University of Pennsylvania
ScholarlyCommons

Technical Reports (CIS) | Department of Computer & Information Science

1-1-2011

Realizing Compositional Scheduling Through Virtualization

Jaewoo Lee
University of Pennsylvania

Sisu Xi
Washington University in St Louis

Sanjian Chen
University of Pennsylvania

Linh T.X. Phan
University of Pennsylvania

Chris Gill
Washington University in St Louis

See next page for additional authors

Follow this and additional works at: http://repository.upenn.edu/cis_reports

Recommended Citation

Jaewoo Lee, Sisu Xi, Sanjian Chen, Linh T.X. Phan, Chris Gill, Insup Lee, Chenyang Lu, and Oleg Sokolsky, "Realizing Compositional Scheduling Through Virtualization", . January 2011.

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-11-13.
Corresponding Conference Paper:
http://repository.upenn.edu/cis_papers/484/

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/cis_reports/955
For more information, please contact libraryrepository@pobox.upenn.edu.

Realizing Compositional Scheduling Through Virtualization

Abstract
We present a co-designed scheduling framework and platform architecture that support compositional scheduling of real-time systems. The architecture is built on Xen virtualization platform, and relies on compositional scheduling theory that uses periodic resource models as component interfaces. We implement resource models as periodic servers and consider enhancements to periodic server design that significantly improve response times of tasks and resource utilization in the system while preserving theoretical schedulability results. We present an extensive evaluation of our implementation using workloads from an avionics case study as well as synthetic ones.

Comments
University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-11-13.

Corresponding Conference Paper:
http://repository.upenn.edu/cis_papers/484/

Author(s)
Jaewoo Lee, Sisu Xi, Sanjian Chen, Linh T.X. Phan, Chris Gill, Insup Lee, Chenyang Lu, and Oleg Sokolsky

This technical report is available at ScholarlyCommons: http://repository.upenn.edu/cis_reports/955

Realizing Compositional Scheduling through Virtualization

Jaewoo Lee[1†] Sisu Xi[2†] Sanjian Chen[1] Linh T.X. Phan[1]
Chris Gill[2] Insup Lee[1] Chenyang Lu[2] Oleg Sokolsky[1]
[1]University of Pennsylvania, USA
[2]Washington University in Saint Louis, USA
E-mail: {jaewoo,sanjian,linhphan,lee,sokolsky}@cis.upenn.edu, {xis,cdgill,lu}@cse.wustl.edu

***Abstract*—We present a co-designed scheduling framework and platform architecture that support compositional scheduling of real-time systems. The architecture is built on Xen virtualization platform, and relies on compositional scheduling theory that uses periodic resource models as component interfaces. We implement resource models as periodic servers and consider enhancements to periodic server design that significantly improve response times of tasks and resource utilization in the system while preserving theoretical schedulability results. We present an extensive evaluation of our implementation using workloads from an avionics case study as well as synthetic ones.**

I. Introduction

Modular development of real-time systems using time-aware components is an important means of reducing complexity of modern real-time systems. Components encapsulate real-time workloads, such as tasks, and are supported by a *local scheduler* that handles those workloads. Components share computational resources with other components. A *higher-level scheduler* is then used to allocate resources to local schedulers, guided by the components' resource needs, which they expose in their interfaces.

Several compositional scheduling frameworks (CSF) have been proposed to support such a component-based approach. Scheduling needs to be compositional to achieve a desirable separation of concerns: on the one hand, the high-level scheduler should not have access to the component internals and should operate only on component interfaces; on the other hand, schedulability analysis of a component's workload and generation of the component interface need to be performed independently from any other components in the system. Further, schedulability analysis at the higher level should be performed only on the basis of component interfaces.

In this paper, we present the Compositional Scheduling Architecture (CSA), which is an implementation of a CSF that relies on periodic resource models as component interfaces. Theoretical background for such an architecture, which provides interface computation for real-time workloads and schedulability analysis, has been laid out in [1], [2]. CSA is built on the virtualization framework provided by Xen, with the VMM being the root component and the guest operating systems (domains) being its subcomponents. Each domain interface is implemented as a periodic server [3], which behaves like a periodic task. The virtual machine monitor (VMM) allocates resources to the domains by scheduling the corresponding servers in the same manner as scheduling a set of tasks.

We also discuss challenges encountered during our implementation of the CSA and our approach to overcome those challenges. In particular, we discovered that CSF theory needed to be modified because of the fixed scheduling quantum imposed by Xen. This precludes direct use of the interface computation algorithm described in [4], since the resource bandwidth computed for the interface has to be an integer multiple of the quantum. Moreover, we discovered that a naive implementation of the periodic server is not work conserving and may lead to significant underutilization of the available computational resources.

Contributions. This paper makes the following distinct contributions to the state of the art in component-based real-time systems:

- We present CSA, a platform architecture for a CSF based on a periodic resource model, using virtualization in Xen. CSA enables timing isolation among virtual machines and supports timing guarantees for real-time tasks running on each virtual machine. Our implementation comes with a wide range of real-time scheduling algorithms at the VMM level, and it is easily extensible with new scheduling algorithms.
- We introduce several enhancements to the periodic server design in CSA to optimize the performance of both hard and soft real-time applications. Our enhancements preserve conservative CSF schedulability analysis, while yielding substantial improvements in observed response times and resource utilization, which are desirable for not only soft real-time but also many classes of hard real-time applications.
- We provide an extension of the CSF theory for quantum-based platforms and fixed-priority scheduling.
- We offer an extensive evaluation of the performance of CSA with respect to a variety of workloads, some of which originate from the avionics system reported in [5] and others that are synthetic.

To the best of our knowledge, CSA is the first open source implementation of a real-time virtualization platform with support for compositional scheduling.

†The first two authors have made equal contributions to this work.

II. Background

A. Compositional Scheduling Framework (CSF)

In a CSF, the system consists of a set of *components*, where each component is composed of either a set of subcomponents or a set of tasks. Each component is defined by $C = (W, \Gamma, A)$, where: W is a workload, i.e., a set of tasks (components); Γ is a resource interface; and A is a scheduling policy used to schedule W, which in our setting is Rate Monotonic (RM). All tasks are periodic, where each task T_i is defined by a period (and deadline) p_i and a worst-case execution time e_i, with $p_i \geq e_i > 0$ and $p_i, e_i \in \mathbb{N}$. Interface Γ is a periodic resource model (described below).

Periodic resource model. A periodic resource model (PRM) is defined by $\Gamma = (\Pi, \Theta)$, where Π is the resource period and Θ is the execution budget guaranteed by Γ in every period. The *bandwidth* of Γ is defined by $\mathsf{bw}(\Gamma) = \Theta/\Pi$. A PRM is *(bandwidth) optimal* for W iff it has the smallest bandwidth among all PRMs that can feasibly schedule W. A workload W is *harmonic* iff the periods of its tasks (subcomponents' interfaces) are pairwise divisible.

The minimum resource guaranteed by a PRM Γ is captured by a *supply bound function* (SBF) [1], written as $\mathsf{sbf}_\Gamma(t)$, which gives the minimum number of execution units provided by Γ over any time interval of length t, for all $t \geq 0$. The SBF of $\Gamma = (\Pi, \Theta)$ for a workload W is thus given by [1], [5]:

$$\mathsf{sbf}_\Gamma(t) = \begin{cases} y\Theta + \max\big(0,\, t - x - y\Pi\big), & \text{if } t \geq \Pi - \Theta \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where

- $x = (\Pi - \Theta)$ and $y = \lfloor \frac{t}{\Pi} \rfloor$, if W is harmonic; and
- $x = 2(\Pi - \Theta)$ and $y = \lfloor \frac{t-(\Pi-\Theta)}{\Pi} \rfloor$, otherwise.

Schedulability condition. Given $C = (W, \Gamma, RM)$ with $W = \{T_1, T_2, \cdots, T_n\}$, $T_i = (p_i, e_i)$, and $p_1 \leq p_2 \leq \cdots \leq p_n$. Here, T_i is a periodic task or a PRM interface of a subcomponent of C. Resource demands of C are characterized by the request bound functions (RBFs) of W, given by $\mathsf{rbf}_{W,i}(t) = \sum_{k \leq i} \left(\left\lceil \frac{t}{p_i} \right\rceil e_i \right)$ for all $1 \leq i \leq n$ [6]. Lemma 1 states its schedulability condition based on $\mathsf{rbf}_{W,i}$ and sbf_Γ [1].

Lemma 1: Given a component $C = \big(W, \Gamma, RM\big)$ with $W = \{T_1, T_2, \cdots, T_n\}$ and $T_i = (p_i, e_i)$ for all $1 \leq i \leq n$. Then, C is schedulable (Γ can feasibly schedule W) iff

$$\forall\ 1 \leq i \leq n,\ \exists t \in [0, p_i]\ s.t.\ \ \mathsf{sbf}_\Gamma(t) \geq \mathsf{rbf}_{W,i}(t). \quad (2)$$

From Lemma 1, a necessary schedulability condition for C is $\mathsf{bw}(\Gamma) \geq \mathsf{bw}(W)$, where $\mathsf{bw}(\Gamma) = \Theta/\Pi$ and $\mathsf{bw}(W) = \sum_{i=1}^{n} e_i/p_i$. The difference, $\mathsf{bw}(\Gamma) - \mathsf{bw}(W)$, is called the *interface overhead* of C. Thus, Γ is optimal for W iff it has the smallest interface overhead compared to all interfaces that can feasibly schedule W. It can be implied from Lemma 1 and Eq. (1) that the interface computed assuming a harmonic workload has a smaller (possibly zero) interface overhead than that of an interface computed assuming a general workload.

PRMs as periodic servers. Each PRM interface $\Gamma = (\Pi, \Theta)$ is implemented as a periodic server [3] with period Π and execution budget Θ, i.e., the server is ready for execution periodically every Π time units and its execution time is at most Θ time units. Thus, interfaces of components can be scheduled in the same manner as periodic tasks are. Further, a component is schedulable iff its interface (i.e., periodic server) is feasibly scheduled by its parent component.

B. Overview of Xen

Xen [7], the most widely used open source virtual machine monitor (VMM), allows a set of guest operating systems (OS), called *domains*, to run concurrently. To guarantee that every guest OS receives an appropriate amount of CPU time, Xen provides a scheduling framework within which developers can implement different scheduling policies. In this framework, every core in a guest OS is instantiated as a Virtual CPU (VCPU), and a guest OS can have as many VCPUs as there are underlying physical cores. Xen schedules VCPUs in the same manner as a traditional OS schedules processes, except that its pluggable scheduling framework allows different scheduling policies to be used. An IDLE VCPU is also created for each physical core to represent an IDLE task in a traditional OS. When the IDLE VCPU is scheduled, the specific physical core becomes idle.

In our earlier work [8], we have developed RT-Xen, a real-time virtual machine manager that supports hierarchical real-time scheduling in Xen. The compositional scheduling architecture (CSA) presented in this paper builds on and complements RT-Xen with a compositional scheduling capability. It differs from RT-Xen in four important aspects: (1) while RT-Xen instantiates hierarchical real-time scheduling in Xen, it was not designed to support the CSF model where the resource demand of a component is encapsulated by its interface; (2) RT-Xen focuses on the implementation and evaluation of different existing server algorithms, including Polling Server, Deferrable Server, Sporadic Server, as well as the classical Periodic Server that is used as a baseline in this work - in contrast, this work proposes two new work-conserving Periodic Server algorithms to improve soft real-time performance; (3) this work presents a new method to select the optimal interface parameters for a given scheduling quantum for RM scheduling, an issue not addressed by RT-Xen or the earlier work; (4) this work introduces an integrated scheduler architecture that allows different periodic servers to be instantiated through component reuse and enables the schedulers to be swapped online.

C. Challenges

Despite the availability of considerable *theoretical* results on CSF for real-time systems, those results have yet to be implemented in a virtualization platform such as Xen. The gap between theory and systems results in two significant problems. First, real-time system integrators cannot take advantage of the body of CSF theory in practice due to a lack of system

support. We have addressed that issue by developing a novel *Compositional Scheduling Architecture (CSA)* within the Xen virtual machine monitor (VMM). This unified scheduling architecture supports different scheduling policies at the VMM level, while preserving the modularity and extensibility of the scheduler implementation.

