the system is schedulable. Partitioned scheduling allows to use well-known single processor schedulability tests, which make the analysis simpler and reduce the overhead due to thread migration compared to global scheduling. In this paper, we restrict to preemptive-EDF as it has recently been made available for CPUs and for GPUs [2]. However, the analysis presented here is modular, so engines may have different scheduling policies as described in Section IV-D.

##### A. Alternative patterns

Given a specification task  $\tau$ , we have to select one of the possible concrete tasks before proceeding to the allocation and scheduling of the sub-tasks on the computing engine. Since the number of combinations can be very large, in this paper we propose an heuristic algorithm based on a *greedy* strategy (see Section V). In particular, we explore the set of concrete tasks in a certain order. The order relation  $\succ$  sorts concrete tasks according to their total execution time.

**Definition 1.** Let  $\bar{\tau}, \bar{\tau}'$  be two concrete tasks of specification task  $\tau$ . The partial order relation  $\succ$  is defined as:

$$\bar{\tau} \succ \bar{\tau}' \implies \text{vol}(\bar{\tau}) \geq \text{vol}(\bar{\tau}') \quad (1)$$

In the next section, we will define a second order relationship  $\gg$  that sorts concrete tasks based on their engine tags.

When considering the example in Figure 2, the concrete task obtained by selecting the left side successor of each alternative node is on the top of the  $\succ$  relation order, therefore it is the preferred choice for allocation as it reduces the overall task utilization.

## B. Tagged Tasks

One concrete task may contain sub-tasks with different tags which will be allocated on different engines. Before proceeding to allocation, we need to select only sub-tasks pertaining to a given tag. We call this operation *task filtering*.

We start by defining an *empty sub-task* as a sub-task with null computation time.

**Definition 2** (Tagged task). Let  $\bar{\tau} = \{\mathcal{T}, \mathcal{D}, \bar{\mathcal{V}}, \bar{\Gamma}, \bar{\mathcal{E}}\}$  be a concrete task. Task  $\bar{\tau}(\text{tag}_i)$  is a tagged task of  $\bar{\tau}$  iff

- $\bar{\tau}(\text{tag}_i) = \{\mathcal{T}, \mathcal{D}, \mathcal{V}_i, \Gamma_i, \mathcal{E}_i\}$  is isomorphic to  $\bar{\tau}$ , that is the graph has the same structure, the same number of nodes of the same type, and the same edges between corresponding nodes;
- let  $v \in \bar{\mathcal{V}}$  be a sub-task of  $\bar{\tau}$ , and let  $v' \in \mathcal{V}_i$  be the corresponding sub-task of  $\bar{\tau}(\text{tag}_i)$  in the isomorphism. If  $\text{tag}(v) = \text{tag}_i$ , then  $C(v') = C(v)$ , else  $C(v') = 0$ ;
- $\Gamma_i = \bar{\Gamma}$ .

We denote with  $\mathcal{S}(\bar{\tau}) = \{\bar{\tau}(\text{tag}_1), \dots, \bar{\tau}(\text{tag}_K)\}$  the set of all possible tagged tasks of  $\bar{\tau}$ .

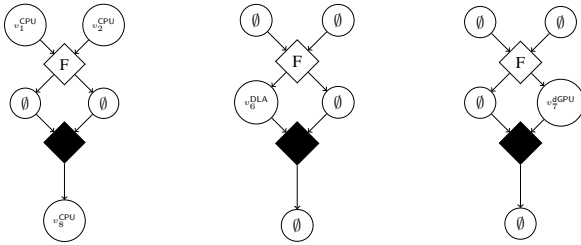


Fig. 3: Tagged tasks for the concrete task of Figure 1b.

Each concrete task generates as many *tagged tasks* as there are tags in the target architecture. Figure 3 shows the three tagged tasks for the concrete task in Figure 1b.

**Definition 3** ( $\gg$  order relationship). Assume the architecture supports  $K$  different tags. Let  $n(\text{tag})$  denote the number of computing engines labeled with  $\text{tag}$ . Assume that tags are ordered by increasing  $n(\text{tag})$ , that is  $n(\text{tag}_i) < n(\text{tag}_j) \implies i < j$ .

Let  $\bar{\tau}', \bar{\tau}''$  be two concrete tasks of specification task  $\tau$ , and let  $\mathcal{S}(\bar{\tau}') = \{\bar{\tau}'(\text{tag}_1), \dots, \bar{\tau}'(\text{tag}_K)\}$  and  $\mathcal{S}(\bar{\tau}'') = \{\bar{\tau}''(\text{tag}_1), \dots, \bar{\tau}''(\text{tag}_K)\}$  be the respective tagged tasks.

The order relation  $\bar{\tau}' \gg \bar{\tau}''$  is defined as follows:

$$\bar{\tau}' \gg \bar{\tau}'' \implies \exists 0 \leq i \leq K \begin{cases} \text{vol}(\bar{\tau}'(\text{tag}_j)) = \text{vol}(\bar{\tau}''(\text{tag}_j)) & \forall j < i \\ \text{vol}(\bar{\tau}'(\text{tag}_i)) < \text{vol}(\bar{\tau}''(\text{tag}_i)) \end{cases}$$

Relationship  $\gg$  gives priority to concrete tasks that allocate less load on scarce resources: if there are few execution engines with a certain tag, and there is a large number of sub-tasks requiring allocation on that specific engine, the relation order prefers alternative patterns with lower workload for those engines.

If we consider the example of Figure 2, the concrete task obtained by selecting the right side successor at each alternative node will be selected. This choice allows to relax the pressure on scarce resources as the number of GPUs is less than the number of CPUs. In this way, GPUs will be available for sub-tasks tagged with GPUs and for which no alternative is available.

