On Real-Time STM Concurrency Control for Embedded Software with Improved Schedulability

Mohammed Elshambakey, Binoy Ravindran
ECE Dept., Virginia Tech
Blacksburg, VA 24060, USA
shambake@vt.edu, binoy@vt.edu

Abstract— We consider software transactional memory (STM) concurrency control for embedded multicore real-time software, and present a novel contention manager for resolving transactional conflicts, called PNF. We upper bound transactional retries and task response times, and show when PNF has better schedulability than previous contention managers and lock-free synchronization. Our implementation in the Rochester STM framework/real-time Linux reveals that PNF yields comparable retry costs than competitors.

I. Introduction

Embedded systems sense physical processes and control their behavior, typically through feedback loops. Since physical processes are concurrent, computations that control them must also be concurrent, enabling them to process multiple streams of sensor input and control multiple actuators, all concurrently. Often, such computations need to concurrently read/write shared data objects. They must also process sensor input and react, while satisfying time constraints.

Lock-based concurrency control has significant programmability, scalability, and composability challenges [11]. Software transactional memory (STM) is an alternative synchronization model for shared memory data objects that promises to alleviate these difficulties. With STM, code that read/write shared objects is organized as transactions, which execute speculatively, while logging changes made to objects. Two transactions conflict if they access the same object and one access is a write. When that happens, a contention manager (CM) [9] resolves the conflict by aborting one and allowing the other to commit, yielding (the illusion of) atomicity. Aborted transactions are re-started, after rolling back the changes. In addition to a simple programming model, STM provides performance comparable to lock-free approach and is composable [10].

Given STM's programmability, scalability, and composability advantages, it is a compelling concurrency control technique also for multicore embedded real-time software. However, this requires bounding transactional retries, as real-time threads, which subsume transactions, must satisfy time constraints. Retry bounds under STM are dependent on the CM policy at hand.

Past real-time CM research has proposed resolving transactional contention using dynamic and fixed priorities of parent threads, resulting in Earliest Deadline First CM (ECM) and Rate Monotonic CM (RCM), respectively [8, 7, 6]. In particular, [7] shows that ECM and RCM achieve higher schedulability – i.e., greater number of task sets meeting their time constraints – than lock-free synchronization only under some ranges for the maximum atomic section length. That range is significantly expanded with the Length-based CM (LCM) in [6], increasing the coverage of STM's timeliness superiority. However, these works restrict to *one* object access per transaction, which is a major limitation (Section III).

To allow multiple objects per transaction, we design a novel contention manager called PNF (Section IV), which can be used with global EDF (G-EDF) and global RMA (G-RMA) multicore real-time schedulers [2]. We upper bound transactional retry costs and task response times (Section V), and formally compare PNF's schedulability with ECM, RCM, LCM, and lock-free synchronization (Section **??**). We show that PNF achieves better schedulability than lock-free synchronization for larger atomic section length range than ECM and RCM. Our implementation reveals that PNF yields comparable retry costs to competitors (Section VI).

PNF's superior timeliness properties thus allow embedded real-time programmers to reap STM's significant programmability and composability advantages for a broader range of multicore embedded real-time software than what was previously possible – paper's contribution.

II. Preliminaries

We consider a multiprocessor system with m identical processors and n sporadic tasks $\tau_1, \tau_2, \ldots, \tau_n$. The k^{th} instance (or job) of a task τ_i is denoted τ_i^k. Each task τ_i is specified by its worst case execution time (WCET) c_i, its minimum period T_i between any two consecutive instances, and its relative deadline D_i, where $D_i = T_i$. Job τ_i^j is released at time r_i^j and must finish no later than its absolute deadline $d_i^j = r_i^j + D_i$. Under a fixed priority scheduler such as G-RMA, p_i determines τ_i's (fixed) priority and it is constant for all instances of τ_i. Under a dynamic priority scheduler such as G-EDF, a job τ_i^j's priority, p_i^j, differs from one instance to another. A task τ_j may interfere with task τ_i for a number of times during an interval L, and this number is denoted as $G_{ij}(L)$.

Shared objects. A task may need to read/write shared, in-memory data objects while it is executing any of its atomic sections (transactions), which are synchronized using STM. The set of atomic sections of task τ_i is denoted s_i. s_i^k is the k^{th} atomic section of τ_i. $p(s_i^k)$ is the priority of transaction s_i^k. Each object, θ, can be accessed by multiple tasks. The set of distinct objects accessed by τ_i is θ_i without repeating objects. The set of atomic sections used by τ_i to access θ is $s_i(\theta)$, and the sum of the lengths of those atomic sections is $len(s_i(\theta))$. $s_i^k(\theta)$ is the k^{th} atomic section of τ_i that accesses θ. s_i^k can access one or more objects in θ_i. So, s_i^k refers to the transaction itself, regardless of the objects accessed by the transaction. We denote the set of all accessed objects by s_i^k as Θ_i^k. While $s_i^k(\theta)$ implies that s_i^k accesses an object $\theta \in \Theta_i^k$, $s_i^k(\Theta)$ implies that $\bar{s_i^k}$ accesses a set of objects $\Theta = \{\theta : \theta \in \Theta_i^k\}$. $\bar{s_i^k} = \bar{s_i^k}(\Theta)$ refers only once to s_i^k, regardless of the number of objects in Θ. So, $|\bar{s_i^k}(\Theta)|_{\forall\theta\in\Theta} = 1$. $s_i^k(\theta)$ executes for a duration $len(s_i^k(\theta))$. $len(s_i^k) = len(s_i^k(\theta)) = len(s_i^k(\Theta)) = len(s_i^k(\Theta_i^k))$ The set of tasks sharing θ with τ_i is denoted $\gamma_i(\theta)$. Atomic sections are non-nested (supporting nested STM is future work). The maximum-length atomic section in τ_i that accesses θ is denoted $s_{i_{max}}(\theta)$, while the maximum one among all tasks is $s_{max}(\theta)$, and the maximum one among tasks with priorities lower than that of τ_i is $s_{max}^i(\theta)$.

