

UNIVERSITY OF CALGARY

Dynamic Task Allocation in Asynchronous Shared Memory

by

Longsheng Zhou

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF Master of Science

DEPARTMENT OF Computer Science

CALGARY, ALBERTA

monthname, 2016

© Longsheng Zhou 2016

# Abstract

TBD

# Acknowledgements

TBD

# Table of Contents

<b>Abstract</b>	i
<b>Acknowledgements</b>	ii
Table of Contents	iii
1 Introduction	1
1.1 Related Work	1
1.2 Statement of Results	1
2 Model of Computation and Definitions	2
2.1 Asynchronous Shared Memory Model	2
2.2 Base Objects	5
2.3 Adversary Models for Randomized Algorithms	6
2.4 The Dynamic Task Allocation Problem	8
2.4.1 Task	8
2.4.2 The Type DTA	9
2.4.3 Progress Conditions	10
3 Data Structure and Implementation	11
4 Analysis	14
4.1 Correctness Proof	14
4.2 Performance Analysis	14
4.3 Competitive Analysis	14
5 Conclusions and Future Work	15

# Chapter 1

## Introduction

Asynchronous *task allocation* problem, also called *do-all* problem[5], is defined informally as the problem of  $n$  processes in the network, cooperatively performing  $m$  independent tasks, in the presence of adversity.

Such cooperation problems consisting of large numbers of tasks by multiple processes is highly related to a broad range of distributed computing problems, such as mutual exclusion [4], consensus problem [11] distributed clocks [3], and shared-memory collect [1].

In shared memory models, the task allocation problem is known as *Write-All* problem, introduced and studied by Kanellakis and Shvartsman [10] and defined as follows: Given a zero-valued array of  $m$  elements and  $n$  processors, write value 1 into each array location in the presence of adversity.

Following the initial work [10], the task allocation problem was studied in a variety of shared memory settings e.g., [5, 7, 14, 51, 65, 68, 69, 82, 87, 88, 89].

### 1.1 Related Work

### 1.2 Statement of Results

## Chapter 2

### Model of Computation and Definitions

#### 2.1 Asynchronous Shared Memory Model

In this chapter, we will describe our model of computation and give the definitions, which are based on Herlihy and Wing's [8] and Golab, Higham and Woelfel's [7].

The computational model we consider is the standard asynchronous shared memory model with a set  $\mathcal{P}$  of  $n$  processes, denoted as  $\mathcal{P} = [p] = \{0, 1, 2, \dots, n - 1\}$ , where up to  $n - 1$  processes may fail by crashing. A process may crash at any moment during the computation and once crashed it does not restart, and does not perform any further actions.

**Type and Object.** A *type*  $\tau$  is defined as an automaton as follows [6],

$$\tau = (\mathcal{S}, s_{init}, \mathcal{O}, \mathcal{R}, \delta)$$

where  $\mathcal{S}$  is a set of states,  $s_{init} \in \mathcal{S}$  is the initial state,  $\mathcal{O}$  is a set of operations,  $\mathcal{R}$  is the set of responses, and  $\delta : \mathcal{S} \times \mathcal{O} \rightarrow \mathcal{S} \times \mathcal{R}$  is a state transition mapping.

An *object* is an implementation of a type. For each type  $\tau$ , the transition mapping  $\delta$  captures the behaviour of objects of type  $\tau$ , in the absence of concurrency, as follows: if a process applies an operation  $opt$  to an object of type  $\tau$  which is in state  $s$ , the object may return to the process a response  $rsp$  and change its states to  $s'$  if and only if  $(s', rsp) \in \delta(s, opt)$ .

**History.** A *history*  $H$ , obtained by processes executing operations on objects, is a sequence of invocation and response events.

