On the Defectiveness of SCHED_DEADLINE w.r.t. Tardiness and Affinities, and a Partial Fix

Stephen Tang
James H. Anderson
{syting, anderson}@cs.unc.edu
University of North Carolina at Chapel Hill
Chapel Hill, USA

Luca Abeni
luca.abeni@santannapisa.it
Scuola Superiore Sant’Anna
Pisa, Italy

ABSTRACT

SCHED_DEADLINE (DL for short) is an Earliest-Deadline-First (EDF) scheduler included in the Linux kernel. A question motivated by DL is how EDF should be implemented in the presence of CPU affinities to maintain optimal bounded tardiness guarantees. Recent works have shown that under arbitrary affinities, DL does not maintain such guarantees. Such works have also shown that repairing DL to maintain these guarantees would likely require an impractical overhaul of the existing code. In this work, we show that for the special case where affinities are semi-partitioned, DL can be modified to maintain tardiness guarantees with minor changes. We also draw attention to the fact that admission control is already broken in several respects in the existing DL implementation.

1 INTRODUCTION

SCHED_DEADLINE (DL for short) is an Earliest-Deadline-First (EDF) implementation included in the mainline Linux kernel since version 3.14. Its inclusion has been significant because it has lowered the barrier to entry for real-time EDF scheduling and it has inspired many publications [4, 8, 9, 11, 13, 15, 16].

Admission control. One feature of DL is its admission-control (AC) system. AC prevents the system from being over-utilized by tasks managed by DL. If the total CPU bandwidth required by DL tasks is near some threshold, then AC will reject any requests to create new DL tasks until existing DL tasks reduce their bandwidth consumption. Note that preventing over-utilization is only a necessary condition for guaranteeing that tasks meet deadlines.

Instead of guaranteeing that deadlines will be met, AC satisfies two goals [3]. The first goal is to provide performance guarantees. Specifically, AC guarantees to each DL task that its long-run consumption of CPU bandwidth remains within a bounded margin of error from a requested rate. The second goal is to avoid the starvation of lower-priority non-DL workloads. Specifically, AC guarantees that some user-specified amount of CPU bandwidth will remain for the execution of non-DL workloads. This goal of AC has been an important aspect of Linux real-time scheduling for years, and predates the merging of DL into the kernel.

The theoretical basis of AC is that global EDF scheduling of sporadic real-time tasks on identical multiprocessors guarantees bounded tardiness to jobs if the system is not over-utilized [6]. The DL documentation [2] cites [6]. Bounded tardiness ensures AC’s first goal as it implies that any execution of a job occurs in a finite window around its release and deadline, meaning a task’s rate of execution stays consistent with its bandwidth. Bounded tardiness also prevents any set of DL tasks from having an unbounded number of tardy jobs which, if allowed to execute uninterrupted, could starve non-DL workloads for an unbounded amount of time. Thus, bounded tardiness also supports AC’s second goal.

Affinities. Specifying tasks’ affinities, subsets of the CPUs on which given tasks are permitted to execute, is useful for maintaining cache locality, as tasks whose execution times depend heavily on whether they are cache hot or cold at a certain cache level may benefit from having their affinities restricted to CPUs that share cache at said level. Affinities can also be useful in reducing scheduling-overheads, as a CPU need only consider tasks with affinity for that CPU when deciding what to execute.

Unfortunately, the tardiness result in [6] heavily relies on the fact that scheduling is global, and does not hold when tasks have affinities. Thus, with the exception of clustered scheduling (in which each cluster resembles its own global subsystem), the setting of affinities is forbidden in DL unless AC is explicitly disabled.

It was shown in [15] that with arbitrary affinities, under-utilized systems (whose definition becomes more complex under arbitrary affinities) may have unbounded tardiness under DL. Though an EDF variant called SAPA-EDF presented in [5] was proven in [16] to guarantee bounded tardiness with arbitrary affinities for identical CPUs, this variant was not implemented by the DL maintainers due to the complexity of the required modifications to both the scheduler and AC (as the meaning of over-utilization is changed).
This variant would likely have also required higher overheads and increased task migrations in practice.

While [15] showed that the current DL implementation cannot support arbitrary affinities, that work suggested that this may not be the case for special cases of affinities. In this paper, we consider the special case of semi-partitioned (SP) systems, wherein each task has affinity for either one CPU (the task is partitioned) or for all CPUs in its cluster (the task is migrating). While not as flexible as arbitrary affinities, SP affinities still reduce overheads by limiting the number of migrating tasks and can still be used to maintain cache locality for partitioned tasks.

**In practice.** In practice, AC does not guarantee bounded tardiness under DL, even under an idealized abstraction of the scheduler. This is because affinities are far from the only way that DL differs from the system model considered in [6], from which the soundness of AC originates. In particular, DL has grown to consider dynamic voltage and frequency scaling (DVFS) and asymmetric CPU capacities, both of which break the assumption of identical CPUs in [6]. Besides these changes in the considered platform model, tasks can also be much more dynamic under DL than in [6].

Unlike affinities, AC need not be disabled when using such features. This calls into question the role of AC, as there is no theoretical basis for its goals without tardiness guarantees. Instead of by analysis, these features have often been validated empirically by demonstrating that deadline-miss frequencies are acceptable. As we demonstrate herein, empirical validation is insufficient for showing that AC is not broken.

**Contributions.** We present a DL variant that supports AC for SP systems. This contribution is divided into three parts.

First, we list features supported by DL that were not considered in the analysis in [6]. For a subset of these features, we show that usage of these features in the existing DL implementation can cause unbounded tardiness with AC. For the remaining features, though usage may not cause unbounded tardiness with AC, we demonstrate that usage can lead to other undesirable effects, detailed later.

Second, we show that the existing DL implementation is broken under SP affinities in the sense that unbounded tardiness may result in under-utilized systems. We present a patch that modifies DL’s migration logic and AC such that tardiness is bounded for all such systems. This patch is intended for Linux 5.4.69, as 5.4 was the most recent LTS release at time of writing; we generally refer to this version as the current implementation unless specified otherwise, though the details discussed in this work do not seem to have changed with recent releases. This patch was designed to be as minimally invasive as possible such that the development of existing features would not be hindered by including this patch.