## C. Deadlines and offsets assignment

Meeting timing constraints of a concrete task depends on the allocation of the sub-tasks onto the different execution engines. As these sub-tasks communicate through shared buffers, they are forced to respect the execution order dictated by the precedence constraints imposed by the graph structure.

To reduce the complexity of dealing with precedence constraints directly, we impose intermediate offsets and deadlines on each sub-task. In this way, precedence constraints are respected automatically if every sub-task is activated after its offset and completes no later than its deadline.

Many authors have proposed techniques to assign intermediate deadlines and offsets to task graphs. In this paper we use techniques similar to those proposed in [5] and [6].

Most of the deadline assignment techniques are based on the computation of the execution time of the critical path. A path  $P_x = \{v_1, v_2, \dots, v_l\}$  is a sequence of sub-tasks of task  $\bar{\tau}$  such that:

$$\forall v_l, v_{l+1} \in P_x, \exists e(v_l, v_{l+1}) \in \mathcal{E}.$$

Let  $\mathcal{P}$  denote the set of all possible paths of task  $\bar{\tau}$ . The critical path  $P_{crit}(\bar{\tau}) \in \mathcal{P}$  is defined as the path with the largest cumulative execution time of the sub-tasks.

We define the slack  $Sl(P, D)$  along path  $P$  as:

$$Sl(P, D) = D - \sum_{v \in P} C(v)$$

The assignment algorithm starts by assigning an intermediate relative deadline to every sub-task along a path by distributing the path's slack as follows:

$$D(v) = C(v) + \text{calculate\_share}(v, P)$$

The *calculate\_share* function computes the slack for sub-task  $v$  along the path. This slack can be shared according to two alternative heuristics:

- **Fair distribution:** assigns slack as the ratio of the original slack by the number of sub-tasks in the path:

$$\text{calculate\_share}(v, P) = \frac{Sl(P, D)}{|P|} \quad (2)$$

- **Proportional distribution:** assigns slack according to the contribution of the sub-task WCET in the path:

$$\text{calculate\_share}(v, P) = \frac{C(v)}{C(P)} \cdot Sl(P, D) \quad (3)$$

Once the relative deadlines of the sub-tasks along the critical path have been assigned, we select the next path in order of decreasing cumulative execution time, and assign the deadlines

to the remaining sub-task by appropriately subtracting the already assigned deadlines. The complete procedure has been described in [6], and due to space constraints we do not report it here.

Let  $O(v)$  be the offset of a subtask with respect of the arrival time of the task's instance. The sum of the offset and of the intermediate relative deadline of a subtask is called *local deadline*  $O(v) + D(v)$ , and it is the deadline relative to the arrival of the task's instance.

The offset of a subtask is set equal to 0 if the subtask has no predecessors; otherwise, it can be computed recursively as the maximum between the local deadlines of the predecessor sub-tasks.

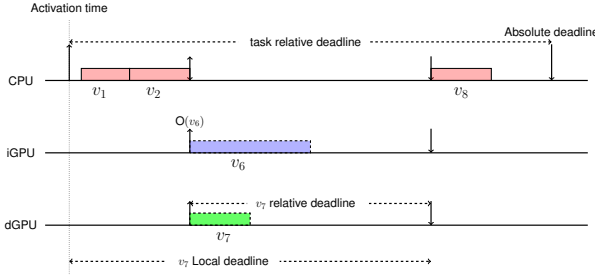


Fig. 4: Example of offset and local deadline

Figure 4 illustrates the relationship between the activation times, the intermediate offsets, relative deadlines and local deadlines of the sub-tasks of the concrete task of Figure 1b. We assume that  $v_1, v_2, v_8$  have been allocated on the same CPU whereas  $v_6$  and  $v_7$  each on a different engine. The activation time is the absolute time of the arrival of the sub-task instance. The activation time of a source sub-task corresponds to the activation time of the task graph. The offset is the interval between the activation of the task graph and the activation of the sub-task. The local deadline is the interval between the task graph activation and the sub-task absolute deadline.

**Definition 4.** Sub-task  $v \in \overline{\mathcal{V}_\tau}$  is feasible if for each task instance arrived at  $a_j$ , sub-task  $v$  executes within the interval bounded by its arrival time  $a(v) = a_j + O(v)$  and its absolute deadline  $a(v) + D(v)$ .

**Lemma 1.** A concrete task (resp. tagged task) is feasible if all its sub-tasks are feasible.

*Proof.* By the definition, the local deadline of the sink sub-tasks is equal to the deadline of the task  $D$ . Moreover, the offset of a sub-task is never before the local deadline of a preceding sub-task. Therefore 1) the precedence constraints are respected and 2) if sink sub-tasks are feasible then the concrete task (tagged task, respectively) is feasible.  $\square$

#### D. Single engine analysis

In this section, we assume that sub-tasks have been already been assigned offsets and deadlines, and they have been allocated on the platform's engines, and we present the

schedulability analysis to test if all tasks respect their deadlines when scheduled by the Earliest Deadline First (EDF) algorithm.

**Theorem 1.** Let  $\mathcal{T}$  a set of task graphs allocated onto a single-core engine. Task set  $\mathcal{T}$  is schedulable by EDF if and only if:

$$\sum_{\bar{\tau} \in \mathcal{T}} \text{dbf}(\bar{\tau}, t) \leq t, \forall t \leq t^* \quad (4)$$

The dbf is the demand bound function [7] for a task graph  $\bar{\tau}$  in interval  $t$ . The demand bound function is computed as the worst-case cumulative execution time of all jobs (instances of sub-tasks) having their arrival time and deadline within any interval of time of length  $t$ . For a task graph, the dbf can be computed as follows:

$$\text{dbf}(\tau, t) = \max_{v \in \tau} \sum_{v' \in \tau} \left\lfloor \frac{t - \tilde{O}(v') - D(v') + T(\tau)}{T(\tau)} \right\rfloor C(v') \quad (5)$$

where<sup>6</sup>:  $\tilde{O}(v') = (O(v') - O(v)) \bmod T(\tau)$

In our model, a task graph may contain *conditional nodes*, which model alternative paths that are selected non-deterministically at run-time. To compute the dbf for a tagged task that contains conditional nodes, we must first enumerate all possible conditional graphs by using the same procedure as the one used for generating concrete tasks from specification tasks. Hence, the dbf of a tagged task in interval  $t$  can be computed as the largest dbf among all the possible conditional graphs.