An invocation event is a 5-tuple,

$$\text{INV} = (\text{invocation}, p, \text{obj}, \text{opt}, t)$$

where *invocation* is the event type, *p* is the process executing the operation, *obj* is the object on which the operation is executed, *opt* is the operation and *t* is the *time* when INV happens which is defined as the position of event INV in history *H*. We also say the event INV is the invocation event of operation *opt*.

A response event is also a 5-tuple,

$$\text{RSP} = (\text{response}, p, \text{obj}, \text{rsp}, t)$$

where *response* is the event type, *p* is the process receiving response *rsp* from an operation on object *obj* and *t* is the time when RSP happens which is defined as the position of event RSP in history *H*.

In the following discussion, we suppose in a history *H*, the situation that an invocation event  $(\text{invocation}, p_i, \text{obj}_p, \text{opt}_0, t_0)$  is followed immediately by another invocation event  $(\text{invocation}, p_j, \text{obj}_q, \text{opt}_1, t_1)$  where  $i = j$  and  $p = q$  will not happen.

We say response event  $(\text{response}, p_j, \text{obj}_q, \text{rsp}, t_1)$  *matches* invocation event  $(\text{invocation}, p_i, \text{obj}_p, \text{opt}_0, t_0)$  in history *H*, if the two events are applied by the same process to the same object, i.e,  $i = j$  and  $p = q$ . In this case, the response event is also called the *matching* response of the invocation event.

An *operation execution* in *H* is a pair  $oe = (\text{INV}, \text{RSP})$  consisting of an invocation event INV and its matching response event RSP, or just an invocation event INV with no matching response event, denoted as  $oe = (\text{INV}, \text{null})$ .

In the latter case, we say the operation execution is *pending*. In the former case, we say the operation execution is *complete*.

A history *H* is *complete* if all operation executions in *H* are *complete*. Otherwise, it is

*incomplete*.

If events INV and RSP are applied by process  $p$ , then we say operation execution  $oe = (INV, RSP)$  is *performed* by process  $p$ . Thus, two operation executions performed by the same process on the same project will not interleave in a history  $H$ .

We say that an operation  $opt$  is *atomic* in history  $H$ , if  $opt$ 's invocation event is either the last event in  $H$ , or else is followed immediately in  $H$  by a matching response event.

If history  $H$  is a prefix of history  $H'$ , then we say history  $H'$  is an extension of  $H$ . History  $H'$  is a *completion* of history  $H$  if  $H'$  contains all events in  $H$  and  $H'$  is an extension of  $H$ , and each operation execution in  $H'$  is complete.

$H|obj$  of history  $H$  is the subsequence of all invocation and response events in  $H$  on object  $obj$ . If all invocation and response events in a history  $H$  have the same object name  $obj$ , then the  $H|obj = H$ .

Let  $H$  be a complete history. We associate a time interval  $I_{oe} = [t_0, t_1]$  with each operation execution  $oe = (INV, RSP)$  in  $H$ , where  $t_0$  and  $t_1$  are the points in time when INV and RSP happen. Similarly, for an incomplete history, we denote the time interval  $I_{oe}$  with respect to a pending operation execution  $oe = (INV, null)$  by  $I_{oe} = [t_0, \infty]$ .

Operation execution  $oe_0$  *precedes* operation execution  $oe_1$  in  $H$  if the response event of  $oe_0$  happens before the invocation event of  $oe_1$  in  $H$ . We say that  $oe_0$  and  $oe_1$  are *concurrent* in  $H$  if neither precedes the other.

A history is *sequential* if its first event is an invocation event, and each invocation event, except possibly the last one, is immediately followed by a matching response event.

A *sequential specification* of an object is the set of all possible sequential histories for that object.

A sequential history  $S$  is *valid*, if for each object  $obj$ ,  $S|obj$  is in the sequential specification



of *obj*.