STM retry cost. If two or more atomic sections conflict, the CM will commit one section and abort and retry the others, increasing the time to execute the aborted sections. The increased time that an atomic section $s_i^p(\theta)$ will take to execute due to a conflict with another section $s_j^k(\theta)$, is denoted $W_i^p(s_j^k(\theta))$. If an atomic section, s_i^p, is already executing, and another atomic section s_j^k tries to access a shared object with s_i^p, then s_j^k is said to "interfere" or "conflict" with s_i^p. The transaction s_j^k is the "interfering transaction", and the transaction s_i^p is the "interfered transaction".

Due to *transitive retry* (introduced in Section III), an atomic section $s_i^k(\Theta_i^k)$ may retry due to another atomic section $s_j^l(\Theta_j^l)$, where $\Theta_i^k \cap \Theta_j^l = \emptyset$. θ_i^* denotes the set of objects not accessed directly by atomic sections in τ_i, but can cause transactions in τ_i to retry due to transitive retry. $\theta_i^{ex}(=\theta_i+\theta_i^*)$ is the set of all objects that can cause transactions in τ_i to retry directly or through transitive retry. γ_i^* is the set of tasks that accesses objects in θ_i^*. $\gamma_i^{ex}(=\gamma_i+\gamma_i^*)$ is the set of all tasks that can directly or indirectly (through transitive retry) cause transactions in τ_i to retry.

The total time that a task τ_i's atomic sections have to retry over T_i is denoted $RC(T_i)$. The additional amount of time by which all interfering jobs of τ_j increases the response time of any job of τ_i during L, without considering retries due to atomic sections, is denoted $W_{ij}(L)$.

III. Limitations of ECM, RCM, and LCM

ECM and RCM [7] use dynamic and fixed priorities, respectively, to resolve conflicts. ECM uses G-EDF, and allows the transaction whose job has the earliest absolute deadline to commit first [8]. RCM uses G-RMA, and commits the transaction whose job has the shortest period. To use STM with real-time systems, retry cost should be bounded in order to satisfy timing constraints. Retry cost under ECM is bounded in [7] as follows:

Claim 1 (from [7]): Under ECM, a task τ_i's maximum retry cost during T_i is upper bounded by:

$$\begin{aligned} RC\left(T_i\right) \quad &\leq \quad \sum_{\theta\in\theta_i}\Bigg(\Big(\sum_{\tau_j\in\gamma_i(\theta)}\Big(\left\lceil\frac{T_i}{T_j}\right\rceil\sum_{\forall s_j^l(\theta)}len\big(s_j^l(\theta)\\ &+ \quad s_{max}(\theta)\big)\Big)\Big)-s_{max}(\theta)+s_{i_{max}}(\theta)\Bigg) \qquad (1)\end{aligned}$$

Retry cost under RCM is similar to ECM, except that only higher priority tasks to τ_i can interfere with any job τ_i^x of τ_i. Retry cost under RCM is bounded by Claim 3 in [7].

G-EDF/LCM [6] and G-RMA/LCM default to ECM and RCM, respectively, with some difference. Under LCM, a higher priority transaction $s_i^k(\theta)$ cannot abort a lower priority transaction $s_j^l(\theta)$ if $s_j^l(\theta)$ has already consumed α percentage of its execution length. G-EDF/LCM's retry cost is bounded in [6] as follows:

Claim 5 (from [6]): $RC(T_i)$ for a task τ_i under G-EDF/LCM is upper bounded by:

$$\begin{aligned} RC(T_i) \quad &= \quad \Bigg(\sum_{\forall\tau_h\in\gamma_i}\sum_{\forall\theta\in\theta_i\wedge\theta_h}\Bigg(\left\lceil\frac{T_i}{T_h}\right\rceil\sum_{\forall s_h^l(\theta)}len\Big(s_h^l(\theta)\Big)\\ &+ \quad \alpha_{max}^{hl}len\Big(s_{max}^h(\theta)\Big)\Bigg)\Bigg)\\ &+ \quad \sum_{\forall s_i^y(\theta)}\Big(1-\alpha_{max}^{iy}\Big)len\Big(s_{max}^i(\theta)\Big) \qquad (2)\end{aligned}$$

where α_{max}^{hl} is the α value that corresponds to ψ due to the interference of $s_{max}^h(\theta)$ by $s_h^l(\theta)$. α_{max}^{iy} is the α value that corresponds to ψ due to the interference of $s_{max}^i(\theta)$ by $s_i^y(\theta)$.

G-RMA/LCM's retry cost is similar to G-EDF/LCM's, except that only higher priority tasks to τ_i can interfere with any job τ_i^x of τ_i. G-RMA/LCM's retry cost is bounded by Claim 8 in [6].

As mentioned before, [7, 6] assumes that each transaction accesses only one object. This assumption simplifies the retry cost (Claims 2 and 3 in [7], and Claims 5, 8 in [6]) and response time analysis (Sections 4 and 5 in [7], and Sections 4.2, 4.5 in [6]). Besides, it enables comparison with lock-free synchronization [4]. With multiple objects per transaction, ECM, RCM and LCM will face transitive retry, which we illustrate with an example.

Example 1. Consider three atomic sections s_1^x, s_2^y, and s_3^z belonging to jobs τ_1^x,τ_2^y, and τ_3^z, with priorities $p_3^z > p_2^y > p_1^x$, respectively. Assume that s_1^x and s_2^y share objects, s_2^y and s_3^z share objects. s_1^x and s_3^z do not share objects. s_3^z can cause s_2^y to retry, which in turn will cause s_1^x to retry. This means that s_1^x may retry transitively because of s_3^z, which will increase the retry cost of s_1^x.

Assume another atomic section s_4^f is introduced. Priority of s_4^f is higher than priority of s_3^z. s_4^f shares objects only with s_3^z. Thus, s_4^f can make s_3^z to retry, which in turn will make s_2^y to retry, and finally, s_1^x to retry. Thus, transitive retry will move from s_4^f to s_1^x, increasing the retry cost of s_1^x. The situation gets worse as more tasks of higher priorities are added, where each task shares objects with its immediate lower priority task. τ_3^z may have atomic sections that share objects with τ_1^x, but this will not prevent the effect of transitive retry due to s_1^x.