#### E. Anticipating the activation of sub-tasks

Given an instance of sub-task  $v$  with arrival at  $a(v)$  and local deadline at  $D(v)$ , at run-time it may happen that all instances of the preceding sub-tasks have already completed their execution before  $a(v)$ . In this case, we activate the sub-task as soon as the preceding sub-tasks have finished with the same local deadline  $D(v)$ .

**Lemma 2.** Consider a feasible set of sub-tasks allocated on a set of engines and scheduled by EDF. If a sub-task is activated as soon as all predecessor sub-tasks have finished, with the same local deadline, the set remains schedulable.

*Proof.* Descends directly from the sustainability property of EDF [8]. In fact, by anticipating the activation of the sub-task without modifying its local deadline, the sub-task will be scheduled with a longer relative deadline, and the demand bound function will not increase.  $\square$

From an implementation point of view, this technique avoids the need to set-up activation timers for intermediate tasks; moreover, it allows us to reduce the pessimism of the analysis in the presence of high preemption costs, as we will see in the next section.

<sup>6</sup>We remind that the remainder of  $a/b$  is by definition a positive number  $r$  such that  $a = kb + r$ .

### F. Preemption-aware analysis

In recent GPUs, preempting an executing task can be a costly operation (see Section VII-C). In particular, the cost of preemption may significantly vary depending on the preempted task and the engine. For example, preempting a graphical kernel induces a larger cost compared to preempting a computing CUDA kernel. Therefore, we need to account for the cost of preemption in the analysis. From now on, a sub-task will be also characterized by  $pc(v)$ , the timing cost of preempting the sub-task  $v$ .

A simple (although pessimistic) approach is to always consider the worst-case preemption cost as part of the worst-case computation time of the preempting task. Let  $pc(v_j)$  denote the cost of preempting sub-task  $v_j$ .

**Lemma 3.** *Let  $\mathcal{V} = \{v_1, v_2, \dots, v_K\}$  be a set of sub-tasks to be scheduled by EDF (resp. fixed priority FP) on a single engine.*

*Consider  $\mathcal{V}^{pc} = \{v'_1, v'_2, \dots, v'_K\}$ , where  $v'_i$  has the same parameters as  $v_i$ , except for the wcet that is computed as  $C(v'_i) = C(v_i) + pc^i$  and  $pc^i = \max\{pc(v) | v \in \mathcal{V} \wedge D(v) > D(v_i)\}$  (respectively  $P(v) > P(v_i)$  resp. for FP).*

*If  $\mathcal{V}^{pc}$  is schedulable by EDF (respectively FP) when considering a null preemption cost, then  $\mathcal{V}$  is schedulable when considering the cost of preemption.*

*Proof.* The Lemma directly follows from the simple observation that the cost of preemption can never exceed  $pc^i$  for sub-task  $v_i$ .  $\square$

Lemma 3 is safe but pessimistic. We can further improve the analysis by observing that a sub-task cannot preempt another sub-task belonging to the same task graph (we remind the reader that we assume constrained deadline tasks). Furthermore, it may be impossible for two consecutive sub-task of a task graph to both preempt the same sub-task as demonstrated by Theorem 2.

We will now further reduce the preemption cost estimation by considering only EDF and DM. Let start by observing that, in the case of EDF and deadline monotonic DM scheduling, a job of a considering only EDF and DM<sup>7</sup>. Let us start by observing that, in the case of EDF and DM scheduling, a job of a sub-task  $v_i$  can preempt a job of sub-task  $v_j$  at most once, and only if its relative deadline is shorter:  $D(v_i) < D(v_j)$ .

**Definition 5** (Maximal sequential subset). *A maximal sequential subset  $\mathcal{V}^M$  of task  $\tau$  is a maximal subset of  $\mathcal{V}_\tau$  such that:*

- 1) *it is weakly-connected;*
- 2)  *$\forall v \in \mathcal{V}^M, v' \in \text{pred}(v)$  is either null and does not belong to  $\mathcal{V}^M$ , or non null and belongs to  $\mathcal{V}^M$ .*

*We denote by  $\text{cand}(\mathcal{V}^M)$  the set of all sub-tasks in  $\mathcal{V}^M$  that are either sources, or have a null predecessor. Further, we denote by  $v^M$  the sub-task with the shortest local deadline in  $\text{cand}(\mathcal{V}^M)$*

We observe that, since all the sub-tasks in  $\mathcal{V}^M$  are allocated on the same engine and since they do not have any predecessor sub-task allocated on a different engine (no empty predecessor), they can be activated as soon as the predecessor sub-tasks have finished.

Now, suppose  $v_1, v_2 \in \mathcal{V}^M$  and that  $v_1$  is an immediate predecessor of  $v_2$ . If  $v_1$  preempts a sub-task  $v_j$ , and  $D(v_2) \leq D(v_j)$ , then  $v_j$  can be executed only after  $v_2$  has finished. This means that the cost of preempting  $v_j$  can be accounted to only  $v_1$ . We assign this preemption cost to the sub-task  $v^M$  with the shorter local deadline among all sub-tasks not having a predecessor in  $\mathcal{V}^M$ , whereas the others do not pay any preemption cost. The preemption cost of any other sub-task in  $\mathcal{V}'$  is set equal to 0. For all sub-tasks that have a null predecessor, we compute a preemption cost as in Lemma 3.