Moreover, without implementation and experimentation on a real system, it is not possible to explore crucial system design tradeoffs and practical issues involved in realizing a particular CSF in a given virtualization platform, such as the following important practical issues we faces in realizing the PRM-based CSF in Xen.

Non-work-conserving scheduling. The periodic server policy was proposed as an effective mechanism for implementing scheduler interfaces in CSF. However, the classical periodic server algorithm [3], referred to as a *Purely Time-driven Periodic Server (PTPS)* in this paper, adopts a non-work-conserving policy. Specifically, when a higher-priority component has no work to do, it simply idles away its budget while lower-priority component are not allowed to run. RT-Xen [8] emulates this feature by scheduling the IDLE VCPU to run while a high-priority domain idles away its budget. This scenario arises when a high-priority domain underutilizes its budget, e.g., due to an interface overhead or an over-estimation of tasks' execution times when configuring the domains' budgets. While the non-work-conserving nature does not affect the *worst-case* guarantees provided by PTPS, it wastes CPU cycles while increasing the response times of low-priority domains. This is particularly undesirable for soft real-time systems, as well as many hard real-time systems where short response times are beneficial in addition to meeting deadlines.

Scheduling quantum. While previous interface calculation techniques assume real values for interface budgets, a real system such as Xen must deal with quantized scheduling. For example, experimental results with RT-Xen showed that 1ms is a suitable scheduling quantum within Xen [8] in order to balance scheduling overhead and temporal granularity of scheduling. To deal with quantized scheduling, new techniques are needed to compute the bandwidth optimal interface for a guest OS and the maximum value of the optimal period when using the RM scheduling algorithm.

III. Solution Approach

Real-time guarantees in Xen can be achieved via compositional schedulability analysis in our *Compositional Scheduling Architecture (CSA)*. As is shown in Figure 1, the Xen VMM corresponds to a root component, and each Xen domain corresponds to a subcomponent of the root component in the CSA. The Xen VMM's scheduler (extended from the original RT-Xen interfaces) schedules domains based on their PRM interfaces, which are implemented as periodic servers (described in Section III-A). Each server's period and budget are computed using our quantum-based extension of compositional scheduling theory (described in Section III-B) to ensure schedulability of tasks in the underlying domain. The system is hence schedulable iff all servers are feasibly scheduled by the VMM's scheduler.

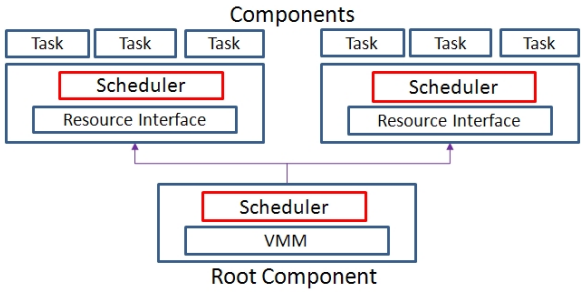


Fig. 1: A Compositional Scheduling Architecture

A. Periodic Server Design

In this section, we present two enhanced variations of the *purely time-driven periodic server* to optimize run-time performance and resource-use efficiency, namely the *work-conserving periodic server* and the *capacity-reclaiming periodic server*. These variations differ in how a server budget changes when the server has remaining budget but is idle (i.e., has no unfinished jobs), or when it is non-idle but has no budget left. Recall that in the classical purely time-driven periodic server, a server's budget is replenished to full capacity every period. The server is eligible for execution only when it has non-empty budget, and its budget is always consumed at the rate of one execution unit per time unit, even if the server is idle. In the work-conserving periodic server variant, whenever the currently scheduled server is idle, the VMM's scheduler let another lower-priority non-idle server to start early; as a result, the system is never left idle if there are unfinished jobs in a lower-priority domain. Finally, the capacity-reclaiming periodic server variant further utilizes the unused resource budget of an idle server to execute jobs of any other non-idle servers, effectively adding extra budget to the other non-idle servers. In what follows, "the scheduler" refers to the VMM's scheduler, unless explicitly mentioned otherwise.

Purely Time-driven Periodic Server (PTPS). As is mentioned above, the budget of a PTPS is replenished at every period and its budget is always consumed whenever it is executed. As Xen is an event-triggered virtual platform, we introduce a mechanism to allow this time-driven budget replenishment and scheduling approach in CSA. Note that the PTPS approach is not work-conserving since the system resource is always left unused if the currently scheduled server (Xen domain) is idle.

Work-Conserving Periodic Server (WCPS). The budget of a WCPS is replenished in the same fashion as that of a PTPS. However, if the currently scheduled server (C_H) is idle, the scheduler picks a lower-priority non-idle server to execute,

according to the following work conserving rules:

(1) Choose a lower-priority server, C_L, with the highest priority among all non-idle lower-priority servers.

(2) Start executing C_L and consuming the budgets of both C_L and C_H, each at the rate of one unit per time unit.

(3) Continue running C_L until one of the following occurs: (a) C_L has no more jobs to execute; (b) C_L has no more budget; (c) Some jobs in C_H become ready and C_H has remaining budget; or (d) C_H has no more budget. In the case of (a) or (b), the scheduler goes back to Step 1 where it selects another lower-priority non-idle server. In the case of (c), C_L immediately stops its execution and budget consumption, whereas C_H resumes its execution. In the case of (d), C_L immediately stops its execution and budget consumption; a new server will be chosen for execution by the scheduler.

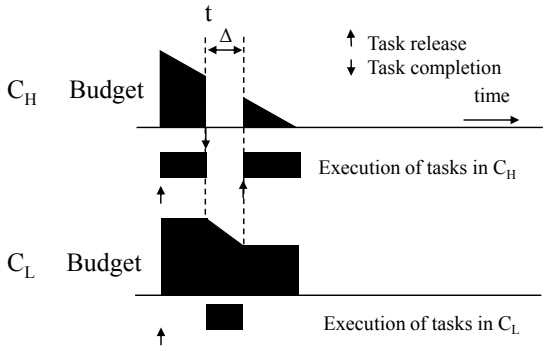


Fig. 2: Execution of Servers in the WCPS Approach.

Figure 2 illustrates a general scenario under the work conserving rule. In this scenario, C_H becomes idle at time t and thus, a lower-priority server C_L is selected for execution. At time $t+\Delta$, some jobs in C_H become ready (i.e., case (c) in Step 3); therefore, C_H preempts C_L and resumes its execution. By allowing C_L to start early (if C_H is idle) and maintaining the same execution for C_H, the WCPS achieves shorter overall response times of tasks compared to PTPS while preserving conservative CSF schedulability.

Lemma 2: A CSF system is schedulable under the WCPS approach if it is schedulable under the PTPS approach.

Proof: Suppose the servers are schedulable under the PTPS approach. Since the period and budget of a WCPS is the same as that of a PTPS, the parameters of all interfaces and workloads within the CSF are the same in both approaches. Additionally, the execution patterns of jobs in both PTPS and WCPS are the same, except when the work conserving rule is applied. It thus suffices to show that for any server C_H and any selected C_L during C_H's idle time – the only servers that are affected – remain schedulable.

As was discussed in Section II-A, analyzing schedulability of the CSF is twofold: (i) C_H and C_L are schedulable by their parent component; and (ii) C_H and C_L receive sufficient resource to schedule their own workloads. Under the work conserving rule, condition (i) is always guaranteed because WCPS does not introduce any additional workloads. To show condition (ii), we first note that C_L receives the same amount of resource as it would get under the PTPS, and hence, it is schedulable. Since C_H has no jobs to execute during the time C_L is executed, the available resource (e.g., during $[t, t+\Delta]$ in Fig. 2) would be wasted if assigned to C_H. Therefore, allocating such available resource to other domains while C_H is idle does not affect the schedulability of C_H. Hence, the work conserving preserves the schedulability of the CSF. ■

Capacity Reclaiming Periodic Server (CRPS). Like the WCPS, the CRPS is also work conserving and the budget of a server is replenished to full capacity every period. However, the CRPS improves tasks' response times by allowing idle time of the currently running server to be utilized by any other server (including higher-priority ones). Specifically, we define the *residual capacity* of a server to be the time interval during which the server consumes its budget but is idle (e.g., C_H has a residual capacity of $[t, t+\Delta]$ in Figure 2). At run time, the server budget is modified using the following capacity-reclaiming rule: during a residual capacity interval of a server C_H, the resource budget of C_H is re-assigned to any other non-idle server C_L and only this budget is consumed (e.g., the budget of C_L remains intact).

Similarly, we can show that the CRPS also preserves conservative CSF schedulability. Since each CRPS server gets not only its own resource budget but also the extra budgets of idle servers, it can potentially finish its jobs earlier than a corresponding WCPS or PTPS can. This results in an overall improvement in tasks' response times compared to the WCPS and PTPS approaches, as is also validated in our evaluation (see Section IV). Note that due to the capacity reclaiming capability, the CRPS is most difficult to implement among the three server variants.

B. Interface Computation for Quantum-based Platforms

In the existing CSF theory [4], the optimal PRM interface of a component is computed by iterating the resource period from 1 to a manually chosen value, while assuming rational values for the resource budget. For this approach to be implementable, given a particular time granularity of a Xen platform, the resource budget needs to be scaled to a multiple of the time unit. Consider the following example. Let the optimal PRM for a component be (1,0.54). Rounding the result up, we obtain the PRM (1,1). However, this may not be the minimum bandwidth that can be obtained with integer values, as a PRM (4,3) may be able to schedule the component.

Further, a naive choice of the period's bound can also result in sub-optimality. To address these shortcomings, in this section we introduce an algorithm for computing the optimal PRM interface for quantum-based platforms under RM scheduling.

Upper bound on the optimal interface period. Theorem 1 gives an upper bound on the resource period of the optimal interface of a given workload $W = \{(p_i, e_i) \mid 1 \leq i \leq n\}$ under RM. Intuitively, a PRM interface Γ is schedulable *only if* its upper supply bound function (USBF) (i.e., the minimum

sloped upper linear curve of the interface's SBF) meets each $\mathsf{rbf}_{W,i}$ at a step-point of $\mathsf{rbf}_{W,i}$ and is below $\mathsf{rbf}_{W,i}$ at all other points in $[0, p_i]$. We call these meeting points *critical points*, with $\mathsf{CrT}_{W,i}$ denoting the set of time-coordinates of the critical points of $\mathsf{rbf}_{W,i}$. Thus, the optimal resource bandwidth is lower bounded by the minimum slope of all linear curves f_i^t that are equal to $\mathsf{rbf}_{W,i}$ at time $t \in \mathsf{CrT}_{W,i}$ and smaller than $\mathsf{rbf}_{W,i}$ at all other times. As a result, the optimal resource period is upper bounded by the minimum of all P_i $(1 \le i \le n)$, where P_i is the maximum of the periods P_i^t of the PRMs with USBFs f_i^t for all $t \in \mathsf{CrT}_{W,i}$. Theorem 1 computes this upper bound based on an initial feasible PRM Γ_c for W.