Third, we provide a high-level overview of our proof of soundness for our modified AC under an abstraction of our patched DL (due to space constraints, the full proof is deferred to App. A, available online [14]). Our proof is based on existing proof techniques from [16]. Our proof modifies these techniques because SAPA-EDF satisfies an invariant that our patched DL does not. Our patched scheduler maintains a weaker invariant that we show is sufficient for bounded tardiness under SP scheduling. Our patched DL compromises between schedulers proposed in theory, which migrate tasks aggressively, and DL, which is unable to make guarantees.

**Organization.** The rest of this paper is organized as follows. After covering needed background in Sec. 2, we present and categorize the list of DL features not considered by [6] in Sec. 3. We demonstrate the non-optimality of DL under SP affinities and present our patch in Sec. 4. We provide an overview of the proof of correctness of this patch under an idealized abstraction of our patched DL in Sec. 5 (again, the formal proof is deferred to an appendix). We evaluate the overheads of our patch relative to the original implementation in Sec. 6. We conclude in Sec. 7.

## 2 BACKGROUND

We start by presenting our system model, mostly derived from [15] with some modifications to consider dynamic task systems. We then discuss how DL fits this system model.

### 2.1 Task Model

We consider a system of $N$ implicit-deadline sporadic tasks $\tau = \{\tau_1, \tau_2, \ldots, \tau_N\}$ running on $M$ unit-speed CPUs $\pi = \{\pi_1, \pi_2, \ldots, \pi_M\}$. We assume basic familiarity with the sporadic task model. We denote the $j$th job released by task $\tau_i$ as $\tau_{ij}$, where $j \geq 1$. Job $\tau_{ij}$ must be completed before job $\tau_{i,j+1}$ is allowed to execute. We let $C_i$ denote the worst-case execution time (WCET) of $\tau_i$ over all its jobs. We let $T_i$ denote the period of task $\tau_i$. The utilization $u_i$ of task $\tau_i$ is given by $C_i/T_i$. We denote the release time, deadline, and completion time of job $\tau_{ij}$ by $r_{ij}$, $\hat{d}_{ij}$, and $\hat{f}_{ij}$, respectively, where $\hat{d}_{ij} = r_{ij} + T_i$ (implicit deadlines).

At time $t$, a job $\tau_{ij}$ is either unreleased ($t < r_{ij}$), pending ($r_{ij} \leq t < \hat{d}_{ij}$), or complete ($t \geq \hat{f}_{ij}$). If a task $\tau_i$ has pending jobs at $t$, then its ready job at $t$ is its earliest-released pending job at $t$.

For a job $\tau_{ij}$, its response time is given by $\hat{f}_{ij} - r_{ij}$, and its tardiness by $\max(0, \hat{d}_{ij} - \hat{f}_{ij})$. The tardiness of task $\tau_i$ is the supremum of the tardiness of its jobs. If the tardiness of all tasks in $\tau$ is bounded under a given scheduler, then $\tau$ is soft real-time (SRT)-schedulable under that scheduler. $\tau$ is SRT-feasible if it is SRT-schedulable under some scheduler. A scheduler is SRT-optimal if any SRT-feasible task system is SRT-schedulable under it. Analogous terms apply to hard real-time (HRT) systems, in which no tasks can be tardy. Feasible without a prefix denotes both SRT- and HRT-feasible.

At any time, a task is either active or inactive. A task is initially inactive, and must become active in order to release jobs. Each task becomes active at most once, and after becoming inactive, remains so indefinitely (this is for simplicity of notation, as a task becoming active a second time is equivalent to an unrelated task becoming active). At time $t$, the set of active tasks is denoted $\tau(t) \subseteq \tau$, and the set of CPUs they may execute upon is denoted $\pi(t) \subseteq \pi$. The set of tasks partitioned onto CPU $\pi_j$ is denoted $\tau(t, \pi_j) \subseteq \tau(t)$.

These functions are defined such that any changes that occur at time instant $t$ are reflected in $\tau(t)$, $\pi(t)$, and $\pi(t, \pi_j)$.

### 2.2 From Threads to Tasks

Linux threads may not adhere to the sporadic task model. Thus, any Linux thread under DL is encapsulated in a Constant Bandwidth Server (CBS). A CBS is conceptually similar to a sporadic task, with a maximum budget $Q$, period $T$, and bandwidth $Q/T$ that are

---

1Note that "active" and "inactive" in this work are distinct from their definitions in DL’s documentation and code.
With this assumption, a CBS behaves identically to a sporadic task, \( q \) was not considered in [6] and [16]. If at the time \( t \) a CBS wakes, thereby removing the dissimilarities between CBSs and sporadic tasks. Because they have identical semantics under this assumption, we refer to CBSs and their encapsulated threads as tasks unless specified otherwise.

2.3 CPUSets and AC

A task’s WCET, period, and relative deadline are provided as arguments when it first enters DL from another scheduling class via the `sched_setattr` system call.

DL schedules tasks using clustered EDF, wherein each task assigned to a cluster is only permitted to migrate between CPUs within said cluster. Both global and partitioned EDF are special cases of clustered EDF. In Linux, clusters are created and managed via the cpuset subsystem under the cgroup virtual file system (a `cpuset` is Linux’s notion of a cluster\(^2\)). Tasks and CPUs are added to and removed from a cpuset dynamically with the cpuset subsystem. \( \tau(i) \) and \( \pi(i) \) in Sec. 2.1 are abstractions for the set of DL tasks and CPUs assigned to a cpuset by the cpuset subsystem.