**Theorem 2** (Limited preemption cost). *Let  $\mathcal{V} = \{v_1, v_2, \dots, v_K\}$  be a set of sub-tasks scheduled by EDF/DM on a single processor. Consider  $\mathcal{V}^{pc} = \{v'_1, v'_2, \dots, v'_K\}$  where  $v'_i$  has the same parameters as  $v_i$ , except for the wcet that is computed as  $C(v'_i) = C(v_i) + pc^i$ , and  $pc^i$  is computed as in Equation (6) or (7).*

- *If  $v_i = v^M$ , then*

$$pc^i = \max\{pc(v) | v \in \mathcal{V} \setminus \mathcal{V}_\tau \wedge D(v) > D(v_i)\}; \quad (6)$$

*where  $\mathcal{V}_\tau$  is the set of sub-tasks of task  $\tau$  where  $v_i$  belongs.*

- *otherwise,*

$$pc^i = 0 \quad (7)$$

*If  $\mathcal{V}^{pc}$  is schedulable by EDF when considering a null preemption cost, then  $\mathcal{V}$  is schedulable when considering the cost of preemption.*

*Proof.* For space constraints, we report here a proof sketch.

Consider any sub-task  $v_i \in \mathcal{V}^M$  not belonging to  $\text{cand}(\mathcal{V}^M)$ : it is activated as soon as the preceding sub-tasks have finished executing their corresponding instances. Then, if one of the preceding task of  $v_i$  preempted a sub-task  $v_j$ , the preemption cost has already been accounted in the worst-case execution time of the preceding task; as discussed above  $v_j$  can only resume execution after  $v_i$  has completed. Thus, no further preemption cost need to be accounted.

If instead none of the preceding sub-task of  $v_i$  has preempted  $v_j$ , then  $v_j$  cannot start executing before  $v_i$  completes because its deadline is not smaller than  $D(v_i)$ , hence no preemption will occur.

In any case, no cost of preemption needs to be accounted for to  $v_i$ .

Similarly, sub-tasks belonging to  $\text{cand}(\mathcal{V}^M)$  and different than  $v^M$  are not subject to preemptions.  $\square$

In this paper, we restrict to EDF. Therefore, the preemption cost is computed using Theorem 2. For other scheduling policies the estimation of Lemma 3 can be used.

<sup>7</sup>Deadline Monotonic

## V. ALLOCATION

### A. Allocation of task specifications

The goal of our methodology is to allocate a set of task specifications into a set of engines, by selecting alternative implementations, so that all tasks complete before their deadlines. From an operational point of view, this is equivalent to finding a feasible solution to a complex Integer Linear Programming problem. In facts, given the large number of combinations (due to alternative nodes, condition-control nodes, and allocation decisions), an ILP formulation of this problem fails to produce feasible solutions in an acceptable time. Therefore, in this section we propose a set of greedy heuristics to quickly explore the space of solutions.

Algorithm 1 describes the basic approach. The algorithm can be customised with four parameters: order is the sorting order of the concrete task sets (see Sections IV-A and IV-B); parameter slack concerns the way the slack is distributed when assigning intermediate deadlines and offsets (see Section IV-C); parameter alloc can be best-fit (BF) or worst-fit (WF); parameter omit concerns the strategy to eliminate sub-tasks when possible (see Section V-C).

At each step, the algorithm tries to allocate one single task specification (the for loop at line 5). For each task, it first generates all concrete tasks (line 6), and sorts them according to one relationship order ( $\succ$  or  $\gg$ ). Then, for each concrete task, it first assigns the intermediate deadlines and offsets according to the methodology described in Section IV-C (line 9), using one between the fair or the proportional slack distributions. Then, it separates the concrete tasks into tagged tasks according to the corresponding tags (line 10).

Then, the algorithm tries to allocate every tagged task onto single engines having the corresponding tag (line 14) (this procedure is described below in Algorithm 2). If a feasible allocation is found, the allocation is generated, and the algorithm goes to the next specification task (lines 15-16). If no feasible sequential allocation can be found, the next concrete task is tested.

The algorithm gives priority to single-engine allocations because they reduce preemption cost, as discussed in Section IV-F. In particular, by allocating an entire tagged task onto a single engine, we reduce the number of null sub-task to the minimum necessary, and so we can assign the cost of preemption to fewer sub-tasks.

If none of the concrete tasks of a specification task can be allocated (line 17), this means that one of the tagged tasks could not be allocated on a single engine. Therefore, the algorithms tries to break some of the tagged tasks of a concrete task into parallel tasks to be executed on different engines of the same type. This is done by procedure `parallelize`, which will be described in Section V-C. In particular, one part of the concrete task will be allocated, while the second part will be put back in the list of not-yet-allocated task graphs (line 24). If this process is also unable to find a feasible concrete task, the analysis fails (line 29).

---

### Algorithm 1 Allocation algorithm

---

```

1: input :  $\mathcal{T}$ : set of task specifications
2: parameters : order ( $\succ$  or  $\gg$ ), slack (fair or proportional),
3:               alloc (BF or WF), omit (parallel or random)
4: output : SUCCESS or FAIL
5: for  $\tau \in \mathcal{T}$  do
6:    $\Omega = \text{generate\_concrete\_task}(\tau)$ 
7:    $\text{sort}(\Omega, \text{order})$ 
8:   for  $(\bar{\tau} \in \Omega)$  do
9:      $\text{assign\_deadlines\_offsets}(\bar{\tau}, \text{slack})$ 
10:     $\mathcal{S}(\bar{\tau}) = \text{filter\_tagged\_task}(\bar{\tau})$ 
11:   end for
12:   allocated = false
13:   for  $(\bar{\tau} \in \Omega)$  do
14:     if ( $\text{feasible\_sequential}(\mathcal{S}(\bar{\tau}), \text{alloc})$ ) then
15:       allocated = true; assign sub-tasks to engines
16:       break;
17:     end if
18:   end for
19:   if (not allocated) then
20:     for  $(\bar{\tau} \in \Omega)$  do
21:        $(\bar{\tau}', \bar{\tau}'') = \text{parallelize}(\bar{\tau}, \text{alloc}, \text{omit})$ 
22:       if  $(\bar{\tau}' \neq \emptyset)$  then
23:         allocate  $\bar{\tau}'$  to selected engines
24:         add back  $\bar{\tau}''$  to  $\mathcal{T}$ 
25:         allocated = true
26:         break
27:       end if
28:     end for
29:     if (not allocated) then return FAIL
30:   end if
31: end for
32: return SUCCESS