Definition 1 ***Transitive retry:*** *A transaction s_i^k suffers from transitive retry when s_i^k retries due to a higher priority transaction s_z^h, yet $\Theta_z^h \cap \Theta_i^k = \emptyset$.*

Claim 1 *ECM, RCM and LCM suffer from transitive retry for multi-object transactions.*

Proof 1 Example 1 applies for any transactions under ECM, RCM and LCM. Claim follows.

Therefore, the analysis in [7] and [6] must extend the set of objects that can cause an atomic section of a lower priority job to retry. This can be done by initializing the set of conflicting objects, γ_i, to all objects accessed by all transactions of τ_i. We then cycle through all transactions belonging to all other higher priority tasks. Each transaction s_j^l that accesses at least one of the objects in γ_i adds all other objects accessed by s_j^l to γ_i. The loop over all higher priority tasks is repeated, each time with the new γ_i, until there are no more transactions accessing any object in γ_i[1].

In addition to the *transitive retry* problem, retrying higher priority transactions can prevent lower priority tasks from running. This happens when all processors are busy with higher priority jobs. When a transaction retries, the processor time is wasted. Thus, it would be better to give the processor to some other task.

Essentially, what we present is a new contention manager that avoids the effect of transitive retry. We call it, Priority contention manager with Negative values and First access (or PNF). PNF also tries to enhance processor utilization. This is done by allocating processors to jobs with non-retrying transactions if any exists. PNF is described in Section IV.

IV. The PNF Contention Manager

Algorithm 1 describes PNF. It manages two sets. The first is the m-set, which contains at most m non-conflicting transactions, where m is the number of processors, as there cannot be more than m executing transactions (or generally, m executing jobs) at the same time. When a transaction is entered in the m-set, it executes non-preemptively and no other transaction can abort it. A transaction in the m-set is called an *executing transaction.* This means that, when a transaction is executing before the arrival of higher priority conflicting transactions, then the one that started executing first will be committed (Step 8) (hence the term "First access" in the algorithm's name).

Algorithm 1: PNF Algorithm

Data: *Executing Transaction:* is one that cannot be aborted by any other transaction, nor preempted by a higher priority task;
m-set: m-length set that contains only non-conflicting executing transactions;
n-set: n-length set that contains retrying transactions for n tasks in non-increasing order of priority;
n(z): transaction at index z of the n-set;
s_i^k: a newly released transaction;
s_j^l: one of the executing transactions;
Result: atomic sections that will commit

1 **if** s_i^k *does not conflict with any executing transaction* **then**
2   Assign s_i^k as an executing transaction;
3   Add s_i^k to the m-set;
4   Select s_i^k to commit
5 **else**
6   Add s_i^k to the n-set according to its priority;
7   Assign temporary priority -1 to the job that owns s_i^k ;
8   Select transaction(s) conflicting with s_i^k for commit;
9 **end**
10 **if** s_j^l *commits* **then**
11   **for** *z=1 to size of n-set* **do**
12   **if** *n(z) does not conflict with any executing transaction* **then**
13   **if** *processor available*[2] **then**
14   Restore priority of task owning n(z);
15   Assign n(z) as executing transaction;
16   Add n(z) to m-set and remove it from n-set;
17   Select n(z) for commit;
18   **else**
19   Wait until processor available
20   **end**
21   **end**
22   **end**
23 **end**

[1]However, note that, this solution may over-extend the set of conflicting objects, and may even contain all objects accessed by all tasks.

[2]An idle processor or at least one that runs a non-atomic section task with priority lower than the task holding $n(z)$.

The second set is the n-set, which holds the transactions that are retrying because of a conflict with one or more of the executing transactions (Step 6), where n stands for the number of tasks in the system. Transactions in the n-set are known as *retrying transaction*. It also holds transactions that cannot currently execute, because processors are busy, either due to processing executing transactions and/or higher priority jobs. Any transaction in the n-set is assigned a temporal priority of -1 (Step 7) (hence the word "Negative" in the algorithm's name). A negative priority is considered smaller than any normal priority, and a transaction continues to hold this negative priority until it is moved to the m-set, where it is restored its normal priority.

A job holding a transaction in the n-set can be preempted by any other job with normal priority, even if that job does not have transactions conflicting with the preempted job. Hence, this set is of length n, as there can be at most n jobs. Transactions in the n-set whose jobs have been preempted are called preempted transactions. The n-set list keeps track of preempted transactions, because as it will be shown, all preempted and non-preempted transactions in the n-set are examined when any of the executing transaction commits. Then, one or more transactions are selected from the n-set to be executing transactions. If a retrying transaction is selected as an executing transaction, the task that owns the retrying transaction regains its priority.

When a new transaction is released, and if it does not conflict with any of the executing transactions (Step 1), then it will allocate a slot in the m-set and becomes an executing transaction. When this transaction is released (i.e., its containing task is already allocated to a processor), it will be able to access a processor immediately. This transaction may have a conflict with any of the transactions in the n-set. However, since transactions in the n-set have priorities of -1, they cannot prevent this new transaction from executing if it does not conflict with any of the executing transactions.

When one of the executing transactions commits (Step 10), it is time to select one of the n-set transactions to commit. The n-set is traversed from the highest priority to the lowest priority (priority here refers to the original priority of the transactions, and not -1) (Step 11). If an examined transaction in the n-set, s_h^b, does not conflict with any executing transaction (Step 12), and there is an available processor for it (Step 13) ("available" means either an idle processor, or one that is executing a job of lower priority than s_h^b), then s_h^b is moved from the n-set to the m-set as an executing transaction and its original priority is restored. If s_h^b is added to the m-set, the new m-set is compared with other transactions in the n-set with lower priority than s_h^b. Hence, if one of the transactions in the n-set, s_d^g, is of lower priority than s_h^b and conflicts with s_h^b, it will remain in the n-set.

The choice of the new transaction from the n-set depends on the original priority of transactions (hence the term "Priority" in the algorithm name). The algorithm avoids interrupting an already executing transaction to reduce its retry cost. In the meanwhile, it tries to avoid delaying the highest priority transaction in the n-set when it is time to select a new one to commit, even if the highest priority transaction arrives after other lower priority transactions in the n-set.