Theorem 1: Suppose $\Gamma_c = (\Pi_c, \Theta_c)$ is the minimum bandwidth PRM among all PRMs that can feasibly schedule a workload W and whose period is at most Π_c. Then, the optimal PRM $\Gamma_{\mathsf{opt}} = (\Pi_{\mathsf{opt}}, \Theta_{\mathsf{opt}})$ for W satisfies $\Pi_c \le \Pi_{\mathsf{opt}} \le \mathsf{MaxResPeriod}(\kappa, W)$ where $\kappa = \frac{\Theta_c}{\Pi_c}$ and

$$\mathsf{MaxResPeriod}(\kappa, W) \stackrel{\mathsf{def}}{=} \min_{1 \le i \le n} \left(\max_{t \in \mathsf{CrT}_{W,i}} \frac{\kappa \cdot t - \mathsf{rbf}_{W,i}(t)}{\kappa(1-\kappa)} \right).$$

Before presenting the proof of Theorem 1, we explicly define some notations and provide basic ideas. An upper bound of the optimal resource period for a given workload W can be derived based on the *upper supply bound functions* (USBFs) [2] of all PRMs that can potentially schedule W. Recall that the USBF of a resource model Γ, denoted by usbf_Γ, is the minimum-slopped linear function that upper bounds sbf_Γ. The USBF of a PRM $\Gamma = (\Pi, \Theta)$ is

$$\forall t \ge 0,\ \mathsf{usbf}_\Gamma(t) = \max\big(\frac{\Theta}{\Pi}(t - (\Pi - \Theta)), 0\big).$$

One can easily prove that a necessary condition for a workload $W = \{T_1, T_2, \cdots, T_n\}$, with $T_i = (p_i, e_i)$ for all i, to be schedulable by Γ under RM is

$$\forall\ i, \exists t \in [0, p_i]\ s.t.\ \ \mathsf{usbf}_\Gamma(t) \ge \mathsf{rbf}_{W,i}(t). \qquad (3)$$

We say that Γ can *potentially* schedule W if it satisfies Eq. (3).

For each task T_i and each $\Pi \in \mathbb{N}$, let $\Gamma_{i,\Pi}$ be the smallest bandwidth PRM among all PRMs with period Π that can *potentially* schedule W. To find an upper bound of the optimal resource period Π_{opt}, we need at least one feasible PRM for a given workload. Let κ be the bandwidth of the feasible PRM. Then, we can derive an upper bound of the period Π_{opt} in the optimal PRM $\Gamma_{\mathsf{opt}} = (\Pi_{\mathsf{opt}}, \Theta_{\mathsf{opt}})$ based on the fact that κ is greater than or equal to the smallest bandwidth among that of all PRMs with period Π_{opt} that can *potentially* schedule W. That is, $B_{\min}(\Pi_{\mathsf{opt}}) \le \kappa$, where $B_{\min}(\Pi) = \max_{1 \le i \le n} B^i_{\min}(\Pi)$ and $B^i_{\min}(\Pi)$ is the bandwidth of $\Gamma_{i,\Pi}$.

By definition of $\Gamma_{i,\Pi}$, we imply that its USBF, $\mathsf{usbf}_{\Gamma_{i,\Pi}}(t)$, meets $\mathsf{rbf}_{W,i}(t)$ at exactly one time point, $t = X_{i,\Pi}$, called the *critical time point*, which is defined by

$$X_{i,\Pi} = \underset{(\Pi - \Theta_{i,\Pi}) < t \le p_i}{\mathsf{argmin}} \left\{ \frac{\mathsf{rbf}_{W,i}(t)}{t - (\Pi - \Theta_{i,\Pi})} \right\},$$

and $\mathsf{usbf}_{\Gamma_{i,\Pi}}(t) < \mathsf{rbf}_{W,i}(t)$ for all $t \ne X_{i,\Pi}$. Note that $\frac{\mathsf{rbf}_{W,i}(t)}{t-(\Pi-\Theta_{i,\Pi})}$ is the slope of $\mathsf{usbf}_{\Gamma_{i,\Pi}}(t)$.

Example 1: Consider a workload W={(7,2), (8,1), (10,1)}. For $i = 2$ and $\Pi = 5$, $X_{2,5}$ is 7, which is shown in Figure 3.

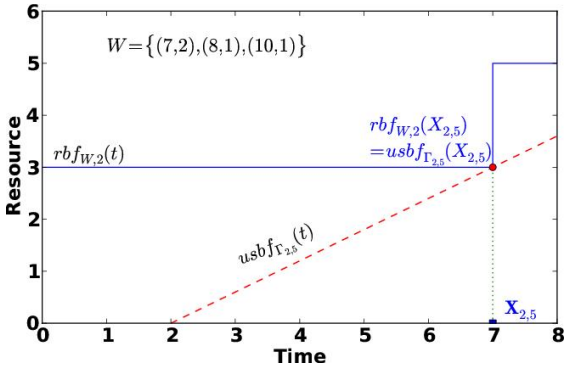


Fig. 3: $X_{i=2,\Pi=5}$ for a workload W={(7,2), (8,1), (10,1)}

Eliminating the dependency of Π, the set of all critical time points for a workload W and any given task T_i in W is formally given by (independent of Π):

$$\begin{aligned}\mathsf{CrT}_{W,i} &= \{X_{i,\Pi} \mid \Pi \in \mathbb{N}\} \\ &= \left\{ \underset{s < t \le p_i}{\mathsf{argmin}} \left\{ \frac{\mathsf{rbf}_{W,i}(t)}{t-s} \right\} \ \mid\ s \in \{0, 1, \cdots, p_i - 1\} \right\}.\end{aligned}$$

Therefore, $\Gamma_{i,\Pi}$ is the minimum bandwidth PRM of all Γ for which usbf_Γ equals $\mathsf{rbf}_{W,i}$ at (exactly) one critical time point in $\mathsf{CrT}_{W,i}$ and is smaller than $\mathsf{rbf}_{W,i}$ at all other time points. Lemma 3 shows that if some PRM is the smallest bandwidth PRM which *potentially* schedule $\mathsf{rbf}_{W,i}$ and its usbf_Γ meet a time point t with $\mathsf{rbf}_{W,i}$, then t belong to $\mathsf{CrT}_{W,i}$.

Lemma 3: Suppose Γ is the smallest bandwidth PRM which *potentially* schedule $\mathsf{rbf}_{W,i}$. Then, if $\exists t'$ s.t. $0 \le t' \le p_i$ and $\mathsf{usbf}_\Gamma(t') = \mathsf{rbf}_{W,i}(t')$ and $\forall t$ s.t. $t \ne t'$, $0 \le t \le p_i$, and $\mathsf{usbf}_\Gamma(t) < \mathsf{rbf}_{W,i}(t)$, then $t' \in \mathsf{CrT}_{W,i}$.

Proof: We will prove the lemma by contradiction. Its details are available in the Appendix. ∎

From the above result, $B^i_{\min}(\Pi)$ can now be computed by examining all USBFs that go through a critical point in $\mathsf{CrT}_{W,i}$. In other words, $B^i_{\min}(\Pi) = \min_{t \in \mathsf{CrT}_{W,i}} B^i_{\min}(\Pi, t)$, where $B^i_{\min}(\Pi, t)$ is the bandwidth of the PRM $\Gamma^* = (\Pi, \Theta^*)$ that has $\mathsf{usbf}_{\Gamma^*}(t) = \mathsf{rbf}_{W,i}(t)$ at time point t:

$$\begin{aligned}B^i_{\min}(\Pi) &= \min_{t \in \mathsf{CrT}_{W,i}} B^i_{\min}(\Pi, t) \\ &= \min_{t \in \mathsf{CrT}_{W,i}} \{ \frac{\Theta^*}{\Pi} \ \mid\ \mathsf{usbf}_{\Gamma^*}(t) = \mathsf{rbf}_{W,i}(t) \}\end{aligned}$$

where $\Gamma^* = (\Pi, \Theta^*)$ and $\mathsf{CrT}_{W,i}$ is the set of all critical time points of W and task index i. Note that f_i^t in the intuition before stating Theorem 1 is defined to usbf function of Γ^* satisfying $\mathsf{usbf}_{\Gamma^*}(t) = \mathsf{rbf}_{W,i}(t)$ at time $t \in \mathsf{CrT}_{W,i}$.

Lemma 4 can compute $B^i_{\min}(\Pi, t)$. For any given Π, any given task index i, and any given $t \in \mathsf{CrT}_{W,i}$, let $B^i_{\min}(\Pi, t)$ be the bandwidth of the PRM with period Π such that its USBF only intersects $\mathsf{rbf}_{W,i}$ at time point t. Then, $B^i_{\min}(\Pi, t)$ is unique.

Lemma 4: Given any $\Pi \in \mathbb{N}$, any task index $i \leq n$, and any $t \in \mathsf{CrT}_{W,i}$. Let $\Gamma^* = (\Pi, \Theta^*)$ be the PRM such that $\mathsf{usbf}_{\Gamma^*}(t) = \mathsf{rbf}_{W,i}(t)$. Then, the bandwidth of the PRM Γ^*,

$$B^i_{\min}(\Pi, t) \stackrel{\mathsf{def}}{=} \frac{\Theta^*}{\Pi} = \frac{\Pi - t + \sqrt{(\Pi - t)^2 + 4\Pi \cdot \mathsf{rbf}_{W,i}(t)}}{2\Pi}.$$

Proof: The proof is based on an observation that there exists a unique solution for Γ^* such that $\mathsf{usbf}_{\Gamma^*}(t) = \mathsf{rbf}_{W,i}(t)$. The details are available in the Appendix. ∎

We also know that the function $B^i_{\min}(\Pi, t)$ is increasing on the domain of Π.

Lemma 5: The function $B^i_{\min}(\Pi, t)$ defined in Lemma 4 is increasing on the domain of Π.

Proof: The proof is established based on the property that $dB^i_{\min}(\Pi, t)/d\Pi \geq 0$. Its details can be found in the Appendix. ∎

Thus, the upper bound of Π_{opt} can now be derived from the relation between $B_{\min}(\Pi_{\mathsf{opt}})$ and κ. We know that $B_{\min}$ depends on Π and the workload W. Since W is given, we can derive an inequality on Π_{opt} by the following transformation:

$$\begin{aligned} B_{\min}(\Pi_{\mathsf{opt}}) \leq \kappa \;&\Rightarrow\; \max_{1 \leq i \leq n} B^i_{\min}(\Pi_{\mathsf{opt}}) \leq \kappa \\ &\Rightarrow\; \max_{1 \leq i \leq n} \left(\min_{t \in \mathsf{CrT}_{W,i}} B^i_{\min}(\Pi_{\mathsf{opt}}, t) \right) \leq \kappa \end{aligned}$$

where $P^t_i \stackrel{\mathsf{def}}{=} B^i_{\min}(\Pi_{\mathsf{opt}}, t)$ and $P_i \stackrel{\mathsf{def}}{=} B^i_{\min}(\Pi_{\mathsf{opt}})$ in the intuition before stating Theorem 1. For a given task T_i and a given time point $t \in \mathsf{CrT}_{W,i}$, we have $B^i_{\min}(\Pi_{\mathsf{opt}}, t) \leq \kappa$.

As $B^i_{\min}(\Pi, t)$ is increasing on the domain of Π in Lemma 5, the following holds for Π_{opt}:

$$B^i_{\min}(\Pi_{\mathsf{opt}}, t) \leq \kappa \Leftrightarrow \quad \Pi_{\mathsf{opt}} \leq \frac{\kappa \cdot t - \mathsf{rbf}_{W,i}(t)}{\kappa(1-\kappa)}. \tag{4}$$

For a given task T_i, Eq. (4) needs to hold for at least one $t \in \mathsf{CrT}_{W,i}$ and thus, Π_{opt} is less than or equal to the maximum of the RHS values of the equation for all critical points t in $\mathsf{CrT}_{W,i}$. Since Eq. (4) needs to hold for all T_i in W, Π_{opt} is less than or equal to the minimum the RHS values computed for all $1 \leq i \leq n$. As a result, an upper bound for the optimal resource period can be computed as is given in Theorem 1. We can now provide the formal proof of Theorem 1.

Proof of Theorem 1: Since algorithm for the optimal interger PRM finds the optimal resource period in an increasing manner, we know that $\Pi_{\mathsf{opt}} \geq \Pi_c$. Further, that Γ_{opt} is optimal implies

$$\mathsf{bw}(\Gamma_{\mathsf{opt}}) \leq \mathsf{bw}(\Gamma_c) = \kappa. \tag{5}$$

where $\mathsf{bw}(\Gamma)$ is the bandwidth of Γ, i.e. Θ/Π if $\Gamma = (\Pi, \Theta)$. Let $\Gamma_{i,\mathsf{opt}} = (\Pi_{i,\mathsf{opt}}, \Theta_{i,\mathsf{opt}})$ be the optimal PRM with $rbf_{W,i}$. Then,

$$\mathsf{bw}(\Gamma_{\mathsf{opt}}) = \max_{1 \leq i \leq n} (\mathsf{bw}(\Gamma_{i,\mathsf{opt}})). \tag{6}$$

since Lemma 1 should be hold for all $0 \leq i \leq n$.