```

---

### B. Sequential allocation

Algorithm 2 tries to allocate a concrete task on a minimal number of engines. It takes as input a set of tagged tasks. For each tagged task, it selects the corresponding engines, and sorts them according to the alloc parameter, that is in decreasing order of utilization in the case of Best-Fit, or in increasing order of utilization in case of Worst-Fit. Then, it tests the feasibility of allocating the tagged task on each engine in turn. If the allocation is successful, the next tagged task is tested, otherwise the algorithm tries the next engine. If the tagged task cannot be allocated on any engine, the algorithm fails. If all tagged tasks have been allocated, the corresponding allocation is returned.

### C. Parallel allocation

When the sequential allocation fails for a given task specification, the algorithm tries to allocate one or more of its tagged tasks onto multiple engines having the same tag. Algorithm 3 takes as input a concrete task and two parameters, alloc for BF or WF heuristics, and omit to select which sub-task to remove first.



---

**Algorithm 2** feasible\_sequential

---

```
1: input:  $\mathcal{S}(\bar{\tau})$ : set of tagged tasks, alloc
2: output: feasibility: SUCCESS or FAIL
3: for ( $\bar{\tau}(\text{tag}) \in \mathcal{S}(\bar{\tau})$ ) do
4:   engine_list=select_engine(tag)
5:   sort_engines(engine_list, alloc)
6:    $f = \text{false}$ 
7:   nfeas = 0
8:   for ( $e \in \text{engine\_list}$ ) do
9:      $f = \text{dbf\_test}(\bar{\tau} \cup \mathcal{T}_e)$ 
10:    if ( $f$ ) then
11:      save_allocation( $\bar{\tau}$ ,  $e$ )
12:      nfeas ++
13:      break
14:    end if
15:  end for
16:  if (not  $f$ ) then return FAIL;
17: end for
18: if ( $\text{nfeas} = |\mathcal{S}(\bar{\tau})|$ ) then
19:   return SUCCESS, saved_allocations
20: end if
```

---

For each tagged task of the concrete task (line 5), the algorithm selects the list of engines corresponding to the selected tag, and sorts them according to BF or WF (line 7). Then, it tries to test the feasibility of the tagged task on each engine (line 9). If the test fails, it removes one sub-task from the tagged task and adds it to list of non allocated sub-tasks  $\bar{\tau}''$  (line 11). We propose two heuristics:

- 1) **Random** heuristic: it selects a random sub-task and adds it to the omitted list.
- 2) **Parallel** heuristic: to be feasible, the critical path of each tagged task must be feasible even on a unlimited number of engines. Thus, we are interested in sub-tasks that do not belong to the critical path because they are the ones causing the non-feasibility. Thus, they are omitted one by one until finding a feasible schedule.

The feasibility test is repeated until a feasible subset of  $\bar{\tau}(\text{tag})$  is found. The omitted tasks are tried on the next engine with the same tag (line 16). At the end of the procedure, two concrete tasks are produced,  $\bar{\tau}'$  is the feasible part that will be allocated, while  $\bar{\tau}''$  will be tried again in the following iteration of Algorithm 1.

## VI. RELATED WORK

Many authors [5], [9]–[16] have proposed real-time task models based on DAGs. However, to the best of our knowledge, none of the existing models supports alternative implementations of the same functionality on different computing engines.

Authors of [5] studied the deadline assignment problem in distributed real-time systems. They formalize the problem and identify the cases where deadline assignment methods have a strong impact on system performances. They propose Fair Laxity Distribution (FLD) and Unfair Laxity Distribution (ULD) and study their impact on the schedulability. In [12],

---

**Algorithm 3** parallelize

---

```
1: input:  $\bar{\tau}$ : concrete task, alloc (BF or WF),
2:   omit (parallel or random)
3: output: concrete tasks ( $\bar{\tau}'$ ,  $\bar{\tau}''$ )
4:  $\bar{\tau}' = \emptyset, \bar{\tau}'' = \emptyset$ 
5: for ( $\bar{\tau}(\text{tag}) \in \mathcal{S}(\bar{\tau})$ ) do
6:   engine_list=select_engines(tag)
7:   sort(engine_list, alloc)
8:   for ( $e \in \text{engine\_list}$ ) do
9:      $f = \text{dbf\_test}(\bar{\tau}(\text{tag}) \cup \mathcal{T}_e)$ 
10:    while (not  $f$ ) do
11:       $\bar{\tau}'' = \bar{\tau}'' \cup \text{remove}(\bar{\tau}(\text{tag}), \text{omit})$ 
12:       $f = \text{dbf\_test}(\bar{\tau}(\text{tag}) \cup \mathcal{T}_e)$ 
13:    end while
14:    if ( $\bar{\tau}(\text{tag}) \neq \emptyset$ ) then
15:       $\bar{\tau}' = \bar{\tau}' \cup \text{save\_allocation}(\bar{\tau}(\text{tag}), e)$ 
16:       $\bar{\tau}(\text{tag}) = \bar{\tau}'', \bar{\tau}'' = \emptyset$ 
17:      allocated = true
18:    break
19:  end if
20: end for
21: if (not allocated) return  $\emptyset, \bar{\tau}$ 
22: end for
23: return  $\bar{\tau}', \bar{\tau}''$ 
```

---

authors analyze the schedulability of a set of DAGs using global EDF, global rate-monotonic (RM), and federated scheduling. In [17], the authors present a general framework for partitioning real-time tasks onto multiple cores using resource reservations. They propose techniques to set activation times and deadlines of each task, and they use ILP formulation to solve the allocation and assignment problems. However, when applying such approaches on large applications consisting of hundred of sub-tasks, the analysis can be very time consuming.