A. Properties

Claim 2 *Transactions scheduled under PNF do not suffer from transitive retry.*

Proof 2 Proof is by contradiction. Assume three transactions s_i^k, s_j^l and s_z^h. $p(s_z^h) > p(s_j^l) > p(s_i^k)$. $\Theta_i^k \cap \Theta_j^l \neq \emptyset$, $\Theta_j^l \cap \Theta_z^h \neq \emptyset$, but $\Theta_i^k \cap \Theta_z^h = \emptyset$. Assume s_i^k is transitively retrying because of s_z^h. This means s_i^k, s_j^l and s_z^h are executing concurrently. $\Theta_i^k \cap \Theta_j^l \neq \emptyset$, so s_i^k and s_j^l cannot be executing transactions at the same time by definition of PNF. $\Theta_j^l \cap \Theta_z^h \neq \emptyset$, so s_j^l and s_z^h cannot be executing transactions at the same time by definition of PNF. Only s_i^k and s_z^h can be executing transactions at the same time because $\Theta_i^k \cap \Theta_z^h = \emptyset$. Hence, the three transactions cannot be running concurrently. So, s_i^k cannot be transitively retrying because of s_z^h which contradicts with the first assumption. Claim follows.

From Claim 2, PNF does not increase the retry cost of multi-object transactions. However, this is not the case for ECM and RCM as shown by Claim 1.

Claim 3 *Under PNF, any job τ_i^x is not affected by the retry cost in any other job τ_j^l.*

Proof 3 As explained in Section 1, PNF assigns a temporary priority of -1 to any job that includes a retrying transaction. So, retrying transactions have lower priority than any other normal priority. When τ_i^x is released and τ_j^l has a retrying transaction, τ_i^x will have a higher priority than τ_j^l. Thus, τ_i^x can run on any available processor while τ_j^l is retrying one of its transactions. Claim follows.

V. Retry Cost under PNF

We now derive an upper bound on the retry cost of any job τ_i^x under PNF during an interval $L \leq T_i$. Since all tasks are sporadic (i.e., each task τ_i has a minimum period T_i), T_i is the maximum study interval for each task τ_i.

Claim 4 *Under PNF, the maximum retry cost suffered by a transaction s_i^k due to a transaction s_j^l is $len(s_j^l)$.*

Proof 4 By PNF's definition, s_i^k cannot have started before s_j^l. Otherwise, s_i^k would have been an executing transaction and s_j^l cannot abort it. So, the earliest release time for s_i^k would have been just after s_j^l starts execution. Then, s_i^k would have to wait until s_j^l commits. Claim follows.

Claim 5 *The retry cost for any job τ_i^x due to conflicts between its transactions and transactions of other jobs under PNF during an interval $L \leq T_i$ is upper bounded by:*

$$RC(L) \leq \sum_{\tau_j \in \gamma_i} \left(\sum_{\theta \in \theta_i} \left(\left(\left\lceil \frac{L}{T_j} \right\rceil + 1 \right) \sum_{\forall s_j^{\bar{l}}(\theta)} len\left(s_j^{\bar{l}}(\theta) \right) \right) \right) \tag{3}$$

Proof 5 Consider a transaction s_i^k belonging to job τ_i^x. By definition of PNF, higher and lower priority transactions than s_i^k can become executing transaction before s_i^k. The worst case scenario for s_i^k occurs when s_i^k has to wait in the n-set, while all other conflicting transactions with s_i^k are chosen to be executing transactions. Executing transactions are not aborted. This is why $s_j^{\bar{l}}$ is included only once in (3) for all shared objects with s_i^k.

The maximum number of jobs of any task τ_j that can interfere with τ_i^x during interval L is $\left\lceil \frac{L}{T_j} \right\rceil + 1$. From the previous observations and Claim 4, Claim follows.

Claim 6 *The blocking time for a job τ_i^x due to lower priority jobs during an interval $L \leq T_i$ is upper bounded by:*

$$D(\tau_i^x) \leq \left\lfloor \frac{1}{m} \sum_{\forall \tau_j^{\bar{l}}} \left(\left(\left\lceil \frac{L}{T_j} \right\rceil + 1 \right) \sum_{\forall s_j^{\ddot{h}}} len\left(s_j^{\ddot{h}} \right) \right) \right\rfloor \tag{4}$$

where $D(\tau_i^x)$ is the blocking time suffered by τ_i^x due to lower priority jobs. $\tau_j^{\bar{l}} = \{\tau_j^l : p_j^l < p_i^x\}$ and $s_j^{\ddot{h}} = \{s_j^h : \left(\Theta_j^h \cap \Theta_i^k = \emptyset\right) \wedge \left(\forall \Theta_i^k \in \theta_i\right)\}$. During this blocking time, all processors are unavailable for τ_i^x.

Proof 6 Under PNF, executing transactions are non preemptive. So, a lower priority executing transaction can delay a higher priority job τ_i^x if no other processors are available. Lower priority executing transactions can be conflicting or non-conflicting with any transaction in τ_i^x. If lower priority transactions are conflicting with any transaction in τ_i^x, then (3) already covers the increase in retry cost of transactions in τ_i^x due to lower priority transactions. Otherwise, lower priority non-conflicting transactions can be executing transactions that block τ_i^x.

Lower priority non-conflicting transactions can block τ_i^x when τ_i^x is newly released, or after that:

Lower priority non-conflicting transactions when τ_i^x is newly released: τ_i^x is delayed if there are no available processors for it. Otherwise, τ_i^x can run in parallel with these non-conflicting lower priority transactions. Each lower priority non-conflicting transaction $s_j^{\ddot{h}}$ will delay τ_i^x for $len(s_j^{\ddot{h}})$.

Lower priority non-conflicting transactions after τ_i^x is released: This situation can happen if τ_i^x is retrying one of its transactions s_i^k. So, τ_i^x is assigned a priority of -1. τ_i^x can be preempted by any other job. When s_i^k is checked again to be an executing transaction, all processors may be busy with lower priority non-conflicting transaction and/or higher priority jobs. Otherwise, τ_i^x can run in parallel with these lower priority non-conflicting transactions.

Each lower priority non-conflicting transaction $s_j^{\ddot{h}}$ will delay τ_i^x for $len(s_j^{\ddot{h}})$. From the previous cases, lower priority non-conflicting transactions act as if they were higher priority jobs interfering with τ_i^x. So, the blocking time can be calculated by the interference workload given by Theorem 7 in [1].