AC prevents the starvation of non-DL workloads and guarantees bounded tardiness to DL tasks within a cpuset. Generally, AC maintains the invariant that for any cpuset,

\[
\forall \tau: \sum_{t \leq \tau(i)} u_i \leq \frac{\text{sched\_rt\_runtime\_us}}{\text{sched\_rt\_period\_us}}[\pi(i)],
\]

where `sched\_rt\_runtime\_us` and `sched\_rt\_period\_us`\(^4\) are set with the proc virtual file system and have default values of 950000 and 100000, respectively. `sched\_rt\_runtime\_us` should always be set to be less than `sched\_rt\_period\_us`. If so, (1) guarantees non-DL workloads in the cpuset are given time to execute because the system is not over-utilized [6], though this guarantee may be invalid for reasons discussed in Sec. 3.

AC must verify that (1) is maintained on changes to \( \tau(i) \) (tasks entering DL), changes to \( \pi(i) \) (changes to the CPUs in the cpuset), and changes to `sched\_rt\_period\_us` or `sched\_rt\_runtime\_us`. Requests for changes fail if doing so would violate (1) for any cpuset. These changes occur immediately otherwise.

The summation in (1) must be decreased when a thread leaves DL. Though the thread may leave immediately upon request, the server must remain until an event called its 0-lag time passes. This implementation detail exists to support HRT guarantees. As we are concerned with SRT, we do not consider this rule in our model.

2.4 Fine-Grained Affinities

In Linux, the `sched\_setaffinity` system call is an auxiliary method to the cpuset subsystem of setting affinities. `sched\_setaffinity` differs to the cpuset subsystem in that calls that contain CPUs outside of a task’s cpuset will ignore said CPUs. Additionally, when \( \pi(i) \) is modified, all tasks in the cpuset gain affinity for all CPUs in \( \pi(i) \). This overwrites any changes made with `sched\_setaffinity`, even if the modified set of CPUs is a superset of the original.

The usage of `sched\_setaffinity` is not permitted for DL tasks unless AC is disabled by writing -1 to `sched\_rt\_runtime\_us`. AC must be disabled as bounded tardiness to tasks is not guaranteed under arbitrary affinities.

\(^2\)Though the cpuset subsystem actually organizes cpusets hierarchically. DL requires that cpusets with DL tasks to be exclusive from each other, which we also assume.

\(^3\)These file names derive from `SCHED_RT`, whose accounting code DL intertwines with. We assume no RT tasks to focus on DL.
The scripts that create these traces are available online [14].

Example never suspend. Consider a cpuset containing one CPU $\pi_1$ and begins with no tasks. At time $t = 0$, task $r_1$ requests to enter DL with $(C_1, T_1) = (94, 100)$. This task is accepted because doing so will not violate (1). $r_1$ releases its first job at time $0$ with $C_{1,1} = 94$ and $d_{1,1} = 100$, before immediately requesting $C_1$ be changed to 1. $C_{1,1}$ and $d_{1,1}$ remain as 94 and 100 after this change. At time $t = 10$, task $r_2$ requests to enter with $(C_2, T_2) = (94, 100)$. This request is accepted by AC because $r_1$ reduced its utilization. $r_2$ releases its first job with $C_{2,1} = 94$ and $d_{2,1} = 110$. After this point, both tasks release jobs periodically.

However, prior to when $r_1$ returns from the throttled state (see Fig. 1), $r_2$ changes $C_2$ to 1 and $r_1$ changes $C_1$ to 94. When the next job of $r_1$ becomes ready at $t = 100$, this job’s budget and deadline are determined entirely by $r_1$’s parameters at the instant it becomes ready. Thus, the budget of the next job is set to 94. Likewise, prior to when $r_2$ is replenished at $t = 188$ (recall from Fig. 1 that a tardy task is replenished once its prior job completes), $r_1$ sets $C_1$ to 1 such that $r_2$ can set $C_2$ to 94. As the total utilization of the system technically never exceeds 0.95, all requests are accepted by AC.

If $r_1$ and $r_2$ continue taking turns with having $C = 94$, then every released job in this system has parameters as if both tasks always had utilizations of 0.94. Because the CPU only has a capacity of 1.0, the system is over-utilized and results in unbounded tardiness even though (1) is never violated.

**Defect 1. Dynamic tasks can break AC.**

**Mitigation 1.** We reject requests to modify DL tasks’ parameters. Any change in parameters after server creation must be implemented by leaving DL and reentering with a fresh server.

**DVFS.** The goal of DVFS is to reduce power consumption by scaling down CPU frequency. DVFS affects DL because DL must ensure that frequencies remain high enough that some level of real-time performance guarantee is maintained for tasks. DL must also account for reduced CPU frequency when depleting tasks’ budgets. This is managed in DL with the GRUB-PA system.

We refer to [13] for a full description of GRUB-PA, which is out of the scope of this work. At a high level, at all times, GRUB-PA tracks the total bandwidth of DL tasks on each CPU’s runqueue in a variable denoted $U_{\text{act}}(t)$. If $U_{\text{act}}(t) < 1$, then the frequency of the corresponding CPU is scaled by $U_{\text{act}}(t)$ of its maximum frequency and the budget of any task executing on said CPU is depleted at rate $U_{\text{act}}(t)$ (tasks’ WCETs are assumed to be provisioned based on CPUs running at maximum frequency). Otherwise, the CPU runs with its maximum frequency and the budget of any executing task is depleted at the standard unit rate. This can be exploited as follows to result in unbounded tardiness.

**Ex. 2.** Consider a cpuset with two CPUs and three tasks with $(C, T) = (63, 100)$. This system is accepted by AC because the total utilization is 1.89 which is less than 95% (recall the default values in (1)) of the capacity of 2 CPUs. However, because any task can be on at most one CPU’s runqueue at any time, there is always a CPU with $U_{\text{act}} \leq 0.63$. Because CPU frequencies are scaled by $U_{\text{act}}$ and the other CPU is unable to scale frequency beyond its maximum, the actual capacity delivered by the CPUs to DL tasks is at most $1.63 < 1.89$, the total utilization. Because the system is over-utilized, at least one task experiences unbounded tardiness.