**Linearization.** A history  $H$  *linearizes* to a sequential history  $S$ , if and only if  $S$  satisfies the following conditions:

- $S$  and any completion of  $H$  have the same operation executions,
- sequential history  $S$  is valid, and
- there is a mapping from each time interval  $I_{oe}$  to a time point  $t_{oe} \in I_{oe}$ , such that the sequential history  $S$  is obtained by sorting the operations in  $H$  based on their  $t_{oe}$  values.

A history is *linearizable* if and only if  $H$  linearizes to some sequential history  $S$ . In this case,  $S$  is called the *linearization* of  $H$ . For each operation *opt* in history  $H$ , we call time point  $t_{oe}$ , which is defined as above, the *linearization point* of *opt*. An object *obj* is linearizable if every history  $H$  on *obj* is linearizable.

## 2.2 Base Objects

In this section, we describe the two base objects, i.e, *read-write register* and *compare-and-swap* (CAS) objects, which will be used in our following discussion. Most implementations of more sophisticated objects use them as the base objects in their implementations and most modern architectures support either read-write registers and CAS objects [9] [12].

**Read-Write Register.** An object that supports only `read()` and `write(x)` operations is called a read-write register (or just *register*).

Operation `read()` returns the current state of register and leaves the state unchanged. Operation `write(x)` changes the state of the register to  $x$  and returns nothing. If the set of

states that can be stored in the register is unbounded then we say the register is *unbounded register*; otherwise the register is *bounded register*.

**CAS Object.** An object that supports `read()` and `CAS(x,y)` operations is called compare-and-swap (CAS) object.

Operation `read()` returns the current state of CAS object, and leaves the state unchanged, while operation `CAS(x,y)` changes the state of the object if and only if the current state is equal to  $x$  and then operation `CAS(x,y)` succeeds, and the state is changed to  $y$  and *true* is returned. Otherwise, operation `CAS(x,y)` fails, the current state remains unchanged and *false* is returned.

## 2.3 Adversary Models for Randomized Algorithms

**Randomness.** A randomized algorithm is an algorithm where processes are allowed to make random decisions for future steps by calling a special operation called *coin-flip operation*. We also say a process *flips a coin* when it calls this operation.

When a process flips a coin, it receives a random value  $c$  from some arbitrary countable set  $\Omega$ , which is the *coin-flip domain*. The process can then use this random value  $c$  in its program for future decisions.

A vector  $\vec{c} = (c_0, c_1, c_2, \dots) \in \Omega^\infty$  is called a *coin-flip vector*. A history  $H$  is said to *observe* the coin-flip vector  $\vec{c}$  if for any integer  $i \in \{0, 1, 2, \dots\}$ , the  $i$ -th coin-flip operation in  $H$  returns value  $c_i \in \Omega$ .

For a history  $H$  that contains  $k$  coin-flip operations, we use  $H[k]$  to denote the prefix of  $H$  that ends with the  $k$ -th invocation of a coin-flip operation. If fewer than  $k$  coin-flips occur during  $H$ , then  $H[k]$  denotes  $H$ .

**Schedule.** In the standard shared memory model, each process executes its program by

applying shared memory operations (`read()`, `write(x)`, `CAS(x,y)`, etc) on objects, as determined by their program. Operation executions of concurrent processes can be interleaved arbitrarily.

A *schedule* with *length*  $k$  is represented by a sequence of process IDs

$$p = (p_0, p_1, p_2, \dots, p_{k-1})$$

where  $k \in \{1, 2, 3, \dots\}$  and for each  $i \in \{0, 1, \dots, k-1\}$ ,  $p_i \in \mathcal{P}$ .

Consider a schedule  $p = (p_0, p_1, p_2, \dots, p_{k-1})$ . A history  $H$  is said to *observe* schedule  $p$  if the number of events in  $H$  is  $k$ , and for each integer  $i \in \{0, 1, \dots, k-1\}$ , the  $i$ -th event is applied by process  $p_i$ .

**Adversary.** In a randomized algorithm, the random choices processes make can influence the schedule. To model the worst possible way that the system can be influenced by the random choices, schedules are assumed to be generated by an adversarial scheduler, called the *adversary*.