Claim 7 *The response time of a job τ_i^x, during an interval $R_i^{up} \leq T_i$, under PNF/G-EDF is upper bounded by:*

$$R_i^{up} = c_i + RC(R_i^{up}) + D_{edf}(\tau_i^x) + \left\lfloor \frac{1}{m} \sum_{\forall j \neq i} W_{ij}(R_i^{up}) \right\rfloor \tag{5}$$

where $RC(R_i^{up})$ is calculated by (3). $D_{edf}(\tau_i^x)$ is the same as $D(\tau_i^x)$ defined in (4). However, for G-EDF systems. $D_{edf}(\tau_i^x)$ is calculated as:

$$D_{edf}(\tau_i^x) \leq \left\lfloor \frac{1}{m} \sum_{\forall \tau_j^{\bar{l}}} \begin{cases} 0 & , R_i^{up} \leq T_i - T_j \\ \sum_{\forall s_j^{\ddot{h}}} len\left(s_j^{\ddot{h}} \right) & , R_i^{up} > T_i - T_j \end{cases} \right\rfloor \tag{6}$$

and $W_{ij}(R_i^{up})$ is calculated by (3) in [7].

Proof 7 Response time for τ_i^x is calculated as (3) in [7] with the addition of blocking time defined by Claim 6. G-EDF uses absolute deadlines for scheduling. This defines which jobs of the same task can be of lower priority than τ_i^x, and which will not. Any instance τ_j^h, released between $r_i^x - T_j$ and $d_i^x - T_j$, will be of higher priority than τ_i^x. Before $r_i^x - T_j$, τ_j^h would have finished before τ_i^x is released. After $d_i^x - T_j$, d_j^h would be greater than d_i^x. Thus, τ_j^h will be of lower priority than τ_i^x. So, during T_i, there can be only one instance τ_j^h of τ_j with lower priority than τ_i^x. τ_j^h is released between $d_i^x - T_j$ and d_i^x. Consequently, during $R_i^{up} < T_i - T_j$, no existing instance of τ_j is of lower priority than τ_i^x. Hence, 0 is used in the first case of (6). But if $R_i^{up} > T_i - T_j$, there can be only one instance τ_j^h of τ_j with lower priority than τ_i^x. Hence, $\left\lceil \frac{R_i^{up}}{T_i} \right\rceil + 1$ in (4) is replaced with 1 in the second case in (6). Claim follows.

Claim 8 *The response time of a job τ_i^x, during an interval $R_i^{up} \leq T_i$, under PNF/G-RMA is upper bounded by:*

$$R_i^{up} = c_i + RC(R_i^{up}) + D(\tau_i^x) + \left\lfloor \frac{1}{m} \sum_{\forall j \neq i, p_j > p_i} W_{ij}(R_i^{up}) \right\rfloor \tag{7}$$

where $RC(L)$ is calculated by (3), $D(\tau_i^x)$ is calculated by (4), and $W_{ij}(R_i^{up})$ is calculated by (2) in [7].

Proof 8 Proof is same as of Claim 7, except that G-RMA assigns fixed priorities. Hence, (4) can be used directly for calculating $D(\tau_i^x)$ without modifications. Claim follows.

There are two sources of retry cost for any τ_i^x under LCM and lock-free. First is due to conflict between τ_i^x's transactions and transactions of other jobs. This is denoted as RC. Second is due to the preemption of any transaction in τ_i^x due to the release of a higher priority job τ_j^h. This is denoted as RC_{re}. Retry due to the release of higher priority jobs do not occur under PNF, because executing transactions are non-preemptive. It is up to the implementation of the contention manager to safely avoid RC_{re}. LCM does not avoid RC_{re}. We introduce RC_{re} for LCM first before comparing PNF with other techniques.

Claim 9 *Under G-EDF/LCM the total retry cost suffered by all transactions in any τ_i^x during an interval $L \leq T_i$ is upper bounded by:*

$$RC_{to}(L) = RC(L) + RC_{re}(L) \tag{8}$$

where $RC(L)$ is the retry cost resulting from conflict between transactions in τ_i^x and transactions of other jobs. $RC(L)$ is calculated by (5) in [6] for G-EDF/LCM. γ_i and θ_i are replaced with γ_i^{ex} and θ_i^{ex}, respectively. $RC_{re}(L)$ is the retry cost resulting from the release of higher priority jobs, which preempt τ_i^x. $RC_{re}(L)$ is:

$$RC_{re}(L) = \sum_{\forall \tau_j \in \zeta_i} \begin{cases} \left\lceil \frac{L}{T_j} \right\rceil s_{i_{max}} & , L \leq T_i - T_j \\ \\ \left\lfloor \frac{T_i}{T_j} \right\rfloor s_{i_{max}} & , L > T_i - T_j \end{cases} \tag{9}$$

where $\zeta_i = \{\tau_j : (\tau_j \neq \tau_i) \wedge (D_j < D_i)\}$.

Proof 9 Two conditions must be satisfied for any τ_j^l to be able to preempt τ_i^x under G-EDF: $r_i^x < r_j^l < d_i^x$, and $d_j^l \leq d_i^x$. Without the first condition, τ_j^l would have been already released before τ_i^x. Thus, τ_j^l will not preempt τ_i^x. Without the second condition, τ_j^l will be of lower priority than τ_i^x and will not preempt it. If $D_j \geq D_i$, then there will be at most one instance τ_j^l with higher priority than τ_i^x. τ_j^l must have been released at most at r_i^x, which violates the first condition. The other instance τ_j^{l+1} would have an absolute deadline greater than d_i^x. This violates the second condition. Hence, only tasks with shorter relative deadline than D_i are considered. These jobs are grouped in ζ_i.

The total number of released instances of τ_j during any interval $L \leq T_i$ is $\left\lceil \frac{L}{T_i} \right\rceil + 1$. The "carried-in" jobs (i.e., each job released before r_i^x and has an absolute deadline before d_i^x [1]) are discarded as they violate the first condition. The "carried-out" jobs (i.e., each job released after r_i^x and has an absolute deadline after d_i^x [1]) are also discarded because they violate the second condition. Thus, the number of considered higher priority instances of τ_j during the interval $L \leq T_i - T_j$ is $\left\lceil \frac{L}{T_j} \right\rceil$. The number of considered higher priority instances of τ_j during interval $L > T_i - T_j$ is $\left\lfloor \frac{T_i}{T_j} \right\rfloor$.