**Defect 2. DVFS causes unbounded tardiness in DL with AC.**

**Mitigation 2.** We require that DVFS be disabled for AC.

GRUB-PA is the latest entry in a long line of GRUB variants implemented in DL. The code for M-GRUB, the GRUB variant used before GRUB-PA, actually still remains in the current DL implementation and can be triggered with the `SCHED_FLAG_RECLAIM` flag, though we have not considered it in this work because it is incompatible with the more recent GRUB-PA and DVFS. Nevertheless, it has also not been proven that M-GRUB does not violate bounded tardiness under AC, though we conjecture that this is the case.

**Asymmetric capacities.** Support for asymmetric CPU capacities, in which different CPUs complete work at different rates, was added...
to DL in Linux 5.9. The fastest CPU in the system is assigned a capacity\(^2\) of 1.0 and all other CPUs are assigned capacities relative to the fastest CPU. Let \(\text{cap}(\pi_t)\) denote the capacity of CPU \(\pi_t\). AC was also modified in 5.9 such that the r.h.s. of (1) considers \(\Sigma_{\pi_t \in \pi(t)} \text{cap}(\pi_t)\) instead of \(|\pi(t)|\).

In the real-time literature, a large body of work considers platforms with asymmetric-capacity CPUs. In these works, such platforms are often called uniform multiprocessors. One such work is [18], which presents an SRT-optimal EDF variant for uniform multiprocessors. DL’s capacity-aware version of (1) admits a superset of the systems accepted by the feasibility condition in [18], proving that AC permits systems with unbounded tardiness.

**Defect 3.** Asymmetric capacities can cause AC to admit task systems with unbounded tardiness.

**Mitigation 3.** We assume CPUs are identical.

### 3.2 Features that Negatively Affect Tardiness

For the following features, we conjecture that their usage does not break bounded tardiness under AC. Nevertheless, though these features may not technically break AC, they can have undesirable effects on tardiness. We chose to omit these features from our system model due to such negative effects and difficulties with the complexities caused by considering them in proofs.

**Inter-cpuset migration.** A dynamic behavior of DL not considered in [6] is the migration of a task from one cpuset to another. Such a migration will be accepted by AC so long as doing so will not violate (1) for the target cpuset. When the migration occurs, the task’s utilization is immediately subtracted from its original cpuset and added to that of its target. The server may keep its prior budget and deadline after the migration. As far as we are aware, it is not proven whether or not such migrations can result in unbounded tardiness, though we conjecture that it does not.

Even if such migrations between cpusets do not result in tardiness becoming unbounded, such migrations can cause higher tardiness than possible when tasks’ cpusets are static.

**Ex. 3.** (Corresponds with Fig. 3.) Consider two cpusets, one with a single CPU \(\pi_1\) and a task \(\tau_5\) with \((C_5, T_5) = (2, 10)\) and another with three CPUs \(\pi_2, \pi_3,\) and \(\pi_4\) and four tasks \(\tau_1, \tau_2, \tau_3,\) and \(\tau_4\) all with \((C, T) = (70, 100)\).

At time \(t = 92\), \(\tau_4\) requests to migrate to the cpuset containing \(\pi_1\) and \(\tau_5\). This is accepted by AC as \(\tau_4\) and \(\tau_5\)’s combined utilization is less than 0.95. However, \(\tau_4\) keeps its current deadline and budget when it migrates, thereby causing \(\tau_5\) to become tardy when it otherwise would not have been.

**Obs. 1.** Inter-cpuset migration can cause higher tardiness than expected under DL.

**CBS suspensions.** As discussed in Sec. 2.2, a CBS may not adhere to the sporadic task model considered in [6] by suspending and resuming before a job’s deadline, resulting in either a self-suspending job or a separation between CBS job releases being less than a period. It is not obvious that this does not break bounded tardiness under AC, as arbitrary self-suspicions are known to cause capacity loss [12].

\(^2\)The actual value assigned is 1024. We discuss relative to 1.0 for simplicity.
We describe our proposed patch in four subsections, each addressing a distinct aspect of the implementation: bypassing the throttled state, pushing to the latest CPU, AC, and dynamic affinities. We begin each subsection by demonstrating how each aspect causes problems in SP systems. Next, we summarize how this behavior arises from the current implementation. Last, we explain how our patch modifies the implementation. Our patch is available online [14].

4 ADDING SP SCHEDULING TO DL

We describe our proposed patch in four subsections, each addressing a distinct aspect of the implementation: bypassing the throttled state, pushing to the latest CPU, AC, and dynamic affinities. We begin each subsection by demonstrating how each aspect causes problems in SP systems. Next, we summarize how this behavior arises from the current implementation. Last, we explain how our patch modifies the implementation. Our patch is available online [14].

4.1 Bypassing the Throttled State

Problem. In DL, a task that completes a job by exhausting its budget or calling sched_yield after its deadline will remain eligible, bypassing the throttled state (as in Fig. 1). If so, this task continues executing its next job on the same CPU if it is not preempted by a different task with an earlier deadline. This is desirable under clustered scheduling because a task that continues to execute on the same CPU retains its cache hotness. This behavior can cause unbounded tardiness when partitioned tasks exist.

* Ex. 5. (Ex. 4 of [15])⁶ corresponds with Fig. 5.) Consider a cpuset with CPUs p₁, p₂ and t₁ with tasks t₁, t₂, t₃, t₄, and t₅. Let (C₁, T₁) = (C₅, T₅) = (2, 6), (C₃, T₃) = (2, 2), and (C₃, T₃) = (1, 6), t₁, t₃, and t₅ are partitioned on p₁, p₂, and p₃, respectively, while t₂ and t₄ are migrating.

t₂ and t₄ release jobs periodically. Initially, t₂ and t₄ execute on p₁ and p₃, respectively. At t = 6, fixed tasks t₁ and t₅ preempt t₂ and t₄, respectively. The only other processor available to both t₂ and t₄ is p₂. t₂ cannot preempt t₅ on p₃ and t₄ cannot preempt t₁ on p₁), which they cannot both use. We assume the tiebreak here favors t₂ and it is scheduled, while t₄ does not execute until t = 8 when it resumes execution on p₃. t₂ is also forced to migrate off of p₂ by fixed task t₃ at t = 12. This repeats at t = 18, except here t₄ is scheduled over t₂ because it is tardy by 2.0 time units due to not being scheduled over [6, 8]. As a result, t₂ also becomes tardy by 2.0 time units by t = 20. This pattern can repeat indefinitely, and with each occurrence, the maximum tardiness experienced by either t₂ or t₄ increases by 2.0.