Mathematically, an adversary is defined as a mapping [7]:

$$\mathcal{A} : \Omega^\infty \rightarrow \mathcal{P}^\infty$$

Given an algorithm  $\mathcal{M}$ , an adversary  $\mathcal{A}$ , and a coin-flip vector  $\vec{c} \in \Omega^\infty$ , a unique history  $H_{\mathcal{M}, \mathcal{A}, \vec{c}}$  is generated, such that all processes apply events as dictated by algorithm  $\mathcal{M}$ , and history  $H_{\mathcal{M}, \mathcal{A}, \vec{c}}$  observes the schedule  $\mathcal{A}(\vec{c})$  and the coin flip vector  $\vec{c}$ .

There are several adversary models with different strengths [2]. In our thesis, we only consider the *adaptive adversary*.

Informally, the adaptive adversary makes scheduling decisions as follows: At any point, it can see the entire history up to that point. This includes all coin-flip operations and their return values up to that point. Depending on this, the adversary decides which process takes

the next step.

Adversary  $\mathcal{A}$  is *adaptive for algorithm  $\mathcal{M}$*  [7] if, for any two coin-flip vectors  $\vec{c} \in \Omega^\infty$  and  $\vec{d} \in \Omega^\infty$  that have a common prefix of length  $k$  (i.e, the first  $k$  elements of  $\vec{c}$  and  $\vec{d}$  are the same), then we have

$$H_{\mathcal{M}, \mathcal{A}, \vec{c}}[k+1] = H_{\mathcal{M}, \mathcal{A}, \vec{d}}[k+1]$$

In this case, we say adversary  $\mathcal{A}$  is an *adaptive adversary*.

From the above definition, we can see an adaptive adversary cannot use future coin flips to make current scheduling decisions.

## 2.4 The Dynamic Task Allocation Problem

### 2.4.1 Task

A *task* is a computation which is assumed to be performed by a single process in constant time[5]. In this thesis, we consider a finite or infinite set of tasks, denoted as  $\mathcal{L} = [m] = \{0, 1, \dots, m-1\}$ , where  $m \in \{0, 1, 2, \dots\}$ .

We assume that each task  $\ell \in \mathcal{L}$  to be performed is associated with a *location*  $M$  in the data structure. Over time, one location can be associated with multiple tasks.

In this section, we are going to specify the dynamic task allocation problem in terms of a type DTA which supports two types of operations **DoTask** and **InsertTask**, and the properties that an implementation of type DTA must satisfy. But before that, we firstly fix an interface by which processes could perform a task, or insert a new task to data structure.

**Operation TryTask( $M$ ).** A process can perform a task  $\ell$  atomically by calling a special **TryTask( $M$ )** operation where  $M$  is a location which task  $\ell$  is associated with.

If the location  $M$  is associated with a task  $\ell$ , then notification *success* will be returned by

**TryTask( $M$ )**. Otherwise, if there is no task associated with location  $M$ , then *failure* will be returned.

**Operation PutTask( $M, \ell$ )**. A process can associate a task  $\ell$  with a location  $M$  in the data structure atomically by calling a special **PutTask( $M, \ell$ )** operation.

If location  $M$  is not associated with any other task, then **PutTask( $M, \ell$ )** will return *success*. Otherwise, if location  $M$  is already associated with another task  $\ell'$ , then *failure* is returned by **PutTask( $M, \ell$ )** call.

#### 2.4.2 The Type DTA

The type DTA supports two types of operations. The **DoTask()** operation performs a task and returns the identifier of that task, while the **InsertTask( $\ell$ )** operation associates task  $\ell$  with a location in the data structure to be performed. Now we describe the sequential specification as follows.

**Operation DoTask()**. The aim of the operation **DoTask()** is to find a location  $M$  which is associated with a task  $\ell$  in the data structure and then perform task  $\ell$  by calling the atomic operation **TryTask( $M$ )**.