Next, for any given task index i and any given $t \in \mathsf{CrT}_{W,i}$, let $\Gamma^* = (\Pi_{i,\mathsf{opt}}, \Theta^*)$ where $\Theta^* = \Pi_{i,\mathsf{opt}} \cdot B^i_{\min}(\Pi_{i,\mathsf{opt}}, t)$. That is, by Lemma 4, the USBF of Γ^* intersects $\mathsf{rbf}_{W,i}$ at time point t.

$\Gamma_{i,\mathsf{opt}}$ must have a bandwidth which is greater or equal to at least one PRM which intersects $\mathsf{rbf}_{W,i}$ at $t \in \mathsf{CrT}_{W,i}$ [1]. Then,

$$\exists\, t \in \mathsf{CrT}_{W,i} :\; B^i_{\min}(\Pi_{i,\mathsf{opt}}, t) \leq \mathsf{bw}(\Gamma_{i,\mathsf{opt}}). \tag{7}$$

Combining Eq. (5), (6), and (7) and letting r_t be $\mathsf{rbf}_{W,i}(t)$, we obatin: for $\exists t \in \mathsf{CrT}_{W,i}$,

$$\begin{aligned} & B^i_{\min}(\Pi_{i,\mathsf{opt}}, t) \leq \kappa \\ \Leftrightarrow\; & \sqrt{(\Pi_{i,\mathsf{opt}} - t)^2 + 4\Pi_{i,\mathsf{opt}} \cdot r_t} \leq 2\Pi_{i,\mathsf{opt}} \cdot \kappa + t - \Pi_{i,\mathsf{opt}} \\ \Leftrightarrow\; & (\Pi_{i,\mathsf{opt}} - t)^2 + 4\Pi_{i,\mathsf{opt}} \cdot r_t \leq \big((2\kappa - 1)\Pi_{i,\mathsf{opt}} + t\big)^2 \\ \Leftrightarrow\; & \Pi_{i,\mathsf{opt}} \leq \frac{\kappa \cdot t - r_t}{\kappa(1-\kappa)} = \frac{\kappa \cdot t - \mathsf{rbf}_{W,i}(t)}{\kappa(1-\kappa)} \end{aligned} \tag{8}$$

Since Eq. (8) should be satisfied for $\exists t \in \mathsf{CrT}_{W,i}$, it can be rewritten as

$$\Pi_{i,\mathsf{opt}} \leq \max_{t \in \mathsf{CrT}_{W,i}} \frac{\kappa \cdot t - \mathsf{rbf}_{W,i}(t)}{\kappa(1-\kappa)}. \tag{9}$$

Since Eq. (9) shows the bound of optimal period for task T_i in W and should be satisfied for all tasks in W by Lemma 1,

$$\Pi_{\mathsf{opt}} \leq \min_{1 \leq i \leq n} \left(\max_{t \in \mathsf{CrT}_{W,i}} \frac{\kappa \cdot t - \mathsf{rbf}_{W,i}(t)}{\kappa(1-\kappa)} \right)$$

or $\Pi_{\mathsf{opt}} \leq \mathsf{MaxResPeriod}(\kappa, W)$. ∎

Algorithm 1 Optimal integer-valued interface computation.

Input: A workload W
Output: The optimal integer-valued PRM Γ_{opt} for W

```
1:  Θ' = MinExec(p_n, W)
2:  κ = Θ'/p_n
3:  Γ_opt = (p_n, Θ')
4:  Π_max = MaxResPeriod(κ, W)
5:  for Π = 1 to Π_max do
6:      Θ = MinExec(Π, W)
7:      if Θ/Π < κ then
8:          κ = Θ/Π
9:          Γ_opt = (Π, Θ)
10:         Π_max = min(Π_max, MaxResPeriod(κ, W))
11:     end if
12: end for
13: return Γ_opt
```

Optimal integer-valued PRM period computation. Algorithm 1 computes the optimal integer-valued PRM of a given workload W by incorporating the above upper bound of the resource period $\mathsf{MaxResPeriod}(\kappa, W)$.

In Lines 1-2, $\mathsf{MinExec}(p_n, W)$ gives the minimum budget for the period p_n such that the resulting PRM can feasibly schedule W (i.e., satisfies Lemma 1), and κ denotes the corresponding bandwidth. The initial bound on the resource period

[1] Otherwise, $\Gamma_{i,\mathsf{opt}}$ does not satisfy the USBF-schedulability condition for $\mathsf{rbf}_{W,i}$.

is given by $\Pi_{\max}$ in Line 4. The function MaxResPeriod(κ,W) in Lines 4 and 10 computes the upper bound on the optimal PRM as defined in Theorem 1. Finally, the minimum bandwidth acquired during the algorithm execution is stored in κ, and it is used to re-evaluate $\Pi_{\max}$ (Lines 7–11).

C. System Architecture

This section presents details about CSA and the implementation of the PTPS, WCPS, and CRPS in Xen.

Compositional Scheduling Architecture. In CSA, at the top level, an existing Xen scheduling framework provides interfaces to a specific scheduler. Each scheduler has its own data structure but must implement several common functions including *wake*, *do_schedule*, *sleep*, and *pick_cpu*. Since the three CSA schedulers mainly differ in how the budget is consumed, we provide a real-time sub-framework which abstracts common functions and data structures among the CSA schedulers. The scheduling-related functions such as *do_schedule* are implemented as pointers to functions in sub-schedulers. Under the real-time sub-framework, we implement PTPS, WCPS, and CRPS separately. Figure 4 shows a high-level view of CSA.

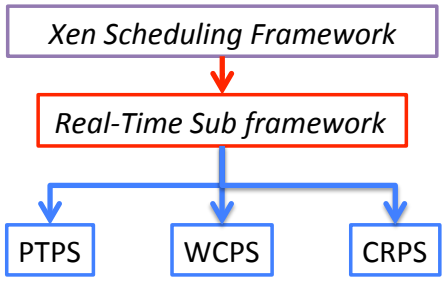


Fig. 4: The CSA Schedulers Architecture

Data structure. Each VCPU has three parameters: *budget*, *period*, and *priority*. In CSA, the priority is determined based on the VCPU's period according to the RM scheduling policy. Each VCPU is implemented as a periodic server, where its *budget* is set to full every *period* units of time, and the VCPU consumes its budget whenever it is executed. Each physical core has two queues: a *Run Queue* (RunQ) and a *Ready Queue* (RdyQ). Both queues are used to store active VCPUs. The IDLE VCPU, which has the lowest priority, is always located at the end of the RunQ. These queues operate as follows.

The RunQ stores VCPUs that have jobs to run (regardless of their budgets), sorted by priority. Whenever the *do_schedule* function is triggered, it first returns the current VCPU to the RunQ or the RdyQ. It then selects an appropriate VCPU from the RunQ based on the scheduling decision, and runs the selected VCPU for one millisecond (which has been shown to be a suitable scheduling quantum in [8]).

The RdyQ holds the VCPUs that have no jobs to run, sorted by their priorities. Because one VCPU can consume other VCPUs' *budget* (e.g., via capacity reclaiming in the CRPS, or scheduling the IDLE VCPU while consuming other VCPUs' *budget* in PTPS), all active VCPUs' information is needed to enable work conserving and capacity reclaiming enhancement.

Implementation of *do_schedule* function. The *do_schedule* function is responsible for updating VCPUs' information and making scheduling decisions. We first present the PTPS algorithm, and then explain the WCPS and CRPS extensions.

Algorithm 2 *do_schedule* function for PTPS

```
Input: currentVCPU, RunQ, RdyQ
Output: nextVCPU to run next
1:  rdyVCPU = queuePick(RdyQ)
2:  if currentVCPU = IDLE_VCPU then
3:      consume budget of rdyVCPU
4:  else
5:      consume budget of currentVCPU
6:  end if
7:  if currentVCPU has jobs to run then
8:      insert currentVCPU into RunQ
9:  else
10:     insert currentVCPU into RdyQ
11: end if
12: nextVCPU = queuePick(RunQ)
13: if priority(rdyVCPU) > priority(nextVCPU) then
14:     nextVCPU = IDLE_VCPU
15: else
16:     remove nextVCPU from RunQ
17: end if
18: return nextVCPU
```

In this algorithm, the function queuePick returns the highest priority VCPU with positive *budget* in the RunQ or in the RdyQ. Whenever a higher priority VCPU has a positive budget, we consume its budget by either scheduling it (if it has jobs to execute) or scheduling the IDLE VCPU (otherwise). Lines 1–6 demonstrate how we consume the highest priority VCPU's budget. Lines 12–17 show how the next VCPU is selected.

For the WCPS, if a higher-priority VCPU has budget but is idle, instead of scheduling the IDLE VCPU, we schedule in advance the next highest priority VCPU among all the non-idle lower-priority ones (so as to improve the responsiveness of jobs belonging to that VCPU); in this case, we consume both *budgets* in parallel. As a result, in Lines 12–17, the algorithm always returns the VCPU from the RunQ, denoted by queuePick(RunQ). Further, in Lines 1–6, if queuePick(RdyQ) has a higher priority than that of queuePick(RunQ), their *budgets* will be both consumed.

For CRPS, only one budget is consumed at a time and "Capacity Reclaiming" is enabled between active VCPUs. In Lines 1–6, the CRPS always consumes the highest *priority* VCPU's budget among currentVCPU, queuePick(RunQ), and queuePick(RdyQ). In Lines 12–17, if the function queuePick(RunQ) returns a VCPU that is different from the IDLE VCPU, that VCPU will be scheduled. Otherwise, the

IDLE VCPU is returned. There are two cases for this: either the RunQ is empty, or all active VCPUs on RunQ have no budget left. In the former case, the IDLE VCPU will be scheduled. In the latter case, if queuePick(RdyQ) returns a valid VCPU (i.e., other VCPUs have *budget*), the returned VCPU will be executed; otherwise, all active VCPUs have no budget left and thus, the IDLE VCPU will be scheduled (even if the active VCPUs still have jobs to execute). In other words, we do not allow budget to be *stolen* from the IDLE VCPU. The implementations of all the above algorithms, along with the hot-swap tool and the periodic tasks, are open source and can be found in [9].

IV. Evaluation

This section presents our evaluation of the PTPS, WCPS, and CRPS approaches that are implemented in our CSA. We focus on the run-time performance of real-time tasks, considering the following two evaluation criteria: (1) *responsiveness*, which is the ratio of a job's response time to its relative deadline; and (2) *deadline miss ratio*. Our evaluation consists of two types of workloads: synthetic workloads (Section IV-B) and ARINC workloads obtained from an avionics system (Section IV-C).

A. Experiment Setup

We implemented CSA in Xen version 4.0. Fedora 13 with para-virtualized kernel 2.6.32 is used for all domains. We pinned *Domain 0* to core 0 with 1 GB memory, and pinned all the guest operating systems to core 1 with 256 MB memory each, to emulate a single core environment. During the experiments, we shut down the network service as well as other inessential applications to avoid potential interference. The experiments for synthetic workloads were done on a Dell Q9400 quad-core processor while the experiments for ARINC workloads were performed on a Dell Vostro 430 quad-core processor, neither with hyper-threading. During the experiments, SpeedStep was disabled and all cores constantly ran at 2.66 GHz.

We assume all tasks are CPU intensive and independent of each other. Every task is characterized by three parameters: *worst case execution time (WCET)*, *period* (equals *deadline*), and *execution time factor (ETF)*. Here, the *ETF* represents the variance of each job's actual execution time (uniformly distributed in the interval $[WCET*ETF, WCET]$). An *ETF* of 100% indicates that every job of the task takes exactly *WCET* units of time to finish. The task model fits typical soft real-time applications (e.g., multimedia decoding applications where frames' processing times are varied but are always below an upper limit).

In the rest of the paper, U_W denotes the total utilization of all tasks in the system (utilization of the workload); U_{RM} denotes the total bandwidth of interfaces (utilization of resource models); $U_{RM} - U_W$ denotes the interface overhead.

