Doing-it-All with Bounded Work and Communication

Bogdan S. Chlebus † Leszek Gąsieniec ‡ Dariusz R. Kowalski § Alexander A. Shvartsman ¶

Abstract

We consider the Do-All problem, where \( p \) cooperating processors need to complete \( t \) similar and independent tasks in an adversarial setting. Here we deal with a synchronous message passing system with processors that are subject to crash failures. Efficiency of algorithms in this setting is measured in terms of work complexity (also known as total available processor steps) and communication complexity (total number of point-to-point messages). When work and communication are considered to be comparable resources, then the overall efficiency is meaningfully expressed in terms of effort defined as work + communication. We develop and analyze a constructive algorithm that has work \( O(t + p \log p (\sqrt{\log p} + \sqrt{\log t})) \) and a nonconstructive algorithm that has work \( O(t + p \log^2 p) \). The latter result is close to the lower bound \( \Omega(t + p \log p / \log \log p) \) on work. The effort of each of these algorithms is proportional to its work when the number of crashes is bounded above by \( c p \), for some positive constant \( c < 1 \). We also present a nonconstructive algorithm that has effort \( O(t + p^{1.77}) \).

Key words: distributed algorithm, message passing, crash failures, fault tolerance, scheduling tasks, load balancing, Ramanujan graphs.

*The results of this paper were announced in a preliminary form in [10].
†Department of Computer Science and Engineering, University of Colorado Denver, Denver, Colorado 80217, USA. The work of this author was supported by NSF Grants 0310503 and 1016847, and by the Engineering and Physical Sciences Research Council [grant number EP/G023018/1].
‡Department of Computer Science, University of Liverpool, Liverpool L69 3BX, UK.
§Department of Computer Science, University of Liverpool, Liverpool L69 3BX, UK. The work of this author was supported by the Engineering and Physical Sciences Research Council [grant numbers EP/G023018/1, EP/H018816/1].
¶Department of Computer Science and Engineering, University of Connecticut, Storrs, Connecticut 06269, USA. The work of this author was supported by NSF Grants 0311368 and 1017232.
1 Introduction

Performing a collection of tasks by processors prone to failures is among the fundamental problems in fault-tolerant distributed computing. We consider the problem called Do-All, where the processors cooperate on tasks that are similar, independent, and idempotent. Simple instances of this include checking all the points in a large solution space, attempting either to generate a witness or to refute its existence, and scheduling collections of tasks admitting “at least once” execution semantics. We consider the synchronous message-passing setting, in which $t$ tasks need to be performed by $p$ processors subject to $f$ crash failures, where $f$ can be as large as $p - 1$.

Synchronous message-passing solutions for Do-All were first developed by Dwork et al. \[26\] who estimated the performance of their algorithms in terms of effort defined as the sum of task-oriented work and communication. Task-oriented work counts only the processing steps expended on performing tasks and it discounts any steps spent idling or waiting for messages. Communication costs are measured as the message complexity, calculated as the total number of point-to-point messages. Note that the time of algorithm executions may be very large, for example, when $f$ can be $p - 1$, time may be at least linear in $t$ for the cases where one remaining processor must perform all tasks. Thus time is not normally used to describe the efficiency of Do-All algorithms when large number of failures is allowed. We also measure algorithm performance in terms of effort using the approach of Dwork et al. \[26\], who defined effort as the sum of task-oriented work and communication complexities. In our work we use the more conservative effort measure where instead of task-oriented work we use available processor steps (or total work) of Kanellakis and Shvartsman \[36\] that accounts for all steps taken by each processor, including idling, until the processor either halts or crashes. Thus we define effort of an algorithm to be $W + M$, where $W$ is its total work complexity and $M$ is its message complexity. Other prior research focused on developing Do-All algorithms that are efficient in terms of work, and next dealt with communication efficiency as a secondary goal, e.g., using the lexicographic complexity \[23\].

Interdependence between work and communication in solutions of Do-All is to be expected, because communication in unreliable systems improves coordination among processors, which results in less duplicated work. In message-passing systems, there are two direct ways in which a processor can get to know that a certain task is complete: the processor can either perform the task by itself or it can receive a message telling that the task was completed by another processor. Note that Do-All can be solved without any communication: simply have each processor perform each task. The work of such an algorithm is $O(p \cdot t)$, and this is quadratic when $p$ and $t$ are comparable. On the other hand, an algorithm may cause each processor to always share its knowledge with all other processors in an attempt to reduce work. The may result in efficient work, but the communication complexity of such an algorithm is at least quadratic. Thus it is highly desirable to develop algorithms that have both subquadratic work and subquadratic communication. This makes it meaningful to balance work and communication, and to consider work and communication to be comparable resources. Motivated by these observations, the quantity $W + M$ provides a simple unifying performance metric.

There are three parameters in terms of which the performance bounds of our algorithms are expressed. These are the number of processors $p$, the number of tasks $t$, and finally the number of crashes $f$ that may occur in the course of an execution. Our algorithms place no constraints on the relationship between $t$ and $p$; these parameters are independent in our algorithms and the complexity bounds. The only restriction on $f$ is that the inequality $f < p$ holds, unless stated
We say that a parameter is known when it can be used in the code of an algorithm. The parameters $p$ and $t$ are always known in this paper. When the parameter $f$ is explicitly referred to in performance bounds, then this indicates that the algorithm is designed to optimize performance for the number of crashes that is assumed to be at most $f$. When $f < p$ is used as a parameter then $f$ is known. (Observe that when $f$ is not used in the code of an algorithm as a parameter then the only thing that is known about $f$ is that $f \leq p - 1$.) It so happens that if $f$ appears in the code of an algorithm in this paper, then the corresponding communication complexity depends on $f$, whereas the bounds on the work complexity involve only $p$ and $t$.

We find it useful to consider crashes in an adversarial setting, in which a malicious adversary, causes processors to crash. An adversary is $f$-bounded if $f < p$ is an upper bound on the number of crashes in any execution. When stating a performance bound for a known $f < p$ we normally state that the bound holds “for an $f$-bounded adversary for a known $f < p$.” The $(p - 1)$-bounded adversary is called unbounded. An $f$-bounded adversary is called linearly bounded if $f \leq cp$, for some positive constant $c < 1$. Our algorithms always solve Do-All when exposed to the unbounded adversary, but their message complexity may sometimes be especially efficient when the adversary is linearly-bounded.

We call a deterministic algorithm constructive if its code can be produced by a deterministic sequential algorithm in time polynomial in $t$ and $p$. This is in contrast with nonconstructive algorithms that may rely on combinatorial objects that are only known to exist. Methodologically, starting with a constructive algorithm, we trade constructiveness for better effort bounds in producing nonconstructive algorithms.

In this work, we explore algorithmic solutions for Do-All where good effort performance is attained, rather than seeking just work-efficient solutions. With this as a goal, it turns out that a key challenge, besides tolerating processor crashes and controlling work, is to ensure that communication costs do not exceed work complexity. Whereas any two processors can communicate in any round of computation, we limit the amount of communication in our algorithms by allowing messaging to take place over certain constant-degree subnetworks. To this end, we use constructive graphs with good “expansion” properties, and this contributes to the emerging understanding of how expansion-related properties of the underlying communication schemes can be used to improve fault tolerance.

Our results. We present a new way of structuring algorithms for the $p$-processor, $t$-task Do-All problem that allows for both work and communication to be controlled in the presence of adaptive adversaries. We give a generic algorithm for performing work in systems with crash-prone processors, and we parameterize it by (i) task-assignment rules and (ii) virtual overlay graphs superimposed on the underlying communication medium.

We now overview our contributions in more detail.

1. We present a deterministic constructive algorithm, called BALANCE-LOAD, that uses a balancing task allocation policy (Section 5). This algorithm solves the Do-All problem in any execution with at least one non-crashed processor. It has work $W = \mathcal{O}(t + p \log p (\sqrt{p \log p} + \sqrt{t \log t}))$ in the presence of $f$-bounded adversary, for known $f$, where $f < p$. When the adversary is additionally constrained to be linearly-bounded, the message complexity of the algorithm is $M = \mathcal{O}(W)$. 

2
No prior algorithms attained work sub-quadratic in both $p$ and $t$ (measured as the available processor steps), against the $f$-bounded adversary with unrestricted crashes, for any known $f < p$. By using embedded graphs whose properties depend on the adversaries, each processor in our algorithm sends $O(1)$ messages in each round in the case of linearly-bounded adversaries, making communication comparable to work. (Section 3 discusses the use of embedded graphs in communication.)

II. We develop a deterministic algorithm, called DETERMINISTIC-PERMUTATIONS, that is more efficient than algorithm BALANCE-LOAD, but is nonconstructive (Section 6). This algorithm has work $W = O(t + p \log^2 p)$ against a known $f$-bounded adversary, for any $f < p$. When the adversary is additionally constrained to be linearly-bounded, the message complexity of the algorithm is $M = O(W)$.

We note that the upper bound on work for this algorithm differs from the lower bound $\Omega(t + p \log p / \log \log p)$ given in [12, 32] by a small factor of $\log p \log \log p$.

III. We give a deterministic nonconstructive algorithm, called EFFORT-PRIORITY, that is effort-efficient (sub-quadratic) against the unbounded adversary (Section 7). The effort of this algorithm is $W + M = O(t + p^{1.77})$, for any unknown $f < p$.

This algorithm is obtained by combining algorithm DETERMINISTIC-PERMUTATIONS with the algorithm of De Prisco et al. [23]. This was the first algorithm that obtained both work and communication sub-quadratic in $\max\{p, t\}$ against the unbounded adversary. Subsequently, Georgiou et al. [31] developed an algorithm with effort $O(t + p^{1+\varepsilon})$ against the unbounded adversary, for any known constant $\varepsilon$, using an approach based on gossiping.

Previous and related work. Dwork et al. [26] were the first to study the Do-All problem in the message-passing setting. The paper [26] introduced the idea of measuring the effort complexity defined as the sum of task-oriented work and communication, and gave three algorithms of optimal task-oriented work. De Prisco et al. [23] were the first to use the available processor steps as the work measure for message-passing algorithms. They gave an algorithm with work $O((t + (f + 1)p)$ and communication $O((f+1)p)$. Galil et al. [29] gave an algorithm with better communication cost, $O(fp^\varepsilon + \min\{f + 1, \log p\}p)$, for any $\varepsilon > 0$. A popular algorithmic paradigm used in prior research was to disseminate local knowledge among groups of processors by implementing coordinators, who first collect and then spread information. Chlebus et al. [9] developed algorithms based on aggressive coordination, where the number of appointed coordinators grows exponentially following crashes of all previously appointed coordinators. Their algorithms rely on reliable multicasting, where if a sender crashes while multicasting, then either none or all of the messages are delivered to the non-faulty recipients. One of their algorithms has work $O((t + p \log p / \log \log p) \log f)$ and communication $O(t + p \log p / \log \log p + fp)$; these bounds were improved by Georgiou et al. [32] for $f \leq p / \log \log p$. Another algorithm in [9] incorporates restarted processors. It is the only known algorithm able to deal with restarts efficiently; it has work $O((t + p \log p + f) \cdot \min\{\log p, \log f\})$ and its message complexity is $O(t + p \log p + fp)$. Chlebus et al. [11] gave a randomized solution with work $O(t + p \log^2 p / \log \log p)$ for the $f$-bounded adversary, for any known $f < p$. This work by Chlebus et al. [11] is a companion paper that considers the same generic algorithm and uses the same overlay graphs as in the current paper, but with a randomized rule for tasks selection by processors. Georgiou et
Table 1: A summary of known solutions for Do-All in the message-passing model. Algorithms are deterministic and constructive, unless stated otherwise. Broadcast is not assumed to be reliable, unless stated otherwise. Work is called task-oriented when idling and waiting is not accounted for, but each instance of performing of a task contributes one unit of work. Neither the upper bound $f$ on the number of crashes, nor $\varepsilon$ appear in the code, unless it is stated in the remarks that these quantities are known. The dates are of the first conference announcements of the results.
al. \cite{31} developed an algorithm with effort $O(t + p^{1+\varepsilon})$ against the unbounded adversary, for any known $\varepsilon$, using an approach based on gossiping. A summary of solutions for Do-All related to the results of this paper is given in Table 1. The book \cite{34} by Georgiou and Shvartsman gives a representative account of the topics in Do-All computing.

Table 1 does not include results dealing with specialized adversaries and models of interprocessor communication. We now give an overview of such work. Chlebus and Kowalski \cite{12} studied Do-All with crash failures in the presence of a weakly-adaptive linearly-bounded adversary. They developed a randomized algorithm with $O(t + p(1 + \log^* (p/d)))$ expected effort. This performance is provably better than that of any deterministic algorithm in such a setting that is subject to the lower bound on work of $\Omega(t + p \log p/ \log \log p)$ \cite{12,32}. The paper \cite{12} shows that a deterministic algorithm that schedules tasks by balancing them perfectly has work $O(t + p \log p/ \log \log p)$ against the linearly-bounded adversary; this implies that $\Theta(t + p \log p/ \log \log p)$ is precisely the optimum amount of work in such a setting. Kowalski and Shvartsman \cite{42} studied the problem in the message-passing model with restricted asynchrony, where every message delay is at most $d$, for some value $d$ unknown to the algorithm. They showed that $\Omega(t + p d \log_p d)$ is a lower bound on the expected work and developed a deterministic algorithm with work $O((t + pd) \log p)$.

Georgiou et al. \cite{32} considered an iterated version of Do-All. The same authors studied in \cite{33} a problem, called Omni-Do, that is an on-line version of Do-All in asynchronous systems with partitionable networks. They gave a randomized algorithm achieving optimal competitive ratio against oblivious adversaries. Fernández et al. \cite{28} considered Do-All with Byzantine failures. Chlebus et al. \cite{16} and Clementi et al. \cite{19} studied Do-All in a model where communication is performed over a multiple-access channel. The former \cite{16} considered the impact of collision detection and randomization on the complexity of protocols. The latter \cite{19} gave tight bounds on work and time for such deterministic algorithms against $f$-bounded adversaries.

Gossiping in a crash-prone message-passing system is a paradigm applied in solutions to Do-All and related problems that emerged after the conference version of this paper was published. It was first applied by Chlebus and Kowalski \cite{14} to obtain message-efficient solution to consensus in synchronous message passing. The paper \cite{14} reformulated gossiping in a crash-prone environment as a problem in which each processor starts with a rumor to be ideally learned by all processors but after a processor $v$ crashes then any other processor needs either to learn the rumor of $v$ or the fact that $v$ has already crashed. That paper proposed solutions to gossiping with $O(p \log p)$ messages with $p$ processors and suitably restricted adversaries, and with $O(p^{1.77})$ messages for unbounded adversaries, with the underlying communication overlay topologies the same as used in the current paper. Georgiou et al. \cite{31} showed how to obtain communication $O(p^{1+\varepsilon})$ for the unbounded adversary while maintaining suitably good time performance, for any fixed $\varepsilon > 0$. Based on this technique, Georgiou et al. \cite{31} gave an algorithm for Do-All with effort $O(t + p^{1+\varepsilon})$ against the unbounded adversary, again for any fixed $\varepsilon > 0$. Kowalski et al. \cite{11} showed how to obtain a constructive solution of gossiping with crashes with communication $O(p^{1+\varepsilon})$ for the unbounded adversary, for any fixed $\varepsilon > 0$. Chlebus and Kowalski \cite{15} developed a constructive gossiping solution working in $O(\log p)$ time while sending $O(p \log p)$ point-to-point messages, and demonstrated its efficiency when applied to consensus.

Problems related to Do-All were studied in distributed systems with shared memory of read-write registers. Bridgland and Watro \cite{6} considered tasks in asynchronous systems with processors prone to crashes under the constraint that each processor can perform at most one asks. The problem to collect values by reading shared registers is about processes who want to learn the
values stored in a collection of read-write registers by performing reads, each followed by a write to the own register. Such a read from register $x$ and write to register $y$ adds the contents of register $x$ to the contents of register $y$. Each register is assumed to be of a sufficient capacity to store all the original values deposited among the registers. Since each register storing a value needs to be read at least once, this resembles the property that a task needs to be performed at least once, but the difference is that the order of reads affects the proliferation of values. A solution to this problem to collect by reading was first used by Saks et al. \[48\] in their consensus algorithm. Chlebus et al. \[17\] showed that $n$ processors in an asynchronous system can collect $n$ values using $O(n)$ shared read-write registers by performing $O(n \text{ polylog } n)$ total work. Problem Write-All, introduced by Kanellakis and Shvartsman \[36\], is about writing at least once to each among a collection of dedicated shared read-write registers. Here writing a new value to a shared register can be interpreted as a task in the same sense as in Do-All. Chlebus and Kowalski \[18\] showed that $n$ processors using $O(n)$ shared read-write registers can can perform writes to each of $n$ registers with $O(n \text{ polylog } n)$ total work. Alistarh et al. \[2\] studied Do-All in asynchronous shared-memory model and showed that a deterministic non-constructive algorithm that performs $t$ tasks on $p$ processors with work $O(t + p \text{ polylog } (p + t))$. Kentros et al. \[39\] introduced At-Most-Once problem in which there is a set of tasks and a number of processors such that we want to perform as many tasks as possible subject to the constraint that a task is performed at most once. This problem was studied in \[39\] in asynchronous shared memory with processors prone to crashes and efficiency of solutions was measured by effectiveness which means the number of tasks completed. Kentros et al. \[39\] showed a lower bound on the number of performed tasks, which implied that it was impossible to perform all the tasks, and next gave an algorithm with effectiveness close to the lower bound. Kentros and Kiayias \[38\] gave algorithms for At-Most-Once improving effectiveness. Kentros et al. \[37\] studied the Strong-At-Most-Once problem, in which an algorithmic solution is required to have all the tasks performed when there are no crashes, and showed that it had the consensus number 2. Censor-Hillel \[8\] applied a randomized asynchronous wait-free multi-valued consensus solution as a building block in an algorithm for At-Most-Once of optimal effectiveness.

The solutions to Do-All given in \[23\] and \[29\] employed a group membership routine (called checkpointing) to repeatedly identify a set of processors that are still available, consistently over all available processors. In this paper, we also use a solution to a membership problem, which is integrated into the algorithm described in Section 7. Group membership abstractions were already proposed by Birman and Joseph \[5\], and intensely researched since then. Solutions to problems related to that of group membership are discussed in books by Cachin et al. \[7\] and Birman \[4\].

In recent years Internet supercomputing started becoming an increasingly popular means for harnessing the power of a vast number of interconnected computers. With this come the challenges of marshaling distributed resources and dealing with failures. Traditional centralized approaches employ a master processor and many worker processors that execute a collection of tasks on behalf of the master. Despite the simplicity and advantages of centralized schemes, the master processor is a performance bottleneck and a single point of failure. Additionally, a phenomenon of increasing concern is that workers may return incorrect results. Thus completely decentralized fault-tolerant solutions are of interest. Representative works in this area include the papers by Davtyan et al. \[22, 21\], also formulated for synchronous, crash-prone, message-passing settings, but with different failure models than in this work. We also note that the goal in those works is for all non-crashed processors to learn the results of all tasks, while in our work it is sufficient for the processors to know that the tasks have been performed.
Graphs called “expanders” are small-degree graphs that have strong connectivity-related properties. The “expansion” properties used to define expanders usually mean either that “small” sets of nodes have “many” neighbors, or that any two “large” disjoint sets of nodes are connected by at least one edge. The former approach to expanders is discussed by Chung [18]. The latter approach is considered by Pippenger [47], who defined an \( a \)-expander to be a graph such that, for any two disjoint sets \( X_1 \) and \( X_2 \) of nodes of size \( a \) each, there is an edge \((x_1, x_2)\) in the graph, where \( x_1 \in X_1 \) and \( x_2 \in X_2 \).

There is a substantial body of research on fault-tolerance in networks with expansion properties. Here we cite the work that is most relevant to this paper. Alon and Chung [3] demonstrated that, for every \( 0 < \varepsilon < 1 \) and every integer \( m > 0 \), there are constructive graphs with \( O(m/\varepsilon) \) vertices and maximum degree \( 1/\varepsilon^2 \) that have the property that after removing any \( 1 − \varepsilon \) fraction of the vertices or edges the remaining graph contains a path of length \( m \); they showed this by using the expanders introduced by Lubotzky et al. [44] with suitably chosen parameters. An \( h(f) \)-agreement is said to be achieved by a protocol if, in each execution with at most \( f \) processors failures, at least \( p − h(f) \) non-faulty processors eventually decide on a common value. A protocol achieves an almost-everywhere agreement if it achieves \( h(f) \)-agreement for some function \( h(f) \leq \mu f \), where the constant \( \mu \) is independent of the number of failures and of the size of the network. Upfal [50] showed how an almost-everywhere agreement can be achieved with a linear number of faults, extending the result by Dwork et al. [27]. This result uses the same family of expanders from [44], and it is shown that a linear-sized component of logarithmic diameter remains if some constant fraction of arbitrary nodes is removed. The graph construction in Section 3, in which we consider powers of Ramanujan graphs, is an extension of that approach.

Magnifying “expansion” by taking a power of a graph was already suggested by Margulis [45], and used, for instance by Diks and Pelc [24], in the context of broadcasting and fault diagnosis. Goldberg et al. [35] developed a deterministic distributed algorithm to reconfigure an \( n \)-input multi-butterfly network with faulty switches so that it still can route any permutation between some sets of \( n − O(f) \) inputs and outputs, for any number of \( f \) faulty switches. Chlebus et al. [17] showed that \( a \)-expander networks have the property that for any set of \( 3a \) non-faulty nodes there are \( 2a \) of these nodes that can communicate in \( O(\log a) \) time only among themselves.

Linearly-bounded adversaries have been used in other contexts. For example, Lamport et al. [43] showed that the consensus problem for Byzantine faults in synchronous settings can be solved against \( f \)-bounded adversaries for \( f < p/3 \). Diks and Pelc [24] studied deterministic broadcast and fault diagnosis in complete networks, where processor faults controlled by linearly-bounded adversaries. Kontogiannis et al. [40] studied a fault-prone BSP (bulk synchronous parallel) computer model, where processor faults are controlled by a variant of oblivious linearly-bounded adversaries.

**Document structure.** The rest of the paper is structured as follows. We define the models and performance measures in Section 2. We specify the communication schemes based on specific overlay graphs and give their relevant properties in Section 3. Our generic algorithm is given in Section 4 along with the basic relationships between work and communication. In the sequel we instantiate the generic algorithm and analyze the complexity of specific algorithm derivations. Section 5 presents a constructive algorithm based on load balancing. Section 6 describes an algorithm optimized for work. Section 7 gives an algorithm optimized for effort. We conclude with a discussion in Section 8.
2 Models and Technical Preliminaries

We consider a system of $p$ processors with unique processor identifiers in the interval $[1, p]$. Each processor knows the value of $p$, and $p$ may appear in the code of the algorithms.

Communication. Processors communicate by sending messages. The size of a message is assumed to be sufficiently large to contain the local state of a processor, as determined by the executed distributed algorithm. Communication is modeled at the “transport layer” level of abstraction, meaning that the network appears to the processors to be completely connected. Thus we abstract away the physical network, where they may or may not be direct links between processors, with routing provided by the “network layer.”

Synchrony. The system is synchronous, where the local steps of all processors are governed by a global clock. An execution is structured as a sequence of rounds. A round is defined to be a fixed number of clock cycles that is sufficiently large but minimal for the following to occur: (1) a processor can send messages to several other processors in one round, (2) messages sent during a round are delivered and received later in that round, (3) the duration of a round is sufficient to perform the needed local processing.

Failures and adversaries. Processors fail by crashing. We denote by $f$ an upper bound on the number of failures that may occur in an execution. We do not consider processor restarts and so $f$ never exceeds $p$. Multicast messaging is not assumed to be atomic with respect to failures: if a processor crashes while attempting to multicast a message, then some arbitrary subset of the destinations may receive the message. We assume that no messages are lost or corrupted in transit.

The occurrence of failures is governed by adversarial models that determine upper bounds on the number of processor failures and the specific subsets of processors that may crash at some instant. An adversary is called $f$-bounded when at most $f$ crashes may occur in an execution. An adversary is linearly bounded when it is $f$-bounded and $f$ satisfies $f \leq cp$, for a positive constant $c < 1$, which means that $p - f = \Omega(p)$. The $(p - 1)$-bounded adversary is also called unbounded. The unbounded adversary is subject only to the weakest restriction that at least one processor does not crash in any execution. Adversaries are always adaptive in this paper, in the sense that the decisions about the failures of specific processors are made on-line. The adversaries are subject only to the upper bound on the number of crashes allowed to occur in an execution, unless stated otherwise (as in Section 6.2).

The Do-All problem and tasks. The Do-All problem is to perform a given set of $t$ tasks in the presence of an adversary. The tasks have unique identifiers in the interval $[1, t]$. Each processor knows the value of $t$, and $t$ may appear in the algorithm code. Any processor can perform any task when given the task’s identifier. Tasks are similar in terms of time needed to perform one; we assume that any task can be performed within a single round and that one round is needed to perform one task by one processor. Tasks are idempotent, obeying at-most-once execution semantics: any task can be performed multiple times and concurrently. Tasks are also independent: they can be performed in any order.

Correctness and termination. A processor may either voluntarily stop working in a given execution, or it may be forced to do so by an adversary that causes it to fail. In the former case, we say that the processor halts, and in the latter case the processor crashes. Processors may halt
in different rounds. Halted processors are considered non-faulty. We say that an algorithm solves the Do-All problem against a given adversary, if the following two conditions are satisfied in any execution consistent with the adversary:

1. Each task is eventually performed by some processor (or processors).
2. Each processor eventually halts, unless it crashes.