Implementation. Explaining why tasks migrate as in Ex. 5 and giving context to this and later subsections requires an explanation of DL’s migration code. Because a task can only run on a given CPU while on said CPU’s dl_rq’s (the DL-specific runqueue data structure), tasks must be exchanged between dl_rq’s to migrate.

The operations for exchanging tasks between dl_rq’s is called pull (from all other CPUs to the pulling CPU) and push (from the pushing CPU to a target CPU). Pulls are used to emulate a per-cluster runqueue during scheduling decisions. Prior to any scheduling decision, a CPU pulls the unscheduled task with earliest deadline and affinity for said CPU from each other CPU’s dl_rq. This allows the scheduling CPU to select the task with earliest deadline from any dl_rq’s in the cpuset as if they were stored in a single runqueue.

Pushes occur when a task is preempted or becomes eligible. In a push, the pushing CPU sends the unscheduled migrating task with earliest deadline on its dl_rq to the dl_rq of a target CPU.

Let the deadline of a CPU be the deadline of the eligible task with earliest deadline on its dl_rq. The target CPU of a push is either a CPU with no DL tasks on its dl_rq (called a free CPU) or, if the cpuset has no free CPUs, the CPU with the latest deadline among the other CPUS in the cpuset. Note that the latest CPU may be the pushing CPU, in which case the task stays on the same dl_rq.

A task that exhausts its budget or calls sched_yield is dequeued from its CPU’s dl_rq. What occurs next depends on whether this task bypasses the throttled state (its deadline has passed) or enters it (its deadline is in the future).

A task that bypasses the throttled state is immediately enqueued back onto its CPUs dl_rq with replenished budget and updated deadline. If another task on this CPU’s dl_rq now has an early enough deadline to preempt the replenished task, it does so. Otherwise, the replenished task continues executing on the same CPU.

Figure 5: (Fig. 2(b) of [15]) Unbounded tardiness under DL. (P for partitioned and M for migrating.)

Fig. 5: (Fig. 2(b) of [15]) Unbounded tardiness under DL. (P for partitioned and M for migrating.)

⁶Ex. 4 of [15] actually does not consider a SP system, as t₂ and t₄ were not allowed to migrate to p₃ and p₁, respectively in [15]. Observe when considering Ex. 5 that the same schedule occurs even when such migrations are allowed given specific tie-breaking assumptions.
Recall that tasks bypassing the throttled state and continuing to execute on the same CPUs resulted in unbounded tardiness in Ex. 5. For example, at its job boundaries at times 2 and 4 in Fig. 5, \(\tau_2\) does not migrate to free CPU \(\pi_2\) even though its current CPU \(\pi_1\) has a partitioned task \(\tau_1\) with a ready job. \(\tau_2\) does not migrate at these job boundaries because it bypasses the throttled state.

On the other hand, a task that enters the throttled state is unscheduled from the CPU that runs it. When this task returns from the throttled state to the eligible state (in function \(\text{dl_task_timer}\), an hrtimer callback), it is enqueued back onto this CPU’s \(\text{dl_rq}\) with a replenished budget and updated deadline. Because the replenished task became eligible, this CPU attempts to push a thread.

**Patch.** Ex. 5 would not have unbounded tardiness if successive jobs of tasks \(\tau_2\) and \(\tau_1\) would migrate when a free CPU or CPU with later deadline is available, thereby allowing partitioned tasks to execute. For example, if \(\tau_2\) had migrated to free CPU \(\pi_2\) at time 2 instead of 6 in Fig. 5, then the first job of \(\tau_1\) would have been able to execute on \(\pi_1\) over the interval \([2, 4]\). This would have freed \(\pi_1\) over \([6, 8]\), allowing migrating task \(\tau_2\) to execute and preventing its tardiness. Recall that successive jobs of tardy tasks do not migrate because they bypass the throttled state, skipping the push that occurs when a task returns from the throttled state.

Our patch addresses this by removing the branch in which a task bypasses the throttled state. This causes successive jobs of tardy tasks that would have otherwise continued to execute on the same CPU to be pushed from that CPU due to the tardy task becoming eligible from the throttled state. For example, at time 2 in Fig. 5, \(\tau_2\) completes its job. Under our patch, \(\tau_2\) would be throttled and immediately unthrottled because its next job already ready at time 2 (instead of not being throttled at all, as in the original implementation). Due to being unthrottled, \(\tau_2\) would be pushed from \(\pi_1\) to free CPU \(\pi_2\). This is the schedule described in the previous paragraph that reduces tardiness.

Note that callback function \(\text{dl_task_timer}\) will not push a scheduled task, as it is not safe to migrate an executing task. This is problematic because throttled tasks may still be scheduled because rescheduling is not instantaneous. To guarantee that a tardy task is pushed, \(\text{dl_task_timer}\) must wait for the tardy task to be unscheduled. Unfortunately, it does not help to wait within \(\text{dl_task_timer}\) for the relevant CPU to reschedule, as both \(\text{dl_task_timer}\) and the rescheduling code both must acquire this CPU’s \(r_q\) lock (the \(r_q\) struct contains generic runqueue fields as well as scheduler-specific data structures such as the \(\text{dl_rq}\)). In our patch, when \(\text{dl_task_timer}\) executes and observes that the task to be pushed is still scheduled, the callback releases the \(r_q\) lock and retries in the future, giving the relevant CPU the chance to reschedule. This is done with \(\text{hrtimer_forward_now}\).

To validate that our patch forces tasks to be pushed, consider the traces in Fig. 6. In Fig. 6a, the topmost task attempts to migrate whenever it completes a job (the vertical lines in the trace). Meanwhile, in Fig. 6b, this same task does not migrate unless preempted.