Task Implementation and Guest OS Scheduler. We now describe the implementation for the periodic task and how to set up the guest OS scheduler.

We assume all tasks are CPU intensive and independent of each other. Every task is characterized by five parameters: *worst case execution time (wcet)*, *period*, *deadline* (which is equal to *period*), *priority*, and *execution time factor (etf)*, where the *etf* represents the variance of each job's actual execution time (uniformly distributed in the interval $[wcet * etf, wcet]$). Here, an *etf* of 100% indicates that every of its jobs takes exactly *wcet* units of time to finish. Our task model fits typical soft real-time applications, for example, in multimedia decoding application, the processing times of different frames of a video/audio stream may vary but are always below an upper limit.

We implement the task same way as in [8]. To implement the above task model in Linux, we first calibrate the exact amount of computation that needs 1 microsecond CPU time under the native Linux, and scale it to generate tasks with different execution times. We use the *SIGEV_SIGNAL* to release jobs periodically, and record the first job's release time as the *start time* for the task. During each run, every job's dispatch time and finish time are also recorded using the RDTSC instruction, which provides a nanosecond resolution with minimal overhead.

We note that in a *virtualized* environment, when a *domain* is not scheduled, the *signal* is not received until the *domain* is resumed, and then the timer for the next signal is set. This makes a periodic task behave like a sporadic task, where the job interval is not only determined by the *period* but also by the actual scheduling decision of the VMM scheduler. Although this may not be an issue when the domain's *period* is relatively small and the entire system is schedulable, it could make task deadline miss ratio calculation imprecise when the domain's *share* ($\frac{budget}{period}$) is relatively small or the system is overloaded. To avoid this problem, whenever the interrupt is received, we compare the current time with the *start_time*, and then calculate how many jobs should be released. By doing this, we ensure that jobs of a task are released periodically and executed in order. Further, the job execution is not affected by deadline misses.

For the data collection, we store the dispatch time and finish time of every job in locked memory to avoid memory paging overhead. Based on the job's *start time* and these records, we can calculate the job's *responsiveness* as $\frac{ResponseTime}{Deadline}$. The source code of our implementation can be found in [9].

Real-time scheduling of domains. We first determined the domains' resource needs by computing an optimal PRM interface for each domain. These interfaces were implemented as PTPS, WCPS, or CRPS variants of periodic servers, which were then scheduled by the VMM. For synthetic workloads, we applied Algorithm 1 to compute the optimal integer-valued PRM interfaces for the domains. The PRM interfaces of ARINC domains were computed based on Eq. (1) using the harmonic workload case. Since the domain periods are pre-specified in the ARINC workloads, the quantum-based interface computation technique in Algorithm 1 cannot be applied. Therefore, we resorted to computing optimal rational-valued interfaces, and then rounding up the budgets to the

closest integer values. Although the real-valued interfaces may have interface overheads of zero, rounding may introduce additional overheads, effectively allocating extra budget to the corresponding domains.

B. Synthetic Workloads

The purpose of this set of experiments is to compare the *soft real-time* performance of the three different periodic servers. The PTPS, WCPS, and CRPS servers differ primarily in how idle time is utilized within the system. The idle time comes from two main sources: the interface overhead due to theoretical pessimism [1]; and over-estimation of tasks' execution times (also called *slack*). Hence, we design two sets of experiments to show the effect of different idle times: (1) The range for the workload periods is varied to create different interface overheads; (2) The *Execution Time Factor (ETF)* for the jobs is varied so that if a job executes less than its *WCET*, it would potentially give some *slack* to other domains.

For *soft real-time* systems, we are interested not only in schedulable situations but also in overloaded situations. As a result, we ranged the U_W from 0.7 to 1.0, with a step of 0.1, to create different U_W conditions.

All the experiments were conducted as follows. We first defined a particular U_W, and then generated tasks (utilization uniformly distributed between 0.2% and 5%) until the U_W was reached. In this way, the generated real U_W is usually larger than the desired one, but would only be 0.05 more in the worst case. After all the tasks were generated, we randomly distributed the tasks among five domains.

Each experiment ran for 5 minutes. We collected data from all tasks and calculated the $\frac{ResponseTime}{Deadline}$ for all the task sets within each domain and within each experiment. For clarity of presentation, any job whose $\frac{ResponseTime}{Deadline}$ is greater than 3 is clipped at 3.

Impact of Task Period. We varied the task period range in this experiment to create different interface overheads, and evaluated the three schedulers for the generated task sets. For each different U_W (from 0.7 to 1.0), we generated three different task sets whose periods are uniformly distributed between [550ms, 650ms], [350ms, 850ms], and [100ms, 1100ms], respectively. From the calculated interfaces, the [350ms, 850ms] task period range gives the most interface overhead, followed by [100ms, 1100ms], and then [550ms, 650ms]. For all the experiments, the *ETF* value was set to 100%. In other words, we let all jobs execute at their worst case execution times, so that the idle time comes only from the interface overheads. Note that when the U_W is the same, we were scheduling different task sets under different task periods.

Figure 5, Figure 7, Figure 9 and Figure 11 shows the results for all domains under different U_W, where DMR means Deadline Miss Ratio. Since we are using rate monotonic scheduling, the higher priority domains have shorter periods, and thus have a larger number of jobs. The data in Figure 9 are therefore dominated by the results for higher priority domains. Lower priority domains, though having fewer jobs, suffer most from the *overloaded* situation. As such, we plot the data for the lowest priority domain (domain 5) in Figure 10 with the interface parameters given in the format of (*period*, *budget*). Figure 9 and Figure 10 clearly show that the CRPS outperforms the WCPS, which in turn outperforms the PTPS. Notably, with an interface overhead of 24% (Figure 10c), while all jobs miss their deadlines under the PTPS ($\frac{ResponseTime}{Deadline} > 1$), 60.5% and 6.2% of the jobs in domain 5 missed their deadlines under the WCPS and CRPS, respectively. These results demonstrate the effectiveness of the work-conserving and capacity-reclaiming mechanisms in exploiting the interface overhead to improve the performance of low-priority domains. The CRPS is the most effective approach for implementing the interfaces in CSA. The results for other U_W are also included.

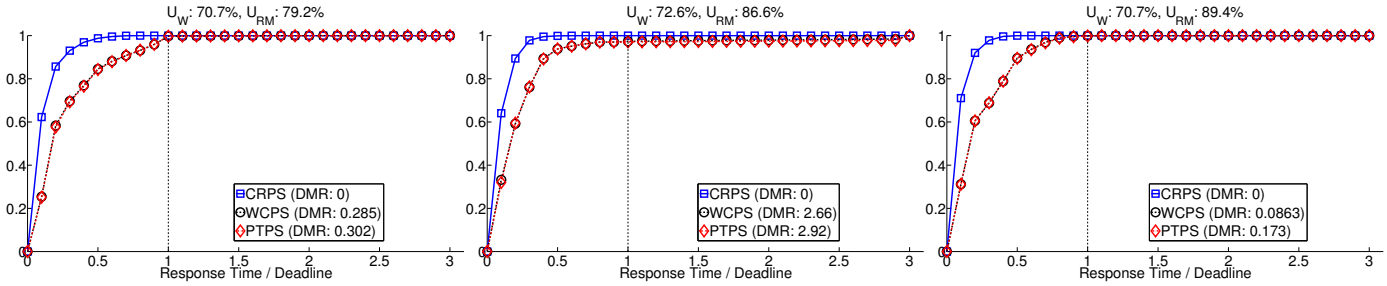
Impact of *Execution Time Factor (ETF)*.

In *real-time* applications such as multimedia frame decoding, every frame may take a different amount of time to finish. Traditionally, the *WCET* is used to represent every task's execution time. This usually results in a relatively large interface, giving more idle time for the domain.

In this set of experiments, the same U_W ranging from 0.7 to 1.0 were used. Under each U_W, we only generated one task set. Then, for each particular task set, three *ETF* values (100%, 50%, 10%) were configured for the three highest priority domains, while leaving the two low priority ones with an *ETF* of 100%. A lower *ETF* value means a lower "actual" U_W for that domain; for example, if an *ETF* of 10% is applied, all jobs' execution time uniformly distributes between 10% and 100% of *WCET*. On average, the actual U_W is 0.55 ($\frac{100\%+10\%}{2}$). All task periods were uniformly distributed between 550 ms and 650 ms. We note that the idle time comes not only from the interface overhead but also from the over-estimation of jobs' execution times.

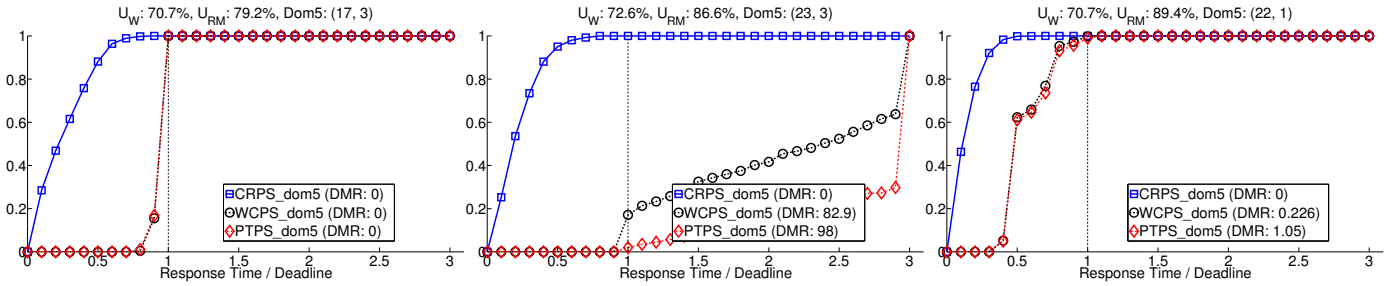
Figure 14 shows the box plot results for all U_W for the lowest priority domain. Results for all domains exhibit the same behavior. On each box, the central mark represents the median value, whereas the upper and lower box edges show the 25th and 75th percentiles separately. If the data values are larger than $q_3 + 1.5 * (q_3 - q_1)$ or smaller than $q_1 - 1.5 * (q_3 - q_1)$ (where q_3 and q_1 are the 75th and 25th percentiles, respectively), they are considered outliers and plotted via individual markers. Within one subfigure, the boxes are divided into three sets, from left to right, corresponding to the results under the *ETFs* of 100%, 50%, and 10%, respectively.