Our algorithms solve Do-All in the presence of the unbounded adversary. Observe that this form of correctness implies that when a processor $v$ halts, then all the tasks must have been performed. This is because if a non-faulty processor halts when not all tasks are complete, then the unbounded adversary can immediately crash all other processors, and thus there will remain tasks that have not been completed.

We say that an algorithm terminates in a round if this is the first round where each processor either halted or crashed. It is not a requirement that algorithms terminate by having all correct processors halt simultaneously.

Performance metrics. We consider work and communication as metrics of performance of algorithms. Work $W$ is the maximum over all executions of the sum, for all processors, of the number of rounds a processor performs in an execution until the processor crashes or the algorithm terminates, whichever occurs first. This means that we count the available processor steps [36]. In particular, halted processors continue contributing to work until the algorithm terminates.

Communication $M$ is the maximum over all executions of the total number of point-to-point messages sent in an execution; multicasts are accounted for in terms of the point-to-point messages according to the number of destinations for each multicast. We define effort as $E = W + M$.

3 Graphs and Communication

To achieve our efficiency goals we control the communication costs of our algorithms by allowing processors to send messages only to a selected subset of processors during a round. In this section, we specify graphs used by the algorithms to limit communication. Each such graph forms a conceptual overlay network over the assumed complete network.

We consider simple graphs, that is, undirected graphs with at most one edge between any two nodes. A graph $G = (V, E)$ is defined in terms of its set of nodes $V$ and the set of edges $E$. We denote by $|G|$ the number of nodes in the graph. A graph $H = (V', E')$ is a subgraph of $G = (V, E)$, denoted by $H \subseteq G$, when $V' \subseteq V$ and $E' \subseteq E$. A subgraph $H = (V', E')$ of $G = (V, E)$ is an induced subgraph of $G$ when, for any pair of nodes $x$ and $y$ in $V'$, if $(x, y)$ is an edge in $E$ then it is also an edge in $E'$. An induced subgraph $H = (V', E')$ is said to be induced by $V'$, since the set of nodes $V'$ determines the edges in $E'$.

If $(x, y)$ is an edge in a graph, we say that node $x$ is a neighbor of $y$ (and $y$ is a neighbor of $x$). For a subset $V_0 \subseteq V$, we define the set of neighbors of $V_0$, denoted $N_G(V_0)$, to be the set of all nodes of $G$ that have neighbors in $V_0$. Extending the neighborhood notation, we denote by $N_G^k(V_0)$ for a positive integer $k$, to be the set of all nodes in $G$ connected to at least one node in $V_0$ by a path of length at most $k$. The graph $G^k = (V, E')$ is called the $k$th power of graph $G$ if for any pair of nodes $u$ and $v$ in $V$, $(u, v)$ is an edge in $E'$ if and only if there is a path between $u$ and $v$ in $G$ of length at most $k$. Observe that for any $V_0 \subseteq V$ we have $N_G^k(V_0) = N_G^k(V_0)$. 


We model each processor as a node in a graph, and any pair of processors that are to communicate regularly is modeled as an edge of an overlay graph. For $p$ processors and a bound $f < p$ on the number of crashes, the suitable overlay graph is denoted by $G(p, f)$. Each crashed or halted processor is removed from this graph, causing the remaining portion of the overlay graph to evolve through a sequence of subgraphs, until the graph vanishes altogether when the algorithm terminates.

Definitions 1 and 2 below are parameterized by three constants $\alpha$, $\beta$ and $\gamma$, where $\alpha$ is real such that $0 < \alpha < 1$, and $\beta$ and $\gamma$ are positive integers, and also by the number of failures $f$ that are tolerated in an execution; we will choose the specific constants $\alpha$, $\beta$ and $\gamma$ later. We use the following notation for logarithms: $\lg x$ means $\log_2 x$, while $\log x$ is used when the constant base of the logarithm does not matter, as in the asymptotic notation $O(\log n)$.

**Definition 1** Let $f$ be an upper bound on the number of crashes. We say that a subgraph $H \subseteq G(p, f)$ is compact if $|H| \geq \alpha(p - f)$ and the diameter of $H$ is at most $\beta \lg p + \gamma$. □

For a graph $G$, a subgraph function $P$ on $G$ assigns an induced subgraph $P(G_0) \subseteq G$ to any induced subgraph $G_0 \subseteq G$.

**Definition 2** Graph $G(p, f)$ has the subgraph property for $f$ if there exists a subgraph function $P$ on $G(p, f)$ that satisfies the following conditions for induced subgraphs $G_0$ and $G_1$, each of at least $p - f$ nodes:

1. $P(G_0) \subseteq G_0$;
2. $|P(G_0)| \geq \alpha|G_0|$;
3. The diameter of $P(G_0)$ is at most $\beta \lg p + \gamma$;
4. If $G_0 \subseteq G_1$ then $P(G_0) \subseteq P(G_1)$. □

**Graphs with good expansion properties.** We denote by $L(n, d)$ the constructive Ramanujan graph on $n$ nodes and of node degree $d$ as given by Lubotzky et al. [44], for positive integers $n$ and $d$. A self-contained exposition of this construction and a discussion of the expansion-type properties of the graphs are given by Davidoff et al. [20]; Theorem 4.2.2 therein is relevant in particular. Any number $d$ such that $d - 1$ is an odd prime can be chosen as a node degree of $L(n, d)$. The number of nodes $n$ is either of the form $r(r^2 - 1)/2$ or of the form $r(r^2 - 1)$, where $r > 2\sqrt{d - 1}$ is an odd prime, depending on whether or not the prime $d - 1$ is a quadratic residue modulo $r$, respectively. We will use a fixed node degree denoted by $\Delta_0$.

Given a number $p$ of processors sufficiently large with respect to $\Delta_0$, we set $x$ to the smallest positive integer with the property that $x(x^2 - 1)/2 \geq p$. Let $r_p$ be a prime number between $x$ and $2x$ that must exist by Chebyshev’s Theorem on the distribution of prime numbers. Consider the corresponding Ramanujan graph of node degree $\Delta_0$ and of either $n_p = r_p(r_p^2 - 1)/2$ or $n_p = r_p(r_p^2 - 1)$ nodes. This graph $L(n_p, \Delta_0)$ has the properties that $n_p \geq p$ and $n_p = \Theta(p)$. If $n_p = p$ then nothing else needs to be done to model the $n_p$ processors. Otherwise, we let each processor simulate $O(1)$ nodes, so that $n_p = \Theta(p)$ nodes of $L(n_p, \Delta_0)$ are simulated. The simulating $p$ processors are prone
to crashes, and when such a simulating processor crashes, then this results in $\Theta(1)$ crashes of the simulated nodes. For the simulation to provide the given results, it is sufficient for the presented constructions to have the property that replacing $p$ by $cp$ and $f$ by $cf$, where $c > 1$ denotes the number of processors simulated by one simulating process, preserves the validity of the used arguments and their conclusions. It is actually only Lemmas 1 and 2 (below) that directly refer to the properties of the graphs $L(p, \Delta_0)$, and they pass the test, by inspection. This allows us to simplify the notation, so that henceforth we assume that $n_p = p$ and refer to the graphs $L(p, \Delta_0)$ by $L(p)$.

The expansion of a graph can be estimated by the eigenvalues of its adjacency matrix. The matrix is symmetric, hence its eigenvalues are real. Let the eigenvalues be denoted by

$$\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{p-1},$$

for a graph of $p$ vertices. The largest eigenvalue $\lambda_0$ of a regular graph is equal to its degree, which in the case of the graphs $L(p)$ is $\Delta_0$. The graphs $L(p)$ have an additional property that their second largest in absolute value eigenvalue $\lambda = \max\{|\lambda_1|, |\lambda_{p-1}|\}$ satisfies the inequality

$$\lambda \leq 2\sqrt{\Delta_0 - 1}, \quad (1)$$

as shown in [20, 44]; this is the property that defines Ramanujan Graphs among regular connected graphs of the node degree $\Delta_0$.

We will refer to the following two facts known from the literature.

**Fact 1 ([49])** For a regular graph $G$ with $p$ nodes and a subset of nodes $V_0$, the size of $N_G(V_0)$ can be estimated by the two largest eigenvalues of $G$ as follows:

$$|N_G(V_0)| \geq \frac{\lambda_0^2 |V_0|}{\lambda_0^2 + (\lambda_0^2 - \lambda^2)|V_0|/p}. $$

**Fact 2 ([50])** If $\Delta_0 = 74$ is used in the specification of graph $L(p)$ then there exists a function $P'$ such that if $L_0$ is a subgraph of $L(p)$ of size at least $71p/72$, then the subgraph $P'(L_0)$ of $L_0$ is of size at least $|L_0|/6$ and its diameter is at most $30\log p$.

In terms of Definition 2, Fact 2 means that if $\Delta_0 = 74$ is used in the construction of graph $L(p)$, and if $f \leq p/72$, then $L(p)$ has the subgraph property for $f$ with $\alpha = 1/6$, $\beta = 30$ and $\gamma = 0$. Next we will show how to extend this fact so that the restriction $f \leq p/72$ can be abandoned, for appropriately adjusted constants $\alpha$, $\beta$ and $\gamma$. To this end, we use the overlay graphs $L(p)^j$, for suitably large $j$, that have stronger connectivity properties than those of $L(p)$. Observe that, for a given exponent $j$ in this construction, the maximum node degree $\Delta$ of $L(p)^j$ is at most $\Delta_0^j$. We will press into service the function $P'$ from Fact 2 several times in the sequel.

**The specification of the overlay graphs.** The definitions of graph $L(p)$ and the notions of compactness and subgraph property are parametrized by the constants $\Delta_0$, $\alpha$, $\beta$, and $\gamma$. From now on, we fix these constants to the values $\Delta_0 = 74$, $\alpha = 1/7$, $\beta = 30$, and $\gamma = 2$. Let $N(R)$ denote $N_{L(p)}(R)$, for any set $R$ of nodes of $L(p)$.

The overlay graphs $G(p, f)$ are defined as $L(p)^\ell$ for a suitable exponent $\ell$. If $f \leq p/72$, then we may take $\ell = 1$ and rely on Fact 2. Let us suppose that $f > p/72$. Let $R$ be a set of fewer
than $71p/72$ nodes. We estimate the size of the neighborhood $N(R)$ by the inequality of Fact 1 and property (1) by which $L(p)$ is a Ramanujan graph, to obtain the following bounds:

$$|N(R)| \geq \frac{\Delta^2_0}{\lambda^2 + (\Delta^2_0 - \lambda^2)p|R|} |R|^\ell \\
\geq \frac{5476}{292 + 5184 \cdot \frac{|R|}{p}} |R|^\ell .$$  

(2)

If the size $|R|$ is at most $p/50$, then $|N(R)| \geq \rho |R|$ holds for the constant $\rho = 13.5$ that is obtained by substituting $p/50$ for $|R|$ in the estimate (2). If $|R|$ is between $p/50$ and $71p/72$, then $|N(R)| \geq \rho_1 |R|$ for the constant $\rho_1 = 1.013$ that is obtained by substituting $71p/72$ for $|R|$ in the estimate (2). A benefit brought in by replacing the estimate (2) by the products $\rho |R|$ and $\rho_1 |R|$ is to be able to express the estimates on the sizes of neighborhoods of the form $N^i(R)$ in a power $L(p)^i$ of graph $L(p)$ by the powers of $\rho$ and $\rho_1$, with the corresponding exponents suitably related to $i$.

We want to take a power $k$ of $L(p)$ that is sufficiently large to expand any neighborhood of set $R$ of at least $p - f$ nodes to $71p/72$ nodes, then use function $P'$ from Fact 2 and finally rely on the exponent $\ell$ to be sufficiently large to translate this action into the suitable properties of graph $L(p)^\ell$. Let us define the three positive integers $r$, $k$, and $\ell$ as follows:

a) Let $r$ be the smallest positive integer such that $(p - f)\rho^r > p/50$.

b) Given this $r$, let $k > r$ be the smallest integer that is large enough so that the inequality $(p - f)\rho^r(\rho_1)^{k-r} > 71p/72$ holds.

c) Set $\ell = 2k + 1$.

The existence of $r$ is by the choice of $\rho = 27/2$, while $k$ is well defined by the choice of $\rho_1$, as specified just after the estimate (2), which means that graph $L(p)^\ell$ is well defined. We make this graph $L(p)^\ell$ to be $G(p, f)$.

Given a set of nodes $R$ in $G(p, f)$, we may estimate the size of the set $N^k(R)$ in $G(p, f)$ as follows. If $R$ is a set of nodes such that $|R| \geq p - f$ then the following inequality holds:

$$|N^k(R)| > \frac{71}{72} p .$$  

(3)

This is because parts a) and b) of the specification of $L(p)^\ell$ provide

$$|R|\rho^r(\rho_1)^{k-r} \geq (p - f)\rho^r(\rho_1)^{k-r} .$$

If $R$ is a set of nodes such that $|R| < p - f$, then the following inequality holds:

$$|N^k(R)| > \frac{|R|}{p - f} \cdot \frac{71}{72} p .$$  

(4)

This is because in such a case we have that

$$|R|\rho^r(\rho_1)^{k-r} = \frac{|R|}{p - f}(p - f)\rho^r(\rho_1)^{k-r} > \frac{|R|}{p - f} \cdot \frac{71p}{72} ,$$

by the inequality used in part b) of the construction of the overlay graph.
Two properties of overlay graphs. Now we discuss two properties of the overlay graphs \( G(p, f) \), as defined above in the form \( L(p) \ell \) for a suitable \( \ell \), that will be key for applications in algorithm design in subsequent sections.

**Lemma 1** For any \( f < p \), the maximum degree of \( G(p, f) \) is \( O\left(\left(\frac{p}{p-f}\right)^{3.31}\right) \).

**Proof:** We estimate \( r \) and \( k - r \) as used in the specification of \( G(p, f) = L(p) \ell \), where \( \ell = 2k + 1 \). It follows from part a) of the construction that

\[
\log_\rho \frac{p}{50(p-f)} < r \leq \log_\rho \frac{p}{50(p-f)} + 1 .
\]

(5)

Similarly, since the inequality

\[
(p-f)\rho^r \cdot \rho_1^{k-r-1} \leq \frac{71p}{72}
\]

holds by part b) in the construction, and the inequality

\[
\frac{1}{(p-f)\rho^r} < \frac{50}{p}
\]

holds by part a), we obtain that

\[
\rho_1^{k-r-1} \leq \frac{71p}{72} \cdot \frac{1}{(p-f)\rho^r} < \frac{71 \cdot 50}{72} .
\]

(6)

Combining the estimates (5) and (6) yields the following bound on \( k \):

\[
k = r + (k-r) \\
\leq \log_\rho \frac{p}{50(p-f)} + 1 + \log_\rho_1 \frac{71 \cdot 50}{72} + 1 \\
= \log_{\rho_0} \frac{p}{p-f} + \left(2 - \log_\rho_0 50 + \log_\rho_1 \frac{71 \cdot 50}{72}\right) \\
= \log_\rho \frac{p}{p-f} + c ,
\]

where \( c \) is a constant \( (c = 2 - \log_\rho 50 + \log_\rho_1 \frac{71 \cdot 50}{72}) \). This leads to the following bound on \( \Delta \), which is the maximum degree of \( L^\ell(p) \):

\[
\Delta \leq \Delta_0^{2k+1} \\
\leq \Delta_0^{2\log_\rho \frac{p}{p-f} + 2c + 1} \\
= O\left(\Delta_0^{2\log_\rho \frac{p}{p-f}}\right) \\
= O\left(\left(\frac{p}{p-f}\right)^{2\log_\rho \Delta_0}\right).
\]

For \( \rho = 27/2 \) and \( \Delta_0 = 74 \), the exponent \( 2\log_\rho \Delta_0 \) in the bound on the maximum degree is less than 3.31, by direct inspection.

Let \( f < p \) and consider a subgraph \( R \) of the graph \( G(p, f) = L(p) \ell \) of at least \( p - f \) nodes.
Definition 3 The subgraph function \( P(R) \) denotes the subgraph of \( L(p)^\ell \) induced by the vertices in \( N^k(P(N^k(R))) \cap R \), where \( \ell = 2k + 1 \) and \( P' \) is the subgraph function from Fact 2.

Lemma 2 For any \( f < p \), graph \( G(p,f) \) has the subgraph property with \( P(R) \) as the subgraph function.

Proof: We verify the four parts of the definition of the subgraph property.

Part 1) states that \( P(R) \subseteq R \): this follows directly from the definition of \( P(R) \).

Part 2) states that \( |P(R)| \geq |R|/7 \). Suppose, to the contrary, that \( |R \setminus P(R)| > 6|R|/7 \). We proceed by making three preliminary observations. The first one is that the following inequality holds:

\[
|N^k(R \setminus P(R))| > \frac{6}{7} \cdot \frac{71}{72} p .
\]

(7)

To show (7), we consider two cases. If \( |R \setminus P(R)| \geq p - f \) then estimate (3) suffices. Otherwise, when \( |R \setminus P(R)| < p - f \), then we resort to (4) and additionally verify that

\[
\frac{|R \setminus P(R)|}{p - f} \geq \frac{|R \setminus P(R)|}{|R|} > \frac{6}{7} .
\]

The second observation is that the following inequalities hold

\[
|P'(N^k(R))| \geq \frac{|N^k(R)|}{6} \geq \frac{1}{6} \cdot \frac{71}{72} p ,
\]

(8)

which follows from bound (3) and Fact 2. Finally, observe that the inequality

\[
\frac{6}{7} \cdot \frac{71}{72} + \frac{1}{6} \cdot \frac{71}{72} > 1
\]

holds, which can be verified directly.

The inequalities (7), (8) and (9) together imply that the intersection of \( N^k(R \setminus P(R)) \) with \( P'(N^k(R)) \) is nonempty, because the sum of the sizes of sets is larger than \( p \), which is the total number of nodes. Let \( v \) be a node that belongs to both \( N^k(R \setminus P(R)) \) and \( P'(N^k(R)) \). There is a path in \( L(p) \) of length at most \( k \) from \( v \) to some \( w \) in \( R \setminus P(R) \). According to the specification \( P(R) = N^k(P(N^k(R))) \cap R \), this means that \( w \) is in \( P(R) \) because the node \( v \) is in \( P'(N^k(R)) \). This results in a contradiction, since the sets \( R \setminus P(R) \) and \( P(R) \) are disjoint. The contradiction yields that \( |P(R)| \geq |R|/7 \).

Part 3) states that the diameter of \( P(R) \) is at most \( 30 \log p + 2 \). Let \( v_1 \) and \( v_2 \) be two vertices in the subgraph \( P(R) \) of \( L(p)^\ell \).

Suppose first that both \( v_1 \) and \( v_2 \) are in \( P'(N^k(R)) \). By Fact 2 there is a path from node \( v_1 \) to node \( v_2 \) of length at most \( 30 \log p \) that traverses nodes in \( P'(N^k(R)) \) via edges in \( L(p) \). Let \( \langle w_1, w_2, \ldots, w_m = v_2 \rangle \) be the sequence of consecutive nodes on this path. We show that there is a path \( \langle w'_1, w'_2, \ldots, w'_m = v_2 \rangle \) between \( v_1 \) and \( v_2 \) in the subgraph \( P(R) \) of \( L(p)^\ell \) that has the following property: for \( 1 \leq i < m \), either \( w'_{i+1} = w_{i+1} \) or \( w'_{i+1} \) is a neighbor of \( w_{i+1} \) in \( L(p)^k \).

The construction is by induction on the length of the path. The base of induction is for the path of length zero, which holds by the assumption \( w'_1 = v_1 \). Suppose that we have found an initial segment of this path up to a node \( w'_i \), where \( 1 \leq i < m - 1 \). The inductive step is accomplished for each of the following four possible cases:
a) The case when \( w_i = w_i' \) and \( w_{i+1} \) is in \( P(R) \):

Set \( w_{i+1}' = w_{i+1} \). The nodes \( w_i' \) and \( w_{i+1}' \) are neighbors in \( L(p) \) and so also in \( L(p)^\ell \).

b) The case when \( w_i = w_i' \) and \( w_{i+1} \) is not in \( P(R) \):

Hence \( w_{i+1} \in P'(N^k(R)) \setminus P(R) \). Let \( C \) denote \( P'(N^k(R)) \setminus R \). We have the following equalities:

\[
P'(N^k(R)) \setminus P(R) = P'(N^k(R)) \setminus (N^k(P'(N^k(R))) \cap R) = P'(N^k(R)) \setminus R = C,
\]

because

\[
P'(N^k(R)) \cap R \subseteq P'(N^k(R))
\]

and so the inclusion

\[
P'(N^k(R)) \cap R \subseteq N^k(P'(N^k(R))
\]

holds. This means that \( w_{i+1} \in C \).

Let \( w_{i+1}' \) be any neighbor of \( w_{i+1} \) that is in the subgraph \( (N^k(C) \setminus C) \cap R \) of \( L(p)^k \), and hence in \( P(R) \); such a node exists since otherwise \( N^k(w_{i+1}) \cap R \) would be empty, which contradicts the fact that

\[
w_{i+1} \in C \subseteq P'(N^k(R)) \subseteq N^k(R).
\]

The nodes \( w_i' \) and \( w_{i+1}' \) are neighbors in \( L(p)^\ell \), as their distance in \( L(p) \) is at most \( 1 + k < \ell \); namely, the distance in \( L(p) \) between \( w_i' \) and \( w_{i+1} \) is 1 and the distance in \( L(p) \) between \( w_{i+1} \) and \( w_{i+1}' \) is at most \( k \).

c) The case when \( w_i \neq w_i' \) and \( w_{i+1} \) is in \( P(R) \):

Take \( w_{i+1} \) to be \( w_{i+1}' \). The distance between \( w_i' \) and \( w_{i+1}' = w_{i+1} \) in \( L(p) \) is \( k + 1 < \ell \), since the node \( w_i \) is a neighbor of \( w_{i+1} \) in \( L(p) \) and, by the inductive assumption, the node \( w_i \) is a neighbor of \( w_i' \) in \( L(p)^k \). Hence there is an edge between \( w_i' \) and \( w_{i+1}' \) in \( L(p)^\ell \).

d) The case when \( w_i \neq w_i' \) and \( w_{i+1} \) is not in \( P(R) \):

Let \( C \) denote \( P'(N^k(R)) \setminus R \). By the inductive assumption, the node \( w_i \) is a neighbor of \( w_i' \) in \( L(p)^k \) and \( w_{i+1} \in C \). Let \( w_{i+1}' \) be any neighbor of \( w_{i+1} \) in the subgraph \( (N^k(C) \setminus C) \cap R \subseteq P(R) \) of \( L(p)^k \); such a node exists since otherwise \( N^k(w_{i+1}) \cap R \) would be empty, which contradicts the fact that \( w_{i+1} \in P'(N^k(R)) \subseteq N^k(R) \). The distance between \( w_i' \) and \( w_{i+1}' \) in \( L(p) \) is at most \( \ell = k + 1 + k \), since the distance between \( w_i' \) and \( w_i \) is at most \( k \), the distance between \( w_i \) and \( w_{i+1} \) is 1, and the distance between \( w_{i+1} \) and \( w_{i+1}' \) is at most \( k \).

Therefore there is an edge between \( w_i' \) and \( w_i \) in \( L(p)^\ell \).

Note that if a path \( \langle v_1 = w_1', w_2', \ldots, w_{m-1}' \rangle \) has been built then the node \( w_{m-1}' \) is in \( N^k(\{w_{m-1}\}) \), by the inductive assumption. Hence the distance in \( L(p) \) between the nodes \( w_{m-1}' \) and \( v_2 \) is at most \( k + 1 \). This completes the inductive construction, and hence the case when both \( v_1 \) and \( v_2 \) are in \( P'(N^k(R)) \).

The case when either \( v_1 \) or \( v_2 \) is not in \( P'(N^k(R)) \) follows from the observation that both these nodes have neighbors in \( P'(N^k(R)) \) in \( L(p)^k \), by the inclusion \( P(R) \subseteq N^k(P'(N^k(R))) \). Therefore the distance between them in the subgraph \( P(R) \) of \( L(p)^\ell \) is at most \( 2 + 30 \log p \).

Part 4) states that \( P(R_1) \subseteq P(R_2) \) for any induced subgraphs \( R_1 \subseteq R_2 \) of \( L^\ell(p) \), each of size at least \( p - f \).
Let \( V_1 \) be the set of nodes of \( R_1 \) and \( V_2 \) the set of nodes of \( R_2 \). Clearly \( N^k(V_1) \subseteq N^k(V_2) \). Moreover, \( |N^k(V_2)| \geq |N^k(V_1)| \geq \frac{71p}{72} \) by (3). It follows that \( P'(N^k(V_1)) \subseteq P'(N^k(V_2)) \), by Fact 2 which implies
\[
N^k(P'(N^k(V_1))) \subseteq N^k(P'(N^k(V_2)))
\]
and next
\[
N^k(P'(N^k(V_1))) \cap V_1 \subseteq N^k(P'(N^k(V_2))) \cap V_2.
\]
Since \( P(R_1) \) is the subgraph of \( L^f(p) \) induced by the nodes in \( N^k(P'(N^k(V_1))) \cap V_1 \), and \( P(R_2) \) is the subgraph of \( L^f(p) \) induced by the nodes in \( N^k(P'(N^k(V_2))) \cap V_2 \), the inclusion \( P(R_1) \subseteq P(R_2) \) follows. □

As a part of the statement of Lemma 2, the diameter of a compact graph is at most \( 30 \log p + 2 \); we will use the notation \( g(p) \) as a shorthand for \( 30 \log p + 2 \).

4 Algorithm Design

In this section we present the design template for our algorithms formulated as the generic algorithm. This algorithm design leaves the following open: 1) the assignment of tasks for processors to perform, and 2) the overlay graphs used for sending messages. The generic algorithm solves the Do-All problem against the unbounded adversary, as we show in Proposition 1. In the subsequent sections we deal with specific algorithms as instantiations of the generic algorithm. The purpose of the instantiations is to achieve suitable efficiency.