The worst RC_{re} for τ_i^x occurs when τ_i^x is always interfered at the end of execution of its longest atomic section, $s_{i_{max}}$. τ_i^x will have to retry for $len(s_{i_{max}})$. The total retry cost suffered by τ_i^x is the combination of RC and RC_{re}.

Claim 10 *Under G-RMA/LCM, the total retry cost suffered by all transactions in any τ_i^x during an interval $L \leq T_i$ is upper bounded by:*

$$RC_{to}(L) = RC(L) + RC_{re}(L) \tag{10}$$

where $RC(L)$ and $RC_{re}(L)$ are defined in Claim 9. $RC(L)$ is calculated by (8) in [6] for G-RMA/LCM. $RC_{re}(L)$ is calculated by:

$$RC_{re}(L) = \sum_{\forall \tau_j \in \zeta_i^*} \left(\left\lceil \frac{L}{T_j} \right\rceil s_{i_{max}} \right) \tag{11}$$

where $\zeta_i^* = \{\tau_j : p_j > p_i\}$.

Proof 10 The proof is the same as that for Claim 9, except that G-RMA uses static priority. Thus, the carried-out jobs will be considered in the interference with τ_i^x. The carried-in jobs are still not considered because they are released before r_i^x. Claim follows.

Claim 11 *Consider lock-free synchronization. Let $r_{i_{max}}$ be the maximum execution cost of a single iteration of any retry loop of τ_i. RC_{re} under G-EDF with lock-free synchronization is calculated by (9), where $s_{i_{max}}$ is replaced by $r_{i_{max}}$. RC_{re} under G-RMA with lock-free synchronization is calculated by (11), where $s_{i_{max}}$ is replaced by $r_{i_{max}}$.*

Proof 11 The interference pattern of higher priority jobs to lower priority jobs is the same in G-EDF/LCM, and G-EDF with lock-free. The pattern is also the same in G-RMA/LCM, and G-RMA with lock-free.

$$\begin{aligned} g(\tau_i) \quad = \quad & \left(\sum_{\forall \tau_j \in \gamma_i^*} \sum_{\theta \in \theta_i^*} \left(\left\lceil \frac{T_i}{T_j} \right\rceil \sum_{\forall s_j^{\bar{k}}(\theta)} len\Big(s_j^{\bar{k}}(\theta) \right.\right. \\ & \left.\left. + \quad s_{max}(\theta)\Big) \right) \right) + RC_{re}(T_i) \end{aligned}$$

where RC_{re} is given by (9). $g(\tau_i)$ includes effect of transitive retry. Let:

$$\eta_1(\tau_i) = \sum_{\forall \tau_j \in \gamma_i} \sum_{\forall \theta \in \theta_i} \left(\sum_{\forall s_j^{\bar{k}}(\theta)} len\left(s_j^{\bar{k}}(\theta)\right) \right)$$

$$\eta_2(\tau_i) = \sum_{\forall \tau_j \in \gamma_i} \sum_{\forall \theta \in \theta_i} \left(\left\lceil \frac{T_i}{T_j} \right\rceil \sum_{\forall \bar{s_j^k}(\theta)} len\left(s_{max}^j(\theta)\right) \right)$$

$$\eta_3(\tau_i) = \sum_{\forall \tau_j \in \gamma_i} \sum_{\forall \theta \in \theta_i} \left(\left\lceil \frac{T_i}{T_j} \right\rceil \sum_{\forall \bar{s_j^k}(\theta)} len\left(\bar{s_j^k}(\theta)\right) \right)$$

By substitution of $g(\tau_i)$, $\eta_1(\tau_i)$, and $\eta_2(\tau_i)$, and subtraction of $\sum_{\forall \tau_i} \frac{\eta_3(\tau_i)}{T_i}$ from both sides of (**??**), we get:

$$\sum_{\forall \tau_i} \frac{\eta_1(\tau_i)}{T_i} \leq \sum_{\forall \tau_i} \frac{\eta_2(\tau_i) + g(\tau_i)}{T_i} \tag{12}$$

Assume that $g(\tau_i)_{\forall \tau_i} \to 0$. From (12), we note that by keeping every $len(\bar{s_j^k}(\theta)) \leq len(s_{max}^j(\theta))$ for each τ_i, $\tau_j \in \gamma_i$, and $\theta \in \theta_i$, (12) holds. Due to G-EDF's dynamic priority, $s_{max}^j(\theta)$ can belong to any task other than τ_j. By keeping $len(\bar{s_j^k}(\theta)) \leq len(s_{max}^j(\theta))$, then 12 holds. By generalizing this condition to any $s_j^k(\theta)$ and $s_{max}^j(\theta)$, then (12) holds if all atomic sections in all tasks have equal lengths. Claim follows.

A. PNF versus G-EDF/LCM

Claim 12 *In the absence of transitive retry, PNF/EDF's schedulability is comparable to G-EDF/LCM's if the conflicting atomic section lengths are approximately equal and all α terms approach 1.*

Proof 12 Assume that $\eta_1(\tau_i)$ and $\eta_3(\tau_i)$ are the same as that defined in the proof of Claim **??**. Let:

$$\begin{aligned} g(\tau_i) \quad &= \quad \left(\sum_{\forall \tau_j \in \gamma_i^*} \sum_{\theta \in \theta_i^*} \left(\left\lceil \frac{T_i}{T_j} \right\rceil \sum_{\forall \bar{s_j^k}(\theta)} len\Big(\bar{s_j^k}(\theta) \right.\right. \\ &+ \quad \left.\left. \alpha_{max}^{ji} s_{max}(\theta)\Big) \right) \right) + RC_{re}(T_i) \end{aligned}$$

$$\eta_2(\tau_i) = \sum_{\forall \tau_j \in \gamma_i} \sum_{\forall \theta \in \theta_i} \left(\left\lceil \frac{T_i}{T_j} \right\rceil \sum_{\forall \bar{s_j^k}(\theta)} len\left(\alpha_{max}^{jl} s_{max}^j(\theta)\right) \right)$$

where α_{max}^{jl} is defined in (2). Following the same steps in the proof of Claim **??**, we get:

$$\sum_{\forall \tau_i} \frac{\eta_1(\tau_i)}{T_i} \leq \sum_{\forall \tau_i} \frac{\eta_2(\tau_i) + g(\tau_i)}{T_i} \tag{13}$$