DAG fixed-priority partitioned scheduling has been presented in [14]. The authors propose methods to compute a response time with tight bounds. They present partitioned DAGs as a set of self-suspending tasks, and proposed an algorithm to traverse a DAG and characterize the worst-case scheduling scenario.

Unlike previous models, Melani et al [10] proposed to model conditional branches in the code in a way similar to our conditional nodes, however their model is not able to express off-line alternative patterns. They proposed different methods to compute an upper-bound on the response-time under global scheduling algorithms. In [18], alternative on-line execution patterns can be expressed using *digraphs*. However, the digraph model cannot express parallelism and only supports sequential tasks.

In this paper we assume preemptive EDF scheduling. Typically, preemption on classical CPUs is assumed to be a negligible percentage of the task execution. However, this is not always the case with GPUs processors. Depending on the computing architecture and on the nature of the workload, GPU tasks present different degrees of preemption granularity

and related preemption costs. Initial work on preemptive scheduling on GPUs assumed preemption was viable at the *kernel* granularity [19]. A finer granularity for computing workloads is represented by CTA (Cooperative Thread Array) level preemption, hence, preemption occurs at the boundaries of group of parallel threads that execute within the same GPU computing cluster [20], [21]. In such a scenario, the cost of preempting an executing context on a GPU might present significant differences as it will involve saving and restoring contexts of variable size and/or reaching the next viable preemption point. Overhead measurements operated in the cited contributions calls for modeling each GPU sub-task with a specific non-negligible preemption cost that can be in the same order of magnitude of the execution time of the sub-task.

## VII. RESULTS AND DISCUSSIONS

In this section, we evaluate the performance of our scheduling analysis and allocation strategies. We compare against the model cp-DAG proposed by Melani et al. [10]. Please notice that in [10] the authors proposed an analysis for cp-DAGs in the context of global scheduling, whereas our analysis is based on partitioned scheduling. Therefore, we extended the cp-DAG model to support multiple engines. Moreover we applied the same allocation heuristics of Section V and the same scheduling analysis of Section IV to HPC-DAGs and to cp-DAG.

In the following experiments, we considered the NVIDIA Jetson AGX Xavier platform<sup>8</sup>. It features 8 CPU cores and four different kinds of accelerators: one discrete and one integrated GPU, one DLA and one PVA. Each accelerator is treated as a single computing resource. In this way, we can exploit task level parallelism as opposed to allowing the parallel execution of more than one sub-task to partitions of the accelerators (e.g. at a given time instant, only one sub-task is allowed to execute in all the computing clusters of a GPU).

### A. Task Generation

We apply our heuristics on a large number of randomly generated synthetic task sets.

The task set generation process takes as input an engine/tag utilization for each tag on the platform. First, we start by generating the utilization of the  $n$  tasks by using the UUniFast-Discard [22] algorithm for each input utilization. Graph sub-tasks can be executed in parallel, thus task utilization can be greater than 1. The sum of every per-tag utilization is a fixed number upper bounded by the number of engines per tag.

The number of nodes of every task is chosen randomly between 10 and 30. We define a probability  $p$  that expresses the chance to have an edge between two nodes, and we generate the edges according to this probability. We ensure that the graph depth is bounded by an integer  $d$  proportional to the number of sub-tasks in the task. We also ensure that the graph is *weakly connected* (i.e. the corresponding undirected graph is connected); if necessary, we add edges between non-connected portions of the graph. Given a sub-task node, one

of its successors is an alternative node or a conditional node with probability of 0.7.

To avoid untractable hyper-periods, the period of every task is generated randomly according to values taken from a list, where the minimum is 120 and the maximum is 120,000. For every sub-task, we randomly select a tag. Further, for each tag, we use algorithm UUNIFAST discard again to generate single sub-task utilization. Thus, the sub-task utilization can never exceed 1. Further, we multiply the utilization of each sub-task by the task period to generate the worst case execution time of every vertex.

A cp-DAG is generated from a HPC-DAG by selecting one of the possible concrete tasks at random.

### B. Simulation results and discussions

We varied the baseline utilization from 0 to the number of engines per engine tag in 16 steps. Therefore, the step size vary from one engine tag to the other: the step size is 0.5 for CPUs, and 0.0625 for the others. For each utilization, we generated a random number of tasks between 20 and 25.

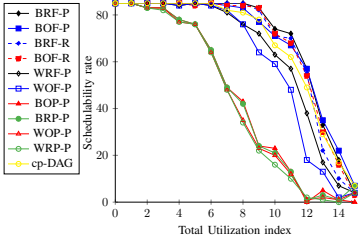
The results are presented as follows. Each algorithm is described using 3 letters: (i) the first letter is either B for best fit or W for worst first allocation techniques; (ii) the second is either  $O$  for the  $\succ$  order relation, or  $R$  for the  $\gg$  order relation; (iii) the third character describes the deadline assignment heuristic,  $F$  for fair and  $P$  for proportional. The algorithm name may also contain either option  $P$  for the parallel allocation heuristic that eliminates parallel nodes first, or  $R$  the random heuristic which randomly selects the sub-task to remove. For Figures 5a, 5b, 5c, 6, we run 85 simulations per utilization step.