4.1 The generic algorithm

The generic algorithm is given in Figure 1, and it contains the initialization followed by the repeat loop. The activity of each processor consists of two stages. In the first stage, the processor works to perform all tasks and to share its knowledge with other processors. This is implemented by procedure Main given in Figure 2. After the processor learns that there are no outstanding tasks, it switches to the second stage, where it propagates its knowledge to all processors. This is implemented by procedure Closing given in Figure 3. The processor halts when it learns that every processor knows that there are no outstanding tasks.

A processor is called active if it neither halted nor crashed. An active processor is called busy when it executes procedure Main, and so is still “busy performing tasks.” A processor executing procedure Closing is no longer busy performing tasks but it remains active. An active processor that is not busy may halt for two reasons in the course of executing procedure Closing. (1) The processor halts after it learns that there are no busy processors, meaning that all active processors know that there are no outstanding tasks. (2) The processor may halt after it learns that there are “too few” active processors located sufficiently close in the overlay communication network to cooperate with productively; this is a technical condition designed to avoid scenarios when isolated islands of processors remain active, thereby unnecessarily contributing to the work and message complexities.

Local states and information propagation. Each processor \( v \) maintains its local state consisting of three ordered lists and additional variables. The list \( \text{Tasks}_v \) contains the identifiers of tasks
Generic Algorithm

/* Initialization */
Tasks,v ← the sorted list of all tasks
Processors,v, Busy,v ← the sorted list of all processors
Phase,v ← Main /* Phase,v is set to point to procedure Main */
Done,v ← false

/* Algorithm code */
repeat call procedure Phase,v

Figure 1: The generic algorithm code for processor v; the algorithm starts with the initialization, followed by iterating either procedure Main or procedure Closing based on the value of Phase,v.

assumed to be outstanding. The list Processors,v contains the identifiers of processors assumed to be active. The list Busy,v contains the identifiers of processors assumed to be busy performing tasks. These lists are sorted in the order of the identifiers contained in them. The position of an item in a list is called this item’s rank in the list. Items may be removed from lists; this affects the ranks of the remaining items.

During the initialization, processor v sets the list Tasks,v to the sorted list of all task identifiers. Both lists Busy,v and Processors,v are initialized to the sorted list of all processor identifiers.

Processors share their knowledge by sending their Tasks, Processors, and Busy lists to their neighbors in the current overlay graph. Each active processor v updates its lists Tasks,v, Processors,v, and Busy,v after it receives messages containing such lists. If an item is not on the list received from some neighbor, then the processor removes the item from the corresponding private list. The processor removes from its Tasks,v list the tasks that it knows to be complete based on the received information. Similarly, the processor removes from its Processors,v list any processors that crashed or halted, and finally, it removes from its Busy,v list the processors that known that there are no outstanding tasks. If no message is received by v from a neighboring processor u, then u is removed from the list Processors,v, as this indicates that u is no longer active. In this way the changes in the lists are propagated through flooding broadcasts over the overlay graph.

The knowledge contained in the private lists may be out of date in a given round. In the case of the lists Processors,v and Busy,v, the sets of processor identifiers on these lists may be the supersets of the actual sets of active and busy processors, respectively. This happens if some processors are no longer active or no longer busy, but processor v has not learned this yet. Similarly, the set of task identifiers on the list Tasks,v may be a superset of the actual set of outstanding tasks, since some tasks may have been performed, but processor v does not know this yet.

Each processor maintains the following additional variables. The variable Phase,v points to the procedure to be performed by v in the next iteration of the main loop; this procedure is either Main or Closing. Initially Phase,v is set to point to Main. The variable Selected,Processor,v is used to store the identifier of the processor to whom the Stop signal is sent by v in the current execution of
Procedure Main

ROUND 1: receive messages

ROUND 2: perform local computation:

a. update the private lists using the messages just received:
   1. remove processor identifier \(x\) from \(\text{Processors}_v\) if either \(x\) is missing from some \(\text{Processors}\) list received, or \(x\) is a neighbor of \(v\) and no message from \(x\) was received
   2. remove processor identifier \(x\) from \(\text{Busy}_v\) if either \(x\) is missing from some \(\text{Busy}\) list received, or \(x\) is not in \(\text{Processors}_v\)
   3. remove task identifier \(y\) from \(\text{Tasks}_v\) if \(y\) is missing from some \(\text{Tasks}\) list received

b. if \(\text{Tasks}_v\) is nonempty then
   1. select a task from \(\text{Tasks}_v\) by applying \(\text{Selection\_Rule}\)
   2. perform the selected task
   3. remove the selected task from \(\text{Tasks}_v\)

   else \(\text{Done}_v \leftarrow \text{true}\)

c. if \(\text{Busy}_v \neq \text{Processors}_v\) then \(\text{Done}_v \leftarrow \text{true}\)

d. if Stop signal received then \(\text{Done}_v \leftarrow \text{true}\)

e. if \(\text{Done}_v = \text{true}\) then
   1. remove \(v\) from \(\text{Busy}_v\)
   2. set \(\text{Tasks}_v\) to an empty list
   3. \(\text{Phase}_v \leftarrow \text{Closing} /\!* \text{Phase}_v\) is set to point to procedure \(\text{Closing}\) */

ROUND 3: multicast a message containing \(\text{Tasks}_v\), \(\text{Processors}_v\), and \(\text{Busy}_v\) to those neighbors in the overlay graph that are in \(\text{Processors}_v\)

Figure 2: Procedure \(\text{Main}\) of the generic algorithm; the code for processor \(v\). The code is parameterized by the procedure \(\text{Selection\_Rule}\) that is used to select an item from a list.

the procedure \(\text{Closing}\). The boolean variable \(\text{Done}_v\) facilitates switching processor \(v\) from executing \(\text{Main}\) to executing \(\text{Closing}\) in the next iteration. The variable \(\text{Done}_v\) is initially set to false, and it remains false until \(v\) learns that all tasks are complete.

\(\text{Done}_v\) is set to true in the following three situations:

(1) When the list \(\text{Tasks}_v\) becomes empty: this means that all tasks are complete.

(2) When \(v\) receives the Stop signal from some processor \(u\): this means that \(u\) knows that all tasks are complete and so it switches to executing \(\text{Closing}\).

(3) When \(v\) determines that the lists \(\text{Busy}_v\) and \(\text{Processors}_v\) are different: this happens when there is an active processor in \(\text{Processors}_v\), that is missing from \(\text{Busy}_v\), meaning that this
Procedure Closing

ROUND 1: receive messages

ROUND 2: perform local computation:

a. update the private lists using the messages just received:
   1. remove processor identifier \( x \) from \( \text{Processors}_v \) if either \( x \) is missing from some received list \( \text{Processors} \), or \( x \) is a neighbor of \( v \) and no message from \( x \) was received
   2. remove processor identifier \( x \) from \( \text{Busy}_v \) if either \( x \) is missing from some received list \( \text{Busy} \), or \( x \) is not in \( \text{Processors}_v \)

b. if \( v \) does not consider itself compact then halt after this phase is over

c. if \( \text{Busy}_v \) is nonempty then
   1. select a processor from \( \text{Busy}_v \) by applying \text{Selection_Rule}
   2. set \( \text{Selected}_v \) to the selected processor
   3. remove \( \text{Selected}_v \) from \( \text{Busy}_v \)
   else
   4. set \( \text{Selected}_v \) to \( v \)
   5. halt after this phase is over

ROUND 3: send messages:

a. send Stop signal to \( \text{Selected}_v \)

b. multicast a message containing \( \text{Tasks}_v \), \( \text{Processors}_v \), and \( \text{Busy}_v \) to those neighbors in the overlay graph that are in \( \text{Processors}_v \)

Figure 3: Procedure Closing of the generic algorithm; the code for processor \( v \). The property of a processor to consider itself compact is defined in terms of the subgraph of the overlay graph induced by the nodes in \( \text{Processors}_v \). The processor \( v \) halts after completing the full three rounds of the last phase it performs, and so in particular after sending all the messages in Round 3 of the phase. The code is parameterized by the procedure \text{Selection_Rule} that is used to select an item from a list.

Once \( \text{Done}_v \) is set to true, processor \( v \) removes its identifier \( v \) from \( \text{Busy}_v \) so that the knowledge that \( v \) is no longer busy is propagated via messages carrying copies of lists.

We refer to each iteration of the repeat loop, where either Main or Closing is invoked, as a phase. Correspondingly, we call a phase either main or closing, depending on which procedure is invoked. Each phase consists of three rounds: 1) receiving messages, 2) performing local computation, and 3) multicasting messages. The goal of main phases is to complete all tasks. Pseudocode for procedure Main, and so for each main phase, is given in Figure 2. A processor whose list Tasks becomes empty switches to executing procedure Closing. Now the goal is to inform all active processors that all tasks are complete. This is accomplished in a manner similar to that of
performing tasks, in that informing a processor is now treated as a “task.” To perform a “task” of this kind, first an item from the Busy list is selected, next the Stop signal is sent to it, and finally the item is removed from the list. A processor halts when its Busy list becomes empty. Pseudocode for procedure Closing, and so for each closing phase, is given in Figure 3.

When a processor needs to select an item from a list, it does so according to a procedure that implements a selection rule denoted by the name Selection_Rule in Figures 2 and 3.

We instantiate the generic algorithm by specifying the following:

1) the overlay graphs governing sending messages to neighbors and what it means for a processor to consider itself compact, and

2) the rules used to select tasks to be performed in a round.

We use the overlay graphs $G(p, f)$ as specified in Section 3 for any $f < p$.

Lemma 3 If $W$ is the work complexity of the generic algorithm that uses the overlay graph $G(p, f)$, for any $f < p$, then its communication complexity is $O\left(\left(\frac{p-f}{p} \right)^{3.31} W\right)$.

Proof: Let $\Delta$ be the maximum degree of $G(p, f)$. The algorithm sends up to $\Delta$ messages for each unit of work, hence the communication cost is $M = O(\Delta \cdot W)$. The maximum degree $\Delta$ of the graph $G(p, f)$ is $O\left(\left(\frac{p-f}{p} \right)^{3.31}\right)$ by Lemma 1. □

Compactness. As stipulated in Section 3 a subgraph of the overlay graph is compact if its size is at least $(p-f)/7$ and its diameter is at most $g(p) = 30 \lg p + 2$. Let the range of a processor $v$ be the subgraph of the overlay graph that contains each non-faulty processor whose distance from $v$ in the overlay graph is at most $g(p)$. A processor is said to be compact if its range includes at least $(n-f)/7$ nodes. Any processor $v$ in a compact subgraph $H$ is compact since $H$ is included in the range of $v$.

The processors estimate distances to other processors in the overlay graph using their local views of this graph. The distances may change during an execution due to crashes and halts, so it is possible for a graph to stop being compact at some point in time. Each processor $v$ computes the subgraph of the overlay graph $G(p, f)$ induced by the processors in $\text{Processors}_v$, that is, by the processors that $v$ still considers being active. Next $v$ computes the distances from each node of $G(p, f)$ to itself in this subgraph. If the estimated size of the range of processor $v$ is at least $(p-f)/7$, then $v$ is said to consider itself compact. This property is used as a criterion for halting, see Figure 3. Observe that, if a processor $v$ is compact, but sufficiently many processors in the range of $v$ have halted, thus no longer being active, then $v$ may consider itself not compact at this point.

Correctness of the generic algorithm. Recall that the algorithm starts with initialization, followed by the repeat loop which, depending on the value of Phase$_v$, invokes either procedure Main or procedure Closing (see Figure 4).

Proposition 1 Any instantiation of the generic algorithm solves the Do-All problem against the unbounded adversary.
Proof: First we show that if a task is not in \( \text{Tasks}_v \) of processor \( v \) in some round, then the task is complete. Note that processor \( v \) removes a task from \( \text{Tasks}_v \) only if either \( v \) performs the task itself (line 2.b.3 in Figure 2), or \( v \) receives a copy of a \( \text{Tasks} \) list without this task in it (line 2.a.3 in Figure 2). The proof is by induction on the length of the path traversed by a sequence of messages that brings the information that the task is not on the list \( \text{Tasks}_u \) of some processor \( u \).

A processor may halt only while executing a closing phase. In order to enter this phase, the private variable \( \text{Done} \) needs to be set to true in the pseudocode of Figure 2. We show that if \( \text{Done}_v \) is true at some \( v \), then \( \text{Tasks}_u \) is empty at some \( u \); this in turn implies that all tasks are complete. Inspecting the code of \text{Main} in Figure 2, we see that there are only three ways for \( \text{Done}_v \) to be set to true. (1) When \( \text{Tasks}_v \) becomes empty (line 2.b in Figure 2); this provides the base of induction. (2) When \( v \) receives \text{Stop} signal (line 2.d), and (3) when some processor on \text{Processors} list is missing from the \text{Busy} list (line 2.c) because it was missing from some received \text{Busy} list. Processor \( v \) removes its identifier from \( \text{Busy}_v \) (line 2.e.1 in Figure 2) and sends \text{Stop} signal (line 3.a in Figure 3) only when its variable \( \text{Done}_v \) is true. This provides the inductive step, since each message that triggers setting \( \text{Done} \) to true increments by one the length of the path originated at a processor whose \( \text{Tasks} \) list became empty.

We need to show that the correctness and termination conditions are satisfied in any execution. Every non-faulty processor performs a task in a main phase, therefore a processor performs this phase at most \( t \) times. Similarly, every non-faulty processor removes at least one processor from its \text{Busy} list in each iteration of the closing phase, except possibly in its last iteration (if processor \( v \) halts at line 2.b in Figure 3 then it does not remove a processor from \text{Busy}_v), hence a processor performs this phase at most \( p \) times. Therefore each processor halts by round \( t + p \), unless it crashes earlier. There is at least one processor that never crashes according to the definition of the unbounded adversary. This processor halts eventually. If a processor \( v \) halts, it happens in the closing phase, so \( \text{Done}_v \) is true. This means that all the tasks are complete. \( \Box \)

Task selection rules. What remains to specify to make the generic algorithm fully instantiated is the task selection rules. We will consider four such rules. The resulting instantiations are named as follows:

- **Balance-Load**: a constructive deterministic algorithm based on load balancing;
- **Randomized-Permutations**: a constructive randomized algorithm using random permutations to select tasks;
- **Deterministic-Permutations**: a nonconstructive deterministic algorithm using permutations that are part of the code to select tasks;
- **Effort-Priority**: a hybrid deterministic algorithm that uses a modification of algorithm Deterministic-Permutations combined with algorithm DMY from [23].

The purpose of considering the randomized algorithm Randomized-Permutations is only to prove the existence of permutations that make Deterministic-Permutations efficient.

4.2 Epochs

We partition an execution of the generic algorithm into disjoint **epochs**, denoted \( \mathcal{E}_i \), for \( i \geq 0 \). Epoch \( \mathcal{E}_0 \) denotes the initialization. An epoch \( \mathcal{E}_i \), for \( i \geq 1 \), is defined to be a segment of \( g(p) = \)
30 \log p + 2$ consecutive phases; the index $i$ in $\mathcal{E}_i$ denotes the $i$th such segment of phases. Epochs include the phases occurring after termination.

We use $K_i$, for $i \geq 0$, to denote the subgraph of the overlay graph $G(p, f)$ induced by the nodes that are non-faulty at the beginning of the epoch $\mathcal{E}_i$. Similarly, we use $G_i$, for $i \geq 0$, to denote the subgraph of the overlay graph $G(p, f)$ induced by the nodes that are non-faulty through the end of the epoch $\mathcal{E}_i$. This means that epoch $\mathcal{E}_i$ begins with $K_i$ as the set of non-faulty nodes and ends with $G_i$ as the set of non-faulty nodes. We have that $|K_i| \geq p - f$ and $|G_i| \geq p - f$ and that the equalities $G_i = K_{i+1}$ and the inclusions $G_i \subseteq K_i$ hold for $i \geq 0$.

Some epochs $\mathcal{E}_i$ may have the property that among the processors that start the epoch as active (these make the set $K_i$) relatively many crash in the course of the epoch. The effect is such that those processors that remain non-faulty through the end of the epoch (these constitute the set $G_i$) make a small fraction of the original set $K_i$. Formally, an epoch $\mathcal{E}_i$ is defined to be stormy when $|G_i| < |K_i|/2$. An epoch $\mathcal{E}_i$ is calm if it is not stormy, that is, when $|G_i| \geq |K_i|/2$. In the analysis of work performance in the following sections, we are mostly concerned with calm epochs, because stormy epochs are taken care of by the following Lemma.

**Lemma 4** Stormy epochs contribute $O(p \log^2 p)$ to work.

**Proof:** There are $O(\log p)$ stormy epochs in any execution. An epoch lasts for $g(p) = O(\log p)$ rounds. In order to calculate the upper bound on work, i.e., the number of the available processor steps, it is sufficient to multiply the number $p$ of processors by the number $O(\log p)$ of stormy epochs and by the duration $O(\log p)$ of an epoch. □

**Core processors.** We define a sequence of subgraphs $H_i$ in $G(p, f)$, for $i \geq 0$, by stipulating that $H_i = P(G_i)$, where $P$ is the subgraph function of the overlay graph $G(p, f)$ from Definition 3 in Section 3. Lemma 2 in that section states that the construction satisfies the requirements of Definition 2 of subgraph property. Graph $H_i$ is called the core graph for $\mathcal{E}_i$, and the processors that belong to $H_i$ are the core processors for $\mathcal{E}_i$. When a core processor halts, it remains core, as we do not consider halted processors to be faulty.

The graphs $K_i$, $G_i$ and $H_i$, for $i \geq 0$, are uniquely determined by how crashes occur in an execution. It follows by the definition of subgraph property that each $H_i$ is compact and that $H_{i+1} \subseteq H_i$ and $|H_i| \geq |G_i|/7 \geq (p - f)/7$, for $i \geq 0$. These properties guarantee that some processors remain core throughout the whole execution.

The diameter of $H_i$, for $i \geq 0$, is at most $g(p)$, which is the reason why epoch is defined to be a segment of $g(p)$ consecutive phases. This definition of epoch guarantees that any epoch is of a sufficient duration for all the core processors to propagate their knowledge among themselves by flooding across the overlay graph. Here the knowledge of processors can be restricted to mean the contents of their lists, as this is what the algorithms uses. When a processor $v \in H_i$, that is core for an epoch $\mathcal{E}_i$, halts in $\mathcal{E}_i$, then $v$ cannot participate in such flooding. The next lemma helps to argue about flooding in such situations.

**Lemma 5** If a processor $v$ that is core for an epoch $\mathcal{E}_i$ halts in that epoch, then, in the last phase before $v$ halts, each of the lists $\text{Busy}_v$ and $\text{Tasks}_v$ is empty.

**Proof:** First we argue about the list $\text{Busy}_v$. A processor can halt only while executing procedure
Closing. By the pseudocode of this procedure in Figure 3, there are two possible triggers to halt: one is when the list Busy is empty and another is when the processor does not consider itself compact. We argue that the list Busy <i>v</i> is empty in the phase just after which <i>v</i> halts by induction on this phase’s number.

First, let us consider the base of induction. Any processor in <i>H</i> <i>i</i> is compact in <i>E</i> <i>i</i>, because <i>H</i> <i>i</i> is included in the <i>v</i>’s range. Therefore a processor in <i>H</i> <i>i</i> considers itself compact as long as no processor in <i>H</i> <i>i</i> has halted yet. The base of induction holds because when a processor <i>v</i> halts and no other core processor halted before then <i>v</i> considers itself compact and so its list Busy <i>v</i> is empty.

Next consider the inductive step. Let <i>v</i> halt in a phase such that any other core processor, say, <i>x</i> that halted before <i>v</i> had its list Busy <i>x</i> empty. Observe that this empty list Busy <i>x</i> got forwarded to the <i>x</i>’s neighbors just before <i>x</i> halted. This is by the pseudocode in Figure 3 which specifies that a processor halts after completing a phase, and so after sending all the messages in that phase. An alternative for the list Busy <i>v</i> to be empty, as a trigger making <i>v</i> halt, is that <i>v</i> does not consider itself compact. Let us consider this possibility to show that it will not occur. For <i>v</i> not to consider itself compact requires removing sufficiently many core processors from its list Processors <i>v</i>. A processor <i>w</i> is removed from the lists Processors of its neighbors for the first time after a phase with no communication from <i>w</i>, and later this information is reflected in the contents of the forwarded lists Processors. When <i>w</i> is core that halted before <i>v</i> then the empty list Busy <i>w</i> was received by the <i>w</i>’s neighbors before <i>w</i> stopped communicating with them, as we already argued. Therefore the information about the empty Busy lists of the core processors that halted before <i>v</i> reaches <i>v</i> before the information that would trigger removing these processors from the list Processors <i>v</i>. When <i>v</i> obtains the first such a message with an empty Busy list, then <i>v</i> makes its list Busy <i>v</i> empty in the same phase. The list Busy <i>v</i> stays empty, once set to be such, until <i>v</i> halts, because entries from this list may be deleted but are never added after initialization. This completes the proof of the inductive step.

Next we reason about the list Tasks <i>v</i>. A transition from Main to Closing occurs after line 2.e.3 of Figure 2 is executed. This line is triggered by the condition Done <i>v</i> = true in line 2.e that also causes list Tasks to be emptied in line 2.e.2. It follows that any processor executing procedure Closing has its list Tasks empty. A processor can halt only while executing procedure Closing, so its list Tasks is empty when it halts.

The properties of epochs stated as Lemmas 6 and 7 below are shown by referring to Lemma 3.

**Lemma 6** If a processor that is core in epoch <i>E</i> <i>i</i> switches to closing phases by the beginning of <i>E</i> <i>i</i>, then every core processor in <i>E</i> <i>i</i> switches to closing phases by the end of epoch <i>E</i> <i>i</i>, for <i>i</i> ≥ 1.

**Proof:** Let <i>v</i> be a processor in <i>H</i> <i>i</i> that switches to closing phases prior to the beginning of <i>E</i> <i>i</i>. Processor <i>v</i> removes its identifier from Busy <i>v</i> in the round when it switches to closing, per line 2.e.1 in Figure 2. Next <i>v</i> sends its lists to the neighbors, resulting in each of them removing <i>v</i> from their respective Busy lists, per line 2.a.2 in Figures 2 and line 2.a.2 in Figure 3. The fact that <i>v</i> is missing in a Busy list propagates through all core processors by flooding within one epoch. Node <i>v</i> stays in the lists Processors throughout epoch <i>E</i> <i>i</i>, since <i>v</i> belongs to <i>G</i> <i>i</i>. Suppose a processor <i>u</i> receives a copy of Busy list without <i>v</i> in it in <i>E</i> <i>i</i> and <i>u</i> is still executing main phases. The lists Busy <i>u</i> and Processors <i>u</i> are compared in line 2.e in Figure 2. The fact that Busy <i>u</i> ≠ Processors <i>u</i> causes <i>u</i> to set Done <i>u</i> to true and next to switch to closing phases in line 2.e.3 in Figure 2.
If such a flooding chain, as referred to above, fails, it is because some core processor \( w \) in the chain has already halted. By Lemma 5 when a core processor \( w \) halts then its list \( \text{Busy}_w \) is empty. Copies of this empty list \( \text{Busy}_w \) are forwarded to the \( w \)'s neighbors in the phase just before \( w \) halts, by the pseudocode in Figure 3. Such forwarding of the empty \( \text{Busy} \) list to the neighbors concludes flooding from the \( w \)'s perspective, because this is the ultimate information to be forwarded in a flooding chain, as the lists \( \text{Busy} \) may shrink but they never grow. It is sufficient to forward an empty \( \text{Busy} \) list to the neighbors only once. This means that a halted processor does not disrupt a flooding chain of copies of \( \text{Busy} \) list.

\[ \square \]

**Counting outstanding tasks.** We make use of the following notation for sets: \( T_{v,i} \) is a set that contains all items in \( \text{Tasks}_v \) at the end of epoch \( \mathcal{E}_i \), for \( i \geq 0 \). When an epoch's index \( i \) is understood from the context, then we may simply write \( T_v \) for \( T_{v,i} \). When considering how much progress has been made by the core processors, we use the following shorthand notation: \( U_i \) stands for the set \( \bigcup_{v \in H_i} T_{v,i} \), and \( S_i \) denotes \( \bigcap_{v \in H_i} T_{v,i} \), for \( i \geq 0 \). We use the following notations for numbers denoting sizes of sets: \( u_i = |U_i| \) and \( s_i = |S_i| \), for \( i \geq 0 \). For convenience, we let \( s_{-1} = u_{-1} = t \).

Epoch \( \mathcal{E}_i \) begins with \( u_{i-1} \) tasks such that each of them belongs to at least one \( \text{Tasks} \) list of a core processor, and it ends with \( u_i \) tasks such that each of them belongs to at least one \( \text{Tasks} \) list of a core processor. Similarly, epoch \( \mathcal{E}_i \) begins with the \( \text{Tasks} \) list of the core processors such that each list has at least \( s_{i-1} \) tasks in it, and it ends with \( \text{Tasks} \) list of the core processors such that each list has at least \( s_i \) tasks in it.

**Lemma 7** The inequality \( u_i \leq s_{i-1} \) holds, for any \( i \geq 0 \).

**Proof:** We argue that \( U_i \subseteq S_{i-1} \). It is sufficient to show that \( T_{v,i} \subseteq T_{w,i-1} \), for any two processors \( v \) and \( w \) in \( H_i \). If a task is not in \( \text{Tasks}_w \) at the beginning of \( \mathcal{E}_i \), then this information will propagate through all of \( H_i \) through flooding by the end of this epoch; this results in processor \( v \) removing the task from its list \( \text{Tasks}_v \) in the epoch, unless the task was removed earlier.