Assume that $g(\tau_i)_{\forall \tau_i} \to 0$. Thus, we ignore the effect of transitive retry and retry cost due to the release of higher priority jobs. Let $len(\bar{s_j^k}(\theta)) = s_{max}^j(\theta) = s$, and $\alpha_{max}^{jl} = \alpha_{max}^{iy} = 1$ in (13). Then, PNF/EDF's schedulability equals LCM/EDF's schedulability if $\left\lceil \frac{T_i}{T_j} \right\rceil = 1, \forall \tau_i, \tau_j$ (which means equal periods for all tasks). If $\left\lceil \frac{T_i}{T_j} \right\rceil > 1, \forall \tau_i, \tau_j$, PNF/EDF's schedulability is better than LCM/EDF's. PNF/EDF's schedulability becomes more better than LCM/EDF's schedulability if $g(\tau_i)$ is not zero. Claim follows.

B. PNF versus G-RMA/LCM

Claim 13 *In the absence of transitive retry, PNF's schedulability is comparable to G-RMA/LCM's if: 1) lower priority tasks suffer increasing number of conflicts from higher priority tasks, 2) the lengths of the atomic sections increase as task priorities increase, and 3) α terms increase.*

Proof 13 Proof is similar to proof of Claim 12. It is omitted here for brevity.

C. PNF versus Lock-free Synchronization

Lock-free synchronization [4, 7] accesses only one object. Thus, the number of accessed objects per transaction in PNF is limited to one. This allows us to compare the schedulability of PNF with the lock-free algorithm.

$RC_B(T_i)$ in (**??**) is replaced with:

$$\sum_{\forall \tau_j \in \gamma_i} \left(\left(\left\lceil \frac{T_i}{T_j} \right\rceil + 1 \right) \beta_{i,j} r_{max} \right) + RC_{re}(T_i) \tag{14}$$

where $\beta_{i,j}$ is the number of retry loops of τ_j that access the same object as accessed by some retry loop of τ_i [4]. r_{max} is the maximum execution cost of a single iteration of any retry loop of any task [4]. $RC_{re}(T_i)$ is defined in Claim 11. Lock-free synchronization does not depend on priorities of tasks. Thus, (14) applies for both G-EDF and G-RMA systems.

Claim 14 *Let r_{max} be the maximum execution cost of a single iteration of any retry loop of any task [4]. Let s_{max} be the maximum transaction length in all tasks. Assume that each transaction under PNF accesses only one object for once. The schedulability of PNF with either G-EDF or G-RMA scheduler is comparable to the schedulability of lock-free synchronization if $s_{max}/r_{max} \leq 1$.*

Proof 14 The assumption in Claim 14 is made to enable a comparison between PNF and lock-free. Let $RC_A(T_i)$ in (**??**) be replaced with (3) and $RC_B(T_i)$ be replaced with (14). To simplify comparison, (3) is upper bounded by:

$$RC(T_i) = \sum_{\tau_j \in \gamma_i} \left(\left(\left\lceil \frac{T_i}{T_j} \right\rceil + 1 \right) \beta_{i,j}^* s_{max} \right)$$

where $\beta_{i,j}^*$ is the number of times transactions in τ_j accesses shared objects with τ_i. Thus, $\beta_{i,j}^* = \beta_{i,j}$, and (**??**)

will be:

$$\sum_{\forall \tau_i} \frac{\sum_{\tau_j \in \gamma_i} \left(\left(\left\lceil \frac{T_i}{T_j} \right\rceil + 1 \right) \beta_{i,j} s_{max} \right)}{T_i} \leq$$

$$\sum_{\forall \tau_i} \frac{\sum_{\forall \tau_j \in \gamma_i} \left(\left\lceil \frac{T_i}{T_j} \right\rceil + 1 \right) \beta_{i,j} r_{max} + RC_{re}(\tau_i)}{T_i} \qquad (15)$$

From (15), we note that if $s_{max} \leq r_{max}$, then (15) holds.

VI. Experimental Evaluation

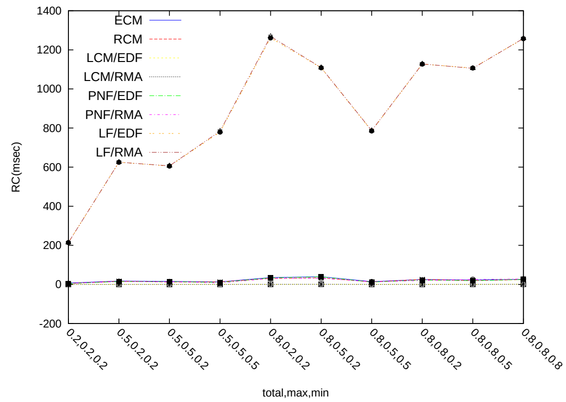
We now would like to understand how PNF's retry cost compares with competitors in practice (i.e., on average). Since this can only be understood experimentally, we implement PNF and the competitors and conduct experiments.

We used the ChronOS real-time Linux kernel [3] and the RSTM library [12] in our implementation. We implemented G-EDF and G-RMA schedulers in ChronOS, and modified RSTM to include implementations of ECM, RCM, LCM, and PNF. For the retry-loop lock-free synchronization, we used a loop that reads an object and attempts to write to it using a CAS instruction. The task retries until the CAS succeeds. We used an 8 core, 2GHz AMD Opteron platform. The average time taken for one write operation by RSTM on any core is $0.0129653375 \mu s$, and the average time taken by one CAS-loop operation on any core is $0.0292546250\ \mu s$.