Every **DoTask()** operation may perform several **TryTask** operations with different locations as the arguments. Once a **TryTask** operation succeeds, **DoTask()** terminates and the task identifier  $\ell$  is returned. Otherwise, **DoTask()** never terminates and keeps calling **TryTask( $M$ )** repeatedly. If there is no task in the data structure, then **DoTask()** returns  $\perp$ .

A task  $\ell$  is said to be *performed* by a process if the process has completed a **DoTask()** call which returns the task identifier  $\ell$ .

**Operation InsertTask( $\ell$ )**. The goal of the **InsertTask( $\ell$ )** operation is to find a location  $M$  in the data structure and associates task  $\ell$  with  $M$  by executing operation **PutTask( $M, \ell$ )**.

Every `InsertTask( $\ell$ )` operation may perform several `PutTask` operations with different locations as the arguments. Operation `InsertTask( $\ell$ )` terminates once a `PutTask` operation succeeds and then location  $M$  will be returned by `InsertTask( $\ell$ )`. Otherwise, `InsertTask( $\ell$ )` never terminates and keeps calling `PutTask` operations repeatedly.

We say task  $\ell$  is associated with a location  $M$  or inserted into the data structure or task  $\ell$  is *available* to perform if a process has completed the `InsertTask( $\ell$ )` call and the location  $M$  is returned, but task  $\ell$  has not been performed yet.

### 2.4.3 Progress Conditions

## Chapter 3

### Data Structure and Implementation

Data Structure

The Implementation of Type DTA

---

**Method 1: DoTask()**

---

```
1 while true do
2    $v \leftarrow \text{root};$ 
3   if  $v.\text{surplus}() \leq 0$  then
4     return  $\perp$ ;
5   end
6   /* Descent */;
7   while  $v$  is not a leaf do
8      $(x_L, y_L) \leftarrow v.\text{left}.\text{read}();$ 
9      $(x_R, y_R) \leftarrow v.\text{right}.\text{read}();$ 
10     $s_L \leftarrow \min(x_L - y_L, 2^{\text{height}(v)});$ 
11     $s_R \leftarrow \min(x_R - y_R, 2^{\text{height}(v)});$ 
12     $r \leftarrow \text{random}(0, 1);$ 
13    if  $(s_L + s_R) = 0$  then
14      Mark-up( $v$ );
15    else if  $r < s_L / (s_L + s_R)$  then
16       $v \leftarrow v.\text{left};$ 
17    else
18       $v \leftarrow v.\text{right};$ 
19    end
20  end
21  /* v is a leaf */;
22   $(x, y) \leftarrow v.\text{read}();$ 
23   $(\text{flag}, \ell) \leftarrow v.\text{TryTask}(\text{task}[y + 1]);$ 
24  /* Update Insertion Count */;
25   $v.\text{CAS}((x, y), (x, y + 1));$ 
26   $v \leftarrow v.\text{parent};$ 
27  Mark-up( $v$ );
28  if  $\text{flag} = \text{success}$  then
29    return  $\ell$ 
30  end
31 end
```