If such flooding fails, it is because some core processor \( w \) in the chain have already halted. By Lemma 5 when a core processor \( w \) halts then its list \( \text{Task}_w \) is empty. Copies of this empty list \( \text{Task}_w \) are forwarded to the \( w \)'s neighbors in the phase just before \( w \) halts, by the pseudocode in Figure 3. Such forwarding of the empty \( \text{Task} \) list to the neighbors concludes flooding from the \( w \)'s perspective, because this is the ultimate information to be forwarded in a flooding chain, since the lists \( \text{Task} \) may only become smaller (they never grow). It is sufficient to forward an empty \( \text{Tasks} \) list to the neighbors only once. This means that a halted processor does not disrupt a flooding chain of copies of \( \text{Tasks} \) lists.

\[ \square \]

The sequences \( \langle u_i \rangle_{i \geq 0} \) and \( \langle s_i \rangle_{i \geq 0} \) are nonincreasing, and \( s_i \leq u_i \), for \( i \geq 0 \), directly from their definitions. These properties combined with Lemma 7 allow us to use either one of the sequences \( \langle u_i \rangle_{i \geq 0} \) and \( \langle s_i \rangle_{i \geq 0} \) when measuring progress in the number of tasks completed.

**Main, mixed, and closing epochs.** Recall that phases of an execution of an instantiation of the generic algorithm are partitioned into main and closing phases. This partitioning of phases is independent for each processor, and is determined by which corresponding procedure the processor performs in a phase. In parallel to having this partitioning of phases, we also partition the epochs of an execution into the related categories, for two of which we use the same terms. Epochs are partitioned into the following groups, which make three disjoint contiguous segments of an
execution:

(1) **main epochs** are those in which all core processors are busy,

(2) **mixed epochs** are those in which some core processors are busy while some are not, and

(3) **closing epochs** are those in which no core processors are busy.

For calm main epochs $E_i$, it is sufficient to estimate total work by the contribution from core processors, as is shown in the following lemma.

**Lemma 8** If $w_i$ is the amount of work accrued by the core processors during a calm main epoch $E_i$, then the total work accrued during $E_i$ is $O(w_i)$.

**Proof:** No core processor halts in epoch $E_i$, as this is possible only in the course of executing Closing, by the pseudocode in Figures 2 and 3. Therefore graph $H_i$ stays intact throughout epoch $E_i$.

Graph $H_i$ is defined as the compact subgraph of $G_i$ of the form $H_i = P(G_i)$, where $P$ is a subgraph function by Lemma 2. It follows that $|G_i| \leq 7|H_i|$. We can combine this fact with the inequality $|K_i| < 2|G_i|$, given by the assumption, to obtain $|K_i| < 14|H_i|$. Therefore the total work accrued during $E_i$ is less than $14w_i$. □

When estimating work in closing epochs, we need to be cautious if we want to restrict our attention only to core processors, by analogy with Lemma 8, because there may still be compact processors with nonempty Busy lists after all core processors have halted. Should this be the case, we might carve out a compact graph from the subgraph of $G(p, f)$ induced by these nodes that have not halted yet. We observe now that a bound on work by core processors applies to the work performed by the processors in this carved graph, because the notion of core processors is conceptual only and used for the sake to calculate the amount of work which these processors contribute; in particular, no processor knows whether it is core or not in an epoch. The newly carved graph may behave similarly as the core graph, in the sense that eventually each of its nodes halts while there are still some other compact processors with nonempty Busy lists. This phenomenon can repeat itself a constant number of times (actually at most six times, because of the definition of compact graph). Therefore, with proper caution, we may consider core processors, and even resort to Lemma 8 while estimating closing work, as we show in the next lemma.

**Lemma 9** Suppose that an instantiation of the generic algorithm is such that if $p$ processors aim to perform $p$ tasks, then this is accomplished with $B(p)$ work in main epochs, for some function $B(p)$. Then the work in closing epochs of such an instantiation with $p$ processors is $O(B(p) + p \log^2 p)$.

**Proof:** Stormy epochs contribute $O(p \log^2 p)$ to work, by Lemma 4. We include this amount, as a component of the bound we seek to justify, and in what follows consider calm epochs only.

Observe that, starting from a round when there are no busy processors, the activity of the system can be interpreted as having at most $p$ processors perform at most $p$ tasks, since sending a Stop signal to processor $u$ can be interpreted as “performing task $u$.” The only difference between this interpretation of procedures Closing and Main is that some processors executing closing phases
may halt, while a processor that is done with all tasks in a main phase subsequently executes at least one closing phase.

The idea of the proof is to relate the closing work performed by all processors to the closing work of the nodes of some compact graph. To avoid a situation when each node of this compact graph will have already halted, while there are still some other compact processors with nonempty Busy lists, we may consider a compact subgraph of $G(p, f)$ that stays compact longest. The choice of the core graph for an epoch, from among all compact graphs surviving this epoch, is needed only to be able to consistently refer to specific compact nodes. Therefore we may assume, without loss of generality, that it is the core processors that stay compact longest. This approach allows us to refer to the properties already shown for core processors, such as that no processor $v$ that is core for an epoch ever halts as long as $\text{Busy}_v$ is nonempty and no core processor has halted yet, shown in Lemma 5 (this avoids the need to reason about such properties holding for other compact subgraphs).

Let us therefore assume that it is the core processors that stay compact longest in the execution. We consider first the closing rounds when each core processor $v$ has a nonempty Busy list. We may treat Busy lists in procedure Closing as playing the role of Tasks in procedure Main, which makes Lemma 8 applicable. The core processors perform at most $B(p)$ work during these rounds, by the assumption defining the quantity $B(p)$. The total contribution to work during these rounds is $O(B(p))$, by Lemma 8.

When some processor $v$ that is core for epoch $E_k$ observes in $E_k$ that Busy$_v$ is empty, then $v$ sends copies of this empty list to its neighbors in the same phase and halts after the phase is over. All core processors for $E_{k+1}$ halt by the end of $E_{k+1}$, because the information about the empty lists of busy processors is propagated by flooding; a processor that halts forwards this information to its neighbors just before halting, by the pseudocode in Figure 3. At the end of epoch $E_{k+1}$ there are no more compact processors, by the assumption we made about core processors. It takes at most one more epoch $E_{k+2}$ for each processor, that has not halted yet, to stop considering itself compact and halt. These three epochs $E_k, E_{k+1},$ and $E_{k+2}$, contribute $O(p \log p)$ extra work to the total closing work. $\square$

5 A Constructive Algorithm

We obtain a specific algorithm by instantiating the generic algorithm with a task-selection rule that assigns tasks to processors with the aim of balancing the load among the processors. The parameter $f < p$ determines the overlay graph $G(p, f)$ used for communication, as defined in Section 3 and discussed in Section 4. We call the resulting constructive algorithm BALANCE-LOAD.

The precise selection rule employed in algorithm BALANCE-LOAD is as follows. Consider processor $v$ and let $k > 0$ be the number of items in Tasks$_v$ list in a certain round, that is $k = |\text{Tasks}_v|$. We let $r(v)$ stand for $\left\lceil \frac{k \cdot v}{p} \right\rceil$. The positions on the list are numbered starting with 1. Processor $v$ selects the item at position $r(v)$ in its list of tasks to perform in this round; this selection is well-defined since $k \cdot \frac{v}{p} \leq k$. Number $r(v)$ is the largest integer $j$ satisfying the inequalities

$$k \cdot \frac{v}{p} - 1 < j < k \cdot \frac{v}{p} + 1. \quad (10)$$

Observe that $v$ as a number is the rank of $v$ on list Processors$_v$ after the initialization. As the
execution proceeds, the rank of $v$ in this list may change due to crashes, but this does not affect the choices of tasks by $v$, as $v$ does not use the size of $\text{Processors}_v$ when deciding which task to perform.

Note also that the distances between the positions of tasks selected by different processors in an epoch need to be at least $pg(p)$ if tasks are not to be duplicated in this epoch. For this to be possible during the initial epochs, $t$ needs to be at least $p^2g(p)$, when the position of a selected task determined as $r(v) = \left\lceil \frac{k_v}{p} \right\rceil$, which is equivalent to $\sqrt{t} \geq p\sqrt{g(p)}$. In view of this, we may expect expressions involving $\sqrt{t}$ to appear in a bound on work of algorithm $\text{Balance-Load}$ (this is indeed that case, as we state later in Theorem 11). Lemma 10 below covers the epochs during which the work performed by core processors is not duplicated and so such epochs contribute $O(t)$ to a bound on work. In the other case, when $\sqrt{t} < p\sqrt{g(p)}$, duplication of work can occur in any epoch. In such a situation, computing work is more involved as progress during an epoch in performing tasks depends on how many crashes occurred during the epoch and how much progress has been made in the previous epoch. (Later in this section we address the involved cases leading to Theorem 11)

Next we formally analyze the algorithm’s performance. Recall the following notation: $U_i = \bigcup_{v \in H_i} T_{v,i}$ and $u_i = |U_i|$, for $i \geq 0$, and similarly $S_i = \bigcap_{v \in H_i} T_{v,i}$ and $s_i = |S_i|$, for $i \geq 0$.

**Lemma 10** If $u_{i-1} \geq 11p^2g(p)$ then $u_{i-1} - u_i \geq |H_i|g(p)$, for $i \geq 1$.

**Proof:** We consider two cases, based on the value of $u_{i-1} - s_{i-1}$.

Case 1:

$$u_{i-1} - s_{i-1} \geq pg(p).$$

In this case, we combine (11) with Lemma 7 to obtain the estimates

$$u_{i-1} - u_i \geq u_{i-1} - s_{i-1} \geq pg(p) \geq |H_i|g(p).$$

Case 2: $u_{i-1} - s_{i-1} < pg(p)$. For any processor $v$, we calculate the decrease in the size of the list Tasks$_v$ during epoch $E_i$. At most $u_{i-1} - s_{i-1} < pg(p)$ tasks can be removed from this list due to discrepancy of the knowledge of processor $v$ and the respective knowledge by other processors in the range of $v$ about the outstanding tasks. Clearly, at most $g(p)$ tasks are performed by processor $v$ during epoch $E_i$.

We want to show that during epoch $E_i$ each task is performed by at most one processor, as this property implies that the number of outstanding tasks decreases by at least $|H_i|g(p)$ during epoch $E_i$. To this end, consider the ranks of tasks in the ordered set $U_{i-1} = \bigcup_{v \in H_{i-1}} T_{v,i-1}$. In the first phase of epoch $E_i$, processor $v$ performs a task whose rank in Tasks$_v$ is $r(v)$; let $w(v)$ be the rank of this task in $U_{i-1}$. Consider any other task performed by $v$ during $E_i$ and let $w'(v)$ be its rank in $U_{i-1}$. List Tasks$_v$ is a dynamic data structure and the ranks of its elements may change during $E_i$, while the rank of a task with respect to $U_{i-1}$ stays the same during $E_i$. Let $T_v$ denote the set $T_{v,i-1}$. We can combine (10), for $k = |T_v|$, the number of elements in Tasks$_v$ at the beginning of $E_i$, with the fact that (11) does not hold to obtain that

$$|T_v| \cdot \frac{v}{p} - 1 < w(v) < (|T_v| + pg(p)) \cdot \frac{v}{p} + 1.$$  \hspace{1cm} (12)

The list Tasks$_v$ may shed at most $2pg(p)$ items during epoch $E_i$; this implies

$$(|T_v| - 2pg(p)) \cdot \frac{v}{p} - 1 < w'(v) < (|T_v| + pg(p)) \cdot \frac{v}{p} + 1.$$ \hspace{1cm} (13)
By comparing the lower and upper boundaries of the ranges $[12]$ and $[13]$, we obtain that

$$|w'(v) - w(v)| < 3pg(p) + 2.$$  

(14)

For each processor $v$, we define $B_v \subseteq U_{i-1}$ to be the set of tasks whose ranks differ from $w(v)$ by at most $3pg(p) + 2$. It follows from (14) that all tasks performed by $v$ during $\mathcal{E}_i$ are in $B_v$.

Consider two different processors $v_1$ and $v_2$, where $v_1 > v_2$. We show that their corresponding sets $B_{v_1}$ and $B_{v_2}$ are disjoint. We first estimate the distance in $U_{i-1}$ of the tasks they both perform in the first phase of $\mathcal{E}_i$. We use the following inequalities

$$r(v_1) > |T_{v_1}| \cdot \frac{v_1}{p} - 1 \quad \text{and} \quad r(v_2) < |T_{v_2}| \cdot \frac{v_2}{p} + 1$$

(15)

to estimate the ranks in the relevant Tasks lists. We then combine (15) with the assumption that (11) does not hold to obtain the following estimates of the ranks in $U_{i-1}$:

$$w(v_1) > |T_{v_1}| \cdot \frac{v_1}{p} - 1 \quad \text{and} \quad w(v_2) \leq r(v_2) + u_{i-1} - s_{i-1} < |T_{v_2}| \cdot \frac{v_2}{p} + 1 + pg(p).$$

This in turn leads to the following estimate:

$$w(v_1) - w(v_2) > |T_{v_1}| \cdot \frac{v_1}{p} - 1 - |T_{v_2}| \cdot \frac{v_2}{p} - 1 - pg(p)$$

$$\geq (|T_{v_2}| - pg(p)) \cdot \frac{v_1}{p} - |T_{v_2}| \cdot \frac{v_2}{p} - 2 - pg(p)$$

$$= |T_{v_2}| \cdot \frac{v_1 - v_2}{p} - (v_1 + p)g(p) - 2$$

$$\geq (u_{i-1} - pg(p)) \cdot \frac{v_1 - v_2}{p} - (v_1 + p)g(p) - 2$$

$$\geq 10pg(p) - 2pg(p) - 2$$

$$> 7pg(p).$$

Here we used the assumption that $u_{i-1} \geq 11p^2g(p)$ and the fact that $p \geq v_1 > v_2$. Notice that $7pg(p) > 2[3pg(p) + 2]$; together with (14) this shows that $B_{v_1}$ and $B_{v_2}$ are disjoint.

Lemma 11 Let $\mathcal{E}_k$ be an epoch and consider another epoch $\mathcal{E}_i$ such that $i > k$. If all processors in $H_i$ are in their main phases during $\mathcal{E}_i$ and $x$ is a number such that $u_{i-1} \geq x \geq 7$, then $u_{i-1} - u_i$ is either at least $\min\{\frac{1}{14p}, 1\}|H_i|$ or at least $\frac{u_{i-1}}{x}$.

Proof: Suppose that the inequality

$$u_{i-1} - u_i < \frac{u_{i-1}}{x}$$

(16)

holds. We need to show that $u_{i-1} - u_i$ is either at least $\frac{x}{14p}|H_i|$ or at least $|H_i|$. Let $v_1$ and $v_2$ be the identifiers of some two processors in $H_i$ that perform the same task in the first phase of $\mathcal{E}_i$, where $v_1 > v_2$; let $r(v_1)$ and $r(v_2)$ be the ranks of this task in Tasks$_{v_1}$ and Tasks$_{v_2}$, respectively, in this phase. We first show that (16) implies

$$v_1 - v_2 < \frac{7p}{x}.$$  

(17)
Lemma 7 gives the estimate \( u_{i-1} - s_{i-1} \leq u_{i-1} - u_i \). We combine this with (16) to obtain the inequality \( u_{i-1} - s_{i-1} < \frac{u_{i-1}}{x} \), which yields

\[
    s_{i-1} > u_{i-1}\left(1 - \frac{1}{x}\right).
\]  

(18)

Let \( T_{v_1} = T_{v_1,i-1} \) and \( T_{v_2} = T_{v_2,i-1} \). Substituting the estimates \( s_{i-1} \leq |T_{v_j}| \) and \(|T_{v_j}| \leq u_{i-1}\) in (18), one obtains

\[
    u_{i-1}\left(1 - \frac{1}{x}\right) \leq |T_{v_j}| \leq u_{i-1},
\]

for both \( j = 1 \) and \( j = 2 \), which next yields

\[
    \max\{|T_{v_1} \setminus T_{v_2}|, |T_{v_2} \setminus T_{v_1}|\} \leq \frac{u_{i-1}}{x}.
\]

(20)

Inequality (20) together with the estimate \(|r(v_1) - r(v_2)| \leq \max\{|T_{v_1} \setminus T_{v_2}|, |T_{v_2} \setminus T_{v_1}|\}\) imply

\[
    |r(v_1) - r(v_2)| \leq \frac{u_{i-1}}{x},
\]

(21)
as \( r(v_1) \) and \( r(v_2) \) are the ranks of the same task in \( \text{Tasks}_{v_1} \) and \( \text{Tasks}_{v_2} \), respectively. We bound the difference \( v_1 - v_2 \) from above using (14) as follows:

\[
    v_1 - v_2 < \frac{p}{|T_{v_1}|} \left( r(v_1) + 1 \right) - \frac{p}{|T_{v_2}|} \left( r(v_2) - 1 \right).
\]

(22)

We continue bounding \( v_1 - v_2 \) from above by substituting the estimates given by (19) and (21) into the right-hand side of the inequality (22):

\[
    v_1 - v_2 < p \cdot \left( \frac{r(v_2) + u_i/x + 1}{|T_{v_1}|} - \frac{r(v_2) - 1}{|T_{v_1}| + u_i/x} \right)
    = p \cdot \frac{r(v_2)u_{i-1}/x + (u_{i-1}/x)^2 + u_{i-1}/x + (u_{i-1}/x)|T_{v_1}| + 2|T_{v_1}|}{|T_{v_1}|(|T_{v_1}| + u_{i-1}/x)} 
    \leq p \cdot \frac{u_{i-1}^2}{x} \cdot \frac{r(v_2)/u_{i-1} + 1/x + 1/u_{i-1} + |T_{v_1}|/u_{i-1} + 2x|T_{v_1}|/u_{i-1}^2}{u_{i-1}(1 - 1/x)(u_{i-1}(1 - 1/x) + u_{i-1}/x)}.
\]

(23)

Note that the numbers \( r(v_2) \), \( |T_{v_1}| \) and \( x \) are at least 1, and at most \( u_{i-1} \), each. We obtain the following upper bound

\[
    v_1 - v_2 \leq p \cdot \frac{u_{i-1}^2}{x} \cdot \frac{6}{u_{i-1}(1 - 1/x)} = \frac{6p}{x - 1} \leq \frac{7p}{x},
\]

by estimating the expression (23) and using the fact that \( x \geq 7 \). This completes the proof of (17).

The concluding argument is as follows. Suppose first that \( 7p < x \). Then at least \( |H_i| \) processors perform different tasks. Otherwise, if \( 7p \geq x \), then at least

\[
    \frac{|H_i|}{7p/x} + 1 \leq \frac{|H_i|}{2 \cdot 7p/x} = |H_i| \cdot \frac{x}{14p}
\]

processors in \( H_i \) perform different tasks. \( \square \)
For any execution, the indices $i$ of main epochs $E_i$ form a contiguous interval, call it $\mathcal{I}$. For the epochs $E_i$ where $i \in \mathcal{I}$, the processors in $H_i$ are busy executing main phases. We first define two subsets $I_1$ and $I_2$ of $\mathcal{I}$ as follows:

Set $I_1$ contains all $i$ from $\mathcal{I}$ for which $E_i$ is stormy, that is, $|G_i| < |K_i|/2$.

Set $I_2$ contains all $i$ from $\mathcal{I}$ with the property $u_{i-1} \geq 11p^2g(p)$.

Let $k + 1$ be the smallest among all epoch indices $i$ such that $u_i < 11p^2g(p)$. We will use Lemma 11 for $x$ equal to $\sqrt{\frac{u_k}{\log u_k}}$; in what follows we use $x$ as a shorthand for $\sqrt{\frac{u_k}{\log u_k}}$. Let us define three subsets $I_3$, $I_4$, and $I_5$ of $\mathcal{I} \setminus (I_1 \cup I_2)$ as follows:

Set $I_3$ contains all $i$ from $\mathcal{I} \setminus (I_1 \cup I_2)$ such that $u_{i-1} < x$.

Set $I_4$ contains all $i$ from $\mathcal{I} \setminus (I_1 \cup I_2)$ such that $u_{i-1} \geq x$ and $u_{i-1} - u_i \geq \frac{u_{i-1}}{x}$.

Set $I_5$ contains all $i$ from $\mathcal{I} \setminus (I_1 \cup I_2)$ such that $u_{i-1} \geq x$ and

\[
  u_{i-1} - u_i \geq \min\left\{\frac{x}{\sqrt{\log u_k}}, 1\right\}|H_i|.
\]

It follows from Lemma 11 that $I_3 \cup I_4 \cup I_5 = \mathcal{I} \setminus (I_1 \cup I_2)$, and hence $I_1 \cup I_2 \cup I_3 \cup I_4 \cup I_5 = \mathcal{I}$.

The following theorem summarizes the properties of algorithm Balance-Load.

**Theorem 1** Algorithm Balance-Load is a constructive deterministic solution for the Do-All problem with $p$ processors and $t$ tasks that can be instantiated, for any known $f < p$, to have work $O(t + p \log p (\sqrt{p \log p} + \sqrt{t \log t}))$.

**Proof:** We estimate the amount of work in main, mixed and closing epochs separately. We first consider main epochs, whose indices are in $\mathcal{I}$. We consider the contribution of each set from the cover $I_1 \cup I_2 \cup I_3 \cup I_4 \cup I_5$ of the interval $\mathcal{I}$ one by one.

The epochs with indices in $I_1$ contribute $O(p \log^2 p)$ to work, by Lemma 11. The amount of work during epochs $E_i$ for $i \in \mathcal{I} \setminus I_1$ that are calm, is estimated by calculating the contribution by the respective core processors, using Lemma 8.

The work performed by the core processors, during epochs $E_i$ whose indices are in $I_2$, is $O(t)$, as during these epochs the number of the outstanding tasks decreases proportionately to the number of executed available processor steps, per Lemma 10.

The work performed by the core processors, during the epochs whose indices are in $I_3$, is $O(p \log p \sqrt{\frac{t}{\log t}})$, since $x \leq \sqrt{\frac{t}{\log t}}$ and an epoch consists of $g(p) = O(\log p)$ phases.

If $i \in I_4$, then $u_i \leq u_{i-1}(1 - \frac{1}{x})$, so each such epoch contributes to decreasing $u_{i-1}$ by a factor of at least $1 - \frac{1}{x}$. The number of such epochs $E_i$ is $O(\frac{x}{\sqrt{\log u_k}})$, because

\[
  u_k \left(1 - \frac{1}{x}\right)^{x \log u_k} = O(1).
\]

It follows that

\[
|I_4| = O(x \log u_k) = O\left(\sqrt{\frac{u_k}{\log u_k}} \log u_k\right) = O(\sqrt{u_k \log u_k}) = O(\sqrt{t \log t}).
\]
as \( u_k \leq t \). Therefore the work accrued by the core processors, during the epochs whose indices \( i \) are in \( I_4 \), is \( \mathcal{O}(p \log p \sqrt{t \log t}) \).

Next we consider the work by the core processors during the epochs whose indices are in \( I_5 \). We estimate the sum of the decrements \( u_{i-1} - u_i \) for \( i \in I_5 \) by the telescoping series:

\[
\sum_{i \in I_5} (u_{i-1} - u_i) \leq \sum_{i \geq k+1} (u_{i-1} - u_i) \leq u_k.
\]

After combining this with (27), we obtain that

\[
u_k \geq \sum_{i \in I_5} (u_{i-1} - u_i) \geq \sum_{i \in I_5} \min\left\{ \frac{\sqrt{t}}{14p}, 1 \right\} |H_i|.
\]

This implies that

\[
\sum_{i \in I_5} |H_i| \leq u_k \max\left\{ \frac{14p}{\sqrt{t}}, 1 \right\} \leq \max\left\{ \frac{14p u_k \sqrt{\log u_k}}{u_k^{1/3}}, u_k \right\}.
\]

Since \( u_k \) is less than \( 11p^2 g(p) \) and at most \( t \), we obtain that

\[
u_k < p \sqrt{11u_k g(p)} \leq p \sqrt{11 t g(p)}.
\]

This makes the right-hand side of (25) to be at most

\[
\max\{14p \sqrt{\log u_k}, u_k\} \leq \max\{14p \sqrt{t \log t}, p \sqrt{11 t g(p)}\}.
\]

By (25) and (26), the work contributed by the processors in the core graphs, during epochs \( \mathcal{E}_i \) for \( i \in I_5 \), is

\[
\sum_{i \in I_5} |H_i| g(p) \leq p g(p) \max\{14 \sqrt{t \log t}, \sqrt{11 t g(p)}\} = \mathcal{O}(p \log p (\sqrt{t \log t} + \sqrt{t \log p})).
\]

The work contributed by the core processors during the epochs \( \mathcal{E}_i \) with \( i \in \mathcal{I} \), is obtained by summing up the contributions of \( I_1, I_2, I_3, I_4, \) and \( I_5 \). We obtain

\[
\mathcal{O}(p \log^2 p) + \mathcal{O}(t) + \mathcal{O}(p \log p \sqrt{\frac{t}{\log t}}) + \mathcal{O}(p \log p \sqrt{t \log t}) + \mathcal{O}(p \log p (\sqrt{t \log t} + \sqrt{t \log p})) = \mathcal{O}(t + p \log^2 p + p \log p \sqrt{\log p})
\]

as the bound on total work performed during main epochs.

Two consecutive mixed epochs are sufficient for every core processor to learn that there are no outstanding tasks, by Lemma 6. These mixed epochs contribute only \( \mathcal{O}(p \log p) \) to work.