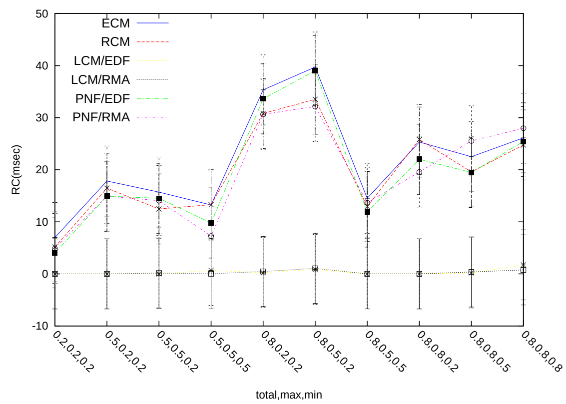
We used four task sets consisting of 4, 5, 8, and 20 periodic tasks. Each task runs in its own thread and has a set of atomic sections. Atomic section properties are probabilistically controlled using three parameters: the maximum and minimum lengths of any atomic section within a task, and the total length of atomic sections within any task. Since lock-free synchronization cannot handle more than one object per atomic section, we first compare PNF's retry cost with that of lock-free (and other CMs) for one object per transaction. We then compare PNF's retry cost with that of other CMs for multiple objects per transaction.

Figures 1 and 2 show the average retry cost for the 5 task and 4 task case, respectively, under 1 and 5 shared objects, respectively. On the x-axis of the figures, we record 3 parameters x, y, and z. x is the ratio of the total length of all atomic sections of a task to the task WCET. y is the ratio of the maximum length of any atomic section of a task to the task WCET. z is the ratio of the minimum length of any atomic section of a task to the task WCET. Confidence level of all data points is 0.95.

While Figure 1(a) includes all methods, Figure 1(b) excludes lock-free. From these figures, we observe that lock-free has the largest retry cost, as it provides no conflict resolution. LCM is better than the others. PNF's retry cost closely approximates ECM's and RCMs, as there is



(a) ECM, RCM, LCM, PNF, Lock-Free



(b) ECM, RCM, LCM, PNF

Fig. 1. Avg. retry cost (one object/transaction).

no transitive retry. From Figure 2, we observe that PNF has shorter or comparable retry cost than ECM, RCM, and LCM. Similar trends were observed for the other task sets. Those are omitted here for brevity, and are available in [5].

VII. Conclusions

Transitive retry increases transactional retry cost under ECM, RCM, and LCM. PNF avoids transitive retry by avoiding transactional preemptions. PNF reduces the priority of aborted transactions to enable other tasks to execute, increasing processor usage. Executing transactions are not preempted due to the release of higher priority jobs. On the negative side of PNF, higher priority jobs can be blocked by executing transactions of lower priority jobs.

EDF/PNF's schedulability is equal or better than ECM's when atomic section lengths are almost equal. RMA/PNF's schedulability is equal or better than RCM's when lower priority jobs suffer greater conflicts from higher priority ones. Similar conditions hold for the schedulability comparison between PNF and LCM, in ad-

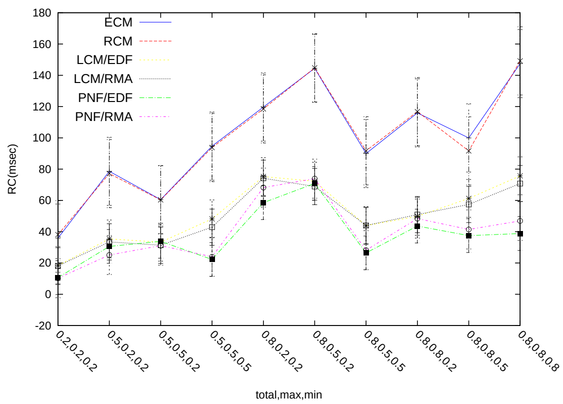


Fig. 2. Avg. retry cost (5 shared objects, 4 tasks).

dition to the increase of α terms to 1. This is logical as LCM with G-EDF (G-RMA) defaults to ECM (RCM) with $\alpha \to 1$. For PNF's schedulability to be equal or better than lock-free, the upper bound on s_{max}/r_{max} must be 1, instead of 0.5 under ECM and RCM.

References

[1] M. Bertogna and M. Cirinei. Response-time analysis for globally scheduled symmetric multiprocessor platforms. In *RTSS*, pages 149–160, 2007.

[2] R. I. Davis and A. Burns. A survey of hard real-time scheduling for multiprocessor systems. *ACM Comput. Surv.*, 43(4):35:1–35:44, Oct. 2011.

[3] M. Dellinger, P. Garyali, and B. Ravindran. Chronos linux: a best-effort real-time multiprocessor linux kernel. In *Proceedings of the 48th Design Automation Conference*, DAC '11, pages 474–479, New York, NY, USA, 2011. ACM.

[4] U. Devi, H. Leontyev, and J. Anderson. Efficient synchronization under global edf scheduling on multiprocessors. In *18th Euromicro Conference on Real-Time Systems*, pages 10 pp. –84, 0-0 2006.

[5] M. El-Shambakey and B. Ravindran. On the design of real-time stm contention managers. Technical report, ECE Department, Virginia Tech, 2011. Available as `http://www.real-time.ece.vt.edu/tech-report-rt-stm-cm11.pdf`.

[6] M. El-Shambakey and B. Ravindran. Stm concurrency control for embedded real-time software with tighter time bounds. In *Proceedings of the 49th Annual Design Automation Conference*, DAC '12, pages 437–446, New York, NY, USA, 2012. ACM.

[7] M. El-Shambakey and B. Ravindran. Stm concurrency control for multicore embedded real-time software: time bounds and tradeoffs. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 1602–1609, New York, NY, USA, 2012. ACM.

[8] S. Fahmy and B. Ravindran. On stm concurrency control for multicore embedded real-time software. In *International Conference on Embedded Computer Systems*, SAMOS, pages 1 –8, July 2011.

[9] R. Guerraoui, M. Herlihy, and B. Pochon. Toward a theory of transactional contention managers. In *PODC*, pages 258–264, 2005.

[10] T. Harris, S. Marlow, S. P. Jones, and M. Herlihy. Composable memory transactions. *Commun. ACM*, 51:91–100, Aug 2008.

[11] M. Herlihy. The art of multiprocessor programming. In *PODC*, pages 1–2, 2006.

[12] V. Marathe, M. Spear, C. Heriot, A. Acharya, D. Eisenstat, W. Scherer III, and M. Scott. Lowering the overhead of nonblocking software transactional memory. In *Workshop on Languages, Compilers, and Hardware Support for Transactional Computing*, TRANSACT.