### 4.2 Pushing to the Latest CPU

**Problem.** The target CPU of a push is the CPU with latest deadline. In DL, the deadline of the pushing CPU may be from the task being pushed. This can result in priority inversions under SP scheduling.

- **Ex. 6.** (Corresponds with Fig. 7) Consider a cpuset with two CPUs \(\pi_1\) and \(\pi_2\) and three tasks \(\tau_1\), \(\tau_2\), and \(\tau_3\) with \((C_1, T_1) = (10, 70), (C_2, T_2) = (10, 50), (C_3, T_3) = (5, 10)\). \(\tau_1\) and \(\tau_2\) are partitioned on \(\pi_1\) and \(\pi_2\), respectively, while \(\tau_3\) is migrating. \(\tau_3\) releases its first job at \(t = 0\) and executes on CPU \(\pi_3\) until \(t = 5\), at which point \(\tau_3\) is throttled. At time \(t = 7\), both \(\tau_1\) and \(\tau_2\) release their first jobs and begin executing. At time \(t = 10\), \(\tau_1\) returns from the throttled state (recall Fig. 1). This results in \(\tau_1\) being placed back onto \(\pi_1\)'s \(d_l_r_q\) (as \(\pi_1\) was \(\tau_1\)'s last CPU) and \(\tau_1\) attempting to push \(\tau_2\).

Because, at the instant \(\tau_3\) is pushed, \(\tau_1\) executes \(\tau_1\) with deadline 77 and \(\pi_2\) executes \(\tau_2\) with deadline 57, \(\tau_3\) should remain on \(\pi_2\) and preempt \(\tau_1\), whose deadline is later than that of \(\tau_2\). The actual behavior exhibited by DL is that \(\tau_1\) will push \(\tau_3\) to \(\pi_2\), preempting \(\tau_2\). This is confirmed in Fig. 7b.

**Implementation.** The search for the target CPU is facilitated by a per-cpuset cpu1 struct. A cpu1d keeps a bitmap of the free CPUs in the cpuset and organizes the other CPUs into a max heap by deadline. This bitmap and heap are updated with functions \(\text{inc}\_d_l\_deadline\) and \(\text{dec}\_d_l\_deadline\), which are called whenever a task is enqueued or dequeued off a CPU’s \(d_l_r_q\), respectively.
In order to be pushed to a target CPU, a task must be on the pushing CPU’s \( d\_l\_rq \). Because this task was enqueued onto this \( d\_l\_rq \), the cpudl heap considers the pushed task’s deadline when evaluating the pushing CPU’s deadline. This may cause the cpudl to identify another CPU as a push target, even if the other tasks on the pushing CPU’s \( d\_l\_rq \) all have later deadlines than this target CPU. This is not a problem under clustered scheduling because a task that is wrongfully preempted on the target CPU can compensate by migrating to the pushing CPU, but this is not an option for partitioned tasks such as in Ex. 6.

After a target CPU is identified with cpudl, the \( rq \) lock of the target CPU is acquired with function \( double\_lock\_balance \) (\( push\_dl\_task \), the function that implements pushes in DL, is already holding the \( rq \) lock of the pushing CPU). The pushed task is then dequeued from the pushing CPU’s \( d\_l\_rq \) and enqueued onto that of the target CPU. Finally, both CPU’s \( rq \) locks are released.

**Patch.** The cpudl would not mistakenly target the wrong CPU in a push if it did not consider the deadline of the task being pushed. Our patch accomplishes this by dequeuing any task in the process of being pushed before the cpudl is accessed to determine the target CPU (in function \( find\_later\_rq \)). Dequeuing ahead of accessing the cpudl ensures that \( dec\_dl\_deadline \) is called prior to identifying a target CPU, preventing the pushing CPU from being represented by the pushed task’s deadline in the cpudl.

Pushes may fail due to race conditions that seem fundamental in DL, as \( rq \) locks may be released in \( double\_lock\_balance \) in order to acquire them in order (to prevent deadlock). Care must be taken so that an inconsistent state is not exposed to concurrently executing kernel code during a push while no \( rq \) lock is held. For this reason, we immediately enqueue a pushed task back onto the pushing CPU’s \( d\_l\_rq \) once \( find\_later\_rq \) (the function that checks the cpudl) returns. This enqueue is redundant if a target CPU is successfully identified and the push does not fail due to race conditions, as the task must then be dequeued once again to be enqueued onto the target CPU’s \( d\_l\_rq \).

### 4.3 AC

**Problem.** The existing AC does not prevent a user from overloading a given CPU in a cpuset by partitioning tasks with combined utilization exceeding 1.0 onto that CPU.

**Implementation.** As discussed in Sec. 2.3, the existing AC maintains (1) for all cpusets. Each cpuset in DL is represented by a \( root\_domain \) struct (e.g., the cpudl is a field of the \( root\_domain \)). Each \( root\_domain \) contains a pointer to a \( dl\_bw \) struct that stores the total utilization of all DL tasks assigned to the corresponding cpuset (i.e., the summation in (1)) in its \( \text{total}\_bw \) field. AC rejects any requests that falsify (1). Such requests may include adding tasks to \( \tau(t) \), changing the CPUs in \( \pi(t) \), or modifying either of \( \text{sched}\_rt\_runtime\_us \) or \( \text{sched}\_rt\_period\_us \), (1) is checked in function \( \text{sched}\_dl\_global\_validate \).

Unfortunately, there is no common helper function used to check (1) for all of these events. For adding tasks to \( \tau(t) \), (1) is checked in function \( _\_dl\_overflow \) (this is also where (1) is checked when task parameters are changed or a task from another cpuset is added to this cpuset, but we assume these events do not occur for the reasons discussed in Sec. 3). For changing the CPUs in \( \pi(t) \), (1) is checked in function \( \text{dl}\_cpuset\_cpumask\_can\_shrink \). For modifying \( \text{sched}\_rt\_runtime\_us \) or \( \text{sched}\_rt\_period\_us \), (1) is checked in function \( \text{sched}\_dl\_global\_validate \).