To estimate the work during closing epochs, we apply Lemma 9. The bound on this work is obtained by substituting \( p \) for \( t \) in (27) and adding \( \mathcal{O}(p \log^2 p) \), yielding

\[
\mathcal{O}(p + p \log^2 p + p \log p \sqrt{\log p}) = \mathcal{O}((p \log p)^{3/2})
\]

This bound on work in closing epochs, combined with \( \mathcal{O}(p \log p) \) work in mixed epochs, and with bound (27) on work in main epochs, gives

\[
\mathcal{O}(t + p \log^2 p + p \log p \left(\sqrt{t \log t} + \sqrt{t \log p}\right) + (p \log p)^{3/2}) = \mathcal{O}(t + p \log p (\sqrt{p \log p} + \sqrt{t \log t}))
\]

as the bound on total work. \( \square \)
Note that the bound in Theorem 1 does not depend on \( f \), even though the instantiations of the generic algorithm use overlay graphs of degrees determined by \( f \), as spelled out in Lemma 1. The degrees of overlay graphs affect communication, as stipulated in Lemma 3.

**Corollary 1** Algorithm \textsc{Balance-Load} is a constructive deterministic solution for the Do-All problem with \( p \) processors and \( t \) tasks that can be instantiated, for a number of crashes \( f \) bounded by \( f \leq cp \), for any known constant \( 0 < c < 1 \), to have effort \( O(t + p \log p (\sqrt{p \log p} + \sqrt{t \log t})) \).

**Proof:** Since the adaptive known adversary is restricted to be linearly bounded, the assumption about failures implies \( p - f \geq (1 - c)p \). This makes the bound on communication given in Lemma 3 to become \( M = O(W) \). The bound on work is taken from Theorem 1. We then combine the bounds on work \( W \) and communication \( M \) to obtain a bound on effort \( E = W + M \). \( \square \)

6 An Algorithm Optimized for Work

We explore a selection paradigm where processors choose items from lists in the order determined by their private permutations. We consider two ways for establishing such suitable permutations. First, we consider a randomized algorithm, called \textsc{Randomized-Permutations}, where processors use private permutations generated randomly at the beginning of an execution. (Recall that the purpose of considering the randomized algorithm \textsc{Randomized-Permutations} is only to prove the existence of certain permutations that are then used in a deterministic algorithm.) Next, we consider a deterministic algorithm, called \textsc{Deterministic-Permutations}, where the code is parameterized by private permutations assigned to processors. We show that algorithm \textsc{Deterministic-Permutations} can be equipped with suitable permutations that make it more efficient than the deterministic algorithm \textsc{Balance-Load}. All our algorithms use the same constructive overlay graphs for communication. The difference between \textsc{Balance-Load} and \textsc{Deterministic-Permutations} is that the former is constructive, whereas the latter is nonconstructive, as it resorts to a selection rule that uses nonconstructive permutations.

We now specify how permutations are used to select items from lists. Each processor \( v \) has two private permutations: \( \pi_1 \) over the set \( \{1, \ldots, p\} \), and \( \pi_2 \) over the set \( \{1, \ldots, 11p^2g(p)\} \). Processor \( v \) permutes its list \( \text{Busy}_v \) according to the permutation \( \pi_1 \). To explain how the permutation \( \pi_2 \) is used, we consider two cases.

The first case is for \( t \geq 11p^2g(p) \). Processors select tasks to perform according to a load balancing rule. This rule is the same as the one used in algorithm \textsc{Balance-Load}, as long as the size of \( \text{Tasks}_v \) is greater than \( 11p^2g(p) \). As soon as exactly \( 11p^2g(p) \) tasks remain in the list \( \text{Tasks}_v \), processor \( v \) permutes \( \text{Tasks}_v \) according to the permutation \( \pi_2 \).

In the second case we have \( t < 11p^2g(p) \). Here the list \( \text{Tasks}_v \) consisting of \( t \) entries is rearranged by the permutation \( \pi_2 \) as if it consisted of \( 11p^2g(p) \) entries. More precisely this means the following: we pad the list \( \text{Tasks}_v \) to the length \( 11p^2g(p) \) by appending “dummy” items, then permute the resulting list, and finally remove the dummy items to compact the list to its original size.

Each time processor \( v \) needs to select an item from \( \text{Tasks}_v \), it selects the first item from the rearranged list, then removes it from the list. If \( v \) needs to select a processor from \( \text{Busy}_v \) in a closing phase, then it selects the first processor from the rearranged list, then removes it from the list.

32
The randomized algorithm `RANDOMIZED-_PERMUTATIONS` starts by each processor selecting two permutations uniformly at random and independently across all processors. The deterministic algorithm `DETERMINISTIC-_PERMUTATIONS` equips each processor with its individual pair of permutations. We show the existence of permutations that guarantee work $O(t + p \log^2 p)$ of `DETERMINISTIC-_PERMUTATIONS`. A technical challenge we encounter here is to demonstrate that the complexity of `DETERMINISTIC-_PERMUTATIONS` can be made suitably small for some permutations. Algorithm `RANDOMIZED-_PERMUTATIONS` is used only as an intermediate step to facilitate the analysis of the deterministic counterpart. The idea is to estimate the probability that algorithm `RANDOMIZED-_PERMUTATIONS` deviates from its expected performance, in order to next apply the probabilistic method to argue that there exists a family of permutations that achieves a comparable performance in the worst case.

We use the following Chernoff bound on the probability that a sum of independent random variables deviates from its expectation: For $0 < r < 1$, let $X_1, \ldots, X_n$ be a sequence of independent Bernoulli trials with $\Pr(X_j = 1) = r$ and $\Pr(X_j = 0) = 1 - r$, for each $1 \leq j \leq n$. If $S = \sum_{j=1}^{n} X_j$ then, for any $0 < \varepsilon < 1$, the following inequality holds [46]:

$$\Pr(S \leq (1 - \varepsilon)nr) \leq \exp(-n\varepsilon^2/2) . \quad (28)$$

### 6.1 Extended epochs

Recall that an epoch $E_i$ begins with $K_i$ as the set of non-faulty nodes, that it ends with $G_i$ as the set of non-faulty nodes, and that the core graph $H_i$ is a compact subgraph of $G_i$, for integer $i \geq 0$. We define an extended epoch $D_j$ of an execution, for integer $j \geq 0$, to be a contiguous segment of (regular) epochs of the execution, subject to the following additional restrictions.

The first extended epoch $D_0$ consists of the first regular epoch $E_0$, which denotes initialization. Suppose that $E_k$ is the first regular epoch of an extended epoch $D_j$, for some $j \geq 0$. If either $|G_k| < |K_k|/2$ or $|H_k|g(p) \geq u_{k-1}$ hold, then the extended epoch $D_j$ consists of only this one regular epoch $E_k$.

Otherwise, when both $|G_k| \geq |K_k|/2$ and $|H_k|g(p) < u_{k-1}$ hold, then the extended epoch $D_j$ consists of a contiguous segment of epochs $\langle E_k, \ldots, E_t \rangle$ such that $\ell \geq k$ is the largest integer for which both $|G_{\ell}| \geq |K_{\ell}|/2$ and $(\ell - k + 1)|H_{\ell}|g(p) < u_{k-1}$ hold.

The number $\ell$ is well defined because the quantity $(\ell - k + 1)(p - f)g(p)$ is unbounded as a function of $\ell$. When the extended epoch $D_j$ is the segment $\langle E_k, \ldots, E_{\ell} \rangle$ of regular epochs, then the next extended epoch $D_{j+1}$ begins with the regular epoch $E_{\ell+1}$.

We define the number $m(j)$ to be the index of the first regular epoch in $D_j$, and, similarly, the number $l(j)$ to be the index of the last regular epoch in $D_j$, so that $D_j = \langle E_{m(j)}, \ldots, E_{l(j)} \rangle$. We also define $m(-1) = -1$ and $l(-1) = -1$ for convenience of notation, letting $u_{m(-1)} = t$ and $u_{l(-1)} = t$.

**Stormy extended epochs.** If $|G_{l(j)}| \geq |K_{m(j)}|/2$, then the extended epoch $D_j$ is called calm, otherwise, when the inequality $|G_{l(j)}| < |K_{m(j)}|/2$ holds, then this extended epoch $D_j$ is stormy.

**Lemma 12** An extended epoch is stormy if and only if it consists of one stormy regular epoch.

**Proof:** Let us consider the first regular epoch $E_k$ of an extended epoch $D_j$, so that $m(j) = k$. If epoch $E_k$ is stormy then $|G_k| < |K_k|/2$ and so, by the definition of an extended epoch, $D_j$ consists of only the epoch $E_k$, and also $D_j$ is clearly a stormy extended epoch. If $E_k$ is calm then the inequality
\[ |G_k| \geq |K_k|/2 \] holds. Now \( D_j \) consists of a contiguous segment of epochs \( \langle E_m(j), \ldots, E_l(j) \rangle \) such that both \( |G_l(j)| \geq |K_m(j)|/2 \) and \( (l(j) - m(j) + 1)|H_l(j)|g(p) < u_{m(j) - 1} \) hold, so \( D_j \) is calm.

**Lemma 13** Work performed during stormy extended epochs is \( O(p \log^2 p) \).

**Proof:** Each stormy extended epoch consists of one stormy regular epoch, by Lemma 12. The stormy regular epochs, that make also stormy extended epochs, contribute \( O(p \log^2 p) \) to work, by Lemma 4.

**Core processors and main extended epochs.** We define \( Z_j \) to denote \( H_l(j) \), for an extended epoch \( D_j \). The subgraph \( Z_j \) is called the core subgraph for \( D_j \), and the processors in \( Z_j \) are the core processors for \( D_j \).

Extended epochs in which all core processors are busy are called main extended epochs.

**Lemma 14** If \( w_j \) work is accrued by the core processors during a calm main extended epoch \( D_j \), then the total work accrued during \( D_j \) is \( O(w_j) \).

**Proof:** No core processor halts in the extended epoch \( D_j \), as this is possible only in the course of executing Closing, by the pseudocode in Figures 2 and 3. Therefore graph \( Z_j \) stays intact throughout the extending epoch \( D_j \).

Graph \( Z_j \) is defined as the compact subgraph of \( G_l(j) \) of the form \( Z_j = P(G_l(j)) \), where \( P \) is a subgraph function by Lemma 2. It follows that \( |G_l(j)| \leq 7|H_j| \). We can combine this fact with the inequality \( |K_m(j)| < 2|G_l(j)| \), given by the assumption, to obtain \( |K_m(j)| < 14|Z_j| \). Therefore the total work accrued during \( D_j \) is less than \( 14w_j \).

We denote by \( c_j \) the number of selections of tasks made by the core processors (those in \( Z_j \)) during the extended epoch \( D_j \); this number \( c_j \) is called the core number for \( D_j \). We have that

\[ c_j = (l(j) - m(j) + 1)|Z_j|g(p), \]

as there are \( l(j) - m(j) + 1 \) regular epochs in \( D_j \), each taking \( g(p) \) rounds, with one selection of a task by a processor in \( Z_j \) per round. An extended epoch \( D_j \) begins with \( u_{l(j - 1)} \) tasks still present in some lists of the processors, and ends with \( s_l(j) \) tasks that occur in every list.

We partition the extended epochs of an execution into two categories as follows. One of them consists of extended epochs \( D_j \) with \( c_j < u_{l(j - 1)} \), for \( j \geq 0 \); such extended epochs are called task-rich extended epochs. Another category consists of extended epochs \( D_j \) with \( c_j \geq u_{l(j - 1)} \), for \( j \geq 0 \); such extended epochs are called task-poor extended epochs.

**Lemma 15** An extended epoch that is task-poor consists of one regular epoch.

**Proof:** Consider an extended epoch \( D_j = \langle E_m(j), \ldots, E_l(j) \rangle \). If it is task-poor then this means \( |H_l(j)|g(p) \geq u_{l(j - 1)} \), because \( Z_j = H_l(j) \). Therefore, by the definition of extended epoch, extended epoch \( D_j \) consists of only the regular epoch \( E_l(j) = E_m(j) \).
**Productive extended epochs.** An extended epoch \( D_j \) is called *productive* when the inequality

\[
  u_{t(j-1)} - s_{t(j)} \geq \min\{u_{t(j-1)}, c_j\}/4
\]

holds at the completion of \( D_j \). The right-hand side of (29) can be simplified when we know if the extended epoch is task-poor or task-rich. Task-rich extended epochs \( D_j \) are defined by \( c_j < u_{t(j-1)} \), therefore the inequality

\[
  \min\{u_{t(j-1)}, c_j\}/4 \geq c_j/4
\]

holds for \( j \in J_4 \). Task-poor extended epochs \( D_j \) are defined by \( c_j \geq u_{t(j-1)} \), therefore the inequality

\[
  \min\{u_{t(j-1)}, c_j\}/4 \geq u_{t(j-1)}/4
\]

holds for \( j \in J_5 \).

**Lemma 16** There are \( \mathcal{O}(\log p) \) productive task-poor main extended epochs, in any execution of the generic algorithm.

**Proof:** We use the sequence \( \langle u_i \rangle_{i \geq 0} \) as reflecting progress in completing tasks (see the discussion just after Lemma 7). If \( D_j \) is a productive task-poor extended epochs then we can combine (29) with (31) to obtain that

\[
  u_{t(j-1)} - u_{m(j+1)} \geq u_{t(j-1)}/4 ,
\]

as \( u_{m(j+1)} \leq s_{t(j)} \). Any such extended epochs \( D_j \) consist of only one regular epoch \( E_m(j) = E_{t(j)} \), by Lemma 15. The inequality (32) means that two consecutive regular epochs \( E_{m(j)} \) and \( E_{m(j)+1} \) contribute to decreasing the number of outstanding tasks by at least \( u_{m(j)}/4 \leq u_{t(j-1)/4} \). The number of tasks in the first productive task-poor main extended epoch is less than \( 11p^2g(p) \), so there are \( \mathcal{O}(\log(11p^2g(p)) = \mathcal{O}(\log p) \) productive task-poor main extended epochs.

For any execution, the indices \( j \) of main extended epochs \( D_j \) form a contiguous interval, call it \( \mathcal{J} \). We first identify two subsets of \( \mathcal{J} \) as follows:

- Set \( J_1 \) consists of these \( j \in \mathcal{J} \) for which \( D_j \) is stormy, that is, \( |G_{t(j)}| < |K_{m(j)}|/2 \).
- Set \( J_2 \) consists of these \( j \in \mathcal{J} \) for which \( u_{t(j-1)} \geq 11p^2g(p) \).

There is at most one extended epoch \( D_j \) such that \( u_{m(j)} \geq 11p^2g(p) \) and \( u_{t(j)} < 11p^2g(p) \). We make \( J_3 \) to be the singleton set \( \{j\} \), when such \( D_j \) exists, otherwise \( J_3 = \emptyset \).

The extended epochs \( D_j \) for \( j \in \mathcal{J} \setminus (J_1 \cup J_2 \cup J_3) \) are called *key*. We define two subsets \( J_4 \) and \( J_5 \) of the set of indices of the key extended epochs \( \mathcal{J} \setminus (J_1 \cup J_2 \cup J_3) \) as follows:

- Set \( J_4 \) consists of these \( j \in \mathcal{J} \setminus (J_1 \cup J_2 \cup J_3) \) for which the key extended epoch \( D_j \) is a task-rich extended epoch.
- Set \( J_5 \) consists of these \( j \in \mathcal{J} \setminus (J_1 \cup J_2 \cup J_3) \) for which the key extended epoch \( D_j \) is a task-poor extended epoch.

These definitions are such that \( \mathcal{J} = J_1 \cup J_2 \cup J_3 \cup J_4 \cup J_5 \).

**Lemma 17** For any known \( f < p \), the work accrued in any execution of the generic algorithm is \( \mathcal{O}(t + p\log^2 p) \) during the main extended epochs \( D_j \), provided that all the key extended epochs among them are productive.
Lemma 7). With this interpretation of the numbers \( \text{Lemma 10} \). The contribution of the segment \( \langle E_k \rangle \) against the general adaptive adversary.

We use the sequence \( \langle u_i \rangle_{i \geq 0} \) as reflecting progress in completing tasks (see the discussion just after Lemma 7). With this interpretation of the numbers \( u_{(j-1)} \) and \( u_{m(j+1)} \), we see that bound (33) means that the progress in completing tasks during \( D_j \) and \( D_{j+1} \), assuming \( D_j \) is task-rich, is proportional to \( c_j \), which is the work expended by the processors that are core for \( D_j \). It follows that the work performed by core processors during productive task-rich extended epochs is \( O(t) \).

As the next step, we estimate the amount of work by the core processors during task-poor key extended epochs \( D_j \); their indices make the set \( J_4 \). These extended epochs are productive, by the assumption. We combine (29) and (30) with \( u_{m(j+1)} \leq s_{l(j)} \) to obtain

\[
 u_{(j-1)} - u_{m(j+1)} \geq c_j / 4. \tag{33}
\]

We use the sequence \( \langle u_i \rangle_{i \geq 0} \) as reflecting progress in completing tasks (see the discussion just after Lemma 7). With this interpretation of the numbers \( u_{(j-1)} \) and \( u_{m(j+1)} \), we see that bound (33) means that the progress in completing tasks during \( D_j \) and \( D_{j+1} \), assuming \( D_j \) is task-rich, is proportional to \( c_j \), which is the work expended by the processors that are core for \( D_j \). It follows that the work performed by core processors during productive task-rich extended epochs is \( O(t) \).

As the next step, we estimate the amount of work by the core processors during task-poor key extended epochs \( D_j \); their indices make the set \( J_5 \). There are \( O(\log p) \) such extended epochs \( D_j \), by Lemma 16 as each of them is productive, by the assumption. Therefore task-poor extended epochs contribute \( O(p \log^2 p) \) work, since each such extended epoch consists of a single regular epoch, by Lemma 16.

There is at most one extended epoch \( D_j \) such that \( u_{m(j)} \geq 11p^2g(p) \) and \( u_{l(j)} < 11p^2g(p) \); if it exists then \( J_3 = \{ j \} \), otherwise \( J_3 \) is empty. We can break this extended epoch \( D_j \) into two segments, determined by some integer \( k \), for \( m(j) \leq k < l(j) \), such that \( u_k \geq 11p^2g(p) \) and \( u_{k+1} < 11p^2g(p) \). The first segment begins with \( E_{m(j)} \) and ends with \( E_k \). The second segment begins with \( E_{k+1} \) and ends with \( E_{l(j)} \). The regular epochs in the segment \( \langle E_{m(j)}, \ldots, E_k \rangle \) contribute \( O(t) \) to work, by Lemma 16. The contribution of the segment \( \langle E_{k+1}, \ldots, E_{l(j)} \rangle \) can be treated either similarly to the contribution of a productive extended epoch that is task-rich, when \( (l(j) - k) |H_{l(j)}| g(p) \leq u_k \), or similarly to the contribution of a productive extended epoch that is task-poor, when the inequality \( (l(j) - k) |H_{l(j)}| g(p) > u_k \) holds. This is because \( (l(j) - k) |H_{l(j)}| g(p) \) is the number of the respective selections of tasks, so it represents the same quantity as the core number. Therefore the contribution of work during the extended epoch \( D_j \), for \( j \in J_3 \), is bounded from above by the work obtained for the extended epochs whose indices are in \( J_1 \cup J_2 \cup J_4 \cup J_5 \).

We obtain that \( O(t + p \log^2 p) \) is the bound on work, by summing up the contributions of all \( D_j \), for \( j \in \mathcal{J} \), where \( \mathcal{J} = J_1 \cup J_2 \cup J_3 \cup J_4 \cup J_5 \). \[ \square \]

### 6.2 Restricted adversaries and randomization

We expect that if permutations are assigned to the processors in a random and independent manner then work will be efficient with positive probability for crashes happening at arbitrary times. If so, this will be sufficient to claim the existence of permutations that can be fixed in the code of the algorithm so that work will be efficient in a deterministic instantiation of the generic algorithm against the general adaptive adversary.
We consider restricted adversaries as technical means for approximating the behavior of the general adaptive adversaries in the probabilistic analysis of a randomized instantiation of the generic algorithm. Random selections of tasks by core processors make the dominating contribution to progress in performing tasks in the course of an extended epoch. These selections are made independently, which facilitates the probabilistic analysis. For such analysis to hold, we need to be certain that the processors that make the selections have not crashed yet. To this end, we simply specify what processors are core at the beginning of the extended epoch, so that these processors will not crash through the end of the extended epoch. This is presented as if the adversary were constrained not to be able to crash these particular processors in this execution. This is only a conceptual constraint, because ultimately we will consider all the possible sets of processors that do not crash through the end of an extended epoch. The adaptive adversary is restricted by the number of crashes anyway, so in retrospect we can interpret any specific execution, in which some processors end up as operational at the end of an extended epoch, as if the adversary were restricted at the outset of the extended epoch to spare these processors.

Bars on crashes. Recall that we denote by $K_{m(j)}$ the set of processors that are non-faulty at the beginning of the extended epoch $D_j$, and by $G_{l(j)}$ the set of non-faulty processors at the end of $D_j$.

A set of processors $B \subseteq K_{m(j)}$ is said to be a bar for crashes for $D_j$ if the adversary is constrained at the beginning of $D_j$ by the requirement that $B = G_{l(j)}$.

This definition means that the processes in the set $B$ cannot fail and that all the processes in the set $K_{m(j)} \setminus B$ must fail during the extended epoch $D_j$. Alternatively, given a set of processors $B$, we could consider extended epochs $D_j$ that happen to result in $B = G_{l(j)}$, rather than present this as a constraint on the adversary. This is because we consider all possible bars on crashes in the proof of Lemma 19, where we use Lemma 18. For each extended epoch $D_j$ the respective set $G_{l(j)}$ is well defined, so in retrospect we can consider $G_{l(j)}$ as bar for crashes when $D_j$ ends.

Lemma 18 For any known $f < p$, if a bar for crashes is specified at the beginning of a main extended epoch $D_j$ of an execution of algorithm Randomized-Permutations, then $D_j$ is not productive with probability at most $\exp(-c_j/16)$.

Proof: Let us consider a specific extended epoch $D_j$. If the inequality

$$s_{l(j-1)} \leq \frac{3u_{l(j-1)}}{4}$$

is satisfied, then the following estimates

$$u_{l(j-1)} - s_{l(j-1)} \geq u_{l(j-1)} - s_{l(j-1)} \geq u_{l(j-1)} - \frac{3u_{l(j-1)}}{4} = \frac{u_{l(j-1)}}{4}$$

hold as well. This implies that the inequality [29], which defines productive extended epochs, holds with certainty. Next consider the case when [29] does not hold. Let $b$ denote $\min\{u_{l(j-1)/4}, c_j\}$; this is the right-hand side of [29]. What needs to be shown is that the inequality $u_{l(j-1)} - s_{l(j)} \geq b$ holds with the claimed probability.

Let us order one by one the $c_j$ random selections of tasks by the core processors in $D_j$. These selections determine a sequence of operations performed, as specified by the algorithm. The adversary does not crash any of the core processors in the course of $D_j$, as each of them belongs to $G_{l(j)}$. 37
Let $S^{(0)}$ denote $S_{l(j-1)}$, and let $S^{(k)}$ be the set of tasks from $S_{l(j-1)}$ that are not completed by the $k$th selection, in this one-by-one ordering.

We define the random variables $X_k$ as follows, for $1 \leq k \leq c_j$:  
\[
X_k = \begin{cases} 
1 & \text{if either } u_{l(j-1)} - |S^{(k)}| \geq b \text{ or } S^{(k)} \subsetneq S^{(k-1)} \\
0 & \text{otherwise.}
\end{cases}
\]

The intended meaning of this definition is that $X_k$ equals 1 when either

(a) the target inequality $u_{l(j-1)} - s_{l(j)} \geq b$ is already satisfied in the round when the $k$th task is selected, or

(b) the $k$th task selected from $S_{l(j-1)}$ increases the set of the tasks already selected to be performed in this extended epoch, that is, this task has not been selected to be performed in this extended epoch prior to this $k$th selection.

We define $X = \sum_{1 \leq k \leq c_j} X_k$. This random variable has the property that $X \geq b$ if and only if $u_{l(j-1)} - s_{l(j)} \geq b$. The random variables $X_1, \ldots, X_k$ are well-defined and mutually independent, as long as $u_{l(j-1)} - |S^{(k)}| < b$. This is because these random variables are determined by random and mutually independent selections of tasks by the core processors, while no such processor may crash in $D_j$, due to the assumed restriction on the adversary.

Suppose that processor $v$ is to perform its $k$th operation in $D_j$. We want to estimate how many tasks in $\text{Tasks}_v$ are outstanding. Regarding the contents of $\text{Tasks}_v$ at the end of the extended epoch $D_{j-1}$, at most $u_{l(j-1)} - |S^{(k-1)}|$ tasks in $\text{Tasks}_v$ are performed during the first $k-1$ operations, since the size of $\text{Tasks}_v$ at the end of $D_{j-1}$ is at most $u_{l(j-1)}$. Let us recall two inequalities now. One is $s_{l(j-1)} > \frac{3u_{l(j-1)}}{4}$; it holds because the inequality (34) does not. The second is $b \leq \frac{u_{l(j-1)}}{4}$; it holds because $b$ denotes $\frac{\min\{u_{l(j-1)}, c_j\}}{4}$. By combining these two inequalities, we obtain that if $u_{l(j-1)} - |S^{(k-1)}| < b$, then
\[
s_{l(j-1)} - (u_{l(j-1)} - |S^{(k-1)}|) > \frac{3u_{l(j-1)}}{4} - \frac{u_{l(j-1)}}{4} = \frac{u_{l(j-1)}}{2}.
\]

It follows that at least half of the tasks that are in the list $\text{Tasks}_v$, in the round when the $k$th selection is made, have not been performed yet. Therefore the probability that $v$ chooses an outstanding task in the $k$th selection is at least $1/2$, and this outcome is independent from the preceding choices. Selecting an outstanding task to perform in the $k$th operation results in $S^{(k)} \subsetneq S^{(k-1)}$, which means that $X_k = 1$.