Figure 5a represents the schedulability rate of each combination of heuristics cited above as a function of the total utilization. The fair deadline assignment technique presents better schedulability rates compared to proportional deadline assignment. In general, BF heuristic combinations outperform WF heuristic: this can be explained by observing that BF tries to pack the maximum number of sub-tasks into the minimum number of engines, and this allows for more flexibility to schedule *heavy* tasks on other engines.

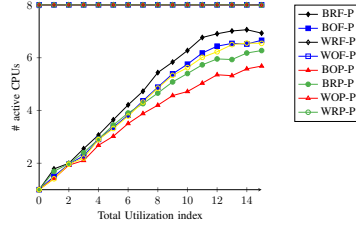
In the figures, the cp-DAG model proposed in [10] is shown in yellow. Since the cp-DAG has no alternative implementations, the algorithm has less flexibility in allocating the sub-tasks, therefore *by construction* the results for HPC-DAG dominate the corresponding results for cp-DAG. However, it is interesting to measure the difference between the two models: for example in Figure 5a the difference in the schedulability rate between the two models is between 10% and 20% for utilization rates between 6 and 14.

When the system load is low, all combinations of heuristics allow having high schedulability rates. BRF shows better results because it is aimed at relaxing the utilization of scarce engines, thus avoiding the unfeasibility of certain task sets due to a high load on a scarce engines (DLA and PVA/ GPUs). However, when dealing with a highly loaded system, BOF presents better

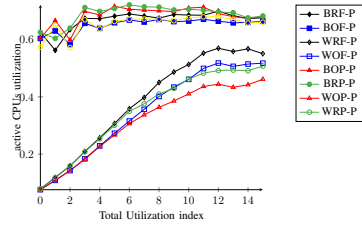
<sup>8</sup> [https://elinux.org/Jetson\\_AGX\\_Xavier](https://elinux.org/Jetson_AGX_Xavier)



(a) Schedulability rate VS utilization.



(b) #Active CPUs vs utilization.



(c) Active CPU utilization VS utilization

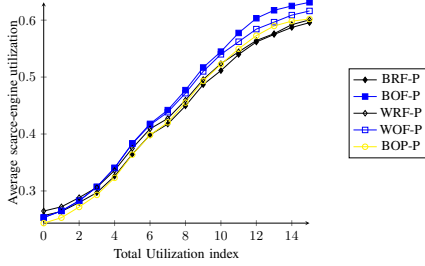


Fig. 6: DLA, GPUs, PVA utilizations vs total utilization.

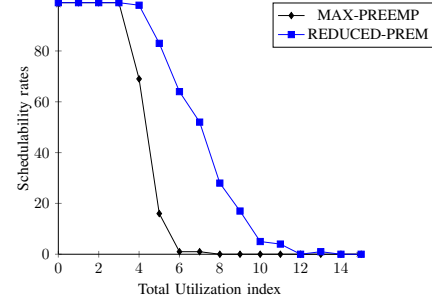


Fig. 7: Preemption cost Theorem vs max

schedulability rates, as it reduces the execution overheads on all engines.

Figure 5b reports the average number of active cores (CPUs) as a function of the total utilization. WF-based heuristics always use the highest number of CPU cores because our task generator outputs at least 15 CPU subtasks. Hence, the number of tasks is larger than the available number of CPU cores (which is 8, in our test platform). BF heuristics allows to pack the maximum number of sub-tasks on the minimum number of engines, thus the utilization increases quasi-linearly. This occurs until the maximum schedulability limit is reached (i.e. number of cores). BRF heuristic uses more CPU cores because it *preserves* the scarce resources, thus it uses more CPU engines. As BOF privileges reducing the overall load, it reduces the load on the CPUs compared to BRF.

Figure 5c shows the average active utilization for CPUs. Average utilization of BF-based heuristics is higher compared to WF. In fact, the latter distributes the work on different engines thus the per-core utilization is low in contrast to BF. Again, BRF has higher utilization than BOF because it schedules more workload on CPU cores than the other heuristics. As the workload is equally distributed on different CPUs, the WF heuristics may be used to reduce the CPUs operating frequency to save dynamic energy. Regarding BF heuristics, we see that BRF is not on the top of the average load because it uses more cores than the others.

Figure 6 shows the average utilization of the scarce resources. As you may notice, order relation  $\gg$  based heuristics allows to reduce the load on the scarce resources compared to  $\succ$ . In fact, the higher is the load, the less loaded are the scarce resources.

### C. Preemption cost simulation

In all previous experiments, we applied the analysis described in Section IV-F to account for preemption costs. In particular,

we applied the technique of Theorem 2, by assuming that the cost of preempting a sub-task is 30% of the sub-task execution time on a GPU, 10% on DLA and PVA, and 0.02% on the CPUs. DLA and PVA are non-preemptable engines, however longer jobs might be split into smaller chunks and this translates in a splitting overhead as we submit many kernel calls as opposed of a single batch of commands.

To highlight the importance of a proper analysis of the cost of preemption, in Figure 7 we report the schedulability rates obtained by BRF-P in two different cases: when considering the analysis of Lemma 3 (where the maximum preemption cost is charged to all preempting sub-tasks) and that of Theorem 2, where the cost is only charged to one of the sub-tasks in the maximal sequential subset.

As the utilization increases, schedulability drastically falls for the first method, while the improved analysis of Theorem 2 keeps tight schedulability rates.

## VIII. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the HPC-DAG real-time task model, which allows to specify both off-line and on-line alternatives, to fully exploit the heterogeneity of complex embedded platforms. We also presented a scheduling analysis and a set of heuristics to allocate HPC-DAGs on heterogeneous computing platforms. The analysis takes into account the cost of preemption that may be non-negligible in certain specialized engines. As for future work, we are considering extending our framework to account for memory interference between the different compute engines, as it is known to cause significant variations in execution times [23], [24].