As is shown in Figure 14, the CRPS again outperforms the WCPS and PTPS. In Figure 14c, the deadline miss ratio under the PTPS stays constant when the *ETF* is varied (26.9%, 27.3%, and 27.3% respectively), while performance improvement is seen under the WCPS (11.7%, 8.51%, and 0.49%) and CRPS (0.02%, 0%, and 0%). In an extremely *overloaded* situation (Figure 14d), all jobs miss their deadlines under the PTPS, whereas (75.6%, 32.7%, and 31.3%) of jobs missed their deadlines under the WCPS, and (36.1%, 0%, and 0%) of jobs missed their deadlines under the CRPS. This again



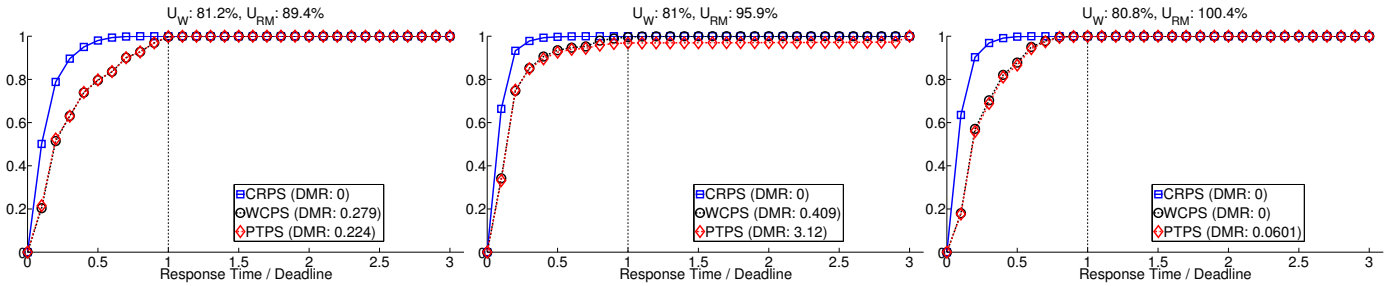
(a) Period: (550, 650), Overhead: 8.6% (b) Period: (100, 1100), Overhead: 13.9% (c) Period: (350, 850), Overhead: 18.7%

Fig. 5: CDF Plot of $\frac{ResponseTime}{Deadline}$ for All Tasks in Five Domains under Designed $U_W = 0.7$ varying Task Periods



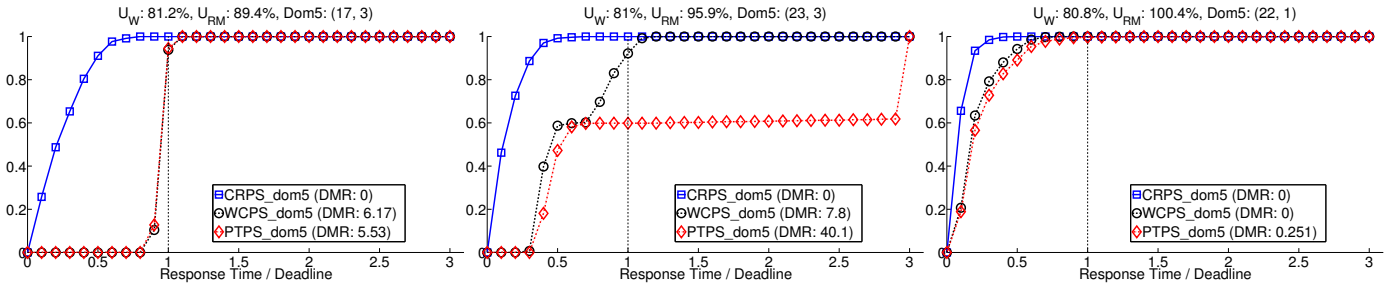
(a) Period: (550, 650), Overhead: 8.6% (b) Period: (100, 1100), Overhead: 13.9% (c) Period: (350, 850), Overhead: 18.7%

Fig. 6: CDF Plot of $\frac{ResponseTime}{Deadline}$ for Tasks in the Lowest Priority Domain under Designed $U_W = 0.7$ varying Task Periods



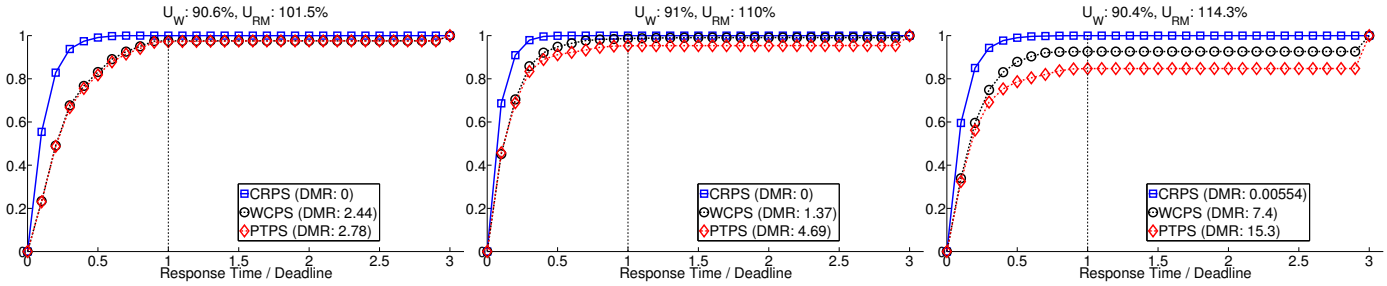
(a) Period: (550, 650), Overhead: 8.2% (b) Period: (100, 1100), Overhead: 19.6% (c) Period: (350, 850), Overhead: 14.9%

Fig. 7: CDF Plot of $\frac{ResponseTime}{Deadline}$ for All Tasks in Five Domains under Designed $U_W = 0.8$ varying Task Periods



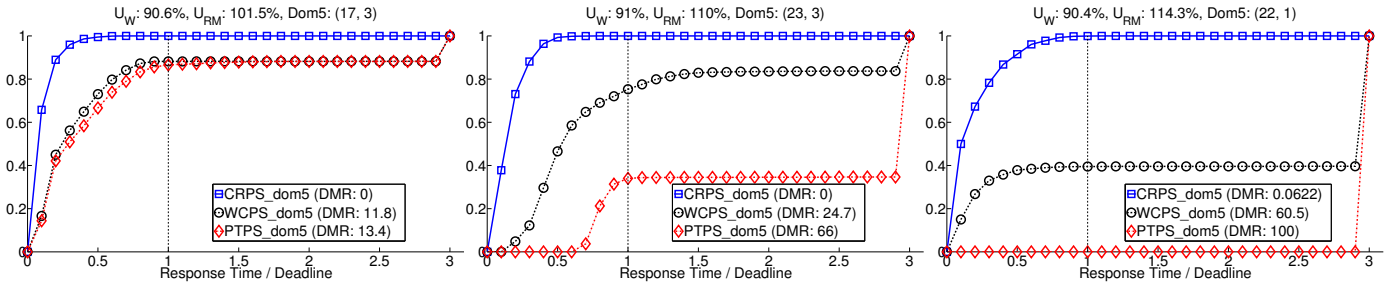
(a) Period: (550, 650), Overhead: 8.2% (b) Period: (100, 1100), Overhead: 19.6% (c) Period: (350, 850), Overhead: 14.9%

Fig. 8: CDF Plot of $\frac{ResponseTime}{Deadline}$ for Tasks in the Lowest Priority Domain under Designed $U_W = 0.8$ varying Task Periods



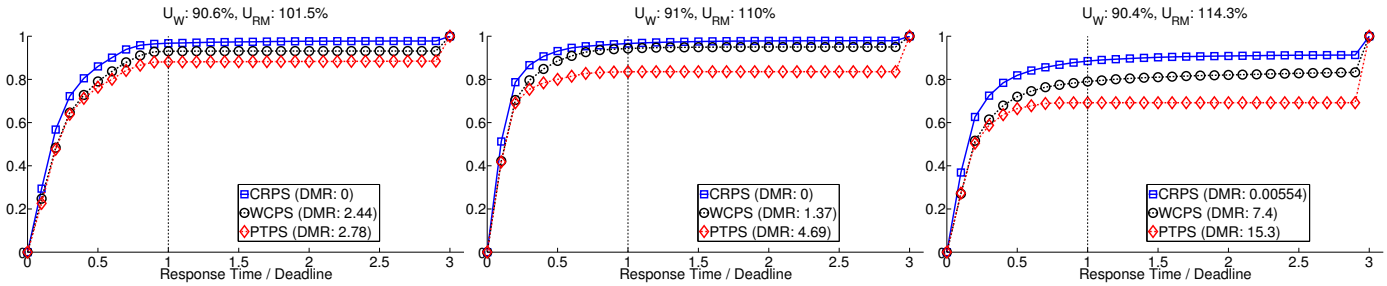
(a) Period: (550, 650), Overhead: 10.9% (b) Period: (100, 1100), Overhead: 19.1% (c) Period: (350, 850), Overhead: 24%

Fig. 9: CDF Plot of $\frac{ResponseTime}{Deadline}$ for All Tasks in Five Domains under Designed $U_W = 0.9$ varying Task Periods



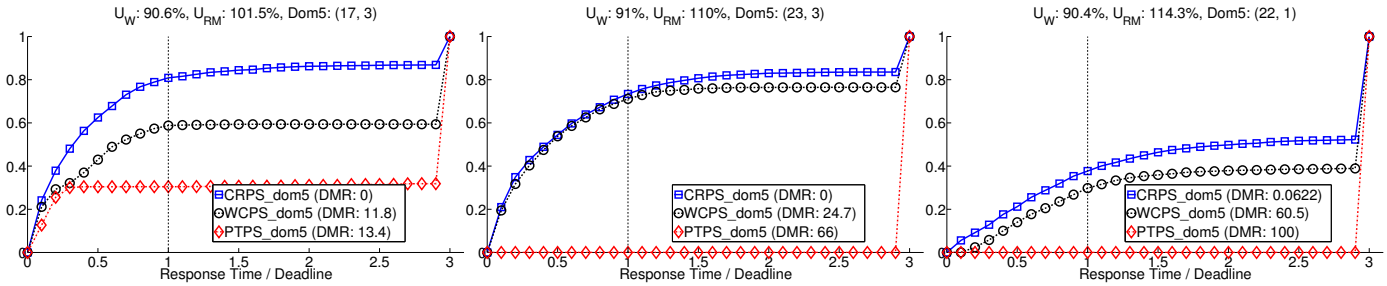
(a) Period: (550, 650), Overhead: 10.9% (b) Period: (100, 1100), Overhead: 19.1% (c) Period: (350, 850), Overhead: 24%

Fig. 10: CDF Plot of $\frac{ResponseTime}{Deadline}$ for Tasks in the Lowest Priority Domain under Designed $U_W = 0.9$ varying Task Periods



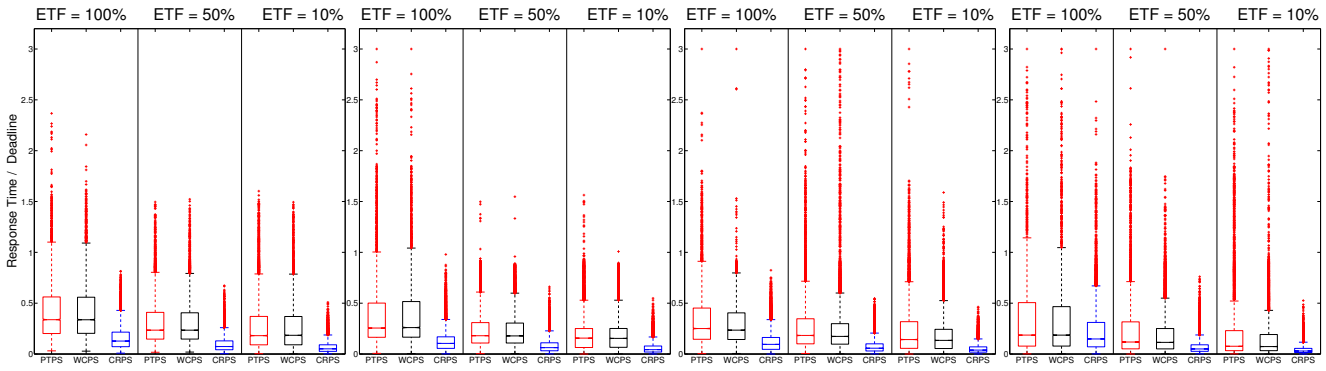
(a) Period: (550, 650), Overhead: 13% (b) Period: (100, 1100), Overhead: 20% (c) Period: (350, 850), Overhead: 26.5%

Fig. 11: CDF Plot of $\frac{ResponseTime}{Deadline}$ for All Tasks in Five Domains under Designed $U_W = 1.0$ varying Task Periods



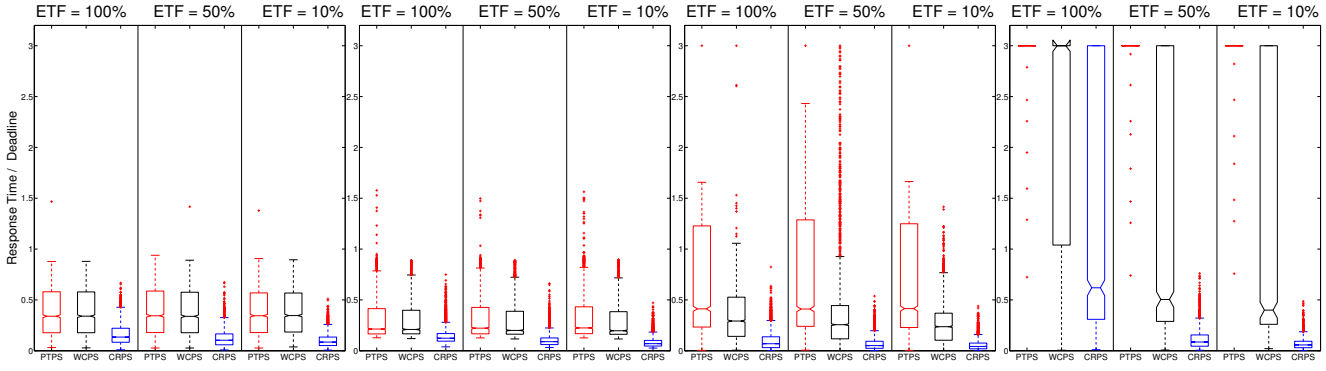
(a) Period: (550, 650), Overhead: 13% (b) Period: (100, 1100), Overhead: 20% (c) Period: (350, 850), Overhead: 26.5%

Fig. 12: CDF Plot of $\frac{ResponseTime}{Deadline}$ for Tasks in the Lowest Priority Domain under Designed $U_W = 1.0$ varying Task Periods



(a) $U_W = 0.703$ (b) $U_W = 0.832$ (c) $U_W = 0.91$ (d) $U_W = 1.021$

Fig. 13: Box Plot of $\frac{ResponseTime}{Deadline}$ for All Tasks under Different U_W and *ETF* Values



(a) $U_W = 0.703$ (b) $U_W = 0.832$ (c) $U_W = 0.91$ (d) $U_W = 1.021$

Fig. 14: Box Plot of $\frac{ResponseTime}{Deadline}$ for Tasks in the Lowest Priority Domain under Different U_W and *ETF* Values

demonstrates that the WCPS and CRPS benefit from the idle time introduced by interface overheads and over-estimations of jobs' execution times. The full results for all domains is shown in Figure 13.