Consider a sequence $\langle Y_k \rangle_{1 \leq k \leq c_j}$ of independent Bernoulli trials, with $\Pr(Y_k = 1) = 1/2$. Define $Y = \sum_{1 \leq k \leq c_j} Y_k$. Now we claim $X$ statistically dominates $Y$, in the sense that, for any $d > 0$:
\[
\Pr(X \leq d) \leq \Pr(Y \leq d) .
\]

This is because of how the constituent random variables $X_k$ and $Y_k$ contribute to $X$ and $Y$, respectively. Namely, the random variables $X_k$ of $X$ have the property that each $X_k = 1$ with probability
at least $1/2$ independently from the prior values, as long as either $k = 1$ or

$$\sum_{j=1}^{k-1} X_j < u_{l(j-1)} - s_{l(j-1)} - b,$$

while $X_k = 1$ when the following is satisfied

$$\sum_{i=1}^{k-1} X_i \geq u_{l(j-1)} - s_{l(j-1)} - b.$$

Comparing $X$ to $Y$, the constituent random variables $Y_i$ of $Y$ are mutually independent Bernoulli trials, in which $Y_i = 1$ with probability $1/2$, for all $1 \leq i \leq c_j$.

We resort to (35) and $b \leq c_j$ to estimate the probability $\Pr(X \leq b)$. These inequalities yield

$$\Pr(Y \leq b) \leq \Pr\left(Y \leq \frac{c_j}{4}\right) = \Pr\left(Y \leq \left(1 - \frac{1}{2}\right) \cdot c_j \cdot \frac{1}{2}\right).$$

The right-hand side of this bound is at most $\exp(-c_j/16)$, by the Chernoff bound (28), in which we substitute $n = c_j$, $\varepsilon = 1/2$, and $r = 1/2$.

**Adversaries and $f$-chains.** We make the following definitions towards finalizing the randomized analysis.

An infinite sequence of sets of processors $\mathcal{C} = \langle V_k \rangle_{k \geq 0}$ is called an $f$-chain if, for each $k \geq 0$:

1. $V_0 = V$ is the set of all processors and $V_{k+1} \subseteq V_k$,
2. $|V_k| \geq p - f$,
3. either $|V_k| = p - f$ or $|V_k| = \lceil p/2^i \rceil$, for some $i \geq 0$.

Recall that we denote by $G_{l(j)}$ the set of non-faulty processors at the end of $D_j$, for $j \geq 0$. An adversary is said to be $f$-chain-constrained when any execution of the algorithm, in which the adversary controls the timing of crashes, has the property that the sequence $\langle G_{l(j)} \rangle_{j \geq 0}$ of sets of processors is an $f$-chain.

The following Lemma 19 extends the estimate on the probability of one key extended epoch to be productive, as given in Lemma 18 in the case when one extended epoch is restricted by a bar on crashes, to the estimate on the probability of an execution to include only productive key extended epochs, in the case when the adversary is $f$-chain constrained.

**Lemma 19** For any known $f < p$ and for sufficiently large $p$, the probability that some key extended epochs in an execution of algorithm RANDOMIZED-PERMUTATIONS are not productive is $O(p^{-1/6} \log p)$ when the algorithm is executed against the $f$-chain-constrained adversary.

**Proof:** When the adversary is $f$-chain-constrained, then the sequence of sets of processors $\langle G_{m(j)} \rangle_{j \geq 0}$ that stay non-faulty through the end of an extended epoch $D_j = \langle E_{m(j)}, \ldots, E_{l(j)} \rangle$ make an $f$-chain. This means that the numbers in the corresponding sequence $\langle |G_{m(j)}| \rangle_{j \geq 0}$ are of a restricted form.
We partition key extended epochs into groups $C_k$, for $0 \leq k \leq \lfloor \lg \frac{p}{f} \rfloor$, determined by these numbers. Namely, group $C_k$ consists of these key extended epochs $\mathcal{D}_j$, for which $|G_{m(j)}| = \lfloor p/2^k \rfloor$, for $0 \leq k < \lfloor \lg \frac{p}{f} \rfloor$, and $|G_{m(j)}| = p - f$ in the case of $k = \lfloor \lg \frac{p}{f} \rfloor$.

The core graph $Z_j$, for an extended epoch $\mathcal{D}_j$ in $C_k$, has at least $\frac{p}{f \cdot 2^{k+1}}$ processors, by Lemma 2. It follows that the core number $c_j$ is at least $\frac{pg(p)}{7 \cdot 16 \cdot 2^{k+1}}$, even when the extended epoch $\mathcal{D}_j$ consists of just a single regular epoch. The probability that an extended epoch $\mathcal{D}_j$ with a specific set $G_{l(j)}$ is not productive is at most $\exp(-pg(p)/(7 \cdot 16 \cdot 2^{k+1}))$, by Lemma 18.

If $\mathcal{D}_j$ is in $C_k$, for $0 \leq k \leq \lfloor \lg \frac{p}{f} \rfloor$, then there are at most $2^{p/2^k}$ sets $G_{l(j)}$. By the union bound, the probability that an extended epoch in $C_k$ is not productive is thus at most

$$2^{p/2^k} \exp \left(- \frac{pg(p)}{7 \cdot 16 \cdot 2^{k+1}} \right) = \exp \left(- \frac{p}{2^k} \left( \frac{g(p)}{224} - \ln 2 \right) \right) \leq \exp(-g(p)/225),$$

for sufficiently large $p$. There are at most $1 + \log p$ groups $C_k$, since $k \leq \lfloor \lg \frac{p}{f} \rfloor$. We can sum the right-hand sides of (35), over all these $k$, to obtain the following estimate

$$\mathcal{O}(\log p) e^{-g(p)/225} = \mathcal{O}(\log p) p^{-30/(226\ln 2)} < \mathcal{O}(\log p) p^{-1/6},$$

that holds for sufficiently large $p$. This is a bound on the probability that some key extended epoch is not productive. \qed

### 6.3 General adversaries and deterministic computation

Next we show that the generic algorithm can be instantiated with fixed permutations at processors such that $\mathcal{O}(t + p \log^2 p)$ is the worst-case bound on work.

**Lemma 20** Algorithm Deterministic-Permutations can be instantiated, for any known $f < p$, to have work $\mathcal{O}(t + p \log^2 p)$ against the $f$-chain-constrained adversary during main extended epochs.

**Proof:** Let us consider algorithm Randomized-Permutations executed against the $f$-chain-constrained adversary. The probability of the event that some key extended epochs in an execution of algorithm Randomized-Permutations are not productive is bounded above by Lemma 19.

That bound is less than 1 for sufficiently large $p$, and then the probability that all key extended epochs are productive is positive. By the principle of the probabilistic method, we obtain that there exist such permutations that, when algorithm Deterministic-Permutations is instantiated with these permutations, then all key extended epochs in any execution of this instantiation against the $f$-chain-constrained adversary are productive. This allows to make use of Lemma 17 to obtain that the work accrued by such an instantiation of algorithm Deterministic-Permutations against the $f$-chain-constrained adversary during main extended epochs is $\mathcal{O}(t + p \log^2 p)$. \qed

**Lemma 21** Suppose that a deterministic instantiation of algorithm Deterministic-Permutations is such that if $p$ processors aim to perform $t$ tasks against an $f$-chain-constrained adversary, for a known $f < p$, then this is accomplished with at most $C(p, t)$ work and within $D(p, f, t)$
rounds during main regular epochs, for some functions $C(p,t)$ and $D(p,f,t)$. If this instantiation is executed against the general $f$-bounded adversary, then its work is $O(C(p,t) + p\log^2 p)$ and time is $D(p,f,t)$ during main regular epochs.

**Proof:** Let us consider an execution $\mathcal{S}$ of the given deterministic instantiation of algorithm **Deterministic-Permutations** against the general $f$-bounded adversary. This adversary is adaptive and constrained by $f$ only as the bound on the number of crashes. We will compare the main work performed in $\mathcal{S}$ to the work performed during a specific execution $\mathcal{T}$ of the algorithm against the $f$-chain-constrained adversary.

Recall that the $f$-chain-constrained adversary is restricted such that extended epochs $D_j$ begin with the set of processors $K_{m(j)}$ whose sizes have the property that $|K_{m(j)}| = \lceil p/2^k \rceil$, for $0 \leq k < \lfloor \log_p \frac{p}{m-j} \rfloor$, or $|K_{m(j)}| = p - f$. For a given execution in which crashes are controlled by an $f$-chain-constrained adversary, let us call a natural number $f$-critical (or just critical if $f$ is clear from the context) if it is the cardinality of a set of processors $K_{m(j)}$ in the execution.

The execution $\mathcal{T}$ is obtained from $\mathcal{S}$ in such a way that processors in $\mathcal{T}$ are crashing in the same order as in $\mathcal{S}$ and not later than in $\mathcal{S}$, but some processors that crash in $\mathcal{S}$ crash possibly earlier in $\mathcal{T}$ than in $\mathcal{S}$. We maintain two invariants regarding timing of crashes. One is that a processor that crashes in $\mathcal{T}$ does it at the very end of some epoch. The other is that the number of processors that crash by the end of each epoch of $\mathcal{T}$ is a critical number.

We consider the consecutive regular epochs of $\mathcal{S}$ one by one to determine crashes that determine the corresponding epochs of $\mathcal{T}$. We denote by $\mathcal{E}_i$ and by $\mathcal{E}'_i$ the $i$th epoch of $\mathcal{S}$ and $\mathcal{T}$, respectively. Similarly, we denote by $\mathcal{G}_i$ and $\mathcal{G}'_i$ the sets of processors that are non-faulty at the end of epoch $\mathcal{E}_i$ and $\mathcal{E}'_i$, respectively. The first epoch $\mathcal{E}_0$ of $\mathcal{S}$ consists of initialization, and this same epoch starts $\mathcal{T}$ as $\mathcal{E}'_0$. Suppose we have considered the epochs $\mathcal{E}_0$ through $\mathcal{E}_j$ of $\mathcal{S}$ to determine the corresponding epochs $\mathcal{E}'_0$ through $\mathcal{E}'_j$ of $\mathcal{T}$. Now consider $\mathcal{E}_{j+1}$. If some processors crash in $\mathcal{E}_{j+1}$ and they have not crashed in $\mathcal{S}$ by the end of $\mathcal{E}_j$ then we crash all these processors at the very end of $\mathcal{E}'_j$ and additionally as many additional processors as needed so that the size of the set $\mathcal{G}'_j$ is a critical number, but the minimum such a number.

All the processors terminate in the execution $\mathcal{S}$ not later than in $\mathcal{T}$ because in each round the processors that are non-faulty in $\mathcal{T}$ make a subset of the processors that are non-faulty in $\mathcal{S}$. So if $D(p,f,t)$ is the bound on rounds in executions against $f$-chain-constrained adversary then it is a bound for the general adversary constrained by $f$, the bound on the number of crashes.

Moreover, if epoch $\mathcal{E}_i$ is calm, then in any round of $\mathcal{E}_i$ the ratio of the cardinalities of these sets is at most 4; this is because the number of non-faulty processors decreases by at most a factor of 2 in $\mathcal{E}_i$ and the number of non-faulty processors in $\mathcal{E}_i$ drops below a critical number in a round of $\mathcal{E}_i$ then this contributes another factor of 2. We conclude that the amount of work performed in $\mathcal{S}$ during calm epochs is at most four times the corresponding amount of work in $\mathcal{T}$. The stormy epochs contribute $O(p\log^2 p)$ work to any execution, by Lemma 4. It follows that the main work in $\mathcal{S}$ is at most $4C(p,t) + O(p\log^2 p)$.

The following Theorem summarizes the properties of algorithm **Deterministic-Permutations**.

**Theorem 2** Algorithm **Deterministic-Permutations** is a deterministic nonconstructive solution for the Do-All problem that, for any known numbers of $p$ processors, $t$ tasks, and bound on the
number of crashes \( f < p \), can be instantiated to have work \( \mathcal{O}(t + p \log^2 p) \).

Proof: Let us take an instantiation of the algorithm as provided by Lemma 20. Its work against the \( f \)-chain-constrained adversary during main extended epochs is always \( \mathcal{O}(t + p \log^2 p) \), as stated in Lemma 20. By Lemma 21 we obtain that the amount of work accrued by this instantiation during main extended epochs against the general \( f \)-bounded adversary, constrained only by \( f \) as a bound on the number of crashes, is also always \( \mathcal{O}(t + p \log^2 p) \).

Let us consider an execution against the general \( f \)-bounded adversary and its extended epochs after the main ones. We may restrict our attention only to calm extended epochs, as stormy extended epochs contribute \( \mathcal{O}(p \log^2 p) \) work, by Lemma 13. Let \( D_j \) be the first such extended epoch that is not main. This means that some core processor starts executing procedure Closing during this extended epoch. There are two cases, depending on whether \( D_j \) is task-poor or task-rich.

If \( D_j \) is task-poor, then it consists of just one regular epoch, say epoch \( E_k \), by Lemma 15. All the core processors switch to performing closing phases by the end of \( E_k+1 \), by Lemma 6. These two epochs \( E_k \) and \( E_{k+1} \) contribute \( \mathcal{O}(p \log p) \) work, so the total work during main regular epochs is \( \mathcal{O}(t + p \log^2 p) \).

If \( D_j \) is task-rich, then the inequality \( c_j < u_l(j-1) \) holds. The number \( c_j \) is the amount of work spent by the core processors in \( D_j \). This work may be charged to the number of tasks \( t \), because \( u_l(j-1) \leq t \). All the core processors switch to performing closing phases by the end of the first regular epoch after the extended epoch \( D_j \), by Lemma 6. This one regular epoch contributes \( \mathcal{O}(p \log p) \) work, so the total work during main regular epochs is \( \mathcal{O}(t + p \log^2 p) \).

We apply Lemma 9 to estimate the work performed during closing regular epochs. We need to take the amount obtained by substituting \( p \) for \( t \) in the estimate \( \mathcal{O}(t + p \log^2 p) \) on the work during main regular epochs, which is \( \mathcal{O}(p \log^2 p) \), and add \( \mathcal{O}(p \log^2 p) \) to it. This yields \( \mathcal{O}(p \log^2 p) \) as the bound on the amount of work performed during closing epochs.

Observe that the bound on work given in Theorem 2 does not depend on \( f \), as it is also the case in Theorem 1. These algorithms use the respective overlay graphs that depend on \( f \). Hence the amount of communication the algorithms generate depends on \( f \), as estimated in Lemma 3 in Section 4.

Corollary 2 Algorithm Deterministic-Permutations is a deterministic nonconstructive solution for the Do-All problem with \( p \) processors and \( t \) tasks that can be instantiated, for a number of crashes \( f \) bounded by \( f \leq cp \), for any known constant \( 0 < c < 1 \), to have effort \( \mathcal{O}(t + p \log^2 p) \).

Proof: The argument is similar to that used in the proof of Corollary 1 in Section 5. The difference is in referring to the bound in Theorem 2.

7 An Algorithm Optimized for Effort

In this section we present an algorithm, called Effort-Priority, that is optimized for effort complexity against the unbounded adversary. This algorithm is structured to trade work for effort efficiency, and so it is not as work-efficient as Deterministic-Permutations, but it curbs the
amount of communication when the number of crashes becomes large. Recall that the communication complexity of Deterministic-Permutations is estimated using Lemma 3, where the expression in the big-O bound can be as high as $\omega(p^3)$ for $f$ sufficiently close to $p$.

In optimizing effort, it is natural to balance the bounds on work and communication. To do so, we employ a hybrid approach to algorithm design, where we first use Deterministic-Permutations for some duration, then we use some other algorithm to complete the remaining work. We use Deterministic-Permutations for as long as the number of failures $f$ is sufficiently small, namely $f < f_1$ for a certain $f_1 < p$ that is used by the code. Additionally, Deterministic-Permutations is modified so that its communication is comparable to its work, which is accomplished by having processors send messages only after sufficiently many tasks have been performed. We use $T_1$ to denote the number of rounds that is sufficient for algorithm Deterministic-Permutations to complete all tasks if there are at most $f_1$ crashes in an execution; this number $T_1$ is used in the code. If first running Deterministic-Permutations for $T_1$ rounds results in completion of all tasks, then we are done. Otherwise we set stage for transitioning to another algorithm by renaming the processors that are still non-faulty. Specifically, after renaming we invoke the algorithm of De Prisco et al. [23], referred to as algorithm DMY.

Fact 3 ([23]) Algorithm DMY has work $O(p(f + 1) + t)$ and message complexity $O(p(f + 1))$.

Renaming the non-faulty processors is necessary in order to use the bounds given in Fact 3 because algorithm DMY assumes that the processors commencing the computation are consecutively numbered (e.g., if there are $k$ processes, then they have unique identifiers from the interval $[1,k]$).

Specifying algorithm Effort-Priority involves determining how $f_1$ and $T_1$ depend on $p$ and $t$. These parameters are chosen so that the work in the invocation of Deterministic-Permutations is approximately the same as the work in the invocation of DMY. This balancing is done by means of the bounds given in Theorem 2 and Fact 3. What we will obtain is that the parameter $f_1$ needs to be close to $p$, namely, $f_1 = p(1 - p^{-c})$, for a constant $0 < c < 1$ (the details will be given in Section 7.2). It turns out that with as few as $p^{1-c}$ processors remaining non-faulty at the invocation of algorithm DMY we can afford to restore the lists of tasks to their original form, thus redoing some of the tasks, without compromising the performance bounds.

7.1 The design of the algorithm

Algorithm Effort-Priority is structured in terms of three parts, named Part-One, Part-Two, and Part-Three. Part-One executes a suitably modified algorithm Deterministic-Permutations. Part-Two provides a transition from to algorithm DMY, its purpose is to coordinate the renaming of the presumed active processes. Part-Three executes algorithm DMY, with the (presumed) active processes referred to by their new names. The three parts are given in more detail as follows.

Part-One invokes the modified version of Deterministic-Permutations for up to $T_1$ rounds. We call this modified algorithm Modified-Det-Perm, and it is obtained as follows. The algorithm uses the overlay graph $L(p)^{\ell}$ with the threshold $f_1$ playing the role of the number of crashes and with $\ell$ as determined by Lemma 1; this means that $G(p, f_1)$ is used as the overlay graph. We use parameter $\Delta_1$ to denote the maximum degree of $G(p, f_1)$. The input set of tasks is partitioned into chunks, with each chunk consisting of $\Delta_1$ tasks. Each such chunk is now treated as a single “task.”
in the sense that messages are sent precisely after a chunk of tasks is performed. After \( T_1 \) rounds of Modified-Det-Perm, the algorithm continues its execution with Part-Two.

Part-Two begins with each processor \( v \) having only its list \( \text{Processors}_v \) as the current approximation of the set of active processors. The active processors need to agree on a set of processors that is simultaneously a subset of the processors that finish Part-One as active, and a superset of the set of processors that are still active when Part-Three begins; we call such a set of processors checkpoint, following [23]. Finding a checkpoint is the goal of Part-Two, which is accomplished as follows.

At the start of Part-Two, every processor \( v \) that has not halted yet sends a message to each processor in its list \( \text{Processors}_v \), including itself. We refer to this round, and the messages circulated in it, as preparatory. Upon receiving the messages sent in the preparatory round, processor \( v \) removes from list \( \text{Processors}_v \) the identifiers of the processors from whom preparatory messages did not arrive. Each processor \( v \) creates a new private list \( \text{Coordinators}_v \) and initializes it to \( \text{Processors}_v \) once the preparatory messages have been processed. Each private copy of Coordinators is ordered in the increasing order of processor identifiers. The next \( 4(p - f_1) \) rounds are structured in groups of four rounds; we call each such group the checkpointing phase. The pseudocode for the preparatory round and a checkpointing phase is given in Figure 4.

When a checkpointing phase begins, each processor \( v \) that is ranked first in its list \( \text{Coordinators}_v \), broadcasts its proposal to serve as the phase coordinator to each processor in \( \text{Processors}_v \). It may happen that multiple processors send such messages in Round 1 of the phase, while we want to have only one coordinator. If processor \( v_1 \) proposes itself as coordinator in Round 1, but receives a proposal from processor \( v_2 \), such that \( v_1 > v_2 \), it does not consider itself coordinator for this phase; otherwise \( v_1 \) stays as the coordinator for the whole phase. A processor that acts as a coordinator through the whole phase is called the successful coordinator for this phase.

In Round 2 of a phase, any processor \( v \) adds the identifiers of all proposers in the previous round to its list \( \text{Coordinators}_v \), in case these identifiers are missing, and responds to the successful coordinator by sending a copy of its list \( \text{Processors}_v \).

In Round 3, the successful coordinator \( v \) removes from \( \text{Processors}_v \) any identifiers that are missing from any list \( \text{Processors} \) just received from other processors. In case the successful coordinator \( v \) does not receive a message from some processor \( w \) whose name occurs in \( \text{Processors}_v \) in the beginning of Round 3, processor \( v \) does not remove \( w \) from \( \text{Processors}_v \) as long as \( w \) occurs in each list \( \text{Processors} \) received by \( v \) in this round. This is done by the coordinator to “stabilize” the lists \( \text{Processors} \) to the eventual checkpoint, so that the “stabilized” lists are not modified even when additional crashes occur. The successful coordinator \( v \) disseminates its updated list \( \text{Processors}_v \) at the end of Round 3 among all processors with identifiers in \( \text{Processors}_v \).

In Round 4, each recipient of a copy of the list \( \text{Processors} \) from the coordinator replaces its list \( \text{Processors} \) with the received one. Next, each processor deletes the first entry on its list \( \text{Coordinators} \). The round may include additional deletions from these lists as follows. It may happen that processor \( v \), that has not acted as a successful coordinator yet, receives in the first round of the phase a message from a would-be coordinator \( w \) such that \( w > v \). This indicates a discrepancy between the lists \( \text{Coordinators}_v \) and \( \text{Coordinators}_w \); in particular, processor \( w \) has already removed \( v \) from \( \text{Coordinators}_w \), while \( w \) is still listed behind \( v \) in \( \text{Coordinators}_w \). In this situation, processor \( v \) removes any identifier in \( \text{Coordinators}_v \) ahead of its own identifier \( v \) to become the first on the list, subsequently proposing itself as the coordinator in the next phase.
**Preparatory Round:** Each processor $v$ sends a preparatory message to all processors in its list $\text{Processors}_v$, including itself.

Upon receiving preparatory messages, processor $v$ removes from $\text{Processors}_v$ the identifiers of processors from whom such messages are not received. The list $\text{Coordinators}_v$ is set to $\text{Processors}_v$, sorted in the order of identifiers.

**Checkpointing Phase**

**Round 1:** If $v$’s name occurs as the first entry in $\text{Coordinators}_v$ then $v$ sends a message to each processor in $\text{Processors}_v$, including itself, proposing to become coordinator.

**Round 2:** If $v$ does not receive coordinator proposal messages from Round 1, then this Round 2 is void.

Otherwise, let $v_1, \ldots, v_k$ be the identifiers of the processors from which $v$ received coordinator proposals, sorted according to the identifiers, i.e., $v_1 < \ldots < v_k$.

First, processor $v$ responds to $v_1$ by sending its list $\text{Processors}_v$. Next, processor $v$ adds any identifiers from $\{v_1, \ldots, v_k\}$ that are not already in $\text{Coordinators}_v$ to $\text{Coordinators}_v$, maintaining the sorted order of entries.

If $v$ proposed itself in Round 1 as the coordinator, but $v$ is not the smallest identifier in $\text{Coordinators}_v$, then $v$ no longer considers itself coordinator for the current phase.

**Round 3:** If $v$ still considers itself coordinator for the phase then, upon receiving response messages sent in Round 2, $v$ removes from its list $\text{Processors}_v$ the identifiers that are missing from any list $\text{Processors}_w$ just received. Next $v$ sends its updated list $\text{Processors}_v$ to each processor in $\text{Processors}_v$.

**Round 4:** Upon receiving list $\text{Processors}_w$ in Round 3 from some coordinator $w$, processor $v$ replaces its list $\text{Processors}_v$ with list $\text{Processors}_w$.

Processor $v$ then removes the first entry in its list $\text{Coordinators}_v$. If there are still some identifiers in $\text{Coordinators}_v$ that are smaller than both $v$ and the identifier of the processor that served as coordinator in Round 3, then $v$ removes these identifiers from $\text{Coordinators}_v$ as well.

Figure 4: Preparatory round and Checkpointing phase of PART-TWO of algorithm EFFORT-PRIORITY; the code for processor $v$.  

45
The rationale for this is to help converge the lists Coordinators, while deleting from them the identifiers of crashed processors. This works because any active processor $y$, that is ahead of $v$ in Round 1, and whose identifier is deleted by $v$ in Round 4 of this phase, will propose itself as coordinator along with $v$ in Round 1 of the next phase, so both identifiers $v$ and $y$ will be restored in the lists Coordinators in Round 2 of the next phase.

After Part-Two concludes, the processors proceed to Part-Three, where algorithm DMY is invoked (given in [23]). In Part-Three the algorithm reverts to using individual tasks instead of chunks of tasks. Processor $v$ cooperates only with the processors whose identifiers are in the list Processors$_v$, at the conclusion of Part-Two. However, instead of using the original processor identifiers, each processor $v$ refers to any active processor $w$ in the list Processors$_v$ by $w$’s rank in the list. The lists Processors are not used by algorithm DMY; these lists only serve to provide new processor identifiers. Finally, the lists of tasks are reverted to their original contents of $t$ tasks. This is sufficient to obtain the desired complexity bounds and simplifies the analysis.

This concludes the specification of algorithm Effort-Priority. Next we discuss its correctness. We first show that the set of processor identifiers in at least one list Coordinators shrinks in each checkpointing phase of Part-Two.

**Lemma 22** The least identifier that occurs in some list Coordinators in Round 1 of a phase is removed from all lists where it is present in Round 4 of the phase.