---



---

**Method 2: InsertTask( $\ell$ )**

---

```
32 while true do
33    $v \leftarrow \text{root};$ 
34   /* Descent */;
35   while  $v$  is not a leaf do
36      $(x_L, y_L) \leftarrow v.\text{left.read}();$ 
37      $(x_R, y_R) \leftarrow v.\text{right.read}();$ 
38      $s_L \leftarrow 2^{\text{height}(v)} - \min(x_L - y_L, 2^{\text{height}(v)});$ 
39      $s_R \leftarrow 2^{\text{height}(v)} - \min(x_R - y_R, 2^{\text{height}(v)});$ 
40      $r \leftarrow \text{random}(0, 1);$ 
41     if  $(s_L + s_R) = 0$  then
42       |  $\text{Mark-up}(v);$ 
43     else if  $r < s_L / (s_L + s_R)$  then
44       |  $v \leftarrow v.\text{left};$ 
45     else
46       |  $v \leftarrow v.\text{right};$ 
47     end
48   end
49   /* v is a leaf */;
50    $(x, y) \leftarrow v.\text{read}();$ 
51    $\text{flag} \leftarrow v.\text{PutTask}(\text{task}[x + 1]);$ 
52   /* Update Insertion Count */;
53    $v.\text{CAS}((x, y), (x + 1, y));$ 
54    $v \leftarrow v.\text{parent};$ 
55    $\text{Mark-up}(v);$ 
56   if  $\text{flag} = \text{success}$  then
57     | return success
58   end
59 end
```

---

---

**Method 3: Mark-up( $v$ )**

---

```
60 if  $v$  is not null then
61   for  $(i = 0; i < 2; i++)$  do
62     |  $(x, y) \leftarrow v.\text{read}();$ 
63     |  $(x_L, y_L) \leftarrow v.\text{left.read}();$ 
64     |  $(x_R, y_R) \leftarrow v.\text{right.read}();$ 
65     |  $v.\text{CAS}((x, y), (\max(x, x_L + x_R), \max(y, y_L + y_R)));$ 
66   end
67 end
```

---

## Chapter 4

### Analysis

#### 4.1 Correctness Proof

#### 4.2 Performance Analysis

DoTask Analysis

InsertTask Analysis

#### 4.3 Competitive Analysis

## Chapter 5

### Conclusions and Future Work

# Bibliography

- [1] Miklos Ajtai, James Aspnes, Cynthia Dwork, and Orli Waarts. A theory of competitive analysis for distributed algorithms. In *Foundations of Computer Science, 1994 Proceedings., 35th Annual Symposium on*, pages 401–411. IEEE, 1994.
- [2] James Aspnes. Randomized protocols for asynchronous consensus. *CoRR*, cs.DS/0209014, 2002.
- [3] Yonatan Aumann and Michael O Rabin. Clock construction in fully asynchronous parallel systems and pram simulation. In *Foundations of Computer Science, 1992. Proceedings., 33rd Annual Symposium on*, pages 147–156. IEEE, 1992.
- [4] Michael Bender, Seth Gilbert, et al. Mutual exclusion with  $o(\log^2 \log n)$  amortized work. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 728–737. IEEE, 2011.
- [5] Chryssis Georgiou. *Do-All Computing in Distributed Systems: Cooperation in the Presence of Adversity*. Springer Science & Business Media, 2007.
- [6] Wojciech Golab, Vassos Hadzilacos, Danny Hendler, and Philipp Woelfel. Constant-rmr implementations of cas and other synchronization primitives using read and write operations. In *Proceedings of the 26th Annual ACM Symposium on Principles of Distributed Computing (PODC)*, pages 3–12, 2007.
- [7] Wojciech Golab, Lisa Higham, and Philipp Woelfel. Linearizable implementations do not suffice for randomized distributed computation. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 373–382. ACM, 2011.
- [8] Maurice P. Herlihy and Jeannette M. Wing. Linearizability: A correctness condition for concurrent objects. *ACM Trans. Program. Lang. Syst.*, 12(3):463–492, July 1990.

- [9] Itanium. *Intel Itanium Architecture Software Developers Manual Volume 1: Application Architecture Revision 2.1*. Intel Corporation, October 2002.
- [10] Paris C Kanellakis and Alex A Shvartsman. Efficient parallel algorithms can be made robust. *Distributed Computing*, 5(4):201–217, 1992.
- [11] Nancy A Lynch. *Distributed algorithms*. Morgan Kaufmann, 1996.
- [12] David L Weaver and Tom Gremond. *The SPARC architecture manual*. PTR Prentice Hall Englewood Cliffs, NJ 07632, 1994.