C. ARINC-653 Workloads

In this section, we evaluate the performance of CSA implementation using ARINC-653 data sets obtained from an avionics system [10]. These data sets contain 7 harmonic workloads, each of which represents a set of domains (components) scheduled on a single processor, with each domain consisting of a set of periodic tasks. The descriptions of the workloads are available in the appendix of [5].

The evaluation goals are threefold: (1) to validate the effectiveness of the CSA implementation on real workloads; (2) to evaluate the relative performance of the PTPS, WCPS, and CRPS approaches under harmonic workloads and under different workload conditions; and (3) to quantify the impact of extra bandwidth available in the implemented interfaces (introduced by rounding up interface budgets, which are required in the implementation as the ARINC interface periods are fixed).

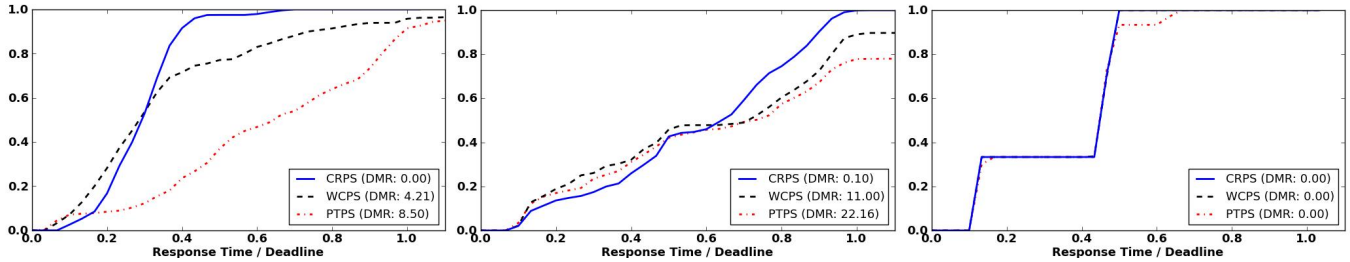
Implementation of ARINC domains. Table I shows the interface overheads introduced for all 7 workloads. Here, U_{RM} and U_W denote the total bandwidth of the interfaces and the total utilization of the tasks in the workload, respectively. $TotOv$ denotes the total overhead of the interface with respect to the tasks in terms of utilization, i.e., $TotOv = U_{RM} - U_W$.

W. ID	1	2	3	4	5	6	7
U_{RM}	0.460	0.570	0.560	0.450	0.583	0.465	0.020
U_W	0.378	0.511	0.481	0.389	0.537	0.426	0.003
$TotOv$	0.082	0.059	0.079	0.061	0.046	0.039	0.017

TABLE I: Interface Overheads in ARINC Workloads.

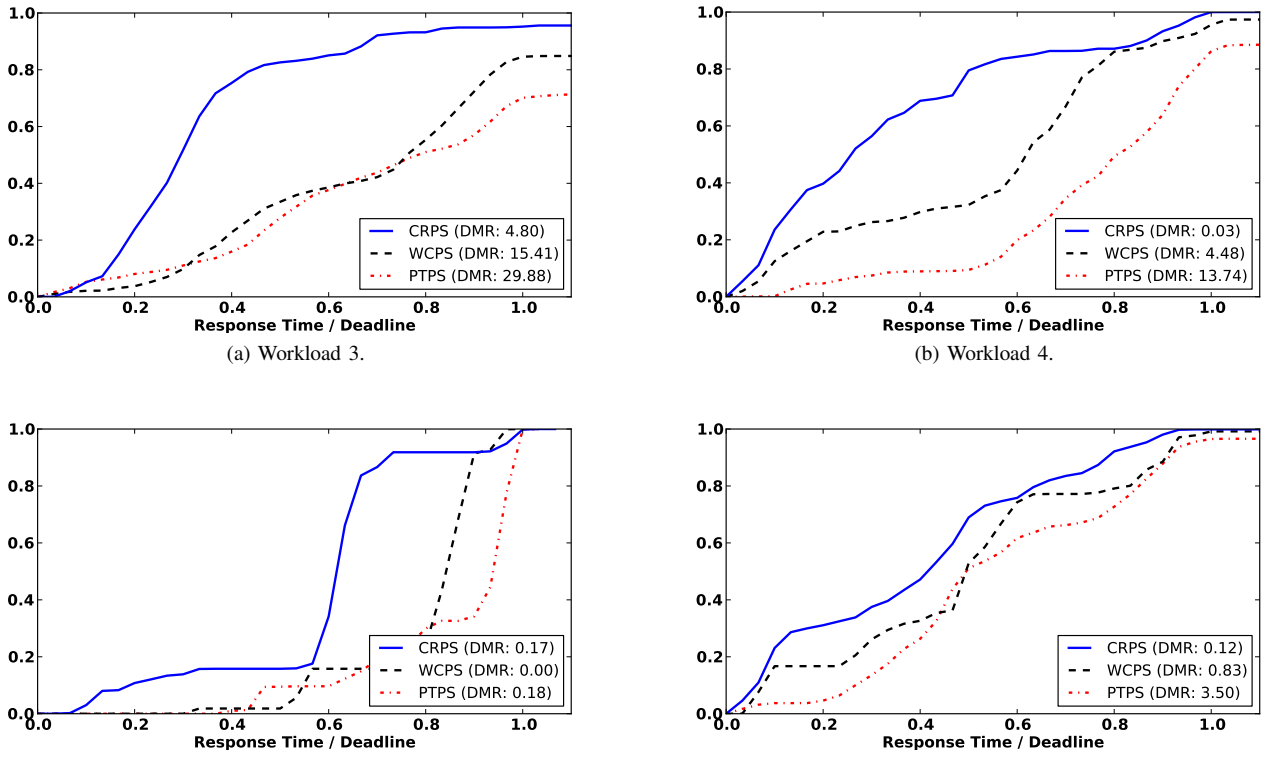
The obtained interfaces were implemented in CSA as periodic servers, following the PTPS, WCPS and CRPS server design approaches. We ran each workload for 10 one-minute runs. The obtained results are then averaged across the 10 runs.

Experimental results and observations. Figure 15 shows the responsiveness ($\frac{ResponseTime}{Deadline}$) distribution of the three server designs for three representative types of workloads: (a) having a high total interface overhead; (b) having a low total interface overhead; and (c) having a minimum number of domains. The CDF plots for all the workloads can be found in Figure 16.



(a) High Total Interface Overhead (W. 1). (b) Low Total Interface Overhead (W. 2). (c) Minimum Number of Domains (W. 7).

Fig. 15: CDF Plots of $\frac{ResponseTime}{Deadline}$ for Different Types of ARINC Workloads.



(a) Workload 3. (b) Workload 4.

(c) Workload 5. (d) Workload 6.

Fig. 16: CDF Plots of $\frac{ResponseTime}{Deadline}$ for Different Types of ARINC Workloads (W. 3- W.6)

We observed the following behaviors:

(1) The WCPS and CRPS approaches consistently outperformed the PTPS approach for all three types of workloads in terms of miss rates and responsiveness, as shown in Figure 15. Note that in this experiment, we did not include all sources of system overheads in the components' parameters; as a result, the computed interfaces cannot account for all system overheads, resulting in potential deadline misses.

(2) In terms of responsiveness, the CRPS approach achieves the greatest improvement over the PTPS and WCPS approaches when the total interface overhead is high, as is illustrated in Figure 15a. Conversely, when the total overhead is low, the responsiveness improvement is less, as is shown in Figure 15b. These behaviors can be explained by the fact that a higher total overhead potentially leads to more available residual-capacity that can be utilized by the scheduler.

(3) In terms of deadline miss rates, the CRPS approach improves significantly over the other two approaches regardless of the interface overheads, as is shown in both Figure 15a and 15b. For example, when the total overhead is high, the CRPS approach incurs no deadline misses whereas both PTPS and WCPS experience deadline misses. Similarly, when the total overhead is low, the PTPS (WCPS) approach incurs a miss rate of at least 200 times (100 times) more than the CRPS approach.

(4) As is illustrated in Figure 15c, the PTPS, WCPS, and CRPS approaches show similar distributions of responsiveness (with some small differences due to different system inter-

ference during different runs). This is expected because all approaches should behave identically if the workload contains a single domain, as is the case for Workload 7.

We also examined the performance of the PTPS approach with respect to the individual interface overheads of different domains within a component. Table II shows a typical example of the interface overhead versus deadline miss rate of different domains. It can be observed from the experimental results that, in general, a domain with a smaller interface overhead often incurs a higher miss rate and vice versa. However, the effect of interface overhead on the domain's miss rate is less prominent when using the CRPS approach. This is expected because in the CRPS approach, the domains with lower overheads (smaller extra budgets) are allowed to reclaim capacity from domains with higher overheads (larger extra budgets).

Dom. ID	2	4	6	3	5	1	Total
Overhead	0.000	0.002	0.004	0.006	0.012	0.035	0.059
DMR	0.844	0.400	0.459	0.000	0.141	0.001	0.222

TABLE II: Relation between Interface Overhead and Deadline Miss Ratio of PTPS in Workload 2.

V. Related Work

In terms of system architecture for compositional scheduling, only a few implementations exist and none of those considers Xen virtualization platform. For example, Behnam et al. [11] and Heuvel et al. [12] provided an implementation of a CSF on VxWorks and on μC/OS-II, respectively. However, neither approach considered virtualization. Recently, Yang et al. [13] developed a two-level CSF for virtualization using the L4/Fiasco. This work differs from ours in several aspects: (1) it builds on L4/Fiasco, which has a different system architecture than Xen; (2) it does not provide different work conserving enhancements to periodic server; and (3) its interface computation is not optimal as it assumes identical periods for all domains and is based on a lower-bound of the SBF instead of the actual SBF.

Hierarchical real-time scheduling frameworks (HSFs) for closed systems also have been implemented in different OS kernels (e.g., [14]–[18]). These approaches, however, are non-compositional. Further, they implement all levels of the scheduling hierarchy within the same operating system. HSF implementations through virtualization also have been explored lately. For instance, [19] proposed a bare VMM which uses virtualization and dedicated device techniques with a fixed cyclic scheduling policy. Cucinotta et al. [20] used the KVM with a hard reservation behavior variant of the Constant Bandwidth Server (CBS). Our work is different from these in that (1) the architecture supports compositional scheduling, which they do not; and (2) ours builds on Xen, which has a very different system architecture from their virtualization platforms.