**Proof:** Consider a phase, and let $v$ be the least identifier occurring in any list Coordinators in the beginning of the phase; thus $v$ is the first entry in any list Coordinators it occurs in. Suppose first that no processor announces itself coordinator in Round 1 of the phase. Then no new entries are added to the lists during the phase and the first entries are removed from each list Coordinators; this eliminates $v$ from any list it occurs in. Now consider the case when some processors announce themselves coordinators in Round 1 of the phase. If $v$ is not among them, then the proposed names of coordinators are greater than $v$, so no entry is added in Round 2 of the phase as new first entry to any list in which $v$ occurs. Therefore all occurrences of $v$ remain first in the respective lists, and so they are removed in Round 4. If, on the other hand, $v$ announces itself coordinator, then $v$ is added in Round 2 as the new first entry to any list Coordinators where $v$ does not occur, only to be removed in Round 4 of the phase. \qed

**Lemma 23** If there are fewer than $p - f_1$ active processors at the preparatory round of Part-Two, then the lists Processors of all active processors become identical at some checkpointing phase in Part-Two, and remain invariant through the remaining checkpointing phases.

**Proof:** We deal only with the processors that are active at the preparatory round of Part-Two. Let us assume that the list Coordinators$_v$ of any such processor $v$ is modified in Round 4 of each phase according to the specification in Figure 4, even after processor $v$ crashes. These modifications do not affect the course of execution, as processors do not send any messages after crashing, but this simplifies the structure and exposition of the argument.

Let $v$ be a processor that stays non-faulty through the end of Part-Two. We claim that there is a phase in which $v$ acts as a successful coordinator. To show this, observe the following three facts:
(1) the set of identifiers occurring in lists \textit{Coordinators} shrinks with each checkpointing phase, by Lemma \ref{lemma:coordinator-shrinkage}.

(2) there are fewer than \( p - f_1 \) identifiers in all lists \textit{Coordinators} when checkpointing phases begin, by the assumption;

(3) there are \( p - f_1 \) checkpointing phases in total, by the code of Part-Two.

It follows, from these three facts combined, that eventually a phase \( P \) occurs such that the identifier of processor \( v \) is removed from its own list \textit{Coordinators}_v in Round 4 of \( P \). If \( v \) is first on its list \textit{Coordinators}_v in Round 4 of \( P \) then \( v \) was also such in Round 1 of \( P \), since no entries are removed from lists \textit{Coordinators} in the first three rounds of a phase. Therefore, processor \( v \) announces itself to be coordinator in Round 1 of phase \( P \). Processor \( v \) is least among all processors that announce themselves coordinators for phase \( P \), as otherwise a smaller identifier would be added to \textit{Coordinators}_v in Round 2 of phase \( P \), and so \( v \) would not be removed from \textit{Coordinators}_v in Round 4 of \( P \). Therefore \( v \) stays as coordinator for the entire phase \( P \).

We have just shown that there exists a phase in which some processor acts as a successful coordinator. Consider the first such a phase, and let \( v \) be the processor that acts as a successful coordinator in this phase. Every active processor \( w \) receives list \textit{Processors}_v from \( v \) in Round 4 of the phase. Now, \( w \) adopts the obtained list as its new list \textit{Processors}_w, per Round 4 specification (see Figure 4). From this round on, all lists \textit{Processors} of active processors stay invariant through the last checkpointing phase, as no element is ever removed from them, per Rounds 3 and 4 specification (see Figure 4).

\textbf{Lemma 24} \textit{For any} \( f_1 \) \textit{and} \( T_1 \), \textit{such that Part-One completes all tasks in time} \( T_1 \) \textit{when at most} \( f_1 < p \) \textit{processors crash, algorithm Effort-Priority that uses} \( f_1 \) \textit{and} \( T_1 \) \textit{solves the Do-All problem against the unbounded adversary.}

\textbf{Proof:} The specification of Round 4 in Figure 4 refers to receiving a message from “some coordinator,” allowing for the possibility of multiple coordinators. We show that there is at most one coordinator in Round 3 of any checkpointing phase in any execution of Part-Two. Suppose that there are two such coordinating processors with identifiers \( v \) and \( w \), and let \( v < w \). These two processors broadcast in Round 1 of the phase. This results in processor \( w \) restoring identifier \( v \) in its list \textit{Coordinators}_w ahead of its own identifier \( w \), thereby losing the status of coordinator for the current phase, per Round 2 of Figure 4.

Next, it must be ensured that the identifiers of processors and tasks are consistent among participating processors at the invocation of algorithm DMY. This algorithm is invoked only when more than \( f_1 \) processors crash while executing Part-One, by the selection of \( T_1 \). Therefore when the checkpointing phases are over, the lists \textit{Processors} are identical among the active processors, by Lemma \ref{lemma:identifier-consistency}. This ensures that each active processor has the same rank in each such sorted list, providing consistent new identifiers to all processors that execute algorithm DMY.

The lists of tasks are restored to their original form of \( t \) entries at each processor just before algorithm DMY is invoked; this provides consistent referencing of the tasks by the processors executing algorithm DMY. Since algorithm DMY does its job in Part-Three, algorithm Effort-Priority solves the Do-All problem as required. \( \square \)
7.2 Specifying algorithm parameters

Next we specify the numeric parameters that are present in the code of algorithm EFFORT-PRIORITY. The idea is to select the parameter $f_1$ and the corresponding $T_1$ so as to make the work of PART-ONE approximately equal to the work of PART-THREE. Let $\rho = 27/2$ and $\Delta_0 = 74$ be the constants used in the specification of the overlay graphs in Section 3. Choose $a$ such that $0 < a < 1$ and

$$a < \frac{1}{1 + 2 \log_\rho \Delta_0}.$$  \hfill (36)

We set $f_1 = p - p^{1-a}$ as the bound on the number of crashes. From this we get $p/(p - f_1) = p^a$. By Lemma 11, the degree of the corresponding overlay graph $L(p)\ell$ satisfies

$$\Delta_1 = \mathcal{O}\left(\left(\frac{p}{p - f_1}\right)^{2 \log_\rho \Delta_0}\right) = \mathcal{O}(p^{2a \log_\rho \Delta_0}).$$  \hfill (37)

We instantiate $T_1$ to be sufficiently large for Lemma 24 to be applicable, but such that

$$T_1 = \mathcal{O}\left(\frac{t + p}{p - f_1} + \Delta_1 \log^2 p\right).$$  \hfill (38)

We next show that it is possible to set $T_1$ so that bound [35] is satisfied. This is ultimately accomplished in Lemma 28.

The first step is computing the time needed to perform tasks when the key extended epochs are productive. To this end, we revisit the proof of Lemma 17 (see Section 6.1) and verify its time aspects. Recall the following definitions given in Section 6.1. The integers $j$ that are indices of main extended epochs $D_j$ make a contiguous interval $J$, which is represented in the form $J = J_1 \cup J_2 \cup J_3 \cup J_5$. Set $J_1$ consists of indices $j$ of stormy extended epochs $D_j$, set $J_2$ consists of these $j \in J$ for which $u_{t(j-1)} \geq 11p^2g(p)$, set $J_3$ consists of the index of the extended epoch $D_j$ such that $u_{m(j)} \geq 11p^2g(p)$ and $u_{l(j)} < 11p^2g(p)$, set $J_4$ consists of these $j \in J \setminus (J_1 \cup J_2 \cup J_3)$ for which the (key) extended epoch $D_j$ is task-rich, and set $J_5$ consists of these $j \in J \setminus (J_1 \cup J_2 \cup J_3)$ for which the (key) extended epoch $D_j$ is task-poor.

Lemma 25 For any known $f < p$, all processors executing the generic algorithm halt after performing $\mathcal{O}\left(\frac{t}{p-f} + \log^2 p\right)$ rounds during the main extended epochs $D_j$, if only all key extended epochs among them are productive.

Proof: We need to sum up the durations of $D_j$, for $j \in J$. We estimate this sum by considering the indices of the extended epochs that fall into the four subsets $J_1$, $J_2$, $J_3$, $J_4$, and $J_5$.

First we consider the set $J_1$. There are $\mathcal{O}(\log p)$ extended epochs with indices in $J_1$, because each of them consists of a single regular stormy epoch, by Lemma 12. So their contribution to the total time is $\mathcal{O}(\log^2 p)$. Next consider extended epochs $D_j$ for $j \in J_2$. The time spent in the extended epochs whose regular epochs $E_i$ satisfy $u_i \geq 11p^2g(p)$ is $\mathcal{O}\left(\frac{t}{p-f}\right)$, by Lemma 10.

The key extended epochs $D_j$ have their indices $j$ in $J_4 \cup J_5$. We assume that each of them is productive. For the extended epochs $D_j$ such that $j \in J_4$, we use the estimate [33], which is derived in the proof of Lemma 17 (see Section 6.1). It follows that the decrease in the number of tasks in two consecutive extended epochs is proportional to the work expended by the core processors, which is $c_j$, and so this takes $\mathcal{O}\left(\frac{t}{p-f}\right)$ time. There are $\mathcal{O}(\log p)$ extended epochs $D_j$ for $j \in J_5$, 48
by Lemma 16, since each such \( D_j \) is productive. Each such extended epoch \( D_j \) for \( j \in J_5 \) consists of a single regular epoch, by Lemma 15. The time contribution of such extended epochs is \( O(\log^2 p) \).

There is at most one extended epoch \( D_j \) such that \( u_m(j) \geq 11p^2g(p) \) and \( u_l(j) < 11p^2g(p) \); if it exists then \( J_3 = \{j\} \). Observe that the duration of this extended epoch is at most the time of all the extended epochs whose indices are in \( J_1 \cup J_2 \cup J_4 \cup J_5 \). This means that is suffices to double the time derived from the extended epochs whose indices are in \( J_1 \cup J_2 \cup J_4 \cup J_5 \) to cover the case of this extended epoch.

The time of the main extended epochs \( D_j \) such that \( j \in J_1 \cup J_2 \cup J_4 \cup J_5 \) is \( O\left(\frac{t}{p-j} + \log^2 p\right) \), if only all key extended epochs among them are productive. It follows that \( O\left(\frac{t}{p-j} + \log^2 p\right) \) is the bound on time of all the main extended epochs with this property. \( \square \)

The next step is computing the time needed to perform the closing work. To this end, we revisit the proof of Lemma 9 (see Section 4.2) to extract time bounds.

**Lemma 26** Suppose an instantiation of the generic algorithm has the property that if \( p \) processors aim to perform \( p \) tasks and up to \( f \) of them may fail, for a known \( f < p \), then this is accomplished with \( O(T(p,f)) \) main rounds, for some function \( T(p,f) \). Then the algorithm terminates after \( O(T(p,f) + \log^2 p) \) closing rounds.

**Proof:** There are \( O(\log p) \) stormy regular epochs, so they contribute \( O(\log^2 p) \) to time. Next we consider only calm epochs. By the definition of \( T(p,f) \), core processors perform \( O(T(p,f)) \) closing rounds while each core processor \( v \) has a nonempty \( \text{Busy}_v \) list. When some processor \( v \), that is core for \( E_k \), obtains an empty list \( \text{Busy}_v \) in \( E_k \), then \( v \) sends copies of this empty list to its neighbors, and halts. All processors that are core for epoch \( E_{k+1} \) halt by the end of \( E_{k+1} \). These two epochs, \( E_k \) and \( E_{k+1} \), contribute \( O(\log p) \) to time. It follows that the core processors halt within \( O(T(p,f) + \log^2 p) \) closing rounds.

It may happen that after all core processors halt, there are still compact processors with nonempty \( \text{Busy} \) lists. In this case we may carve compact graphs from the subgraph of \( G(p,f) \) induced by these nodes. The nodes of any such a graph terminate within time bounds we obtained for the core processors, because the notion of core processor is conceptual only, used for the sake of arguments, while no processor knows whether it is core or not. Let us choose such a compact subgraph of \( G(p,f) \) that stays compact longest; these could be core processors, without loss of generality. The processors in this subgraph halt within \( O(T(p,f) + \log^2 p) \) closing rounds. Afterwards, at most two epochs occur such that by their end no processor considers itself compact, thus all processors halt in these epochs. This again contributes only \( O(\log p) \) additional time. \( \square \)

We revisit the proof of Theorem 2 to examine time performance, for the same instantiation that provides good work performance.

**Lemma 27** For any known \( f < p \), algorithm Deterministic-Permutations can be instantiated so that its work is \( O(t + p^2 \log^2 p) \), with all processors halting in time \( O\left(\frac{t + p}{p-f} + \log^2 p\right) \).

**Proof:** We take an instantiation of the algorithm as provided by Lemma 20. The bound on work performance given in Theorem 2 applies. Next we consider the time performance. The algorithm instantiation has the property that all key extended epochs in any execution against the \( f \)-chain-constrained adversary are productive. This lets us use Lemma 25 that gives \( O\left(\frac{t}{p-f} + \log^2 p\right) \) as
the time spent by such an instantiation of algorithm \textsc{Deterministic-Permutations} during main extended epochs against the $f$-chain-constrained adversary. By Lemma \ref{lem:main-rounds}, it takes $O\left(\frac{t}{p-f} + \log^2 p\right)$ time for this instantiation to complete the main epochs against the general $f$-bounded adversary.

To estimate the number of closing rounds, we apply Lemma \ref{lem:final}. This contribution is obtained by adding $O(\log^2 p)$ to the amount obtained by substituting $p$ for $t$ in the estimate $O\left(\frac{t}{p-f} + \log^2 p\right)$ that we have obtained for the main rounds. These two contributions together make up $O(\frac{p}{p-f} + \log^2 p)$ closing rounds.

The performance bounds of algorithm \textsc{Modified-Det-Perm} can be obtained by resorting to the bounds already derived for algorithm \textsc{Deterministic-Permutations}. To this end, we apply the following scaling. We substitute the number of chunks for the number of tasks and modify the time scale by extending the time of each by the time sufficient to perform one chunk. Observe first that the number of chunks is $O(t/\Delta_1)$ and a chunk is performed in $\Delta_1$ rounds. This is used in the proof of the next lemma.

\textbf{Lemma 28} Algorithm \textsc{Effort-Priority} can be initialized with $T_1$ satisfying Equation (\ref{eq:time-bound}) so that all processors halt by round $T_1$ while executing Part-One provided that the number of crashes in the execution is at most $f_1$.

\textbf{Proof:} We estimate the time by which all tasks are performed when there are $f_1$ crashes. To this end we apply scaling of tasks to chunks. Partitioning the tasks into chunks of size $\Delta_1$ scales down the term $t + p$ in Lemma \ref{lem:main-rounds} to $(t + p)/\Delta_1$. We substitute this term for $t + p$ in the bound on time in Lemma \ref{lem:main-rounds} and simultaneously extend the time bound by a factor of $\Delta_1$ that is due to the time sufficient to complete a chunk, to obtain

$$O\left(\Delta_1\left(\frac{t + p}{\Delta_1(p - f_1)} + \log^2 p\right)\right) = O\left(\frac{t + p}{p - f_1} + \Delta_1 \log^2 p\right)$$

as a bound on the number of rounds by which all processors halt. \hfill $\square$

This concludes the justification that $T_1$ can be initialized such that Lemma \ref{lem:final} is applicable, and estimate (\ref{eq:time-bound}) is satisfied.

\section{The bound on effort}

We proceed with the analysis of effort for algorithm \textsc{Effort-Priority}. The parameters in the code of the algorithm are set as specified in the beginning of Section \ref{sec:effort}.

\textbf{Lemma 29} The effort of Part-One is $O(t + \Delta_1p\log^2 p)$.

\textbf{Proof:} Each processor executing the algorithm sends $\Delta_1$ messages precisely after completing $\Delta_1$ tasks. Thus communication complexity is the same as work complexity, and so it suffices to estimate work. We consider two cases depending on the number of processors staying non-faulty until round $T_1$.

Case 1: the number of such non-faulty processors is at least $p - f_1$. All processors halt by round $T_1$ executing Part-One, by Lemma \ref{lem:final}. Their work is as estimated in Theorem \ref{thm:main} by Lemma \ref{lem:main-rounds}.
suitably adjusted by scaling. Applying the scaling, we obtain that the work during Part-One is

\[ O\left( \Delta_1 \left( \frac{t}{\Delta_1} + p \log^2 p \right) \right) = O(t + \Delta_1 p \log^2 p) \tag{39} \]

Case 2: the number of such non-faulty processors is less than \( p - f_1 \). Since there are few surviving processors, algorithm Modified-Det-Perm used in Part-One may not complete all tasks. Using the same derivation as for bound (39), the effort is \( O(t + \Delta_1 p \log^2 p) \) up to the round when the number of active processors falls below \( p - f_1 \). The remaining rounds of Part-One contribute

\[ O((p - f_1) T_1) = O\left( (p - f_1) \left( \frac{t + p}{p - f_1} + \Delta_1 \log^2 p \right) \right) \]
\[ = O(t + p + \Delta_1 (p - f_1) \log^2 p) \]
\[ = O(t + \Delta_1 p \log^2 p) \]

to the work, by (38) and Lemma 28.

Next we deal with message complexity for Part-Two. We begin by discussing a situation in which a processor announces itself coordinator in a checkpointing phase and next abandons this status within the same phase. This generates \( O(p) \) messages each time for each such processor. In these scenarios, some processors \( v_1, \ldots, v_k \), where \( k > 1 \) and \( v_1 < \ldots < v_k \), propose, in Round 1 of some checkpointing phase \( P \), that each of them intends to become coordinator for the phase. Then \( v_1 \) remains as the successful coordinator for phase \( P \), while \( v_2, \ldots, v_k \) give up the status of coordinator for phase \( P \). Observe that each processor \( v_i \), among \( v_1, \ldots, v_k \), has the identifiers \( v_1, \ldots, v_k \) in its list \( \text{Coordinators}_{v_i} \) just after the preparatory round, because these processors are still non-faulty later in checkpointing phase \( P \). Still, each processor \( v_i \) appears first on its own list \( \text{Coordinators}_{v_i} \) in the first round of phase \( P \). An explanation of what must have occurred prior to phase \( P \) is as follows, where, for simplicity, we assume that no crashes after the preparatory round affect the considered scenario.

Each processor \( v_i \), for \( 1 < i \leq k \), must have deleted \( v_1, \ldots, v_{i-1} \) from list \( \text{Coordinators}_{v_i} \) prior to phase \( P \) to become first in \( \text{Coordinators}_{v_i} \) in Round 1 of phase \( P \), as \( v_j < v_i \) for \( 1 \leq j < i \). Additionally, any processor \( v_j \), for \( j < i \), has had to delete some \( i - j \) entries from its list \( \text{Coordinators}_{v_j} \) (to become first in this list) that \( v_i \) did not need to delete (to become first in its own list \( \text{Coordinators}_{v_i} \)), to make up for the entries \( v_j, \ldots, v_{i-1} \) that \( v_j \) did not need to delete (in order to become first in its list \( \text{Coordinators} \)) as compared to \( v_i \). Specializing this argument to \( i = k \) and \( j = 1 \), we obtain that processor \( v_1 \) deleted \( k - 1 \) processors that \( v_k \) did not delete, prior to phase \( P \). If \( x_1, \ldots, x_{k-1} \) are some \( k-1 \) processors that \( v_1 \) has deleted by the first round of phase \( P \), but \( v_k \) has not, then \( x_i < v_1 \), for \( 1 \leq i \leq k-1 \), as \( v_1 \) has removed \( x_i \) from \( \text{Coordinators}_{v_1} \) prior to announcing itself coordinator. Therefore, these are the discrepancies among their lists \( \text{Coordinators} \) that make processors \( v_1, \ldots, v_k \) broadcast simultaneously in Round 1 of phase \( P \). Next we explain how such disparities could have occurred.

The preparatory messages sent by each \( x_i \), for \( 1 \leq i \leq k - 1 \), were received by \( v_1 \), since \( v_1 \) removes \( x_i \) from \( \text{Coordinators}_{v_1} \), so \( x_i \) is placed in list \( \text{Coordinators}_{v_1} \) at the end of the preparatory round. On the other hand, the preparatory messages sent by each \( x_i \) are not received by \( v_k \). This is because otherwise \( x_i \) would be in the list \( \text{Processors}_{v_k} \) after the preparatory round, and therefore also in list \( \text{Coordinators}_{v_k} \), so that \( v_k \) would need to remove \( x_i \) eventually from \( \text{Coordinators}_{v_k} \) to become first in this list, as \( x_i < v_1 < v_k \), while we know that \( v_k \) does not do
it. It follows that each processor \( x_i \), for \( 1 \leq i \leq k - 1 \), crashes in the preparatory round, as some of the messages that \( x_i \) was to send in that round are not received. We refer to such a scenario of broadcasts in Round 1 of phase \( P \) by saying that processor \( v_1 \) is delayed with respect to \( v_k \) by \( k - 1 \) crashes, where these crashes occur in the preparatory round. We also say that the broadcasts of processors \( v_2, \ldots, v_k \) in Round 1 of checkpointing phase \( P \) are futile broadcasts in this phase. When the identifiers of processors \( v_2, \ldots, v_k \) are added in Round 4 of a phase to each list \( \text{Coordinators} \), in which any of them are absent at that point, then we say that the crashes of processors \( x_1, \ldots, x_{k-1} \) become neutralized thereby.

Now, we are ready to tackle the message complexity of PART-TWO. This complexity is estimated by careful accounting of futile broadcasts caused by delays due to crashes on one hand, and by neutralizing the same number of crashes, on the other.

Let \( p_2 \) denote the number of processors that stay non-faulty and do not halt by round \( T_1 \), that is, they start executing PART-TWO.

**Lemma 30** There are \( \mathcal{O}(p(p-f_1)) \) messages sent during PART-TWO.

**Proof:** It is sufficient to consider only the case when more than \( f_1 \) processors crash during PART-ONE, as otherwise PART-TWO is not invoked at all. We have \( p_2 < p-f_1 \), so that \( p_2 \cdot p = \mathcal{O}(p(p-f_1)) \).

The preparatory round contributes \( \mathcal{O}(p_2 p) \) messages, since at most \( p_2 \) processors broadcast messages, each to at most \( p \) recipients. Next we consider the messages generated in the checkpointing phases. Any processor \( v \) acts as a successful coordinator at most once, since afterwards \( v \) removes its identifier from list \( \text{Coordinators} \) to never announce itself as a coordinator again. A processor that acts as a successful coordinator for a phase broadcasts messages twice, in Round 1 and Round 3 (rounds refer to Figure 4), and collects messages once, in Round 2. Therefore such activity contributes \( \mathcal{O}(p_2 \cdot p_2) = \mathcal{O}(p_2 p) \) messages.

To account for futile broadcasts, we charge these broadcasts to some crashes that occur either in the preparatory round, or in subsequent checkpointing phases. For this accounting to work, it is sufficient to assure that at most one futile broadcast is charged to any one crash. The next part of the proof deals with showing this fact. The argument is broken into two major parts, named Claims 1 and 2, based on whether crashes occur only in the preparatory round or also after it.

We define **configuration** to be a snapshot of the contents of all lists \( \text{Coordinators} \) just after the preparatory round. A configuration is determined by crashes in the preparatory round and the messages that are to be sent by the processors that crash in the preparatory round; each such message may be sent and delivered, or sent and not delivered, or not sent at all. We consider executions of PART-TWO in which no crashes occur, except in the preparatory round. For any configuration \( C \) that may come into existence at the end of the preparatory round as the result of such an execution, we define \( \mathcal{E}(C) \) as the fragment of this execution during the rest of PART-TWO. Since no new crashes occur in PART-TWO, execution \( \mathcal{E}(C) \) is unique with respect to configuration \( C \).

**Claim 1:** For any configuration \( C \), each futile broadcast in \( \mathcal{E}(C) \) can be charged to a unique crash in the preparatory round.

If there are futile broadcasts in \( \mathcal{E}(C) \), then let \( P_C \) be the first round in which simultaneous broadcasts by multiple processors occur in \( \mathcal{E}(C) \). We consider two cases for execution \( \mathcal{E}(C) \): (1) there is a single
coordinator that broadcasts in Round 1 of the previous phase $P_C - 1$, and (2) there is no coordinator in phase $P_C - 1$. The case when $P_C$ is the very first checkpointing phase in $E(C)$, i.e., there is no checkpointing phase $P_C - 1$, falls under case (2) for $E(C)$. Next, we consider these cases in detail.

Case (1): For execution $E(C)$ in this case, let $v$ be the coordinator of phase $P_C - 1$. The processors that broadcast simultaneously in Round 1 of the next phase $P_C$ include all processors whose identifiers are less than $v$, per Round 4, and possibly processors with identifiers greater than $v$. Let $w_1, \ldots, w_\ell$ be those processors with identifiers less than $v$, where $w_1 < \ldots < w_\ell$, and let $v_1, \ldots, v_k$ be those processors with identifiers greater than $v$. Processor $w_1$ acts as coordinator for this phase, and in the following $\ell - 1$ phases processors $w_2, \ldots, w_\ell$ act as coordinators one by one. Following this, all processor identifiers in lists Coordinators are greater than $v$. Processors $v$ and $v_1, \ldots, v_k$, where $v < v_1 < \ldots < v_k$, may not constitute a segment of consecutive identifiers of processors that are still active. If there are some additional $j$ active processors whose identifiers are between $v$ and $v_k$, then this means that $w_1$ is delayed with respect to $v_k$ by $k + \ell + j - 1$ crashes. It is also possible that some processors delete identifiers greater than $v_k$ from their lists Coordinators prior to phase $P_C$. If $m$ is the maximum number of such identifiers deleted from the list of some processor $z$, then $w_1$ is delayed by at least $k + \ell + j + m - 1$ crashes with respect to this processor $z$. We have already used $k + \ell - 1$ crashes to account for futile broadcasts by $w_2, \ldots, w_\ell$ and by $v_1, \ldots, v_k$. The identifiers of $w_2, \ldots, w_\ell$ and $v_1, \ldots, v_k$ are restored in phase $P_C$ to any Coordinators list in which some of them were missing, per Round 2 specification, neutralizing these $k + \ell - 1$ crashes. After this restoration, processor $w_1$ is still delayed by at least $j + m$ crashes with respect to processor $z$. This means that at most $j + m$ futile broadcasts may occur in $E(C)$ after phase $P_C$, and any such broadcast can be charged to a unique crash in the preparatory round.