Adding a task to \( \tau(t) \) using \( \text{sched}\_setattr \) will fail if the server to add does not have affinity for every CPU in \( \pi(t) \). This is checked by comparing field span in the \( root\_domain \) (bitmask of CPUs in the cpuset) and field \( cpus\_ptr \) in the task’s \( task\_struct \) (bitmask of CPUs the task has affinity for) in helper function \( \text{__sched\_setscheduler} \). This check is skipped when AC is disabled.

**Patch.** Partitioned tasks are created by giving a task affinity for a single CPU in its cpuset prior to entering DL (as discussed in the next subsection, DL tasks are not allowed to change their affinities). We modify \( \text{__sched\_setscheduler} \) so that \( \text{sched\_setattr} \) will not fail if a task has affinity for a single CPU.

Besides the condition in (1), we modify AC to also maintain,

\[
\forall t : \forall \pi_j \in \pi(t) : \sum_{\tau_j \in \tau(t, \pi_j)} u_i \leq \frac{\text{sched\_rt\_runtime\_us}}{\text{sched\_rt\_period\_us}} \tag{2}
\]

This is done by adding checks for (2) where \( _\_dl\_overflow \) is called and in \( \text{sched\_dl\_global\_validate} \), which can cause (2) to be violated by adding a new task to \( \tau(t, \pi_j) \) or modifying its r.h.s. Similar to how the l.h.s. of (1) is tracked in \( \text{total}\_bw \) stored in the \( root\_domain \), the l.h.s. of (2) is tracked in a new variable \( \text{partitioned}\_bw \) stored in each \( \pi_j \)’s \( dl\_rq \).

(2) is not checked in \( \text{dl}\_cpuset\_cpumask\_can\_shrink \), which is called when \( \pi(t) \) changes. This is because the expected behavior under Linux when the CPUs in a cpuset are changed is that any affinity changes made with \( \text{sched\_setaffinity} \) are lost. Thus, all partitioned DL tasks in the cpuset will become migrating, making (2) irrelevant. Because all partitioned tasks become migrating, \( \text{partitioned}\_bw \) must be set to 0 for any CPUs in the prior \( \pi(t) \).

### 4.4 Dynamic Fine-Grained Affinities

**Problem.** We neglected to mention any checks made for (2) when \( \text{sched\_setaffinity} \) is used to change a partitioned task’s CPU, nor did we mention how we modified \( \text{sched\_setaffinity} \) to accept such requests. As it turns out, allowing DL tasks to dynamically change their affinity can actually starve such tasks.
6.1 Experimental Setup

Our experiments and the execution traces presented prior to this section were conducted on a 16 CPU Intel Xeon Silver 4110 multiprocessor. Measured workloads were restricted to a cluster composed of eight CPUs, as these compose a single socket and NUMA node, and clusters should reflect the hardware topology in practice. Periodic workloads were generated for these experiments using taskgen [7, 10] and rt-app [1]. SP task systems were created by applying worst-fit packing to the task sets generated by taskgen and determining any unpacked tasks to be migrating.

We are interested in how our modifications to DL’s migration code described in Sec. 4 affect overheads. Changes to overheads are due to forcing tardy tasks to enter the throttled state, thereby requiring that such tasks wait for an hrtimer callback (dl_task_timer) to complete before becoming eligible again, and due to the added dequeue and enqueue operations required for every push. For measuring the additional latency caused by this hrtimer callback, we inserted ftrace event tracepoints into our patched kernel that are triggered whenever a tardy task is forced into the throttled state and when dl_task_timer returns said task to a runqueue. To measure the duration of pushes, we also inserted tracepoints around push_dltask. To get a more holistic view of how these changes to

5 SOUNDNESS ARGUMENT FOR AC

Recall from Sec. 1 that EDF variant SAPA-EDF was proven to be SRT-optimal in [16], but was not implemented in DL due to its increased migrations and scheduler complexities. Our DL variant is a compromise between SAPA-EDF and the existing DL implementation for SP systems. The soundness of our modified AC derives from the behaviors of SAPA-EDF that our patched DL emulates. We begin with an overview of these behaviors.

SAPA-EDF is SRT-optimal because it prevents affinity-related priority inversions (API). Under SP systems, an API occurs when an unscheduled partitioned task is unable to preempt a migrating task on the partitioned task’s CPU; meanwhile, some other CPU

Ex. 7. This example corresponds with Fig. 8. Consider a SP system on a cpuset with two CPUs \(\pi_1\) and \(\pi_2\) and two tasks \(t_1\) (initially partitioned on \(\pi_1\)) and \(t_2\) migrating with \((C_1, T_1) = (1, 2)\) and \((C_2, T_2) = (1, 1)\). Both tasks release jobs at time \(t = 0\).

\(t_2\), whose deadline is earlier than \(t_1\)’s deadline, executes on \(\pi_1\) until preempted by \(t_1\) at \(t = 2\). \(t_2\) migrates to \(\pi_2\) once preempted. However, \(t_1\) requests to change its affinity from being partitioned on \(\pi_1\) to \(\pi_2\). Similar to how a task’s deadline is updated based on the current time when returning from the idle state after said task’s original deadline (recall Fig. 1), \(t_1\)’s deadline is updated from 2 to 4 when it migrates at \(t = 2\). Thus, \(t_1\) does not have an early enough deadline to preempt \(t_2\), and so it continues to be unscheduled.

Repeating this pattern of dynamic affinity requests can prevent \(t_1\) from executing indefinitely, as in Fig. 8a (the reason that the workload corresponding to \(t_1\) in Fig. 8b executes briefly every two time units is because the request to change affinity does not occur instantaneously at the moment of preemption).

Implementation. When AC is enabled, sched_setaffinity performs the same check between the cpuset’s CPUs and the requested affinity as __sched_setscheduler. Thus, effectively sched_setaffinity did nothing for DL tasks, as such tasks needed to already have affinity for all the cpuset’s CPUs to enter DL. When AC is disabled, this check is bypassed and the task is given the requested affinity. If the task no longer has affinity for the CPU whose rq the task is currently on, it is immediately moved to the rq of a CPU in it has affinity for. If the task is tardy during this migration, it is immediately replenished based on the current time \(t\).