## REFERENCES

- [1] N. Capodieci, R. Cavicchioli, and M. Bertogna, "Work-in-progress: Nvidia gpu scheduling details in virtualized environments," in *2018 International Conference on Embedded Software (EMSOFT)*. IEEE, 2018, pp. 1–3.
- [2] N. Capodieci, R. Cavicchioli, M. Bertogna, and A. Paramakuru, "Deadline-based scheduling for gpu with preemption support," in *2018 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, 2018, pp. 119–130.
- [3] S. Baruah, "The federated scheduling of systems of conditional sporadic dag tasks," in *Proceedings of the 12th International Conference on Embedded Software*. IEEE Press, 2015, pp. 1–10.
- [4] S. Kato, E. Takeuchi, Y. Ishiguro, Y. Ninomiya, K. Takeda, and T. Hamada, "An open approach to autonomous vehicles," *IEEE Micro*, vol. 35, no. 6, pp. 60–68, 2015.
- [5] D. Marinca, P. Minet, and L. George, "Analysis of deadline assignment methods in distributed real-time systems," *Comput. Commun.*, vol. 27, no. 15, pp. 1412–1423, Sep. 2004. [Online]. Available: <http://dx.doi.org/10.1016/j.comcom.2004.05.006>
- [6] Y. Wu, Z. Gao, and G. Dai, "Deadline and activation time assignment for partitioned real-time application on multiprocessor reservations," *Journal of Systems Architecture*, vol. 60, no. 3, pp. 247 – 257, 2014, real-Time Embedded Software for Multi-Core Platforms. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S138376211300266X>
- [7] S. K. Baruah, L. E. Rosier, and R. R. Howell, "Algorithms and complexity concerning the preemptive scheduling of periodic, real-time tasks on one processor," *Real-Time Systems*, vol. 2, no. 4, 1990.
- [8] A. Burns and S. Baruah, "Sustainability in real-time scheduling," *Journal of Computing Science and Engineering*, vol. 2, no. 1, pp. 74–97, 2008.
- [9] M. Qamhieh, F. Fauberteau, L. George, and S. Midonnet, "Global edf scheduling of directed acyclic graphs on multiprocessor systems," in *Proceedings of the 21st International conference on Real-Time Networks and Systems*. ACM, 2013, pp. 287–296.
- [10] A. Melani, M. Bertogna, V. Bonifaci, A. Marchetti-Spaccamela, and G. C. Buttazzo, "Schedulability analysis of conditional parallel task graphs in multicore systems," *IEEE Trans. Computers*, vol. 66, no. 2, pp. 339–353, 2017.
- [11] A. Saifullah, K. Agrawal, C. Lu, and C. Gill, "Multi-core Real-Time Scheduling for Generalized Parallel Task Models," pp. 217–226, Nov. 2011.
- [12] J. Li, J. J. Chen, K. Agrawal, C. Lu, C. Gill, and A. Saifullah, "Analysis of federated and global scheduling for parallel real-time tasks," in *Real-Time Systems (ECRTS), 2014 26th Euromicro Conference on*. IEEE, 2014, pp. 85–96.
- [13] A. Saifullah, D. Ferry, C. Lu, and C. Gill, "Real-time scheduling of parallel tasks under a general dag model," 2012.
- [14] J. Fonseca, G. Nelissen, V. Nélis, and L. M. Pinho, "Response time analysis of sporadic dag tasks under partitioned scheduling," in *Industrial Embedded Systems (SIES), 2016 11th IEEE Symposium on*. IEEE, 2016, pp. 1–10.
- [15] H.-E. Zahaf, A. E. H. Benyamina, R. Olejnik, and G. Lipari, "Energy-efficient scheduling for moldable real-time tasks on heterogeneous computing platforms," *Journal of Systems Architecture*, vol. 74, pp. 46 – 60, 2017.
- [16] —, "Modeling parallel real-time tasks with di-graphs," in *Proceedings of the 24th International Conference on Real-Time Networks and Systems*, ser. RTNS '16. ACM, 2016, pp. 339–348.
- [17] Y. Wu, Z. Gao, and G. Dai, "Deadline and activation time assignment for partitioned real-time application on multiprocessor reservations," *Journal of Systems Architecture*, vol. 60, no. 3, pp. 247–257, 2014.
- [18] M. Stigge, P. Ekberg, N. Guan, and W. Yi, "The digraph real-time task model," in *Real-Time and Embedded Technology and Applications Symposium*, April 2011.
- [19] J. Zhong and B. He, "Kernelet: High-throughput gpu kernel executions with dynamic slicing and scheduling," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 6, pp. 1522–1532, 2014.
- [20] T. Amert, N. Otterness, M. Yang, J. H. Anderson, and F. D. Smith, "Gpu scheduling on the nvidia tx2: Hidden details revealed," in *2017 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, 2017, pp. 104–115.
- [21] N. Capodieci, R. Cavicchioli, P. Valente, and M. Bertogna, "Sigamma: Server based integrated gpu arbitration mechanism for memory accesses," in *Proceedings of the 25th International Conference on Real-Time Networks and Systems*. ACM, 2017, pp. 48–57.
- [22] P. Emberson, R. Stafford, and R. I. Davis, "Techniques for the synthesis of multiprocessor tasksets," in *WATERS*, 2010.
- [23] W. Ali and H. Yun, "Protecting real-time gpu applications on integrated cpu-gpu soc platforms," *arXiv preprint arXiv:1712.08738*, 2017.
- [24] R. Cavicchioli, N. Capodieci, and M. Bertogna, "Memory interference characterization between cpu cores and integrated gpus in mixed-criticality platforms," in *Emerging Technologies and Factory Automation (ETFA), 2017 22nd IEEE International Conference on*. IEEE, 2017, pp. 1–10.