In terms of server designs, the general idea of 'capacity reclaiming' has been explored earlier in other contexts. For instance, Lehoczky et al. [21] provided a 'slack stealing' algorithm that allows aperiodic tasks to steal slack from periodic tasks. Caccamo et al. [22] and Nogueira et al. [23] provided CBS algorithms that allow one server to 'steal' another server's budget under EDF scheduling. These approaches, however, do not support compositional scheduling. In addition, their 'reclaiming capacity' includes only idle budget due to an overestimation of tasks' execution times, whereas ours includes the idle budget due to interface overheads as well.

In terms of theoretical computation of server parameters for quantum-based platforms, the only existing technique we are aware of was developed by Yoo et al. [24]. That work, however, assumes a manually chosen bound on the server period, which cannot guarantee an optimal resource period. In this paper, we provide a method for computing the maximum optimal server period, thus avoiding such non-optimality.

VI. Conclusion

In this paper we have presented CSA, an architecture platform with system support for compositional scheduling of real-time systems. CSA realizes the key concepts and important results of a PRM-based CSF within the Xen virtualization platform, bringing the benefits of existing CSF theory to practical application. We discuss several challenges faced in the development of CSA, and propose theoretical extensions and server design enhancements to address these challenges. We also present an extensive evaluation to demonstrate the utility and effectiveness of CSA in optimizing real-time performance. Our implementation provides a number of scheduling policies; at the same time, it is modular and easily extensible with new server-based scheduling algorithms. CSA is released as open-source and is available at
http://sites.google.com/site/realtimexen.

An important direction of future work is to extend CSA to support compositional multicore processor scheduling and dependent tasks, which undoubtedly will present additional challenges.

References

[1] I. Shin and I. Lee, "Compositional Real-time Scheduling Framework with Periodic Model," *ACM Transactions on Embedded Computing Systems (TECS)*, 2008.

[2] A. Easwaran, M. Anand, and I. Lee, "Compositional Analysis Framework Using EDP Resource Models," in *RTSS*, 2007.

[3] L. Sha, J. P. Lehoczky, and R. Rajkumar, "Solutions for Some Practical Problems in Prioritized Preemptive Scheduling," in *RTSS*, 1986.

[4] A. Easwaran, I. Lee, I. Shin, and O. Sokolsky, "Compositional Schedulability Analysis of Hierarchical Real-Time Systems," in *ISORC*, 2007.

[5] A. Easwaran, I. Lee, O. Sokolsky, and S. Vestal, "A Compositional Framework for Avionics (ARINC-653) Systems," *Tech Report MS-CIS-09-04*, 2009, University of Pennsylvania. [Online]. Available: http://repository.upenn.edu/cis_reports/898

[6] J. Lehoczky, L. Sha, and Y. Ding, "The Rate Monotonic Scheduling Algorithm: Exact Characterization and Average Case Behavior," in *RTSS*, 1989.

[7] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, "Xen and the Art of Virtualization," in *SOSP*, 2003.

[8] S. Xi, J. Wilson, C. Lu, and C. Gill, "RT-Xen: Real-Time Virtualization Based on Hierarchical Scheduling," in *EMSOFT*, 2011.

[9] S. Xi, J. Lee, C. Lu, C. D. Gill, S. Chen, L. T. X. Phan, O. Sokolsky, and I. Lee, "RT-Xen: Real-Time Virtualization Based on Hierarchical Scheduling," http://sites.google.com/site/realtimexen/.
[10] Avionics Electronic Engineering Commitee(ARINC), "Avionics Application Software Standard Interface: Part 1 - required services (ARINC specification 653-2)," in *TechReport*, 2006.
[11] M. Behnam, T. Nolte, I. Shin, M. Åsberg, and R. J. Bril, "Towards Hierarchical Scheduling on Top of VxWorks," in *OSPERT*, 2008.
[12] M. M. van den Heuvel, R. J. Bril, J. J. Likkien, and M. Behnam, "Extending a HSF-enabled Open-source Real-time Operating System with Resource Sharing," in *OSPERT*, 2010.
[13] J. Yang, H. Kim, S. Park, C. Hong, and I. Shin, "Implementation of Compositional Scheduling Framework on Virtualization," *SIGBED Rev.*, 2011.
[14] J. Regehr and J. Stankovic, "HLS: A Framework for Composing Soft Real-Time Schedulers," in *RTSS*, 2001.
[15] M. Danish, Y. Li, and R. West, "Virtual-CPU Scheduling in the Quest Operating System," in *RTAS*, 2011.
[16] Y. Wang and K. Lin, "The Implementation of Hierarchical Schedulers in the RED-Linux Scheduling Framework," in *ECRTS*, 2000.
[17] G. Parmer and R. West, "HIRES: a System for Predictable Hierarchical Resource Management," in *RTAS*, 2011.
[18] T. Aswathanarayana, D. Niehaus, V. Subramonian, and C. Gill, "Design and Performance of Configurable Endsystem Scheduling Mechanisms," in *RTAS*, 2005.
[19] A. Crespo, I. Ripoll, and M. Masmano, "Partitioned Embedded Architecture Based on Hypervisor: the XtratuM Approach," in *EDCC*, 2010.
[20] T. Cucinotta, G. Anastasi, and L. Abeni, "Respecting Temporal Constraints in Virtualised Services," in *COMPSAC*, 2009.
[21] J. Lehoczky and S. Ramos-Thuel, "An Optimal Algorithm for Scheduling Soft-aperiodic Tasks in Fixed-priority Preemptive Systems," in *RTSS*, 1992.
[22] M. Caccamo, G. Buttazzo, and D. Thomas, "Efficient Reclaiming in Reservation-based Real-time Systems with Variable Execution Times," *IEEE Transactions on Computers*, 2005.
[23] L. Nogueira and L. Pinho, "Capacity Sharing and Stealing in Dynamic Server-based Real-time Systems," in *IPDPS*, 2007.
[24] S. Yoo, Y.-P. Kim, and C. Yoo, "Real-time Scheduling in a Virtualized CE Device," in *ICCE*, 2010.

APPENDIX

In this section, we provide the complete proofs for Lemmas 3, 4, and 5.

Proof of Lemma 3: We will prove the lemma by contradiction. Suppose there exists $t' \notin \mathsf{CrT}_W$ such that $\mathsf{usbf}_\Gamma(t') = \mathsf{rbf}_{W,i}(t')$. Let $s = \Pi - \Theta$. Then, by definition $\mathsf{CrT}_{W,i}$,

$$\exists\, t_0 \in \mathsf{CrT}_{W,i} :\ \frac{\mathsf{rbf}_{W,i}(t_0)}{t_0 - s} < \frac{\mathsf{rbf}_{W,i}(t')}{t' - s} \tag{10}$$

On the other hand, since $\forall t \neq t', \mathsf{usbf}_{\Gamma_i}(t) < \mathsf{rbf}_{W,i}(t)$ and $t_0 \neq t'$,

$$\begin{aligned} & \mathsf{usbf}_\Gamma(t_0) < \mathsf{rbf}_{W,i}(t_0) \\ \Rightarrow & \frac{\Theta}{\Pi}(t_0 - (\Pi - \Theta)) < \mathsf{rbf}_{W,i}(t_0) \\ \Rightarrow & \frac{\mathsf{rbf}_{W,i}(t')}{t' - s}(t_0 - s) < \mathsf{rbf}_{W,i}(t_0) \\ \Rightarrow & \frac{\mathsf{rbf}_{W,i}(t')}{t' - s} < \frac{\mathsf{rbf}_{W,i}(t_0)}{t_0 - s} \end{aligned} \tag{11}$$

Since Eq. (11) contradicts Eq. (10), the lemma holds. ∎

Proof of Lemma 4: Let r_t be $\mathsf{rbf}_{W,i}(t)$. We have:

$$\begin{aligned} & \mathsf{usbf}_{\Gamma^*}(t) = \mathsf{rbf}_{W,i}(t) \\ \Leftrightarrow\ & \frac{\Theta^*}{\Pi}(t - (\Pi - \Theta^*)) = r_t \\ \Leftrightarrow\ & (\Theta^*)^2 + \Theta^*(t - \Pi) - \Pi \cdot r_t = 0 \end{aligned}$$

Since $\Theta^* \geq 0$, the above equation has a unique solution:

$$\Theta^* = \frac{-(t - \Pi) + \sqrt{(t - \Pi)^2 + 4\Pi \cdot r_t}}{2}.$$

As a result, the bandwidth of Γ^* is $B^i_{\min}(\Pi, t) = \frac{\Theta^*}{\Pi}$. Hence the lemma. ∎

Proof of Lemma 5: Let $r_t = \mathsf{rbf}_{w,i}(t)$. Since $\frac{dB^i_{\min}(\Pi,t)}{d\Pi} \geq 0$ implies $B^i_{\min}(\Pi, t)$ is increasing, we would like to show

$$\begin{aligned} & \frac{dB^i_{\min}(\Pi, t)}{d\Pi} \geq 0 \\ \Leftrightarrow\ & \Big(\frac{\Pi - t + \sqrt{(\Pi - t)^2 + 4r_t\Pi}}{\Pi}\Big)' \geq 0 \\ \Leftrightarrow\ & 0 + \frac{t}{\Pi^2} + \frac{1}{2}\frac{(\Pi - t)^2 + 4r_t\Pi}{\Pi^2}^{(-\frac{1}{2})}\Big(\frac{(\Pi - t)^2 + 4r_t\Pi}{\Pi^2}\Big)' \geq 0 \\ \Leftrightarrow\ & \frac{t}{\Pi^2} + \frac{1}{2}\sqrt{\frac{\Pi^2}{(\Pi - t)^2 + 4r_t\Pi}}\Big(\frac{(2t - 4r_t)\Pi - 2 \cdot t^2}{\Pi^3}\Big) \geq 0 \\ \Leftrightarrow\ & 2t \cdot \Pi + \sqrt{\frac{\Pi^2}{(\Pi - t)^2 + 4r_t\Pi}}\big((2t - 4r_t)\Pi - 2 \cdot t^2\big) \geq 0 \end{aligned}$$

$$\begin{aligned} \Leftrightarrow\ & \sqrt{\frac{\Pi^2}{(\Pi - t)^2 + 4r_t\Pi}} \\ & \Big(\Big(\frac{\Pi^2}{(\Pi - t)^2 + 4r_t\Pi}\Big)^{(-\frac{1}{2})} 2t \cdot \Pi + (2t - 4r_t)\Pi - 2 \cdot t^2\Big) \geq 0 \\ \Leftrightarrow\ & \sqrt{\frac{(\Pi - t)^2 + 4r_t\Pi}{\Pi^2}} 2t \cdot \Pi + (2t - 4r_t)\Pi - 2 \cdot t^2 \geq 0 \\ & \qquad \textit{by dividing } \sqrt{\frac{\Pi^2}{(\Pi - t)^2 + 4r_t\Pi}} \geq 0 \\ \Leftrightarrow\ & \sqrt{\frac{(\Pi - t)^2 + 4r_t\Pi}{\Pi^2}} 2t \cdot \Pi + (2t - 4r_t)\Pi - 2 \cdot t^2 \geq 0 \\ \Leftrightarrow\ & \sqrt{\frac{(\Pi - t)^2 + 4r_t\Pi}{\Pi^2}} \geq \frac{t^2 - (t - 2r_t)\Pi}{t \cdot \Pi} \\ \Leftrightarrow\ & \frac{(\Pi - t)^2 + 4r_t\Pi}{\Pi^2} \geq \Big(\frac{t^2 - (t - 2r_t)\Pi}{t \cdot \Pi}\Big)^2 \end{aligned}$$

by applying square in both sides (since the left side≥ 0, the equation trivailly holds if the right side< 0).

$$\begin{aligned} \Leftrightarrow\ & t^2(\Pi - t)^2 + t^2 \cdot 4r_t\Pi \geq \big(t^2 - (t - 2r_t)\Pi\big)^2 \\ \Leftrightarrow\ & 4r_t \cdot (t - r_t) \geq 0 \end{aligned}$$

which is obvious since $t \geq r_t = \mathsf{rbf}_{W,i}(t)$. ∎