Case (2): For execution $E(C)$ in this case, there is no coordinator in phase $P_C - 1$. Let processors $v_1, \ldots, v_k$ broadcast together in Round 1 of phase $P_C$. Processor $v_2$ also broadcasts in Round 1 of phase $P_C + 1$. Some other processors $w_1, \ldots, w_\ell$, where $w_1 < \ldots < w_\ell < v_1$, may become first in their lists Coordinators in phase $P_C$, per Round 4 specification. These processors also broadcast in Round 1 of phase $P_C + 1$. If this is the case, then $w_1$ is delayed with respect to crashes of $v_k$ by $k + \ell - 1$. In the next $\ell - 1$ phases of execution $E(C)$, processors $w_2, \ldots, w_\ell$, that perform futile broadcasts in phase $P_C + 1$, act as coordinators. Following this, all identifiers in lists Coordinators are greater than $v_1$. The identifiers of processors $v_1, \ldots, v_k$, where $v_1 < \ldots < v_k$, may not form a segment of consecutive identifiers. If there are some additional $j$ active processors whose identifiers are between $v_1$ and $v_k$, then this means that processor $w_1$ is delayed with respect to $v_k$ by $k + \ell + j - 1$ crashes. It is also possible that some processors delete identifiers greater than $v_k$ from their lists Coordinators prior to phase $P_C$. If $m$ is the maximum number of such identifiers deleted from the list of some processor $z$, then $w_1$ is delayed by at least $k + \ell + j + m - 1$ crashes with respect to this processor $z$. We have already used $k - 1$ crashes to account for the futile broadcasts by $v_2, \ldots, v_k$, and $\ell$ crashes for similar broadcasts by processors $w_2, \ldots, w_\ell$ and $v_2$. The identifiers of processors $v_2, \ldots, v_k$ are restored in all lists Coordinators in phase $P_C$, and the identifiers of processors $w_2, \ldots, w_\ell$ and $v_2$ are restored in all lists Coordinators in phase $P_C + 1$, per Round 2 specification, which neutralizes $k + \ell - 1$ crashes. After this restoration, processor $w_1$ is still delayed by at least $j + m$ crashes with respect to processor $z$. This means that at most $j + m$ futile broadcasts may occur in $E(C)$ after phase $P_C$, and any such broadcast can be charged to a unique crash in the preparatory round. This completes the proof of Claim 1.

Let us associate credits with configuration $C$. We define the number of credits to be equal to $k + \ell + j + m - 1$, where these integers are determined by $E(C)$. We will use credits to account...
for futile broadcasts. In particular, \( j + m \) credits that are still available after phase \( P_C \), when execution \( \mathcal{E}(C) \) falls under Case 1, or after phase \( P_C + 1 \) in Case 2, respectively, are sufficient to account for futile broadcasts after phase \( P_C \) or \( P_C + 1 \), respectively, in execution \( \mathcal{E}(C) \), as at most \( j + m \) of them may possibly occur.

In a general scenario, crashes after the preparatory round may also contribute to multiple simultaneous broadcasts. Crashes during checkpointing phases matter only when they occur in Round 1 of a checkpointing phase, because discrepancies among lists \textit{Coordinators} may be created when some recipients receive messages sent in these rounds and some not. In what follows we restrict our attention only to crashes of some processors \( x \) when these crashes have one of the following properties: (1) processor \( x \) crashes while broadcasting messages in Round 1 of some phase, and (2) some recipient of a message sent by \( x \) receives the message and some does not. Now, consider arbitrary executions of \textit{Part-Two}.

\textbf{Claim 2:} The following invariant is maintained through all checkpointing phases \( P \):

the number of credits available in the beginning of \( P \) is sufficiently large to account for all futile broadcasts in \( P \) and thereafter if crashes occur only prior to phase \( P \).

The proof of the invariant is by induction on the phase number \( P \). Let an execution of \textit{Part-Two} begin with configuration \( C \). The basis of the induction deals with the case when \( P = 1 \), which means that no crashes occur after the preparatory round; this is taken care of by Claim 1. Next we show the inductive step. Let \( P \) be a phase number, where \( P > 1 \), and assume that the invariant is established by phase \( P - 1 \). If there is no crash in \( P \), then the invariant extends automatically, so suppose at least one crash occurs in \( P \). There are two cases, depending on whether there is a single or multiple broadcasts in Round 1 of phase \( P \).

\textbf{Case 1, single broadcast:} There is one coordinator \( v \) that broadcasts in Round 1 of \( P \), and \( v \) crashes in this round so that some processors receive a message from \( v \) and some do not. There are two sub-cases: (1.a) a successful broadcast by \( v \) results in some processors removing at least two entries from their lists \textit{Coordinators} in Round 4, and (1.b) this does not happen.

The latter sub-case (1.b) is such that there is no difference whether \( v \) crashes or not, as every processor removes just the first entry from its list \textit{Coordinators}, so the inductive assumption is applicable directly. It remains to consider the former sub-case (1.a) that may result in multiple simultaneous broadcasts in phase \( P + 1 \).

Suppose that if \( v \) does not crash, then \( w_1, \ldots, w_\ell \) and \( v_1, \ldots, v_k \) broadcast simultaneously in Round 1 of phase \( P + 1 \), where the identifiers \( w_1, \ldots, w_\ell \) are less than \( v \), and the identifiers \( v_1, \ldots, v_k \) are greater than \( v \). (This is a situation similar to the one we considered before for executions of the form \( \mathcal{E}(C) \), when there is a single coordinator in phase \( P_C - 1 \) prior to phase \( P_C \) with multiple broadcasts.) By the inductive assumption, there are enough credits to pay for the futile broadcasts among those by the processors \( w_1, \ldots, w_\ell \) and \( v_1, \ldots, v_k \). Each of the processors \( v_1, \ldots, v_k \) removes just the first entry from its respective list \textit{Coordinators}, which is not the result of the crash by \( v \), so we account for all their futile broadcasts using \( k \) of the available credits. The processors \( w_1, \ldots, w_\ell \) are those that remove from \textit{Coordinators} some identifiers that are less than both \( v \) and their own name, per Round 4 specification. A crash of \( v \) may result in some among \( w_1, \ldots, w_\ell \) not receiving the message of \( v \) in phase \( P \), and so they do not shorten their lists \textit{Coordinators}, and do not broadcast at all in phase \( P + 1 \). But if this occurs, then we will use correspondingly fewer credits to account for the broadcasts of processors among \( w_1, \ldots, w_\ell \) in the next phase \( P + 1 \), so
the accounting breaks even.

Case 2, multiple broadcasts: Here some processors perform futile broadcasts in phase $P$ and some of those that broadcast crash in the process. (This is a situation similar to the one we considered before for executions of the form $E(C)$, when there were multiple broadcasts by would-be coordinators in phase $P_C$, while there was no coordinator in the immediately preceding phase $P_{C-1}$.) Suppose that $v_1, \ldots, v_k$ broadcast together in Round 1 of phase $P$, and some of them crash in the process. If the crashes did not occur, then processor $v_2$ would broadcast in Round 1 of phase $P+1$, and some other processors $w_1, \ldots, w_\ell$ might become first in their lists $\text{Coordinators}$ in phase $P$, per Round 4 specification, and they will broadcast in Round 1 of phase $P+1$.

We consider how the crashes of some of the processors $v_1, \ldots, v_k$ affect the accounting, in which multiple broadcasts are charged to previous crashes, by relating to the situation without these crashes. By the inductive assumption, there are enough credits to account for the futile broadcasts by any processors $w_1, \ldots, w_\ell$ and $v_1, \ldots, v_k$.

First, consider futile broadcasts by processors $w_1, \ldots, w_\ell$. It could be the effect of the crashes that some among $w_1, \ldots, w_\ell$ do not broadcast in phase $P+1$; but if this occurs, then we use correspondingly fewer credits to account for the broadcasts of processors among $w_1, \ldots, w_\ell$ that actually take place in $P+1$, saving them for possible future futile broadcasts.

Next, consider futile broadcasts by processors $v_1, \ldots, v_k$. When a message from processor $v_i$ is received in Round 1 of a phase, for $1 \leq i \leq k$, then identifier $v_i$ is restored in any list $\text{Coordinators}$ from which it was deleted before. Such a restoration neutralizes the crash that contributes to the discrepancies among the lists $\text{Coordinators}$ that in turn results in delays that caused $v_i$ to broadcast along with other processors. A crash by such $v_i$ may prevent this neutralization to occur, so that $v_i$ appears in some lists $\text{Coordinators}$ while it does not in others. The accounting is based on the principle that when we associate a credit with a futile broadcast, then the underlying crash is simultaneously neutralized by restoring the names of all such broadcasters, per Round 2 specification. It follows that the inductive assumption is not immediately applicable in this situation. Instead, we charge the futile broadcast by $v_i$ to its very own crash. Thereby we save one credit available to pay for the futile broadcast by $v_i$ if it does not crash. Discrepancies among the lists $\text{Coordinators}$ created by the presence or absence of $v_i$ in them may still occur, which may result in one more futile broadcast after phase $P+1$, but one credit was saved, which means that the accounting breaks even. This concludes the proof of Claim 2.

To complete accounting for futile broadcasts, consider the very last phase $P$ of Part-Two. By Claim 2, there are enough credits through phase $P$ if no crashes occur in $P$. But crashes in $P$ may only diminish the number of messages sent in $P$, while they have no effect on future phases, as there are no phases after $P$. This implies that each futile broadcast proclaiming intent to become a coordinator can be charged to a unique processor that crashed during Part-Two. As fewer than $p_2$ processors may crash in Part-Two, these broadcasts contribute $O(p_2 \cdot p_2) = O(p_2 p)$ messages.

Summing up all messages sent during Part-Two as considered above, we obtain that in total there are $O(p_2 p) = O(p(p - f_1))$ messages.

Next we calculate the effort complexity of algorithm Effort-Priority.

**Lemma 31** The effort of algorithm Effort-Priority is $O(t + p^{2-a})$. 

□
Proof: The overall effort of the algorithm is composed of the contributions by Part-One, Part-Two, and Part-Three.

The effort of Part-One, by Lemma 29 and Equation (37), is

\[ O(t + \Delta_1 p \log^2 p) = O\left(t + p^{1+2a \log p} \Delta_0 \log^2 p\right). \tag{40} \]

To take care of the factor \( \log^2 p \) in (40), we use the identity \( \log^2 p = p^{2\log p/\log p}. \) Substituting this into the right-hand side of (40), we obtain that the effort is

\[ O(t + p^{a^2 \log p} \Delta_0 \log^2 p) = O(t + p^{1+2a \log p} \Delta_0 + \frac{2 \log \log p}{\log p}). \tag{41} \]

By the property \( 2 \log p/\log p = o(1) \) and by the specification of the parameter \( a \) in (36), making \( 2a \log p \Delta_0 < 1 - a \), the following inequality holds for sufficiently large \( p \):

\[ a \log p \Delta_0 + \frac{2 \log \log p}{\log p} < 1 - a. \]

This shows that the effort from bound (41), and so of Part-One, is \( O(t + p^{2-a}) \).

Next we consider Part-Two and Part-Three. Let \( p_2 \) be the number of processors that begin Part-Two. There are two cases depending on the magnitude of \( p_2 \).

Case 1 is for \( p_2 \geq p - f_1 \). This means that all chunks of tasks are performed, and the non-faulty processors halt in the execution of Modified-Det-Perm, by Lemma 28. Therefore there is no contribution to the effort by both Part-Two and Part-Three in this case.

Case 2 is for \( p_2 < p - f_1 \). Now algorithms Part-Two and Part-Three are executed.

We calculate work complexity first. Part-Two contributes

\[ p_2 (p - f_1) < (p - f_1)^2 = O(p^{2-2a}) = O(t + p^{2-a}) \]

to work, as \( p - f_1 = p^{1-a} \). Let \( f_2 \) be the number of crashes that occur during Part-Two and Part-Three. By Fact 3 along with \( p - f_1 = p^{1-a} \), we obtain that the work accrued during Part-Three is

\[ O(t + p_2(p_2 + 1)) = O(t + (p - f_1)^2) = O(t + p^{2-2a}) = O(t + p^{2-a}). \]

Next we calculate the message complexity of Part-Two and Part-Three. The contribution of Part-Two is \( O(p(p - f_2)) \) by Lemma 30. From Fact 3, the message complexity of Part-Three is

\[ O((f_2 + 1)p_2) = O(p_2(p_2 + 1)). \]

Summing up the two contributions we obtain message complexity that is

\[ O(p(p - f_1)) + O(p_2(p_2 + 1)) = O(p(p - f_1)) = O(p^{2-a}). \]

Therefore the effort of Part-Two and Part-Three together is \( O(t + p^{2-a}) \).

Finally, we give the main result for algorithm Effort-Priority.

Theorem 3 Algorithm Effort-Priority can be instantiated, for any \( p \) processors and \( t \) tasks, to be a deterministic nonconstructive solution for the Do-All problem, for any unknown \( f < p \), with effort \( O(t + p^{1.77}) \).
Proof: We initialize algorithm Effort-Priority with the numeric parameters as specified in the beginning of Section 7.2 and such that Lemma 24 is applicable. The work and communication of the algorithm against the unbounded adversary is $O(t + p^{2-a})$, by Lemma 31. The parameter $a$ was selected in (36) so as to satisfy the inequality $a < (1 + 2 \log \rho \Delta_0)^{-1}$. We can directly verify that $0.23 < (1 + 2 \log \rho \Delta_0)^{-1}$, for $\rho = 27/2$ and $\Delta_0 = 74$. Hence the parameter $a$ can be set equal to 0.23, yielding 1.77 as the numeric value of the exponent. □

8 Conclusion

We give a constructive algorithm for the Do-All problem that achieves $O(n^{3/2} \text{ polylog } n)$ work, for $n = \max\{p, t\}$, against the unbounded adversary. No deterministic algorithms for the Do-All problem with work complexity $o(n^2)$, for $n = \max\{p, t\}$, were known to exist prior to this research, even against linearly-bounded adversaries. Regarding nonconstructive solutions, we present an algorithm with work $O(t + p \text{ polylog } p)$, against arbitrary but known adversaries.

Developing Do-All solutions that are efficient in terms of effort, defined as the work plus communication complexities, was among the ultimate goals of this work. We have shown that Do-All can be solved with effort $O(t + p^{1.77})$ against the unbounded adversary. This was the first known algorithm with both work and communication sub-quadratic in $p$ in the worst case, at the time it was announced in [10]. Recently Georgiou et al. [31] developed an algorithm with effort $O(t + p^{1+\varepsilon})$, for any constant $\varepsilon > 0$ being part of the algorithm code. On the other hand, the best known general lower bound on work is $\Omega(t + p \log p/\log \log p)$. (The corresponding lower bound is slightly higher for shared-memory Do-All, being $\Omega(t + p \log p)$. Thus a gap remains between the lower and upper bounds.

Our nonconstructive algorithm, optimized for work against arbitrary, but known, adversaries attains effort $O(t + p \text{ polylog } p)$ against arbitrary, but known, linearly-bounded adversaries. It is an open problem whether there exists a constructive Do-All solution with effort $O(t + p \text{ polylog } p)$ against the unbounded adversary.

We envision directions of future work on problems related to Do-All depending on whether static or dynamic variants of the problem are to be investigated. Regarding static variants, work may encompass aspects of performing tasks related to the possibility of partial information that each processor has about a set of tasks. Recently, Drucker et al. [25] studied the communication complexity of a distributed task allocation problem in which each processor receives as input a subset of the set of all tasks and the goal is to assign each task to a unique processor. Regarding dynamic variants, research effort may pursue adversarial models of task generation and competitive performance. Recently, Alistarh et al. [1] proposed a formalization of a dynamic version of Do-All in asynchronous shared-memory systems and gave a solution that is work optimal within polylogarithmic factors. Georgiou and Kowalski [30] proposed a competitive-performance framework to study dynamic task arrivals into a message-passing system with processors prone to crashes and restarts.
References

[1] D. Alistarh, J. Aspnes, M. A. Bender, R. Gelashvili, and S. Gilbert. Dynamic task allocation in asynchronous shared memory. In Proceedings of the 25th ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 416–435, 2014.

[2] D. Alistarh, M. A. Bender, S. Gilbert, and R. Guerraoui. How to allocate tasks asynchronously. In Proceedings of the 53rd IEEE Symposium on Foundations of Computer Science (FOCS), pages 331–340, 2012.

[3] N. Alon and F. R. K. Chung. Explicit construction of linear sized tolerant networks. Discrete Mathematics, 72(1-3):15–19, 1988.

[4] K. P. Birman. Guide to Reliable Distributed Systems. Springer, 2012.

[5] K. P. Birman and T. A. Joseph. Reliable communication in the presence of failures. ACM Transactions on Computer Systems, 5(1):47–76, 1987.

[6] M. F. Bridgland and R. J. Watro. Fault-tolerant decision making in totally asynchronous distributed systems (preliminary version). In Proceedings of the 6th ACM Symposium on Principles of Distributed Computing (PODC), pages 52–63, 1987.

[7] C. Cachin, R. Guerraoui, and L. Rodrigues. Reliable and Secure Distributed Programming. Springer, 2nd edition, 2011.

[8] K. Censor-Hillel. Multi-sided shared coins and randomized set-agreement. In Proceedings of the 22nd ACM Symposium on Parallelism in Algorithms and Architectures (SPAA), pages 60–68, 2010.

[9] B. S. Chlebus, R. De Prisco, and A. A. Shvartsman. Performing tasks on synchronous restartable message-passing processors. Distributed Computing, 14(1):49–64, 2001.

[10] B. S. Chlebus, L. Gasieniec, D. R. Kowalski, and A. A. Shvartsman. Bounding work and communication in robust cooperative computation. In Proceedings of the 16th International Symposium on Distributed Computing (DISC), Lecture Notes in Computer Science, vol. 2508, pages 295–310. Springer, 2002.

[11] B. S. Chlebus, L. Gasieniec, D. R. Kowalski, and A. A. Shvartsman. A robust randomized algorithm to perform independent tasks. Journal of Discrete Algorithms, 6(4):651–665, 2008.

[12] B. S. Chlebus and D. R. Kowalski. Randomization helps to perform independent tasks reliably. Random Structures and Algorithms, 24(1):11–41, 2004.

[13] B. S. Chlebus and D. R. Kowalski. Cooperative asynchronous update of shared memory. In Proceedings of the 37th ACM Symposium on Theory of Computing (STOC), pages 733–739, 2005.

[14] B. S. Chlebus and D. R. Kowalski. Robust gossiping with an application to consensus. Journal of Computer System and Sciences, 72(8):1262–1281, 2006.
[15] B. S. Chlebus and D. R. Kowalski. Time and communication efficient consensus for crash failures. In *Proceedings of the 20th International Symposium on Distributed Computing (DISC)*, Lecture Notes in Computer Science, vol. 4167, pages 314–328. Springer, 2006.

[16] B. S. Chlebus, D. R. Kowalski, and A. Lingas. Performing work in broadcast networks. *Distributed Computing*, 18(6):435–451, 2006.

[17] B. S. Chlebus, D. R. Kowalski, and A. A. Shvartsman. Collective asynchronous reading with polylogarithmic worst-case overhead. In *Proceedings of the 36th ACM Symposium on Theory of Computing (STOC)*, pages 321–330, 2004.

[18] F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[19] A. E. F. Clementi, A. Monti, and R. Silvestri. Optimal F-reliable protocols for the Do-All problem on single-hop wireless networks. In *Proceedings of the 13th International Symposium on Algorithms and Computation (ISAAC)*, Lecture Notes in Computer Science, vol. 2518, pages 320–331. Springer, 2002.

[20] G. Davidoff, P. Sarnak, and A. Valette. *Elementary Number Theory, Group Theory, and Ramanujan Graphs*. Cambridge University Press, 2003.

[21] S. Davtyan, K. Konwar, A. Russell, and A. Shvartsman. Dealing with undependable workers in decentralized network supercomputing. In *Proceedings of the 14th International Conference on Distributed Computing and Networking (ICDCN)*, Lecture Notes in Computer Science, vol. 7730, pages 27–41. Springer, 2013.

[22] S. Davtyan, K. Konwar, and A. Shvartsman. Robust network supercomputing without centralized control. In *Proceedings of the 15th International Conference on Principles of Distributed Systems (OPODIS)*, Lecture Notes in Computer Science, vol. 7109, pages 435–450. Springer, 2011.

[23] R. De Prisco, A. J. Mayer, and M. Yung. Time-optimal message-efficient work performance in the presence of faults. In *Proceedings of the 13th ACM Symposium on Principles of Distributed Computing (PODC)*, pages 161–172, 1994.

[24] K. Diks and A. Pelc. Optimal adaptive broadcasting with a bounded fraction of faulty nodes. *Algorithmica*, 28(1):37–50, 2000.

[25] A. Drucker, F. Kuhn, and R. Oshman. The communication complexity of distributed task allocation. In *Proceedings of the 31st ACM Symposium on Principles of Distributed Computing (PODC)*, pages 67–76, 2012.

[26] C. Dwork, J. Y. Halpern, and O. Waarts. Performing work efficiently in the presence of faults. *SIAM Journal on Computing*, 27(5):1457–1491, 1998.

[27] C. Dwork, D. Peleg, N. Pippenger, and E. Upfal. Fault tolerance in networks of bounded degree. *SIAM Journal on Computing*, 17(5):975–988, 1988.

[28] A. Fernández, C. Georgiou, A. Russell, and A. A. Shvartsman. The Do-All problem with Byzantine processor failures. *Theoretical Computer Science*, 333(3):433–454, 2005.
[29] Z. Galil, A. J. Mayer, and M. Yung. Resolving message complexity of Byzantine agreement and beyond. In Proceedings of the 36th IEEE Symposium on Foundations of Computer Science, pages 724–733, 1995.

[30] C. Georgiou and D. R. Kowalski. Performing dynamically injected tasks on processes prone to crashes and restarts. In Proceedings of the 25th International Symposium on Distributed Computing (DISC), Lecture Notes in Computer Science, vol. 6950, pages 165–180. Springer, 2011.

[31] C. Georgiou, D. R. Kowalski, and A. A. Shvartsman. Efficient gossip and robust distributed computation. Theoretical Computer Science, 347(1-2):130–166, 2005.

[32] C. Georgiou, A. Russell, and A. A. Shvartsman. The complexity of synchronous iterative Do-All with crashes. Distributed Computing, 17(1):47–63, 2004.

[33] C. Georgiou, A. Russell, and A. A. Shvartsman. Work-competitive scheduling for cooperative computing with dynamic groups. SIAM Journal on Computing, 34(4):848–862, 2005.

[34] C. Georgiou and A. A. Shvartsman. Do-All Computing in Distributed Systems. Springer Verlag, 2008.

[35] A. V. Goldberg, B. M. Maggs, and S. A. Plotkin. A parallel algorithm for reconfiguring a multibutterfly network with faulty switches. IEEE Transactions on Computers, 43(3):321–326, 1994.

[36] P. C. Kanellakis and A. A. Shvartsman. Efficient parallel algorithms can be made robust. Distributed Computing, 5(4):201–217, 1992.

[37] S. Kentros, C. Kari, and A. Kiayias. The strong at-most-once problem. In Proceedings of the 26th International Symposium on Distributed Computing (DISC), Lecture Notes in Computer Science, vol. 7611, pages 386–400. Springer, 2012.

[38] S. Kentros and A. Kiayias. Solving the at-most-once problem with nearly optimal effectiveness. Theoretical Computer Science, 496:69–88, 2013.

[39] S. Kentros, A. Kiayias, N. C. Nicolaou, and A. A. Shvartsman. At-most-once semantics in asynchronous shared memory. In Proceedings of the 23rd International Symposium on Distributed Computing (DISC), Lecture Notes in Computer Science, vol. 5805, pages 258–273. Springer, 2009.

[40] S. C. Kontogiannis, G. E. Pantziou, P. G. Spirakis, and M. Yung. Robust parallel computations through randomization. Theory of Computing Systems, 33(5/6):427–464, 2000.

[41] D. R. Kowalski, P. M. Musial, and A. A. Shvartsman. Explicit combinatorial structures for cooperative distributed algorithms. In Proceedings of the 25th International Conference on Distributed Computing Systems (ICDCS), pages 49–58, 2005.

[42] D. R. Kowalski and A. A. Shvartsman. Performing work with asynchronous processors: Message-delay-sensitive bounds. Information and Computation, 203(2):181–210, 2005.

[43] L. Lamport, R. E. Shostak, and M. C. Pease. The Byzantine generals problem. ACM Transactions on Programming Languages and Systems, 4(3):382–401, 1982.
[44] A. Lubotzky, R. Phillips, and P. Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988.

[45] G. A. Margulis. Explicit constructions of concentrators. *Problems of Information Transmission*, 9(4):325–332, 1973.

[46] M. Mitzenmacher and E. Upfal. *Probability and Computing*. Cambridge University Press, 2005.

[47] N. Pippenger. Sorting and selecting in rounds. *SIAM Journal on Computing*, 16(6):1032–1038, 1987.

[48] M. E. Saks, N. Shavit, and H. Woll. Optimal time randomized consensus - making resilient algorithms fast in practice. In *Proceedings of the 2nd ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 351–362, 1991.

[49] R. M. Tanner. Explicit concentrators from generalized N-gons. *SIAM Journal on Algebraic and Discrete Methods*, 5(3):287–293, 1984.

[50] E. Upfal. Tolerating a linear number of faults in networks of bounded degree. *Information and Computation*, 115(2):312–320, 1994.