This sets the task’s deadline \(d = t + T\), lowering the task’s priority.

Patch. We forbid DL tasks from changing their affinities. We do this by modifying sched_setaffinity to automatically reject any requests for DL tasks. If a user desires to change the affinity of a DL task, the task must first leave DL, change its affinity as a non-DL workload, and reenter DL with a new server. This new server’s deadline will be set based on its time of creation. This ultimately has the same effect as the original implementation, but forcing tasks to leave DL emphasizes to the user that real-time performance guarantees may be invalid under changes to affinity.

Rejecting all requests to sched_setaffinity may be heavy handed, but it is non-trivial to determine what restrictions are necessary to both prevent race conditions and account for such requests in proofs of bounded tardiness.

 Altogether, our patch is fairly minor, modifying roughly 200 LOC (for context, the main DL file is roughly 3000 LOC).

6 EVALUATION

In this section, we evaluate the performance of our patched DL variant against the original implementation.

Ex. 8. Consider a time instant such that an ARPI exists between partitioned task \(t_1\), migrating task \(t_2\) executing on \(t_1\)’s CPU, and task \(t_3\) with later deadline than \(t_1\).

Under our patched DL, this priority inversion lasts for at most the execution time of \(t_2\)’s job. When \(t_2\)’s job completes, \(t_2\) will always be throttled (recall from Sec. 4 that all tasks that complete jobs are throttled in our patch), and will migrate to \(t_3\)’s CPU when it becomes eligible and is pushed (recall that we’ve modified pushes to always target the CPU that executes the task with latest deadline).

Note that it is impossible for \(t_2\) to be preempted before its job completes for the duration of the ARPI, as any migrating task will preempt \(t_3\) before \(t_2\) under our patched DL because \(t_3\) has a later deadline. If a partitioned task preempts \(t_2\), then the ARPI ends by definition (recall from the above paragraphs that for an ARPI, it is required that an executing task be scheduled on \(t_1\)’s CPU).

Because these ARPIs are transient (i.e., bounded), any resulting increase in tardiness relative to SAPA-EDF is also bounded. As tardiness is bounded under SAPA-EDF for any feasible system (because SAPA-EDF is SRT-optimal [16]) and because our AC conditions (1) and (2)) guarantee that the system is feasible [17], it follows that tardiness is bounded under our patched DL and AC. As a reminder, the formal version of this proof is available online [14].
migration code affect performance, we also measured the tardiness tasks experience scaled by their periods.

Taskgen is configured such that each generated task set has a total utilization of 7.52. This is slightly below 95% of eight CPUs’ worth of capacity to guarantee that AC will not reject tasks due to potential rounding in taskgen in our patched kernel (AC must be disabled for SP scheduling in the original). We focus on heavily utilized systems as we are primarily interested in SRT and lightly utilized systems produce less, if any, tardiness. We considered task systems composed of 16 and 40 tasks to consider systems with both heavy and light per-task utilizations. Ten different task systems were measured for each number of tasks. Timestamps for each task system were collected over an interval of ten minutes on both the original and our patched kernel.

Latency of forced throttles. Note that we do not compare against the original DL implementation when considering forced task throttling because this overhead is unique to our variant. The distribution of sampled durations during which our patched DL forced a task to be throttled when it would not have in the original implementation is presented in Fig. 9. The average latency caused by a forced throttle was 44 us for systems with 16 tasks and 34 us for systems with 40 tasks. The multimodal distribution of throttle times in Fig. 9a is likely due to our usage of hr_timer_forward_now with the dl_task_timer callback to ensure that tasks forced into the throttle state are unscheduled before dl_task_timer attempts to push a task. The time difference between peaks roughly correlates with the time interval the hrtimer is forwarded by in our patched DL. As each usage of hr_timer_forward_now requires the task to wait an additional timer interval, each peak in this histogram likely corresponds with a different number of calls to this function.

These latencies may not be acceptable for servers whose workloads require sub ms response times. However, if the size of these latencies is being caused by waiting for the hrtimer as we suspect, an alternative method for forcing tasks that complete jobs to migrate that avoids using the hrtimer may be more practical. As stated earlier in Sec. 1, one of the goals of this work was to make as few adjustments to the existing implementation as possible. In future work, we will loosen this restriction.

Duration of pushes. The distribution of sampled push durations for both DL and our patched variant is presented in Fig. 10. These distributions are multimodal because push_dl_task contains a retry loop in case the state of the system changes while rq locks are dropped in double_lock_balance. The effect of the added enqueue and dequeue operations on push overheads is minor, entailing a change of about 1 us to the average duration of a push.

Tardiness. The distribution of samples of tasks’ tardiness levels is presented in Fig. 11. Average tardiness is negligibly lower under our patched variant compared to the original implementation.

Altogether, the changes made by our patch do not seem to increase overheads by a substantial amount. The most concerning overhead is the latency caused by forwarding the dl_task_timer callback function, and even this overhead occurs relatively infrequently. This can be observed by comparing the y-axes of Fig. 9 and 10. The number of pushes performed by the system vastly outnumbers the count of forced throttles.

7 CONCLUSION

We have presented a patch to DL to support semi-partitioned affinities with theoretically sound AC. In looking to prove the soundness of our modified AC, we have highlighted several instances of AC being broken in the existing DL implementation. We believe this has been caused by implementation decisions being driven by heuristics and empirical validation over theoretical analysis. While heuristic-driven development certainly outperforms being oblivious to features such as DVFS or asymmetric capacities, we have demonstrated that intuition is unreliable when it comes to guaranteeing bounded tardiness under AC.

In future work, we would like to consider supporting AC for some of the features we omitted from our model in Sec. 3. We believe that similarly to how our patch compromises between the current DL implementation and SAPA-EDF, a similar compromise can be made between DL and the EDF variants proposed in the theory for asymmetric CPUs.
REFERENCES


